Boletín de la Sociedad Española de Matemática Aplicada SëMA

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

Dirección Editorial: Boletín de SëMA. 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:

Fieles a nuestra cita trimestral, nos ponemos de nuevo en contacto con todos vosotros a través del presente boletín, que recoge cinco interesantes artículos de investigación, otro de Educación Matemática y uno más en la sección de Matemáticas e Industria, además de las secciones habituales de Resúmenes de Libros, Noticias, etc.

Como todos sabéis, en el boletín anterior se produjo un lamentable error tipográfico, que no pudimos detectar a tiempo, en el artículo *Codificación de información mediante códigos de barras*, escrito por Luis Hernández y Ángel Martín del Rey, a quienes les pedimos disculpas. Con el fin de subsanar en la medida de lo posible esta anomalía, hemos decidido reproducir íntegramente su artículo aunque ello suponga un aumento de volumen.

Se incluye en este boletín el artículo *EDPs de difusión y transporte óptimo de masa* de José Antonio Carrillo de la Plata, ganador del VI Premio S$\vec{e}$MA al Joven Investigador. J.A. Carrillo, Profesor Investigador de la Institució Catalana de Recerca i Estudis Avançats y Profesor Titular (en excedencia voluntaria) de la Universidad de Granada, desarrolla su investigación, fundamentalmente, en los modelos de difusión no lineal abarcando tanto aspectos cualitativos como numéricos. El trabajo que se recoge aquí es una buena muestra de esto y del excelente nivel de su investigación. Nuestra enhorabuena y muchas gracias por su contribución a situar la investigación matemática española en el alto nivel que ocupa en la actualidad.

En la sección de Noticias incluimos una semblanza en memoria del Profesor Miguel de Guzmán, recientemente fallecido, escrita por el que fue su discípulo y amigo, Ireneo Peral. Desde aquí nos unimos al sentimiento de pesar de toda la comunidad matemática española y deseamos transmitir nuestro sentido pésame a su familia, discípulos y amigos.

También nos hacemos eco de un hecho muy relevante como es la aparición, por primera vez, de las Matemáticas como una de las áreas prioritarias del Programa Nacional de Ciencia y Tecnología. Además del texto íntegro del Programa Nacional de Matemáticas, se incluyen a modo de prólogos sendos escritos firmados, por un lado, por Alfonso Beltrán García-Echániz (Subdirector General de Planificación del Ministerio de Ciencia y Tecnología en el momento de la elaboración de las ponencias que dieron lugar al Plan Nacional de Investigación 2004-2007) y, por otro lado, por José Manuel Fernández de Labastida y Enrique Zuazua (Presidente y Secretario de la ponencia del Programa Nacional de Matemáticas).

Un cordial saludo,

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

Bol. Soc. Esp. Mat. Apl. nº 28(2004), 7–50

Modelado y simulación numérica de la dinámica y termomecánica de grandes masas de hielo *

N. Calvo[1], J. Durany[1] y C. Vázquez[2]

[1] Departamento de Matemática Aplicada II. Universidad de Vigo.
[2]Departmento de Matemáticas. Universidade da Coruña.

nati@dma.uvigo.es, durany@dma.uvigo.es, carlosv@udc.es

Resumen

En este trabajo planteamos un modelo matemático complejo de tipo hielo poco profundo que gobierna los distintos procesos hidrodinámicos y termomecánicos acoplados que tienen lugar en grandes masas de hielo como, por ejemplo, la Antártida. Además, se proponen técnicas numéricas adecuadas para la resolución de los distintos tipos de submodelos (térmicos, de determinación del espesor, de velocidades, de magnitudes basales, etc.) que rigen procesos específicos. Finalmente, se presentan algunos resultados de simulación numérica sobre ejemplos con datos reales para ilustrar tanto la validez de los modelos como el buen funcionamiento de los algoritmos numéricos implementados.

Palabras clave: *Modelos de masas de hielo, Fronteras libres, Problemas no lineales, Elementos finitos, Método de las características, Métodos de dualidad.*
Clasificación por materias AMS: *35K60, 65K20, 65N30.*

1 Introducción

En una fecha relativamente reciente como el 10 de mayo de 2002, un satélite de órbita polar detectaba la presencia de una nueva masa de hielo gigantesca desprendida de la Antártida, en concreto de la región del Mar de Ross (situado al sur de Nueva Zelanda). Su tamaño, de aproximadamente 200 km de longitud, podría suponer un serio peligro para la navegación en la zona. Se cree que el hielo que lo forma se ha estado deslizando lentamente sobre la plataforma antártica durante los últimos 30 años. Los repetidos desprendimientos de enormes masas de hielo en la Antártida (en marzo de 2000 fue noticia un caso parecido al citado) podrían tener relación con los fenómenos de calentamiento global y cambio

*Investigación parcialmente financiada por los Proyectos de Investigación M.C.Y.T. (BFM2001-3261-C02) y Xunta de Galicia (SXID-2002-PR405A y PGIDIT02PXIC10503PN)
Fecha de recepción: 23 de noviembre de 2003

climático.

En efecto, las investigaciones realizadas en los últimos años constatan el gran interés en la respuesta que el casquete polar Antártico puede proporcionar al anticipado calentamiento de la atmósfera terrestre producido, entre otras cosas, por las emisiones de dióxido de carbono. Aunque las predicciones de dicho calentamiento térmico no puedan todavía confirmarse con seguridad por mediciones metereológicas, parece existir un consenso entre los climatólogos de que la temperatura en la segunda mitad del siglo XXI será de dos a cuatro grados centígrados superior a los niveles de la época preindustrial. Si tales especulaciones se confirman y se inicia una fusión de las capas de hielo polares las consecuencias podrían ser desastrosas para todas las zonas costeras del planeta. Como ejemplo basta decir que una disminución de un uno por cien en el volumen de hielo de la Antártida produciría un aumento del nivel del mar de más de setenta centímetros. Predicciones de este tipo ilustran la relevancia que pueden tener para una región costera los estudios de fenómenos que tienen lugar a miles de kilómetros de distancia. Es un hecho asumido por los investigadores el carácter global que tiene el cambio climático que se está produciendo en nuestro planeta. Por otro lado, la interacción de las grandes masas de hielo con la atmósfera y con los océanos y el hecho de que almacenan más del 60 % de los recursos de agua dulce de la Tierra motiva la necesidad de un mejor conocimiento de su comportamiento básico.

Con este objetivo, los recientes avances en la tecnología de los supercomputadores hacen posible redirigir los estudios de algunas cuestiones hacia los modelos matemáticos y numéricos del flujo de hielo. Al mismo tiempo que se han obtenido nuevos e importantes resultados en el campo de la geoquímica y climatología, en los últimos años se ha producido un considerable aumento del conocimiento de datos y condiciones de contorno que determinan la configuración presente de las grandes masas de hielo y de su fluir. Este material parece suficiente para que la investigación actual de los modelos físico-matemáticos de flujos de hielo en la Antártida o en casquetes polares, propuestos de manera global y a gran escala, sea además de necesaria también posible, a pesar de las muchas dudas que existen todavía en la comunidad científica sobre algunos aspectos de la dinámica del hielo.

En este marco de los modelos matemáticos de los casquetes polares, de su análisis y resolución numérica, se centra el contenido de este trabajo. Una de sus principales motivaciones se concreta en el hecho de que para la formulación de las ecuaciones matemáticas que rigen la dinámica del hielo, a gran escala, es necesario tener en cuenta que el dominio del flujo no está definido “a priori” y forma parte de la solución de los problemas. Otra motivación importante radica en la necesidad de considerar en los modelos globales la coexistencia de hielo y agua en zonas de fusión o solidificación. Ambas motivaciones desembocan en ejemplos típicos de problemas de frontera libre o móvil.

La formulación matemática de modelos globales que establecen el comportamiento mecánico y termodinámico de grandes masas de hielo es complicado, ya que, como se ha mencionado anteriormente, todavía existen ciertas dudas sobre algunas condiciones en el flujo de hielo, tanto en el dominio como en el contorno. Todo ello se pone de manifiesto al observar las distintas simplificaciones y desacoplamientos de los fenómenos físicos que se han utilizado desde la aparición de los primeros modelos matemáticos globales a finales de los años cincuenta. Modelos termodinámicos, modelos isotérmicos de la dinámica del hielo, modelos termomecánicos, modelos específicos para la Antártida, son claros ejemplos de la dificultad que estos problemas físicos entrañan. Nosotros remitimos al lector interesado al trabajo de Huybrechts [17], en donde se referencian las particularidades de los diferentes modelos matemáticos que se han manejado en los últimos cuarenta años y al libro de Hutter [20] donde se recogen los avances conseguidos por Nye, Glen, Lliboutry, Weertman y Fowler, entre otros. En consecuencia, dirigimos nuestra atención a ciertos modelos ya simplificados que marcan la pauta de este trabajo.

2 Modelos matemáticos del comportamiento de una gran masa de hielo

En principio, los glaciares y los casquetes polares son objetos físicos similares, pero algunas particularidades de los mismos, fundamentalmente su tamaño, obligan a realizar una distinción en su estudio. Así, los glaciares tienen una longitud entre diez y cincuenta kilómetros, una anchura de centenares de metros y un espesor de decenas o centenas de metros y se mueven principalmente en una dirección a una velocidad típica de cien metros/año, debido a las fuerzas de gravedad y a la inclinación del lecho en el que reposan. Sin embargo, los casquetes polares son grandes masas de hielo que pueden alcanzar más de mil kilómetros de largo y ancho y más de dos mil metros de espesor, descansan en lechos horizontales y tienen cumbres desde donde fluye el hielo en todas las direcciones.

En una primera aproximación, la idea intuitiva de un glaciar es la de un río de hielo drenado desde las zonas en las que se acumula la nieve. Así, comienzan a formarse cuando la nieve se dispone en estratos, y, con su peso y el paso de los años, la más profunda se hace más compacta y dura y acaba convirtiéndose en hielo. Aunque el hielo es sólido, está formado por capas internas que se rompen al estar sometidas a presiones. De este modo, la gran masa de hielo se comporta como un material de una viscosidad elevada, del orden de 10^{12} Pa s, es decir, 10^{15} veces la viscosidad del agua. Como consecuencia de sus dimensiones (profundidad de cientos de metros, anchura de kilómetros y longitud de decenas de kilómetros), al alcanzar cierto espesor, el peso del hielo provoca la deformación y el movimiento del glaciar. No obstante, su elevada viscosidad provoca que el movimiento sea lento, del orden de 10 a 100 metros por año. En cuanto al emplazamiento actual de los glaciares, existen ejemplos en Alaska, las montañas Rocosas, los Alpes, Spitsbergen, etc.

Como se ha mencionado anteriormente, los glaciares ocupan actualmente el 10 por cien de la tierra y el 12 por cien de los océanos y la mayor parte están localizados en las regiones polares como la Antártida y Groenlandia, aunque en casi todos los continentes hay algunos, y su existencia requiere ciertas condiciones climáticas, como por ejemplo que haya grandes nevadas en invierno y bajas temperaturas en verano.

La idea intuitiva de un casquete polar es la de una gran gota de hielo. Sin embargo, mientras que la configuración estacionaria de la gota de agua es consecuencia de la tensión superficial, en el caso del casquete polar el equilibrio se consigue entre la acumulación del hielo en la parte alta del mismo y la ablación en los márgenes o parte más baja. El casquete polar surge en grandes zonas con clima polar (un continente o parte del mismo). Cuando en dicha zona la nieve se acumula y se comprime, fluye lentamente igual que una gota de agua en una mesa bajo la acción de la gravedad. Los casquetes polares actuales más representativos son la Antártida y Groenlandia. En la última glaciación (hace diez mil años) fueron famosos Laurentide y Fennoscadian.

En esta sección planteamos las ecuaciones matemáticas que constituyen el modelo que gobierna el comportamiento de una gran masa de hielo como la Antártida, por ejemplo. Para ello tenemos en cuenta que en los casquetes polares el flujo es casi bidimensional. Así, nos restringiremos, por tanto, a masas de hielo cuyos perfiles son los mismos para las diferentes secciones longitudinales (Hutter [20]).

2.1 Modelo inicial

Para plantear el modelo inicial, basado en la mecánica de los medios continuos, introducimos las coordenadas reales (X, Z), que toman valores en un dominio asociado a una sección longitudinal del casquete polar. Suponemos que la frontera superior está definida en el plano XZ mediante la función $Z = \eta^*(t^*, X)$, mientras que la frontera inferior es plana ($Z = 0$). La consideración de superficies no planas y con una cierta pendiente como es el caso de los glaciares alpinos requiere sólamente ligeras modificaciones en los razonamientos empleados, complicando la notación.

Cuando modelamos un medio continuo consistente en hielo por debajo de la temperatura de fusión (hielo frío), los balances de la masa, cantidad de movimiento y energía interna del fluido se expresan a través de leyes de conservación: ecuaciones de continuidad, conservación del momento y de la energía interna. Las ecuaciones resultantes se completan con leyes constitutivas que caracterizan el hielo y su respuesta dinámica o estática a los procesos termomecánicos de carga y deformación.

Por lo que respecta a la conservación de masa y momento, el sistema de ecuaciones que describe el flujo de hielo, que se considera en la literatura como un fluido incompresible, viscoso, lento y movido por efecto de la gravedad, está gobernado por las ecuaciones de Stokes estacionarias e incompresibles:

$$U_X + V_Z = 0$$

$$\begin{aligned} 0 &= -P_X + \tau^*_{11X} + \tau^*_{12Z} \qquad (1) \\ 0 &= -P_Z + \tau^*_{12X} + \tau^*_{22Z} - \rho\overline{g} \end{aligned}$$

donde (U, V) son las componentes del campo de velocidades en las direcciones X y Z, respectivamente; P es la presión del hielo; τ^*_{11}, τ^*_{22} y τ^*_{12} son las tensiones longitudinales y transversal, respectivamente; ρ es la densidad; $\overline{g}$ es la fuerza de gravedad y los subíndices X y Z representan las derivadas parciales respecto a las variables X y Z, respectivamente.

Clásicamente, en el caso de los casquetes polares, otras formas de energía interna distintas de la térmica se desprecian de modo que la distribución de temperaturas está gobernada por un balance entre los mecanismos de reacción, conducción y convección. La temperatura verifica la siguiente ecuación:

$$\rho c_p D\Theta/Dt^* = k\nabla^2\Theta + \tau^*_{ij} e^*_{ij}, \qquad (2)$$

con c_p el calor específico, Θ la temperatura en grados Kelvin, k la conductividad térmica, $D/Dt^* = \partial/\partial t^* + U\partial/\partial X + V\partial/\partial Z$ la derivada material con respecto al campo de velocidades (U, V) y t^* el tiempo real en años.

En la ecuación (2), el tensor e^*_{ij} representa la tasa de deformación del material y se define en mecánica de fluidos como la parte simétrica del tensor gradiente de velocidades, es decir,

$$e^*_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \quad \forall i, j = 1, 2, \quad (x_1 = X, \;\; x_2 = Z, \;\; u_1 = U, \;\; u_2 = V). \quad (3)$$

En un modelo clásico de un material isótropo ($\tau^*_{ij} = \tau^*_{ji}$, $e^*_{ij} = e^*_{ji}$), incompresible ($\tau^*_{11} + \tau^*_{22} = e^*_{11} + e^*_{22} = 0$) y viscoso se representa la tasa de deformación en el fluido debida a los esfuerzos a través de los invariantes segundos de los tensores de tensiones y deformaciones τ^*_{ij} y e^*_{ij}, respectivamente. Estos invariantes se definen como

$$2(e^*)^2 = e^*_{ij} e^*_{ij} \;,\; 2(\tau^*)^2 = \tau^*_{ij}\tau^*_{ij} \,.$$

La ley de comportamiento no lineal para el flujo relaciona los tensores e^*_{ij} y τ^*_{ij} mediante la expresión

$$e^*_{ij} = A^*(\Theta) Gl^*(\tau^*)\tau^*_{ij}/\tau^* \qquad (4)$$

de modo que

$$e^* = A^*(\Theta) Gl^*(\tau^*).$$

La función Gl^* es una ley de potencias, conocida como ley de Glen, en la forma

$$Gl^*(\tau^*) = (\tau^*)^n \tag{5}$$

donde n se toma generalmente igual a 3 (ver Fowler [14]) y $A^*(\Theta)$ es una función de la temperatura que generalmente se considera del tipo Arrhenius:

$$A^*(\Theta) = A_0^* \exp(-Q/R\Theta) \tag{6}$$

donde A_0^* es una constante característica del material, Q es la energía de activación y R es la constante universal de los gases.

Por tanto, las ecuaciones que modelan el comportamiento de la región de hielo frío son

$$\begin{aligned}
U_X + V_Z &= 0 && (7)\\
0 &= -P_X + \tau^*_{11X} + \tau^*_{12Z} && (8)\\
0 &= -P_Z + \tau^*_{12X} + \tau^*_{22Z} - \rho\overline{g} && (9)\\
\rho c_p D\Theta/Dt^* &= k\nabla^2\Theta + \tau^*_{ij} e^*_{ij} && (10)\\
e^*_{ij} &= A^*(\Theta) Gl^*(\tau^*)\tau^*_{ij}/\tau^* && (11)
\end{aligned}$$

El modelo definido por las ecuaciones (7)-(11) se plantea en el dominio ocupado por la masa de hielo, cuya frontera superior está definida por la función $Z = \eta^*(t^*, X)$, que a su vez define el perfil superior en contacto con la atmósfera. No obstante, en la práctica, la extensión longitudinal sobre la que se levanta la masa de hielo y la ecuación de la frontera superior (o dicho perfil) son desconocidas "a priori" y su determinación dependerá, entre otros factores, de la acumulación-ablación del hielo (las precipitaciones en forma de nieve, la evaporación, etc.) y del movimiento del casquete polar. En este sentido se trata de una frontera cinemática, cuya localización está caracterizada por la ecuación

$$\frac{\partial \eta^*}{\partial t^*} + U\frac{\partial \eta^*}{\partial X} - V = a^* \tag{12}$$

donde (U, V) es el campo de velocidades en los puntos situados en dicha frontera, es decir, $U(X, \eta^*(t^*, X))$, $V(X, \eta^*(t^*, X))$ y a^* es la tasa real de acumulación-ablación. La deducción de la ecuación de frontera cinemática se puede encontrar en Hutter [20], por ejemplo.

Por otra parte, para la resolución de las ecuaciones (7)-(11) es preciso imponer condiciones de contorno.

En cuanto a la frontera superior, determinada por la ecuación de frontera cinemática, siguiendo trabajos anteriores (Fowler [14], por ejemplo) suponemos

que las tensiones totales en la superficie $Z = \eta^*(t^*, X)$ son iguales a cero, es decir,

$$\sigma^* \vec{n}^* = 0 \quad \text{donde} \quad \sigma^* = -PI + \tau^* \quad \text{y} \quad \vec{n}^* = (-\eta_X^*, 1), \tag{13}$$

y consideramos conocida la temperatura en esa superficie,

$$\Theta(t^*, X, \eta^*(t^*, X)) = \Theta_A(t^*, X, \eta^*(t^*, X)).$$

En cuanto a las condiciones térmicas en la frontera inferior, debemos distinguir dos tipos de condiciones según se trate de un caso polar o politérmico. En efecto, en términos de su distribución de temperatura, los glaciares y los casquetes polares se pueden clasificar en tres tipos: fríos, politérmicos y temperados. Esta clasificación resulta relevante para imponer algunas condiciones de contorno. Así, en el primer caso, se asume que la temperatura en la base está por debajo de la temperatura de fusión y se impone que el flujo de calor equilibra el calor geotérmico, es decir,

$$k \nabla \Theta \cdot \vec{n}^* = G_b$$

donde $\vec{n}^* = (0, -1)$ es el vector normal que apunta hacia el exterior del dominio, que suponemos que reposa en una base horizontal, y G_b es el calor geotérmico que se considera constante (un valor típico en la Antártida es $G_b = 5 \times 10^{-2}\ Wm^{-2}$). En este caso el calor geotérmico se invierte en elevar la temperatura del hielo en esa frontera.

En el caso politérmico, una parte de la frontera, desconocida "a priori", se encuentra a la temperatura de fusión (zona temperada) y el resto está por debajo. Éste fenómeno se modela mediante una condición de tipo Signorini de la siguiente forma:

$$\Theta \leq \Theta_m\,, \quad (\Theta - \Theta_m)(k\,\frac{\partial \Theta}{\partial \vec{n}^*} - G_b) = 0, \quad 0 \leq k\,\frac{\partial \Theta}{\partial \vec{n}^*} \leq G_b$$

donde Θ_m denota la temperatura de fusión del hielo ($\Theta_m = 273\,K$, aproximadamente). En la zona temperada el calor geotérmico se invierte en fundir el hielo. En los modelos considerados en secciones posteriores, se considerará la energía generada por el deslizamiento en la base como una fuente de energía térmica adicional al calor geotérmico.

En teoría, a partir de los datos adecuados mediante las ecuaciones, leyes de comportamiento y condiciones de contorno planteadas en esta sección, es posible determinar la geometría de un casquete polar, su temperatura y velocidad y su distribución de tensiones, así como la variación de estas cantidades con el tiempo para condiciones de contorno prescritas. Sin embargo, en la práctica, la aproximación numérica de la solución del sistema de ecuaciones en derivadas parciales obtenido resulta muy complicada. Por otra parte, el modelado del carácter politérmico (sólo adoptado en las condiciones de contorno) aumentaría aún más la complejidad del sistema resultante. Así, pues, se hace necesario considerar algún

tipo de simplificación, física o puesta en evidencia por la observación, para que la obtención de la solución sea factible e incluya el carácter politérmico. Y aún así, el problema matemático es lo suficientemente complicado para no poder resolverlo analíticamente.

Estas simplificaciones se traducen, por una parte, en ignorar ciertos tipos de fenómenos locales en espacio y tiempo como la formación de grietas en la masa de hielo, la acumulación de tensiones en algunas zonas, la aparición de avalanchas o el desprendimiento de icebergs. Por otra parte, la relevancia de las escalas de tiempo en años, y no en segundos o minutos, implican la exclusión de algunos efectos como las ondas sísmicas, entre otros.

La adimensionalización de las ecuaciones bajo la consideración de hielo poco profundo *(shallow ice approximation)*, que permite despreciar algunos términos del modelo matemático, y la imposición de ciertas hipótesis que relacionan los órdenes de magnitud de algunas incógnitas, dan lugar a modelos que pueden ser tratados con técnicas analíticas y numéricas conocidas.

Es por ello por lo que en la siguiente sección se propone una simplificación de las ecuaciones basándose en los órdenes de magnitud de la región ocupada por la masa de hielo. Es una técnica, que se conoce como escalado de hielo poco profundo (*ice sheet scaling*), que conduce a los modelos de hielo poco profundo. Estos presentan ciertas analogías con los modelos de aguas poco profundas (*shallow water models*).

2.2 Escalado de hielo poco profundo

El conjunto de ecuaciones y condiciones de contorno planteadas en la sección anterior, escrito en variables reales, no tiene en cuenta las diferentes escalas temporales y espaciales. Para obtener las ecuaciones adimensionalizadas se introduce el siguiente conjunto de nuevas variables e incógnitas:

$$X = d_1 x, \quad Z = d_2 z, \tag{14}$$

$$\tau_{12}^* = [\tau^*]\tau_{12}, \quad \tau_{11}^* = \epsilon[\tau^*]\tau_{11}, \quad \tau_{22}^* = \varepsilon[\tau^*]\tau_{22}, \tag{15}$$

$$\Theta = \Theta_m + (\Delta\Theta)T, \quad U = [U]u, \quad V = \varepsilon[U]v, \tag{16}$$

$$A^* = [A^*]A, \quad Gl^* = [Gl^*]Gl, \quad t^* = (d_1/[U])t, \tag{17}$$

$$P - \rho\overline{g}(\eta^* - Z) = \varepsilon[\tau^*]p, \quad \eta^* = d_2\,\eta\,, \tag{18}$$

donde d_1 y d_2 representan los órdenes de magnitud para la longitud y el espesor, respectivamente. En el caso de la Antártida los valores típicos son $d_1 = 3000\ km$ y $d_2 = 3\ km$, de forma que el parámetro adimensional $\varepsilon = d_2/d_1$ es suficientemente pequeño para despreciar las potencias de segundo orden y obtener un modelo reducido.

Los órdenes de magnitud $[\tau^*]$, $[U]$, $[A^*]$ y $[Gl^*]$ correspondientes al esfuerzo, velocidad horizontal, término de Arrhenius y ley de Glen se utilizan para obtener las diferentes funciones escaladas. La temperatura de fusión $\Theta_m = 273\,K$ y el orden de la temperatura del hielo prescrita en la superficie, $\Theta_A = 223\,K$, dan lugar a un rango de temperaturas $\Delta\Theta = 50\,K$ lo que motiva la definición de la variable de temperatura de referencia T.

Además, la elección de $[A^*]$ y $[Gl^*]$ es tal que las funciones adimensionalizadas A y Gl son de orden $O(1)$. Por otro lado, como consecuencia de las expresiones de las leyes de Glen y Arrhenius, tenemos las igualdades

$$[Gl^*] = [\tau^*]^n, \quad [A^*] = A_0^* \exp(-Q/R\Theta_m). \tag{19}$$

En este punto de la deducción del modelo simplificado, con la perspectiva de que el modelo tenga en cuenta la posibilidad de que la mayoría de los efectos de "tensiones transversales" tenga lugar en las proximidades de la base del casquete polar, se introduce un grado de libertad mediante el parámetro ν. Esta idea es original de Fowler [14], quien introduce la relación

$$\frac{[U]}{2\nu d_2} = [A^*][Gl^*] \tag{20}$$

para que el efecto cizalla se localice en una capa de espesor de $O(\nu d_2)$ situada en la base del casquete polar (que es la zona de mayor temperatura). Esta hipótesis de relación entre los órdenes de magnitud tiene que ver con la definición de glaciar politérmico y como veremos más adelante, es crucial para deducir la versión adimensional de la ecuación (12) que define la frontera superior.

Por otra parte, se supone que hay una relación entre los órdenes de magnitud de la velocidad ($[U]$), el parámetro ε y el ratio de acumulación ($[a^*]$). Matemáticamente se escribe en la forma

$$[a^*] = \varepsilon[U].$$

Siguiendo los argumentos de Fowler [14] respecto al balance del tensor de esfuerzos impondremos que

$$[\tau^*] = \rho\,\overline{g}\,d_2\,\varepsilon. \tag{21}$$

A continuación introducimos los cambios de variable (14)-(18), o escalado de hielo poco profundo, en las ecuaciones (7)-(11) y obtenemos el siguiente conjunto de ecuaciones adimensionalizadas para el hielo:

$$
\begin{aligned}
u_x + v_z &= 0 && (22)\\
0 &= -\eta_x + \tau_{12z} + \varepsilon^2(-p_x + \tau_{11x}) && (23)\\
0 &= -p_z + \tau_{12x} + \tau_{22z} && (24)\\
DT/Dt &= (\alpha/\nu)\,\tau\, A(T)\, Gl(\tau) \,+\, \beta(T_{zz} + \varepsilon^2 T_{xx}) && (25)\\
u_z + \varepsilon^2 v_x &= A(T)Gl(\tau)\tau_{12}/\nu\tau && (26)\\
2\nu u_x &= A(T)Gl(\tau)\tau_{11}/\tau && (27)\\
2\nu v_z &= A(T)Gl(\tau)\tau_{22}/\tau && (28)\\
\tau^2 &= \tau_{12}^2 + \varepsilon^2(\tau_{11}^2 + \tau_{22}^2)/2 && (29)
\end{aligned}
$$

Siguiendo a Fowler [14] tomamos los siguientes valores típicos $[a^*] = 0{,}1\ my^{-1}$, $d_2 = 3000\ m$, $\kappa = 38\ m^2y^{-1}$, $c_p = 2\times 10^3\ Jkg^{-1}K^{-1}$, $\overline{g} = 10\ ms^{-2}$, $\Delta\Theta = 50\ K$, para obtener los parámetros adimensionales

$$
\begin{aligned}
\alpha &= \frac{10\ ms^{-2}\ 3000\ m}{2\times 10^3\ Jkg^{-1}K^{-1}\ 50\ K} \approx 0{,}3\\
\beta &= \frac{38\ m^2y^{-1}}{3000\ m\ 0{,}1\ my^{-1}} \approx 0{,}12
\end{aligned}
$$

donde las abreviaturas son las típicas del sistema internacional y, abusando de la notación anglosajona, denotamos con "y" el tiempo en años.

Un paso importante hacia modelos más simplificados es la obtención de la clásica aproximación de Frank-Katmeneskii de la ley de Arrhenius para energías de activación elevadas. En concreto, a partir de (6), (17) y (19) se tiene

$$
\begin{aligned}
A(T) &= \exp[Q/R\Theta_m]\exp[-Q/(R\Theta_m) + Q(\Delta\Theta)T/(R\Theta_m^2)]\\
&= \exp[Q(\Delta\Theta)T/(R\Theta_m^2)]\\
&= \exp[\gamma T]\,,
\end{aligned}
$$

donde $\gamma = Q\Delta\Theta/(R\Theta_m^2)$. El valor de γ se obtiene de manera sencilla del valor aproximado de la energía de activación, $Q = 140\ kJmol^{-1}$, a la temperatura de fusión, Θ_m, y de la constante de Boltzman $R = 8{,}3\ Jmol^{-1}K^{-1}$:

$$
\gamma = \frac{140\ kJmol^{-1}\ 50\ K}{0{,}0083\ kJmol^{-1}K^{-1}\ 273^2\ K^2} \approx 11{,}3\,.
$$

En los siguientes párrafos introducimos el escalado en las condiciones sobre la frontera superior.

En primer lugar, la condición de frontera cinemática (12) se escala mediante la función η como frontera del dominio adimensional:

$$
\eta(t,x) = \frac{1}{d_2}\eta^*(t^*, X),\ \ X = d_1\,x,\ \ t^* = \frac{d_1\,t}{[U]}\,.
$$

Así, teniendo en cuenta que

$$\frac{\partial \eta^*}{\partial t^*} + U\frac{\partial \eta^*}{\partial X} - V = \frac{[U]d_2}{d_1}\frac{\partial \eta}{\partial t} + \frac{[U]d_2}{d_1}u\frac{\partial \eta}{\partial x} - \frac{[U]d_2}{d_1}v$$

y la relación

$$a^* = a[a^*] = \frac{a[U]d_2}{d_1}$$

sustituimos en (12) para deducir:

$$\frac{\partial \eta}{\partial t} + u\frac{\partial \eta}{\partial x} - v = a \quad \text{en} \quad z = \eta(t,x). \tag{30}$$

Para escribir la ecuación anterior en términos de conservación del flujo introducimos éste mediante la expresión

$$\overline{Q} = \int_0^\eta (u,v)\, dz\,.$$

De este modo

$$\begin{aligned}
\nabla \cdot \overline{Q} &= \frac{\partial}{\partial x}\int_0^\eta u(x,z)\, dz + \frac{\partial}{\partial z}\int_0^\eta v(x,s)\, ds \\
&= \int_0^\eta u_x(x,z)\, dz + u(x,\eta(t,x))\frac{\partial \eta(t,x)}{\partial x} \\
&= u(x,\eta(t,x))\frac{\partial \eta(t,x)}{\partial x} - \int_0^\eta v_z(x,z)\, dz \\
&= u(x,\eta(t,x))\frac{\partial \eta(t,x)}{\partial x} - v(x,\eta(t,x)).
\end{aligned}$$

Por tanto la ecuación (30) se escribe de modo equivalente como

$$\frac{\partial \eta}{\partial t} + \nabla \cdot \overline{Q} = a$$

o bien,

$$\frac{\partial \eta}{\partial t} + \frac{\partial}{\partial x}\int_0^\eta u(x,s)\, ds = a. \tag{31}$$

Por otra parte, la condición de frontera (13) sobre las tensiones se escribe de forma más detallada como

$$\begin{aligned}
(P - \tau_{11}^*)\,\eta_X^* + \tau_{12} = 0 \quad &\text{en} \quad Z = \eta^*(t^*, X) \\
-\tau_{12}^*\,\eta_X^* + \tau_{22}^* - P = 0 \quad &\text{en} \quad Z = \eta^*(t^*, X),
\end{aligned}$$

que, en variables adimensionales, se traduce en las ecuaciones:

$$\tau_{12} = \varepsilon^2(\tau_{11} - p)\eta_x \quad \text{en} \quad z = \eta(t,x) \tag{32}$$

$$\tau_{22} = \tau_{12}\eta_x + p \quad \text{en} \quad z = \eta(t,x). \tag{33}$$

En cuanto a la frontera inferior, la adimensionalización (14)-(18) conduce a la condición

$$\frac{\partial T}{\partial \vec{n}} = g_b$$

en el caso polar, y a la condición

$$T \leq 0\,, \quad T\left(\frac{\partial T}{\partial \vec{n}} - g_b\right) = 0\,, \quad 0 \leq \frac{\partial T}{\partial \vec{n}} \leq g_b$$

en el caso politérmico. El parámetro g_b representa el calor geotérmico adimensional y está dado por

$$g_b = \frac{G_b\, d_2}{k\Delta\Theta} \approx 1{,}5$$

donde G_b, el flujo geotérmico real, se toma igual a $G_b = 5 \times 10^{-2}\ Wm^{-2}$, $\Delta\Theta = 50\,K$ y $k = 2{,}1\ Jm^{-1}s^{-1}$.

Una vez introducido el escalado, la aproximación de tipo "hielo poco profundo" consiste en despreciar los términos de orden $\varepsilon^2 \approx 10^{-6}$ en todas las ecuaciones en donde intervienen para obtener modelos simplificados. En concreto, para deducir la condición cinemática de frontera libre en términos de la función η, en primer lugar despreciamos los términos de $O(\varepsilon^2)$ en (23), (26) y (29), obteniendo

$$\begin{aligned} 0 &= -\eta_x + \tau_{12z} && (34)\\ u_z &= A(T)Gl(\tau)\tau_{12}/\nu\tau && (35)\\ \tau^2 &= \tau_{12}^2\,. && (36) \end{aligned}$$

Además, despreciando los términos de $O(\varepsilon^2)$ en la condición de frontera (32), se deduce que $\tau_{12} = 0$ y de (33) se sigue $\tau_{22} = p$.

Integrando la ecuación (34) se tiene

$$-z\eta_x + \tau_{12} = f(x)$$

de la cual, junto con $\tau_{12} = 0$ en $z = \eta$, se concluye

$$\tau_{12} = -(\eta - z)\eta_x\,.$$

Por lo tanto, la ecuación (36) es equivalente a

$$\tau = |\tau_{12}| = (\eta - z)|\eta_x|. \tag{37}$$

Por otro lado, la sustitución de (37) en (35) y la aplicación de la ley de Glen (5) dan lugar a

$$u_z = \frac{-A(T)}{\nu}(\eta - z)^n\, |\eta_x|^{n-1}\, \eta_x\,. \tag{38}$$

Así, la expresión para la velocidad longitudinal se obtiene al integrar entre 0 y z la ecuación (38):

$$u - u_b = \int_0^z \frac{-A(T)}{\nu} (\eta - s)^n \, |\eta_x|^{n-1} \, \eta_x \, ds, \tag{39}$$

donde u_b es la velocidad de deslizamiento en la base (i.e. en $z = 0$).

Para avanzar hacia una ecuación para el perfil, introducimos la expresión

$$A(T) \approx e^{-\gamma}, \tag{40}$$

que representa una aproximación isotérmica del problema no isotérmico de partida. De (39) y (40) podemos deducir la siguiente aproximación para la velocidad:

$$\begin{aligned} u &\approx u_b - \frac{e^{-\gamma}}{\nu} \int_0^z (\eta - s)^n \, |\eta_x|^{n-1} \, \eta_x \, ds \\ &= u_b - \frac{e^{-\gamma} \, |\eta_x|^{n-1} \, \eta_x}{\nu(n+1)} \left[\eta^{n+1} - (\eta - z)^{n+1}\right] \end{aligned} \tag{41}$$

Si introducimos la función de flujo transversal

$$\Upsilon(x, \eta(t,x)) = \int_0^\eta u \, dz,$$

utilizando (41), podemos expresarla en la forma

$$\begin{aligned} \Upsilon(x, \eta(t,x)) &= \eta u_b - \frac{e^{-\gamma} \, |\eta_x|^{n-1} \, \eta_x}{\nu \, (n+1)} \int_0^\eta \left[\eta^{n+1} - (\eta - z)^{n+1}\right] dz = \\ &= \eta u_b - \frac{e^{-\gamma} \, |\eta_x|^{n-1} \, \eta_x \, \eta^{n+2}}{\nu \, (n+2)} \, . \end{aligned}$$

Entonces la ecuación de frontera cinemática (31) se escribe

$$\frac{\partial \eta}{\partial t} + \frac{\partial}{\partial x} \left(\Upsilon(x, \eta(t,x))\right) = a$$

o, equivalentemente,

$$\frac{\partial \eta}{\partial t} + \frac{\partial (\eta u_b)}{\partial x} - \frac{e^{-\gamma}}{\nu} \frac{\partial}{\partial x} \left(\frac{\eta^{n+2} |\eta_x|^{n-1} \, \eta_x}{n+2} \right) = a, \tag{42}$$

que es la ecuación fundamental para el cálculo del perfil del casquete polar.

Debemos puntualizar que la aproximación isotérmica (40) evita el cálculo de un modelo integro-diferencial más complejo en lugar de la ecuación parabólica no lineal resultante (42).

A continuación procedemos a obtener la ecuación básica del modelo térmico en el marco de la aproximación de hielo poco profundo. Para ello se desprecian los términos de orden ε^2 en las ecuaciones (22)-(29). Así, la ecuación (29) da lugar a

$$\tau = |\tau_{12}|$$

y si en la ecuación (23) integramos respecto a z resulta

$$\tau_{12} = -(\eta - z)\eta_x$$

y, por tanto, se obtiene una aproximación de τ en la ecuación térmica (25). Además, despreciando los términos de orden ε^2 en esta misma ecuación, se tiene

$$\frac{DT}{Dt} = (\alpha/\nu)\,\tau\,A(T)\,Gl(\tau) \;+\; \beta T_{zz}$$

o, equivalentemente,

$$\frac{\partial T}{\partial t} + \vec{v}\cdot\nabla T - \beta\frac{\partial^2 T}{\partial z^2} - (\alpha/\nu)\,\tau\,A(T)\,Gl(\tau) = 0, \tag{43}$$

que es la ecuación reducida fundamental para la distribución de temperaturas en la región de hielo frío.

Si consideramos la ley de Glen (5), la expresión para las tensiones (37) y la aproximación de Frank-Kamenetskii, entonces la ecuación (43) se puede escribir en la forma

$$\frac{\partial T}{\partial t} + \vec{v}\cdot\nabla T - \beta\frac{\partial^2 T}{\partial z^2} - \left(\frac{\alpha}{\nu}\right)((\eta(t,x) - z)\eta_x)^4 e^{\gamma T} = 0. \tag{44}$$

3 Modelo acoplado de hielo poco profundo

En la sección anterior se ha realizado un deducción detallada de las ecuaciones que se obtienen al introducir el escalado de hielo poco profundo en las ecuaciones de conservación de la masa, momento y energía, así como en las leyes constitutivas y en las condiciones de contorno. Al igual que en el modelo de partida, las incógnitas del problema siguen siendo el perfil superior de la masa de hielo (que, además, identifica la región que ésta ocupa), el campo de velocidades y la distribución de temperatura en el interior del glaciar. También, al igual que en el problema de partida, el cálculo de las incógnitas conlleva la resolución de un problema fuertemente acoplado en el modelo de hielo poco profundo; las incógnitas están presentes en las distintas ecuaciones y no es posible resolverlas separadamente.

Para plantear el modelo acoplado de hielo poco profundo, dado que se desconoce la región ocupada por el hielo, en primer lugar se considera un dominio rectangular fijo Ω que incluye, además de la sección longitudinal

del casquete polar, parte de la atmósfera que rodea a la masa de hielo. Concretamente, se define el dominio:

$$\Omega = \{(x,z)\,/\ -1 \le x \le 1\,,\ \ 0 \le z \le z_{max}\,\}, \tag{45}$$

de modo que la masa de hielo ocupa la región

$$\Omega_I(t) = \{(x,z)\,/\ S_-(t) \le x \le S_+(t)\,,\ \ 0 \le z \le \eta(t,x)\}, \tag{46}$$

donde se tiene la inclusión $\Omega_I(t) \subset \Omega$ y, denotando por $\Omega_A(t)$ el dominio ocupado por la atmósfera, se verifica

$$\Omega = \Omega_I(t) \cup \Omega_A(t). \tag{47}$$

Dado que la extensión longitudinal de la capa de hielo es desconocida, el intervalo $(S_-(t), S_+(t)) \subset (-1,1)$ de las definiciones de los conjuntos anteriores, constituye una incógnita adicional. La determinación de dicho intervalo se realiza al resolver el problema de frontera móvil asociado a la ecuación del perfil.

En las secciones siguientes se plantean los modelos que gobiernan los tres subproblemas fundamentales: cálculo de perfil superior, de las velocidades y de la temperatura. Cada uno de los subproblemas requiere conocer la solución de los otros dos, lo cual sugiere la propuesta de un algoritmo iterativo para la simulación numérica que resuelva secuencialmente los distintos subproblemas.

Además de la novedad que suponen algunos aspectos de modelado de los fenómenos y el algoritmo de resolución propuesto para la simulación numérica con datos reales, creemos conveniente resaltar la importante aportación que supone la inclusión del cálculo de velocidades así como la consideración de los efectos basales en el problema térmico de Stefan-Signorini.

4 Modelo de frontera móvil para el perfil

4.1 Planteamiento del modelo

Tal como se ha deducido en un apartado anterior, la ecuación (42) es válida para la función η que define el perfil superficial del casquete polar en los puntos x donde $\eta > 0$. Sin embargo, el conjunto de estos puntos es una incógnita adicional del problema de partida debido al hecho de que, *a priori*, la extensión longitudinal de la capa de hielo es desconocida. Por lo tanto, estamos ante un problema típico de frontera móvil que planteamos a continuación.

Sea $t_{\text{máx}} > 0$ un instante de tiempo suficientemente grande; se consideran una función de acumulación-ablación $a : (0, t_{\text{máx}}) \times (-1,1) \to \mathbb{R}$ y un perfil inicial $\eta_0 : (-1,1) \to \mathbb{R}$ conocidos. Entonces, el problema se formula como sigue:

Para todo $t \in [0, t_{\text{máx}}]$, encontrar el conjunto $\Gamma_0(t) = (S_-(t), S_+(t)) \subset (-1, 1)$ y la función

$$\eta : \mathcal{Q} = \bigcup_{t\in[0,t_{\text{máx}}]} \Gamma_0(t) \to \mathbb{R}$$

tales que :

$$\begin{array}{l} \dfrac{D\eta}{Dt} \geq \dfrac{e^{-\gamma}}{\nu}\dfrac{\partial}{\partial x}\left(\dfrac{\eta^{n+2}}{n+2}\left|\dfrac{\partial \eta}{\partial x}\right|^{n-1}\dfrac{\partial \eta}{\partial x}\right) + a \quad \text{en } \mathcal{Q} \\ \\ \eta \geq 0 \quad \text{en } \mathcal{Q} \\ \\ \left(\dfrac{D\eta}{Dt} - \dfrac{e^{-\gamma}}{\nu}\dfrac{\partial}{\partial x}\left(\dfrac{\eta^{n+2}}{n+2}\left|\dfrac{\partial \eta}{\partial x}\right|^{n-1}\dfrac{\partial \eta}{\partial x}\right) + a\right)\eta = 0 \ \text{ en } \mathcal{Q} \\ \\ \eta = 0 \quad \text{en } \{S_-(t)\} \cup \{S_+(t)\}, \quad t \in (0, t_{\text{máx}}); \quad \eta(0, x) = \eta_0(x) \quad \text{en } (-1, 1), \end{array} \tag{48}$$

donde se ha utilizado para la derivada material respecto de la velocidad de deslizamiento basal, u_b, la notación

$$\frac{D\eta}{Dt} = \frac{\partial \eta}{\partial t} + \frac{\partial}{\partial x}(u_b\, \eta). \tag{49}$$

Además, la función a representa la tasa de acumulación-ablación.

La solución del problema de complementariedad (48) proporciona, para cada instante de tiempo t, la extensión de la masa de hielo $(S_-(t), S_+(t))$ y la función $\eta(t, .)$ que define la frontera superior de $\Omega_I(t)$. Así, las fronteras del dominio $\Omega_I(t)$ se denotan en la forma

$$\Gamma_0(t) \quad = \quad \{(x, z) \,/\, x \in (S_-(t), S_+(t)),\ z = 0\} \tag{50}$$

$$\Gamma_1(t) \quad = \quad \{(x, z) \,/\, x \in \Gamma_0(t)\,,\ z = \eta(t, x)\}. \tag{51}$$

En los ejemplos reales, la función a es positiva en la región central y negativa cerca de las fronteras de $I(t)$. Estas hipótesis cualitativas sobre la función a son debidas a que la acumulación de nieve tiene lugar fundamentalmente en la región superior del casquete polar y la ablación (evaporación y fusión) ocurre principalmente cerca de los márgenes.

El modelado del deslizamiento basal ha sido objeto de varios trabajos de diferentes autores (ver las referencias en Paterson [19], Hutter [20] o Fowler [15], por ejemplo). La deducción de la expresión de una ley de deslizamiento en función de otras variables, como el tensor de tensiones basal o la presión efectiva, es un tema controvertido entre los glaciólogos teóricos aunque la mayoría de ellos está de acuerdo en la necesidad de considerarla en el caso de casquetes polares politérmicos. En un apartado posterior explicamos nuestra elección a la hora de modelar dicho fenómeno.

4.2 Resolución numérica del modelo del perfil

En cuanto a la resolución numérica, en cada etapa del algoritmo global que proponemos en este trabajo, es necesario resolver el modelo (48) de frontera móvil del perfil para una velocidad basal y una temperatura conocidas. Para ello, utilizamos las técnicas numéricas que combinan el método de las características, elementos finitos y algoritmos de dualidad para los operadores monótonos asociados a la no linealidad del término de difusión y a la que surge de la presencia de una frontera móvil. A continuación describimos dicho método.

En primer lugar, nótese que la resolución numérica mediante un esquema explícito en difusión requiere pasos de tiempo muy pequeños, como se ha puesto de manifiesto en ensayos previos (ver [8]). Para evitar este problema, proponemos una formulación equivalente a (48) con la parte de difusión no lineal en términos de un operador maximal monótono (ver [6], para los detalles). Dado que $n = 3$, se define la nueva incógnita u como

$$u(t,x) = \eta^{8/3}(t,x)$$

de modo que el problema (48) se puede reescribir en la incógnita u en la forma:

$$\begin{cases} \dfrac{D}{Dt}\left(u^{3/8}\right) - \mu\left(|u_x|^2 u_x\right)_x - a \geq 0 & \text{en} \quad (0,t_{\text{máx}}) \times \Omega \\ \left[\dfrac{D}{Dt}\left(u^{3/8}\right) - \mu\left(|u_x|^2 u_x\right)_x - a\right] u = 0 & \text{en} \quad (0,t_{\text{máx}}) \times \Omega \\ u \geq 0 & \text{en} \quad (0,t_{\text{máx}}) \times \Omega \\ u = 0 & \text{en} \quad (0,t_{\text{máx}}) \times \partial\Omega \\ u = u_0(x) = \eta_0^{8/3}(x) & \text{en} \quad \Omega, \end{cases} \tag{52}$$

donde la constante μ toma el valor $\mu = \frac{(3/8)^3}{5}$. La introducción del cambio de variable anterior en las ecuaciones permite obtener una formulación maximal monótona del nuevo término no lineal de difusión, pero da lugar a un término no lineal de convección.

El siguiente paso consiste en la semidiscretización en tiempo de esta nueva formulación. Para ello, sean t_{max}, M_t y Δt dos números reales positivos tales que $t_{max} = M_t \Delta t$. El problema (52) se discretiza en tiempo mediante un esquema de características con paso de tiempo Δt. Este esquema está basado en la aproximación de la derivada total (ver Pironneau [21], para ecuaciones lineales de convección-difusión).

Así, en nuestro caso particular con convección no lineal, para $m = 0, 1, \ldots, M_t$, $(M_t = t_{max}/\Delta t)$, se propone la siguiente aproximación:

$$\frac{D}{Dt}(u^{3/8})((m+1)\Delta t, x) \approx \frac{(u^{m+1})^{3/8}(x) - J^m(x)\,(u^m)^{3/8}(\chi^m(x))}{\Delta t} \qquad (53)$$

donde

$$u^{m+1}(x) = u((m+1)\Delta t, x), \qquad \text{en} \quad \Omega,$$

y $J^m(x)$ se obtiene utilizando cuadratura numérica en la expresión

$$J^m(x) = J(t^{m+1}, x; t^m) = 1 - \int_{t^m}^{t^{m+1}} (u_b(\tau, \chi(x, t^{m+1}; \tau)))_x \, d\tau,$$

donde J es el jacobiano asociado al cambio de variable definido por $x \to \chi(t, x; \tau)$ (ver Bercovier, Pironneau and Sastri [2] para los detalles).

El valor $\chi^m(x)$ está dado por $\chi^m(x) = \chi((m+1)\Delta t, x; m\Delta t)$ donde χ es solución del problema de ecuaciones diferenciales

$$\begin{cases} \dfrac{d\chi(t,x;s)}{ds} &= u_b(s, \chi(x,t;s)), \\ \chi(t,x;t) &= x. \end{cases}$$

Sustituyendo la aproximación (53) en (52), se obtiene la siguiente sucesión de problemas de complementariedad elípticos no lineales:

Para $m = 0, 1, 2, \ldots, M_t$, *hallar* u^{m+1} *tal que:*

$$\begin{cases} \dfrac{(u^{m+1})^{3/8} - J^m((u^m)^{3/8} \circ \chi^m)}{\Delta t} - \mu \dfrac{\partial}{\partial x}\left(|u_x^{m+1}|^2 u_x^{m+1}\right) - a^{m+1} \geq 0 \quad \text{en } \Omega \\ u^{m+1} \geq 0 \quad \text{en } \Omega \\ \left[\dfrac{(u^{m+1})^{3/8} - J^m((u^m)^{3/8} \circ \chi^m)}{\Delta t} - \mu \dfrac{\partial}{\partial x}\left(|u_x^{m+1}|^2 u_x^{m+1}\right) - a^{m+1}\right] u^{m+1} = 0 \quad \text{en } \Omega \\ u^{m+1} = 0 \quad \text{en } \partial\Omega \\ u_0(x) = (\eta_0)^{8/3}(x) \quad \text{en } \Omega \end{cases} \qquad (54)$$

donde $a^{m+1}(x) = a((m+1)\Delta t, x)$ y ` $\circ$ ´ denota el símbolo de la composición.

Para la resolución de los problemas de complementariedad no lineales semidiscretizados en tiempo (54), se propone una formulación variacional utilizando el conjunto convexo

$$K = \{\varphi \in W_0^{1,4}(\Omega) \,/\, \varphi \geq 0 \text{ c.p.d en } \Omega\},$$

donde $W_0^{1,4}(\Omega)$ es el espacio de Sobolev adecuado [1], y, posteriormente, resolver el siguiente problema de desigualdad variacional:

Encontrar $u^{m+1} \in K$ *tal que*

$$\frac{1}{\Delta t}\int_\Omega (u^{m+1})^{3/8}\,(\varphi - u^{m+1})\,dx + \mu\int_\Omega |\,u_x^{m+1}\,|^2\,u_x^{m+1}(\varphi-u^{m+1})_x dx \geq \frac{1}{\Delta t}\int_\Omega J^m\,((u^m)^{3/8}\circ\chi^m)(\varphi-u^{m+1})\,dx + \int_\Omega a^{m+1}\,(\varphi-u^{m+1})dx, \qquad \forall\varphi\in K.$$

Los resultados del cálculo subdiferencial [5] para la función indicatriz I_K del convexo K permiten reescribir el problema con una formulación variacional equivalente: *Encontrar* $u^{m+1}\in W_0^{1,4}(\Omega)$ *tal que*

$$\frac{1}{\Delta t}\int_\Omega (u^{m+1})^{3/8}\,\psi\,dx + \int_\Omega \xi_1^{m+1}\,\psi\,dx + \mu\int_\Omega \xi_2^{m+1}\,\psi_x\,dx - \frac{1}{\Delta t}\int_\Omega J^m((u^m)^{3/8}\circ\chi^m)\,\psi\,dx = \int_\Omega a^{m+1}\,\psi\,dx, \qquad \forall\psi\in W_0^{1,4}(\Omega) \tag{55}$$

$$\xi_1^{m+1} \in \partial I_K\left(u^{m+1}\right) \tag{56}$$

$$\xi_2^{m+1} = \Lambda\left(\frac{\partial u^{m+1}}{\partial x}\right) \tag{57}$$

donde $\Lambda(v) = |\,v\,|^2\,v = v^3$.

El problema no lineal (55)-(57) puede resolverse mediante la aplicación de un método de dualidad propuesto en [3]. La idea consiste en introducir dos nuevas incógnitas q_1^{m+1} y q_2^{m+1} (multiplicadores de Lagrange) definidas por

$$q_1^{m+1} \in \partial I_K\left(u^{m+1}\right) - \omega_1 u^{m+1} \tag{58}$$

$$q_2^{m+1} = \Lambda\left(\frac{\partial u^{m+1}}{\partial x}\right) - \omega_2\frac{\partial u^{m+1}}{\partial x} \tag{59}$$

en términos de dos parámetros positivos ω_1 y ω_2 y escribir (55) en la forma:

$$\frac{1}{\Delta t}\int_\Omega (u^{m+1})^{3/8}\,\psi\,dx + \int_\Omega (q_1^{m+1}+\omega_1 u^{m+1})\,\psi\,dx + \mu\int_\Omega\left(q_2^{m+1}+\omega_2\frac{\partial u^{m+1}}{\partial x}\right)\frac{\partial\psi}{\partial x}\,dx = \int_\Omega a^{m+1}\,\psi\,dx + \frac{1}{\Delta t}\int_\Omega J^m\,((u^m)^{3/8}\circ\chi^m)\psi dx, \qquad \forall\psi\in W_0^{1,4}(\Omega). \tag{60}$$

Dado que ∂I_K y Λ son operadores maximales monótonos, las definiciones (58) y (59) se caracterizan por las respectivas igualdades:

$$q_1^{m+1} = (\partial I_K)_{\lambda_1}^{\omega_1}\left[u^{m+1} + \lambda_1 q_1^{m+1}\right] \tag{61}$$

$$q_2^{m+1} = \Lambda_{\lambda_2}^{\omega_2}\left[\frac{\partial u^{m+1}}{\partial x} + \lambda_2 q_2^{m+1}\right] \tag{62}$$

donde $(\partial I_K)_{\lambda_1}^{\omega_1}$ y $\Lambda_{\lambda_2}^{\omega_2}$ denotan las aproximaciones Yosida (ver Brezis [4, 5], por ejemplo) de los operadores $(\partial I_K - \omega_1 I)$ y $(\Lambda - \omega_2 I)$ con parámetros positivos λ_1 y λ_2.

Para aproximar en espacio las ecuaciones (60), (61) y (62) se considera el espacio de elementos finitos de Lagrange de grado menor o igual que uno. Así, para un parámetro positivo h, se construye una malla uniforme de elementos finitos τ_h con nodos $x_i = (i-1)h$, $i = 1, \ldots, N+1$. Se introducen a continuación los espacios y conjuntos clásicos de elementos finitos

$$V_h = \{\varphi_h \in C^0(\Omega) \,/\, \varphi_h\,|_E \in P_1, \quad \forall E \in \tau_h\},$$

$$V_{0h} = \{\varphi_h \in V_h \,/\, \varphi_h\,|_{\partial\Omega} = 0\},$$

$$K_h = \{\varphi_h \in V_{0h} \,/\, \varphi_h \geq 0 \quad \text{c.p.d. en}\,\Omega\},$$

donde E denota un elemento finito. Entonces, el problema discretizado puede escribirse en la forma:

Encontrar $u_h^{m+1} \in K_h$ *tal que*

$$\begin{aligned}
&\frac{1}{\Delta t}\int_\Omega (u_h^{m+1})^{3/8}\,\psi_h\,dx \;+\; \omega_1\int_\Omega u_h^{m+1}\,\psi_h\,dx \;+\\
&\mu\omega_2\int_\Omega \frac{\partial u_h^{m+1}}{\partial x}\,\frac{\partial \psi_h}{\partial x}\,dx \;=\; \int_\Omega a_h^{m+1}\,\psi_h\,dx \;+\\
&\frac{1}{\Delta t}\int_\Omega J^m\,((u_h^m)^{3/8}\circ\chi^m)\,\psi_h\,dx - \int_\Omega q_{1,h}^{m+1}\,\psi_h\,dx -\\
&\mu\int_\Omega q_{2,h}^{m+1}\,\frac{\partial\psi_h}{\partial x}\,dx, \quad \forall\psi_h \in V_{0h}.
\end{aligned} \tag{63}$$

Finalmente, la no linealidad del primer término de (63) se trata mediante un método de punto fijo que a cada paso del algoritmo actualiza los multiplicadores de Lagrange utilizando (61) y (62). Esto es:

Etapa 0 : Inicializar $(u_h^{m+1})_0$ (igual a u_h^m, por ejemplo)

Etapa j : Dada$(u_h^{m+1})_j$, calcular $(u_h^{m+1})_{j+1} \in V_{0h}$
resolviendo el problema lineal:

$$\omega_1 \int_\Omega (u_h^{m+1})_{j+1}\ \psi_h\ dx + \mu\omega_2 \int_\Omega \frac{\partial (u_h^{m+1})_{j+1}}{\partial x}\frac{\partial \psi_h}{\partial x}\ dx\ =$$

$$-\frac{1}{\Delta t}\int_\Omega (u_h^{m+1})_j^{3/8}\ \psi_h\ dx\ - \int_\Omega (q_{1,h}^{m+1})_j\ \psi_h\ dx$$

$$-\mu \int_\Omega (q_{2,h}^{m+1})_j\ \frac{\partial \psi_h}{\partial x} dx + \int_\Omega a_h^{m+1}\psi_h\ dx$$

$$+\frac{1}{\Delta t}\int_\Omega J^m\ (((u_h^m)_j)^{3/8} \circ \chi^m)\ \psi_h\ dx, \qquad \forall \psi \in V_{0h}.$$

$$(q_{1,h}^{m+1})_{j+1} \quad = \quad (\partial I_K)_{\lambda_1}^{\omega_1}\left[(u_h^{m+1})_{j+1} + \lambda_1 (q_{1,h}^{m+1})_j\right]$$

$$(q_{2,h}^{m+1})_{j+1} \quad = \quad \Lambda_{\lambda_2}^{\omega_2}\left[\frac{\partial}{\partial x}(u_h^{m+1})_{j+1} + \lambda_2 (q_{2,h}^{m+1})_j\right]$$

La convergencia del método de dualidad está garantizada para $\lambda_i\omega_i = 0{,}5$, con i=1,2. Para esta elección de parámetros, las aproximaciones Yosida (61) y (62) se calculan fácilmente como

$$(\partial I_K)_{\frac{1}{2\omega_1}}^{\omega_1}(r) \quad = \quad -2\omega_1\ |\ r\ |\,,$$

$$\Lambda_{\frac{1}{2\omega_2}}^{\omega_2}(r) \quad = \quad 2\,\Lambda_{\frac{1}{\omega_2}}(2r) - 2\,\omega_2\, r\,,$$

donde $\Lambda_\lambda(r) = (r-s)/\lambda$, siendo s solución de la ecuación no lineal $\lambda s^3 + s = r$, que puede resolverse para cada r utilizando las fórmulas de Cardano.

4.3 Validación numérica del modelo y algoritmos

El primer ejemplo trata de validar la respuesta del método numérico comparando los resultados que proporciona el algoritmo numérico con la solución exacta del problema para un caso particular. Para ello, se consideran $(0,T)$ un intervalo suficientemente grande y $\Omega = (-L,L)$ el dominio del problema, que incluye la región sobre la que descansa el hielo. Se define la función de acumulación- ablación constante a trozos:

$$a(x) = \begin{cases} a_1 & \text{si } 0 \le |x| < R \\ \\ -a_2 & \text{si } R \le |x| \le L, \end{cases} \tag{64}$$

donde $L > 1$, $a_1 > 0$, $a_2 > 0$ y $R \in (0,1)$ y verificando la identidad

$$a_1 R = a_2(1-R). \tag{65}$$

Entonces para $a_1 = 1{,}24 \times 10^{-5}$ y $a_2 = 3{,}72 \times 10^{-5}$, el problema con $u_b = 0$ admite la siguiente solución exacta estacionaria:

$$\overline{\eta}(x) = \begin{cases} H\left[1 - \left(1 + \dfrac{a_1}{a_2}\right)^{1/3} \left(\dfrac{|x|}{L}\right)^{4/3}\right]^{3/8} & \text{si } |x| \le R \\ H\left(1 + \dfrac{a_2}{a_1}\right)^{1/8} \left(1 - \dfrac{|x|}{L}\right)^{1/2} & \text{si } R \le |x| \le 1 \\ 0 & \text{si } 1 \le |x| \le L, \end{cases} \tag{66}$$

donde $H = (40\, a_1\, R)^{1/8}$ representa el espesor en $x = 0$ (*divide*) (ver Paterson [19]).

En este primer test (Test 1) se toman $L = 2$, $R = 0{,}75$ y, por tanto, $H = 0{,}86$ y se considera como condición inicial para el problema evolutivo la siguiente función:

$$\eta_0(x) = \begin{cases} c\,(1 - |x|^{4/3})^{3/8} & \text{si} \quad |x| \le 1 \\ 0 & \text{si} \quad 1 \le |x| \le 2\,, \end{cases} \tag{67}$$

con $c = 0{,}5$.

Para una malla uniforme de elementos finitos con $N = 2001$ nodos y un paso de tiempo igual a $\Delta t = 1$, en la Figura 1 se presentan el perfil inicial ($t = 0$), las soluciones numéricas para $t = 5$ y $t = 75$, y la solución exacta estacionaria (66) para el Test 1 (que coincide con la solución numérica para $t = 125$). Los parámetros de dualidad utilizados son $\omega_1 = 15$ y $\omega_2 = 30$.

El Test 2 se ha diseñado para ilustrar la propiedad de *tiempo de espera*. La idea es mostrar numéricamente esta propiedad cualitativa observada en el movimiento de grandes masas de hielo. La propiedad de tiempo de espera consiste en que cuando la condición inicial del problema tiene forma convexa-cóncava suficientemente plana entonces el desplazamiento inicial de la frontera libre ($S_+(t_0)$, por ejemplo) comienza después de un cierto tiempo (*tiempo de espera*), mientras que para una condición inicial cóncava (como (67), por ejemplo) este desplazamiento tiene lugar de forma instantánea (ver [6]).

Se han comparado las soluciones numéricas obtenidas con la condición inicial

dada en (67) y con la siguiente condición inicial convexa-cóncava:

$$\overline{\eta}_0(x) = \begin{cases} c\,(1-|x|^{4/3})^{3/8} & \text{si} \quad -0{,}75 \le x \le 0{,}75 \\ 16{,}77\,c\left(\dfrac{a_2}{2}\right)^{1/3}|x-1|^{4/3} & \text{si} \quad 0{,}75 \le x \le 1 \\ 16{,}77\,c\left(\dfrac{a_2}{2}\right)^{1/3}|x+1|^{4/3} & \text{si} \quad -1 \le x \le -0{,}75 \\ 0 & \text{en otro caso} \end{cases} \tag{68}$$

mostrando en la Figura 2 las correspondientes fronteras móviles para $c = 0{,}75$. En este caso, se han tomado los datos $L = 2$ y $R = 1$ en (64), de modo que la ecuación (65) no se verifica y no conocemos la solución exacta.

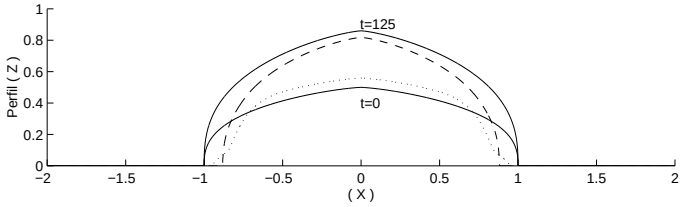


Figura 1: Solución numérica para el Test 1. $t = 0(-)$, $t = 5(\cdots)$, $t = 75(--)$, Estacionaria $(-)$.

Finalmente, para completar el Test 2, introducimos el campo de velocidad basal

$$u_b(t,x) = \begin{cases} C\,x^2 & \text{si} \quad x \ge 0 \\ -C\,x^2 & \text{si} \quad x < 0. \end{cases} \tag{69}$$

y consideramos de nuevo la condición inicial (67) con $c = 0{,}5$. Las Figuras 3, 4 y 5 muestran la evolución de las soluciones para distintas velocidades ($C = 0{,}005$, $C = 0{,}05$, $C = 0{,}1$) y se comprueba la pérdida de los perfiles cóncavos de las soluciones a medida que aumenta la convección. Asimismo, en la Figura 6 se constata la influencia del término convectivo en la propiedad de *tiempo de*

espera cuando inicializamos con la función (68) para $c = 0{,}75$ y utilizamos varios valores de C en el campo de velocidades (69).

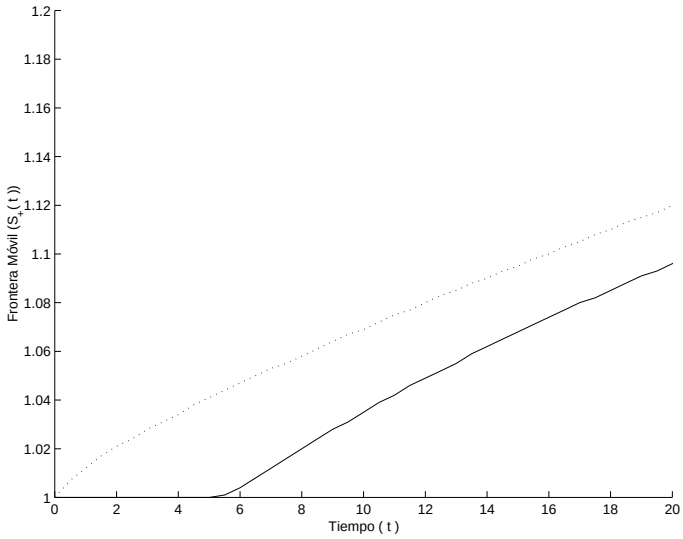


Figura 2: Evolución de la frontera móvil $S_+(t)$ con condición inicial convexa-cóncava $\overline{\eta}_0$ $(-)$ y cóncava η_0 $(\cdots)$ para $c = 0{,}75$.

5 Modelo de hielo poco profundo para la velocidad

En esta sección describimos brevemente una propuesta que permite obtener expresiones para la velocidad del hielo en el interior del glaciar en el marco del modelo de hielo poco profundo, sin necesidad de recurrir a las ecuaciones previas a este escalado. Dicha propuesta original se recoge en [11]. En concreto, al introducir el escalado de hielo poco profundo descrito en la primera sección de este trabajo y despreciar los términos de orden $\varepsilon^2 = 10^{-6}$ en el modelo global, se obtiene la siguiente expresión para la derivada parcial respecto de z de la componente horizontal de la velocidad:

$$\frac{\partial u}{\partial z} = \frac{-A(T)}{\nu}(\eta - z)^n \left|\frac{\partial \eta}{\partial x}\right|^{n-1} \frac{\partial \eta}{\partial x}\,. \tag{70}$$

Anteriormente se ha explicado que la función $A(T) = \mathrm{e}^{\gamma T}$ proviene de la aproximación de Frank-Katmeneskii de la ley de Arrhenius para los efectos térmicos en la disipación viscosa, siendo T la temperatura y el parámetro adimensional $\gamma = 11{,}3$.

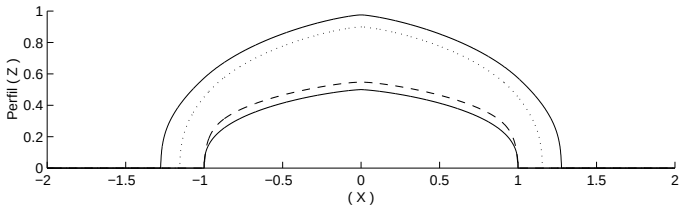


Figura 3: Solución numérica para el Test 2 con $C = 0{,}005$. $t = 0\,(-)$, $t = 5\,(--)$, $t = 50\,(\cdots)$, $t = 90\,(-)$.

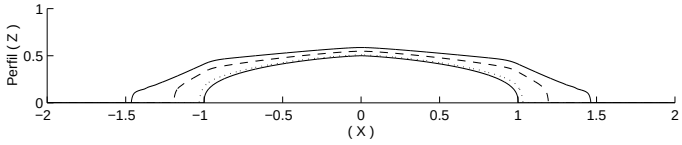


Figura 4: Solución numérica para el Test 2 con $C = 0{,}05$. $t = 0\,(-)$, $t = 1\,(--)$, $t = 5\,(\cdots)$, $t = 9\,(-)$.

Para deducir las expresiones de la componentes de la velocidad que se utilizarán en la simulación numérica del modelo acoplado, en primer lugar, siguiendo el mismo camino que para la deducción de la velocidad longitudinal en (39), integramos la ecuación (70) entre 0 y z, teniendo en cuenta que la velocidad en

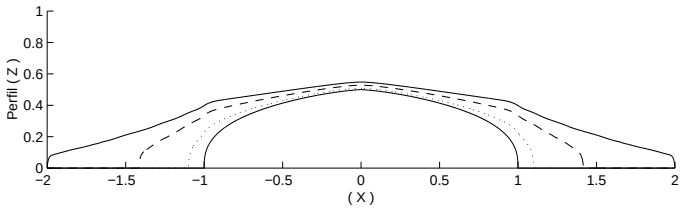


Figura 5: Solución numérica para el Test 2 con $C = 0{,}1$. $t = 0\,(-)$, $t = 1\,(\cdots)$, $t = 3\,(--)$, $t = 5\,(-)$.

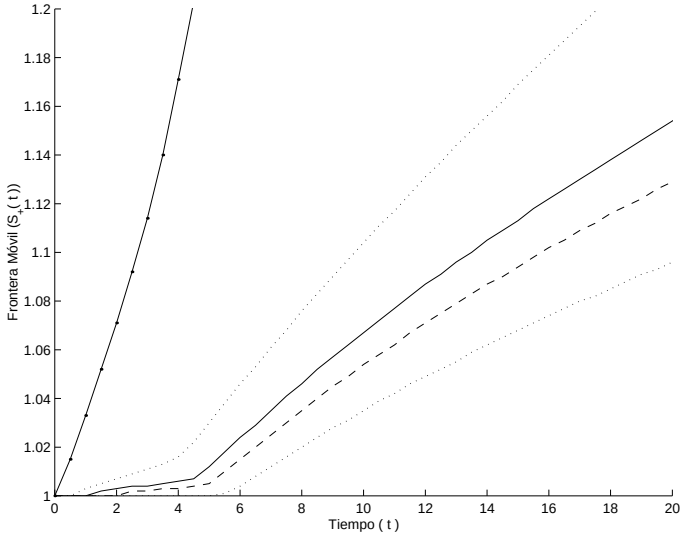


Figura 6: Evolución de la frontera móvil $S_+(t)$ con $\overline{\eta}_0$ y $c = 0{,}75$.
$C = 0(\cdots)$, $C = 0{,}003(--)$, $C = 0{,}005(-)$, $C = 0{,}01(\cdots)$, $C = 0{,}05(.-.-)$.

$z = 0$ es la velocidad basal, u_b.

A continuación, si denotamos por ψ a la función de corriente asociada al campo de velocidades del hielo, podemos extenderla al dominio fijo, Ω, que incluye a la atmósfera, en la forma:

$$\psi(x,z) = \begin{cases} \displaystyle\int_0^z u(x,s)\, ds & \text{si } z \leq \eta \\ \\ \displaystyle\int_0^\eta u(x,s)\, ds & \text{si } z > \eta\,, \end{cases} \tag{71}$$

de modo que la componente vertical de la velocidad, v, viene dada a partir de la función de corriente ψ por la expresión

$$v(x,z) = \begin{cases} -\dfrac{\partial}{\partial x}(\psi(x,z) & \text{si } z \leq \eta \\ \\ 0 & \text{si } z > \eta\,. \end{cases} \tag{72}$$

En la simulación numérica del problema acoplado global, para la obtención del campo de velocidades para una velocidad basal, una distribución de temperaturas y un perfil conocidos, es preciso implementar fórmulas de cuadratura numérica que aproximen secuencialmente las integrales que aparecen en las expresiones de u y ψ. Para ello, se han elegido fórmulas de integración numérica de tipo trapecio compuesto, que llevan asociada una fórmula de trapecio simple en cada triángulo de la malla del problema térmico que interseca con el intervalo de integración. Para obtener v a partir de ψ se emplea una fórmula de derivación numérica con dos puntos. El carácter no estructurado de la malla utilizada supone una dificultad adicional importante a la hora de implementar informáticamente estas técnicas.

6 Modelo de Stefan-Signorini con deslizamiento basal para la temperatura

6.1 Planteamiento del modelo

En cuanto a las ecuaciones que gobiernan la evolución de la temperatura, cabe destacar que estamos interesados en el estudio de casquetes polares politérmicos, es decir, regímenes que incluyen la presencia de hielo temperado (regiones de hielo que están a la temperatura de fusión, $T = 0$). Para tener en cuenta la existencia de hielo temperado, concebido como una mezcla de hielo y agua a cero grados, se propone un modelo realista de tipo Stefan en dos fases. De este modo, el hielo temperado se asocia a la fase *pastosa* de dicho modelo. Introduciendo el modelo de Stefan de dos fases en la aproximación de hielo poco profundo descrita en la primera sección, en cada instante de tiempo t se tiene

el siguiente conjunto de ecuaciones:

$$\begin{aligned} &\frac{\partial T}{\partial t} + \vec{v}\cdot\nabla T - \beta\frac{\partial^2 T}{\partial z^2} + F = 0, \quad T < 0 \quad \text{en} \quad \Omega_C(t) \\ &T \geq 0 \quad \text{en} \quad \Omega_T(t) \\ &\beta\frac{\partial T}{\partial \vec{n}(t)} = L_c\,\frac{ds}{dt} \quad \text{en} \quad \Sigma(t) \end{aligned} \tag{73}$$

donde $\vec{v} = (u, v)$ denota el campo de velocidades y el término no lineal F está asociado a los efectos térmicos de disipación viscosa y está definido por

$$F = F(T, x, z) = -\left(\frac{\alpha}{\nu}\right)((\,\eta(x) - z\,)\,\eta_x)^{n+1} e^{\gamma T}. \tag{74}$$

En las ecuaciones anteriores intervienen los parámetros adimensionales $\beta = 0{,}12\,,\ \alpha = 0{,}3$ y $\nu = 0{,}01$, y los subconjuntos $\Omega_C(t)$ y $\Omega_T(t)$, que denotan la región de hielo frío y de hielo temperado, respectivamente. Es decir,

$$\begin{aligned} \Omega_C(t) &= \{(x,z) \in \Omega_I(t)\,/\,T(t,x,z) < 0\}, \\ \Omega_T(t) &= \{(x,z) \in \Omega_I(t)\,/\,T(t,x,z) \geq 0\}. \end{aligned}$$

Además, la superficie $\Sigma(t)$ representa la frontera móvil entre que separa la regíon fría de la temperada en cada instante t, $\vec{n}(t)$ denota el vector unitario normal a $\Sigma(t)$ apuntando hacia $\Omega_C(t)$ y $L_c = 3{,}4$ es el calor latente adimensionalizado. De este modo la tercera ecuación en (73) es la condición de Stefan en la frontera de cambio de fase, que indica que el flujo de calor se invierte en la fusión del hielo.

En la frontera inferior $\Gamma_0(t)$ se impone una condición de tipo Signorini en términos del flujo geotérmico, g_b, la velocidad de deslizamiento basal, u_b, y las tensiones basales, τ_b, dada por las expresiones siguientes:

$$\begin{aligned} -\frac{\partial T}{\partial z} &= g_b + \tau_b u_b \quad \text{si} \quad T < 0 \\ -\frac{\partial T}{\partial z} &= 0 \quad \text{si} \quad T > 0 \\ 0 < -\frac{\partial T}{\partial z} &< g_b + \tau_b u_b \quad \text{si} \quad T = 0. \end{aligned} \tag{75}$$

En la condición (75) el calor geotérmico adimensional resultante tras el escalado es $g_b = 1{,}5$ que se corresponde a una magnitud real $G_b = 5 \times 10^{-2}\, W\, m^{-2}$, valor típico para la Antártida. Dicha condición modela matemáticamente el hecho de que la energía calorífica producida por el calor geotérmico y el deslizamiento basal se invierte en derretir el hielo en contacto con la superficie

terrestre, cuando este está por debajo de la temperatura de fusión, y en calentarlo, cuando se encuentra a temperatura de fusión.

En cuanto a las expresiones para la tensión y la velocidad basal, su elección forma parte del complicado problema de modelado del deslizamiento basal de grandes masas de hielo. En una primera aproximación al modelado se pueden despreciar las tensiones y velocidades basales lo que simplifica las expresiones de (75). Bajo esta simplificación, en [7] se resuelve numéricamente el modelo térmico resultante y en [9] se proporciona un resultado de existencia para (73)-(75) en el caso $F = 0$. No obstante, en Fowler [13] se menciona la observación experimental de procesos de deslizamiento basal en zonas donde la masa de hielo está por debajo de la temperatura de fusión, aunque próxima a ella (por ejemplo, para valores del orden de $T = -4{,}6\,C$). Este hecho sugiere expresiones para la velocidad de deslizamiento que dependen de la temperatura de forma determinada. En concreto, se trata de leyes del tipo $u_b = f(\tau_b, T_b)$ en las que u_b decrece muy rápidamente cuando la temperatura basal, T_b, decrece por debajo de la de temperatura de fusión ($T = 0$). Es por ello por lo que en este trabajo proponemos una ley de la forma:

$$u_b = c_b|\tau_b|\tau_b \exp(T_b/\delta_b) \tag{76}$$

donde los parámetros c_b y δ_b se eligen verificando $c_b \in (0{,}1, 10)$ y $\delta_b << 1$. De este modo, u_b tiende rápidamente a zero en cuanto T_b disminuye por debajo del valor cero. Por otra parte, la tensión basal esta dada por

$$\tau_b(x, 0) = -\,\eta \frac{\partial \eta}{\partial x}(x)\ , \tag{77}$$

dependiendo del perfil superior y de la pendiente del mismo.

La introducción de los efectos basales supone un aspecto original importante en el presente trabajo. En esta nueva modelización del comportamiento termohidrodinámico basal, al calor geotérmico se suma la contribución del calor generado por los efectos de rozamiento asociados al deslizamiento basal. Así, en la zona de la frontera inferior que se encuentra significativamente por debajo de la temperatura de fusión sólamente el calor geotérmico contribuye a calentar el hielo de la base. En la zona *sub-temperada* de hielo próximo a la temperatura de fusión el calor geotérmico se une al generado por el rozamiento para calentar el hielo. Por último, en la zona *temperada* ambos aportes energéticos se invierten en la fusión del hielo circundante.

En cuanto al tratamiento numérico de esta compleja condición de contorno (75), en cada etapa del algoritmo global, se resuelve el problema de Stefan-Signorini para un campo de velocidades y un perfil conocido, con lo que la velocidad y tensión basal serán un dato en cada paso de tiempo. Una vez determinada la temperatura mediante el algoritmo numérico que describimos en la siguiente sección, ésta se utiliza para actualizar la velocidad basal que interviene

en el cálculo del perfil de la siguiente etapa.

Por otra parte, en la frontera en contacto con la atmósfera, $\Gamma_1(t)$, se considera la temperatura conocida y dada por

$$T(t,x,z) = T_A(t,x,z) \ \text{ en } \ \Gamma_1(t), \tag{78}$$

donde $T_A(t,x,z) = -1$, que se corresponde con una temperatura atmosférica real promedio de $223\,K$. Además, como el problema térmico se resuelve numéricamente en el dominio fijo, se considera la extensión de la temperatura superficial T_A a la región $\Omega_A(t)$:

$$T(t,x,z) = T_A(t,x,z) \text{ en } \Omega_A(t).$$

Por último, dado que el conjunto de ecuaciones (73) define un problema evolutivo, se introduce una condición inicial

$$T(0,x,z) = T_0(x,z) \ \text{ en } \ \Omega_I(t)\,. \tag{79}$$

A continuación pasamos a describir el complejo algoritmo numérico que permite obtener las temperaturas a partir de un perfil y un campo de velocidades conocido.

6.2 Algoritmo de simulación numérica

En primer lugar, previamente a la semidiscretización en tiempo del problema de Stefan-Signorini, definido por (73), (75),(78) y (79), planteamos la formulación variacional del mismo en términos de operadores multívocos. Por una parte, podemos introducir el operador multívoco de Heaviside H para expresar la condición de contorno (75) en la forma:

$$\frac{\partial T}{\partial z} \in (g_b + \tau_b u_b)\,(H(T) - 1) \Longleftrightarrow \frac{\partial T}{\partial z} + (g_b + \tau_b u_b) \in (g_b + \tau_b u_b)\,H(T). \tag{80}$$

y, por otra parte, consideramos el operador de entalpía para la clásica formulación del problema de Stefan de dos fases. Así, el operador entalpía en términos de la temperatura de fusión y el calor latente adimensionalizados está dado por

$$E(T) = \begin{cases} T & \text{si } T < 0 \\ [0,\ L_c] & \text{si } T = 0 \\ L_c & \text{si } T > 0 \end{cases}$$

Por tanto, se plantea la siguiente formulación variacional del problema de Stefan-Signorini:

Encontrar $y(t,.) \in V_A(t)$ tal que

$$\int_\Omega \frac{D^*e}{Dt}\,\varphi\,d\Omega + \int_\Omega \frac{\partial y}{\partial z}\,\frac{\partial \varphi}{\partial z}\,d\Omega + \delta \int_\Omega \frac{\partial y}{\partial x}\,\frac{\partial \varphi}{\partial x}\,d\Omega$$

$$+ \int_\Omega (F \circ \Lambda^{-1})(y)\,\varphi\,d\Omega - \int_{\Gamma_0(t)} g\,\varphi\,d\Gamma \tag{81}$$

$$+ \int_{\Gamma_0(t)} g\,\theta\,\varphi\,d\Gamma = 0, \quad \forall \varphi \in V_0(t)$$

$$e \in (E \circ \Lambda^{-1})(y) \tag{82}$$

$$\theta \in (H \circ \Lambda^{-1})(y)\,, \tag{83}$$

donde se ha introducido el cambio de variable de Kirchoff

$$y = \Lambda(T) = \int_0^T \beta\,ds = \beta T, \tag{84}$$

que permitiría la inclusión en nuestro método de casos con coeficientes de difusión no constantes, un pequeño término de difusión horizontal controlado por el parámetro δ y la notación D^* para la derivada total relativa al campo de velocidades $\vec{v}$:

$$\frac{D^*e}{Dt} = \frac{\partial e}{\partial t} + \vec{v}\,\nabla e. \tag{85}$$

Además, en la formulación anterior se considera la función $g = \beta g_b + \tau_b u_b$ en la frontera inferior, así como los siguientes conjuntos:

$$\begin{aligned} V_0(t) &= \{\varphi \in H^1(\Omega)\,/\,\varphi = 0 \ \text{ en } \ \Gamma_1(t) \cup \Omega_A(t)\} \\ V_A(t) &= \{\varphi \in H^1(\Omega)\,/\,\varphi = \Lambda(T_A) \ \text{ en } \ \Gamma_1(t) \cup \Omega_A(t)\}. \end{aligned}$$

Para la semidiscretización en tiempo de (81)-(83) se ha desarrollado un esquema de características, introduciendo la aproximación de la derivada total:

$$\frac{D^*e}{Dt}((m+1)\Delta t, x, z) \approx \frac{e^{m+1} - e^m \circ \chi^m}{\Delta t}, \tag{86}$$

donde

$$e^{m+1} = e((m+1)\Delta t, x, z)\,, \tag{87}$$

con χ^m definida por

$$\chi^m(x,z) = S((m+1)\Delta t, x, z; m\Delta t),$$

siendo S la trayectoria asociada al campo de velocidades $\vec{v}$. Dicha trayectoria se obtiene resolviendo el problema de valor final para una ecuación diferencial:

$$\begin{cases} \dfrac{dS(t,x,z;s)}{ds} &= \vec{v}(S(t,x,z;s),s) \\ S(t,x,z;t) &= (x,z). \end{cases}$$

La aproximación de la derivada total requiere la resolución del siguiente problema para obtener la temperatura en el instante t^{m+1}:

Encontrar $y^{m+1} \in V_A(t^{m+1})$ tal que:

$$\frac{1}{\Delta t}\int_\Omega e^{m+1}\,\varphi\,d\Omega - \frac{1}{\Delta t}\int_\Omega e^m \circ \chi^m\,\varphi\,d\Omega + \delta\int_\Omega \frac{\partial y^{m+1}}{\partial x}\,\frac{\partial \varphi}{\partial x}\,d\Omega$$

$$+\int_\Omega \frac{\partial y^{m+1}}{\partial z}\,\frac{\partial \varphi}{\partial z}\,d\Omega + \int_\Omega (F\circ\Lambda^{-1})(y^{m+1})\,\varphi\,d\Omega \tag{88}$$

$$+\int_{\Gamma_0(t^{m+1})} g\theta^{m+1}\,\varphi\,d\Gamma - \int_{\Gamma_0(t^{m+1})} g\,\varphi\,d\Gamma = 0, \quad \forall\varphi \in V_0(t^{m+1})$$

$$e^{m+1} \in (E\circ\Lambda^{-1})(y^{m+1}) \tag{89}$$

$$\theta^{m+1} \in (H\circ\Lambda^{-1})(y^{m+1}). \tag{90}$$

Observación 6.1 *Nótese que en la ecuación (88) la función $g = \beta g_b + \tau_b u_b$ se trata de manera explícita en la temperatura, es decir, $g = g(y^m)$.*

De este modo, para obtener y^{m+1} a partir de la formulación (88)-(90), es necesario resolver numéricamente un problema de ecuaciones en derivadas parciales fuertemente no lineal. De hecho, aparecen tres aspectos no lineales: el primero asociado a la condición de contorno de Signorini en la frontera $\Gamma_0(t)$, el segundo originado por el operador de entalpía E, que modela el carácter politérmico de la masa de hielo, y el tercero, relacionado con el aporte térmico debido a la disipación viscosa, definido por la función no lineal F. Dado que los dos primeros términos no lineales se pueden formular mediante operadores monótonos, los tratamos numéricamente mediante algoritmos de dualidad [3]. Así, introducimos las nuevas incógnitas q^{m+1} y p^{m+1} con los respectivos parámetros reales positivos ω_1 y ω_2:

$$q^{m+1} \in (H\circ\Lambda^{-1})(y^{m+1}) - \omega_1 y^{m+1} \tag{91}$$

$$p^{m+1} \in (E\circ\Lambda^{-1})(y^{m+1}) - \omega_2 y^{m+1}. \tag{92}$$

Las relaciones anteriores se pueden caracterizar mediante las identidades

$$q^{m+1} \in (H\circ\Lambda^{-1} - \omega_1 I)(y^{m+1}) \Longleftrightarrow q^{m+1} = (H\circ\Lambda^{-1})^{\omega_1}_{\lambda_1}(y^{m+1} + \lambda_1 q^{m+1})$$

$$p^{m+1} \in (E\circ\Lambda^{-1} - \omega_2 I)(y^{m+1}) \Longleftrightarrow p^{m+1} = (E\circ\Lambda^{-1})^{\omega_2}_{\lambda_2}(y^{m+1} + \lambda_2 p^{m+1}) \tag{93}$$

donde $(H\circ\Lambda^{-1})^{\omega_1}_{\lambda_1}$ denota la aproximación Yosida del operador $((H\circ\Lambda^{-1}) - \omega_1 I)$ con parámetro $\lambda_1 > 0$ y $(E\circ\Lambda^{-1})^{\omega_2}_{\lambda_2}$ denota la aproximación Yosida del operador $((E\circ\Lambda^{-1}) - \omega_2 I)$ con parámetro positivo λ_2 (ver [5], por ejemplo). Por razones de convergencia del algoritmo se han elegido verificando las relaciones

$\lambda_i\omega_i = 0{,}5,\ i = 1, 2.$

A continuación, introduciendo en (88)- (90) las nuevas incógnitas (91) y (92) y sus caracterizaciones (93), se plantea el problema:

Encontrar $y^{m+1} \in V_A$ tal que:

$$\begin{aligned} & \frac{\omega_2}{\Delta t}\int_\Omega y^{m+1}\ \varphi\ d\Omega\ +\ \delta\int_\Omega \frac{\partial y^{m+1}}{\partial x}\ \frac{\partial \varphi}{\partial x}\ d\Omega\ +\ \int_\Omega \frac{\partial y^{m+1}}{\partial z}\ \frac{\partial \varphi}{\partial z}\ d\Omega \\ & +\ \int_\Omega (F\circ\Lambda^{-1})(y^{m+1})\ \varphi\ d\Omega\ +\ \int_{\Gamma_0} g\ \omega_1\ y^{m+1}\ \varphi\ d\Gamma \\ = \ & \frac{1}{\Delta t}\int_\Omega [(E\circ\Lambda^{-1})(y^m)]\circ\chi^m\ \varphi\ d\Omega\ -\ \int_{\Gamma_0} g\ q^{m+1}\ \varphi\ d\Gamma \qquad (94) \\ & +\ \int_{\Gamma_0} g\ \varphi\ d\Gamma\ -\ \frac{1}{\Delta t}\int_\Omega p^{m+1}\ \varphi\ d\Omega, \qquad \forall\varphi \in V_0 \end{aligned}$$

$$q^{m+1} = (H\circ\Lambda^{-1})^{\omega_1}_{\lambda_1}(y^{m+1} + \lambda_1\ q^{m+1}) \qquad (95)$$

$$p^{m+1} = (E\circ\Lambda^{-1})^{\omega_2}_{\lambda_2}(y^{m+1} + \lambda_2\ p^{m+1}). \qquad (96)$$

En la ecuación (94) la no linealidad asociada a la función F se mantiene. Para resolverla se aplica el método de Newton sobre el problema discretizado por elementos finitos y se utiliza una aproximación producto del término no lineal que emplea el propio espacio de discretización. En concreto, a partir de una malla triangular ${\tau_h}^*$ del dominio, se construye el espacio V_h de elementos finitos de Lagrange lineales en cada triángulo y los conjuntos asociados:

$$\begin{aligned} V_h &= \{\varphi_h \in C^0(\bar\Omega)\ /\ \varphi_h|_P \in P_1,\ P \in \tau_h^*\} \\ V_{0h} &= \{\varphi_h \in V_h\ /\ \varphi_h = 0\ \text{ on }\ \Gamma_1(t^{m+1}) \cup \Omega_A(t^{m+1})\} \\ V_{Ah} &= \{\varphi_h \in V_h\ /\ \varphi_h = \Lambda(T_A)\ \text{ on }\ \Gamma_1(t^{m+1}) \cup \Omega_A(t^{m+1})\}. \end{aligned}$$

Denotamos por $\{w_1, \ldots, w_M\}$ la base de V_h formada por las funciones w_i definidas por las condiciones:

$$w_i(p_j) = \delta_{ij},\ j = 1, 2, \ldots, M,$$

donde $\{p_j,\ j = 1, \ldots, M\}$ es el conjunto de vértices de los triángulos de la malla. Para aproximar la integral correspondiente al término de reacción no lineal asociado a F se propone la siguiente aproximación producto de elementos finitos:

$$\int_\Omega (F\circ\Lambda^{-1})(y_h^{m+1})\ \varphi_h\ d\Omega \approx\ \sum_{j=1}^{M}\int_\Omega (F\circ\Lambda^{-1})(y_h^{m+1}(p_j))\ w_j\ \varphi_h\ d\Omega$$

que se basa en una aproximación de la función F en la forma

$$(F \circ \Lambda^{-1})(y_h^{m+1}) \approx \sum_{j=1}^{M} (F \circ \Lambda^{-1})(y_h^{m+1}(p_j))\, w_j\,.$$

De este modo, la discretización de (94) mediante la técnica de elementos finitos propuesta conduce al sistema de ecuaciones no lineales

$$\begin{aligned} &\left(\frac{\omega_2}{\Delta t} M_h + K_h^{\delta}\right) Y_h^{m+1} + M_h G(Y_h^{m+1}) + N_h G_B(Y_h^{m+1}) \\ = \;& \frac{1}{\Delta t} B_h Y_h^m - b_q^{m+1} + b_g - \frac{1}{\Delta t} b_p^{m+1}\,, \end{aligned} \tag{97}$$

donde la expresión para los diferentes vectores y matrices está dada por:

$$(M_h)_{ij} = \int_\Omega w_j w_i \, d\Omega\,, \quad (K_h^{\delta})_{ij} = \delta \int_\Omega \frac{\partial w_j}{\partial x}\frac{\partial w_i}{\partial x}\, d\Omega + \int_\Omega \frac{\partial w_j}{\partial z}\frac{\partial w_i}{\partial z}\, d\Omega\,,$$

$$(N_h)_{ij} = \int_{\Gamma_0} w_j w_i \, d\Gamma\,, \quad (B_h)_{ij} = \int_\Omega w_j \circ \chi_h^m w_i \, d\Omega\,, \quad (b_q^{m+1})_i = \int_{\Gamma_0} g q^{m+1} w_i \, d\Gamma\,,$$

$$(b_g)_i = \int_{\Gamma_0} g w_i \, d\Gamma\,, \quad (b_p^{m+1})_i = \int_\Omega p^{m+1} w_i \, d\Omega\,.$$

y se ha considerado la siguiente notación:

$$Y_h = \begin{pmatrix} y_h(p_1) \\ \vdots \\ y_h(p_M) \end{pmatrix}\,, \quad G(Y_h) = \begin{pmatrix} (F \circ \Lambda^{-1})(y_h(p_1)) \\ \vdots \\ (F \circ \Lambda^{-1})(y_h(p_M)) \end{pmatrix}$$

$$G_B(Y_h) = \begin{pmatrix} \omega_1 g y_h(p_1) \\ \vdots \\ \omega_1 g y_h(p_M) \end{pmatrix},$$

omitiendo, por simplicidad, la obvia dependencia de m.

Las igualdades (95) y (96) nos permiten resolver (97) en la iteración $m+1$ mediante un procedimiento iterativo de la siguiente forma:

- Paso 0: Inicializar $(b_q^{m+1})^0$ y $(b_p^{m+1})^0$ por ejemplo a b_q^m y b_p^m, respectivamente, los cuales han sido calculados en la anterior iteración m.

- Paso j: Calcular $(Y_h^{m+1})^j$ resolviendo el sistema no lineal

$$\begin{aligned} &\left(\frac{\omega_2}{\Delta t} M_h + K_h^{\delta}\right) (Y_h^{m+1})^j + M_h G((Y_h^{m+1})^j) + N_h G_B((Y_h^{m+1})^j) \\ = \;& \frac{1}{\Delta t} B_h Y_h^m - (b_q^{m+1})^{j-1} + b_g - \frac{1}{\Delta t} (b_p^{m+1})^{j-1}\,. \end{aligned} \tag{98}$$

- Actualizar $(b_q^{m+1})^j$ y $(b_p^{m+1})^j$ con las fórmulas

$$(q^{m+1})^j = (H \circ \Lambda^{-1})_{\lambda_1}^{\omega_1}((Y_h^{m+1})^j + \lambda_1(q^{m+1})^{j-1}) \quad (99)$$

$$(p^{m+1})^j = (E \circ \Lambda^{-1})_{\lambda_2}^{\omega_2}((Y_h^{m+1})^j + \lambda_2(p^{m+1})^{j-1}) \quad (100)$$

- Test de convergencia de la sucesión indicada por j.

Como se muestra en el anterior esquema del algoritmo, para cada paso j, debe resolverse el sistema no lineal (98). Para ello proponemos el clásico método de Newton. Con objeto de detallar la aplicación de esta técnica a nuestro caso, introducimos la función vectorial $\vec{f}$, suprimiendo la dependencia en h por simplicidad:

$$\begin{aligned}\vec{f}((Y^{m+1})^j) &= \left(\frac{\omega_2}{\Delta t}M + K^\delta\right)(Y^{m+1})^j + NG_B((Y^{m+1})^j) \\ &\quad + \sigma MG((Y^{m+1})^j) + (1-\sigma)MG((Y^{m+1})^{j-1}) \quad (101) \\ &\quad - \frac{1}{\Delta t}BY^m + (b_q^{m+1})^{j-1} - b_g + \frac{1}{\Delta t}(b_p^{m+1})^{j-1}\end{aligned}$$

con σ parámetro real tal que $0 \leq \sigma \leq 1$. Así, la elección de $\sigma = 1$ corresponde a un esquema implícito mientras que $\sigma = 0$ corresponde a un esquema explícito. Además, el sistema (98) en cada paso j se puede escribir como

$$\vec{f}((Y^{m+1})^j) = \vec{0}$$

de forma que las iteraciones de Newton se plantean como sigue:

- Paso 0: Inicializar $(Y^{m+1})_0^j$ (por ejemplo a $(Y^{m+1})^{j-1}$).

- Paso $s+1$: Calcular $(Y^{m+1})_{s+1}^j$ resolviendo el sistema lineal

$$D\vec{f}((Y^{m+1})_s^j)(Y^{m+1})_{s+1}^j = D\vec{f}((Y^{m+1})_s^j)(Y^{m+1})_s^j - \vec{f}((Y^{m+1})_s^j), \quad (102)$$

donde $D\vec{f}(Y_s)$ denota la matriz jacobiana de la función vectorial $\vec{f}$ en Y_s, dada por

$$D\vec{f}(Y_s) = \left(\frac{\omega_2}{\Delta t}M + K^\delta\right) + N\,DG_B(Y_s) + \sigma M\,DG(Y_s)$$

con las matrices diagonales $DG(Y_s)$ y $DG_B(Y_s)$ definidas por

$$DG(Y_s) = \begin{pmatrix} [D(F \circ \Lambda^{-1})](y_s(p_1)) & & \\ & \ddots & \\ & & [D(F \circ \Lambda^{-1})](y_s(p_M)) \end{pmatrix}$$

$$DG_B(Y_s) = \begin{pmatrix} \omega_1 g & & \\ & \ddots & \\ & & \omega_1 g \end{pmatrix}$$

y el vector $\vec{f}((Y^{m+1})_s^j)$

$$\begin{aligned}\vec{f}((Y^{m+1})_s^j) &= \left(\frac{\omega_2}{\Delta t}M + K^\delta\right)(Y^{m+1})_s^j + NG_B((Y^{m+1})_s^j) \\ &\quad + \sigma MG((Y^{m+1})_s^j) + (1-\sigma)MG((Y^{m+1})^{j-1}) \\ &\quad - \frac{1}{\Delta t}BY^m + (b_q^{m+1})^{j-1} - b_g + \frac{1}{\Delta t}(b_p^{m+1})^{j-1}.\end{aligned} \tag{103}$$

Finalmente, el sistema lineal (102) resultante en cada etapa se resuelve mediante el método del doble gradiente conjugado con precondicionamiento (ver Joly [18] o Golub-Meurant [16] para los detalles). Dicha elección está motivada por no tener garantizado que la matriz del sistema lineal sea definida y bien condicionada.

Observación 6.2 *Un esquema totalmente implícito para (88)-(90) puede verse en [12]. El aumento de complejidad en la formulación se ve compensado con la reducción significativa del tiempo de cálculo (del orden de* $30-40\,\%$*).*

Algoritmo para la simulación numérica

En esta sección presentamos brevemente un esquema del algoritmo numérico que hemos desarrollado para la simulación numérica del problema acoplado asociado a un modelo completo de hielo poco profundo. El objetivo del método numérico es calcular el perfil, el campo de velocidades, las magnitudes basales y la distribución de temperatura de la masa de hielo. Para ello, se resuelven secuencialmente los modelos descritos en los apartados anteriores. En concreto, el pseudocódigo del algoritmo es el siguiente (ver [10, 11], para los detalles):

1. **Paso** 0**:**
   - Mallado de los dominios fijos para los problemas del perfil, $[-1,1]$, y de la temperatura, Ω.
   - Inicialización de la velocidad basal, $u_b^0 = 0$, y de la temperatura, $T^0 = T_0 = -1$.
2. **Paso** $m+1$**:** Cálculo de magnitudes en tiempo $t^{m+1} = (m+1)\Delta t$.
   - A partir de u_b^m y T^m, se resuelve (48) para calcular el perfil superior $\eta^{m+1}(x)$ y las fronteras de la masa de hielo $S_-(t^{m+1})$ y $S_+(t^{m+1})$.
   - Se determinan los conjuntos $\Omega_I(t^{m+1})$, $\Omega_A(t^{m+1})$, $\Gamma_0(t^{m+1})$ y $\Gamma_1(t^{m+1})$ mediante (50) y (51).

- Se calculan las velocidades u^{m+1} y v^{m+1} con las expresiones (39) y (72).
- Se obtiene y^{m+1} y, con ello, la temperatura T^{m+1} resolviendo (81)-(84).
- Se actualizan las magnitudes basales, u_b^{m+1} y τ_b^{m+1} , mediante las expresiones (76) y (77).

7 Aplicación: simulación del casquete polar antár- tico

El algoritmo numérico descrito en la sección anterior se ha aplicado sobre diferentes ejemplos con datos reales del problema acoplado termomecánico. A continuación recogemos uno de ellos, con el objeto de reflejar las propiedades cualitativas de la solución del problema global.

Para la resolución numérica se ha considerado una malla uniforme del intervalo $[-1,1]$ con $N = 2001$ nodos, así como una malla triangular no estructurada del dominio $\Omega = [-1,1] \times [0, z_{max}]$ con 9706 triángulos y 4977 nodos. Se ha tomado el valor $z_{max} = 1{,}2$ porque garantiza la inclusión $\Omega_I(t) \subset \Omega$ en cualquier instante t. Además se ha refinado localmente la malla para conseguir mejores aproximaciones en zonas específicas: zona basal y perfil superior (ver Fig. 7).

Las constantes físicas adimensionalizadas toman los siguientes valores:

$$\gamma = 11{,}3,, \quad \nu = 0{,}01, \quad \alpha = 0{,}3, \quad \beta = 0{,}12, \quad \delta = 0,$$

$$L_c = 3{,}4,, \quad g = 0{,}18, \quad T_A(t,x,z) = -1.$$

Además, como es clásico en glaciología, se considera el exponente de la ley de Glen $n = 3$.

Se ha considerado el perfil inicial definido por

$$\eta_0(x) = 0{,}5(1 - |x|^{4/3})^{3/8}\,, \quad x \in [-1,1].$$

La temperatura inicial que consideramos es de 50^oC bajo cero en la región atmosférica situada por encima de la masa de hielo y la mitad en el interior de la misma (este dato no es muy relevante ya que se busca el estado estacionario). Es decir, se ha tomado

$$T(0,x,z) = \begin{cases} -1 & \text{si } z \geq \eta_0(x) \\ \\ -0{,}5 & \text{si } z < \eta_0(x) \end{cases}$$

Al partir inicialmente de hielo muy por debajo de su temperatura de fusión en la zona en contacto con la superficie, se ha elegido lógicamente una velocidad inicial de deslizamiento nula. Además, se ha elegido la función de acumulación adaptada de la propuesta que figura en [14]:

$$a(t,x) = \begin{cases} \dfrac{a_0(1-1{,}5\,|x|^{\frac{4}{3}})}{\left(1-|x|^{\frac{4}{3}}\right)^{\frac{5}{8}}} & \text{si } 0 \le |x| < L-\varepsilon \\ \dfrac{a_0(1-1{,}5\,(L-\varepsilon)^{\frac{4}{3}})}{\left(1-(L-\varepsilon)^{\frac{4}{3}}\right)^{\frac{5}{8}}} & \text{si } (L-\varepsilon) \le |x| \le L \end{cases}$$

con $L = 1$, $\varepsilon = 10^{-4}$ y $a_0 = 1{,}24 \times 10^{-5}$. Esta función de acumulación es más realista y más suave que la adaptada por Paterson, pero incluso para el problema del perfil aislado no se puede encontrar una solución exacta. Por esta última razón se ha preferido la de Paterson para validar el método numérico del perfil.

Para el cálculo de la velocidad, los valores de las constantes que aparecen en (76) se han ajustado a $c_b = 0{,}1$ y $\delta_b = 0{,}001$.

En las Figuras 8 y 9 se muestran el campo de velocidades y las isolíneas de temperatura, respectivamente, que se han obtenido para el tiempo $t = 50$. La elección de la respuesta numérica en este tiempo pretende poner de manifiesto el aspecto cualitativo de la velocidad y las fronteras libres del espesor del hielo $S_-(t)$, $S_+(t)$ (contracción masa de hielo) y $\eta(t)$ (separación hielo-atmósfera) y la frontera libre del problema térmico que representa la zona de hielo temperado. En la figura de la temperatura se observa como la isoterma asociada a la temperatura de fusión (hielo temperado) es una línea adherida a la base del casquete polar. La gráfica de los campos de velocidades da una idea del desplazamiento que se está produciendo en el hielo en los distintos puntos.

En las Figuras 10 y 11 se muestran la velocidad y la temperatura basal obtenidas para los tiempos $t = 25$, $t = 50$ y $t = 75$. Tras una comparación entre ambas figuras, se puede comprobar que el modelo propuesto permite deslizamiento en zonas de la base con hielo por debajo de la temperatura de fusión, tal y como se ha detectado experimentalmente. En la gráfica de la velocidad basal los valores negativos a la izquierda y positivos a la derecha conllevan un deslizamiento cuyo sentido va siempre desde el interior hacia los márgenes de la masa de hielo.

En el análisis cualitativo del conjunto de las figuras se puede observar la esperada simetría de las magnitudes calculadas con respecto al centro del casquete polar.

Para facilitar la realización de simulaciones numéricas en ordenador se ha

elaborado la aplicación informática **GLANUSIT** (**GL**aciology **NU**merical **SI**mulation **T**oolbox) [22]. Esta herramienta informática, realizada en entorno gráfico de MATLAB, permite una cómoda entrada de datos reales (a partir de los cuales se hace el escalado de tipo hielo poco profundo), la ejecución de los programas FORTRAN originales que implementan los algoritmos numérico descritos y la visualización numérica, gráfica y animada de las distintas magnitudes calculadas. En la Figura 12 se muestra una típica pantalla de datos y en la Figura 13 se presenta un ejemplo de pantalla de resultados.

Por último, es conveniente señalar que los modelos propuestos y el algoritmo señalado permiten obtener una importante información cualitativa sobre la respuesta de las grandes masas de hielo, como la Antártida, frente a las magnitudes que influyen en el comportamiento termoelastohidrodinámico de las mismas. Tal vez un preprocesado más refinado de las medidas experimentales que proporcionan estas magnitudes (medidas térmicas en distintos puntos de la superficie, mediciones de acumulación y ablación más precisas, inclusión de variaciones estacionales, variación espacial del calor geotérmico, perfiles de la base no planos, etc.) permitirían mejorar algunos de los datos de entrada del modelo y, con ello, los resultados en el aspecto cuantitativo (no tanto en lo cualitativo). No obstante, la toma en consideración de dichas mejoras de datos no supone la necesidad de ningún cambio en el trabajo realizado debido a la generalidad del modelo y a la versatilidad del algoritmo de simulación.

Referencias

[1] R.A. Adams, *Sobolev Spaces*, Vol. 65 de *Pure and Applied Mathematics Series*, Academic Press, New York, (1975).

[2] M. Bercovier, O. Pironneau and V. Sastri, Finite elements and characteristics for some parabolic-hyperbolic problems, *Appl. Math. Model.*, 89-96 (1983).

[3] A. Bermúdez, C. Moreno, Duality methods for solving variational inequalities. *Comp. Math. with Appl.*, **7**, (1981), 43–58.

[4] H. Brezis, *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert*, North Holland, Amsterdam, (1973).

[5] H. Brezis, *Analyse functionnelle. Théorie et applications*, Masson, Paris, (1983).

[6] N. Calvo, J.I. Díaz, J. Durany, E. Schiavi, C. Vázquez, On a doubly nonlinear parabolic obstacle problem modelling ice sheet dynamics, *SIAM J. of App. Math.*, **63**, (2002), 683-707.

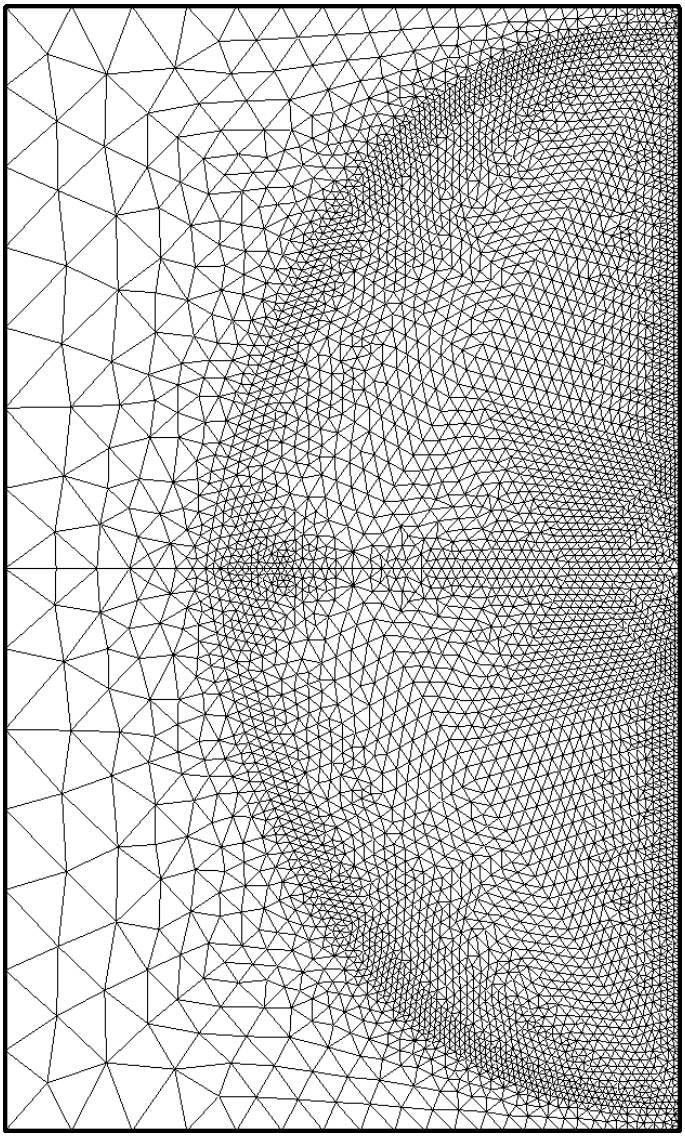


Figura 7: Malla de elementos finitos del dominio Ω con 9706 triángulos y 4977 nodos.

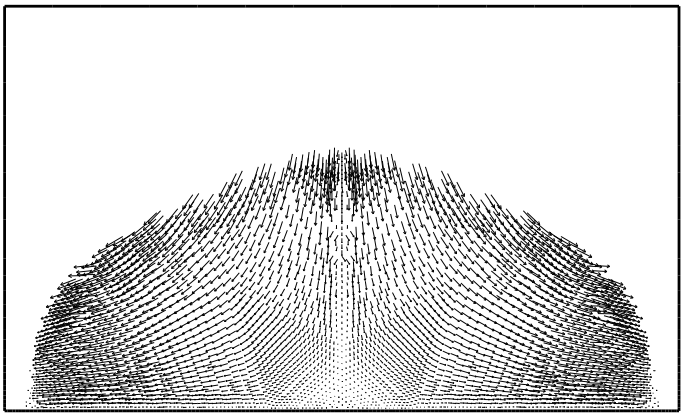


Figura 8: Campo de velocidades en $t = 50$.

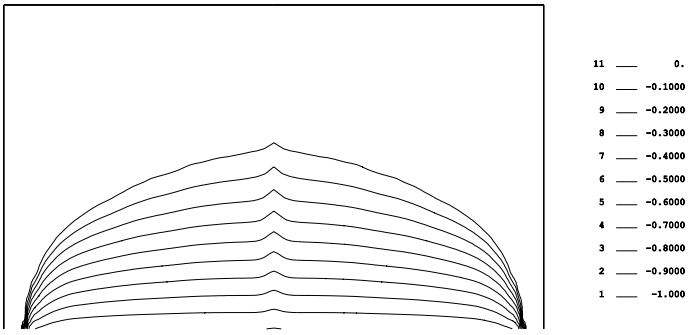


Figura 9: Isotermas en $t = 50$.

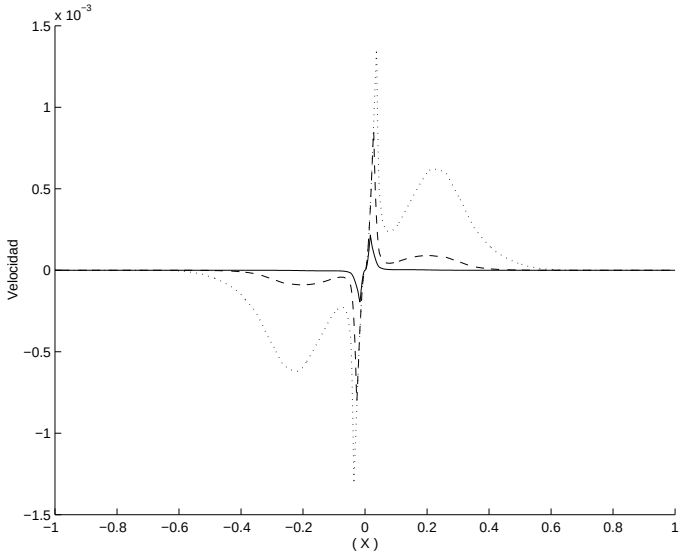


Figura 10: Velocidad basal: $t = 25(-)$, $t = 50(--)$, $t = 75(\cdots)$.

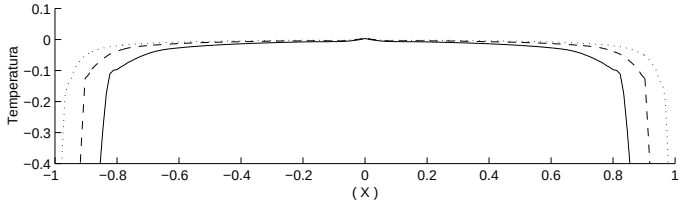


Figura 11: Temperatura basal: $t = 25(-)$, $t = 50(--)$, $t = 75(\cdots)$.

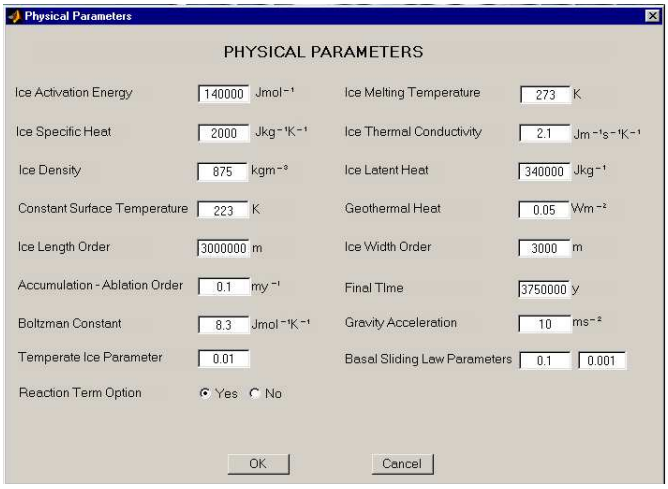


Figura 12: Ejemplo de entrada de datos reales con GLANUSIT.

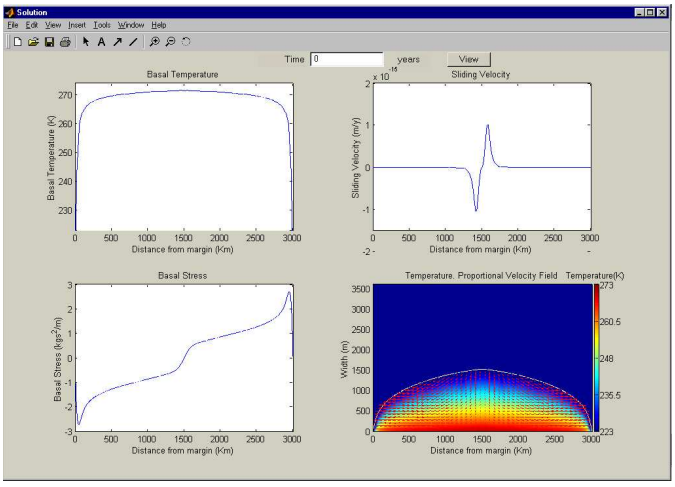


Figura 13: Ejemplo de presentación de magnitudes reales con GLANUSIT.

[7] N. Calvo, J. Durany, C. Vázquez, Numerical approach of thermomechanical coupled problems with moving boundaries in theoretical glaciology, *Numerische Mathematik.*, **83**, (1999), 557-580.

[8] N. Calvo J. Durany C. Vázquez, Numerical computation of ice sheet profiles with free boundary problems, *Appl. Numer. Math.*, **35**, (2000), 111-128.

[9] N. Calvo J. Durany C. Vázquez, Mathematical analysis of a Stefan problem with Dirichlet-Signorini boudary conditions appearing in polithemic ice sheet modelling, *J. Math. Anal. and Appl.*, **262**, (2001), 577-600.

[10] N. Calvo, J. Durany, C. Vázquez, Numerical approach of thermomechanical coupled problems with moving boundaries in theoretical glaciology, *Math. Mod. and Meth. in App. Sci.*, **32**, 2, (2002), 229-248.

[11] N. Calvo J. Durany C. Vázquez, Finite elements numerical solution of a coupled-profile-velocity-temperature shallow ice sheet approximation, *J. Comput. and Appl. Math.*, **158**, (2003), 31-41.

[12] N. Calvo J. Durany C. Vázquez, Un esquema numérico implícito para un modelo térmico de Stefan-Signorini en grandes masas de hielo, *Actas del XVII CEDYA-VIII Congreso de Matemática Aplicada* (2003), CD-ROM.

[13] A.C. Fowler, Sub-temperate basal sliding, *J. Glaciology*, **32**, 110, 3-5 (1986).

[14] A.C. Fowler, Modelling ice sheet dynamics, *Geophys. Astrophys. Fluid Dynamics*, **63**, 29-65 (1992).

[15] A.C. Fowler, *Mathematical Models in the Applied Sciences*, Cambridge University Press, Cambridge, (1997).

[16] G. Golub, G. Meurant, *Résolution numérique des grands systèmes linéaires*, Ed. Eyrolles, (1984).

[17] P. Huybrechts, *The Anctartic ice sheet and environemental change: a three dimensional modelling study*, Ph.D. Thesis, University of Brussels, (1991).

[18] P. Joly, Résolution de systémes linéaires non symétriques par des méthodes de gradient conjugué, Publication du Laboratoire d´Analyse Numérique, Paris VI, (1982).

[19] W.S.B. Patterson, *The Physics of Glaciers*, Pergamon, (Oxford), (1981).

[20] K. Hutter, *Theoretical Glaciology*, Reidel, Dordrecht, (1981).

[21] O. Pironneau, On the transport-diffusion algorithm and its application to Navier-Stokes equation, *Numer. Math.*, **38**, 309-332 (1982).

[22] R. Toja, *GLANUSIT: Aplicación informática para el pre y postprocesado en simulación numérica de grandes masas de hielo*, PFC Ingeniería Informática, Universidade da Coruna, (2003).

Bol. Soc. Esp. Mat. Apl. nº 28(2004), 51–70

Codificación de información mediante códigos de barras*

L. Hernández[1] y A. Martín[2]

[1]Departamento de Tratamiento de la Información y Codificación,
Instituto de Física Aplicada,
Consejo Superior de Investigaciones Científicas
[2]Departamento de Matemática Aplicada,
Universidad de Salamanca

luis@iec.csic.es, delrey@usal.es

Resumen

Se presentan en este artículo varios sistemas de codificación de información mediante códigos de barras. En particular se detalla la forma de elaborar el código de barras EAN13, posiblemente el más utilizado en la actualidad, y su relación con otro de los códigos más empleados: el ISBN para la catalogación de libros. Se incluyen y comentan los procesos matemáticos empleados para elaborar tales códigos.

Palabras clave: *Codificación de la información, códigos de barras, códigos detectores de errores, EAN13, ISBN.*

Clasificación por materias AMS: *94A15, 94B05, 94B60*

1 Introducción

El desarrollo tecnológico actual, la proliferación de los ordenadores y periféricos, así como la facilidad para el establecimiento de redes locales y su rapidez de acceso, ha permitido el desarrollo de nuevas aplicaciones matemáticas e informáticas a diferentes campos de la vida cotidiana. Una de ellas es la codificación de objetos de toda índole; en la que además de las herramientas informáticas que agilizan el tratamiento de la información, intervienen diferentes aspectos matemáticos, como son los algoritmos utilizados en la elaboración de los códigos y la verificación de que éstos son leídos correctamente.

* Parcialmente subvencionado por el Ministerio de Ciencia y Tecnología, TIC2001–0586.
Fecha de recepción: 23 de julio de 2003

Se denomina *código* a todo sistema de signos o señales y reglas que cambia la forma de un mensaje. Esta codificación se lleva a cabo de forma estándar y no secreta, es decir, su finalidad es la de resumir determinados datos y permitir una manipulación electrónica de los mismos. Desde este punto de vista, no es objetivo de un código impedir que la información que contiene sea accesible a personas diferentes de quienes elaboraron dicho código. No se trata pues de códigos secretos, que entran en el campo de la criptografía (para temas criptográficos véanse, por ejemplo, [10, 14]).

En particular, los códigos de barras que se presentan en este artículo son *no significativos* en el sentido de que no proporcionan más información que la contenida en el propio código, es decir, no permiten ocultar información, dado que su finalidad no es otra que la de transformar la información constituida por caracteres en una información gráfica, que puede ser tratada por medios informáticos. Por este motivo, no es conveniente que información confidencial se codifique por medio de códigos de barras.

Hoy en día, la mayor parte de los productos manufacturados, desde libros a latas de tomate, pasando por prendas de vestir, medicamentos, o paquetes de envío urgente, llevan una etiqueta con deteminados símbolos o barras, que codifican información relativa a dicho artículo y que permiten identificarlo de forma unívoca. Así, se entiende por *código de barras* a un conjunto de líneas y números asociados a ellas, que va impreso en los productos de consumo y que se utiliza para su gestión informática. La información contenida en el código de barras hace referencia a datos relevantes del artículo, como el país de fabricación, su tamaño, propiedades, precio, etc. Estos datos son accesibles por medio de un lector óptico que "lee" el contenido del código mediante un rayo láser. La "lectura" es transformada por el software correspondiente y manipulada conforme a determinados requerimientos informáticos. El código leído es enviado a una base de datos que responde con el nombre del artículo, su precio y otros datos. En el caso de una venta, el artículo es dado de baja en el almacén, con lo que es posible gestionar las ventas diarias, el stock, etc.

En el presente trabajo se comentan diferentes métodos de codificación mediante códigos de barras. En §2 se presentan algunas definiciones básicas sobre códigos en su sentido matemático. En la sección 3 se hace un breve repaso a los códigos de barras más empleados. El caso particular del código de barras EAN13 se comenta con detalle en la sección 4. Finalmente, en §5 se presenta el código ISBN, su relación con el EAN13 y algunas de sus propiedades para detectar determinados tipos de errores.

2 Definiciones y Notación

Se llama *alfabeto* a un conjunto finito de q símbolos, $\Sigma_q = \{s_1, s_2, \ldots, s_q\}$. Los elementos (vectores) de Σ_q^n se denominan *palabras de longitud* n, y se llama *código de longitud* n a todo subconjunto $\mathcal{C} \subset \Sigma_q^n$. Los elementos de $\mathcal{C}$ se conocen como *codewords* o *palabras codificadas* (para un estudio más detallado véanse, por ejemplo, [5, 9, 15]). Si $q = 2$, los códigos se llaman *códigos binarios*.

La *distancia de Hamming* entre dos vectores $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ de Σ_q^n se define como el número de posiciones en las que difieren dichos vectores y viene dada por:

$$d(x, y) = |\{i \colon 1 \leq i \leq n \text{ y } x_i \neq y_i\}|\,.$$

Dado un código $\mathcal{C}$, se define su *distancia mínima*, d, por la expresión

$$d = \min\{d(x, y) \colon x, y \in \mathcal{C} \text{ y } x \neq y\}.$$

Si $\mathcal{C}$ tiene k elementos, $|\mathcal{C}| = k$, el código $\mathcal{C}$ suele llamarse un (n, k, d)-*código*. Los códigos con distancia mínima $d \geq 2$ pueden ser empleados para detectar $d - 1$ errores (*códigos detectores de errores*). Sea $d = 2e + 1 \geq 3$, entonces $\mathcal{C}$ puede utilizarse para corregir errores (*código corrector de errores*), de modo que puede corregir $e = (d - 1)/2$ errores. Si $d = 2e \geq 4$, el código podrá corregir $e-1 = (d-2)/2$ errores. La capacidad de detectar y de corregir errores depende del valor de d y del algoritmo que se utilice. Así, un código con distancia $d = 5$, por ejemplo, puede utilizarse para detectar 4 errores, para corregir 2 errores, o para detectar 3 y corregir 1, etc.

Se define la *tasa de información* de un código $\mathcal{C}$ como la razón entre la cantidad de información significativa de cada palabra y la longitud de cada palabra. Esta tasa puede determinarse mediante la expresión:

$$R = \log_q |\mathcal{C}|/n. \tag{1}$$

El concepto de tasa de información es natural en el sentido de que para codificar 4 palabras usando un alfabeto binario, bastaría con emplear las palabras 00, 01, 10 y 11; pero si se emplea un código de 4 palabras de longitud 3, entonces, por la expresión (1), su tasa de información sería 2/3. Se llama *peso de Hamming* de un vector $x \in \Sigma_q^n$, y se representa por $w(x)$, al número de elementos no nulos que contiene, es decir, a la distancia de x al vector 0. Dado que la distancia entre dos vectores, x, y, es el número de posiciones en las que difieren, se tiene $d(x, y) = w(x - y)$. El *peso mínimo* de un código es el mínimo de los pesos de Hamming de las palabras no nulas de dicho código.

Los ejemplos más útiles de alfabetos son $\Sigma_q = \mathbb{Z}_q$ y, si q es la potencia de un número primo, el cuerpo $\Sigma_q = \mathbb{F}_q$ (recuérdese que $\mathbb{Z}_q$ es un cuerpo si y sólo si q es primo). Un código $\mathcal{C}$ es un *código lineal* sobre $\mathbb{F}_q$ si es un subespacio vectorial de $\mathbb{F}_q^n$, y se dice que es un $[n, k]$-*código* si dicho código tiene dimensión k. Si, además, su mínima distancia es d, entonces se llama un $[n, k, d]$-*código*. Es claro que la tasa de información de un $[n, k, d]$-código es $R = k/n$. Además como los códigos lineales son subespacios vectoriales, este peso mínimo coincide con la distancia mínima del código.

Para ilustrar las anteriores definiciones, en el Cuadro 1 se presentan dos ejemplos de códigos binarios.

$\mathcal{C}_1$ es un código binario de longitud 5 y se forma añadiendo un bit de paridad al final de cada una de las palabras de $\mathbb{F}_2^4$; es decir, al final de cada palabra se añade el bit 0 si la suma de sus bits es par, y el 1 en caso contrario. Claramente

Palabra	Código $\mathcal{C}_1$	Código $\mathcal{C}_2$
0000	00000	0000000
0001	00011	1010101
0010	00101	0110011
0011	00110	1100110
0100	01001	0001111
0101	01010	1011010
0110	01100	0111100
0111	01111	1101001
1000	10001	1111111
1001	10010	0101010
1010	10100	1001100
1011	10111	0011001
1100	11000	1110000
1101	11011	0100101
1110	11101	1000011
1111	11110	0010110

Cuadro 1: Ejemplos de dos códigos lineales binarios

este código puede detectar hasta un bit de error, pero no puede corregir errores dado que su distancia mínima es 2. $\mathcal{C}_1$ es un $[5, 4, 2]$-código binario. Por su parte, el código $\mathcal{C}_2$ es de longitud 7 y distancia mínima 3. Dado que es un $[7, 4, 3]$-código, es capaz de corregir hasta un bit de error.

Esta ganancia a la hora de corregir errores y no sólo de detectarlos tiene un precio. En este caso, según la expresión (1), la tasa de información de $\mathcal{C}_2$ es $4/7$; mientras que la tasa correspondiente a $\mathcal{C}_1$ es $4/5$. Como es de suponer, los códigos que atraen más atención son aquellos que poseen una alta tasa de información y proporcionan, además, una alta capacidad para corregir errores.

Los códigos de barras que se presentarán de forma detallada en este artículo son códigos detectores de errores puesto que en su elaboración se incluye un dígito de control que permite llevar a cabo tal detección; pero no permiten corregir errores y no son códigos lineales.

La combinación de dos códigos con distancia $d = 2$ permite obtener códigos con distancia mayor; por lo que pueden corregir errores. Es el caso del código binario del Servicio Postal americano utilizado para indicar el código postal en las cartas, llamado *código Postnet*. El hecho de que permita corregir errores se debe a que se puede determinar la posición en la que se ha producido el error. El código Postnet está formado por dos códigos de distancia $d = 2$: uno decimal de longitud 10 y otro binario de longitud 5 (que permite escribir cada dígito mediante barras). Este segundo código se muestra en el Cuadro 2, donde los ceros se representan por medio de barras cortas y los unos mediante barras largas. Ambos códigos se combinan para obtener un código binario de longitud 50 y distancia $d = 4$.

Palabra	Código Postnet	Representación gráfica
1	00011	
2	00101	
3	00110	
4	01001	
5	01010	
6	01100	
7	10001	
8	10010	
9	10100	
0	11000	

Cuadro 2: Representación binaria y gráfica del Código Postnet

Cada código Postnet está formado por 9 dígitos, $c_1, \ldots, c_9$, a los que se añade un dígito de control, c_{10}, calculado de modo que su suma con los 9 precedentes sea divisible entre 10. A la hora de imprimir el código Postnet se insertan dos barras largas al inicio y al final, que hacen el oficio de separadores. Se puede observar, además, que cada codeword tiene 3 barras cortas y 2 barras largas, lo que permite detectar si se ha producido un error a la hora de imprimirlo o leerlo. En efecto, si se produce un error y un dígito estuviera representado por 3 barras largas o sólo por una, se sabría de la existencia del error y de su posicición —la razón de representar cada dígito mediante 5 barras (o bits) y no con 4 es que en este caso hay exactamente 10 posibles combinaciones de 2 barras largas (unos) y 3 barras cortas (ceros)—. Si el error se ha producido en la posición i-ésima, y $x = c_i$ fuera el dígito desconocido, para corregir el error bastaría con calcular

$$x \equiv -\left(c_1 + \cdots + c_{i-1} + c_{i+1} + \cdots + c_{10}\right) \mod 10.$$

Por ejemplo, si la impresión errónea del código Postnet 8765432195 fuera

,

fácilmente se detectaría que el error se ha producido en el dígito que ocupa la quinta posición, dado que en dicha posición (eliminadas las dos barras laterales del código) aparecen 4 barras cortas y sólo una larga. Ahora bien, si x es el dígito que debería ocupar dicha posición, el error se corrige calculando

$$x \equiv -(8+7+6+5+3+2+1+9+5) \mod 10 = -46 \mod 10 = 4.$$

3 Códigos de barras unidimensionales

La codificación de artículos más extendida se lleva a cabo mediante los llamados códigos de barras unidimensionales. Estos códigos ocupan, en realidad, dos dimensiones; sin embargo se llaman unidimensionales debido a que son leídos

por un láser lineal, y basta con leer cualquier línea transversal al código, dado que todas ellas son iguales. No obstante, existen diferentes formas de llevar a cabo esta codificación. En general, la elección de un determinado tipo de código de barras depende de la aplicación para la que se desee utilizar.

Los códigos de barras suelen tener dos representaciones: una determinada cantidad de dígitos o caracteres en su parte inferior (no siempre presente) y una parte gráfica formada por unas barras verticales de diferente grosor y separadas por unos espacios paralelos. Los códigos codifican determinado *juego de caracteres*, es decir, un conjunto específico de letras, números y símbolos. Todo código de barras posee unos elementos característicos. Así, los *separadores de inicio y de fin* de cada código son combinaciones específicas de barras y espacios que indican al lector óptico dónde empieza y dónde termina el código a leer. Los códigos se llaman *bidireccionales* si pueden ser leídos tanto de derecha a izquierda como de izquierda a derecha. A continuación se presentan los códigos de barras lineales más utilizados en la actualidad, así como una breve descripción de cada uno de ellos (para más detalles ver [2, 3, 16]).

El *Código* 39 (Código 3 de 9) es un código de longitud variable, adecuado para codificar datos alfanuméricos de carácter general. Cada carácter se representa mediante cinco barras y cuatro espacios, con sólo dos anchuras posibles: gruesa y fina (ver Figura 1(a)). El *Código* 2 *de* 5 (Intercalado 2 de 5, ITF ó I-2/5) es un sistema de codificación (véase la Figura 1(b)) de propósito general para datos numéricos.

Figura 1: Ejemplos de Código 39 y Código 2 de 5

Codabar es un código para la codificación de números que incluye algunos caracteres especiales (ver Figura 2(a)). Dispone de 4 separadores de inicio/fin, que pueden llevar información adicional. El *Código* 128, más condensado que el Código 39, permite codificar 128 caracteres ASCII (se pueden utilizar diferentes juegos de caracteres para codificar caracteres extendidos que no pertenezcan al inglés). Cada carácter se representa mediante 3 barras y 3 espacios, que pueden tener cuatro anchuras diferentes (ver Figura 2(b)).

Figura 2: Ejemplos de Codabar y de Código 128

Existen otros códigos, como el *Código* 93 que ofrece una densidad de información alfanumérica mayor que la que ofrecen los Códigos 39 y 128, incluyendo el código ASCII completo. Por su parte, el *Código de canal* es una familia de símbolos de barras unidimensionales diseñada para codificar cadenas de entre 2 y 7 dígitos con la menor longitud de símbolo posible.

Hay muchas aplicaciones en las que no se pueden emplear los códigos de barras presentados hasta ahora. Es el caso, por ejemplo, de la codificación que precise de los 256 caracteres ASCII de forma simultánea. Por este motivo se ha desarrollado la denominada *simbología* 93*i*, que codifica bytes y permite, por tanto, codificar todos los caracteres pertenecientes al código ASCII completo. La simbología 93i ([2]) extiende la del Código 39, incorpora los 65536 caracteres Unicode, su longitud es variable y utiliza 2 caracteres para comprobar los símbolos ó 6 códigos Reed-Solomon correctores de errores ([15]).

Dado que en ocasiones se requiere que el código de barras no supere determinada superficie, se ha desarrollado otra simbología que permite llevar a cabo este objetivo y se conoce como *Simbología para reducir espacio* o RSS (Reduced Space Symbology). Esta simbología consta de tres familias diferentes: *RSS*-14, *RSS Limitada* y *RSS Expandida*, que pueden ser utilizadas por el sistema EAN.UCC (European Article Numbering [7] y Uniform Code Council [17], respectivamente). Existe otra versión de esta simbología, conocida como *RSS*-14 *Apilada*, que corta un RSS-14 demasiado ancho y lo apila en dos filas.

4 El código de barras EAN13

Sin duda, el código de barras más extendido en el mundo es el llamado *EAN13* (existe una versión reducida y menos utilizada llamada *EAN8*). Este código data de 1977 y es para Europa el análogo al código UPC de Estados Unidos (Universal Product Code, [17]) o al código JAN de Japón (Japanese Article Number [13]). El EAN13 es una extensión del UPC, por lo que ha heredado algunas de sus características y arbitrariedades, como se mencionará más adelante. El nombre del código, EAN, procede de las iniciales de European Article Numbering ([7]), mientras que el 13 hace referencia al número de dígitos que van impresos en la parte inferior del código, de los que las barras son su representación gráfica. Como ya se mencionó, el EAN13 es *no significativo*, es decir, no proporciona ninguna información adicional que no esté contenida en los 13 dígitos que lo acompañan.

4.1 Determinación de los dígitos

Los 13 dígitos del código EAN13, $c_1, \ldots c_{13}$, se calculan como sigue:

1. *Código del país*: Son los dos primeros dígitos (algunos países utilizan 3), c_1 y c_2, y son asignados por la organización nacional del sistema EAN a la que se ha adscrito la empresa que elabora el producto (ver [1]). Por ejemplo, los dos dígitos para España son 84, para el Reino Unido, 50, mientras que Alemania usa del 400 al 440.

2. *Código de la empresa*: Los siguientes 5 u 8 dígitos, $c_3, \ldots c_{10}$, se reservan para las diferentes empresas registradas en cada país.
3. *Código del producto*: Son los dígitos que restan hasta 12 (entre 2 y 5), $c_8, \ldots, c_{12}$, y están a disposición del propietario de la marca.
4. *Dígito de control*: El último dígito, c_{13}, se calcula a partir de los 12 dígitos anteriores y permite decidir si el código de barras se ha leído correctamente. Su expresión simplificada es:

$$c_{13} = -\sum_{j=1}^{6} (c_{2j-1} + 3c_{2j}) \mod 10. \tag{2}$$

Una vez que los 13 dígitos han sido determinados, cada uno de ellos es representado mediante una colección de barras y espacios verticales, y todo ello constituye el código EAN13.

4.2 Fundamentos matemáticos

Antes de proceder a la determinación de las barras y espacios del código, se justificarán las razones matemáticas de la elección del dígito de control dado en la expresión (2).

En general, el cálculo de un dígito de control, a_n, para una colección de $n-1$ dígitos dados: $a_1, \ldots, a_{n-1}$, se lleva a cabo de manera que una operación sobre todos los dígitos de identificación: $a_1, \ldots, a_{n-1}, a_n$, permita decidir si el código ha sido leído correctamente. Para ello se considera el conjunto de los enteros módulo k, $\mathbb{Z}_k$, y n aplicaciones $\sigma_i \colon \mathbb{Z}_k \to \mathbb{Z}_k$, y se supone que el código se ha leído correctamente si y sólo si se verifica la siguiente condición:

$$\sum_{i=1}^{n} \sigma_i (a_i) \equiv 0 \mod k.$$

(Nótese que la elección del 0 en la congruencia anterior es arbitraria, es decir, no hay ninguna razón que impida elegir cualquier otro entero de $\mathbb{Z}_k$). La secuencia $\sigma_1, \ldots, \sigma_n$ se denomina *esquema de control de dígitos*. Normalmente, las aplicaciones σ_i se definen por $\sigma_i(a_i) = \omega_i \cdot a_i$, con $\omega_i \in \mathbb{Z}_k$. Un *error de un único dígito* (ED para abreviar), es decir, confundir el dígito a_i con a_i'; es detectable si y sólo si se verifica que $\sigma_i (a_i) \not\equiv \sigma_i (a_i') \mod k$. Por otra parte, un *error de transposición* (ET para abreviar), esto es, intercambiar el dígito a_i por el a_j; es detectable si y sólo si $\sigma_i (a_i) + \sigma_j (a_j) \not\equiv \sigma_i (a_j) + \sigma_j (a_i) \mod k$. En el caso en que se considere $k > 10$, se deben introducir caracteres extra que sustituyan a los valores $10, \ldots, k-1$. El principal resultado sobre la detección de errores se basa en el siguiente

Teorema 1 *Si los números de identificación $a_1, \ldots, a_n$, verifican*

$$(\omega_1, \omega_2, \ldots, \omega_n) \cdot (a_1, a_2, \ldots, a_n) = \sum_{i=1}^{n} \omega_i \cdot a_i \equiv 0 \mod k, \tag{3}$$

entonces, un error de un único dígito es indetectable si y sólo si $\omega_i\,(a_i - a_i') \equiv 0 \mod k$*; mientras que un error de transposición que intercambia los elementos de las posiciones* i *y* j *es indetectable si y sólo si* $(\omega_i - \omega_j)\,(a_i - a_j) \equiv 0 \mod k$.

La demostración de este Teorema es inmediata, basta tener en cuenta que cuando se comete un ED, la diferencia entre el valor correcto (esto es, cuando se utiliza a_i) para la expresión (3) y el valor incorrecto (cuando se emplea a_i') es precisamente $\omega_i\,(a_i - a_i')$. Por otra parte, la diferencia entre el valor correcto y el incorrecto para (3) cuando se lleva a cabo un ET es $(\omega_i - \omega_j)\,(a_i - a_j)$.

A partir del Teorema 1 si un dígito de control, a_n, verifica la condición (3) (en realidad esta condición puede ser contrastada si se consideran varios dígitos de control y no sólo uno), entonces, es inmediato determinar las condiciones de los pesos w_i de cada a_i, de modo que se pueda asegurar qué errores son detectables (ver Cuadro 3). Nótese que las unidades de $\mathbb{Z}_k$ son los elementos $z \in \mathbb{Z}_k$ tales que $\mathrm{mcd}(z, k) = 1$.

Error	Forma	Condición sobre el módulo k
ED	$a_i \to a_i'$	$\mathrm{mcd}(w_i, k) = 1$
ET 1	$a_i \ldots a_j \to a_j \ldots a_i$	$\mathrm{mcd}(w_j - w_i, k) = 1$
ET 2	$a_{i-1}a_ia_{i+1} \to a_{i+1}a_ia_{i-1}$	$\mathrm{mcd}(w_{i+1} - w_{i-1}, k) = 1$
Gemelos 1	$\underset{i\,i+1}{a\ a} \to \underset{i\,i+1}{b\ b}$	$\mathrm{mcd}(w_i + w_{i+1}, k) = 1$
Fonética	$a_i 0 \leftrightarrow 1 a_i$	$jw_{i+1} \not\equiv (j-1)w_i \mod k, 0 \le j \le k-1$
Gemelos 2	$\underset{i-1\ \ i+1}{a\ c\ a} \to \underset{i-1\ \ i+1}{b\ c\ b}$	$\mathrm{mcd}(w_{i-1} + w_{i+1}, k) = 1$

Cuadro 3: Condiciones del módulo k sobre diferentes errores

Si por ejemplo, $k = 10$ y los w_j son impares, y por tanto $w_i - w_j$ es par, las condiciones presentadas en la última columna del Cuadro 3 son incompatibles, es decir, no pueden verificarse todas simultáneamente. Así pues, no se pueden detectar, en el 100 % de los casos, los errores mencionados anteriormente para el caso $k = 10$. De forma más general se verifica el siguiente

Teorema 2 *Si un esquema de detección de errores con un módulo par detecta los errores de un único dígito, entonces para cualesquiera* i *y* j*, existe un error de transposición que afecta a las posiciones* i *y* j*, que no puede ser detectado.*

También en este caso la demostración es inmediata puesto que si el esquema detecta todos los errores de tipo ED, entonces las aplicaciones σ_i son permutaciones de $\mathbb{Z}_{2m}$. Además, para detectar todos los errores ET que involucran a las posiciones i y j, es necesario que $\sigma_i(a) + \sigma_j(b) \neq \sigma_i(b) + \sigma_j(a)$, para todos los $a \neq b \in \mathbb{Z}_{2m}$. De donde se tiene que $\sigma(z) = \sigma_i(z) - \sigma_j(z)$ es una permutación de $\mathbb{Z}_{2m}$. Sumando los elementos de $\mathbb{Z}_{2m}$, módulo $2m$, se tiene que

$$m = m + 0 + 1 + (2m-1) + 2 + (2m-2) + \cdots + (m-1) + (m-1),$$

y además

$$m = \sum z = \sum \sigma(z) = \sum \sigma_i(z) - \sum \sigma_i(z) = m - m = 0,$$

lo que es una contradicción.

De lo dicho anteriormente se puede afirmar que el dígito de control, c_{13}, dado por (2) para el código EAN13 está bien definido puesto que verifica

$$\sum_{j=0}^{6} c_{2j+1} + 3\sum_{j=1}^{6} c_{2j} \equiv 0 \mod 10. \tag{4}$$

Nótese que para el EAN13, $w_j = 1$ si j es impar y $w_j = 3$ si j es par. El hecho de asignar diferentes pesos a las posiciones pares e impares hace que se puedan detectar (ver Teorema 1 y Cuadro 3) el 100 % de los errores de un único dígito dado que $\mathrm{mcd}(w_j, 10) = 1$, aunque no se puedan corregir. Además, no detecta la totalidad de los errores de transposición, tal y como se desprende del Teorema 2. En efecto, si se transponen los dígitos a_i y a_j, con i y j de diferente paridad, se tiene que la diferencia de las sumas dadas por (4) para cada uno de los dos casos es $(3a_i + a_j) - (3a_j + a_i) = 2\,(a_i - a_j)$, y el error es indetectable si $|a_i - a_j| = 5$.

Por ejemplo, si se tienen los 13 dígitos siguientes para un código EAN13: 8412345678905, es inmediato comprobar que tal código es correcto. En efecto, basta con verificar que se cumple la congruencia dada en (4):

$$(8+1+3+5+7+9+5)+3(4+2+4+6+8+0) \mod 10 = 110 \mod 10 \equiv 0.$$

Ahora bien, si en lugar de considerar el código anterior, se ha cometido un ED al considerar el código 8412395678905, se puede detectar que éste es erróneo puesto que se tiene

$$(8+1+3+5+7+9+5)+3(4+2+9+6+8+0) \mod 10 \equiv 5 \mod 10,$$

pero no hay forma de saber dónde se cometió el error, y éste no puede corregirse. El código original podría haber sido 3412395678905 ó 8412895678905, etc.

Por otra parte, si se ha producido un ET entre las posiciones sexta y décimo primera y los 13 dígitos considerados fueran 8412395678405, éste código se daría por correcto, a pesar de haberse producido un error, dado que también se verifica que

$$(8+1+3+5+7+4+5)+3(4+2+9+6+8+0) \mod 10 \equiv 0 \mod 10.$$

Para determinar el porcentaje de errores de transposición de dos dígitos consecutivos que son detectados por el EAN13, basta tener en cuenta que de los 90 posibles ET de esta forma, todos se pueden detectar salvo que se tengan los siguientes pares de dígitos consecutivos: 05, 16, 27, 38, 49 y sus inversos, es decir, se puede detectar el $80/90 = 89{,}89\,\%$ de estos errores. Razonando de forma análoga se obtiene el mismo valor para todos los ET.

Sin embargo, si los pesos dados a las posiciones fueran $w_j = 1$ si j es impar y $w_j = 2$ si j es par, éste esquema detectaría el 100 % de los ET, dado que en este caso si se transponen, por ejemplo, los dígitos consecutivos a y b, se tiene que el error sólo es indetectable si $(2a + b) - (2b + a) = a - b \equiv 0 \mod 10$ (ver [11] para esquemas con tres pesos diferentes). Sin embargo, en este caso no se detectarían todos los ED.

4.3 Determinación de las barras y espacios

Para determinar el código de barras bajo el que se escribirán los 13 dígitos de que consta, se utiliza el *módulo* como unidad básica de representación y corresponde a la mínima anchura que puede tener una barra ($\langle 1 \rangle$) o un espacio ($\langle 0 \rangle$). De esta forma la codificación $\langle 000 \rangle$ hace referencia a un espacio de módulo 3, mientras que la codificación $\langle 11 \rangle$ se refiere a una barra de módulo 2. Además, cada código tiene siempre tres separadores (ver Figura 3), que son los mismos para todos los códigos EAN13, y que indican dónde comienza el código (*separador izquierdo* o de inicio), dónde termina (*separador derecho* o de fin) y cuál es su centro (*separador central*). Los separadores derecho e izquierdo son siempre $\langle 101 \rangle$ mientras que el separador central es $\langle 01010 \rangle$. Estos valores permiten al escáner que lee las barras determinar la anchura utilizada por cada módulo del código, lo que posibilita que un mismo escáner sea capaz de leer códigos de barras de diferentes tamaños. Las barras de dichos separadores son ligeramente más largas que las restantes del código.

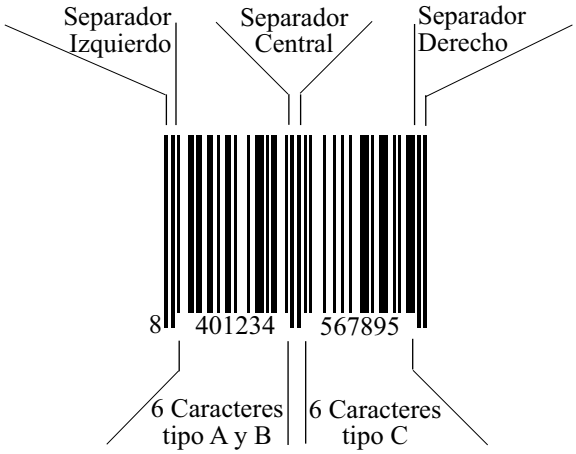


Figura 3: Descripción de la partes de un código EAN13

El primer dígito de los 13 del EAN13 se coloca en la parte exterior izquierda del código, es decir, antes del separador izquierdo y no se representa gráficamente, aunque influye en la representación gráfica de los dígitos que ocupan las posiciones #2 a #7, como se verá posteriormente. Los restantes 12 dígitos se separan en dos grupos de 6 dígitos cada uno, de modo que el primer grupo se coloca y codifica entre el separador izquierdo y el central; mientras que el segundo grupo queda situado entre el separador central y el derecho.

La representación gráfica de cada dígito está formada por dos barras y dos espacios y ocupa siempre 7 módulos, es decir, cada carácter está codificado por un vector binario de longitud 7 y peso de Hamming variable —el número de 1's que contiene—. Así, el vector $\langle 0111001 \rangle$ correspondería a un espacio de módulo

1, una barra de módulo 3, un espacio de módulo 2 y una barra de módulo 1. La codificación anterior también podría haberse llevado a cabo mediante los cuatro dígitos que indican los módulos de cada espacio y barra: 1321.

El número total de módulos de un código EAN13 es, por tanto, de $2 \cdot 3 + 12 \cdot 7 + 5 = 95$, si bien se acostumbran a dejar zonas mudas alrededor del mismo para facilitar su lectura por el escáner. Por otra parte, como la anchura estándar de cada módulo es de 0,33 *mm.* y su altura estándar de 22,85 *mm.*, resulta que el tamaño estándar para un código de este tipo, incluidas las zonas mudas, es de $37,29 \times 26,26$ *mm.* (se permite un escalado entre 0,8 y 2 veces el valor estándar anterior).

Las representaciones gráficas de los 6 primeros dígitos siempre empiezan por un espacio (es decir, por uno o varios 0's) y terminan por una barra (esto es, por uno o varios 1's); mientras que para el segundo grupo es al revés: empiezan por una barra y terminan por un espacio. Nótese que esta distribución es compatible con la codificación de los separadores izquierdo, derecho y central. El primer grupo utiliza dos patrones diferentes (tipos A y B) para codificar cada dígito, mientras que el segundo grupo utiliza un tercer patrón (tipo C). La codificación del primer grupo (posiciones #2 a #7) se hace utilizando el Cuadro 4, según el valor del primer dígito del código. Por esta razón no es necesario representar el primer dígito del código mediante barras, sin que se pierda por ello la información que el mismo proporciona.

Primer dígito	#2	#3	#4	#5	#6	#7
0	A	A	A	A	A	A
1	A	A	B	A	B	B
2	A	A	B	B	A	B
3	A	A	B	B	B	A
4	A	B	A	A	B	B
5	A	B	B	A	A	B
6	A	B	B	B	A	A
7	A	B	A	B	A	B
8	A	B	A	B	B	A
9	A	B	B	A	B	A

Cuadro 4: Patrones de codificación para las posiciones #2 a #7

Así por ejemplo, si el primer dígito es 3, los dígitos de las posiciones #2, #3 y #7 se codifican según el patrón A, mientras que los de las posiciones #4, #5 y #6 utilizan el patrón B. Finalmente, es necesario conocer cómo se codifica cada dígito para cada uno de los patrones. Esta codificación se presenta en el Cuadro 5, de modo que las dos barras (b) y los dos espacios (e) de cada patrón están caracterizados como sigue:

- Patrón A. Su secuencia es *ebeb* y los vectores binarios de este tipo son de longitud 7 y peso impar, esto es, 3 ó 5.

- Patrón B. Su secuencia es *ebeb* y los vectores binarios pertenecientes a este tipo son de longitud 7 y peso par, es decir, 2 ó 4.
- Patrón C. Su secuencia es *bebe* y sus vectores binarios son de longitud 7 y peso par: 2 ó 4 (al ser su paridad diferente de la de A, que es el patrón con el que siempre empieza el código, éste puede leerse bidireccionalmente).

Dígito	Patrón A	Patrón B	Patrón C
0	$3211 \to \langle 0001101 \rangle$	$1123 \to \langle 0100111 \rangle$	$3211 \to \langle 1110010 \rangle$
1	$2221 \to \langle 0011001 \rangle$	$1222 \to \langle 0110011 \rangle$	$2221 \to \langle 1100110 \rangle$
2	$2122 \to \langle 0010011 \rangle$	$2212 \to \langle 0011011 \rangle$	$2122 \to \langle 1101100 \rangle$
3	$1411 \to \langle 0111101 \rangle$	$1141 \to \langle 0100001 \rangle$	$1411 \to \langle 1000010 \rangle$
4	$1132 \to \langle 0100011 \rangle$	$2311 \to \langle 0011101 \rangle$	$1132 \to \langle 1011100 \rangle$
5	$1231 \to \langle 0110001 \rangle$	$1321 \to \langle 0111001 \rangle$	$1231 \to \langle 1001110 \rangle$
6	$1114 \to \langle 0101111 \rangle$	$4111 \to \langle 0000101 \rangle$	$1114 \to \langle 1010000 \rangle$
7	$1312 \to \langle 0111011 \rangle$	$2131 \to \langle 0010001 \rangle$	$1312 \to \langle 1000100 \rangle$
8	$1213 \to \langle 0110111 \rangle$	$3121 \to \langle 0001001 \rangle$	$1212 \to \langle 1001000 \rangle$
9	$3112 \to \langle 0001011 \rangle$	$2113 \to \langle 0010111 \rangle$	$3112 \to \langle 1110100 \rangle$

Cuadro 5: Asignación de los patrones de codificación

Es claro que si se elige otra codificación diferente a la presentada en los Cuadros 4 y 5, es decir, otra reordenación en los patrones para cada dígito inicial, o las asignaciones a dígitos dentro de cada patrón de codificación, se obtendrán diferentes códigos de barras. En este caso no se estaría utilizando el estándar EAN13, sino otra codificación diferente.

4.4 EAN13 y códigos binarios

Después de lo mencionado acerca de la representación mediante espacios y barras (alternados) los dígitos 0–9 en un código EAN13, parece necesario dar una justificación matemática a dicho proceso, aunque el inventor del código UPC, precursor del EAN13, no se basara en conceptos matemáticos.

En primer lugar, es claro que la representación de cada dígito necesita, al menos, de 2 espacios y 2 barras, puesto que como el final de la representación de un dígito debe servir para marcar el inicio de la representación del dígito siguiente, no pueden emplearse 2 espacios y 1 barra (o viceversa). Por tanto, el código binario a emplear debe ser de longitud mayor o igual que 4.

Por otra parte, se ha de garantizar, también, que cada dígito pueda ser codificado de dos maneras diferentes: una para el patrón A (vectores de peso impar) y otra para el B (vectores de peso par), por lo que hace falta un mínimo de 20 codewords. Además, todas las representaciones de los dígitos deben empezar por $\langle 0 \rangle$ (espacio) y terminar por $\langle 1 \rangle$ (barra). Dado que sólo existen $2^4 = 16$, codewords de longitud 6 de la forma

$$\langle 0\ x_1\ x_2\ x_3\ x_4\ 1 \rangle, \qquad x_i \in \mathbb{F}_2,$$

es necesario utilizar codewords de longitud, al menos, 7. Supongamos entonces que los codewords son de longitud 7 y que de los $2^7 = 128$ posibles codewords sólo se consideran los $2^5 = 32$ de la forma preestablecida, es decir,

$$\langle 0\ x_1\ x_2\ x_3\ x_4\ x_5\ 1\rangle, \qquad x_i \in \mathbb{F}_2.$$

De estos 32 codewords se deben eliminar todos aquellos para los que no haya exactamente 3 cambios de la forma 01 ó 10, dado que los 2 espacios y las 2 barras de que consta cada representación (*ebeb*) deben estar alternados. Una vez eliminados los que no verifican esta condición, quedan los siguientes posibles codewords:

0000101,	0001001,	0001011,	0001101,	0010001,
0010011,	0010111,	0011001,	0011011,	0011101,
0100001,	0100011,	0100111,	0101111,	0110001,
0110011,	0110111,	0111001,	0111011,	0111101.

Como se puede observar, la mitad de los codewords anteriores son de peso impar (3 ó 5) y la otra mitad son de peso par (2 ó 4), por lo que dicha división puede utilizarse como criterio para separar los dos patrones y obtener dos códigos binarios de longitud 7, diferentes:

$$\begin{aligned}\mathcal{C}_A = \{&0001101, 0011001, 0010011, 0111101, 0100011,\\ &0110001, 0101111, 0111011, 0110111, 0001011\},\\ \mathcal{C}_B = \{&0100111, 0110011, 0011011, 0100001, 0011101,\\ &0111001, 0000101, 0010001, 0001001, 0010111\}.\end{aligned}$$

La ordenación de los codewords (que corresponde a la presentada en el Cuadro 5) es la misma que la empleada en el código UPC americano, inventado en 1973 por George J. Laurer —4 años antes que el EAN13—, quien, en comunicación personal, nos ha asegurado que la ordenación del patrón A (y por tanto la asignación de los dígitos 0–9 para el primer grupo de 6 caracteres) la llevó a cabo de forma completamente arbitraria (lo mismo que la ordenación de los patrones A y B del Cuadro 4). Por otra parte, la ordenación del patrón B la tomó, de alguna manera, como imagen especular del patrón A. A partir de las asignaciones mencionadas, la correspondiente al patrón C es inmediata de modo que permita decidir al escáner si la lectura del código se lleva a cabo de derecha a izquierda o al revés.

$$\begin{aligned}\mathcal{C}_C = \{&1110010, 1100110, 1101100, 1000010, 1011100,\\ &1001110, 1010000, 1000100, 1001000, 1110100\}.\end{aligned}$$

Así pues, los códigos $\mathcal{C}_A$, $\mathcal{C}_B$ y $\mathcal{C}_C$ son códigos binarios de longitud 7 y distancia mínima 2, es decir son $(7, 10, 2)$-códigos, pero no son códigos lineales, dado que no son subespacios vectoriales de $\mathbb{F}_2^7$. Su tasa de información es:

$$R = \frac{\log_2 10}{7} = 0{,}475.$$

El hecho de añadir barras al código EAN13 para representar dígitos, dota a este de cierta capacidad para corregir errores, de manera análoga a como lo hace el código Postnet, ya comentado, aunque esta posibilidad no se utiliza habitualmente. Por ejemplo, si se pierde un dígito del código o es leído erróneamente —ya sea porque hay manchas o roturas o porque la paridad no es la adecuada—, el error es detectado porque el EAN13 puede detectar un error al ser su distancia mínima $d = 2$. Además, se puede saber la posición del error a partir de la lectura de las barras. Conocida ésta, el dígito puede ser recuperado sin más que utilizar la expresión (4). Por otra parte, si el error se comete en la lectura del primer dígito, que no está representado mediante barras, también puede ser corregido debido a la paridad de los 6 dígitos siguientes: basta con observar la asignación de los patrones del Cuadro 4.

4.5 Ejemplo de código

Para determinar el código EAN13 de un producto ficticio se calculan los 13 dígitos de que consta y se determina el código de barras correspondiente.

Si la empresa utiliza el sistema EAN de la organización AECOC (Asociación Española de Codificación Comercial [1]) de España, los dos primeros dígitos serán 84, es decir, $c_1 = 8$, $c_2 = 4$. Si la AECOC ha asignado a esta empresa los dígitos 765432, éstos serán los siguientes, es decir, $c_3 = 7$, $c_4 = 6$, $c_5 = 5$, $c_6 = 4$, $c_7 = 3$ y $c_8 = 2$. Si 0209 son los 4 dígitos que el propietario de la marca asigna al artículo ficticio, entonces $c_9 = 0$, $c_{10} = 2$, $c_{11} = 0$ y $c_{12} = 9$. Según esto, los dígitos para el código de barras son: $847654320209c_{13}$. Para calcular el dígito de control se procede como se indica en la expresión (2):

$$c_{13} = -\left(23 + 3 \cdot 27\right) \mod 10 = -104 \mod 10 = 6.$$

Por lo que los dígitos para el código de barras son: 8476543202096.

A continuación, se determinará la forma en que los dígitos calculados se representan mediante barras y espacios verticales para dar lugar al código de barras EAN13.

Como el primer dígito del código es 8, la secuencia de los patrones A y B es A, B, A, B, B, A (ver Cuadro 4). Dado que los 6 primeros dígitos son 476543, su codificación, según el Cuadro 5, es la siguiente:

$$\begin{aligned} &4 \to \langle 0100011 \rangle, \quad 7 \to \langle 0010001 \rangle, \quad 6 \to \langle 0101111 \rangle, \\ &5 \to \langle 0111001 \rangle, \quad 4 \to \langle 0011101 \rangle, \quad 3 \to \langle 0111101 \rangle. \end{aligned} \tag{5}$$

Procediendo de forma análoga para el segundo grupo de 6 dígitos, para los que se utiliza el patrón C, se obtiene su codificación:

$$2 \to \langle 1101100 \rangle, \quad 0 \to \langle 1110010 \rangle, \quad 9 \to \langle 1110100 \rangle, \quad 6 \to \langle 1010000 \rangle. \tag{6}$$

Finalmente, la representación del código de barras EAN13 para el artículo que se está considerando es la que se muestra en la Figura 4. Dicha figura se ha elaborado colocando en primer lugar el separador izquierdo $\langle 101 \rangle$, a continuación

los vectores binarios del primer grupo presentado en (5), luego el separador central $\langle 01010 \rangle$, a continuación los vectores binarios del segundo grupo mostrado en (6), y finalmente, el separador derecho $\langle 101 \rangle$.

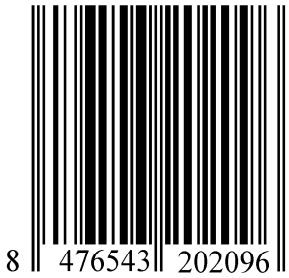


Figura 4: Ejemplo de Código EAN13

Con relación a la posibilidad de corregir errores del EAN13, si en el código anterior se hubiera perdido el primer dígito, observando que los pesos de los 6 dígitos siguientes son; $3, 2, 5, 4, 4, 5$, se deduce que el orden de los patrones es: A, B, A, B, B, A, por lo que a partir del Cuadro 4 se tiene que dicho dígito es 8 y el error se habría corregido. Por otra parte, si, por ejemplo, el dígito que ocupa la sexta posición (el 4), no pudiera leerse o fuera leído como $\langle 0111101 \rangle$, se tendría que, o bien no se conoce su valor o se sabría que es erróneo puesto que su peso es impar, cuando debería ser par. En cualquiera de los dos casos su valor real se podría conocer sin más que utilizar la expresión (4). En efecto, si se denota por x al valor desconocido o erróneo, se tendría que

$$(8 + 7 + 5 + 3 + 0 + 0 + 6) + 3(4 + 6 + x + 2 + 2 + 9) \equiv 0 \mod 10,$$

es decir,

$$\begin{aligned} 3x &\equiv -29 - 69 \mod 10 \equiv -98 \mod 10 = 2, \\ x &\equiv 2 \cdot 3^{-1} \mod 10 \equiv 2 \cdot 7 \mod 10 = 4. \end{aligned}$$

4.6 EAN13 y las particiones de 7

Para concluir esta sección se presentará otra explicacón a la señalada anteriormente para la elaboración de las barras y espacios que forman parte de un código EAN13, que tiene como base las particiones de un entero.

Dado que cada dígito del EAN13 ocupa un módulo de 7 unidades —suma de las anchuras de 2 barras y 2 espacios—, el problema consiste en escribir el número 7 como suma de 4 dígitos de diferentes formas. La descomposición de un número entero como suma de dígitos se conoce como *particiones de un entero*, y es un tema ampliamente estudiado en Combinatoria (ver, por ejemplo, [4, Capítulo 5.3] y [6, Capítulo 2.7]). En particular, las particiones de 7 como suma

de dígitos son las siguientes:

$$\begin{array}{ll} 7 = 1+1+1+1+1+1+1 & = 1+1+1+1+1+2 \\ = 1+1+1+2+2 & = 1+1+1+1+3 \\ = 1+2+2+2 & = 1+1+2+3 \\ = 1+1+1+4 & = 1+3+3 \\ = 1+2+4 & = 1+1+5 \\ = 2+2+3 & = 3+4 \\ = 2+5 & = 1+6. \end{array}$$

De todas ellas, sólo son interesantes para el EAN13 las que contienen exactamente 4 dígitos: $(1,2,2,2)$, $(1,1,2,3)$ y $(1,1,1,4)$. Por otra parte, dado que el orden de colocación de las barras y espacios hace que el código sea visualmente diferente, se deben considerar todas las formas de escribir los tres grupos de cuatro dígitos anteriores, es decir, sus permutaciones:

$$\begin{array}{llllll} [1,2,2,2], & [2,1,2,2], & [2,2,1,2], & [2,2,2,1], & & \\ [1,1,2,3], & [1,1,3,2], & [1,2,1,3], & [1,2,3,1], & [1,3,1,2], & [1,3,2,1], \\ [2,1,1,3], & [2,1,3,1], & [2,3,1,1], & [3,1,1,2], & [3,1,2,1], & [3,2,1,1], \\ [1,1,1,4], & [1,1,4,1], & [1,4,1,1], & [4,1,1,1]. & & \end{array}$$

Se observa que la mitad de las 20 permutaciones obtenidas son simétricas respecto de las restantes. Por ejemplo, si se consideran los 10 grupos:

$$\begin{array}{lllll} [2,1,2,2], & [2,2,2,1], & [1,1,3,2], & [1,2,1,3], & [1,2,3,1], \\ [1,3,1,2], & [3,1,1,2], & [3,2,1,1], & [1,1,1,4], & [1,4,1,1], \end{array} \tag{7}$$

los restantes 10 son su simétricos. Por tanto, se puede considerar como patrón A al determinado por la codificación de las 10 permutaciones dadas en (7), y como patrón B a las restantes 10 permutaciones. Además, como el patrón A comienza por barra y termina por espacio, se puede utilizar el mismo conjunto de permutaciones para el patrón C, dado que éste empieza por espacio y concluye con barra. Así pues, asignando una permutación de las dadas en (7) a cada uno de los 10 dígitos, se llega a la codificación de cada patrón dada en el Cuadro 5.

Nótese que si las 10 permutaciones que definen el patrón A o si su orden fuera diferente del mostrado en el Cuadro 5, se obtendrían códigos de barras diferentes para una misma cadena de dígitos y, consecuentemente, no pertenecerían al estándar EAN13.

5 Los códigos ISBN y EAN13

Como es sabido, el código ISBN (International Standar Book Number, [12]) es un sistema internacional de numeración e identificación de libros surgido en 1965, y aunque no sea un código de barras propiamente dicho, se incluye en este artículo porque es posible elaborar un código EAN13 para cada código ISBN, de modo que pueda ser tratado informáticamente.

El código ISBN contiene 10 dígitos divididos en cuatro grupos por tres guiones. Dado que las posiciones de los guiones son irrelevantes para el código, se hará caso omiso de las mismas. El último dígito es el de control (la letra X si su valor es 10) y se calcula mediante la siguiente expresión:

$$b_{10} = -\sum_{i=1}^{9} p_i \cdot b_i \mod 11, \tag{8}$$

donde $p_i = 11 - i$ es el peso de cada posición. La verificación de la lectura del código se lleva a cabo de forma análoga a como se hizo con el código EAN13, es decir, se comprueba si se verifica la siguiente congruencia:

$$\sum_{i=1}^{10} p_i \cdot b_i = 10b_1 + 9b_2 + \ldots + 2b_9 + b_{10} \equiv 0 \mod 11. \tag{9}$$

Es fácil probar que el código ISBN detecta el 100 % de los ED's. En efecto, si al cambiar el dígito b_i por b_i' el error no fuera detectable, los valores de las dos sumas de comprobación dadas por (9) serían múltiplos de 11, por lo que también lo sería su diferencia: $(11-i)\,(b_i - b_i')$, lo cual es imposible puesto que 11 es primo y $|b_i - b_i'| < 11$. Este código también detecta el 100 % de los ET's. En efecto, razonando como antes, la diferencia de las sumas dadas por (9) sería mútiplo de 11, pero esta diferencia es $(p_i b_i + p_j b_j) - (p_j b_i + p_i b_j) = (j - i)\,(b_i - b_j)$, que tampoco puede ser múltiplo de 11. De forma similar se puede probar que es capaz de detectar el 100 % de los errores Gemelos 2 dados en el Cuadro 3, aunque no detecta todos los errores Gemelos 1 ni Fonéticos. Así pues, desde el punto de vista de la detección de errores, puede afirmarse que el código ISBN es preferible al EAN13 (véase [8] para un estudio sobre la detección de errores dobles en ISBN). Para esta mejora es básico que se utilice el número 11, que es primo, y, por tanto, $\mathbb{Z}_{11}$ es un cuerpo.

Para determinar el código de barras EAN13 a partir de un código ISBN dado b_1b_2-$b_3b_4b_5b_6$-$b_7b_8b_9$-b_{10}, basta con calcular los 13 dígitos que le corresponden en formato EAN13. En primer lugar se eliminan los guiones y el dígito de control: $b_1b_2b_3b_4b_5b_6b_7b_8b_9$. A continuación se añaden los dígitos 978 al inicio del código ISBN de modo que $c_1 = 9$, $c_2 = 7$, $c_3 = 8$, $c_i = b_{i-3}$, para $4 \leq i \leq 12$. Finalmente se determina el dígito de control como se señaló en (2):

$$c_{13} = -\sum_{j=1}^{6} (c_{2j-1} + 3c_{2j}) \mod 10 = -\left(8 + \sum_{j=2}^{4} b_{2j} + 3\sum_{j=1}^{5} b_{2j-1}\right) \mod 10.$$

Si, por ejemplo, el código ISBN fuera 84-7897-421-0, bastará con escribir 978847897421 y calcular su dígito de control según (2) o (5):

$$c_{13} = -\,(8 + 21 + 3 \cdot 29) \mod 10 = -116 \mod 10 = 4.$$

Por lo que su código de barras EAN13 sería: 9788478974214.

Recíprocamente, para convertir un código EAN13, $987c_4, \ldots, c_{13}$, en formato ISBN, se eliminan los tres primeros dígitos, 978, y el dígito de control, c_{13}; y a continuación se determina el dígito de control ISBN según la expresión (8).

Como ejemplo, se puede tomar como código EAN13 el calculado anteriormente: 9788478974214, y proceder como se ha indicado: se eliminan los tres primeros dígitos y el último: 847897421, y se calcula el dígito de comprobación según la expresión (8):

$$-(80+36+56+56+54+35+16+6+2) \mod 11 = -341 \mod 11 = 0.$$

Con lo que el código ISBN es: 8478974210.

Hay otro código basado en la aritmética módulo 11, al igual que el ISBN, que sí puede detectar el 100 % de los errores señalados en el Cuadro 3; aunque no otros errores, como errores dobles de transposición o de dos dígitos. Es el código utilizado por Bancos y Cajas de Ahorros en los números de las cuentas bancarias. El código asigna pesos geométricos con base 2 a las posiciones de los dígitos, en lugar de pesos aritméticos:

$$\left(2, 2^2, 2^3, \ldots, 2^{10}\right) \mod 11 = (2,4,8,5,10,9,7,3,6,1).$$

Las cuentas bancarias se identifican por 20 dígitos: los 4 primeros corresponden a la entidad bancaria, b_1, b_2, b_3, b_4; los 4 siguientes a la sucursal, s_1, s_2, s_3, s_4; el noveno dígito, c_1, es el de control de los 8 primeros; mientras que el décimo, c_2, es el de control de los 10 últimos, $d_1, d_2, \ldots, d_{10}$, que constituyen el número de la cuenta. Los valores de c_1 y c_2 se determinan como sigue:

$$c_1 \equiv -(4b_1 + 8b_2 + 5b_3 + 10b_4 + 9s_1 + 7s_2 + 3s_3 + 6s_4) \mod 11,$$
$$c_2 \equiv -(d_1 + 2d_2 + 4d_3 + 8d_4 + 5d_5 + 10d_6 + 9d_7 + 7d_8 + 3d_9 + 6d_{10}) \mod 11.$$

Sin embargo, si el dígito de control es 10, en lugar de cambiarlo por la letra X, se sustituye por el número 1, lo que supone una ambigüedad y, por tanto, disminuye la capacidad de este código para detectar errores.

Existen otros esquemas basados en la aritmética modular que, sorprendentemente, son bastante utilizados a pesar de que ni siquiera son capaces de detectar el 100 % de los errores de un único dígito. Son sistemas que utilizan los 10 dígitos en su código, pero hacen módulo 7 ó 9 para determinar el dígito de control. Es el caso de los cheques del Servicio Postal de Estados Unidos, los cheques de viaje de American Express, los servicios de mensajería de Federal Express y United Parcel Service, o los billetes de las compañías aéreas. Por ejemplo, el dígito de control de los cheques de viaje de American Express se determina de modo que sumado a los anteriores sea divisible por 9. Esto significa que un ED que involucre a los números 0 y 9 no será detectado. Por su parte, cuando se hace módulo 7 la efectividad para detectar ED's es ligeramente menor dado que no se detectan errores en los que intervengan los pares $0,7$; $1,8$ y $0,9$; aunque aumente la efectividad para detectar los ET's, que en ningún caso llega al 100 %.

Agradecimientos. Los autores agradecen al evaluador anónimo sus sugerencias para la mejora en la versión final de este artículo.

Referencias

[1] Asociación Española de Codificación Comercial (AECOC). http://www.aecoc.es/

[2] Association for Automatic Identification and Data Capture Technologies (AIM Global Network). http://www.aimglobal.org/sitemap/

[3] American National Standard Intitute (ANSI). http://www.ansi.org/

[4] C. Chuan Chong and K. Khee Meng. *Principles and techniques in combinatorics*. World Scientific, Singapore, 1992.

[5] G.C. Clark, Jr. and J.B. Cain. *Error-correction coding for digital communications*. Plenum Press, New York, 1981.

[6] G.M. Constantine. *Combinatorial theory and statistical design*. John Willey & Sons, New York, 1987.

[7] European Article Numbering (EAN). http://www.ean-int.org/index 800.html

[8] L. Egghe and R. Rousseau. *The detection of double errors in ISBN- and ISSN-like codes*. Math. Comput. Modelling 33:943–955, 2001.

[9] D. Hankerson, D.R. Hoffman, D.A. Leonard, C.C. Lindner, K.T. Phelps, C.A. Rodger and J.R. Wall. *Coding theory and cryptography, The essentials*. Marcel Dekker, Pure and Applied Mathematics, 234, 2nd edition, 2000.

[10] A. Fúster Sabater, D. de la Guía Martínez, L. Hernández Encinas, F. Montoya Vitini y J. Muñoz Masqué. *Técnicas criptográficas de protección de datos*. RA-MA, Madrid, 2^a edición, 2000.

[11] J.A. Gallian. *Error detecting methods*. ACM Computing Surveys 28(3):504–517, September 1996.

[12] The International ISBN Agency. http://www.isbn-international.org/

[13] Japanese Article Number (JAN). http://www.n-barcode.com/shurui-en/jan.html

[14] A. Menezes, P. van Oorschot, and S. Vanstone. *Handbook of applied cryptography*. CRC Press, Boca Raton, FL, 1997.

[15] W.W. Peterson and E.J. Weldon, Jr. *Error-correcting codes*. The MIT Press, Cambridge, 2nd edition, 1972.

[16] L. Steen (Ed.). *Las Matemáticas en la vida cotidiana*. Addison-Wesley Iberoamericana España, S.A., Madrid, 3^a edición, 1999.

[17] Uniform Code Council (UCC). http://www.uc-council.org/

Bol. Soc. Esp. Mat. Apl. n°28(2004), 71–98

Quelques équations de transport apparaissant en biologie*

B. Perthame

Département de Mathématiques et Applications
École Normale Supérieure, Paris

benoit.perthame@ens.fr

Résumé

De nombreuses Équations aux Dérivées Partielles interviennent dans diverses théories issues de la biologie. Ce papier fournit à la fois un point de vue général sur le rôle classique que jouent les EDP dans ce domaine des sciences (morphogenèse, écologie), et un point de vue plus particulier sur quelques exemples d'équations de transport qui sont apparues plus récemment. Ces exemples que nous présentons ici proviennent : (i) d'asymptotiques dans des modèles de dynamique adaptative (évolution d'espèces par mutation lors de la naissance), (ii) de la théorie des populations structurées (par exemple vieillissement d'une population, cycle cellulaire) fondée sur les équations hyperboliques, (iii) de la description macroscopique du mouvement cellulaire où des systèmes elliptiques/paraboliques ou hyperboliques/paraboliques ont été introduits depuis quelques décennies et (iv) d'un point de vue microscopique sur le mouvement cellulaire qui mène à des équations cinétiques.

Mots-clés: *Équations aux Dérivées Partielles, chimiotactie, dynamique des populations structurées, équations de transport, biomathématiques.*
Classement AMS: *35-02 ;45-02 ;35a90 ;92c17 ;92c50.*

1 Introduction : aspects des mathématiques en biologie

Quand on consulte la littérature en biologie, pour voir où y interviennent des mathématiques, on remarque tout de suite des domaines bien établis : par exemple, les probabilités et les statistiques jouent un rôle important, et

*Ce texte a été rédigé à l'occasion d'une des Leçons de Mathématiques de Bordeaux et sera publié, dans un ouvrage regroupant ces leçons, en 2004, par les éditions Cassini (Paris).
Fecha de recepción: 4 de agosto de 2003

beaucoup de statisticiens travaillent déjà pour la biologie. La raison en est simple : il s'agit de manipuler des données réelles ou de regarder des phénomènes influencés par les aléas naturels. Un autre exemple de domaine mathématique dont l'application à la biologie est bien connue et reconnue, est celui des systèmes différentiels : systèmes dynamiques, théorie de la bifurcation, dynamique lente-rapide... etc. C'est l'objet de modélisation le plus usuel en biologie car il est simple et permet de rendre compte de situtations "moyennes" et éventuellement de les contôler efficacement. Des travaux de Daniel Bernouilli ([6], 1760) utilisent déjà ce formalisme.

Les Équations aux Dérivées Partielles (ÉDP) et les Équations Intégrales (ÉI) ont été utilisées plus récemment. On les connaît essentiellement dans deux types de problèmes, qu'on peut appeler *type Fisher-KPP* et *type Turing*. Mais bien d'autres types d'ÉDP sont apparues en biologie et servent d'outil courant de modélisation. Une référence classique sur ce sujet est le livre de James D. Murray ([52], 1983). Par exemple on trouve des ÉDP et ÉI issues de modèles antérieurs basés sur des équations différentielles ordinaires, auxquelles on a adjoint ensuite une variable supplémentaire, on parle de modèles structurés et le plus courant est de structurer une population par l'âge ou la taille des individus. En écologie il est courant de structurer une population grâce à une variable d'espace.

L'objet de ce texte est de montrer comment les ÉDP interviennent en biologie à partir d'exemples classiques ou de travaux récents.

2 Équations différentielles ordinaires

2.1 Dynamique des populations : équations de Malthus (1798) et de Verhulst (1838)

Notons $n(t)$ le nombre d'individus d'une population à l'instant t. La question est de savoir comment $n(t)$ évolue au cours du temps.

Commençons par le cas très simple où les ressources seraient illimitées. Thomas Malthus, dans le chapitre 1 de son célèbre *Essay on the principle of population* ([46], 1798), postule qu'en pareil cas le taux de natalité et le taux de mortalité seraient constants. Ce qui conduit à l'équation suivante (que Malthus n'écrit pas, mais qu'on appelle quand même *l'équation de Malthus*) :

$$\frac{dn}{dt} = an, \tag{1}$$

où le paramètre a est la différence entre le taux de natalité et le taux de mortalité.

Dans ce modèle, la population croît exponentiellement sans limite (si $a > 0$), ou s'éteint exponentiellement vite (si $a < 0$). Bien qu'il n'écrive pas d'équation, Malthus ne manque pas de remarquer cette conséquence de son postulat (chapitre 2 de [46] ; en fait il ne considère que le cas où $a > 0$).

Ce modèle décrit bien les périodes d'explosion démographique : colonisation d'une terre vierge aux ressources abondantes (une fois achevée la vague d'immigration : l'équation ne tient compte que de l'accroissement par

reproduction) ou, ce qui revient au même, repeuplement après une catastrophe. Mais, comme le fait remarquer Malthus lui-même, ces périodes de croissance exponentielle ne sont jamais indéfinies : dans la nature, la population est toujours bornée par les conditions environnementales : les ressources sont limitées (ou elles croissent lentement : Malthus pense qu'elles croissent linéairement), et les différents individus se font concurrence. C'est même là le thème central de l'*Essay* de Malthus. Mais il ne donne pas de formulation quantitative de cette situation.

En 1838, Pierre Verhulst [61], intéressé à la question par Adolphe Quételet, propose une modification de l'équation (1) pour tenir compte de cette limitation. L'équation de Verhulst (ou *équation logistique*) est :

$$\frac{d}{dt}n = kn\,(M - n)\,, \tag{2}$$

où M est la population maximale possible dans l'environnement donné, et $a = kM$ la différence entre le taux de natalité et le taux de mortalité. Les solutions sont $n = 0$ (personne !), $n = M$ (saturation) et $n(t) = \frac{M}{1+e^{-a(t-t_1)}}$ avec $t_1 = \text{cte}$ (saturation asymptotique si $a > 0$, extinction si $a < 0$).

2.2 Proies et prédateurs : le modèle de Lotka-Volterra (1925-1926)

Dans le chapitre 2 de son *Essay* ([46], 1798), Malthus décrit aussi (sans équations) un système ressources-population conduisant à des oscillations entre périodes de relatif confort et périodes de disette (pour les classes les moins aisées de la société) : quand les ressources abondent, la population croît (plus vite que les ressources), et le rapport ressources/population diminue ; les ressources viennent alors à manquer : la population stagne ou même décroît ; pour survivre, on cultive de nouvelles terres, on améliore la productivité, jusqu'à ramener le rapport ressources/population au même niveau qu'avant, et le cycle recommence.

C'est une version quantitative de ce genre de modèle que vont proposer Alfred Lotka ([45], 1925) et (indépendamment) Vito Volterra ([62], 1926). Lotka s'intéressait à un système plantes-herbivores (ce qui est très proche du problème de Malthus, mais plus simple car débarrassé des complications économiques), et Volterra à un système proies-prédateurs.

Le problème que se posait Volterra est célèbre : il s'agissait de comprendre pourquoi, à la fin de la Première Guerre mondiale, lorsque les pêcheurs de sardines ont repris leur activité dans la mer Adriatique, un biologiste a constaté sur les marchés que la proportion de sardines avait diminué, et que la proportion de requins avait augmenté. Pour expliquer cet état de fait, il fallait décrire l'évolution spontanée du système sardines-requins, c'est-à-dire son évolution en l'absence de pêcheurs.

Notons respectivement $S(t)$ et $R(t)$ les densités de sardines et de requins dans l'Adriatique à l'instant t. Volterra suppose que s'il n'y avait pas les requins, les sardines se développeraient selon la loi de Malthus $\frac{d}{dt}S = aS$. (L'effet de saturation de Verhulst apparaîtrait au bout d'un certain temps, quand $M - S$

cesserait d'être à peu près égal à M, mais on n'a pas besoin d'en tenir compte car l'effet des requins va se faire sentir bien avant !)

Quant aux requins, s'ils étaient seuls (pas ou peu de sardines) leur taux de mortalité serait supérieur à leur taux de natalité, et leur population décroîtrait exponentiellement (loi de Malthus $\frac{d}{dt}R = -bR$ avec un coefficient $-b < 0$).

La présence simultanée des deux espèces sauve les requins de l'extinction et empêche l'explosion démographique des sardines. Le système couplé décrivant cette situation est le suivant :

$$\begin{aligned} \frac{d}{dt}S &= aS - cSR, \\ \frac{d}{dt}R &= -bR + dSR, \end{aligned} \tag{3}$$

où a, b, c, d sont des constantes positives. Le modèle suppose que les requins n'ont rien d'autre à manger que des sardines, que les sardines n'ont aucun problème pour se nourrir et pas d'autres prédateurs que les requins, et que le nombre de rencontres entre requins et sardines est proportionnel à R et à S.

Ce modèle conduit à des oscillations régulières de S et de R, avec la même période et un léger décalage : un maximum (resp. un minimum) dans la densité de proies est bientôt suivi par un maximum (resp. un minimum) dans la densité de prédateurs. Le décalage entre les deux cycles peut être assez important pour que le maximum de densité des prédateurs coïncide avec une faible densité de proies : c'est la situation qui régnait dans l'Adriatique pour le système sardines-requins à la fin de la Première Guerre mondiale.

Lotka et Volterra eux-mêmes ont considéré d'autres situations, modélisées par d'autres systèmes (espèces qui se font concurrence pour une même nourriture, etc.), et beaucoup de modèles ont été mis au point depuis. Un des domaines actifs d'application, la biochimie, mène à des systèmes beaucoup plus complexes (voir par exemple Albert Goldbeter [33] 1997). L'épidémiologie est aussi un domaine toujours actif (voir par exemple O. Diekmann, J.P. Heesterbeek ([21] 2000).

3 Équations aux dérivées partielles

3.1 Diffusion avec reproduction : Fisher et KPP (1937)

Les équations aux dérivées partielles apparaissent à propos de questions concernant l'écologie et l'épidémiologie : on veut décrire par exemple la propagation des espèces invasives (comme les algues en Méditerranée), ou la diffusion d'un gène dans une population.

Supposons qu'un certain gène confère aux individus qui le possèdent un quelconque avantage dans la lutte pour la vie. La sélection naturelle opérant, ce gène va se répandre dans la population. En 1930, Ronald Fisher [27] montre que sous des hypothèses raisonnables, la croissance de ce gène dans une population

donnée va suivre la loi de Verhulst

$$\frac{d}{dt}u = ku(1-u), \tag{4}$$

où $u(t)$ est la densité du gène dans la population.

Mais Fisher se pose aussi une autre question : décrire la progression spatiale, géographique, du gène (c'est un problème de propagation de front). Son idée est qu'on peut considérer qu'il s'agit d'un phénomène de *diffusion*. On sait depuis Fourier (1807) que la diffusion de la chaleur est décrite par une équation du type

$$\partial_t u - \kappa \Delta u = 0. \tag{5}$$

∂_t est une notation abrégée pour $\frac{\partial}{\partial t}$, et $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ est le *laplacien*. κ est une constante (qu'on peut rendre égale à 1, si on veut, en changeant l'unité de longueur), et $u(t,x)$ est la température à l'instant t au point x. On s'est rendu compte par la suite que cette équation décrivait en fait bien d'autres processus de diffusion, comme celui des molécules dans le phénomène de l'osmose (Adolph Fick, 1855 : cf. [26], 1856) ou des particules animées du mouvement brownien (Einstein et Smoluchowski, 1905), etc.

Fisher ([28], 1937) suppose que les gènes favorables diffusent de la même façon, et il propose donc l'équation :

$$\partial_t u - \Delta u = ku(1-u), \tag{6}$$

qui prend en compte à la fois la diffusion géographique et la reproduction.

Au même moment, Kolmogorov, Petrovskii et Piskunov ([40], 1937), d'ailleurs inspirés par le livre séminal de Fisher [27], et voulant résoudre le même problème de propagation d'un gène favorable, étaient conduits à la même équation, et même plus généralement à ce qu'on appelle aujourd'hui l'équation KPP :

$$\partial_t u - \Delta u = f(u), \tag{7}$$

où f est une fonction non linéaire soumise à certaines conditions : Kolmogorov, Petrovskii, Piskunov envisagent plusieurs systèmes de conditions, par exemple f est définie sur $[0,1]$, dérivable autant de fois que nécessaire, nulle en 0 et en 1, strictement positive sur $]0,1[$ et telle que $f'(0) > 0$ et $f'(u) < f'(0)$ pour $0 < u \leq 1$. L'exemple le plus simple de fonction f satisfaisant à ces conditions est celui de Verhuslt, $f(u) = ku(1-u)$, et l'équation est alors celle de Fisher. Une variante de l'équation (7), de type ÉI, est utilisée pour prendre en compte des propagations par sauts (grains de pollen par exemple) de y à x avec des taux $k(x-y)$,

$$\partial_t u - \int k(x-y)u(t,y)dy = f(u). \tag{8}$$

Du point de vue mathématique, KPP est une équation parabolique semi-linéaire, c'est-à-dire que toute la non-linéarité est dans les termes d'ordre 0. S'il y a plusieurs espèces en compétition on peut avoir un système de plusieurs équations KPP couplées. Des modèles où f admet des points singuliers bistables

décrivent des propagations de signaux (électriques, ioniques) dans les nerfs ou le cerveau par exemple.

Une des problèmatiques récentes consiste ici à faire apparaître un front de colonisation, grâce à un changement d'échelle, et à caractériser sa vitesse. On obtient alors des équations de Hamilton-Jacobi, voir par exemple Guy Barles, Lawrence C. Evans, Patragiotis E. Souganidis [2], 1990,

$$\partial_t \varphi + H(\nabla \varphi) = S. \tag{9}$$

La fonction H, le hamiltonien, prend la forme $H(p) = |p|^2$ dans le cas de l'équation (7). Nous avons étendu la dérivation de fronts au cas de l'équation (8) (cf. [56], 2003) et des hamiltoniens convexes plus généraux apparaissent alors, du type $H(p) = \int k(z) e^{-p\cdot z} dz$. D'autres questions se posent. L'écologie mène naturellement à introduire de fortes inhomogénéités dans les coefficients pour prendre en compte la réalité de la nature. On pourra trouver un exemple dans Fitzgibbon *et al* [29], 2001. Des notions d'ondes progressives pour des milieux périodiques sont étudiées dans Henri Berestycki et François Hamel [5], 2002.

3.2 Turing et les rayures du zèbre (1952)

L'autre type de problèmes où interviennent des ÉDP en biologie, extrêmement classique lui aussi, est celui proposé par Alan Turing en 1952 [60] pour décrire l'apparition des taches sur un pelage d'animal. Fisher et KPP ont mis de la diffusion dans le modèle de Verhlust, Turing met de la diffusion dans un système qui est un peu de la même forme que celui de Lotka-Volterra. Il part d'un système dynamique :

$$\frac{d}{dt}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} f_1(u_1, u_2) \\ f_2(u_1, u_2) \end{pmatrix}, \tag{10}$$

dont O est un point attractif stable. Par exemple, le système pourrait décrire les concentrations u_1, u_2 de deux réactifs chimiques dans un mélange, un blanc un noir, en prenant $u_1 = u_2 = 0$ lorsque le mélange est homogène (un beau gris uniforme).

Puis Turing ajoute un terme diffusif :

$$\partial_t \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} - \begin{pmatrix} D_1 \Delta u_1 \\ D_2 \Delta u_2 \end{pmatrix} = \begin{pmatrix} f_1(u_1, u_2) \\ f_2(u_1, u_2) \end{pmatrix}, \tag{11}$$

où D_1, D_2 sont des coefficients constants en espace et en temps (coefficients de diffusion).

Il paraît évident que la diffusion a un effet stabilisant : si la concentration augmente en un point, par diffusion elle va rebaisser. Or, Turing observe que pour certains choix de f_1, f_2 et des constantes D_1, D_2 dans le système parabolique (11), le point O peut être devenu instable. C'est vraiment un phénomène contre-intuitif que Turing a découvert là : *la diffusion peut déstabiliser les équilibres.*

Concrètement, cela veut dire que dans notre mélange chimique l'état homogène n'est plus stable : les réactifs vont spontanément se séparer en des endroits différents, il va y avoir des zones blanches et des zones noires, parfois selon une structure périodique : Turing pensait que ce phénomène pouvait expliquer la genèse des rayures du zèbre, par exemple. En tout cas, il explique des structures observées aujourd'hui par les *chimistes*, et aussi la forme d'amas de bactéries.

La structure mathématique des systèmes à la Turing (11) est encore celle de systèmes paraboliques semi-linéaires.

Ces deux types de problèmes (Fisher-KPP et Turing) sont classiques et de nombreuses applications (morphogenèse, écologie, génétique, cicatrisation) utilisent toujours ce type de modèles avec des questions nouvelles.

4 Un pas vers la théorie de l'évolution

Cette section s'appuie sur deux articles récents où nous avons utilisé systématiquement des systèmes dynamiques en les structurant et en introduisant directement des opérateurs de mutation (cf. [32] et [23]). Cette modélisation est assez habituelle pour des équations simples (des exemples peuvent être trouvés dans le livre de R. Burger ([10], 2000). Mais il ne semble pas y avoir d'étude systématique pour des sytèmes plus complexes (voir toutefois [11] 2002). En effet, la formalisation d'un opérateur de mutation n'est pas toujours nécessaire. Une grande partie de la dynamique adaptive peut se voir sous l'hypothèse de séparation d'échelle de temps, soit d'un point de vue déterministe suivant le papier fondateur [49] , 1996 (voir aussi Odo Diekmann [22] 2002), soit d'un point de vue stochastique comme dans Ulf Dieckmann ([20]1996), ou dans Régis Ferrière ([25], 2000).

4.1 Un exemple de sélection naturelle

Reprenons le modèle de Verhulst :

$$\frac{d}{dt}n(t) = kn(t)\left(M - n(t)\right), \tag{12}$$

Le problème est de savoir comment les paramètres M et k peuvent être sélectionnés par la nature. Rappelons que le paramètre M qui représente ici la limitation de croissance dûe à des ressources limitées.

Amenons dans un certain environnement plusieurs espèces ayant des 'traits' différents, notés x ci-dessous. Ces espèces sont alors caractérisées par des valeurs M, k dépendant du paramètre x (disons $x \in \mathbb{R}$, pour fixer les idées). Comme elles partagent les resources disponibles, le paramètre M dépend de toute la population, quelquesoit son trait. Dans le cas le plus simple on aboutit à

$$\begin{aligned} \partial_t n(t,x) &= a(x)n(t,x) - \rho(t)n(t,x), \\ \rho(t) &= \int n(t,x)dx. \end{aligned} \tag{13}$$

Remarquez que dans le cas où x vit dans un intervalle borné $[\alpha, \beta]$, et où n et a sont indépendants de x, on a $\rho(t) = (\beta - \alpha)n(t)$, et on retrouve bien le modèle de Verhulst.

Mais, en général, le comportement asymptotique est ici moins trivial que dans le modèle de Verhulst. On peut le déterminer en exprimant la solution $n(t, x)$ par une formule intégrale explicite, à laquelle on peut appliquer la méthode de Laplace. Pour cela on doit supposer que a atteint son maximum en un seul point $\overline{x}$:

$$a(\overline{x}) = \max_{x \in \mathbb{R}} a(x) \tag{14}$$

et, sous des hypothèses adéquates, on obtient que

$$n(t, x) \underset{t \to +\infty}{\longrightarrow} \delta(x - \overline{x}) a(\overline{x}). \tag{15}$$

La nature a donc sélectionné la valeur $\overline{a} = a(\overline{x})$. Autrement dit, au bout d'un certain temps, seule a survécu l'espèce qui a le taux de croissance maximum.

4.2 Mutations

En réalité, les choses ne sont pas aussi simples que cela : lors de la division cellulaire la complexité de transcription de l'ADN mène inévitablement à des erreurs qui peuvent conduire à des mutations. De sorte que même si on amène initialement des individus qui ont tous le même taux $a(x)$, leurs descendants auront des taux $a(x')$ qui ne seront pas tous égaux à ce $a(x)$. Supposons que pour une grosse erreur de transcription, l'organisme ne va plus être viable, donc seuls sont possibles les paramètres très proches du paramètre de départ (en terme de trait). Et la sélection naturelle favorisera les lignées pour lesquelles $a(x') > a(x)$, ce qui va se traduire par un *glissement* progressif du taux a dans la population (sauf, bien sûr, si le trait x des individus qu'on a introduits dans l'environnement donné est déjà celui qui maximise a, c'est-à-dire si $x = \overline{x}$, auquel cas les mutations n'auraient aucun effet sur a). On peut compliquer un peu le modèle précédent pour tenir compte de ce phénomène. Remplaçons la première équation dans (13) par :

$$\partial_t n(t, x) = a(x)n(t, x) - \rho(t)n(t, x) + \int K_\varepsilon(y - x)n(t, y)dy, \tag{16}$$

où $\int K_\varepsilon(y - x)n(t, y)dy$ est le nombre d'individus de trait x qui proviennent d'une mutation. Le ε mesure le support de K_ε, c'est-à-dire que $K_\varepsilon(y - x) = 0$ quand $|y - x| > \varepsilon$ (mutations non viables).

L'analyse asymptotique de ce problème lorsque ε est petit (i.e. sur des grands intervalles de temps) peut devenir très complexe pour des systèmes dynamiques très simples en l'absence de mutations. Une méthode classique qui permet de comprendre la dynamique consiste en la *séparation des échelles* de temps (voir Metz *et al* [49] 1996 ; Odo Diekmann, [22]). On va supposer que la sélection agit plus vite que les mutations (du moins, que les mutations substantielles). On décompose donc le temps en petits intervalles de durée Δt. Initialement, le

paramètre est $a(x)$ avec $x = x_0$. Dans l'intervalle Δt, une mutation se produit, changeant x en tout un spectre de valeurs x' ; très vite, un seul trait va dominer (par le principe de sélection précédent), soit $\overline{x}(\Delta t)$: c'est le paramètre qui rend a maximal parmi ceux qui correspondent à des mutations viables. En général x ne peut pas sauter brutalement à la valeur correspondant au maximum absolu de a, et le processus se poursuit : le trait dominant sera successivement $\overline{x}(2\Delta t)$, $\overline{x}(3\Delta t)$, etc. La question est alors de décrire l'évolution du trait dominant $\overline{x}(t)$ (en passant à la limite continue).

On montre dans [23] que la limite $\varepsilon \to 0$ dans (16) mène à nouveau (voir Section 3.1) à des équations de Hamilton-Jacobi qui décrivent l'évolution du trait dominant qui « *glisse* » lentement pour favoriser les mutants mieux adaptés au nouvel environnement que chaque mutation crée.

Notons que d'autres points de vues mathématiques interviennent également dans la théorie de l'évolution : la théorie des jeux, la théorie des équilibres (si on tient compte de phénomènes d'interaction entre plusieurs espèces) est aussi utilisée, ainsi que la théorie du contrôle : les espèces qui persistent sont celles qui maximisent le taux de croissance de la population compte tenu des ressources disponibles ([16], 2002 ; [11], 2002). Un exemple de phénomène que l'on désire expliquer est le suivant. Dans la nature, on trouve typiquement, mais pas seulement loin de là, deux types de croissances : celle des plantes et celle des animaux. Les plantes donnent des graines très tôt, et plus elles grossissent, plus elles donnent de graines : elles peuvent donner de plus en plus de graines pendant toute leur existence. Au contraire, les animaux connaissent d'abord une phase de croissance pendant laquelle ils ne se reproduisent pas, puis une phase reproductive une fois arrivé à maturité. Du point de vue de la théorie du contrôle ([16], 2002), ce sont deux points optimaux, pour deux différents types de ressources. Les arbres ont des racines qui grossissent en même temps qu'eux et leurs ressources sont ainsi proportionnelles à leur volume, tandis que les animaux ont des ressources plus limitées, et plus ils sont gros, plus il leur est difficile de trouver des ressources adaptées à leur taille.

Concluons cette section par une remarque sur une autre façon dont le trait dominant peut « *glisser* » (cf [32]) en supposant un certain déterminisme dans la mutation. Puisque les mutations changent le trait de moins que ε, on peut faire un changement d'échelle pour *zoomer* sur ces mutations. Pour cela, on suppose que $K_\varepsilon(y-x) = \frac{1}{\varepsilon} K\left(\frac{y-x}{\varepsilon}\right)$, où K est une fonction indépendante de ε ; on a donc, pour tout x :

$$\int K_\varepsilon(y-x)dy = \int K\left(\frac{y-x}{\varepsilon}\right)\frac{1}{\varepsilon}dy = \int K(z)dz \equiv \langle K\rangle. \qquad (17)$$

Donc :

$$\begin{aligned}\int \frac{1}{\varepsilon} K\left(\frac{x'-x}{\varepsilon}\right) n(t,x')dx' &= \langle K\rangle n(t,x) \\ &\quad + \int_{\mathbb{R}} K\left(\frac{x'-x}{\varepsilon}\right)(n(t,x')-n(t,x))\,\frac{1}{\varepsilon}dx' \\ &= \langle K\rangle n(t,x) + \varepsilon \int_{\mathbb{R}} K(z)\left(\frac{n(t,x+\varepsilon z)-n(t,x)}{\varepsilon}\right)dz \\ &\approx \langle K\rangle n(t,x) + \varepsilon \int_{\mathbb{R}} zK(z)dz\;\partial_x n(t,x),\end{aligned} \tag{18}$$

si le coefficient dominant $\int_{\mathbb{R}} zK(z)dz = c$ est non nul. Les mutations, qu'on voyait comme des sauts, sont maintenant décrites comme un phénomène continu de *translation dans l'espace des traits* x, et (16) devient :

$$\partial_t n(t,x) = (\langle K\rangle + a(x))\,n(t,x) - \rho(t)n(t,x) + \varepsilon c\partial_x n(t,x). \tag{19}$$

On est conduit à une ÉDP qui est le point de départ des modèles de virulence en modélisation de l'interaction système immunitaire/parasite par l'école de Turin (cf. [4], 2000 ou [19], 2003 par exemple).

5 Populations structurées et équations aux dérivées partielles

On a vu qu'une population structurée est une population dont l'effectif $n(t,x)$ dépend du temps et d'un paramètre x. On a aussi montré comment on pouvait être conduit naturellement à une équation aux dérivées partielles. Le premier exemple historique (à ma connaissance) d'ÉDP est toutefois apparu en démographie. C'est un exemple très simple également qui conduit à une équation linéaire. Je parlerai ensuite d'une application de ce modèle en biologie cellulaire. Des modèles plus intéressants, que l'on ne peut étudier avec des formules explicites, sont aussi couramment employés, on peut les trouver dans [15].

5.1 Vieillissement des populations : équation de McKendrick (1926) et von Fœrster (1959)

Notons $n(t,x)$ la densité de population d'âge $x \geq 0$ à l'instant t. Ignorant dans un premier temps les décès et les naissances, on obtient une première équation d'évolution, très simple :

$$\partial_t n(t,x) + \partial_x n(t,x) = 0. \tag{20}$$

La solution générale de cette équation est évidente, et bien connue depuis d'Alembert (dans un contexte différent) : c'est $n(t,x) = n_0(x-t)$, où n_0 est la population initiale d'âge x : $n_0(x) = n(0,x)$, fonction supposée ici dérivable.

Cette solution exprime que la population vieillit avec le temps, selon une simple translation.

Maintenant, tenons compte des naissances et des décès. Les taux de natalité et de mortalité dépendent de l'âge : notons-les respectivement $b(x)$ (« *birthrate* ») et $d(x)$ (« *death rate* »).

L'équation proposée par McKendrick est la suivante :

$$\partial_t n(t,x) + \partial_x n(t,x) + d(x)n(t,x) = 0, \tag{21}$$

avec comme condition au bord l'expression évidente du nombre de nouveaux-nés en fonction du taux de natalité :

$$n(t,x)|_{x=0} = \int b(x)n(t,x)dx. \tag{22}$$

avec l'hypothèse suivante

$$b \in L^\infty_+(\mathbb{R}^+), \quad d(x) \geq 0, \quad d(x) \to \infty \text{ pour } x \to +\infty. \tag{23}$$

On pourrait aussi prendre des taux $d(t,x)$ et $b(t,x)$ dépendant du temps.

Le modèle de McKendrick et von Fœrster n'est pas un simple modèle-jouet, il est prédictif et est réellement utilisé dans les études démographiques.

L'équation (21) est linéaire. On peut enrichir le modèle pour essayer de prendre en compte certains phénomènes non linéaires. Par exemple, quand la population augmente, les ressources sont plus difficiles à partager et on s'attend à ce que le taux de mortalité augmente, ce qui conduit à faire dépendre $d(x)$ de n. On peut aussi faire dépendre $b(x)$ de n (le taux de natalité devrait diminuer quand la population augmente). La littérature sur ces modèles non linéaires est assez abondante, mais aussi très disparate, et il serait bon que des mathématiciens reprennent le sujet de façon systématique : une vision un peu unificatrice serait bien utile.

Revenons à l'équation (21) elle-même. Sa linéarité a permi de l'étudier très tôt avec les outils de la théorie spectrale et des semi-groupes. On obtient le théorème suivant :

Théorème 1 *Sous l'hypothèse (23), il existe un réel λ, une fonction $x \mapsto \varphi(x) \geq 0$, et une fonction $x \mapsto N(x) > 0$, tels que*

$$\int \left| e^{-\lambda t} n(t,x) - m_0 N(x) \right| \varphi(x)dx \underset{t\to+\infty}{\longrightarrow} 0, \tag{24}$$

avec m_0 défini par $\int n_0(x)\varphi(x)dx = m_0 \int N(x)\varphi(x)dx$.

Ce phénomène, appelé désynchronisation (cf [14], 2001), exprime une perte rapide de mémoire concernant la distribution initiale, seule la quantité moyenne m_0 reste. Ce comportement est illustré dans la figure 1, par une simulation numérique. En fait une décroissance exponentielle peut être démontrée assez généralement. Le signe de λ dépend du nombre de naissances par rapport au

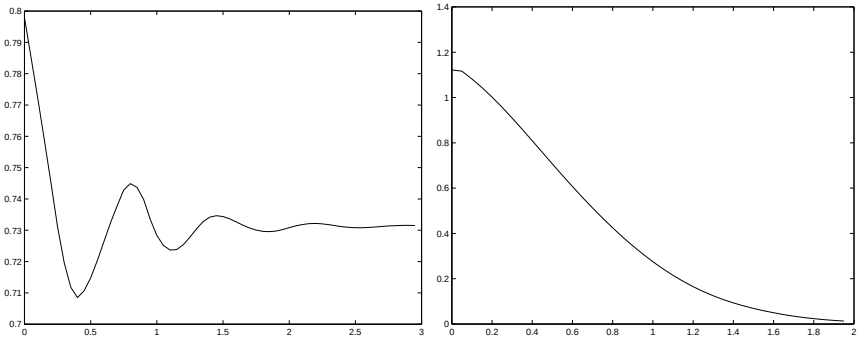


FIG. 1 – Solution de l'équation de renouvellement. Gauche : effectif total en fonction du temps. Droite : l'état stationnaire $N(x)$.

nombre de décès. Derrière ce théorème, il y a le fait que l'opérateur stationnaire $\partial_x + d(x)$ préserve la positivité, donc il a une première valeur propre (c'est λ) associée à une fonction propre positive (qui est $N(x)$)

$$\partial_x N(x) + (\lambda + d(x))N(x) = 0, \qquad x \geq 0,$$
$$N(x)|_{x=0} = \int b(x)N(x)dx.$$

On peut démontrer le théorème directement, mais des opérateurs plus élaborés utilisent des méthodes à la Krein-Milman. Récemment nous avons proposé avec Leonid Rizhik et Stéphane Mischler ([50], 2002) une méthode d'entropie qui permet d'aborder des problèmes plus généraux (structuration en maturité, diffusion). Des exemples plus complexes où ce type de méthode s'applique provient de la division cellulaire, des effets de maturation ou de modèles de cycle cellulaire.

5.2 Transposition du modèle aux cycles cellulaires

Les cellules, lors de la division cellulaire, suivent un certain cycle : dans la première phase, G_1, la cellule grossit. Quand elle atteint une certaine taille critique, un certain nombre de phénomènes se déclenchent et on entre dans la phase S (synthèse d'ADN) : l'ADN se duplique. Ensuite vient la phase G_2, qui est une phase de réparation de l'ADN, puis arrive la phase de mitose proprement dite, dans laquelle les deux brins d'ADN se séparent et la cellule se divise en deux nouvelles cellules.

Pour décrire la population cellulaire dans les différentes phases, on peut appliquer l'équation de McKendrick et von Fœrster, avec quelques modifications. Ici, la variable pertinente ne va pas être l'âge de la cellule, mais son âge *dans la phase* du cycle où elle se trouve. C'est un âge biologique : il dit en quel endroit du cycle se trouve la cellule (par exemple, à mi-parcours). Pour faire le lien avec le temps physique, il faut introduire une nouvelle quantité : la

vitesse de parcours du cycle, qui n'est pas uniforme. Si on note i le numéro de la phase (dans le cycle de la mitose, $i \in \{1, 2, 3, 4\}$), $v_i(t, x)$ la vitesse de parcours du cycle dans la phase i, $K_{i,i+1}(x)$ le taux de passage de la phase i à la suivante pour une cellule qui est dans la phase i depuis le temps x, on a :

$$\partial_t n_i(t, x) + \partial_x \left(v_i(t, x) n_i(t, x)\right) + d_i(x) n_i(t, x) + K_i(x) n_i(t, x) = 0, \qquad (25)$$

avec la condition au bord

$$v_i(0) n_i(t, 0) = \int K_{i-1}(x) n_{i-1}(t, x) dx. \qquad (26)$$

Outre le fait que la vitesse de parcours des cycles $v_i(t, x)$ est maintenant variable, la différence avec l'équation de McKendrick historique, c'est qu'à la place des naissances on a l'arrivée dans la phase depuis la phase précédente (K_{i-1}), et qu'en plus des décès ($d_i(x)$) on a les départs vers la phase suivante (K_i).

L'équation et la condition au bord sont du même type que (21) et (22), c'est-à-dire toujours linéaire, et on peut établir un théorème similaire au précédent (voir [15], 2003).

Là aussi, on peut introduire des non-linéarités. *A priori*, $d_i(x)$ et $v_i(t, x)$ devraient dépendre des ressources disponibles, et donc de la population.

Dans le contexte de l'apparition des cancers, l'une des questions que se posent les biologistes est de caractériser les différences de cycles entre cellules saines et cancéreuses. Les cellules cancéreuses ont des cycles plus rapides, et qui semblent beaucoup moins contrôlés par les divers signaux émis par le système nerveu central. On aimerait comprendre par exemple pourquoi les cellules saines et les cellules cancéreuses réagissent de façon différente à l'horloge circadienne. C'est un phénomène qui est observé cliniquement : les chimiothérapies sont plus efficaces quand elles sont administrées à certaines heures qu'à d'autres (c'est le principe de la chronothérapeutique). Ces thérapeutiques sont généralement toxiques pour les cellules et à certains moments de la journée les cellules saines sont dans une phase où elles se défendent mieux alors que les cellules cancéreuses sont désynchronisées. L'idée des praticiens (mais elle n'est pas démontrée) est que les cycles cellulaires des cellules saines sont réglés par l'horloge circadienne, à travers un certain nombre de points de contrôle. Dans (25), les noyaux K_i décrivant les transitions d'une phase à la suivante dépendraient non seulement de l'âge x dans la phase, mais aussi du temps t, *via* certaines protéines, enzymes ou hormones qui vont contrôler les transitions, et il y aurait donc une interaction entre l'horloge et le cycle. On pourra consulter le volume édité par Francis Lévi à ce propos ([43], 2002). Notons enfin que les méthodes expérimentales permettent maintenant de mesurer les contenus en ADN des cellules au cours du cycle ce qui change la variable âge en une variable de contenu en ADN (voir Basse *et al*, [3] , 2003).

6 Mouvements cellulaires : le modèle chimiotactique de Keller-Segel (1970)

6.1 La chimiotaxie

Presque toutes les cellules (des plus simples, comme les bactéries, aux plus complexes, commes les amibes ou les cellules endothéliales humaines) sont naturellement capables de se mouvoir.

Un cas particulièrement intéressant est celui du *dictyostelium discoideum* : c'est une amibe, qui vit dans l'humus des forêts. Elle se nourrit de bactéries et autres micro-organismes. Tant que la nourriture est abondante, elle évolue en solitaire. Mais quand la nourriture vient à manquer, *dictyostelium discoideum* émet un chemo-attractant, et elle se déplace en projettant un pseudopode vers les concentrations les plus fortes de chemo-attractant, pour la tirer. Comme le chemo-attractant attire aussi bien l'amibe qui l'a émis que ses voisines, *les amibes s'attirent les unes les autres*. Ce qui conduit à la création d'amas tridimensionnels, qui ressemblent à de gros vers. Ces vers se dirigent vers les sources de lumière et de chaleur, comme le ferait un organisme unique, et ils atteignent la surface du sol. (Ce comportement curieux fait que *dictyostelium discoideum* est qualifiée d'*amibe sociale*.)

Ce mode de locomotion, par l'intermédiaire d'un produit chemo-attractant, s'appelle la chimiotaxie.

6.2 Le modèle de Keller-Segel

Evelyn Fox Keller et Lee Segel ([41], 1970 ; [42], 1971) ont proposé un modèle de chimiotaxie (justement, à propos du *dictyostelium discoideum*) qui, sous sa forme la plus simplifiée, est le suivant. Notons $n(t,x)$ la densité de cellules (au point $x \in \mathbb{R}^3$, à l'instant t), et $c(t,x)$ la concentration de chemo-attractant. Alors :

$$\begin{aligned} \partial_t n - \Delta n &= -\operatorname{div}\left(n \nabla_x c\right), \\ -\Delta c &= n. \end{aligned} \tag{27}$$

La première équation décrit le mouvement des cellules : elles vont un peu dans toutes les directions (termes de diffusion : $\partial_t n - \Delta n$), mais elles détectent le gradient de concentration du chemo-attractant, ce qui crée une direction privilégiée, d'où le terme de dérive $\operatorname{div}(n\nabla_x c)$. Dans des versions plus sophistiquées du modèle, le terme de dérive est $-\operatorname{div}(n\nabla_x \varphi(c))$, où la « *sensibilité* » φ est une fonction lisse et strictement croissante sur $]0,+\infty[$, caractérisant le potentiel attractif du chemo-attractant.

La seconde équation décrit la diffusion du chemo-attractant, quand on suppose qu'il est émis pas les cellules elles-mêmes. On devrait avoir $\varepsilon\partial_t c - \Delta c = n$, mais comme c'est une diffusion moléculaire, elle est beaucoup plus rapide que celle des cellules (qui sont des objets beaucoup plus gros), et on peut négliger le terme en $\partial_t c$. On néglige aussi, dans le second membre, un terme de

dégradation du chemo-attractant, $-\alpha c$, qui ne jouerait un rôle que sur des durées plus importantes que celles pendant lesquelles on peut observer les populations.

Les conditions au bord sont celles de Neumann (flux nul pour n et c).

Le système (27) est un système parabolique non linéaire, qui rappelle un peu celui de Turing :

$$\begin{aligned} \partial_t u_1 - D_1 \Delta u_1 &= f_1(u_1, u_2), \\ \partial_t u_2 - D_2 \Delta u_2 &= f_2(u_1, u_2), \end{aligned} \tag{28}$$

sauf qu'ici il est de la forme

$$\begin{aligned} \partial_t u_1 - \Delta u_1 &= -\operatorname{div}(u_1 \nabla_x u_2), \\ \varepsilon \partial_t u_2 - \Delta u_2 &= u_1, \end{aligned} \tag{29}$$

(avec $u_1 = n$, $u_2 = c$ et $\varepsilon \partial_t u_2$ négligeable). La différence est que dans les systèmes à la Turing, les couplages se font par des termes d'ordre 0 ($f_1(u_1, u_2)$ et $f_2(u_1, u_2)$), tandis qu'ici l'un des termes de couplage ($-\operatorname{div}(u_1 \nabla_x u_2)$) est d'ordre 1. En fait, le système de Keller-Segel est l'un des systèmes paraboliques non linéaires les plus simples qu'on puisse écrire après ceux du type de Turing.

Mais si le système est simple à écrire, ses solutions sont très compliquées, et il y a énormément de questions encore ouvertes. Voyons d'abord ce qu'on sait.

On suppose que $n_0 \geq 0$ et $n_0 \in L^1(\mathbb{R}^d)$. En intégrant (27) sur $\mathbb{R}^d$, on obtient

$$\partial_t \int n(t,x) dx = 0,$$

c'est-à-dire la conservation de la masse (en fait : du nombre de cellules). On pose :

$$m_0 = \int n(t,x) dx = \text{cte} \quad \text{(conservation de la masse)}. \tag{30}$$

En dimension 1, il y a toujours des solutions globales. Les choses difficiles commencent avec la dimension 2.

6.3 Dimension 2. L^1 est l'espace critique

Notons $n_t(x) = n(t,x)$. Soit Ω un ouvert connexe de $\mathbb{R}^2$ dans lequel x va vivre. Je vais me limiter ici au cas où Ω est soit tout $\mathbb{R}^2$, soit borné avec un bord lisse (c'est-à-dire aussi régulier que nécessaire).

Le théorème suivant résume certains des résultats obtenus par Willi Jäger et Stephan Luckhaus ([37], 1992), Toshitaka Nagai ([53], 1995 ; [54], 1997), Miguel Herrero et Juan Velázquez ([34], 1996 ; [35], 1997), et Piotr Biler ([7], 1998) :

Théorème 2 (Jäger-Luckhaus, Nagai, Biler, Herrero-Velázquez) *On considère le système (27). Alors*

1. *Il existe une constante C_* telle que pour tout $m_0 \leq C_*$, il existe une solution globale ayant même régularité que la donnée initiale : pour tout $p > 1$,*

$$n_0 \in L^p(\Omega) \Longrightarrow n_t \in L^p(\Omega). \tag{31}$$

2. *Il existe une constante C^* telle que pour tout $m_0 > C^*$, il y a explosion en temps fini T^* :*
$$\| n_t \|_{L^p(\Omega)} \underset{t \to T^*}{\longrightarrow} +\infty \tag{32}$$
pour tout $p > 1$ (le système est critique pour $\| \cdot \|_{L^1(\Omega)}$).
3. *Si $\Omega = \mathbb{R}^2$, ou si Ω est un disque $D(0,R)$ de $\mathbb{R}^2$, on a $C_* = C^* = 8\pi$: cela provient d'une constante de Sobolev. (Sinon, on ne sait pas si $C_* = C^*$.)*
4. *Dans le cas d'un disque $\Omega = D(0,R)$, ce ne sont pas seulement les normes $\| n_t \|_{L^p(\Omega)}$ qui explosent, il y a en fait convergence vers un pic de Dirac : les solutions radiales vérifient*
$$n(t,x) \underset{t \to T^*}{\longrightarrow} 8\pi\delta(x) + \text{un reste}, \tag{33}$$
où le reste est dans un espace L^p et connu explicitement.

Ce dernier résultat est beaucoup plus précis que celui de l'explosion des normes L^p, et aussi beaucoup plus intéressant, du point de vue biologique, car dans les expériences, on voit effectivement la masse se concentrer dans un tout petit volume : c'est ce qu'on appelle le « *chemotactic collapse* ». Mathématiquement, on n'a pas encore le résultat analogue pour un ouvert Ω arbitraire.

On voit qu'il y a déjà une grande complexité en dimension 2. En dimension 3 le blow-up est bien plus complexe.

6.4 Dimension 3. $L^{3/2}$ est l'espace critique

Soit maintenant Ω un ouvert connexe de $\mathbb{R}^3$. Contrairement à la dimension 2, il peut y avoir explosion même si la masse m_0 est petite. Ce n'est plus $m_0 = \| n_0 \|_{L^1(\Omega)}$ qui joue un rôle discriminant ici, c'est plutôt (essentiellement) $\| n_0 \|_{L^{3/2}(\Omega)}$:

Théorème 3 *Considérons le système (27) en dimension 3. On a*

1. *Si $\| n_0 \|_{L^{3/2}(\Omega)}$ est assez petite, alors il existe des solutions globales.*
2. *Il existe une constante C telle que si*
$$\int |x|^2 n_0(x)dx \leq C \left(\int n_0(x)dx \right)^3 \tag{34}$$
(où $|x|$ désigne la norme euclidienne dans $\mathbb{R}^3$), alors il y a explosion en temps fini. Et il y a de nombreuses formes d'explosion ([8], 1999) .

Remarque 1 *La condition* (34) *implique que l'état initial (n_0) est fortement concentré, donc $\| n_0 \|_{L^{3/2}(\Omega)}$ reste supérieur à une certaine constante : c'est incompatible avec le point* 1 *(garantissant l'existence de solutions globales). Donc il n'y a pas de contradiction entre les deux assertions du théorème. Cela dit, la condition* (34) *n'est pas très naturelle, on aimerait la remplacer par l'hypothèse que $\| n_0 \|_{L^{3/2}(\Omega)}$ n'est pas trop petite, mais sous cette seule hypothèse on n'a pas de démonstration de l'explosion en temps fini.*

6.5 Une idée de la preuve des théorèmes d'existence

Je ne vais pas donner ici la preuve de ces théorèmes, je vais seulement en donner l'idée générale. Il y a une compétition entre la diffusion (liée au laplacien), qui tend à tout étaler, et l'aspect hyperbolique qui tend à tout concentrer dans les zones de grand gradient. On peut préciser cela en regardant l'évolution des normes L^p.

Plaçons-nous dans $\mathbb{R}^d$. On commence par multiplier l'équation (27) par n^{p-1}. On obtient :

$$\frac{d}{dt}\int \frac{n^p}{p}dx = -(p-1)\int n^{p-2}|\nabla_x n|^2 dx + \int n\nabla_x n^{p-1}\ \cdot \nabla_x c\ dx. \tag{35}$$

Dans le membre de droite, le terme négatif provient du laplacien par intégration par parties et correspond à la tendance à la diffusion ; le terme positif correspond à la tendance à la concentration. Pour les comparer, on va réécrire un peu le second :

$$\begin{aligned}
&\int n\nabla_x n^{p-1}\cdot\nabla_x c\ dx = (p-1)\int n^{p-1}\nabla_x n\cdot\nabla_x c\ dx = \frac{p-1}{p}\int \nabla_x n^p\cdot\nabla_x c\ dx\\
&= -\frac{p-1}{p}\int n^p\Delta c\ dx \quad \text{(intégration par parties)}\\
&= \frac{p-1}{p}\int n^{p+1}\ dx \quad \text{(puisque } \Delta c = -n\text{)}.
\end{aligned} \tag{36}$$

Suivant les valeurs de d et de p, c'est le terme de diffusion ou le terme de concentration qui va gagner.

Pour $d = 2$, on a

$$\int n^{p+1}\ dx \le C(p)\int n^{p-2}|\nabla_x n|^2 dx\int n\ dx, \tag{37}$$

d'où l'existence de solutions globales pour des masses petites.

Pour $d \ge 3$, on a

$$\int n^{p+1}\ dx \le C\int n^{p-2}|\nabla_x n|^2 dx\int n^p\ dx, \tag{38}$$

ce qui n'est possible qu'avec $p = \frac{d}{2}$ (alors que pour $d = 2$ l'inégalité était valable pour tout p). L'exposant $p = \frac{d}{2}$ est critique, et le système est critique pour la norme $\|\cdot\|_{L^{d/2}}$.

Ceci donne quelques idées concernant l'existence de solutions globales. Maintenant, pour la partie explosion, on peut utiliser des arguments dûs à Nagai : on montre par l'absurde que, sous des hypothèses de régularité et décroissance à l'infini, le deuxième moment en x de n devient négatif (ce qui est bien sûr impossible), sauf si la masse est assez grande (si $d = 2$), ou si la condition (34) (si $d = 3$) est vérifiée.

En dimension 3, la topologie joue un rôle. (S'il y a des trous dans Ω, la criticité peut changer.)

6.6 Points d'explosion

Pour un ouvert Ω quelconque, très peu de calculs ont été faits (aussi bien en dimension 2 qu'en dimension 3). Biologiquement, on observe une concentration des cellules au cours du temps. Il est naturel de penser que cette condensation se fait au point où c (la concentration du chemo-attractant) est maximale : les cellules devraient aller vers le point qui les attire le plus fortement, et ce point doit être celui où la concentration du chemo-attractant est la plus grande.

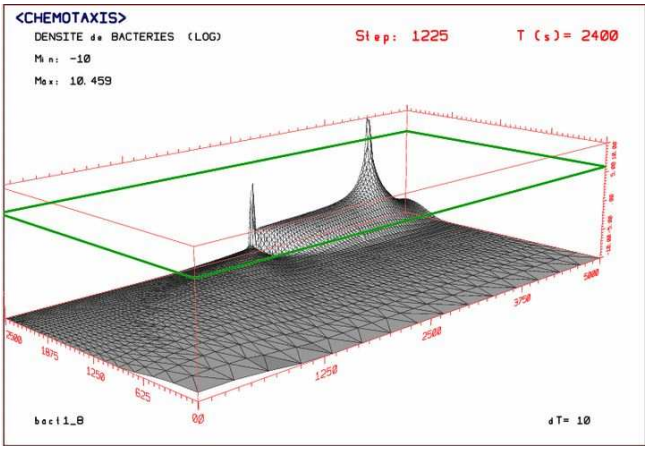


FIG. 2 – Exemple de solution exhibant un collapse chimiotactique pour le modèle de Keller-Segel (27) dans un rectangle. En échelle logarithmique. (Calculs de A. Marrocco [48], INRIA, projet BANG).

Pourtant, il semble bien que ce ne soit pas le cas, même déjà en dimension 2 : Americo Marrocco ([48], 2003) a fait les calculs dans le cas d'un rectangle, et ses résultats montrent que l'explosion se produit en des points du grand axe de symétrie du rectangle (ici deux ; là, avec un rectangle un peu plus grand, quatre), mais pas au centre, qui est le point où il devrait y avoir le plus de chemo-attractant, par raison de symétrie (initialement la densité est uniforme). De plus, il y a de la masse qui se concentre en chacun des points d'explosion (ce n'est pas seulement une explosion en norme L^p, c'est bien du « *chemotactic collapse* »). Apparemment, l'explosion se produit toujours sur des points où on a l'impression d'avoir un manque de régularité du $\nabla_x c$. Une illustration est donnée dans la figure 2. Ces simulations semblent montrer que, en dimension 2, l'explosion dans le cas d'un rectangle est loin d'être aussi simple que dans le cas d'un disque et le point de concentration n'est pas le centre du rectangle.

6.7 Ondes progressives

Un autre phénomène observé par les biologistes est celui des ondes progressives : on voit des migrations de cellules le long d'anneaux. Est-ce que les

modèles du type de Keller-Segel sont capables d'en rendre compte ? La question n'est pas clairement tranchée. Le modèle de Keller-Segel, qui a pourtant été inventé pour cela, ne rend compte correctement des ondes progressives que dans des cas extrêmement particuliers parce que les sensitivités y sont très singulières (consulter [58] par exemple). La tendance actuellement est de penser que le modèle est trop simple pour incorporer ce genre de complexité, et qu'il faut lui adjoindre des termes ou des équations supplémentaires. Une idée possible serait des substances chemorépulsives à courte porté. Une autre idée ([9] , 1998) serait que le chemo-attractant n'est pas émis par la bactérie, mais qu'il résulte d'une réaction chimique entre un produit émis par la bactérie et un produit présent dans son environnement. Serguei Esipov et James Shapiro [24] proposent encore un autre mécanisme. Enfin H. Schwetlick, A. Stevens [58] présentent une étude mathématique détaillée de la question, en considérant des termes de reproduction et de mort.

7 Aspect mésoscopique : le modèle de Othmer-Dunbar-Alt (1988) pour les déplacements d'*Escherichia Coli*

Jusqu'ici seules les équations macroscopiques ont été abordées. Il est parfois utile de regarder certains aspects microscopiques ou mésoscopiques. On va présenter cette approche aussi bien pour la chimiotaxie mais des approches microscopiques pour l'angiogenèse ont aussi été proposée (voir par exemple Tomas Alarcon, Helen Byrne et Philip Maini [1], 2002).

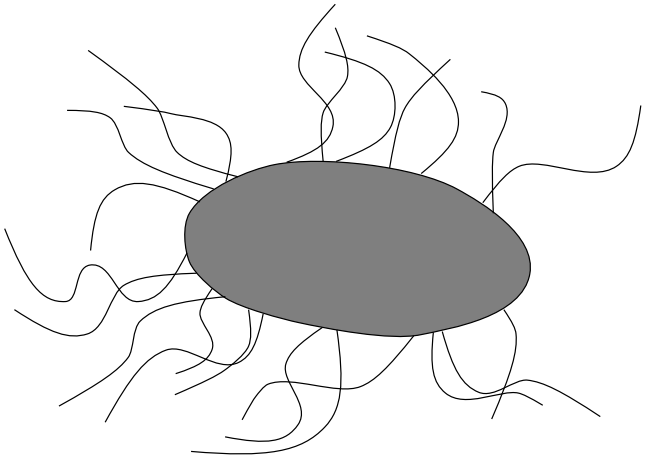


FIG. 3 – La bactérie E. Coli est pourvue de flagelles qui peuvent la propulser comme des hélices.

Observons le déplacement de la célèbre bactérie *Escherichia Coli* (en abrégé *E. Coli*). Elle bouge à l'aide de flagelles extérieures (voir figure 3). Le

mouvement de ces flagelles est très intéressant. Il y a deux temps. Dans le premier temps, toutes les flagelles tournent dans le même sens, un peu comme une hélice, et propulsent la bactérie : elle saute. Au bout d'un temps de l'ordre de une seconde, les récepteurs qui activent les flagelles sont saturés, les flagelles se désynchronisent et tournent dans tous les sens : alors la bactérie tourne sur elle-même. Au bout d'un certain temps (très court par rapport à la phase de saut), les flagelles se remettent en phase, et la bactérie saute de nouveau, dans une autre direction et avec une autre vitesse.

La situation peut être modélisée de la façon suivante : à un moment, la bactérie a une vitesse ξ', puis assez brutalement elle passe à la vitesse ξ, dont la norme et la direction sont à peu près aléatoires, à ceci près qu'il y a un petit biais : une sorte d'effet de mémoire des récepteurs qui activent les flagelles.

Notons $f(t,x,\xi)$ la densité de bactéries qui bougent avec la vitesse ξ au point x à l'instant t, et $k(c;\xi',\xi)$ le taux de transition de ξ' à ξ, pour une concentration c de la substance qui attire la bactérie. Ce taux $k(c;\xi',\xi)$ dépend de x à travers c (qui est une fonction de x). Posons :

$$K[f] = \int k(c;\xi',\xi) f(t,x,\xi') d\xi' - \int k(c;\xi,\xi') d\xi' \; f(t,x,\xi). \tag{39}$$

C'est le nombre de bactéries qui adoptent à l'instant t la vitesse ξ, moins le nombre de bactéries qui l'abandonnent au même instant. Ce nombre doit être égal à $\frac{d}{dt} f(t,x,\xi) = \partial_t f + \xi \cdot \nabla_x f$, d'où l'équation :

$$\partial_t f + \xi \cdot \nabla_x f = K[f]. \tag{40}$$

Ce modèle cinétique est très ancien (il est utlisé couramment pour décrire le transport des neutron et prend souvent le nom d'équation de scattering qui exprime bien sa généralité). Dans le contexte du mouvement bactérien il a été proposé par Hans Othmer, Steven Dunbar et Wolfgang Alt ([55], 1988). Sa spécificité provient de la forme du noyau $k(c;\xi,\xi')$ qui est à l'origine de la non linéarité. Il n'est pas symétrique en (ξ,ξ'). Un exemple simple est

$$k(c;\xi,\xi') = 1 + c(x - \xi\varepsilon). \tag{41}$$

Dans ce cas les sauts de vitesse sont entièrement déterminés par la concentration que la bactérie a vue juste un peu avant (c'est-à-dire la concentration là où elle était il y a un temps ε).

Bien sûr, comme dans le cas de la chimiotaxie, on couple l'équation avec la densité totale de particules (ici, de bactéries). On obtient un système non linéaire :

$$\begin{aligned} &\partial_t f + \xi \cdot \nabla_x f = K[f], \\ &-\Delta c(t,x) = \int f(t,x,\xi) d\xi = n(t,x), \end{aligned} \tag{42}$$

Voici un théorème que j'ai obtenu en collaboration avec Fabio Chalub, Peter Markowich et Christian Schmeiser, [12] (2002) :

Théorème 4 *[12] Considérons le système* (42) *avec le noyau* (41).

1. *Ce système a des solutions globales* $f \in L^\infty(0,T;L^1 \cap L^\infty(\mathbb{R}^d))$, *for all* $T > 0$.
2. *Supposons que* ξ *ne puisse décrire qu'une boule de* $\mathbb{R}^d$ *(en pratique les vitesses des bactéries sont évidemment bornées),* x *décrivant tout l'espace* $\mathbb{R}^d$. *Changeons d'échelles, de façon à transformer* (42) *en :*

$$\begin{aligned} \partial_t f_\varepsilon + \frac{1}{\varepsilon}\xi \cdot \nabla_x f_\varepsilon &= \frac{1}{\varepsilon^2} K[f_\varepsilon], \\ -\Delta c_\varepsilon(t,x) &= \int f_\varepsilon(t,x,\xi)d\xi = n_\varepsilon(t,x), \end{aligned} \tag{43}$$

avec le ε *qui figure dans le noyau* (41). *(On accélère le temps : c'est ce qu'on appelle une approximation de diffusion.) Alors il existe un temps* T^* *tel que pour tout* $t \leq T^*$, *on ait des limites indépendantes de la vitesse :*

$$\begin{aligned} f_\varepsilon(t,x,\xi) &\underset{\varepsilon\to 0}{\longrightarrow} n(t,x), \\ c_\varepsilon(t,x) &\underset{\varepsilon\to 0}{\longrightarrow} c(t,x), \end{aligned} \tag{44}$$

et la densité et la concentration limites (n,c) *satisfont au système de la chimiotaxie* (27) *avec toutefois des coefficients de diffusion et de sensibilité plus complexes.*

Le point le plus intéressant du théorème est l'existence de solutions globales, la partie asymptotique s'appuyant sur des arguments classiques. Il est surprenant puisque l'équation de transport, hyperbolique, n'a pas l'effet régularisant du Laplacien et le modèle de Keller-Segel explose. Comment démontre-t'on donc qu'il n'y a pas d'explosion en temps fini ? La difficulté provient de la nonlinéarité du système : dans $K[f]$, f est multiplié par k, qui est une fonction affine de c vu (41), et c est lié linéairement à f par la seconde équation (42). Il y a donc des termes quadratiques. Ce qui permet de gagner de la régularité par rapport au modèle de Keller-Segel est le fait que la nonlinéarité utilise la quantité c directement et non pas ∇c. Il faut toutefois utiliser des arguments spécifiques pour arriver à conclure et la démonstration n'a encore qu'une portée limitée.

Voici des exemples de problèmes ouverts :

Problème 1

1. *Si au lieu de prendre le noyau* (41) *on suppose seulement que*

$$|k| \leq \text{cte}\,(\,\|\,c\,\|_\infty + 1)\,, \tag{45}$$

existe-t-il encore forcément une solution globale ? La méthode qui a permis de démontrer le théorème 4 ne marche plus dans ce cas, et on ne sait pas s'il existe une solution globale.

2. *Si, au lieu de dépendre directement de* c *comme dans* (41), k *dépendait du gradient* $\nabla_x c$ *(c'est une hypothèse concurrente de celle de l'effet de retard), qu'est-ce qui se passerait ? Ici, la conjecture est qu'il aurait explosion en temps fini, mais on ne sait pas le démontrer.*

8 L'initiation de l'angiogenèse

Une tumeur cancéreuse croît, et quand elle atteint une certaine taille critique (quelques centaines de cellules), elle devient trop grosse pour être normalement alimentée par le réseau sanguin capillaire. La raison en est évidente : l'approvisionnement venant de l'extérieur est proportionnel à la surface de la tumeur, ses besoins sont proportionnels à son volume. Les cellules non approvisionnées meurent. Mais, en mourant, elles émettent des produits chemo-attractants (entre autre les VEGF, vascular endothelial growth factors) qui attirent les vaisseaux sanguins et les détournent vers la tumeur (en les ramifiant). Ces VEGF ont deux effets :

- Ils vont activer les cellules endothéliales des vaisseaux sanguins (qui forment les parois des vaisseaux) et remettre en route leur cycle de division cellulaire. (La plupart de nos cellules endothéliales sont endormies.)
- Ils vont réveiller aussi la mobilité de ces cellules.

Le résultat est la naissance de nouveaux capillaires qui vont se développer vers et dans la tumeur : celle-ci est alors complètement vascularisée, et peut recommencer à se développer.

Les modèles proposés pour rendre compte de ce phénomène sont extrêmement compliqués (voir [13], 1996 ou [44] , 2001 par exemple). Ils contiennent de nombreuses équations. Que peut-on isoler pour essayer d'y voir un peu plus clair ?

D'abord, fondamentalement, le phénomène est à peu près le même que celui de la chimiotaxie. Mais il y a une grosse différence : c'est que le produit chemo-attractant n'est pas émis par les cellules attirées (les cellules endothéliales des vaisseaux sanguins), il est émis par une source extérieure (la tumeur). Si on note $n(t,x)$ la densité de vaisseaux capillaires, un système (simplifié au maximum) peut être :

$$\begin{aligned} &\partial_t n - \Delta n = -\operatorname{div}\left(n\nabla_x c\right), \\ &\partial_t c = -cn. \end{aligned} \tag{46}$$

La seconde équation est assez différente de celle de Keller-Segel. Le second membre n'est plus linéaire, il n'y a plus d'aspect diffusif $(-\Delta c)$ sur c — qui avait un effet régularisant bien commode. Mais, surtout, maintenant, *plus n est grand, plus la concentration du chemo-attractant va diminuer*. C'est le contraire de ce qui se passait dans le système de Keller-Segel, où n était la *source* de c $(-\Delta c = n)$. Cette seconde équation traduit, bien sûr, le fait qu'ici les vaisseaux *consomment* du chemo-attractant. Et cette sorte de changement de signe dans la relation entre n et c a une conséquence importante : c'est que sous des conditions raisonnables *il existe toujours des solutions faibles* :

Théorème 5 *([17], [18]) Supposons que $c_0 \in L^\infty$ et $\nabla_x c_0 \in L^2$. Alors :*

1. *Il existe toujours des solutions faibles, avec $\sqrt{n}\nabla_x c \in L^2_{t,x}$ et $\nabla_x \sqrt{n} \in L^2_{t,x}$.*

2. *En dimension $d = 2$, toutes les normes $\| \cdot \|_{L^p}$ sont propagées (il n'y a pas d'explosion, pas d'extinction). En dimension $d = 3$, si $\| n_0 \|_{L^{3/2}}$ est petite, les solutions propagent cette norme :*

$$\forall t, \quad \| n_t \|_{L^{3/2}} < +\infty. \tag{47}$$

Notons également que la dimension 1 a été traitée, pour des données grandes, par Marco Fontelos, Avner Friedman et Bei Hu, ([31], 2002) et ces auteurs montrent que les solutions sont régulières dans des espaces $C^{2,\alpha}$.

Sans entrer dans les détails, la difficulté principale est de montrer que $\nabla_x c$ existe. Comme il n'y a plus ici d'aspect diffusif sur c, donc pas d'effet régularisant, cela ne peut résulter que de la propagation d'une régularité initiale.

9 La vasculogenèse

Je vais conclure par un problème plus récent, qui va me permettre de préciser ce que les biologistes attendent des mathématiciens actuellement. Ils n'attendent pas des théorèmes d'existence ou des théorèmes asymptotiques. Ils attendent des modèles qui vont reproduire les faits qu'ils observent. Leur objectif est une meilleure compréhension des phénomènes, ou de poser de nouvelles questions pour réaliser de nouvelles expériences qui vont leur permettre d'optimiser l'efficacité des substances qu'ils vont utiliser ou, du moins, comprendre ce qui se passe.

Par exemple, le problème de la vasculogenèse est de comprendre la formation de ces vaisseaux capillaires qui vont se diriger dans certaines directions. Pour cela, l'équipe autour de Massoud Mirshahi (Université Pierre et Marie Curie), isole des cellules endothéliales humaines et ils les font croître en grand nombre sur une substance spéciale, le *matrigel*. Le temps de reproduction est normalement de l'ordre de 24 heures, ou même 48 heures, et par génie génétique ils obtiennent des temps plus courts (c'est seulement tout récemment que cela est devenu techniquement possible). Les cellules s'agglomèrent, et au bout d'environ 6 heures, on voit apparaître des formes, des motifs qui, *in vivo*, seraient des vaisseaux capillaires. Ces formes ne sont pas statiques (elles évoluent). Au début, elles sont constituées de cellules à peu près rondes, ou légèrement oblongues, et bien réparties mais petit à petit les cellules se concentrent et cela crée des structures vides, elles se déforment, s'allongent également.

Les modèles de type *chimiotaxie* ne donnent pas du tout ce genre de structures. Est-il possible de trouver d'autres modèles mathématiques, qui s'appliquent à ce cas et expliquent ces structures ?

Plusieurs réponses ont été proposées récemment. L'école de Turin ([59], 2003) propose un modèle de type hyperbolique : dans l'équation

$$\partial_t n - \Delta n = -\operatorname{div}(n \nabla_x c)$$

de Keller-Segel, ils suppriment le laplacien (donc il n'y a plus de diffusion, c'est de la pure propagation) ; et dans le terme de dérive $-\operatorname{div}(n\nabla_x c)$, ils remplacent

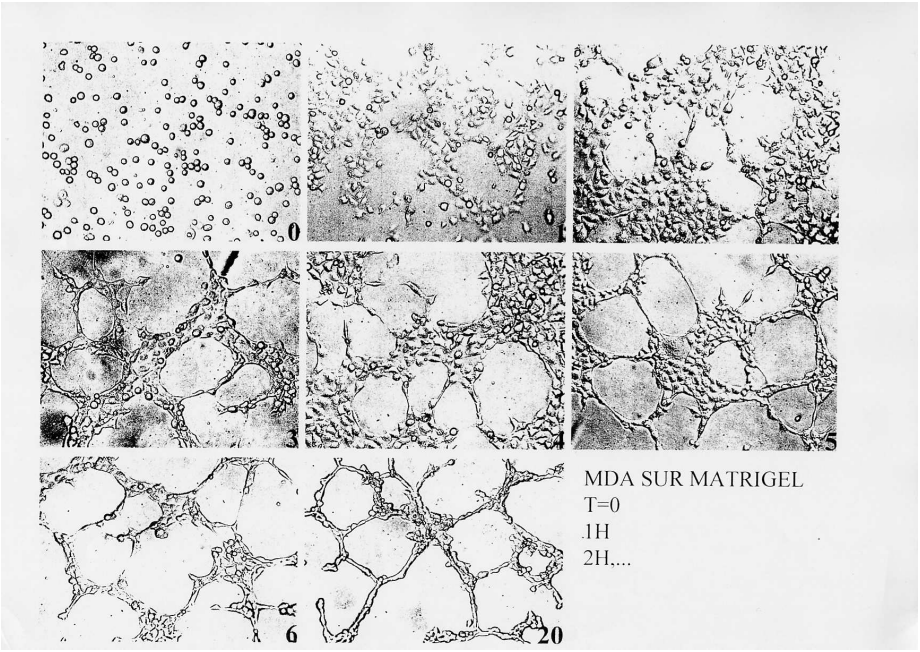


FIG. 4 – Mouvement de cellules endothéliales humaines sur matrigel. (M. Mirshahi, Univ. P. et M. Curie).

$\nabla_x c$ par un champ de vitesses $u(t,x)$. Ils supposent assez naturellement que u est déterminé par c au moyen d'une équation d'Euler : $\partial_t u + (u \cdot \nabla_x)u = \nabla c$. La concentration c reste déterminée par la densité de cellules n comme dans le modèle de Keller-Segel. Ce qui conduit au système :

$$\begin{aligned} &\partial_t n + \operatorname{div}(nu) = 0, \\ &\partial_t u + (u \cdot \nabla_x)u = \nabla c, \\ &\tau c - \Delta c = n. \end{aligned} \tag{48}$$

Et, numériquement, ils obtiennent des formes qui ressemblent à celles qu'observent les biologistes.

Un autre point de vue a été proposé par Daphné Manoussaki [47]. Son idée est que ce qui crée des formes, c'est un phénomène d'élasticité : les cellules, en bougeant, tirent sur le matrigel, un peu comme si on tirait sur une nappe, et cela crée des lignes de tension, qui sont des lignes privilégiées pour le mouvement des cellules. Ses calculs numériques montrent aussi des formes qui ressemblent à celles qu'observent les biologistes.

Il se peut que les deux approches soient correctes. Les formes que l'École de Turin, d'une part, et Manoussaki, d'autre part, cherchaient à reproduire ne sont pas obtenues expérimentalement avec le même type de cellules : les cellules

considérées par Manoussaki sont beaucoup plus grandes (cellules aortiques bovines), et donc on peut avoir un effet d'élasticité beaucoup plus important dans le cas de cellules cappilaires humaines.

Remerciements. L'auteur remercie Maryse Desnous pour plusieurs dessins reproduits dans ce texte, ainsi que Americo Marrocco (INRIA, projet BANG) pour la reproduction ici de simulations numérique du blow-up dans la chimiotactie et M. Mirshahi (INSERM E 355, faculté de Médecine Université P. et M. Curie) pour des données expérimentales sur le mouvement cellulaire. Ce texte a été rédigé, et complété pour les sections 2 et 3, par Éric Charpentier à l'occasion d'une "Leçon de Mathématiques" qu'il organise pour l'École Doctorale de Bordeaux. Je le remercie également très chaleureusement.

Références

[1] T. Alarcon, H.M. Byrne et P.K. Maini, *A design principle for vascular beds : effects of blood rheology and transmural pressure*. J. of Math. Biology. To appear. Preprint (2002).

[2] G. Barles, L. C. Evans, P. E. Souganidis, *Wavefront propagation for reaction diffusion systems of PDE*. Duke Math. J. **61** (1990) p. 835–858.

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

[4] N. Bellomo, L. Preziosi, *Modeling and mathematical problems related to tumors immune system interactions*. Math. Comp. Modelling, 31, (2000) p. 413–452.

[5] H. Berestycki, F. Hamel, *Front propagation in periodic excitable media*. Comm. Pure Appl. Math. 55 (2002), no. 8, p. 949–1032.

[6] Daniel Bernouilli, *Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévention*. Mem. Math. Phys. Acad. Roy., 33 (1760) p. 303–314.

[7] P. Biler, *Local and global solvability of some parabolic systems modelling chemotaxis*, Adv. Math. Sci. Appl. **8**, n^o 2 (1998), p. 715-743.

[8] M. P. Brenner, P. Constantin, L. P. Kadanoff, A. Schenkel, S. C. Venkataramani, *Diffusion, attraction and collapse*, Nonlinearity, 12(4) (1999), p. 1071–1098.

[9] M. P. Brenner, L. S. Levitov, E. O. Budrene, Physical mechanisms for chemotactic pattern formation by bacteria. Biophysical Journal 74 (1998), p. 1677–1693.

[10] R. Bürger, *The mathematical theory of selection, recombination and mutation*. Wiley (2000).

[11] À. Calcina, S. Cuadrado, *Small mutation rate and evolutionarily stable strategies in infinite dimensional adaptive dynamics*. Preprint 2002.

[12] F. Chalub, P. Markowich, B. Perthame, C. Schmeiser, *Kinetic Models for Chemotaxis and their Drift-Diffusion Limits*. To appear in Monat. Math. http ://www.dma.ens.fr/edition/publis/2002/titre02.html

[13] M. A. J. Chaplain. *Avascular growth, angiogenesis and vascular growth in solid tumors : the mathematical modelling of the stages of tumor development*. Math. Comput. Modelling, **23**, (1996), p. 47–87.

[14] G. Chiorino, J. A. J. Metz, D. Tomasoni, P. Ubezio, *Desynchronization rate in cell populations : mathematical modeling and experimental data*, J. Theor. Biol. (2001) 208, p. 185-199.

[15] J. Clairambault, B. Laroche, S. Mischler, B. Perthame, *A mathematical model of cell-cycle*, Proceedings du CANUM 2003, La Grande Motte. À paraître, et Rapport de Recherche INRIA (2003).

[16] M. Cohen-Lara, Lectures given at École Normale Supérieure, Mai 2002.

[17] L. Corrias, B. Perthame et H. Zaag, A chemotaxis model motivated by angiogenesis. C. R. Acad. Sc. Paris. Série I. vol 336/2, p. 141–146 (2003).

[18] L. Corrias, B. Perthame et H. Zaag, article en préparation.

[19] E. De Angelis, P.-E. Jabin, *Analysis of a mean field modelling of tumor and immune system competition*. Math. Models Meth. Appl. Sci., Vol. 13, (2003), No. 2 p. 187–206.

[20] U. Dieckmann, R. Law, *The dynamical theory of coevolution : a derivation from stochastic ecological processes*. J. Math. Biol. 34 (1996), no. 5-6, p. 579–612.

[21] O. Diekmann, J.P. Heesterbeek, *Mathematical Epidemiology of infectious Diseases*, Wiley, New-York (2000).

[22] O. Diekmann, Lectures given at École Normale Supérieure, Jan. 2003.

[23] O. Diekmann, P.-E. Jabin, S. Mischler, B. Perthame, travail en préparation.

[24] S. E. Esipov, J. A. Shapiro, *Kinetic model of* Proteus mirabilis *swarm colony development*, J. Math. Biology **36**, n^o 3 (1998), p. 249-268.

[25] R. Ferrière, *Adaptative responses to environmental threats : evolutionary suicide, insurance, and rescue.* Options, Spring 2000, IIASA ; Laxenburg, Austria, 12–16 (2000).

[26] A. Fick, *Die Medizinische Physik*, Vieweg-Verlag, Braunschweig, 1856.

[27] R. A. Fisher, *The genetical theory of natural selection*, Clarendon Press, 1930. Deuxième éd. : Dover, 1958. Troisième édition, présentée et annotée par Henry Bennett : Oxford Univ. Press, 1999.

[28] R. A. Fisher, *The wave of advance of advantageous genes*, Ann. of Eugen. (London) **7** (1937), p. 355-369.

[29] W. E. Fitzgibbon, M. Langlais, J. J. Morgan, *A mathematical model of the spread of feline leukemia virus (FeLV) through a highly heterogeneous spatial domain*. SIAM J. Math. Anal. 33 (2001), no. 3, p. 570–588

[30] H. von Fœrster, *Some remarks on changing populations* : in *The Kinetics of Cell Proliferation*, F. Stohlman (ed.), Grune and Stratton (New York), 1959 : p. 382-407.

[31] M. Fontelos, A. Friedman, B. Hu, *Mathematical analysis of a model for the initiation of angiogenesis*. SIAM J. Math. Anal. 33(6), (2002), p. 1330–1355.

[32] H. Frid, P.-E. Jabin, B. Perthame, *Global Stability of Steady Solutions in Virus Dynamics*. ESAIM :M2AN, , Vol. 37, No.4, (2003).

[33] A. Goldbeter, *B*iochemical oscillators and cellular rythms, Cambridge University Press (1997).

[34] M. A. Herrero, J. J. L. Velázquez, *Singularity patterns in a chemotaxis model*, Math. Ann. **306**, n^o 3 (1996), p. 583-623.

[35] M. A. Herrero, J. J. L. Velázquez, *A blow-up mechanism for a chemotaxis model*, Ann. Sc. Norm. Super. Pisa, Cl. Sci., IV. Ser. 24, n^o 4 (1997), p. 633-683.

[36] F. Hoppensteadt, *Mathematical theories of populations : Demographics, genetics, and epidemics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1975.

[37] W. Jäger, S. Luckhaus, *On explosions of solutions to a system of partial differential equations modelling chemotaxis*, Trans. Amer. Math. Soc. **329**, n^o 2 (1992), p. 819-824.

[38] E. F. Keller : *Assessing the Keller-Segel model : How has it fared ?*, *in* Biological growth and spread, mathematical theories and applications, Proc. Conf. held at Heidelberg in 1979 : Lect. Notes Biomath., vol. 38 (1980), p. 379-387.

[39] B. L. Keyfitz, N. Keyfitz, *The McKendrick partial differential equation and its uses in epidemiology and population study*, Math. Comput. Modelling **26**, n^o 6 (1997), p. 1-9.

[40] A. N. Kolmogorov, I. G. Petrovsky, N. S. Piskunov : *Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique*, Bull. Université d'État Moscou (*Bjul. Moskowskogo Gos. Univ.*), série internationale, section A **1**, fasc. 6 (1937), p. 1-25.

[41] E. F. Keller, L. A. Segel : *Cells migrate in a self-imposed field of chemoattractant*, J. Theor. Biol. **26** (1970), p. 399-415.

[42] E. F. Keller, L. A. Segel : *A model for chemotaxis*, J. Theor. Biol. **30**(2) (1971), p. 225-234.

[43] F. Lévi, *Cancer chronotherapeutics*, Special issue of Chronobiology international, Vol. 19, No. 1, 2002.

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

[45] A. J. Lotka, *Elements of physical biology*, Williams and Wilkins, Baltimore, 1925. (Réédité sous le titre *Elements of mathematical biology* en 1956, Dover, New York.)

[46] T. R. Malthus : *An Essay on the Principle of Population, as it Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers*, J. Johnson, Londres, 1798. Traduction française par Eric Vilquin : *Essai sur le principe de population en tant qu'il influe sur le progrès de la société, avec des 17s sur les théories de M. Godwin, de M. de Condorcet et d'autres auteurs*, I.N.É.D., Paris, 1980.

[47] D. Manoussaki, *Modeling and simulation of the formation of vascular networks*, ESAIM, Proc. **12** (2002), p. 108-114, et ESAIM :M2AN, Vol. 37, No.4, (2003).

[48] A., Marrocco, *2D simulation of chemotactic bacteria agreggation.* ESAIM :M2AN, Vol. 37, No.4, (2003).

[49] J. A. J. Metz, S. A. H. Geritz, G. Meszéna, F. J. A. Jacobs, J. S. van Heerwaarden, *Adaptive dynamics, a geometrical study of the consequences of nearly faithful reproduction. Stochastic and spatial structures of dynamical systems* (Amsterdam, 1995), 183–231, Konink. Nederl. Akad. Wetensch. Verh. Afd. Natuurk. Eerste Reeks, 45, North-Holland, Amsterdam, 1996.

[50] St. Mischler, B. Perthame et L. Ryzhik, *Stability in a Nonlinear Population Maturation Model.* M3AS Vol. 12, No. 12 (2002), p. 1751–1772.

[51] A. G. McKendrick, *Applications of mathematics to medical problems*, Proc. Edinburgh Math. Soc. **44** (1926), p. 98-130.

[52] James D. Murray, *Mathematical biology.* Springer (1993).

[53] T. Nagai, *Blow-up of radially symmetric solutions to a chemotaxis system*, Adv. Math. Sci. Appl. **5**, n^o 2 (1995), p. 581-601.

[54] T. Nagai, *Global existence of solutions to a parabolic system for chemotaxis in two space dimensions*, Nonlinear Anal. Theory Methods Appl. **30**, n^o 8 (1997), p. 5381-5388.

[55] H. G. Othmer, S. R. Dunbar, W. Alt, *Models of dispersal in biological systems*, J. Math. Biol. **26**, n^o 3 (1988), p. 263-298.

[56] B. Perthame, P.E. Souganidis, *Front propagation for a jump process model arising in spatial ecology*, preprint (2003).

[57] L. A. Segel (editor) : *Biological kinetics*, Cambridge Studies in Mathematical Biology, n^o 12, Cambridge University Press, 1991.

[58] H. Schwetlick, A. Stevens, personal communication (2003) and paper in preparation.

[59] G. Serini, D. Ambrosi, E. Giraudo, A. gamba, L. Preziosi, F. Bussolino : *Modeling the early stages of vascular network assembly*, EMBO 2003.

[60] A. M. Turing : *The Chemical Basis of Morphogenesis*, Phil. Trans. Roy. Soc. London **B 237** (1952), p. 37-72.

[61] P. Verhulst, *Notice sur la loi que la population suit dans son accroissement*, Correspondence Mathématique et Physique **10** (1838), p. 113-121.

[62] V. Volterra, *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi*, Mem. Reale Accad. Naz. dei Lincei. Ser. VI, vol. **2**, 1926.

[63] G. F. Webb, *Theory of nonlinear age-dependent population dynamics*, Monographs and Textbooks in Pure and Applied Mathematics, vol. 89 : M. Dekker (New York), 1985.

Bol. Soc. Esp. Mat. Apl. n°28(2004), 99–128

Recent Progress on Exact Controllability Theory of the Wave and Plate Equations*

Xu Zhang

Departamento de Matemáticas
Universidad Autónoma de Madrid, Spain
and
School of Mathematics, Sichuan University, Chengdu 610064, China.

xu.zhang@uam.es

Abstract

In this paper, we survey some recent results on (global) exact boundary and/or internal controllability problem for the wave and plate equations, for both linear and semi-linear case. For the semi-linear case, the nonlinearity may be globally Lipschitz continuous, or more generally, satisfy some super-linear growth condition at infinity. Via the duality argument, the problem is reduced to the obtention of suitable uniform observability inequalities for the dual linear and/or linearized system, with respect to the coefficients of its lower order terms. The later is solved by means of a delicate point-wise estimate for the related principal operator and a global Carleman-type estimate.

Key words: *Exact controllability, wave equation, plate equation, duality argument, observability inequality, global Carleman estimate.*

AMS subject classifications: *Primary 93B05; Secondary 93B07, 35B37, 70Q05.*

*Supported by the Foundation for the Author of National Excellent Doctoral Dissertation of P. R. China (Project No: 200119), the Grant BFM2002-03345 of the Spanish MCYT, National Natural Science Foundation of China, and The Project-sponsored by SRF for ROCS, SEM.

Fecha de recepción: 9 de septiembre de 2003

1 Introduction

Let $T > 0$ be a given time duration. Let X and U be two Hilbert spaces, $\mathcal{A}$ be the generator of a C_0-semigroup $e^{\mathcal{A}t}$ on X, $f(\cdot,\cdot) : (0,T) \times X \to X$ be a given function. We begin with the following abstract controlled evolution system

$$\begin{cases} \dfrac{dx}{dt} &= \mathcal{A}x + f(t,x) + \mathcal{B}u, \quad t \in (0,T], \\ x(0) &= x_0. \end{cases} \tag{1}$$

In (1), $x = x(t) \in X$ is the state variable, $u = u(t) \in U$ is the control variable, and $\mathcal{B}$ maps the control space U into the state space X. Many control problems for relevant linear and/or semi-linear partial differential equations (PDEs) enter in this context. For instance, the linear and/or semi-linear heat equation, the wave equation, the plate equation, the Schrödinger equation, etc.

Definition 1 *System (1) is said to be exactly controllable in X at time T if for any $x_0, x_1 \in X$, there is a control u belongs to, say $L^2(0,T;U)$, such that the solution of system (1) with this control satisfies $x(T) = x_1$.*

There is a by now well established literature on exact controllability problems. The exact controllability theory for finite dimensional linear systems was introduced by Kalman at the beginning of the 1960s. Thereafter, many authors were devoted to develop it for more general systems including infinite dimensional ones, and its nonlinear and/or stochastic counterparts.

In this paper, we will concentrate on systems governed by linear and semi-linear PDEs. It is well-known that the *exact* controllability property of PDEs is possible only for the time reversible systems. Time reversible systems include the wave, plate and Schrödinger equations, as well as the Maxwell and Lamé systems and so on. In the sequel, we will focus on the linear and/or semi-linear wave and plate equations. We refer to [6], [10] and [21] for recent progresses on the controllability theory of the time irreversible system, especially for the semi-linear heat equations.

Early studies on exact controllability of PDEs can be found in [4], [5], [7], [8], [9], etc. We refer to Russell's survey paper [24] for the available results before 1978. In the recent two decades, stimulated by Lions's book ([20]), great progress have been made there. We mention only an incomplete list of related works, [1], [13], [30], [35] and the rich references cited therein.

The mail tool to solve the exact controllability problem of PDEs is the duality argument, which reduces the problem to the obtention of suitable observability inequalities for the dual linear and/or linearized systems. For the case of semi-linear systems, fixed point techniques are also employed.

In the literature, various methods are developed to derive the observability inequalities of time reversible PDEs, which include the spectral analysis method [24], the classical multiplier method [20], the micro-local analysis method [1], the global Carleman estimate method [16], [25], etc.

In this paper, we will describe mainly the author and his collaborators' works on exact controllability of the wave and plate equations by means of global Carleman estimate.

2 The linear wave equations

In the sequel, we fix a bounded domain $\Omega \subset \mathbb{R}^n$ ($n \in \mathbb{N}$) with a smooth boundary Γ, and set $Q = (0,T) \times \Omega$, $\Sigma = (0,T) \times \Gamma$.

We need some notations. Fix $x_0 \in \mathbb{R}^n$, put

$$\Gamma_0 = \Big\{ x \in \Gamma \;\Big|\; (x - x_0) \cdot \nu(x) > 0 \Big\}, \qquad \Sigma_0 = (0,T) \times \Gamma_0, \tag{2}$$

where $\nu(x)$ denotes the unit outward normal vector of Ω at $x \in \Gamma$. For any $S \in \mathbb{R}^n$ and $\epsilon > 0$, denote the characteristic function of the set S by χ_S, and put

$$\mathcal{O}_\epsilon(S) = \Big\{ y \in \mathbb{R}^n \;\Big|\; |y - x| < \epsilon \text{ for some } x \in S \Big\}.$$

In what follows, we use the notation

$$f_i = f_i(x) \triangleq \frac{\partial f(x)}{\partial x_i} (i = 1, 2, \ldots, n), \quad \sum_i \triangleq \sum_{i=1}^n$$

(on the other hand, x_i is always the ith coordinate of the point x).

This section is addressed to the exact (boundary) controllability of the following linear wave equation:

$$\begin{cases} y_{tt} - \Delta y = p_1(t,x) y + p_2(t,x) y_t + \langle\, p_3(t,x), \nabla y \,\rangle & \text{in } Q, \\ y = \chi_{\Gamma_0}(x) u(t,x) & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } \Omega. \end{cases} \tag{3}$$

In (3), the "state space" and "control space" are chosen respectively to be $L^2(\Omega) \times H^{-1}(\Omega)$ and $L^2(\Gamma_0)$.

We have the following result.

Theorem 1 ([28]) *Let $T > 2 \max\limits_{x \in \Omega} |x - x_0|$, $p_1 \in L^{n+1}(Q)$, $p_2 \in W^{1,\infty}(Q)$, and $p_3 \in W^{1,\infty}(Q; \mathbb{R}^n)$. Then for any given (y_0, y_1), $(z_0, z_1) \in L^2(\Omega) \times H^{-1}(\Omega)$, there is a control $u \in L^2(\Sigma_0)$ such that the weak solution y of (3) satisfies*

$$y(T) = z_0, \quad y_t(T) = z_1 \qquad \text{in } \Omega. \tag{4}$$

Furthermore, concerning the control u, we have the following estimate:

$$|u|_{L^2(\Sigma_0)} \le \mathcal{C}(\ell)(|y_0|_{L^2(\Omega)} + |y_1|_{H^{-1}(\Omega)} + |z_0|_{L^2(\Omega)} + |z_1|_{H^{-1}(\Omega)}), \tag{5}$$

where $\mathcal{C}(\ell)$ is given by

$$\mathcal{C}(\ell) = C \exp(C\ell^2) \tag{6}$$

for some constant $C = C(T, \Omega) > 0$, with

$$\ell \triangleq |p_1|_{L^{n+1}(Q)} + |p_2|_{W^{1,\infty}(Q)} + |p_3|_{W^{1,\infty}(Q;\mathbb{R}^n)}.$$

In order to prove Theorem 1, one needs to establish the (boundary) observability inequality for the following wave equations with lower order terms:

$$\begin{cases} w_{tt} - \Delta w = q_1(t,x)w + q_2(t,x)w_t + \langle\, q_3(t,x), \nabla w \,\rangle & \text{in } Q, \\ w = 0 & \text{on } \Sigma, \\ w(0) = w_0, \quad w_t(0) = w_1 & \text{in } \Omega. \end{cases} \tag{7}$$

In (7), $q_i(\cdot)$ $(i = 1, 2, 3)$ are given functions allowed to be *time-variant and nonsmooth*.

More precisely, we have the following *a priori* estimate for solutions of (7):

Theorem 2 ([28]) *Let $T > 2\max\limits_{x\in\Omega}|x-x_0|$, $q_1 \in L^{n+1}(Q)$, $q_2 \in L^\infty(Q)$, and $q_3 \in L^\infty(Q;\mathbb{R}^n)$. Then for any weak solution $w \in C([0,T];H_0^1(\Omega))\cap C^1([0,T];L^2(\Omega))$ of (7), it holds that*

$$|w_0|^2_{H_0^1(\Omega)} + |w_1|^2_{L^2(\Omega)} \le \mathcal{C}(r)\|\frac{\partial w}{\partial \nu}\|^2_{L^2(\Sigma_0)} \qquad \forall\ (w_0, w_1) \in H_0^1(\Omega)\times L^2(\Omega), \tag{8}$$

where $\mathcal{C}(r)$ is a constant given as in (6) with ℓ replaced by

$$r \triangleq |q_1|_{L^{n+1}(Q)} + |q_2|_{L^\infty(Q)} + |q_3|_{L^\infty(Q;\mathbb{R}^n)}.$$

Remark 1 *The originality of our method consists in the fact that we can give explicit estimates as that in (6) for the constant $\mathcal{C}(r)$ in (8) via the norm of the coefficients of (7). For the case $n = 1$, Zuazua [34] obtained a similar estimate, and such an estimate played a crucial role in the proof of his main result on exact controllability for the subcritical semilinear wave equations in one space dimension. However, we would like to point out that estimate (6) is not sharp. In fact, one may expect an estimate of the order of $e^{Cr^{1/2}}$, as indicated by [34] for the case $n = 1$. How to derive a sharp estimate on (6) is an open problem.*

Remark 2 *It would be interesting to analyze whether the same result in Theorem 2 holds true for more general observer Σ_0, say, for T and Γ_0 which satisfy the Geometric Control Condition introduced in [1]. But this is still an open problem.*

Remark 3 *Our method can be adopted to the same wave equation (7) but with purely homogenous Neumann boundary condition. We refer to [17] for the details. Note however that for this case, the techniques are much more involved.*

Proof of Theorem 1 via Theorem 2. We use the duality argument. First, we solve

$$\begin{cases} v_{tt} - \Delta v = p_1 v + p_2 v_t + \langle\, p_3, \nabla v \,\rangle & \text{in } Q, \\ v = 0 & \text{on } \Sigma, \\ v(T) = z_0, \quad v_t(T) = z_1 & \text{in } \Omega. \end{cases} \tag{9}$$

Next, for any $(\varphi_0, \varphi_1) \in X \stackrel{\triangle}{=} H_0^1(\Omega) \times L^2(\Omega)$, we solve

$$\begin{cases} \varphi_{tt} - \Delta\varphi = [p_1 - (p_2)_t - \nabla \cdot p_3]\varphi - p_2\varphi_t - \langle\, p_3, \nabla\varphi \,\rangle & \text{in } Q, \\ \varphi = 0 & \text{on } \Sigma, \\ \varphi(0) = \varphi_0, \quad \varphi_t(0) = \varphi_1 & \text{in } \Omega, \end{cases} \tag{10}$$

and

$$\begin{cases} \eta_{tt} - \Delta\eta = p_1\eta + p_2\eta_t + \langle\, p_3, \nabla\eta \,\rangle & \text{in } Q, \\ \eta = \big(\partial\varphi/\partial\nu\big)\chi_{\Sigma_0}(t,x) & \text{on } \Sigma, \\ \eta(T) = 0, \quad \eta_t(T) = 0 & \text{in } \Omega. \end{cases} \tag{11}$$

Then, we define a linear and continuous operator $\Lambda : X \to X'(\equiv H^{-1}(\Omega) \times L^2(\Omega))$ by

$$\Lambda(\varphi_0, \varphi_1) = \big(p_2(0)\eta(0) - \eta_t(0), \eta(0)\big), \tag{12}$$

where $\eta \in C([0,T]; L^2(\Omega)) \cap C^1([0,T]; H^{-1}(\Omega))$ is the weak solution of (11). It suffices to prove the existence of some $(\varphi_0, \varphi_1) \in X$ such that

$$\Lambda(\varphi_0, \varphi_1) = \Big(p_2(0)(y_0 - v(0)) - y_1 + v_t(0),\, y_0 - v(0)\Big), \tag{13}$$

where $v \in C([0,T]; L^2(\Omega)) \cap C^1([0,T]; H^{-1}(\Omega))$ is the weak solution of (9). In order to solve (13), we observe that (by (10)–(11))

$$\langle\, \Lambda(\varphi_0, \varphi_1), (\varphi_0, \varphi_1) \,\rangle_{X',X} = \int_{\Sigma_0} \|\frac{\partial\varphi}{\partial\nu}\|^2 d\Sigma_0. \tag{14}$$

However, by Theorem 2 and (14), we have

$$|(\varphi_0, \varphi_1)|_X^2 \le \mathcal{C}(\ell)\langle\, \Lambda(\varphi_0, \varphi_1), (\varphi_0, \varphi_1) \,\rangle_{X',X}, \qquad \forall\, (\varphi_0, \varphi_1) \in X. \tag{15}$$

Therefore $\Lambda : X \to X'$ is an isomorphism. Thus (13) admits a unique solution $(\varphi_0, \varphi_1) \in X$ and $u = \partial\varphi/\partial\nu$ is the desired control such that the weak solution of (3) satisfies (4).

Finally, (5) follows easily from (11)–(15) and the usual energy estimate to (9). □

In order to prove Theorem 2, we need the following key point-wise estimate, which is a special case of [18, Lemma 1, p. 124] and [22, Lemma 5.1].

Lemma 3 *Let $\lambda > 0$ and $\alpha_1, \alpha_2 \in \mathbb{R}$ be constant. Let $x_0 \in \mathbb{R}^n$, $T > 0$, and*

$$\begin{cases} \psi(t,s,x) = \dfrac{1}{2}\Big[|x - x_0|^2 - \alpha_1(t - T/2)^2 - \alpha_2(s - T/2)^2\Big], \\ \ell = \lambda\psi, \quad \beta \stackrel{\triangle}{=} \min(n + \alpha_1 - 1, n + \alpha_2 - 1), \quad \Psi = \beta\lambda. \end{cases} \tag{16}$$

Let $z = z(t,s,x) \in C^2(\mathbb{R}^{2+n})$. Denote

$$v \stackrel{\triangle}{=} \theta z \quad \textit{with} \quad \theta = e^{\ell}. \tag{17}$$

Then

$$\begin{aligned}
&\theta^2|z_{tt}+z_{ss}-\Delta z|^2\\
&\geq\Big[-2\ell_t\Big(v_t^2-v_s^2+\sum_j v_j^2\Big)-4\ell_s v_t v_s+4\sum_j(\ell_j v_t v_j)\\
&\quad+2\Psi v_t v-2\ell_t(A+\Psi)v^2\Big]_t\\
&\quad+\Big[-2\ell_s\Big(v_s^2-v_t^2+\sum_j v_j^2\Big)-4\ell_t v_t v_s+4\sum_j(\ell_j v_s v_j)\\
&\quad+2\Psi v_s v-2\ell_s(A+\Psi)v^2\Big]_s\\
&\quad-2\sum_j\Big[2\sum_i(\ell_i v_i v_j)-\ell_j\sum_i v_i^2-2\ell_t v_t v_j-2\ell_s v_s v_j\\
&\quad+\Psi v_j v+\ell_j(v_t^2+v_s^2)-(A+\Psi)\ell_j v^2\Big]_j+2(n-\alpha_1+\alpha_2-\beta)\lambda v_t^2\\
&\quad+2(n+\alpha_1-\alpha_2-\beta)\lambda v_s^2+2(2-n-\alpha_1-\alpha_2+\beta)\lambda\sum_j v_j^2+Bv^2,
\end{aligned}\tag{18}$$

where

$$A=\lambda^2\Big[\alpha_1^2(t-T/2)^2+\alpha_2^2(s-T/2)^2-|x-x_0|^2\Big]+(n+\alpha_1+\alpha_2-\beta)\lambda,\tag{19}$$

and for large λ*,*

$$\begin{aligned}
B=\;&2\lambda^3\Big[(2+n-\beta+\alpha_1+\alpha_2)|x-x_0|^2\\
&+\alpha_1^2(\beta-n-3\alpha_1-\alpha_2)(t-T/2)^2\\
&+\alpha_2^2(\beta-n-3\alpha_2-\alpha_1)(s-T/2)^2\Big]+O(\lambda^2).
\end{aligned}\tag{20}$$

Proof of Theorem 2. The proof is split into several steps.

Step 1. Following [11], the main idea of our proof is to use the pointwise estimate (18) in Lemma 3. For this purpose, we need to choose suitable parameters x_0, α_1 and α_2 in function ψ.

We choose x_0 as that in (2) and for simplicity, we assume that $x_0\in\mathbb{R}^n\setminus\updownarrow\Omega$. Put

$$R_0\triangleq\min_{x\in\Omega}|x-x_0|,\quad R_1\triangleq\max_{x\in\Omega}|x-x_0|.\tag{21}$$

Then $R_0>0$ and $T>2R_1$. Thus we can choose a constant $\alpha\in(0,1)$ (close to 1) such that

$$R_1^2<\alpha T^2/4.\tag{22}$$

Then, we set

$$\psi=\psi(t,x)\triangleq[|x-x_0|^2-\alpha(t-T/2)^2]/2.\tag{23}$$

Step 2. We need the following notations. First, denote

$$\Lambda_j \triangleq \Big\{(t,x) \in Q \;\Big|\; 2\psi(t,x) > R_0^2/(j+2)\Big\}, \tag{24}$$

where $j = 0, 1, 2$. Next, denote

$$T_i \triangleq T/2 - \varepsilon_i T, \quad T_i' \triangleq T/2 + \varepsilon_i T, \quad Q_i \triangleq (T_i, T_i') \times \Omega, \tag{25}$$

where $i = 0, 1$; $0 < \varepsilon_0 < \varepsilon_1 < 1/2$ will be determined as follows.

First of all, by (21)–(23), one sees that

$$\psi(0,x) = \psi(T,x) = (R_1^2 - \alpha T^2/4)/2 < 0 \quad \forall\, x \in \Omega. \tag{26}$$

Thus, one can find an $\varepsilon_1 \in (0, 1/2)$ such that

$$\Lambda_2 \subset Q_1 \tag{27}$$

and for any $(t,x) \in ((0,T_1) \cup (T_1', T)) \times \Omega$ it holds that

$$\psi(t,x) < 0. \tag{28}$$

Next, noting that since $\{T/2\} \times \Omega \subset \Lambda_0$, one can find a small $\varepsilon_0 \in (0, \varepsilon_1)$ such that

$$Q_0 \subset \Lambda_0. \tag{29}$$

Now, we note that (recall (20) for B)

$$B = B\chi_{\Lambda_2}(t,x) + B\chi_{Q\setminus\Lambda_2}(t,x). \tag{30}$$

By (23), we see that (recall (16) for α_1, α_2, and β)

$$\alpha_1 = \alpha, \quad \alpha_2 = 0, \quad \beta = n - 1, \tag{31}$$

where α is given in (22). Thus, by (20), (24), and (31), one sees easily that there exists a constant $\lambda_1 > 1$ such that for any $\lambda > \lambda_1$, it holds that

$$B\chi_{\Lambda_2}(t,x) \geq c_0 \lambda^3 \chi_{\Lambda_2}(t,x) \tag{32}$$

and

$$\big|B\chi_{Q\setminus\Lambda_2}(t,x)\big| \leq C\lambda^3 \tag{33}$$

for some constants $c_0 > 0$ and $C > 0$, which depend only on T and Ω.

Step 3. We now use Lemma 3. For any given $\tau \in (0, T_1)$ and $\tau' \in (T_1', T)$ (recall (25) for T_1 and T_1'), denote

$$Q_\tau^{\tau'} \triangleq (\tau, \tau') \times \Omega. \tag{34}$$

Let us observe (18), where $z = z(t,s,x)$ is replaced by $w = w(t,x)$, and ψ is given by (23). Integrating (18) on $Q_\tau^{\tau'}$, using integration by parts, and taking (1.1) into account, we arrive at (noting that by (17), $v = \theta w$)

$$\begin{aligned}
&2(1-\alpha)\lambda \int_{Q_\tau^{\tau'}} \left(v_t^2 + \sum_i v_i^2 \right) dxdt + \int_{Q_\tau^{\tau'}} Bv^2 dxdt \\
&\le \int_Q \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w \,\rangle|^2 dxdt + \int_{\Sigma_0} \|\frac{\partial v}{\partial \nu}\|^2 d\Sigma_0 \\
&\quad + C\lambda^3 \left[\int_\Omega \Big(|v(\tau,x)|^2 + |v_t(\tau,x)|^2 + \sum_i |v_i(\tau,x)|^2 \right. \\
&\qquad \left. + |v(\tau',x)|^2 + |v_t(\tau',x)|^2 + \sum_i |v_i(\tau',x)|^2 \Big) dx \right], \qquad \forall\, \lambda > 1.
\end{aligned} \tag{35}$$

However, by $v = \theta w$ and $\theta = e^\ell$, by (16), (23), and (28), we get

$$\begin{aligned}
&\int_\Omega \Big(|v(\tau,x)|^2 + |v_t(\tau,x)|^2 + \sum_i |v_i(\tau,x)|^2 \\
&\qquad + |v(\tau',x)|^2 + |v_t(\tau',x)|^2 + \sum_i |v_i(\tau',x)|^2 \Big) dx \\
&\le C\lambda^2 \Big[\int_\Omega \Big(|w(\tau,x)|^2 + |w_t(\tau,x)|^2 + \sum_i |w_i(\tau,x)|^2 \\
&\qquad + |w(\tau',x)|^2 + |w_t(\tau',x)|^2 + \sum_i |w_i(\tau',x)|^2 \Big) dx \Big].
\end{aligned} \tag{36}$$

Further, by (23)–(24), (16)–(17), (30), and (32)–(34), we get

$$\begin{aligned}
&\int_{Q_\tau^{\tau'}} Bv^2 dxdt = \int_{Q_\tau^{\tau'} \cap \Lambda_2} Bv^2 dxdt + \int_{Q_\tau^{\tau'} \setminus \Lambda_2} Bv^2 dxdt \\
&\ge c_0 \lambda^3 \int_{Q_\tau^{\tau'} \cap \Lambda_2} v^2 dxdt - C\lambda^3 e^{R_0^2 \lambda/4} \int_Q w^2 dxdt, \quad \forall\, \lambda > \lambda_1.
\end{aligned} \tag{37}$$

Note that by (24), (27), and (34), we have $Q_\tau^{\tau'} \supset \Lambda_1$. Thus, by (37), for any $\lambda > \lambda_1$, we have

$$\begin{aligned}
&2(1-\alpha)\lambda \int_{Q_\tau^{\tau'}} \Big(v_t^2 + \sum_i v_i^2 \Big) dxdt + \int_{Q_\tau^{\tau'}} Bv^2 dxdt \\
&\ge c_1 \Big[\lambda \int_{\Lambda_1} \Big(v_t^2 + \sum_i v_i^2 \Big) dxdt + \lambda^3 \int_{\Lambda_1} v^2 dxdt \Big] - C\lambda^3 e^{R_0^2 \lambda/4} \int_Q w^2 dxdt,
\end{aligned} \tag{38}$$

where $c_1 > 0$ and $C > 0$ are two constants which depend only on T and Ω.

Now, combining (35)–(36) and (38), we conclude that for any $\lambda > \lambda_1$, it holds that

$$\begin{aligned}
&\int_{\Lambda_1} \left(v_t^2 + \sum_i v_i^2 \right) dxdt + \lambda^2 \int_{\Lambda_1} \theta^2 v^2 dxdt \\
&\leq C\lambda^{-1} \left\{ \int_Q \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w \,\rangle|^2 dxdt + \int_{\Sigma_0} \|\frac{\partial v}{\partial \nu}\|^2 d\Sigma_0 \right. \\
&\quad + \lambda^5 \left[\int_\Omega \left(|w(\tau,x)|^2 + |w_t(\tau,x)|^2 + \sum_i |w_i(\tau,x)|^2 \right. \right. \\
&\quad \left. \left. + |w(\tau',x)|^2 + |w_t(\tau',x)|^2 + \sum_i |w_i(\tau',x)|^2 \right) dx \right] \\
&\quad \left. + \lambda^3 e^{R_0^2 \lambda/4} \int_Q w^2 dxdt \right\}.
\end{aligned} \tag{39}$$

Integrating (39) with respect to τ and τ' from T_2, T_1 and T_1', T_2', respectively, we get

$$\begin{aligned}
&\int_{\Lambda_1} \left(v_t^2 + \sum_i v_i^2 \right) dxdt + \lambda^2 \int_{\Lambda_1} v^2 dxdt \\
&\leq C\lambda^{-1} \left\{ \int_Q \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w \,\rangle|^2 dxdt + \int_{\Sigma_0} \|\frac{\partial v}{\partial \nu}\|^2 d\Sigma_0 \right. \\
&\quad \left. + \lambda^5 \int_Q \left(w^2 + w_t^2 + \sum_i w_i^2 \right) dxdt + \lambda^3 e^{R_0^2 \lambda/4} \int_Q w^2 dxdt \right\}.
\end{aligned} \tag{40}$$

Consequently, by (16)–(17) and (23), recalling that $w = \theta^{-1} v$ with $\theta = e^\ell$, and using (40) and (7), we see that for any $\lambda > \lambda_1$, it holds that

$$\begin{aligned}
&\int_{\Lambda_1} \theta^2 \left(w_t^2 + \sum_i w_i^2 \right) dxdt + \lambda^2 \int_{\Lambda_1} \theta^2 w^2 dxdt \\
&\leq C\lambda^{-1} \left\{ \int_Q \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w \,\rangle|^2 dxdt + \int_{\Sigma_0} \theta^2 \|\frac{\partial w}{\partial \nu}\|^2 d\Sigma_0 \right. \\
&\quad \left. + \lambda^5 \int_Q \left(w^2 + w_t^2 + \sum_i w_i^2 \right) dxdt + \lambda^3 e^{R_0^2 \lambda/4} \int_Q w^2 dxdt \right\}.
\end{aligned} \tag{41}$$

Step 4. We need to estimate

$$\int_Q \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w \,\rangle|^2 dxdt.$$

By the Hölder inequality, the Sobolev embedding theorem, and the Poincaré inequality, we get (recalling $r \triangleq |q_1|_{n+1} + |q_2|_\infty + |q_3|_\infty$)

$$\begin{aligned}
&\int_Q \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w \,\rangle|^2 dxdt \\
&= \left\{ \int_{\Lambda_1} + \int_{Q\setminus\Lambda_1} \right\} \theta^2 |q_1 w + q_2 w_t + \langle\, q_3, \nabla w \,\rangle|^2 dxdt \\
&\le Cr^2 \left[\int_{\Lambda_1} \theta^2 (w_t^2 + |\nabla w|^2) dxdt + (1+\lambda^2) \int_{\Lambda_1} \theta^2 w^2 dxdt \right. \\
&\quad \left. + e^{R_0^2 \lambda/3} \int_Q (w_t^2 + |\nabla w|^2) dxdt \right].
\end{aligned} \tag{42}$$

Thus, combining (41) and (42), we see that for any $\lambda > \lambda_1$, it holds that

$$\begin{aligned}
&\int_{\Lambda_1} \theta^2 (w_t^2 + |\nabla w|^2) dxdt + \lambda^2 \int_{\Lambda_1} \theta^2 w^2 dxdt \\
&\le C_1 \lambda^{-1} \left\{ r^2 \left[\int_{\Lambda_1} \theta^2 (w_t^2 + |\nabla w|^2) dxdt + \lambda^2 \int_{\Lambda_1} \theta^2 w^2 dxdt \right. \right. \\
&\quad \left. + e^{R_0^2 \lambda/3} \int_Q (w_t^2 + |\nabla w|^2) dxdt \right] + \int_{\Sigma_0} \theta^2 \left| \frac{\partial w}{\partial \nu} \right|^2 d\Sigma_0 \\
&\quad \left. + \lambda^5 \int_Q \Big(w^2 + w_t^2 + \sum_i w_i^2 \Big) dxdt + \lambda^3 e^{R_0^2 \lambda/4} \int_Q w^2 dxdt \right\},
\end{aligned} \tag{43}$$

where $C_1 > 0$ is a constant. Now, taking

$$\lambda_2 \triangleq \max(\lambda_1, 2 + C_1 r^2), \tag{44}$$

by (43)–(44), we see that for any $\lambda > \lambda_2$ it holds that

$$\begin{aligned}
&\int_{\Lambda_1} \theta^2 (w_t^2 + |\nabla w|^2) dxdt \\
&\le C\lambda^{-1} \left\{ \int_{\Sigma_0} \theta^2 \| \frac{\partial w}{\partial \nu} \|^2 d\Sigma_0 + r^2 e^{R_0^2 \lambda/3} \int_Q (w_t^2 + |\nabla w|^2) dxdt \right. \\
&\quad \left. + \lambda^5 \int_Q \left(w^2 + w_t^2 + \sum_i w_i^2 \right) dxdt + \lambda^3 e^{R_0^2 \lambda/4} \int_Q w^2 dxdt \right\}.
\end{aligned} \tag{45}$$

Note that by (24) and (29), we have

$$\begin{aligned}
\int_{\Lambda_1} \theta^2 (w_t^2 + |\nabla w|^2) dxdt &\ge \int_{\Lambda_0} \theta^2 (w_t^2 + |\nabla w|^2) dxdt \\
&\ge e^{R_0^2 \lambda/2} \int_{Q_0} (w_t^2 + |\nabla w|^2) dxdt.
\end{aligned} \tag{46}$$

Thus, by (45)–(46), we see that for any $\lambda > \lambda_2$, it holds that

$$\begin{aligned}
&\int_{Q_0} (|w_t|^2 + |\nabla w|^2) dxdt \\
&\leq C\lambda^{-1} \left\{ e^{C\lambda} \int_{\Sigma_0} \|\frac{\partial w}{\partial \nu}\|^2 d\Sigma_0 + r^2 e^{-R_0^2\lambda/6} \int_Q (|w_t|^2 + |\nabla w|^2) dxdt \right. \\
&\quad \left. + \lambda^5 e^{-R_0^2\lambda/2} \int_Q \Big(w^2 + w_t^2 + \sum_i w_i^2 \Big) dxdt + \lambda^3 e^{-R_0^2\lambda/4} \int_Q w^2 dxdt \right\}.
\end{aligned} \tag{47}$$

Step 5. Let us complete the proof of Theorem 2. Denote the energy of system (7) by

$$\mathcal{E}(t) \triangleq |w_t(t,\cdot)|^2_{L^2(\Omega)} + |w(t,\cdot)|^2_{H_0^1(\Omega)}. \tag{48}$$

By (47) and (44), using Poincaré inequality, we conclude that there is a constant $\lambda_3 > 0$, which depends only on T and Ω, such that for all $\lambda > \lambda_2 + \lambda_3$, it holds (recall (25) for T_0 and T_0')

$$\int_{T_0}^{T_0'} \mathcal{E}(t) dt \leq C \left\{ e^{C\lambda} \int_{\Sigma_0} \|\frac{\partial w}{\partial \nu}\|^2 d\Sigma_0 + e^{-R_0^2\lambda/8} \int_0^T \mathcal{E}(t) dt \right\}. \tag{49}$$

On the other hand, applying the classical energy method to system (7), we find

$$\mathcal{E}(t) \leq C\mathcal{E}(s) e^{Cr}, \qquad \forall\, t, s \in [0, T]. \tag{50}$$

Hence, combining (49) and (50), we arrive at

$$\mathcal{E}(0) \leq C_2 \left\{ e^{C_2\lambda} \int_{\Sigma_0} \|\frac{\partial w}{\partial \nu}\|^2 d\Sigma_0 + e^{-R_0^2\lambda/8 + C_2 r} \mathcal{E}(0) \right\}, \quad \forall\, \lambda > \lambda_2 + \lambda_3, \tag{51}$$

where $C_2 = C_2(T, \Omega)$ is a positive constant. However, it is easy to find a constant $\lambda_4 = \lambda_4(R_0, C_2) > 0$ such that

$$C_2 e^{-R_0^2\lambda_4/8 + C_2 r} \leq 1/2. \tag{52}$$

Thus, by (51)–(52), one gets

$$\mathcal{E}(0) \leq C e^{C\lambda} \int_{\Sigma_0} \|\frac{\partial w}{\partial \nu}\|^2 d\Sigma_0, \qquad \forall\, \lambda > \max(\lambda_2 + \lambda_3, \lambda_4), \tag{53}$$

which is exactly the desired inequality (8). On the other hand, the explicit estimate (6) follows from (44) and (52)–(53) immediately. □

3 The semi-linear wave equations

Fix a nonlinear function $f \in C^1(\mathbb{R})$, and an open (proper) subset ω of Ω. Let us consider the following controlled semi-linear wave equation

$$\begin{cases} y_{tt} - \Delta y + f(y) = \chi_\omega(x)u(t,x) & \text{in} \quad (0,T)\times\Omega, \\ y = 0 & \text{on} \quad (0,T)\times\Gamma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in} \quad \Omega. \end{cases} \tag{54}$$

In (7), the "state space" and "control space" are chosen respectively to be $H_0^1(\Omega)\times L^2(\Omega)$ and $L^2(\omega)$. This section is devoted to analyze the exact (internal) controllability of system (54). The exact (boundary) controllability of (54) can be considered similarly. In that case the control u enters on the system through the boundary conditions. However, in order to avoid unnecessary technical difficulties, we will concentrate on considering only the internal controllability problem.

First of all, let us consider the case of the nonlinearity being global Lipshitz continuous. It was proved in [27] that

Theorem 4 *Let $\omega = \Omega \cap \mathcal{O}_\epsilon(\Gamma_0)$ for some $\epsilon > 0$, $T > 2\max\limits_{x\in\Omega\setminus\omega} |x - x_0|$ and $f(\cdot) \in C^1(\mathbb{R})$ with $f'(\cdot) \in L^\infty(\mathbb{R})$. Then system (54) is exactly controllable in $H_0^1(\Omega)\times L^2(\Omega)$ at time T by means of control $u \in L^2((0,T)\times\omega)$.*

Theorem 4 is a generalization of the main result in [33], where the controller ω is assumed to be a neighborhood of the whole boundary Γ.

By means of the fixed point technique, and using again the duality argument, Theorem 4 is a consequence of the following uniform observability estimate for system (7).

Theorem 5 *Let $\omega = \Omega \cap \mathcal{O}_\epsilon(\Gamma_0)$ for some $\epsilon > 0$, $T > 2\max\limits_{x\in\Omega\setminus\omega} |x - x_0|$, $q_1 \in L^\infty(Q)$, $q_2 = 0$, and $q_3 = 0$. Then for any weak solution $w \in C([0,T];L^2(\Omega)) \cap C^1([0,T];H^{-1}(\Omega))$ of (7), it holds that*

$$|w_0|^2_{L^2(\Omega)} + |w_1|^2_{H^{-1}(\Omega)} \le \mathcal{C}(h)\int_0^T\int_\omega w^2 dxdt, \quad \forall\ (w_0,w_1) \in L^2(\Omega)\times H^{-1}(\Omega) \tag{55}$$

for some constant $\mathcal{C}(h) > 0$ with $h \triangleq |q_1|_{L^\infty(Q)}$. Furthermore, the constant $\mathcal{C}(h)$ in (55) may be bounded as

$$\mathcal{C}(h) = C\exp(Ch^2) \tag{56}$$

for some constant $C = C(T,\Omega) > 0$.

Remark 4 *The first conclusion in Theorem 5 can be found in [27]. However, the explicit estimate on the observability constant C(h) in [27] is much weaker than (56). Indeed, the estimate there reads* $\mathcal{C}(h) = C\exp(\exp(\exp(h)))$. *In order to obtain the estimate (56), we need to use some technique developed in [28].*

Proof of Theorem 5. We shall proceed as in the proof of Theorem 2. However, only a sketch of proof will be given and we refer to [27] and [28] for a complete proof.

We need to introduce the following key integral transformation. Put

$$z(t,s,x) \triangleq \int_s^t w(\xi,x)d\xi, \qquad \forall\ (t,s,x) \in (0,T)\times Q, \tag{57}$$

where w is the weak solution of (7). One sees that z satisfies the following ultra-hyperbolic equation (recall $q_2 = 0$ and $q_3 = 0$):

$$\begin{cases} z_{tt} + z_{ss} - \Delta z = \displaystyle\int_s^t q_1(\xi,x) z_t(\xi,s,x)d\xi & \text{in } (0,T)\times Q, \\ z = 0 & \text{on } (0,T)\times\Sigma. \end{cases} \tag{58}$$

Applying Lemma 3 to (58), similar to Steps 1–4 in the proof of Theorem 2, we conclude that there exist $T_0 \in (0,T/2)$, $T_0' \in (T/2,T)$, $T_1 \in (0,T_0)$, $T_1' \in (T_0',T)$ and a positive constant $\lambda_1 = O(h^2)$ such that for any $\lambda > \lambda_1$ it holds

$$\begin{aligned} \int_{T_0}^{T_0'}\int_{T_0}^{T_0'}\int_\Omega (z_t^2 + z_s^2)dxdt \le C\Bigg\{ & e^{C\lambda}\int_{T_1}^{T_1'}\int_{T_1}^{T_1'}\int_{\Gamma_0} \|\frac{\partial z}{\partial\nu}\|^2 dtdsdx \\ & + \lambda^5 e^{-R_0^2\lambda/6}\int_{T_1}^{T_1'}\int_{T_1}^{T_1'}\int_\Omega \left(z^2 + z_t^2 + z_s^2 + \sum_{i=1}^n z_i^2\right)dxdtds\Bigg\}, \end{aligned} \tag{59}$$

where R_0 is defined by (21).

However, applying the classical energy method to equation (58), one may show that

$$\int_{T_1}^{T_1'}\int_{T_1}^{T_1'}\int_{\Gamma_0} \|\frac{\partial z}{\partial\nu}\|^2 dtdsdx \le C(1+h)\int_0^T\int_0^T\int_\omega (z^2 + z_t^2 + z_s^2)dxdtds \tag{60}$$

and

$$\sum_{i=1}^n \int_{T_1}^{T_1'}\int_{T_1}^{T_1'}\int_\Omega z_i^2 dxdtds \le C(1+h)\int_0^T\int_Q (z^2 + z_t^2 + z_s^2)dxdtds. \tag{61}$$

Combining (59), (60) and (61), we end up with

$$\begin{aligned} \int_{T_0}^{T_0'}\int_{T_0}^{T_0'}\int_\Omega (z_t^2 + z_s^2)dxdt \le C\Bigg\{ & e^{C\lambda}\int_0^T\int_0^T\int_\omega (z^2 + z_t^2 + z_s^2)dxdtds \\ & + \lambda^6 e^{-R_0^2\lambda/6}\int_0^T\int_Q (z^2 + z_t^2 + z_s^2)dxdtds\Bigg\}. \end{aligned} \tag{62}$$

We now return to "w". Define the energy of (7) (with $q_2 = 0$ and $q_3 = 0$) by

$$E(t) \triangleq |w_t(t,\cdot)|^2_{H^{-1}(\Omega)} + |w(t,\cdot)|^2_{L^2(\Omega)}. \tag{63}$$

By (59) and (57), for any $\lambda > \lambda_1$, we get

$$\int_{T_0}^{T_0'} \int_\Omega w^2 dxdt \le C \left\{ e^{C\lambda} \int_0^T \int_\omega w^2 dxdt + \lambda^6 e^{-R_0^2\lambda/6} \int_0^T E(t)dt \right\}. \tag{64}$$

Fix $S_0 \in (T_0, T/2)$ and $S_0' \in (T/2, T_0')$. Then it is easy to check that

$$\int_{S_0}^{S_0'} E(t)dt \le C(1+h) \int_{T_0}^{T_0'} \int_\Omega w^2 dxdt. \tag{65}$$

Thus, by (64)–(65), one can find a constant $\lambda_3 = \lambda_3(R_0) > 0$ such that, for all $\lambda > \lambda_2 + \lambda_3$,

$$\int_{S_0}^{S_0'} E(t)dt \le C \left\{ e^{C\lambda} \int_0^T \int_\omega w^2 dxdt + e^{-R_0^2\lambda/8} \int_0^T E(t)dt \right\}. \tag{66}$$

Finally, from (66) and applying the classical energy estimate to system (7) again, one gets

$$E(0) \le C_2 \left\{ e^{C_2\lambda} \int_0^T \int_\omega w^2 dxdt + e^{-R_0^2\lambda/8 + C_2\sqrt{h}} E(0) \right\}, \forall\, \lambda > \lambda_2 + \lambda_3, \tag{67}$$

where $C_2 = C_2(T, \Omega)$ is a positive constant. However, it is easy to find a constant $\lambda_4 = \lambda_4(R_0, C_2)$ such that

$$C_2 e^{-R_0^2\lambda_4/8 + C_2\sqrt{h}} \le 1/2. \tag{68}$$

Thus, by (67)–(68), we see that for any $\lambda > \max(\lambda_2 + \lambda_3, \lambda_4)$ it holds that

$$E(0) \le C e^{C\lambda} \int_0^T \int_\omega w^2 dxdt. \tag{69}$$

Equation (69) is exactly the desired result. Thus, the proof of Theorem 5 is completed. □

Next, we consider the case of the nonlinearity growing "mildly" at infinity, i.e.,

$$\overline{\lim_{|x|\to\infty}} \left| \int_0^x f(s)ds \right| \left[|x| \prod_{k=1}^{\infty} \log_k(e_k + x^2) \right]^{-2} < \infty, \tag{70}$$

where the iterated logarithm function $\log_j$ is defined by the formulas: $\log_0 s = s$ and $\log_{j+1} s = \log(\log_j s)$, $\quad j = 0, 1, 2, \cdots$, the number e_j is defined by the equations $\log_j e_j = 1$.

For the case of one space dimension, Cannarsa, Komornik and Loreti ([3]) proved that

Theorem 6 *Let $n = 1$, $\Omega = (0,1)$ and $\omega \equiv (a,b)$ be a subinterval of $(0,1)$, $T > 2\max(a, 1-b)$, $f \in C^1(\mathbb{R})$ and (70) hold. Then, (54) is exact controllable in $H_0^1(\Omega) \times L^2(\Omega)$ at time T.*

The approach in [3] is genuinely one dimensional, which is quite different from ours. We refer to [3] for details. Theorem 6 is an improvement of the main result in [34], where the nonlinearity $f(\cdot)$ is assumed to satisfy the growth condition: $\lim_{|x|\to\infty} |f(x)||x|^{-1}\log^{-2}|x| = 0$. The growth condition (70) on f is sharp since solutions of (54) may blow up whenever f grows faster than (70) at infinity and f has the bad sign.

For the case of several space dimensions, by assuming the nonlinearity $f(\cdot)$ satisfies the growth condition

$$\lim_{|x|\to\infty} f(x)|x|^{-1}\log^{-1/2}|x| = 0, \tag{71}$$

Li and Zhang ([19]) obtained the following result:

Theorem 7 *Let $\omega = \Omega \cap \mathcal{O}_\epsilon(\Gamma)$ for some $\epsilon > 0$, $T > \operatorname{diam}(\Omega \setminus \omega)$, $f \in C^1(\mathbb{R})$ and (71) hold. Then (54) is exactly controllable in $H_0^1(\Omega) \times L^2(\Omega)$ at time T.*

Remark 5 *It is easy to see that neither the geometric condition on the controller ω nor the growth condition (71) on the nonlinearity $f(\cdot)$ in Theorem 7 is sharp. We refer to [31] for more unsolved problems in this respect.*

In order to prove Theorem 7, we need the following explicit observability estimate for the wave equation with a potential in the *L^p-classes*.

Theorem 8 *Let $\omega \equiv \Omega \cap \mathcal{O}_{\varepsilon_0}(\Gamma)$ for some $\varepsilon_0 > 0$ and $T > T_0 \triangleq \operatorname{diam}(\Omega \setminus \omega)$, $q_1 \in L^{1+n}(Q)$ (or $q_1 \in L^\infty(0,T;L^n(\Omega))$), $q_2 = 0$, and $q_3 = 0$. Then for any weak solution $w \in C([0,T];L^2(\Omega)) \cap C^1([0,T];H^{-1}(\Omega))$ of (7), it holds that*

$$\begin{aligned} |w_0|^2_{L^2(\Omega)} + |w_1|^2_{H^{-1}(\Omega)} \le L(\ell)\int_0^T\int_\omega |w|^2 dxdt, \\ \forall\ (w_0,w_1) \in L^2(\Omega)\times H^{-1}(\Omega) \end{aligned} \tag{72}$$

for some constant $L = L(\ell)$ with $\ell \triangleq |q_1|_{L^{1+n}(Q)}$ (or $\ell \triangleq |q_1|_{L^\infty(0,T;L^n(\Omega))}$). Furthermore the constant $L(\ell)$ has the following explicit estimate

$$L(\ell) = O\big(\exp(C\ell^2)\big) \text{ as } \ell \to \infty \tag{73}$$

for some positive constant $C = C(T,\Omega)$, independent of ℓ and (w_0,w_1).

Theorem 8 is a consequence of the following Carleman estimate proved by Ruiz in [23]:

Lemma 9 *Let $\eta \in (0,1)$, $\mu > 0$, and $D_\mu \triangleq \Big\{(t,x) \in \mathbb{R}^{1+n} \,\Big|\, \eta^2 t^2 - |x|^2 > \mu\Big\}$. Let K be a compact subset of D_μ. Then there is a $\lambda_0 > 0$ and a constant $C = C(K,\mu)$ such that*

$$\lambda|e^{2\lambda\varphi}v|^2_{L^2(K)} \le C|e^{2\lambda\varphi}(v_{tt} - \Delta v)|^2_{H^{-1}(K)}, \qquad \forall\ \lambda > \lambda_0 \text{ and } v \in C_0^\infty(K).$$

Proof of Theorem 7. Let us fix the initial and final date (y_0, y_1), $(z_0, z_1) \in H_0^1(\Omega) \times L^2(\Omega)$ and let us introduce the continuous function

$$h(s) \triangleq \begin{cases} s^{-1}[f(s) - f(0)], & \text{if } s \neq 0; \\ f'(0), & \text{if } s = 0. \end{cases}$$

For any given $z(\cdot) \in L^\infty(0,T; L^2(\Omega))$, by Theorem 8 and the duality argument, we conclude that there exists a control $u \in L^2((0,T) \times \omega)$ such that the solution $y = y(\cdot\,; z(\cdot))$ of the following equation

$$\begin{cases} y'' - \Delta y + h(z(\cdot))y + f(0) = \chi_\omega(x)u(t,x) & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0, \quad y'(0) = y_1 & \text{in } \Omega \end{cases} \tag{74}$$

satisfies

$$y(T) = z_0, \quad y'(T) = z_1 \qquad \text{in } \Omega;$$

furthermore, concerning the control u, one have the estimate

$$|u|^2_{L^2((0,T)\times\omega)} \leq C \exp\big(C|h(z(\cdot))|^2_{L^\infty(0,T;L^n(\Omega))}\big)$$

for a constant $C = C(T, \Omega, f(0), |y_0|_{H_0^1(\Omega)}, |y_1|_{L^2(\Omega)}, |z_0|_{H_0^1(\Omega)}, |z_1|_{L^2(\Omega)})$. Thus, for any $\varepsilon \in (0, 4]$, we have

$$|u|^{2(1+\varepsilon)}_{L^2((0,T)\times\omega)} \leq C \exp\big(C|h(z(\cdot))|^2_{L^\infty(0,T;L^n(\Omega))}\big). \tag{75}$$

However, by calculus, we have

$$\begin{aligned} \exp\big(C|h(z(\cdot))|^2_{L^\infty(0,T;L^n(\Omega))}\big) &= \sum_{j=0}^{\infty} \frac{C^j}{j!} \operatorname*{ess\,sup}_{t\in(0,T)} \Big(\int_\Omega |h(z(t,x))|^n dx \Big)^{\frac{2j}{n}} \\ &= \sum_{j=0}^{n-1} \frac{C^j}{j!} \operatorname*{ess\,sup}_{t\in(0,T)} \Big(\int_\Omega |h(z(t,x))|^n dx \Big)^{\frac{2j}{n}} \\ &\quad + \sum_{j=n}^{\infty} \frac{C^j}{j!} \operatorname*{ess\,sup}_{t\in(0,T)} \Big(\int_\Omega |h(z(t,x))|^n dx \Big)^{\frac{2j}{n}} \\ &\leq C + \operatorname*{ess\,sup}_{t\in(0,T)} \Big(\int_\Omega |h(z(t,x))|^n dx \Big)^2 \\ &\quad + \sum_{j=n}^{\infty} \frac{C^j}{j!} \operatorname*{ess\,sup}_{t\in(0,T)} \Big(\int_\Omega |h(z(t,x))|^n dx \Big)^{\frac{2j}{n}} \\ &\leq C + 2\sum_{j=n}^{\infty} \frac{C^j}{j!} \operatorname*{ess\,sup}_{t\in(0,T)} \Big(\int_\Omega |h(z(t,x))|^n dx \Big)^{\frac{2j}{n}}. \end{aligned} \tag{76}$$

Note that for any $j \geq n$, it holds

$$\int_\Omega |h(z(t,x))|^n dx \leq C\Big(\int_\Omega |h(z(t,x))|^{2j} dx\Big)^{\frac{n}{2j}}.$$

Thus, by (75)–(76), we have

$$\begin{aligned} &\exp\big(C|h(z(\cdot))|^2_{L^\infty(0,T;L^n(\Omega))}\big) \\ &\leq C\Big[1+\sum_{j=n}^{\infty}\frac{C^j}{j!}\operatorname*{ess\,sup}_{t\in(0,T)}\int_\Omega |h(z(t,x))|^{2j}dx\Big] \\ &\leq C\Big[1+\operatorname*{ess\,sup}_{t\in(0,T)}\int_\Omega e^{C|h(z(t,x))|^2}dx\Big]. \end{aligned} \tag{77}$$

However, by our assumption (71), we have

$$e^{C|h(z(t,x))|^2} \leq C(1+|z(t,x)|^2). \tag{78}$$

Thus, by (75) and (77)–(78), we conclude that

$$|u|^{2(1+\varepsilon)}_{L^2((0,T)\times\omega)} \leq C(1+|z(t,x)|^2_{L^\infty(0,T;L^2(\Omega))}).$$

Thus

$$|u|^2_{L^2((0,T)\times\omega)} \leq C(1+|z(t,x)|^{\frac{2}{1+\varepsilon}}_{L^\infty(0,T;L^2(\Omega))}). \tag{79}$$

Now, applying the usual energy estimate to system (74) and noting (79), we end up with

$$|y|_{C([0,T];H^1_0(\Omega))} \leq C(1+|z(t,x)|^{\frac{4}{1+\varepsilon}}_{L^\infty(0,T;L^2(\Omega))}), \quad \forall\,\varepsilon\in(0,4] \tag{80}$$

for some constant $C = C(T,\Omega,f(0),|y_0|_{H^1_0(\Omega)},|y_1|_{L^2(\Omega)},|z_0|_{H^1_0(\Omega)},|z_1|_{L^2(\Omega)})$. Consequently, if we take $\varepsilon = 4$ in (80), the desired result follows from the fixed point technique. □

Note that in the above theorems, we need T to be greater than some positive time duration whenever the *fixed* controller ω is a proper open-subset of Ω. This is due to the finite speed of propagation of solutions of the wave equation.

Now, one has such a question: What happens for the exact controllability of system (54) with *changing* controller?

In [26], under suitable conditions, the author showed the *rapid* exact controllability of system (54) with changing controller. In order to state this result, without loss of generality, we assume the following:

$$\begin{cases} \inf\Big\{x_1\in\mathbb{R}\;\Big|\;\exists x'\in\mathbb{R}^{n-1}\text{ such that }(x_1,x')\in\Omega\Big\}=0, \\ \sup\Big\{x_1\in\mathbb{R}\;\Big|\;\exists x'\in\mathbb{R}^{n-1}\text{ such that }(x_1,x')\in\Omega\Big\}\equiv\beta>0. \end{cases}$$

Fix any $S > 0$ and $0 < \sigma < S$. Put

$$a = (S - \sigma)/\beta, \quad K_\sigma = \Big\{(t, x_1) \in [0, S] \times [0, \beta] \ \Big| \ ax_1 < t < ax_1 + \sigma\Big\}$$

and

$$D_\sigma = (K_\sigma \times \mathbb{R}^{n-1}) \bigcap ([0, S] \times \Omega).$$

We need the following assumption on $\mathcal{G}$, the class of controllers.

(H) *Class $\mathcal{G}$ is a family of set-valued functions $G : [0, \infty) \to 2^\Omega$ (the set of all subsets in Ω) with the following properties:*

(*i*) *Any $G(\cdot) \in \mathcal{G}$ is continuous with respect to the Hausdorff metric (defined on 2^Ω);*

(*ii*) *For any $S > 0$, there exists a $G(\cdot) \in \mathcal{G}$ and a $\sigma \in (0, S)$, such that $D_\sigma \subset \{(t, x) \in [0, S] \times \Omega \ \big| \ x \in G(t), t \in [0, S]\}$.*

The main result in [26] reads as follows:

Theorem 10 *Let assumption (H) hold and $f(\cdot) \in C^1(\mathbb{R})$ with $f'(\cdot) \in L^\infty(\mathbb{R})$. Then there is a controller $G(\cdot) \in \mathcal{G}$ such that system (54) with ω replaced by $G(t)$ is exactly controllable in $H_0^1(\Omega) \times L^2(\Omega)$ at any given time duration $T > 0$.*

4 The linear plate equations

Put

$$V = \Big\{g \in H^3(\Omega) \cap H_0^1(\Omega) \ \Big| \ \Delta g = 0 \text{ on } \Gamma\Big\}.$$

Denote by V' the dual space of V with respect to the pivot space $L^2(\Omega)$.

This section is addressed to the exact (boundary) controllability of the following linear plate equation:

$$\begin{cases} y_{tt} + \Delta^2 y = p(t, x) y_t + \displaystyle\sum_{|\alpha| \le 2} p_\alpha(t, x) \partial_x^\alpha y(t, x) & \text{in } Q, \\ y = \chi_{\Gamma_0}(x) u_1(t, x), \quad \Delta y = \chi_{\Gamma_0}(x) u_2(t, x) & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } \Omega. \end{cases} \tag{81}$$

In (81), the "state space" and "control space" are chosen respectively to be $H^{-1}(\Omega) \times V'$ and $L^2(\Gamma_0) \times L^2(\Gamma_0)$.

We have the following result.

Theorem 11 *Let $T > 0$, $p \in W^{1,\infty}(Q)$, $p_\alpha \in W^{|\alpha|,\infty}(Q)$ for $|\alpha| \le 2$. Then for any given (y_0, y_1), $(z_0, z_1) \in H^{-1}(\Omega) \times V'$, there are two controls $u_1 \in L^2(\Sigma_0)$ and $u_2 \in H^{-1}(0, T; L^2(\Gamma_0))$ such that the weak solution y of (81) satisfies $y(T) = z_0$ and $y_t(T) = z_1$ in Ω.*

Remark 6 *We refer to [2], [12], [13], [14], [20], [32], and so on for earlier exact controllability results of the linear plate equation.*

Similar to the wave equation, by means of the duality argument, in order to prove Theorem 11, one needs to establish the (boundary) observability inequality for the following plate equations with lower order terms:

$$\begin{cases} w_{tt} - \Delta w = q(t,x)w_t + \sum\limits_{|\alpha|\le 2} q_\alpha(t,x)\partial_x^\alpha w(t,x) & \text{in } Q, \\ w = \Delta w = 0 & \text{on } \Sigma, \\ w(0) = w_0, \quad w_t(0) = w_1 & \text{in } \Omega. \end{cases} \tag{82}$$

More precisely, we need the following *a priori* estimate for solutions of (82):

Theorem 12 *Let $T > 0$, $q, q_\alpha \in L^\infty(Q)$ for $|\alpha| = 2$, $q_\alpha \in L^{n+1}(Q)$ for $|\alpha| \le 1$. Then for any $(w_0, w_1) \in V \times H_0^1(\Omega)$, the weak solution $w \in C([0,T];V) \cap C^1([0,T];H_0^1(\Omega))$ of (82) satisfies*

$$|w_0|_V^2 + |w_1|_{H_0^1(\Omega)}^2 \le C\left[\|\frac{\partial w_t}{\partial \nu}\|_{L^2(\Sigma_0)}^2 + \|\frac{\partial \Delta w}{\partial \nu}\|_{L^2(\Sigma_0)}^2\right]. \tag{83}$$

Theorem 12 is also new. The proof of Theorem 12 is based on the following fundamental point-wise estimate for the Schrödinger operator.

Lemma 13 ([29]) *Let $w \in C^2(\mathbb{R}\times\mathbb{R}\times\mathbb{R}^n;\mathbb{C})$, Ψ and Φ be two real constants. Let $(a^0, a^1, \cdots, a^n) \in \mathbb{R}^{1+n}$ and $(t_0, x_0^1, \cdots, x_0^n) \in \mathbb{R}^{1+n}$ be given. Let*

$$\ell(t,s,x) = \sum_j a^j(x^j - x_0^j)^2 + a^0(t-t_0)^2 + a^0(s-t_0)^2, \quad \theta(t,s,x) = e^{\ell(t,s,x)}, \tag{84}$$

where $(t,s,x) \overset{\triangle}{=} (t,s,x^1,\cdots,x^n) \in \mathbb{R}^{2+n}$. Then for any $\varepsilon > 0$, it holds

$$\begin{aligned} &(1+\varepsilon^{-1})\theta^2|iw_t - iw_s + \Delta w|^2 \\ &\ge \theta^2\Big[4(\Psi + \sum_j \ell_{jj})\sum_k \ell_k^2 - 2|\Psi + \sum_j \ell_{jj}|\sum_k \ell_k^2 - \varepsilon(\Psi + \sum_j \ell_{jj})^2 \\ &\quad -\Psi^2 - \Phi^2 + 2\Phi\sum_j \ell_{jj} + \ell_{tt} + \ell_{ss}\Big]|w|^2 \\ &\quad +2\theta^2\Big\{\sum_{j,k}[\ell_{kj}(w_j\bar{w}_k + w_k\bar{w}_j - |w_k|^2)] - |\Psi + \sum_j l_{jj}|\sum_k(|w_k|^2)\Big\} \\ &\quad -2\sum_j\Big\{\theta^2\ell_j(2\sum_k \ell_k^2 + \Phi)|w|^2 + \theta^2\Big[\sum_k(\ell_k^2 + \frac{1}{2}\ell_{kk})\Big](w_j\bar{w} + \bar{w}_j w) \\ &\quad -\theta^2\ell_j\sum_k|w_k|^2 + \theta^2\sum_k[\ell_k(w_j\bar{w}_k + \bar{w}_j w_k)]\Big\}_j - [(\ell_t - \ell_s)\theta^2|w|^2)_t \\ &\quad +[(\ell_t - \ell_s)\theta^2|w|^2)_s + 2\sum_j\Big[\theta^2(\ell_t - \ell_s)(\xi_j\eta - \xi\eta_j)\Big]_j \\ &\quad +2\sum_j\Big\{[\theta^2\ell_j(\xi_j\eta - \xi\eta_j)]_t - [\theta^2\ell_j(\xi_t\eta - \xi\eta_t)]_j \\ &\quad -[\theta^2\ell_j(\xi_j\eta - \xi\eta_j)]_s + [\theta^2\ell_j(\xi_s\eta - \xi\eta_s)]_j\Big\}, \end{aligned} \tag{85}$$

where $i = \sqrt{-1},\ \xi \triangleq \Re w$ *and* $\eta \triangleq \Im w$.

Proof of Lemma 13. We borrow some ideas from [18, pp. 124]. Denote

$$\mathcal{P}w \triangleq iw_t - iw_s + \Delta w.$$

Introducing the function

$$v(t,s,x) = \theta(t,s,x)w(t,s,x), \qquad (t,s,x) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^n \tag{86}$$

and using the equality

$$\left\{\begin{array}{rcl} w_t &=& \theta^{-1}(v_t - \ell_t v),\\ w_s &=& \theta^{-1}(v_s - \ell_s v),\\ w_{jj} &=& \theta^{-1}[v_{jj} - 2\ell_j v_j + (\ell_j^2 - \ell_{jj})v], \quad j = 1,\dots,n, \end{array}\right. \tag{87}$$

we obtain

$$\begin{aligned} \theta^2|\mathcal{P}w|^2 &\equiv |\theta(iw_t - iw_s + \Delta w)|^2\\ &= |i(v_t - \ell_t v) - i(v_s - \ell_s v) + \sum_j [v_{jj} - 2\ell_j v_j + (\ell_j^2 - \ell_{jj})v]|^2\\ &= \|\Big\{iv_t - iv_s + \sum_j [v_{jj} + (\ell_j^2 - \ell_{jj})v] - \Psi v\Big\}\\ &\quad - \Big[2\sum_j (\ell_j v_j) + \Phi v + i(\ell_t - \ell_s)v\Big] + (\Phi v + \Psi v)\|^2. \end{aligned} \tag{88}$$

Denote

$$\left\{\begin{array}{l} I_1 \triangleq iv_t - iv_s + \sum_j [v_{jj} + (\ell_j^2 - \ell_{jj})v] - \Psi v,\\ I_2 \triangleq 2\sum_j (\ell_j v_j) + \Phi v + i(\ell_t - \ell_s)v,\\ I_3 \triangleq \Phi v + \Psi v. \end{array}\right. \tag{89}$$

Then

$$\begin{aligned} &\theta^2|\mathcal{P}w|^2 = |I_1 - I_2 + I_3|^2\\ &= |I_1|^2 + |I_2|^2 + |I_3|^2 - (I_1\bar{I}_2 + I_2\bar{I}_1) - (I_2\bar{I}_3 + I_3\bar{I}_2) + (I_1\bar{I}_3 + I_3\bar{I}_1)\\ &\geq |I_3|^2 - (I_1\bar{I}_2 + I_2\bar{I}_1) - (I_2\bar{I}_3 + I_3\bar{I}_2) + (I_1\bar{I}_3 + I_3\bar{I}_1). \end{aligned} \tag{90}$$

However, by (84) and noting $ia + \overline{ia} = 0$ ($\forall\ a \in \mathbb{R}$), we have

$$
\begin{aligned}
&I_1\bar{I}_2 + I_2\bar{I}_1\\
&= \Phi(I_1\bar{v} + \bar{I}_1 v) + 2i\sum_j \Big[\ell_j(v_t\bar{v}_j - \bar{v}_t v_j - v_s\bar{v}_j + \bar{v}_s v_j)\Big]\\
&\quad +2\sum_{j,k}[\ell_j(v_{kk}\bar{v}_j + \bar{v}_{kk}v_j)]\\
&\quad +2\sum_j\Big\{\Big[\sum_k(\ell_k^2 - \ell_{kk}) - \Psi\Big]\ell_j(v\bar{v}_j + \bar{v}v_j)\Big\}\\
&\quad +(\ell_t - \ell_s)(v_t\bar{v} + \bar{v}_t v - v_s\bar{v} - \bar{v}_s v) + i(\ell_t - \ell_s)\sum_j(v\bar{v}_{jj} - \bar{v}v_{jj})\\
&= \Phi(I_1\bar{v} + \bar{I}_1 v) + 2i\sum_j\Big\{\ell_j[(v\bar{v}_j)_t - (\bar{v}_t v)_j - (v\bar{v}_j)_s + (\bar{v}_s v)_j]\Big\}\\
&\quad +i(\ell_t - \ell_s)\sum_j(v\bar{v}_j - \bar{v}v_j)_j + 2\sum_{j,k}\Big\{\ell_j[(v_k\bar{v}_j + \bar{v}_k v_j)_k - (|v_k|^2)_j]\Big\}\\
&\quad +2\sum_j\Big\{\Big[\sum_k(\ell_k^2 - \ell_{kk}) - \Psi\Big]\ell_j(|v|^2)_j\Big\} + (\ell_t - \ell_s)\Big[(|v|^2)_t - (|v|^2)_s\Big] \qquad (91)\\
&= \Phi(I_1\bar{v} + \bar{I}_1 v) + i\Big\{2\sum_j\Big[(\ell_j v\bar{v}_j)_t - (\ell_j\bar{v}_t v)_j\\
&\quad +\ell_{jj}v\bar{v}_t - (\ell_j v\bar{v}_j)_s + (\ell_j\bar{v}_s v)_j - \ell_{jj}v\bar{v}_s\Big]\\
&\quad +\sum_j\Big[(\ell_t - \ell_s)(v\bar{v}_j - \bar{v}v_j)\Big]_j\Big\}\\
&\quad +2\sum_j\Big\{\sum_k[\ell_k(v_j\bar{v}_k + \bar{v}_j v_k) - \ell_j|v_k|^2]\\
&\quad +\Big[\sum_k(\ell_k^2 - \ell_{kk}) - \Psi\Big]\ell_j|v|^2\Big\}_j - 2\sum_{j,k}[\ell_{kj}(v_j\bar{v}_k + \bar{v}_j v_k) - \ell_{jj}|v_k|^2]\\
&\quad -2\Big[\sum_k(\ell_k^2 - \ell_{kk}) - \Psi\Big]\sum_j\ell_{jj}|v|^2 - 4\sum_{j,k}(\ell_j\ell_k\ell_{jk})|v|^2\\
&\quad +[(\ell_t - \ell_s)|v|^2]_t - [(\ell_t - \ell_s)|v|^2]_s - (\ell_{tt} + \ell_s)|v|^2.
\end{aligned}
$$

Further,

$$
\begin{aligned}
I_2\bar{I}_3 + I_3\bar{I}_2 &= 2\Phi(\Phi + \Psi)|v|^2 + 2(\Phi + \Psi)\sum_j[\ell_j(v_j\bar{v} + \bar{v}_j v)]\\
&= 2(\Phi + \Psi)\sum_j(\ell_j|v|^2)_j + 2(\Phi + \Psi)(\Phi - \sum_j\ell_{jj})|v|^2.
\end{aligned} \qquad (92)
$$

Further,

$$\begin{aligned}
&I_1\bar{I}_3 + I_3\bar{I}_1\\
&= \Phi(I_1\bar{v} + \bar{I}_1 v) + i\Psi(v_t\bar{v} - \bar{v}_t v - v_s\bar{v} + \bar{v}_s v)\\
&\quad + \Psi\sum_j (v_{jj}\bar{v} + \bar{v}_{jj}v) + 2\Psi\Big[\sum_j(\ell_j^2 - \ell_{jj}) - \Psi\Big]|v|^2\\
&= \Phi(I_1\bar{v} + \bar{I}_1 v) + i\Psi(v_t\bar{v} - \bar{v}_t v - v_s\bar{v} + \bar{v}_s v)\\
&\quad + \Psi\sum_j [(v_j\bar{v} + \bar{v}_j v)_j - 2|v_j|^2] + 2\Psi\Big[\sum_j(\ell_j^2 - \ell_{jj}) - \Psi\Big]|v|^2.
\end{aligned} \tag{93}$$

Combining (90)–(93) and noting that

$$v\bar{v}_t = \frac{1}{2}[(|v|^2)_t - (v_t\bar{v} - \bar{v}_t v)], \qquad v\bar{v}_s = \frac{1}{2}[(|v|^2)_s - (v_s\bar{v} - \bar{v}_s v)], \tag{94}$$

we get

$$\begin{aligned}
&\theta^2|\mathcal{P}w|^2\\
&\geq \Big[2(\Psi + \sum_j \ell_{jj})\sum_k(\ell_k^2 - \ell_{kk}) + 4\sum_{j,k}(\ell_j\ell_k\ell_{kj})\\
&\quad -\Psi^2 - \Phi^2 + 2\Phi\sum_j \ell_{jj} + \ell_{tt} + \ell_{ss}\Big]|v|^2\\
&\quad +2\Big\{\sum_{j,k}[\ell_{jk}(v_j\bar{v}_k + v_k\bar{v}_j)] - (\Psi + \sum_j \ell_{jj})\sum_k |v_k|^2\Big\}\\
&\quad +i(\Psi + \sum_j \ell_{jj})(v_t\bar{v} - v\bar{v}_t - v_s\bar{v} + v\bar{v}_s)\\
&\quad -i\sum_j\Big[(\ell_t - \ell_s)(v\bar{v}_j - \bar{v}v_j)\Big]_j\\
&\quad -i\Big\{\Big(\sum_j \ell_{jj}|v|^2\Big)_t + 2\sum_j[(\ell_j v\bar{v}_j)_t - (\ell_j\bar{v}_t v)_j] - \Big(\sum_j \ell_{jj}|v|^2\Big)_s\\
&\quad -2\sum_j[(\ell_j v\bar{v}_j)_s - (\ell_j\bar{v}_s v)_j]\Big\} - [(\ell_t - \ell_s)|v|^2]_t + [(\ell_t - \ell_s)|v|^2]_s\\
&\quad -2\sum_j\Big\{\Phi\ell_j|v|^2 + \sum_k\Big[\ell_k(v_j\bar{v}_k + \bar{v}_j v_k) - \ell_j|v_k|^2\Big]\\
&\quad +\ell_j|v|^2\sum_k(\ell_k^2 - \ell_{kk}) - \frac{\Psi}{2}(v_j\bar{v} + \bar{v}_j v)\Big\}_j.
\end{aligned} \tag{95}$$

Returning to the function w, we obtain

$$\begin{aligned}
&\sum_{j,k}[\ell_{jk}(v_j\bar{v}_k+v_k\bar{v}_j)]-(\Psi+\sum_j\ell_{jj})\sum_k|v_k|^2\\
&=\theta^2\Big\{\sum_{j,k}\Big[\ell_{jk}\Big((\ell_jw+w_j)(\ell_k\bar{w}+\bar{w}_k)+(\ell_j\bar{w}+\bar{w}_j)(\ell_kw+w_k)\Big)\Big]\\
&\quad-(\Psi+\sum_j\ell_{jj})\sum_k|\ell_kw+w_k|^2\Big\}\\
&=\theta^2\Big\{\sum_{j,k}\Big[\ell_{jk}\Big(2\ell_j\ell_k|w|^2+\ell_j(w\bar{w}_k+w_k\bar{w})+\ell_k(w\bar{w}_j+w_j\bar{w})\\
&\quad+(w_j\bar{w}_k+\bar{w}_jw_k)\Big)\Big]\\
&\quad-(\Psi+\sum_j\ell_{jj})\sum_k\Big[\ell_k^2|w|^2+|w_k|^2+\ell_k(|w|^2)_k\Big]\Big\}.
\end{aligned}\tag{96}$$

Noting that

$$\begin{aligned}
&\theta^2(\Psi+\sum_j\ell_{jj})\sum_k[\ell_k(|w|^2)_k]\\
&=\sum_k\Big[\theta^2(\Psi+\sum_j\ell_{jj})\ell_k|w|^2\Big]_k\\
&\quad-2\theta^2(\Psi+\sum_j\ell_{jj})\sum_k\ell_k^2|w|^2-\theta^2(\Psi+\sum_j\ell_{jj})\sum_k\ell_{kk}|w|^2
\end{aligned}\tag{97}$$

and $\sum_{j,k}[\ell_j\ell_{jk}(w\bar{w}_k+w_k\bar{w})]=\sum_{j,k}[\ell_k\ell_{jk}(w\bar{w}_j+w_j\bar{w})]$, we get

$$\begin{aligned}
&\sum_{j,k}[\ell_{jk}(v_j\bar{v}_k+v_k\bar{v}_j)]-(\Psi+\sum_j\ell_{jj})\sum_k|v_k|^2\\
&=\theta^2\Big\{\sum_{j,k}[\ell_{jk}(w_j\bar{w}_k+\bar{w}_jw_k)]-(\Psi+\sum_j\ell_{jj})\sum_k|w_k|^2\Big\}\\
&\quad+\theta^2\Big[2\sum_{j,k}(\ell_{jk}\ell_j\ell_k)+(\Psi+\sum_j\ell_{jj})\sum_k(\ell_k^2+\ell_{kk})\Big]|w|^2\\
&\quad-\sum_k[\theta^2(\Psi+\sum_j\ell_{jj})\ell_k|w|^2]_k+2\theta^2\sum_{j,k}[\ell_j\ell_{jk}(w\bar{w}_k+w_k\bar{w})].
\end{aligned}\tag{98}$$

However, thanks to the elementary inequality $|a|^2+|b|^2\geq a\cdot\bar{b}+\bar{a}\cdot b\geq-(|a|^2+|b|^2)$, and noting that $|\ell_{jk}|=\ell_{jk}$ (recalling (84)), we get

$$2\theta^2\sum_{j,k}[\ell_j\ell_{jk}(w\bar{w}_k+\bar{w}w_k)]\geq-\theta^2\Big[\sum_{j,k}(\ell_{jk}|w_k|^2)+4\sum_{j,k}(\ell_j^2\ell_{jk})|w|^2\Big].\tag{99}$$

Thus, by (98) and (99), and noting that $\ell_j\ell_k\ell_{jk} = \ell_j^2\ell_{jk}$ $(j,\ k = 1, \cdots, n)$, we get

$$\begin{aligned}
&\sum_{j,k}[\ell_{jk}(v_j\bar{v}_k + v_k\bar{v}_j)] - (\Psi + \sum_j \ell_{jj})\sum_k |v_k|^2 \\
&\geq \theta^2\Big\{\sum_{j,k}[\ell_{jk}(w_j\bar{w}_k + \bar{w}_j w_k - |w_k|^2)] - (\Psi + \sum_j \ell_{jj})\sum_k |w_k|^2\Big\} \\
&\quad +\theta^2\Big\{\Big[(\Psi + \sum_j \ell_{jj})\sum_k(\ell_k^2 + \ell_{kk}) - 2\sum_{j,k}(\ell_j\ell_k\ell_{jk})\Big]|w|^2\Big\} \\
&\quad -\sum_k\Big[\theta^2(\Psi + \sum_j \ell_{jj})\ell_k|w|^2\Big]_k.
\end{aligned} \tag{100}$$

Further (recalling $\mathcal{P}w \triangleq iw_t - iw_s + \Delta w$),

$$\begin{aligned}
&i(\Psi + \sum_j \ell_{jj})(v_t\bar{v} - v\bar{v}_t - v_s\bar{v} + v\bar{v}_s) \\
&= i\theta^2(\Psi + \sum_j \ell_{jj})\Big[(\ell_t w + w_t)\bar{w} - w(\ell_t\bar{w} + \bar{w}_t) \\
&\quad -(\ell_s w + w_s)\bar{w} + w(\ell_s\bar{w} + \bar{w}_s)\Big] \\
&= \theta^2(\Psi + \sum_j \ell_{jj})\Big[(iw_t - iw_s)\bar{w} + w(\overline{iw_t - iw_s})\Big] \\
&= \theta^2(\Psi + \sum_j \ell_{jj})\Big[(\mathcal{P}w)\bar{w} + w(\overline{\mathcal{P}w}) - \sum_k(w_{kk}\bar{w} + \bar{w}_{kk}w)\Big] \\
&= \theta^2(\Psi + \sum_j \ell_{jj})[(\mathcal{P}w)\bar{w} + w(\overline{\mathcal{P}w})] \\
&\quad +2\theta^2(\Psi + \sum_j \ell_{jj})\sum_k |w_k|^2 - \theta^2(\Psi + \sum_j \ell_{jj})\sum_k(w_k\bar{w} + \bar{w}_k w)_k.
\end{aligned} \tag{101}$$

However, we have

$$\begin{aligned}
&\theta^2(\Psi + \textstyle\sum_j \ell_{jj})\sum_k(w_k\bar{w} + \bar{w}_k w)_k \\
&= \sum_k\Big[\theta^2(\Psi + \sum_j \ell_{jj})(w_k\bar{w} + \bar{w}_k w)\Big]_k \\
&\quad -2\theta^2(\Psi + \sum_j \ell_{jj})\sum_k[\ell_k(w_k\bar{w} + \bar{w}_k w)] \\
&\leq \sum_k\Big[\theta^2(\Psi + \sum_j \ell_{jj})(w_k\bar{w} + \bar{w}_k w)\Big]_k \\
&\quad +2\theta^2|\Psi + \sum_j \ell_{jj}|\Big[\sum_k(|w_k|^2) + \sum_k(\ell_k^2|w|^2)\Big].
\end{aligned} \tag{102}$$

Thus, by (101) and (102), we have

$$\begin{aligned}
&i(\Psi+\textstyle\sum_j \ell_{jj})(v_t\bar v - v\bar v_t - v_s\bar v + v\bar v_s)\\
&\geq \theta^2(\Psi+\sum_j \ell_{jj})\Big[(\mathcal{P}w)\bar w + w(\overline{\mathcal{P}w})\Big]\\
&\quad +2\theta^2\Big[(\Psi+\sum_j \ell_{jj})\sum_k |w_k|^2 - |\Psi+\sum_j \ell_{jj}|\sum_k (|w_k|^2)\Big]\\
&\quad -\sum_k \Big[\theta^2(\Psi+\sum_j \ell_{jj})(w_k\bar w + \bar w_k w)\Big]_k\\
&\quad -2\theta^2|\Psi+\sum_j \ell_{jj}|\sum_k (\ell_k^2|w|^2).
\end{aligned} \tag{103}$$

Further,

$$\begin{aligned}
&\Big(\sum_j \ell_{jj}|v|^2\Big)_t + 2\sum_j \Big[(\ell_j v\bar v_j)_t - (\ell_j \bar v_t v)_j\Big]\\
&-\Big(\sum_j \ell_{jj}|v|^2\Big)_s - 2\sum_j \Big[(\ell_j v\bar v_j)_s - (\ell_j \bar v_s v)_j\Big]\\
&= \Big(\theta^2\sum_j \ell_{jj}|w|^2\Big)_t - \Big(\theta^2\sum_j \ell_{jj}|w|^2\Big)_s\\
&\quad +2\sum_j \Big[(\theta^2\ell_j^2|w|^2 + \theta^2\ell_j w\bar w_j)_t - (\theta^2\ell_j\ell_t|w|^2 + \theta^2\ell_j w\bar w_t)_j\Big]\\
&\quad -2\sum_j \Big[(\theta^2\ell_j^2|w|^2 + \theta^2\ell_j w\bar w_j)_s - (\theta^2\ell_j\ell_s|w|^2 + \theta^2\ell_j w\bar w_s)_j\Big].
\end{aligned} \tag{104}$$

Further,

$$\sum_j \Big[(\ell_t-\ell_s)(v\bar v_j - \bar v v_j\Big]_j = \sum_j \Big[\theta^2(\ell_t-\ell_s)(w\bar w_j - \bar w w_j)\Big]_j. \tag{105}$$

Further,

$$\begin{aligned}
&\sum_j \Big\{\Phi\ell_j|v|^2 + \sum_k [\ell_k(v_j\bar v_k + \bar v_j v_k) - \ell_j|v_k|^2]\\
&\quad +\ell_j|v|^2\sum_k(\ell_k^2 - \ell_{kk}) - \frac{\Psi}{2}(v_j\bar v + \bar v_j v)\Big\}_j\\
&= \sum_j \Big\{\theta^2\sum_k \Big[\ell_k(w_j\bar w_k + \bar w_j w_k) - \ell_j|w_k|^2\Big]\\
&\quad +\theta^2\ell_j\Big[\sum_k(2\ell_k^2 - \ell_{kk}) - \Psi + \Phi\Big]|w|^2\\
&\quad +\theta^2\Big(\sum_k \ell_k^2 - \frac{\Psi}{2}\Big)(w_j\bar w + \bar w_j w)\Big\}_j.
\end{aligned} \tag{106}$$

Consequently we obtain that

$$
\begin{aligned}
&\theta^2|\mathcal{P}w|^2\\
&\geq \theta^2(\Psi+\sum_j \ell_{jj})[(\mathcal{P}w)\bar{w}+(\overline{\mathcal{P}w})w]\\
&\quad+\theta^2\Big[4(\Psi+\sum_j \ell_{jj})\sum_k \ell_k^2-2|\Psi+\sum_j \ell_{jj}|\sum_k \ell_k^2\\
&\quad-\Psi^2-\Phi^2+2\Phi\sum_j \ell_{jj}+\ell_{tt}+\ell_{ss}\Big]|w|^2\\
&\quad+2\theta^2\Big\{\sum_{j,k}\Big[\ell_{kj}(w_j\bar{w}_k+w_k\bar{w}_j-|w_k|^2)\Big]-|\Psi+\sum_j l_{jj}|\sum_k(|w_k|^2)\Big\}\\
&\quad-2\sum_j\Big\{\theta^2\ell_j(2\sum_k \ell_k^2+\Phi)|w|^2+\theta^2\Big[\sum_k(\ell_k^2+\frac{1}{2}\ell_{kk})\Big](w_j\bar{w}+\bar{w}_j w)\\
&\quad-\theta^2\ell_j\sum_k|w_k|^2+\theta^2\sum_k[\ell_k(w_j\bar{w}_k+\bar{w}_j w_k)]\Big\}_j\\
&\quad-[(\ell_t-\ell_s)\theta^2|w|^2]_t+[(\ell_t-\ell_s)\theta^2|w|^2]_s\\
&\quad-i\sum_j\Big[\theta^2(\ell_t-\ell_s)(\bar{w}_j w-w_j\bar{w})\Big]_j\\
&\quad-i\Big\{\Big[\theta^2\sum_j \ell_{jj}|w|^2\Big]_t-\Big[\theta^2\sum_j \ell_{jj}|w|^2\Big]_s\\
&\quad+2\sum_j\Big[(\theta^2\ell_j^2|w|^2+\theta^2\ell_j w\bar{w}_j)_t-(\theta^2\ell_j\ell_t|w|^2+\theta^2\ell_j w\bar{w}_t)_j\Big]\\
&\quad-2\sum_j\Big[(\theta^2\ell_j^2|w|^2+\theta^2\ell_j w\bar{w}_j)_s-(\theta^2\ell_j\ell_s|w|^2+\theta^2\ell_j w\bar{w}_s)_j\Big]\Big\}.
\end{aligned}
\tag{107}
$$

Now, recalling that $\xi \triangleq \Re u$ and $\eta \triangleq \Im u$, by means of a direct calculation, we obtain

$$
\begin{aligned}
&-i\sum_j\Big[\theta^2(\ell_t-\ell_s)(\bar{w}_j w-w_j\bar{w})\Big]_j\\
&\quad-i\Big\{\Big[\theta^2\sum_j \ell_{jj}|w|^2\Big]_t-\Big[\theta^2\sum_j \ell_{jj}|w|^2\Big]_s\\
&\quad+2\sum_j\Big[(\theta^2\ell_j^2|w|^2+\theta^2\ell_j w\bar{w}_j)_t-(\theta^2\ell_j\ell_t|w|^2+\theta^2\ell_j w\bar{w}_t)_j\Big]\\
&\quad-2\sum_j\Big[(\theta^2\ell_j^2|w|^2+\theta^2\ell_j w\bar{w}_j)_s-(\theta^2\ell_j\ell_s|w|^2+\theta^2\ell_j w\bar{w}_s)_j\Big]\Big\}\\
&=2\sum_j\Big[\theta^2(\ell_t-\ell_s)(\xi_j\eta-\xi\eta_j)\Big]_j+2\sum_j\Big\{[\theta^2\ell_j(\xi_j\eta-\xi\eta_j)]_t\\
&\quad-[\theta^2\ell_j(\xi_t\eta-\xi\eta_t)]_j-[\theta^2\ell_j(\xi_j\eta-\xi\eta_j)]_s+[\theta^2\ell_j(\xi_s\eta-\xi\eta_s)]_j\Big\}.
\end{aligned}
\tag{108}
$$

Finally, noting that for any $\varepsilon > 0$, we have

$$(\Psi + \sum_j \ell_{jj})[(\mathcal{P}w)\bar{w} + (\overline{\mathcal{P}w})w] \geq -\varepsilon^{-1}|\mathcal{P}w|^2 - \varepsilon(\Psi + \sum_j \ell_{jj})^2|w|^2. \quad (109)$$

Thus the desired estimate (85) follows from (107)–(109) immediately. □

5 The semi-linear plate equations

Finally, let us consider the controlled semi-linear plate equation:

$$\begin{cases} y_{tt} + \Delta^2 y + f(y) = \nabla \cdot \big(\chi_\omega(x)\nabla u(t,x)\big) & \text{in} \quad (0,T)\times\Omega, \\ y = \Delta y = 0 & \text{on} \quad (0,T)\times\Gamma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in} \quad \Omega. \end{cases} \quad (110)$$

In (110), we will choose the "state space" and the "control space" to be $H_0^1(\Omega)\times H^{-1}(\Omega)$ and $C([0,T];H_0^1(\Omega))$, respectively.

The exact boundary controllability problem of system (110) was studied in [15] for nonlinearity $f(\cdot)$ satisfying the following assumptions: f' is absolutely continuous and for some constant L

$$|f'(s)| + |f''(s)| \leq L, \quad \text{for a.e. } s \in \mathbb{R}. \quad (111)$$

Under certain additional conditions, with the help of the global implicit function theorem, the exact boundary controllability of (110) was obtained in [15]. Of course, using the same approach, one may obtain the exact internal controllability of system (110) under similar technical assumptions.

Recently, by means of global Carleman-type inequality, the author proved the following result:

Theorem 14 ([29]) *Let $\omega = \Omega \cap \mathcal{O}_\epsilon(\Gamma_0)$ for some $\epsilon > 0$, $f \in C^1(\mathbb{R})$ and (71) hold. Then system (110) is exactly controllable in $H_0^1(\Omega)\times H^{-1}(\Omega)$ at any given time duration $T > 0$ by means of control $u \in C([0,T];H_0^1(\Omega))$.*

Remark 7 *Obviously, the assumption on the nonlinearity $f(\cdot)$ in Theorem 14 is much weaker than (111). On the other hand, the assumptions in Theorem 14 are very close to that in Theorem 7. However, the controller ω in Theorem 14 is only assumed to be a neighborhood of Γ_0 rather than the whole boundary Γ.*

By the duality argument, Theorem 14 is a consequence of the following observability estimate for the plate equation.

Theorem 15 ([29]) *Let $T > 0$, $q = q_\alpha = 0$ for $2 \geq |\alpha| > 0$, $q_0(\cdot) \in L^\infty(0,T;L^n(\Omega))$. Let ω satisfy assumption (H1). Then there exists a constant $C = C(T,\Omega,\omega) > 0$ such that*

$$|(w_0,w_1)|^2_{H_0^1(\Omega)\times H^{-1}(\Omega)} \leq \mathcal{G}(r)\int_0^T\int_\omega |\nabla w|^2 dxdt, \quad \forall\ (w_0,w_1) \in H_0^1(\Omega)\times H^{-1}(\Omega), \quad (112)$$

where $w \in C([0,T];H_0^1(\Omega)) \cap C^1([0,T];H^{-1}(\Omega))$ *is the weak solution of (82),*

$$\mathcal{G}(r) = O(\exp(Cr^2)) \qquad \text{as } r \triangleq |q|_{L^\infty(0,T;L^n(\Omega))} \to \infty. \tag{113}$$

The proof of Theorem 15 is long, and based on Lemma 13 in an essential way.

References

[1] C. Bardos, G. Lebeau and J. Rauch, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary, SIAM J. Control and Optim.,* **30** (1992), pp. 1024–1065.

[2] N. Burq, *Contrôle de l'équation de schrödinger en présence d'obstacles strictement convexes, Memoires de la Soc. Math. de France,* 1993, No: 55.

[3] P. Cannarsa, V. Komornik, P. Loreti, *One-sided and internal controllability of semilinear wave equations with infinitely iterated logarithms, Discrete Contin. Dyn. Syst.,* **8** (2002), pp. 745–756.

[4] W. C. Chewning, *Controllability of the nonlinear wave equation in several space variables, SIAM J. Control and Optim.,* **14** (1976), pp. 9–25.

[5] M. A. Cirina, *Boundary controllability of nonlinear hyperbolic systems, SIAM J. Control,* **7** (1969), pp. 198–212.

[6] A. Doubova, E. Fernández-Cara, M. González-Burgos and E. Zuazua, *On the controllability of parabolic systems with a nonlinear term involving the state and the gradient, SIAM J. Control Optim.,* to appear.

[7] Yu. V. Egorov, *Some problems in the theory of optimal control, Z. Vycisl Mat. i Mat. Fiz.,* (1963), pp. 887–904.

[8] H. O. Fattorini, *Boundary control systems, SIAM J. Control,* **6** (1968), pp. 349–385.

[9] H. O. Fattorini, *Local controllability of a nonlinear wave equation, Math. Systems Theory,* **9** (1975), pp. 35–40.

[10] E. Fernández-Cara and E. Zuazua, *Null and approximate controllability for weakly blowing up semilinear heat equations. Annales de l'IHP. Analyse non linéaire,* **17** (2000), pp. 583–616.

[11] M. A. Kazemi and M. V. Klibanov, *Stability estimates for ill-posed Cauchy problems involving hyperbolic equations and inequalities, Appl. Anal.,* **50** (1993), pp. 93–102.

[12] J. U. Kime, *Exact semi-internal control of an Euler-Bernoulli equation, SIAM J. Control and Optim.,* **30** (1992), pp. 1001–1023.

[13] V. Komornik, *Exact Controllability and Stabilization (the Multiplier Method), John Wiley & Sons,* Masson, Paris, 1995.

[14] I. Lasiecka and R. Triggiani, *Exact controllability of the Euler-Bernoulli equation with controls in the Dirichlet and Neumann boundary conditions: a non-conservative case, SIAM J. Control & Optim.,* 27 (1989), pp. 330–373.

[15] I. Lasiecka, R. Triggiani, *Exact controllability of semilinear abstract systems with application to waves and plates boundary control problem, Appl. Math. Optim.,* **23** (1991), pp. 109–154.

[16] I. Lasiecka and R. Triggiani, *Carleman estimates and exact boundary controllability for a system of coupled, non-conservative second order hyperbolic equations,* in *Partial Differential Equations Methods in Control and Shape Analysis*, *Lecture Notes in Pure and Applied Mathematics*, Marcel Dekker, New York, **188** (1994), pp. 215–243.

[17] I. Lasiecka, R. Triggiani, X. Zhang, *Nonconservative wave equations with purely Neumann B.C.: Global uniqueness and observability in one shot, Contemp. Math.,* **268** (2000), pp. 227–326.

[18] M. M. Lavrentév, V. G. Romanov, and S. P. Shishataskii, *Ill-Posed Problems of Mathematics Physics and Analysis*, Transl. Math. Monogr. 64, AMS, Providence, RI, 1986.

[19] L. Li, X. Zhang, *Exact controllability for semilinear wave equation, J. Math. Anal. Appl.,* **250** (2000), pp. 589–597.

[20] J. L. Lions, *Contrôlabilité exacte, perturbations et systémes distribués, tome 1, RMA* No: 8, Masson, Paris, 1988.

[21] W. Li, *Observability estimate for the parabolic equations with non-homogeneous terms,* Preprint.

[22] A. López, X. Zhang, E. Zuazua, *Null controllability of the heat equation as singular limit of the exact controllability of dissipative wave equations, J. Math. Pures Appl.,* **79** (2000), pp. 741–808.

[23] A. Ruiz, *Unique continuation for weak solutions of the wave equation plus a potential, J. Math. Pures Appl.,* **71** (1992), pp. 455–467.

[24] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: recent progress and open problems, SIAM Rev.* **20** (1978), pp. 639–739.

[25] X. Zhang, *Exact Controllability of the Semilinear Distributed Parameter System and Some Related Problems,* Ph. D. Thesis, Fudan University, Shanghai, China (1998).

[26] X. Zhang, *Rapid exact controllability of the semilinear wave equation, Chinese Ann. Math. Ser. B,* **20** (1999), pp. 377–384.

[27] X. Zhang, *Explicit observability estimate for the wave equation with potential and its application, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.,* **456** (2000), pp. 1101–1115.

[28] X. Zhang, *Explicit observability inequalities for the wave equation with lower order terms by means of Carleman inequalities, SIAM J. Control Optim.,* **39** (2001), pp. 812–834.

[29] X. Zhang, *Exact controllability of the semilinear plate equations, Asymptot. Anal.,* **27** (2001), pp. 95–125.

[30] X. Zhang, *Exact Controllability of Semi-linear Distributed Parameter Systems,* Chinese Higher Education Press, Beijing, in press. (In Chinese).

[31] X. Zhang and E. Zuazua, *Exact controllability of the semi-linear wave equation,* in *Sixty open problems in the mathematics of systems and control,* edited by V. D. Blondel and A. Megretski, Princeton University Press, to appear.

[32] E. Zuazua, *Contrôlabilité exacte en un temps arbitrairement petit de quelques modèles de plaques,* Appendix I in [20], pp. 465–491.

[33] E. Zuazua, *Exact boundary controllability for the semilinear wave equation,* in *Nonlinear partial differential equations and their applications*, Collége de France Seminar, Vol. X (Paris, 1987-1988), Pitman Res. Notes Math. Ser., **220**, Longman Sci. Tech., Harlow, 1991, pp. 357–391.

[34] E. Zuazua, *Exact controllability for semilinear wave equations in one space dimension, Ann. Inst. H. Poincaré Anal. Non Linéaire,* **10** (1993), pp. 109–129.

[35] E. Zuazua, *Some problems and results on the controllability of partial differential equations,* in *Progress in Mathematics,* Vol: 169, Birkhäuser Verlag, Basel/Switzerland, 1998, pp. 276–311.

Bol. Soc. Esp. Mat. Apl. nº28(2004), 129–154

EDPs de difusión y transporte óptimo de masa

J. A. Carrillo

ICREA-Departament de Matemàtiques,
Universitat Autònoma de Barcelona

carrillo@mat.uab.es.

Resumen

Se describen las principales aplicaciones que la teoría de transporte óptimo de masa tiene en el comportamiento asintótico de EDPs. Se incluyen ecuaciones de difusión no lineal, ecuaciones cinéticas homogéneas para medios granulares, ecuaciones de Fokker-Planck no lineal, ... La estructura de flujo gradiente infinito dimensional, respecto a una estructura riemaniana que induce la distancia de Wasserstein, permite demostrar la contractividad de distancias de Wasserstein. A partir de esta propiedad fundamental se deducen propiedades cualitativas de las soluciones.

1 Introducción

El estudio del comportamiento asintótico de ecuaciones de difusión no lineal ha tomado un impulso renovado en los últimos años debido a las profundas conexiones que se han descubierto entre estas ecuaciones y distintos métodos y teorías aparentemente lejanas. Estaremos interesados en estudiar el comportamiento cuando $t \to \infty$ de soluciones de las ecuaciones:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left[\rho \nabla \left(U'(\rho) + V + W * \rho \right) \right], \tag{1}$$

donde $V, W : \mathbb{R}^d \longrightarrow \mathbb{R}$ son llamados usualmente el potencial de confinamiento y el potencial de interacción respectivamente y la función $U : \mathbb{R} \longrightarrow \mathbb{R}$ es llamada la energía interna. Nótese que si $P(\rho)$ se define tal que $P'(\rho) = \rho U''(\rho)$, entonces el primer término del lado derecho en (1) es $\Delta P(\rho)$. El símbolo $*$ denota el producto convolución.

Estas ecuaciones incluyen como casos particulares a: las ecuaciones de Fokker-Planck lineales y no lineales [23], ecuaciones de medios porosos y difusión

rápida [52], ecuaciones en medios granulares donde aparecen interacciones no locales [6, 7], ...

La relación entre la teoría del comportamiento asintótico para estas ecuaciones y las técnicas e ideas usadas en teoría cinética de gases, tuvo su origen en los trabajos de G. Toscani [45, 47] donde se utilizan las ideas del funcional de entropía clásico en teoría cinética de gases y el teorema-H de Boltzmann, para describir el comportamiento asintótico de la ecuación de Fokker-Planck lineal y la ecuación del calor. Este uso de técnicas de comportamiento asintótico para ecuaciones cinéticas no lineales en las ecuaciones lineales de Fokker-Planck y del calor, supuso el germen de una estrategia de ataque desde un punto de vista "cinético" del comportamiento asintótico para las ecuaciones de difusión no lineal que se plasmó en [23]. Coetáneamente otros dos grupos de investigadores descubrieron esta relación, pero lo más inesperado fue, que con puntos de vista distintos y complementarios, enriquecieron aún más la teoría del comportamiento asintótico para estas ecuaciones [42, 30].

Hoy en día, se ha extendido el método de entropía relativa a aplicaciones en ecuaciones de difusión no lineal generales [19], mejores órdenes de convergencia en ecuaciones de difusión rápida [22], ecuaciones de convección-difusión [16], ecuaciones cinéticas homogéneas para medios granulares [20], ecuaciones cinéticas no homogéneas [29, 13], ... Referimos al informe divulgativo de investigación en [1] para una exposición amena de dicho método y sus raíces históricas. El método de entropía relativa usa un funcional de Liapunov destacado de las ecuaciones (1), para el cual se puede estimar con precisión su disipación debida a la evolución temporal mediante desigualdades de tipo Logarítmico-Sobolev. Estas desigualdades son clásicas en teoría de probabilidades [36] y fueron ya usadas para estudiar las ecuaciones de Fokker-Planck lineal en [5]. El primero en hacer la conexión con las ecuaciones cinéticas homogéneas de Fokker-Planck fue de nuevo G. Toscani en [46], el cual dió una demostración muy sencilla de la convergencia hacia equilibrio basada en esta desigualdad.

Otra estrategia fue introducida por F. Otto en [42], el cual introdujo una estructura riemaniana formal en el espacio de las medidas de probabilidad para ver la ecuación de medios porosos como un flujo gradiente infinito dimensional. Este punto de vista le permitió conectar la entropía relativa y su disipación con la distancia inducida por esta estructura riemaniana formal, que a la postre es la distancia de Wasserstein para medidas de probabilidad. Estas distancias están conectadas con el problema clásico de Monge-Kantorovich de transporte de masa. Esta imponente conexión entre dos áreas aparentemente distantes como el comportamiento asintótico de EDPs en base a funcionales de Liapunov y estimaciones a priori y el campo de transporte óptimo ha dado lugar a una explosión de nuevos resultados y conexiones impensables años atrás.

La estructura riemaniana formal permitió demostrar desigualdades funcionales: Log-Sobolev, Talagrand, HWI, desigualdades Log-Sobolev generalizadas, ... que han tenido influencia en otros campos y en otras direcciones [43, 28, 24]. Dicha estructura está basada en conceptos previos introducidos por R. J. McCann [40] relativos a las geodésicas respecto a la distancia de Wasserstein y la

convexidad de los funcionales de entropía. El tratado reciente de C. Villani [53] hace un extenso e intenso desarrollo de todas las ideas apuntadas anteriormente. La demostración rigurosa de la estructura de flujo gradiente de las ecuaciones (1) es sujeto de hirviente actualidad [21, 4].

Por último, J. Dolbeault y M. del Pino [30] tomaron como estrategia la demostración directa de las desigualdades Log-Sobolev generalizadas por métodos variacionales, los cuales, relacionan dichas desigualdades con la unicidad de solución para ecuaciones elípticas no lineales en $\mathbb{R}^d$.

El objetivo de este trabajo es dar una pincelada de las aplicaciones que las técnicas de transporte óptimo de masa encuentran en este tipo de problemas obviando totalmente el punto de vista del método de entropía relativa. Nos concentraremos en estudiar en profundidad la consecuencia primordial que tiene la estructura de flujo gradiente: la contractividad de las distancias de Wasserstein para las soluciones de (1). Como intentaremos mostrar en lo que sigue, dicha contractividad tiene implicaciones importantes en el comportamiento asintótico y en propiedades cualitativas de las soluciones.

2 Distancias entre medidas de probabilidad y transporte óptimo de masa

La distancia de Wasserstein [37, 54] entre dos medidas de probabilidad en $\mathbb{R}^d$ se define como

$$d_2(\mu,\nu) = \inf\Big\{\sqrt{E|X-Y|^2};\ \mathrm{ley}(X)=\mu,\ \mathrm{ley}(Y)=\nu\Big\}, \tag{2}$$

donde el ínfimo se toma entre todas las parejas de variables aleatorias cuyas leyes vienen dadas por las medidas de probabilidad μ y ν, es decir, en notación más analítica tenemos

$$d_2(\mu,\nu) = \inf\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^2\, d\gamma(x,y);\ \gamma\in\Gamma(\mu,\nu)\right\}^{1/2} \tag{3}$$

donde $\Gamma(\mu,\nu)$ es el conjunto de todas las medidas de probabilidad en $\mathbb{R}^d\times\mathbb{R}^d$ que tienen por marginales a μ y ν, y $|\cdot|$ es la norma euclídea en $\mathbb{R}^d$. Nótese que dicha distancia es finita siempre que ambas medidas de probabilidad tengan momentos de orden dos finitos. Sea $\mathcal{P}_2(\mathbb{R}^d)$ el conjunto de todas las medidas de probabilidad con momento de orden dos finito. Es conocido [31, 34, 53] que la distancia d_2 hace del espacio $\mathcal{P}_2(\mathbb{R}^d)$ un espacio métrico completo. De hecho, la convergencia en distancia de Wasserstein es equivalente a la convergencia débil-$*$ en medidas junto con la convergencia de los momentos de orden dos [53, Capítulo 7]. Si el ínfimo en la definición de la distancia se alcanza para una medida de probabilidad conjunta γ_o, dicha medida es llamada plan óptimo de transporte.

El problema de calcular la distancia de Wasserstein entre dos medidas de probabilidad está íntimamente ligado al problema clásico de transporte óptimo de masa de Monge. Para describir mínimamente dicho problema necesitamos

introducir una notación que será útil en lo que sigue. Diremos que una aplicación medible $T : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ transporta μ en ν, y lo denotaremos por $\nu = T\#\mu$, si tienen la misma ley, es decir si:

$$\nu[K] := \mu[T^{-1}(K)]$$

para cualquier conjunto medible $K \subset \mathbb{R}^d$, o equivalentemente (en formulación débil),

$$\int_{\mathbb{R}^d} f \, d\nu = \int_{\mathbb{R}^d} (f \circ T) \, d\mu \tag{4}$$

para cualquier $f \in C_o(\mathbb{R}^d)$ (funciones continuas con límite cero en infinito).

El problema clásico de Monge consiste en buscar el ínfimo del coste euclídeo pero no mediante medidas conjuntas de probabilidad sino mediante aplicaciones T que transportan una medida en otra, es decir, consiste en resolver el problema variacional

$$I := \inf \left\{ \int_{\mathbb{R}^d} |x - T(x)|^2 \, d\mu(x); \; T \text{ medible tal que } \nu = T\#\mu \right\}^{1/2}.$$

Obsérvese que el problema variacional en la definición de la distancia de Wasserstein (3) generaliza al anterior tomando como medidas de probabilidad en el espacio producto $\gamma_T = (1_{\mathbb{R}^d} \times T)\#\mu$, donde $1_{\mathbb{R}^d}$ denota la identidad en $\mathbb{R}^d$, para cada T que transporta μ en ν. Por tanto, es evidente que $d_2(\mu, \nu) \leq I$. Si el ínfimo en la definición de I se alcanza para una aplicación medible T, dicha aplicación es llamada aplicación óptima de transporte.

De hecho, bajo determinadas hipótesis ambos problemas tienen la misma solución. El teorema de Y. Brénier [11, 12, 39] nos demuestra que dadas $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ existe una función convexa $\varphi(x)$ en $\mathbb{R}^d$ cuyo gradiente transporta de forma óptima μ en ν, es decir, se tiene que $\nu = \nabla\varphi\#\mu$ y que para $T = \nabla\varphi$ se alcanza ambos el ínfimo de I y el ínfimo en la definición (2). Como consecuencia tenemos una expresión "simple" del plan óptimo de transporte $\gamma_o = (1_{\mathbb{R}^d} \times \nabla\varphi)\#\mu$ y de la distancia de Wasserstein

$$d_2(\mu, \nu) = \left(\int_{\mathbb{R}^d} |x - \nabla\varphi(x)|^2 \, d\mu(x) \right)^{1/2} \tag{5}$$

en este caso. Dicha aplicación óptima es única salvo conjuntos de medida cero respecto de μ.

La distancia de Wasserstein se puede generalizar a otros costes distintos del euclídeo para definir distancias de la forma

$$d_p(\mu, \nu) = \inf \Big\{ E|X - Y|^p; \; \text{ley}(X) = \mu, \; \text{ley}(Y) = \nu \Big\}^{1/p}, \tag{6}$$

con $1 \leq p < \infty$. El espacio de las medidas de probabilidad con momentos de orden p finitos $\mathcal{P}_p(\mathbb{R}^d)$ dotado de la distancia d_p es de nuevo un espacio métrico completo. De nuevo el ínfimo se alcanza sobre aplicaciones óptimas, es decir,

el problema variacional de Monge y el anterior problema variacional tienen la misma solución. Sin embargo, la aplicación óptima no tiene una expresión tan sencilla como en los teoremas de representación de Brénier [33, 53].

Por último y para terminar esta pequeña recolección de resultados sobre transporte óptimo, destacar que las anteriores distancias toman una expresión bastante simple en el caso uno dimensional. Observemos que de (4) se deduce que si T transporta μ en ν en una dimensión se tiene

$$\int_{-\infty}^{x} d\mu = \int_{-\infty}^{T(x)} d\nu$$

para todo $x \in \mathbb{R}$. Por tanto, cabe esperar poder tener una expresión de la aplicación optimal entre dos medidas en términos de su funciones de distribución.

Sea μ una medida de probabilidad en $\mathbb{R}$ y $F(x) = \mu((-\infty, x])$ su función de distribución. Consideremos la *inversa generalizada* de F dada por $F^{-1}(\eta) = \inf\{x \in \mathbb{R}/F(x) > \eta\}$. Ambas funciones son continuas por la derecha y no decrecientes en sus respectivos dominios. Sean μ, $\nu \in \mathcal{P}_p(\mathbb{R})$ y $F(x)$, $G(x)$ sus correspondientes funciones de distribución. Se demuestra (véase [53, Teorema 2.18]) que el valor de $d_p(\mu, \nu)$ viene dado por

$$d_p(\mu, \nu) = \left(\int_0^1 |F^{-1}(\eta) - G^{-1}(\eta)|^p d\eta \right)^{1/p}, \quad \forall p \in [1, +\infty), \tag{7}$$

es decir, la distancia d_p en una dimensión no es más que la distancia L^p entre las inversas de las funciones de distribución. De hecho si μ es absolutamente continua respecto a la medida de Lebesgue, entonces la aplicación óptima viene determinada por $T = G^{-1} \circ F$.

La distancia $d_\infty(\mu, \nu)$ se puede definir en términos de la sucesión de distancias $d_p(\mu, \nu)$. De hecho, es fácil observar que la sucesión $d_p(\mu, \nu)$ es creciente en el exponente p. Por tanto, se puede definir la distancia d_∞ entre dos medidas con todos sus momentos finitos mediante

$$d_\infty(\mu, \nu) = \lim_{p \uparrow \infty} d_p(\mu, \nu).$$

Denotando por δ_{x_0} la medida de probabilidad delta de Dirac en el punto x_0, se tiene que $d_p(\delta_{x_0}, \delta_{x_1}) = |x_0 - x_1|$ para cualesquiera $x_0, x_1 \in \mathbb{R}^d$, $1 \leq p \leq \infty$.

Referimos al reciente tratado de C. Villani [53] sobre tranporte óptimo de masa para el desarrollo de los contenidos de esta sección y referencias históricas sobre el problema clásico de Monge.

3 ¿Qué tiene que ver esto con EDPs?

Consideremos el caso particular del conjunto de EDPs (1) en el que tomamos $W = 0$ y $U = 0$, es decir,

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V). \tag{8}$$

Por tanto, estamos estudiando la ecuación de continuidad para la evolución de una densidad de probalidad con campo de velocidades dado por $u = -\nabla V$ donde $V : \mathbb{R}^d \longrightarrow \mathbb{R}$ es una función C^2 estrictamente convexa, coerciva (los conjuntos de subnivel son acotados) y acotada inferiormente y por tanto, supongamos sin pérdida de generalidad que alcanza su único mínimo global en 0.

Es fácil dar sentido a una solución de (8) con datos iniciales medidas de probabilidad. Dada $\rho \in C([0,T], \mathcal{P}(\mathbb{R}^d))$ donde $\mathcal{P}(\mathbb{R}^d)$ es el conjunto de todas las medidas de probabilidad $\mathbb{R}^d$ y la continuidad se considera respecto de la topología débil-*, diremos que es una solución del problema de Cauchy para la ecuación (8) con dato inicial $\mu \in \mathcal{P}(\mathbb{R}^d)$ si para cada $\psi \in C_o^\infty([0,\infty) \times \mathbb{R}^d)$ (el conjunto de funciones infinitamente derivables de soporte compacto en $[0,\infty) \times \mathbb{R}^d$) se verifica que:

$$\begin{aligned} &\int_0^T \int_{\mathbb{R}^d} \frac{\partial \psi}{\partial t}\, d\rho(t)\, dt + \int_{\mathbb{R}^d} \psi(0)\, d\mu \\ &= \int_0^T \int_{\mathbb{R}^d} (\nabla \psi \cdot \nabla V)\, d\rho(t)\, dt + \int_{\mathbb{R}^d} \psi(T)\, d\rho(T). \end{aligned} \tag{9}$$

Consideremos $\Phi_{s,t}(x)$ la solución general del flujo gradiente autónomo finito dimensional:

$$\begin{cases} \dfrac{\partial r}{\partial \tau} = -\nabla V(r) \quad \text{en} \quad s < \tau < t, \\ r(s) = x \in \mathbb{R}^d \end{cases}$$

que forman una familia de difeomorfismos de $\mathbb{R}^d$ en sí mismo por las hipótesis sobre el potencial V. Denotaremos por $\Phi_t(x)$ al caso particular en el que $s = 0$ y sea $u(r) = -\nabla V(r)$ el campo de velocidades.

Uno puede comprobar directamente la unicidad de solución del problema de Cauchy para (8) simplemente mediante el método de dualidad. De hecho consideremos el problema de Cauchy

$$\begin{cases} \dfrac{\partial \psi}{\partial t} - (\nabla V \cdot \nabla \psi) = \dfrac{\partial \psi}{\partial t} + (u \cdot \nabla \psi) = 0 \quad \text{en} \quad t < T, x \in \mathbb{R}^d \\ \psi(T,x) = \varphi(x) \in C_o^\infty(\mathbb{R}^d) \end{cases}$$

que tiene por solución $\psi(t,x) = \varphi(\Phi_{t,T}(x))$. Es claro que tomando $\mu = 0$ y tomando por ψ la solución del problema anterior se deduce de (9) que $\rho(T) = 0$ para cada $T > 0$.

Es más de nuevo tomando $\psi(t,x) = \varphi(\Phi_{t,T}(x))$ en la definición de solución (9) se deduce que

$$\int_{\mathbb{R}^d} \varphi(\Phi_T(x))\, d\mu = \int_{\mathbb{R}^d} \varphi\, d\rho(T)$$

para $\varphi(x) \in C_o^\infty(\mathbb{R}^d)$, y por tanto, usando la definición (4) la única solución del problema de Cauchy para (8) con dato inicial $\mu \in \mathcal{P}(\mathbb{R}^d)$ viene dada por $\rho(t) = \Phi_t \# \mu$.

Obsérvese que la solución para el dato inicial $\mu = \delta_{x_1}$ corresponde a $\rho(t) = \delta_{x_1(t)}$ donde $x_1(t)$ es la solución del flujo gradiente finito dimensional:

$$\begin{cases} \dfrac{\partial x_1}{\partial t} = -\nabla V(x_1(t)) \quad \text{en} \quad t > 0 \\ x_1(0) = x_1 \in \mathbb{R}^d \end{cases}$$

y que se tiene la solución estacionaria $\rho_\infty = \delta_0$, ya que 0 es el mínimo del potencial V.

Parece intuitivo que las propiedades de convexidad del potencial $V(x)$ determinen el orden de convergencia hacia el equilibrio ρ_∞ de las soluciones de la ecuación (8). De hecho es fácil demostrar:

Proposición 1 (Caricatura de los órdenes de convergencia) *Sean $k \in \mathbb{R}^+$ y $V \in C^2(\mathbb{R}^d)$ tal que $D^2V(x) \geq kI$ en $\mathbb{R}^d$. Dadas dos soluciones $x_1(t)$ y $x_2(t)$ de $\frac{\partial r}{\partial \tau} = -\nabla V(r)$ se tiene que*

$$d_2(\delta_{x_1(t)}, \delta_{x_2(t)}) \leq e^{-kt} d_2(\delta_{x_1(0)}, \delta_{x_2(0)}).$$

Demostración. Sea $f(t) = |x_1(t) - x_2(t)|^2/2$. Entonces

$$\begin{aligned} f'(t) &= -(x_1(t) - x_2(t)) \cdot (\nabla V(x_1(t)) - \nabla V(x_2(t))) \\ &= -(x_1(t) - x_2(t)) \cdot \left(\int_0^1 D^2V[(1-s)x_1(t) + sx_2(t)]\ (x_1(t) - x_2(t)) ds \right) \\ &\leq -2kf(t) \int_0^1 ds. \end{aligned}$$

Integrando en t y teniendo en cuenta que $d_2(\delta_{x_1}, \delta_{x_2}) = |x_1 - x_2|$ se tiene el resultado: $f(t) \leq e^{-2kt} f(0)$. □

Es posible demostrar un teorema de convergencia hacia equilibrio para datos iniciales generales en $\mathcal{P}_p(\mathbb{R}^d)$.

Teorema 2 (Comportamiento asintótico $W = U = 0$) *Sean $k \in \mathbb{R}^+$ y $V \in C^2(\mathbb{R}^d)$ tal que $D^2V(x) \geq kI$ en $\mathbb{R}^d$ y $D^2V(x) \leq C(1 + |x|^{p-2})I$ con $p \geq 2$. Dadas dos soluciones $\rho_1(t)$ y $\rho_2(t)$ de (8) con datos iniciales en $\mathcal{P}_p(\mathbb{R}^d)$ se tiene que*

$$d_2(\rho_1(t), \rho_2(t)) \leq e^{-kt} d_2(\rho_1(0), \rho_2(0)).$$

Como consecuencia, dada cualquier solución $\rho_1(t)$ se verifica

$$d_2(\rho_1(t), \rho_\infty) = d_2(\rho_1(t), \delta_0) \leq e^{-kt} d_2(\rho_1(0), \delta_0).$$

Demostración. Consideremos γ_o el plan óptimo de transferencia entre $\rho_1(0)$ y $\rho_2(0)$. Las soluciones $\rho_1(t)$ y $\rho_2(t)$ vienen dadas por $\rho_1(t) = \Phi_t \# \rho_1(0)$

y $\rho_2(t) = \Phi_t \# \rho_2(0)$. Definiendo $\gamma_t = (\Phi_t \times \Phi_t)\#\gamma_o$ tenemos un plan de transferencia entre $\rho_1(t)$ y $\rho_2(t)$ y por tanto

$$d_2^2(\rho_1(t), \rho_2(t)) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \, d\gamma_t(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} |\Phi_t(x) - \Phi_t(y)|^2 \, d\gamma_o(x,y).$$

Como consecuencia, suponiendo que podemos intercambiar el signo integral y la derivada en t se tiene que

$$\left.\frac{d}{dt}\right|_0 d_2^2(\rho_1(t), \rho_2(t))/2 = -\int_{\mathbb{R}^d \times \mathbb{R}^d} (x-y) \cdot (\nabla V(x) - \nabla V(y)) \, d\gamma_o(x,y)$$

y por tanto usando la hipótesis de convexidad uniforme de V tenemos

$$\left.\frac{d}{dt}\right|_0 d_2^2(\rho_1(t), \rho_2(t)) \leq -2k \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \, d\gamma_o(x,y) = -2k \, d_2^2(\rho_1(0), \rho_2(0)).$$

Justifiquemos el intercambio del signo integral con la derivada en t. Primero observemos que la derivada en t del integrando se puede acotar como

$$\begin{aligned}\frac{d}{dt}|\Phi_t(x) - \Phi_t(y)|^2 &= |(\Phi_t(x) - \Phi_t(y)) \cdot (\nabla V(\Phi_t(x)) - \nabla V(\Phi_t(y)))| \\ &\leq C(1 + (|\Phi_t(x)| + |\Phi_t(y)|)^{p-2})|\Phi_t(x) - \Phi_t(y)|^2\end{aligned}$$

por hipótesis sobre V. Como $p \geq 2$ y $|\Phi_t(x)| \leq |x|$ por la convexidad uniforme en V y ser 0 su mínimo global, se deduce que

$$|(\Phi_t(x) - \Phi_t(y)) \cdot (\nabla V(\Phi_t(x)) - \nabla V(\Phi_t(y)))| \leq C(1 + (|x| + |y|)^{p-2})(|x| + |y|)^2$$

cuyo término derecho es integrable con respecto al plan optimal γ_o teniendo en cuenta que las medidas iniciales tienen momentos de orden p acotados. Por tanto, el teorema de la convergencia dominada de Lebesgue y sus consecuencias sobre teoremas de derivación de integrales respecto de parámetros nos permiten el intercambio de la derivada y la integral.

Por último, constatemos que las soluciones se mantienen en $\mathcal{P}_p(\mathbb{R}^d)$ si sus datos iniciales pertenecen a dicho espacio. Esta afirmación se deduce trivialmente de $\rho_1(t) = \Phi_t \# \rho_1(0)$ y $|\Phi_t(x)| \leq |x|$ para $t \geq 0$. Por tanto, el argumento anterior se puede hacer en cada instante de tiempo y se deduce que

$$\frac{d}{dt} d_2^2(\rho_1(t), \rho_2(t)) \leq -2k \, d_2^2(\rho_1(t), \rho_2(t))$$

para cada $t \geq 0$ y concluimos la tesis del teorema.

La última consecuencia del teorema se deduce tomando como una de las soluciones la solución estacionaria $\rho_\infty = \delta_0$. □

En resumen, hemos visto cómo las propiedades de contracción de la distancia de Wasserstein determinan el comportamiento asintótico de las soluciones de la ecuación (8). Por otro lado, consideremos el "semigrupo" generado por la

ecuación (8), es decir, $X_t : \mathcal{P}_p(\mathbb{R}^d) \longrightarrow \mathcal{P}_p(\mathbb{R}^d)$, $p \geq 2$ definida por el hecho de que $X_t(\mu)$ es la única solución del problema de Cauchy para (8) con dato inicial μ. Acabamos de demostrar que X_t es una contracción global en el espacio métrico completo $\mathcal{P}_p(\mathbb{R}^d)$, y por tanto, el estado estacionario $\rho_\infty = \delta_0$ está caracterizado por ser el único punto fijo de dichas aplicaciones para todo $t \geq 0$.

4 Flujos gradientes en espacios de medidas

La estructura de flujo gradiente respecto a un potencial convexo que vimos en la sección anterior, permitió estimar el orden de convergencia en el flujo gradiente finito dimensional, y posteriormente, poder demostrar estimaciones del orden de convergencia en ecuaciones de continuidad cuyas características estaban relacionadas con dicho flujo gradiente finito dimensional. La comparación va mucho más allá, ya que F. Otto nos mostró en [42], que podemos dotar a todas las ecuaciones de la forma (1) de una estructura Riemaniana formal, donde dichas ecuaciones son flujos gradiente infinito dimensionales respecto a funcionales de Liapunov naturales de dichas ecuaciones.

Consideremos el Problema de Cauchy para la ecuación de calor con datos iniciales medidas de probabilidad, es decir,

$$\begin{cases} \dfrac{\partial \rho}{\partial t} = \Delta \rho \qquad \text{en} \quad t > 0, x \in \mathbb{R}^d \\ \rho(0, x) = \mu \end{cases}$$

el cual sabemos que tiene una única solución en dicha clase que conserva la masa y que viene dada por la fórmula de Poisson $\rho(t) = K(t, x) * \mu$ donde $K(t, x)$ es el núcleo del calor. Estructuras clásicas de flujos gradiente para esta ecuación han sido introducidos en el espacio $H^1(\mathbb{R}^d)$ [41]. Sin embargo, veamos que esta ecuación tiene una estructura de flujo gradiente más natural, lo que nos permite esperar propiedades de contracción para las distancias de Wasserstein.

Consideremos el funcional de entropía típico en mecánica estadística y en teoría de la información dado por

$$H(\rho) = \int_{\mathbb{R}^d} \rho \log \rho \, dx$$

definido sobre el espacio de las medidas de probabilidad con momento de orden dos acotado absolutamente continuas respecto Lebesgue $M = \mathcal{P}_2^{ac}(\mathbb{R}^d)$, es decir, funciones integrables positivas con integral unitaria y momento de orden dos finito tales que $H(\rho) < \infty$. Consideremos variaciones en dicho espacio, es decir, tomemos el espacio tangente en ρ definido por

$$T_\rho M = \{v \in L^1(\mathbb{R}^d) \text{ con media cero}\}.$$

Es más tomaremos una representación del espacio tangente en términos de funciones $\psi \in \mathcal{H} = W^{1,2}_\rho := W^{1,2}(\mathbb{R}^d, d\rho)$, es decir, la clausura de $C^\infty_o(\mathbb{R}^d)$ con

respecto a la métrica

$$\langle \psi, \psi \rangle_\rho = \int_{\mathbb{R}^d} |\nabla \psi|^2 d\rho(x).$$

Dicha representación la haremos mediante la resolución de la ecuación elíptica

$$-\text{div}\,(\rho \nabla \psi) = v.$$

De forma que la métrica en el espacio tangente $T_\rho M$ entre dos vectores cualesquiera viene definida por

$$\langle v_1, v_2 \rangle_\rho := \langle \psi_1, \psi_2 \rangle_\rho = \int_{\mathbb{R}^d} \nabla \psi_1 \cdot \nabla \psi_2 \, d\rho(x) = \int_{\mathbb{R}^d} \psi_1 v_2 \, dx$$

donde ψ_i es la representación de v_i, $i = 1, 2$.

Formalmente es fácil deducir que

$$DH_\rho(v) := \lim_{\epsilon \to 0} \frac{H(\rho + \epsilon v) - H(\rho)}{\epsilon} = \int_{\mathbb{R}^d} \frac{\delta H}{\delta \rho} v \, dx = \int_{\mathbb{R}^d} \nabla \frac{\delta H}{\delta \rho} \cdot \nabla \psi \, d\rho,$$

con $\frac{\delta H}{\delta \rho} = \log \rho$. De hecho, si $\frac{\delta H}{\delta \rho} \in H$ y se nos permite poder escribir que $DH_\rho(v) = \langle \frac{\delta H}{\delta \rho}, \psi \rangle_\rho$, se deduce que el gradiente formal del funcional de entropía en ρ viene determinado por

$$\nabla H_\rho = -\text{div} \left(\rho \nabla \frac{\delta H}{\delta \rho} \right).$$

De esta forma, podemos ver la ecuación del calor como el flujo gradiente infinito dimensional asociado a $H(\rho)$ en el espacio M, ya que

$$\frac{\partial \rho}{\partial t} = -\nabla H_\rho = \text{div} \left(\rho \nabla \frac{\delta H}{\delta \rho} \right) = \Delta \rho.$$

La estructura anterior fue introducida formalmente por F. Otto en su estudio de la ecuaciones de medios porosos [42], y después generalizada a todas las ecuaciones de la forma (1) en [20]. De hecho, todas las ecuaciones (1) se pueden escribir como flujos gradiente formales:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\rho \nabla \frac{\delta H}{\delta \rho} \right),$$

con respecto a los funcionales de entropía:

$$H(\rho) = \int_{\mathbb{R}^d} U(\rho) \, dx + \int_{\mathbb{R}^d} V(x) \, d\rho(x) + \frac{1}{2} \int_{\mathbb{R}^d \times \mathbb{R}^d} W(x - y) \, d\rho(x) \, d\rho(y), \quad (10)$$

donde $\frac{\delta H}{\delta \rho} = U'(\rho) + V + W * \rho$.

Notemos que estructuras gradiente en espacios infinito dimensionales se encuentran en la literatura, donde las cartas que definen la variedad Riemaniana infinito dimensional son a la postre difeomorfimos con abiertos de espacios de

Hilbert. Esta estructura no se puede aplicar en nuestro caso pues las cartas locales [21] hacen localmente equivalente M a subconjuntos de H que no son abiertos con la topología de H. Para poder hacer rigurosa la estructura anterior introducimos un espacio de longitudes "riemaniano" que nos permite generalizar los conceptos base de geodésicas y subgradientes a espacios de longitudes [35, 14]. Referimos a [21] para más detalles.

Hasta ahora, se ha introducido una estructura de flujo gradiente infinito dimensional para (1). Para demostrar su utilidad uno puede usar dicha estructura para obtener el comportamiento asintótico de ecuaciones del tipo (1) bajo hipótesis adicionales. Para lograrlo, se deben aclarar dos cuestiones:

- ¿Cuál es la distancia inducida en M por dicha estructura de flujo gradiente? dicha distancia de tener una expresión "sencilla" es la candidata clara para estimar el comportamiento asintótico de las soluciones de (1) vistos los resultados de la sección anterior.
- ¿Cuál es la noción de convexidad sobre el funcional $H(\rho)$ (10) que determina el comportamiento asintótico de las soluciones de (1)?

La respuesta a la primera pregunta fue dada por F. Otto en [42], el cual fue el primero en observar que las ecuaciones de Euler-Lagrange para las geodésicas en M inducidas por $\langle , \rangle_\rho$ son las mismas que las ecuaciones satisfechas por las geodésicas de la distancia de Wasserstein d_2 previamente estudiadas e introducidas en la tesis doctoral de R. J. McCann [40]. Como consecuencia la distancia inducida por $\langle , \rangle_\rho$ en M es la distancia de Wasserstein d_2.

La respuesta a la segunda pregunta fue dada en parte en [40] donde se introduce el concepto de convexidad por desplazamientos para un funcional. Dicho concepto no es más que convexidad de $H(\rho_s)$ en el parámetro s para cualquier geodésica. Este concepto por sí mismo no es suficiente para estimar todas las distintas posibilidades en el comportamiento asintótico de las soluciones de (1). El concepto de ϕ-convexidad uniforme introducido en [21] cuantifica la convexidad del funcional de entropía generalizado $H(\rho)$ en términos de la convexidad de V y W. Este concepto se desarrollará en mayor profundidad en la sección quinta.

Todos los conceptos introducidos en esta sección han sido recientemente introducidos de forma rigurosa en [21] y por tanto, podemos hablar de la estructura de flujo gradiente para esta conjunto de ecuaciones respecto de la distancia de Wasserstein d_2 con total propiedad. Referimos al lector ansioso de demostraciones a dicho trabajo y al libro de próxima aparición [4].

5 Contractividad de distancias: EDPs de difusión no lineal

En esta sección supondremos siempre que $V : \mathbb{R}^d \longrightarrow \mathbb{R}$ es una función C^2. Por otro lado consideremos que $W = 0$ y que la función $U(\rho)$ sea un función convexa en $[0, \infty)$, tal que si $P(\rho) := U'(\rho)\rho - U(\rho)$, se cumple

$$P(\rho) \geq 0 \text{ es creciente y } \frac{P(\rho)}{\rho^{1-1/d}} \text{ es no decreciente en } \rho \in (0, \infty). \qquad (11)$$

La ecuación (1) se reduce en esta caso a la ecuación de Fokker-Planck no lineal:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \Delta P(\rho), \tag{12}$$

de la cual las ecuaciones de medios porosos, difusión rápida y la ecuación del calor en variables autosemejantes son casos particulares con las elecciones $P(\rho) = \rho^m$, $m > \frac{d-2}{d}$ y $V(x) = \frac{|x|^2}{2}$. Más concretamente, es fácil comprobar en estos casos, que el cambio de variable:

$$\rho(x,t) = e^{dt} u(e^t x, k(e^{t/k} - 1))$$

donde $k = (d(m-1)+2)^{-1}$ hace equivalentes el problema de Cauchy para (12) y el problema de Cauchy para la ecuación

$$\frac{\partial u}{\partial t} = \Delta u^m. \tag{13}$$

Además este cambio de variables traduce una traslación temporal de la solución autosemejante de Barenblatt para (13) en la solución estacionaria para (12). Por tanto, mediante este "inocente" cambio de variable reducimos el estudio de la convergencia hacia estados autosemejantes para (13) en el estudio de la convergencia hacia equilibrio para (12). Nótese que las ecuaciones de difusión no lineal general:

$$\frac{\partial u}{\partial t} = \Delta P(u) \tag{14}$$

no están relacionadas en principio con las soluciones de (12) debido a la falta de homogeneidad de la no linealidad. Recordemos que estamos estudiando sólo soluciones de estas ecuaciones con dato inicial funciones integrables positivas con momento de orden dos acotado e integral unitaria. Los problemas de Cauchy en este espacio para estas ecuaciones están bien planteados y referimos a los trabajos de J. L. Vázquez [50, 51, 52] como una estupenda fuente de información al respecto. Recordar que la restricción $m > (d-2)/d$ hace que la solución autosemejante de Barenblatt sea integrable. Si además queremos que dicha solución tenga momentos de orden dos acotados necesitamos que $m \geq d/(d+2)$.

Además, (11) implica la restricción $m > (d-1)/d$. Los resultados que mostraremos a continuación son válidos pues para las ecuaciones (13) con $m > \min((d-1)/d, d/(d+2))$ de forma que la solución autosemejante de Barenblatt para $t > 0$ pertenezca a la clase de datos iniciales que estamos tomando.

Comencemos estudiando el comportamiento asintótico en una dimensión. Sean u_1 y u_2 dos soluciones cualesquiera para la ecuación (14). Supongamos que son lo suficientemente regulares como para que todas las operaciones que se hagan en lo sucesivo tengan sentido; de nuevo lectores ávidos de rigor deberían dirigirse a las referencias [18, 25, 16].

Usemos la expresión de las distancias de Wasserstein en 1d para calcular la evolución de la distancia entre $u_1(t)$ y $u_2(t)$. Consideremos $F_1(t)$ y $F_2(t)$ las

funciones de distribución asociadas a dichas soluciones cuyas funciones pseudo-inversas $F_1^{-1}(t)$ y $F_2^{-1}(t)$ verifican la ecuación

$$\frac{\partial F^{-1}(\eta,t)}{\partial t} = -\frac{\partial}{\partial \eta}\left[P\left(\left(\frac{\partial F^{-1}}{\partial \eta}\right)^{-1}\right)\right]. \tag{15}$$

Por tanto, estudiemos la evolución de la norma L^p, $p \geq 2$, de la diferencia de ambas funciones pseudo-inversas de distribución que puede ser escrita después de integraciones por partes como

$$\begin{aligned}&\frac{1}{p(p-1)}\frac{d}{dt}\int_0^1 |F_1^{-1} - F_2^{-1}|^p d\eta \\ &\quad = \int_0^1 |F_1^{-1} - F_2^{-1}|^{p-2}(F_{1,\eta}^{-1} - F_{2,\eta}^{-1})\left[P\left(\left(F_{1,\eta}^{-1}\right)^{-1}\right) - P\left(\left(F_{2,\eta}^{-1}\right)^{-1}\right)\right] d\eta\end{aligned}$$

cuyo lado derecho es no positivo ya que la nolinealidad $P(\rho)$ es creciente y las funciones $F_1^{-1}(t)$ y $F_2^{-1}(t)$ son no decrecientes. Por tanto, se verifica que

$$d_p(u_1(t), u_2(t)) \leq d_p(u_1(0), u_2(0))$$

para $t \geq 0$, $2 \leq p \leq \infty$, supuesto que el lado derecho es finito, o bien que los datos iniciales tienen los momentos iniciales necesarios acotados. En otras palabras, todas las ecuaciones (14) son contracciones (no estrictas) para las distancias de Wasserstein d_p, $2 \leq p \leq \infty$, en una dimensión. Por supuesto, aquí estamos obviando el hecho de que las funciones pseudoinversas de distribución pueden no ser derivables en η pues la solución puede tener varias componentes conexas en su soporte (véase figura 2). La demostración rigurosa de este resultado se lleva a cabo aproximando el problema de Cauchy anterior por un problema de Cauchy en un intervalo con condiciones de Dirichlet donde la solución es estrictamente positiva en el interior (véanse [18, 50] para los detalles).

Es conocido, que la ecuación de medios porosos (14), $m > 1$, tiene velocidad de propagación finita, es decir, soluciones que inicialmente son de soporte compacto lo continuan siendo para todos los tiempos, y su soporte crece en tiempo con una velocidad que como mucho alcanza la velocidad de la solución autosemejante de Barenblatt con la misma masa. Este resultado se demostró con técnicas de principio del máximo y comparación de soluciones con las soluciones autosemejantes (véanse las referencias en [52]). En las figuras 1 y 2, podemos observar la evolución de un dato inicial de soporte compacto para la ecuación (14), $m = 2$, en 1 y 2 dimensiones espaciales (el esquema numérico usado es un simple Euler implícito con diferencias finitas centradas para las segundas derivadas resuelto mediante Newton-Raphson, agradezco a M. P. Gualdani por las figuras). En ellas se observa claramente la propiedad de propagación finita (amén de la concavidad asintótica [52]) de las soluciones.

Veamos que la contracción de las distancias de Wasserstein en una dimensión implica esta propiedad cualitativa de las soluciones de (14):

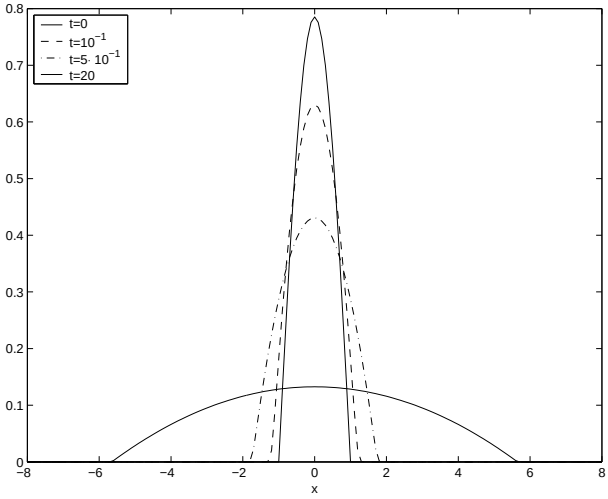


Figura 1: Evolución de la ecuación de medios porosos con dato inicial $\frac{\pi}{4}\cos(\frac{\pi}{2}x)$ con $x \in [-1,1]$ para $m=2$.

Teorema 3 (Estimación de la velocidad de propagación en 1d) *[18] Sean* $u_1(x,t)$, $u_2(x,t)$ *soluciones fuertes [52] de* (13), $d=1$ $m>1$, *con condiciones iniciales* $u_{01}(x)$ *y* $u_{02}(x)$ *respectivamente, donde* $u_{0i} \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$, $\|u_{0i}\|_{L^1(\mathbb{R})}=1$, $u_{0i} \geq 0$ *y* u_{0i} *son de soporte compacto,* $i=1,2$. *Sea*

$$\Omega_i(t) = \{x \in \mathbb{R} / u_i(x,t) > 0\}, \quad i=1,2,$$

y $\xi_i(t) = \inf[\Omega_i(t)]$, $\Xi_i(t) = \sup[\Omega_i(t)]$, *para* $t \geq 0$, $i=1,2$. *Entonces*

$$\text{máx}\left\{|\xi_1(t)-\xi_2(t)|, |\Xi_1(t)-\Xi_2(t)|\right\} \leq d_\infty(u_{01},u_{02}), \quad \forall t \in [0,+\infty). \tag{16}$$

Dicho resultado, nos permite obtener una estimación de la velocidad de propagación del soporte, tomando como una de las soluciones en (16) una solución autosemejante de masa unidad, y por tanto, la velocidad de crecimiento del soporte de cualquier solución con las anteriores condiciones en los datos iniciales está limitada por la velocidad de propagación explícita de la Barenblatt cuyo soporte crece como $t^{1/m+1}$. En los trabajos [15, 25] se desarrollan más aplicaciones de las contracciones de distancias en el caso uno dimensional a través de la expresión simplificada de d_2.

Problema abierto 4 (Contracción de las distancias en $d \geq 2$) *¿Son las distancias* d_p, $p \geq 2$, *contracciones para las ecuaciones* (13) *en dimensión* $d \geq 2$*? De serlas darían una buena forma de atacar resultados sobre estimaciones de la distancia entre los soportes de las soluciones de la ecuación en* $d \geq 2$ *las cuales son sólo conocidas de forma óptima en el caso radial.*

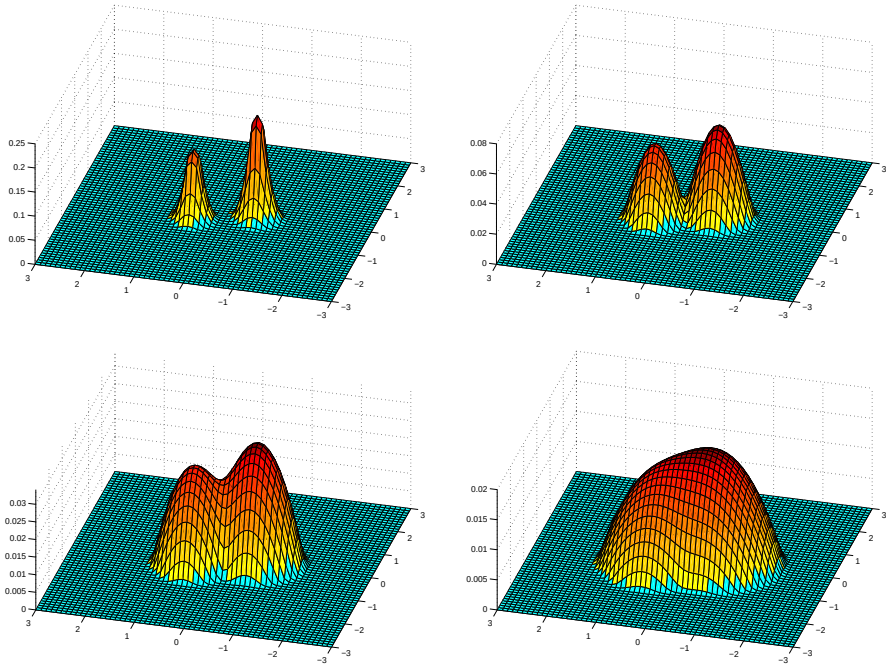


Figura 2: Evolución de la ecuación de medios porosos para $m = 2$.

Para reafirmar la conjetura del problema abierto planteado demostremos que es el caso para la ecuación del calor en cualquier dimensión (esta observación se la debo a R. J. McCann al cual estoy agradecido).

Proposición 5 (Contracción para la ecuación del calor) *Sean* $u_1(x,t)$,$u_2(x,t)$ *soluciones de la ecuación del calor con condiciones iniciales* $u_{01}(x)$ *y* $u_{02}(x)$ *respectivamente, donde* $u_{0i} \in L^1(\mathbb{R}^d)$, $\|u_{0i}\|_{L^1(\mathbb{R}^d)} = 1$, $u_{0i} \geq 0$, *y con momentos de orden* $p \geq 2$ *acotados, entonces*

$$d_p(u_1(t), u_2(t)) \leq d_p(u_1(0), u_2(0))$$

para $t \geq 0$, $2 \leq p \leq \infty$.

Demostración. Consideremos γ_o el plan óptimo de transferencia entre $u_1(0)$ y $u_2(0)$ para la distancia d_p, $2 \leq p \leq \infty$. Las soluciones $u_1(t)$ y $u_2(t)$ vienen dadas por $u_1(t) = K(t,x) * u_1(0)$ y $u_2(t) = K(t,x) * u_2(0)$. Definiendo γ_t como funcional actuando sobre funciones continuas de la forma:

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} g(x,y)\, d\gamma_t(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \int_{\mathbb{R}^d} g(x+z, y+z) K(t,z)\, dz\, d\gamma_o(x,y)$$

tenemos que γ_t es un plan de transferencia entre $u_1(t)$ y $u_2(t)$, y por tanto,

$$d_p^p(u_1(t), u_2(t)) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^p \, d\gamma_t(x,y)$$
$$= \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^p \, d\gamma_o(x,y) = d_p^p(u_1(0), u_2(0)),$$

para $2 \leq p < \infty$. La igualdad intermedia se debe a la definición de γ_t que implica que sobre funciones de la forma $g(x,y) = h(x-y)$ actúa como γ_o. Como d_p es monótona en p, el resultado se extiende a $p = \infty$. □

Nota 1 *El resultado anterior implica que la distancia d_∞ controla de alguna forma el comportamiento de las colas de las soluciones para la ecuación del calor. Cuantificar de forma analítica este control es un problema abierto.*

La respuesta a la conjetura 4 es sí en general para la distancia d_2. La demostración de este hecho está de alguna forma ímplicita en alguno de los cálculos en [42] pero en su forma más general se ha demostrado recientemente en [21]:

Teorema 6 (Contracción para difusión no lineal en d_2) *[21] Supongamos que el potencial V es semiconvexo, es decir, $D^2V(x) \geq kI$ para $x \in \mathbb{R}^d$, con $k \in \mathbb{R}$, y $P(\rho)$ satisfaciendo* (11)*, entonces dadas dos soluciones $\rho_1(t)$ y $\rho_2(t)$ de* (12) *con datos iniciales $\rho_i(0) \in L^1(\mathbb{R}^d)$, $\|\rho_i(0)\|_{L^1(\mathbb{R}^d)} = 1$, $\rho_i(0) \geq 0$, y con momentos de orden 2 acotados,*

$$d_2(\rho_1(t), \rho_2(t)) \leq e^{-kt} d_2(\rho_1(0), \rho_2(0))$$

para $t \geq 0$.

Nota 2

- *El teorema anterior se puede aplicar a las ecuaciones de difusión no lineal generales* (14) *con $V = 0$, $k = 0$, demostrando pues el no crecimiento de la distancia d_2.*
- *En el caso de que V sea un potencial confinante, es decir, $V : \mathbb{R}^d \longrightarrow \mathbb{R}$ sea una función C^2 uniformemente convexa, es decir, $D^2V(x) \geq kI$ con $k > 0$, se tiene que la distancia d_2 es una contracción estricta:*

  $$d_2(\rho_1(t), \rho_2(t)) \leq L(t)\, d_2(\rho_1(0), \rho_2(0))$$

  para $t \geq 0$, con $L(t) = e^{-kt} < 1$. Por tanto, esto permite obtener directamente el orden de convergencia exponencial hacia equilibrio sin más que tomar una de las soluciones como la solución de equilibrio. Es más, permite utilizar este hecho para demostrar la existencia de dicho equilibrio, para ello hay que extender la aplicación flujo de la solución de manera continua al espacio de medidas de orden dos para completar el espacio,

lo cual es trivial al ser una contracción, y aplicar el teorema del punto fijo de Banach. Posteriormente, se puede comprobar que de hecho este equilibrio es absolutamente continuo respecto de Lebesgue. Nótese que la caracterización del equilibrio en este caso tiene relativo interés pues se puede calcular de forma explícita.

- *Esta propiedad de contracción permite un estudio del comportamiento asintótico para la ecuación de difusión no lineal general* (14)*, sobre la cual no se conocen prácticamente resultados de comportamiento asintótico, debido precisamente a la no existencia de soluciones autosemejantes que permitan comparar las soluciones. Esta línea es un trabajo de investigación que actualmente estamos llevando a cabo [17] y que permite dar un sustituto de la solución autosemejante general en base a una curva de puntos fijos de una familia de aplicaciones asociadas al flujo de* (14)*.*

- *Resultados de decrecimiento de distancias d_p (que no de contracción) han sido demostrados por M. Agueh para el caso de ecuaciones de la forma:*

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \Delta_p P(\rho),$$

donde Δ_p es el operador p-Laplaciano [2, 3].

La demostración del resultado anterior está basada, por un lado en la teoría rigurosa de flujos gradientes apuntada en la sección anterior, junto con una noción de convexidad cuantificada para funcionales en el espacio de medidas, que desarrollaremos en la sección 6 en una situación mucho más general.

6 Contractividad de distancias: EDPs cinéticas en gases granulares

En primer lugar, cuando hablamos de modelos cinéticos tenemos que cambiar nuestra intuición sobre las ecuaciones (1). La incógnita $\rho(t, x)$ representa en estos modelos la densidad de probabilidad de encontrar partículas con una determinada velocidad $x \in \mathbb{R}^d$ en un determinado instante de tiempo $t \geq 0$. La ecuación (1) se suele denominar una ecuación cinética homogénea pues en ella se han eliminado los términos correspondientes a las variables espaciales y nos hemos quedado sólo en el espacio de velocidades. Hay que recordar que los modelos cinéticos suelen ser EDPs de evolución de densidades de probabilidad en el espacio de fases (r, x) donde $r \in \mathbb{R}^d$ representa una posición y $x \in \mathbb{R}^d$ representa una **velocidad**.

Un gas granular corresponde al estudio de la dinámica de un conjunto muy grande de esferas rígidas todas ellas homogéneas y de tamaño pequeño que cuando colisionan lo hacen de forma inelástica. La inelasticidad quiere decir que en cada colisión las partículas pierden parte de su energía de forma que la velocidad relativa después de la colisión es más pequeña que previamente a ella. En una dimensión se han desarrollado modelos de mecánica estadística [55]

para estudiar el comportamiento global de este conjunto de partículas que han dado lugar a modelos basados en ecuaciones cinéticas. En dimensión 2 y 3 se han considerado modelos basados en aproximaciones de las ecuaciones de Boltzmann-Enskog [26, 27] para partículas inelásticas y gases densos [10, 15]. Dichos modelos han sido utilizados por grupos de físicos experimentales en sus aplicaciones, referimos a [9] y las referencias allí incluidas.

Si se considera el caso particular de (1) en una dimensión con $V = 0$, $W = |x|^3/3$, $U = 0$ uno recupera el modelo simplificado uno dimensional para medios granulares introducido por D. Benedetto, E. Caglioti y M. Pulvirenti en [6]; tomando $V = x^2/2$, $W = |x|^3/3$, $U = \rho \log \rho$ o bien $V = 0$, $W = |x|^3/3$, $U = \rho \log \rho$ uno recupera los modelos simplificados uno dimensionales para medios granulares en un baño térmico introducidos y estudiados por el método de entropía relativa en [7]. Modelos uno dimensionales más generales fueron introducidos por G. Toscani [48], cuyo comportamiento asintótico usando la distancia d_2 en términos de las inversas de las funciones de distribución fue desarrollado en [38].

Estos dos modelos son casos particulares de la estrategia apuntada en la sección anterior, siendo la gran diferencia que típicamente en los modelos en medios granulares W no es uniformemente convexo. Por tanto, es el momento de introducir la nueva noción de convexidad que permite el estudio de los órdenes de convergencia en caso de que la convexidad uniforme falle.

Definición 1 *Dadas dos medidas de probabilidad $\rho_0, \rho_1 \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ diremos que la curva $\rho_s : [0,1] \longrightarrow \mathcal{P}_2^{ac}(\mathbb{R}^d)$ es una geodésica para la distancia d_2 si $d_2(\rho_s, \rho_{s+t}) = t\, d_2(\rho_0, \rho_1)$ para $0 \leq s \leq s+t \leq 1$.*

El teorema de Y. Brénier [11, 12, 39] nos demuestra que dadas $\rho_0, \rho_1 \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ existe una función convexa $\varphi(x)$ en $\mathbb{R}^d$ con $\rho_1 = \nabla\varphi \# \rho_0$ para la que la distancia d_2 se alcanza (véase la segunda sección). A partir de la aplicación óptima, R. J. McCann [40] introdujo las curvas

$$\rho_s = [(1-s)1_{\mathbb{R}^d} + s\nabla\varphi]\#\rho_0$$

con $0 \leq s \leq 1$, que se demuestran son geodésicas para la distancia d_2 (basta utilizar la parte de unicidad del teorema de Brénier). Las geodésicas satisfacen de forma débil la ecuación de continuidad con campo de velocidad $\nabla\varphi(x) - x$, donde las características son rectas con pendiente fijada por el campo de velocidad anterior.

Definición 2 *Diremos que un funcional $H : \mathcal{P}_2^{ac}(\mathbb{R}^d) \longrightarrow \mathbb{R}$ es convexo por desplazamientos, si para cualquier par de medidas de probabilidad $\rho_0, \rho_1 \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$, la función $H(\rho_s)$ es una función convexa respecto de $s \in [0,1]$, donde ρ_s es una geodésica que une $\rho_0, \rho_1 \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$.*

R. J. McCann [40] demostró en su tesis doctoral, que todos los fucionales de la forma (10) bajo las hipótesis de convexidad para V y W donde $P(\rho)$ verifica (11), son convexos por desplazamientos. Ahora bien, como hemos dicho

varias veces en este artículo necesitamos cuantificar el tipo de convexidad de los funcionales. El término *módulo de convexidad* se refiere a cualquier función ϕ satisfaciendo:

(ϕ_0) $\quad \phi : [0,\infty) \longrightarrow \mathbb{R}$ es continua, positiva y anulándose sólo en $\phi(0) = 0$;

(ϕ_1) $\quad \chi_s(x) := \dfrac{1}{2}\displaystyle\int_{|1-2s|\sqrt{x}}^{\sqrt{x}} \phi(t)\,dt$ es convexa en $x \geq 0$ para cada $s \in [0,1]$.

Ejemplos típicos de dichas funciones son $\phi(s) = ks^{q-1} \geq 0$ con $q \geq 2$. Es importante notar que si ϕ es convexa y satisface ϕ_0, entonces satisface ϕ_1.

Definición 3 *Diremos que un funcional $H : \mathcal{P}_2^{ac}(\mathbb{R}^d) \longrightarrow \mathbb{R}$ es ϕ-uniformemente convexo si para cualquier par de medidas de probabilidad $\rho_0, \rho_1 \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$, la función $H(\rho_s)$ es una función convexa respecto de $s \in [0,1]$ donde ρ_s es una geodésica que une $\rho_0, \rho_1 \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ y además:*

$$H(\rho_0) - H(\rho_s) - H(\rho_{1-s}) + H(\rho_1) \geq \frac{1}{2}\int_{|1-2s|L}^{L} \phi(t)\,dt \quad , 0 \leq s \leq 1 \ ,$$

donde $L = d_2(\rho_0, \rho_1)$. Diremos que una función $V : \mathbb{R}^d \longrightarrow \mathbb{R}$ es ϕ-uniformemente convexa si verifica lo anterior sobre rectas uniendo cualesquiera dos puntos de $\mathbb{R}^d$.

La ϕ-convexidad de los potenciales V y W se traduce en la ϕ-convexidad para el funcional H, remitimos a [21] para los detalles del resultado, pero con esta noción de convexidad se puede demostrar el siguiente resultado:

Teorema 7 (Contracción general en d_2) *[21] Supongamos que $P(\rho)$ satisface (11), entonces dadas dos soluciones $\rho_1(t)$ y $\rho_2(t)$ de (1) con datos iniciales $\rho_i(0) \in L^1(\mathbb{R}^d)$, $\|\rho_i(0)\|_{L^1(\mathbb{R}^d)} = 1$, $\rho_i(0) \geq 0$, y con momentos de orden 2 acotados, se verifica que:*

(A) Si $V : \mathbb{R}^d \longrightarrow \mathbb{R}$ es uniformemente convexo, es decir, $D^2V(x) \geq kI$ con $k > 0$ y W es convexo, entonces

$$d_2(\rho_1(t), \rho_2(t)) \leq e^{-kt} d_2(\rho_1(0), \rho_2(0))$$

se verifica para $t \geq 0$.

(B) Sea $\phi(s) = (k/r)s^{r+1}$, $k, r > 0$, y supongamos que $V : \mathbb{R}^d \longrightarrow \mathbb{R}$ y $W : \mathbb{R}^d \longrightarrow \mathbb{R}$ son convexas satisfaciendo:

(i) $V(x)$ es ϕ-uniformemente convexa en $\mathbb{R}^d$, o

(ii) $W(x)$ es ϕ-uniformemente convexa en $\mathbb{R}^d$, y las soluciones tienen igual centro de masas (igual velocidad media en modelos cinéticos), es decir, $\langle x \rangle_{\rho_1(t)} = \langle x \rangle_{\rho_2(t)} = 0$ para $t \geq 0$.

Entonces para $t \geq 0$

$$d_2^2(\rho_1(t), \rho_2(t)) \leq \frac{d_2^2(\rho_1(0), \rho_2(0))}{(1 + tkd_2^r(\rho_1(0), \rho_2(0)))^{2/r}}.$$

Nota 3

- *La hipótesis de igual centro de masas en el anterior teorema se debe a que en caso de que* $V = 0$ *en la ecuación* (1) *el centro de masas se preserva a lo largo de la evolución, y por tanto el funcional* H *no es estrictamente convexo por desplazamientos a menos que nos restringamos al subconjunto de* $\mathcal{P}_2^{ac}(\mathbb{R}^d)$ *con centro de masa nulo.*
- *Al igual que comentamos en el caso de las ecuaciones de difusión no lineal, las propiedades de contracción estricta se pueden usar para demostrar la existencia y unicidad de soluciones de equilibrio para* (1). *Dichos resultados se han demostrado haciendo uso del funcional de entropía como funcional de Liapunov [20].*

Por último, veamos que la idea básica de contractividad de distancias en espacios de medidas de probabilidad, es muy útil en diversas circunstancias que puedan parecer lejanas de las ecuaciones de la forma (1). Consideremos la ecuación de Boltzmann homogénea para partículas inelásticas en la aproximación maxwelliana introducida en [10], en la cual suponemos que nuestras partículas se encuentran excitadas térmicamente y modelamos dicha excitación con un movimiento Browniano entre colisiones. Este modelo introducido en [15] da lugar a la ecuación

$$\frac{\partial f}{\partial t} = \sqrt{\theta(t)}\, Q(f, f) + \Delta_v f \tag{17}$$

donde $f(t, v)$ representa la probabilidad de encontrar partículas con velocidad $v \in \mathbb{R}^3$ en tiempo $t \geq 0$ y $3\theta(t)$ es el momento de orden dos de f; $\theta(t)$ representa la temperatura del medio granular.

El operador de colisión lo definimos de forma débil, de forma que

$$\langle \varphi, Q(f, f) \rangle = \frac{1}{4\pi} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_{S^2} f(v) f(w) \Big[\varphi(v') - \varphi(v) \Big] dn\, dv\, dw \tag{18}$$

con

$$v' = \frac{1}{2}(v + w) + \frac{1 - e}{4}(v - w) + \frac{1 + e}{4}|v - w| n \tag{19}$$

donde $0 < e \leq 1$ es el coeficiente de inelasticidad y $\varphi \in C(\mathbb{R}^3)$.

La ecuación de Boltzmann (17) se puede escribir de una forma muy sencilla en Fourier, de forma que

$$\frac{\partial \hat{f}}{\partial t} = \sqrt{\theta(t)} \frac{1}{4\pi} \int_{S^2} \Big\{ \hat{f}(t, k_-) \hat{f}(t, k_+) - \hat{f}(t, 0) \hat{f}(t, k) \Big\} dn - |k|^2 \hat{f} \tag{20}$$

con

$$\begin{aligned} k_- &= \frac{1+e}{4}(k - |k|n), \\ k_+ &= \frac{3-e}{4}\, k + \frac{1+e}{4}\, |k|n\,, \end{aligned} \tag{21}$$

donde $\hat{f}$ es la transformada de Fourier de f. El operador $Q(f,f)$ formalmente preserva la masa y la velocidad media, es decir, los momentos de orden cero y uno de f, y sin embargo, disipa la temperatura, es decir, el momento de orden dos.

Consideremos $\mathcal{P}_p(\mathbb{R}^3)$ el conjunto de las medidas de probabilidad con momento de orden $p > 0$ finito con la distancia basada en sus transformadas de Fourier definida por:

$$\tilde{d}_p(f,g) = \sup_{k\in\mathbb{R}^3} \frac{|\hat{f}(k) - \hat{g}(k)|}{|k|^p}$$

para cada par de medidas de probabilidad f, g en $\mathcal{P}_p(\mathbb{R}^3)$. Esta distancia es finita siempre que las medidas de probalidad tengan momentos iguales hasta orden $[p]$, donde $[p]$ es la parte entera de p. En caso de que $p \geq 1$ sea entero, basta con que los momentos hasta orden $p-1$ sean iguales para que $\tilde{d}_p$ sea finita.

De hecho, la topología inducida por $\tilde{d}_2$ es equivalente a la topología débil-* más convergencia de momentos de orden 2, y por tanto a la topología inducida por la distancia d_2 en $\mathcal{P}_2(\mathbb{R}^3)$ [49]. La distancia $\tilde{d}_2$ fue usada en el caso de la ecuación de Boltzmann elástica $e = 1$ para moléculas maxwellianas para demostrar unicidad de solución [32, 49].

Teorema 8 (Contracción uniforme de la distancia $\tilde{d}_2$) *Sean $\hat{f}_1$ y $\hat{f}_2$ dos soluciones de* (20) *con masa unidad, velocidad media cero y temperatura inicial igual, entonces $\tilde{d}_2(\hat{f}_1(0), \hat{f}_2(0)) < \infty$ y existe una constante $C > 0$ tal que*

$$\tilde{d}_2(\hat{f}_1(t), \hat{f}_2(t)) \leq \tilde{d}_2(\hat{f}_1(0), \hat{f}_2(0))\, \mathrm{e}^{-(1-\gamma_0)\mathrm{Ct}},$$

para cada $t \geq 0$ con $\gamma_0 = \frac{e^2+3}{4} < 1$.

Como consecuencia inmediata de la contracción uniforme obtenemos el siguiente resultado de existencia, unicidad y regularidad de los estados estacionarios de (17):

Teorema 9 (Existencia y unicidad de equilibrios difusivos) *La ecuación* (17) *tiene una única solución estacionaria f_∞ en el conjunto de las medidas de probabilidad con velocidad media cero. Además, todos los momentos de f_∞ son finitos, f_∞ es regular ($f_\infty \in H^\infty(\mathbb{R}^3)$) y nos describe el comportamiento asintótico de* (17)*:*

$$d_2(\hat{f}(t), \hat{f}_\infty) \leq d_2(\hat{f}_0, \hat{f}_\infty)\, \mathrm{e}^{-(1-\gamma_0)\mathrm{C}_1\mathrm{t}} + \mathrm{C}_2\mathrm{e}^{-\mathrm{C}_3\mathrm{t}},$$

para cada $t \geq 0$ y cada solución $\hat{f}(t)$ de (20) *con masa unidad, velocidad media cero y temperatura inicial finita con C_i, $i = 1, 2, 3$, constantes positivas.*

Aquí observamos otra de las grandes ventajas de la prueba de contracciones de distancias, nos demuestra de una manera sencilla la existencia y unicidad de las soluciones estacionarias y como consecuencia el comportamiento asintótico de las soluciones. El desarrollo de esta parte se puede ver en el trabajo en preparación [8].

Problema abierto 10 *Es conocido que la distancia de Wasserstein d_2 no crece para soluciones de la ecuación de Boltzmann para moléculas maxwellianas ($e = 1$) [44], resultado probado con técnicas probabilísticas. Es un problema abierto demostrar que dicha distancia es una contracción uniforme para la ecuación de Boltzmann* (17)*, así como dar una demostración alternativa analítica del resultado de Tanaka.*

Agradecimientos: Dedico este trabajo a la memoria de mi madre. Agradezco al SEMA por la concesión del Sexto Premio al Joven Investigador, del cual quiero hacer partícipe a todos aquellos que contribuyeron a formarme como matemático y como persona; especialmente en: Departamento de Matemática Aplicada de la Universidad de Granada, Department of Mathematics de la University of Texas at Austin, Departament de Matemàtiques de la Universitat Autònoma de Barcelona y de la Institució Catalana de Recerca i Estudis Avançats (ICREA). Agradezco también al Fields Institute (Toronto) por su hospitalidad durante el período en que este artículo fue escrito y la financiación recibida en el proyecto DGI-MCYT/FEDER BFM2002-01710.

Referencias

[1] Arnold, A., Carrillo, J. A., Desvillettes, L., Dolbeault, J., Jüngel, A., Lederman, C., Markowich, P. A., Villani, C., y Toscani, G., Entropies and equilibria of many-particle systems: an essay of recent research. Aparecerá en *Monatsh. Math.*

[2] Agueh, M., Existence of solutions to degenerate parabolic equations via the Monge-Kantorovich theory. PhD Thesis, Georgia Institute of Technology, 2002.

[3] Agueh, M. Asymptotic behavior for doubly degenerate parabolic equations. *C. R. Math. Acad. Sci. Paris, Ser. I 337*, (2003), 331-336.

[4] Ambrosio, L. A., Gigli, N., y Savaré, G., Gradient flows in metric spaces and in the Wasserstein spaces of probability measures. Aparecerá en Lecture Notes in Mathematics.

[5] Bakry, E. y Emery, M., Diffusions hypercontractives in *Sem. Probab. XIX* LNM 1123 Springer, New York, 1985, pp. 177–206.

[6] Benedetto, D., Caglioti, E., y Pulvirenti, M., A kinetic equation for granular media. *RAIRO Modél. Math. Anal. Numér. 31*, 5 (1997), 615–641.

[7] Benedetto, D., Caglioti, E., Carrillo, J. A., y Pulvirenti, M., A non-maxwellian steady distribution for one-dimensional granular media. *J. Stat. Phys. 91*, (1998), 979–990.

[8] Bisi, M., Carrillo, J. A., y Toscani, G., Contractive Metrics for a Boltzmann equation in granular gases: Diffusive equilibria. Trabajo en preparación.

[9] Bizon, C., Shattuck, M. D., Swift, J. B., y Swinney, H. L., Transport coefficients for granular media from molecular dynamics simulations. *Phys. Rev. E 60*, (1999), 4340–4351.

[10] Bobylev, A. V., Carrillo, J. A., y Gamba, I., On some properties of kinetic and hydrodynamics equations for inelastic interactions. *J. Statist. Phys. 98*, (2000), 743–773.

[11] Brenier, Y., The least action principle and the related concept of generalized flows for incompressible perfect fluids. *J. Amer. Math. Soc. 2*, (1989), 225–255.

[12] Brenier, Y., Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math. 44*, (1991), 375–417.

[13] Cáceres, M. J., Carrillo, J. A., y Goudon, T., Equilibration rate for the linear inhomogeneous relaxation-time Boltzmann equation for charged particles. *Comm. in PDEs 28*, (2003), 969–989.

[14] Carlen, E. y Gangbo, W., Constrained steepest descent in the 2-Wasserstein metric. *Annals Math. 157*, (2003), 807–846.

[15] Carrillo, J. A., Cercignani, C., y Gamba, I., Steady states of a Boltzmann equation for driven granular media. *Phys. Rev. E 62*, (2000), 7700–7707.

[16] Carrillo, J. A. y Fellner, K., Long time asymptotics via entropy methods for diffusion dominated equations. HYKE preprint (2003) (www.hyke.org).

[17] Carrillo, J. A., DiFrancesco, M., y Toscani, G., Asymptotic profiles for general nonlinear diffusion equations. Trabajo en preparación.

[18] Carrillo, J. A., Gualdani, M. P., y Toscani, G., Finite speed of propagation for the porous medium equation by mass transportation methods. HYKE preprint (2003) (www.hyke.org), aparecerá en Compte Rendus Acad. Sci. Paris.

[19] Carrillo, J. A., Jüngel, A., Markowich, P. A., Toscani, G., y Unterreiter, A., Entropy dissipation methods for degenerate parabolic systems and generalized Sobolev inequalities. *Monatsh. Math. 133*, (2001), 1–82.

[20] Carrillo, J. A., McCann, R. J., y Villani, C., Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Rev. Matemática Iberoamericana 19*, (2003), 1–48.

[21] Carrillo, J. A., McCann, R. J., y Villani, C., Contractions in the 2-Wasserstein length space and thermalization of granular media. HYKE preprint (2004) (www.hyke.org).

[22] Carrillo, J. A. y Vázquez, J. L., Fine asymptotics for fast diffusion equations. *Comm. PDE 28*, (2003), 1023–1056.

[23] Carrillo, J. A. y Toscani, G., Asymptotic L^1-decay of solutions of the porous medium equation to self-similarity. *Indiana Univ. Math. J. 49*, (2000), 113–141.

[24] Carrillo, J. A. y Toscani, G., Long-time asymptotics for strong solutions of the thin film equation. *Comm. Math. Phys. 225*, (2002), 551–571.

[25] Carrillo, J. A. y Toscani, G., Wasserstein metric and large–time asymptotics of nonlinear diffusion equations. HYKE preprint (2003) (www.hyke.org).

[26] Cercignani, C., The Boltzmann equation and its applications. Springer series in Applied Mathematical Sciences 67, Springer-Verlag, 1988.

[27] Cercignani, C., Recent developments in the mechanics of granular materials. Fisica matematica e ingegneria delle strutture, 119–132, Pitagora Editrice, Bologna, 1995.

[28] Cordero-Erausquin, D., Gangbo, W., y Houdre, C., Inequalities for generalized entropy and optimal transportation. To appear in Proceedings of the Workshop: Mass Transportation Methods in Kinetic Theory and Hydrodynamics.

[29] Desvillettes, L. y Villani, C., On the trend to global equilibrium in spatially inhomogeneous entropy-dissipating systems: the linear Fokker-Planck equation. *Comm. Pure Appl. Math. 54*, 1 (2001), 1–42.

[30] Dolbeault, J. y del Pino, M., Best constants for Gagliardo-Nirenberg inequalities and application to nonlinear diffusions. *J. Math. Pures Appl. 81*, (2002), 847–875.

[31] Dudley, R. M., Probabilities and metrics - Convergence of laws on metric spaces, with a view to statistical testing. Universitet Matematisk Institut, Aarhus, Denmark, 1976.

[32] Gabetta, E., Toscani, G., y Wennberg, W., Metrics for Probability Distributions and the Trend to Equilibrium for Solutions of the Boltzmann Equation. *J. Statist. Phys. 81*, (1995), 901–934.

[33] Gangbo, W. y McCann, R. J., The geometry of mass transportation. *Acta Math. 177*, (1996), 113–161.

[34] Givens, C. R. y Shortt, R. M., A class of Wasserstein metrics for probability distributions. *Michigan Math. J. 31*, (1984), 231–240.

[35] Gromov, M., Metric Structures for Riemannian and non-Riemannian Spaces. J. Lafontaine and P. Pansu, eds. With appendices by S. Semmes. Birkhauser, Boston, 1999.

[36] Gross, L., Logarithmic Sobolev inequalities. *Amer. J. of Math. 97*, (1975), 1061–1083.

[37] Kantorovich, L. V. y Rubinstein, G. S., On a space of completely additive functions. *Vestnik Leningrad. Univ. 13*, (1958), 52–59.

[38] Li, H. y Toscani, G., Long–time asymptotics of kinetic models of granular flows. Aparecerá en *Arch. Rat. Mech. Anal.*

[39] McCann, R. J., Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J. 80*, (1995), 309–323.

[40] McCann, R. J., A convexity principle for interacting gases. *Adv. Math. 128*, 1 (1997), 153–179.

[41] Neuberger, J., Sobolev gradients and differential equations. Lecture Notes in Mathematics 1670, Springer-Verlag, 1997.

[42] Otto, F., The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations 26*, (2001), 101–174.

[43] Otto, F. y Villani, C., Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal 173*, (2001), 361–400.

[44] Tanaka, H., Probabilistic treatment of the Boltzmann equation of Maxwellian molecules. *Z. Wahrsch. Verw. Gebiete 46*, 1 (1978/79), 67–105.

[45] Toscani, G., Kinetic approach to the asymptotic behaviour of the solution to diffusion equations. *Rendiconti di Matematica Serie VII 16*, (1996), 329–346.

[46] Toscani, G., Sur l'inégalité logaritmique de Sobolev. *C. R. Acad. Sci. Paris 324*, (1997), 689–694.

[47] Toscani, G., Entropy production and the rate of convergence to equilibrium for the Fokker-Planck equation. *Quarterly of Appl. Math. 57*, (1999), 521–541.

[48] Toscani, G., One-dimensional kinetic models of granular flows. *RAIRO Modél. Math. Anal. Numér. 34*, 6 (2000), 1277–1291.

[49] Toscani, G. y Villani, C., Probability Metrics and Uniqueness of the Solution to the Boltzmann Equation for a Maxwell Gas. *J. Statist. Phys. 94*, (1999), 619–637.

[50] Vázquez, J. L., An introduction to the mathematical theory of the porous medium equation. Shape optimization and free boundaries (Montreal, PQ, 1990). *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 380*, 347–389, Kluwer Acad. Publ., 1992.

[51] Vázquez, J. L., Las ecuaciones de filtración de fluidos en medios porosos. *Bul. SEMA 14*, (1999), 37–83.

[52] Vázquez, J. L., Asymptotic behaviour for the porous medium equation posed in the whole space. *J. Evol. Eq. 3*, (2003), 67–118.

[53] Villani, C., *Topics in optimal transportation.* Graduate Studies in Mathematics Vol. 58, Amer. Math. Soc, Providence, 2003.

[54] Wasserstein, L. N., Markov processes over denumerable products of spaces describing large systems of automata. *Problems of Information Transmission 5*, (1969), 47–52.

[55] Williams, D. R. M. y MacKintosh, F. C., Driven granular media in one dimension: Correlations and equation of state. *Phys. Rev. E 54*, (1996), R9–R12.

Bol. Soc. Esp. Mat. Apl. nº 28(2004), 155–168

Método de producción diaria de huevos para la anchoa del Golfo de Bizkaia

L. Ibaibarriaga, M. Santos y A. Uriarte

Fundación AZTI[1], Pasajes, Gipuzkoa

libaibarriaga@pas.azti.es

Resumen

El método de producción diaria de huevos es un método directo de evaluación de poblaciones de peces y viene siendo aplicado por AZTI para la evaluación de la anchoa del Golfo de Bizkaia desde hace más de 12 años. El presente artículo trata por un lado de describir las bases de este método tal y como se ha utilizado tradicionalmente, y por otro, de presentar las últimas líneas de investigación orientadas al uso de modelos aditivos generalizados.

1 Introducción

AZTI es un centro tecnológico especializado en recursos pesqueros, medio ambiente marino y tecnología de los alimentos que está situado en la Comunidad Autónoma del País Vasco (CAPV). La actividad principal de AZTI se centra en proyectos de Investigación y Desarrollo Tecnológico (I+DT) que realiza para empresas y administraciones públicas de la CAPV, nacionales y europeas. Asimismo, ofrece servicios de transferencia de tecnología, asesoramiento y consultoría técnica, junto con labores de formación y difusión tecnológica.

En particular, en el ámbito de los recursos pesqueros, el objetivo principal de AZTI es investigar para conseguir una actividad pesquera sostenible, llevada a cabo por una flota económicamente competitiva, con prácticas de pesca responsable. Sus principales áreas de actuación son:

Biología pesquera. Determinación de los parámetros biológicos relevantes en la evaluación de los stocks, como crecimiento, mortalidad, reproducción, migración y distribución espacio-temporal.

[1]AZTI es una Fundación sin ánimo de lucro, comprometida con el desarrollo social y económico del sector pesquero y alimentario, así como con la protección del medio ambiente marino y los recursos naturales.

Seguimiento de pesquerías. Obtención de índices indirectos y directos de abundancia. Los índices indirectos son aquellos derivados de las capturas comerciales (desembarcos y descartes), mientras que los índices directos provienen de campañas científicas específicamente diseñadas con el fin de evitar los posibles sesgos derivados de la actividad pesquera. Ambos tipos de índices se incorporan en modelos de evaluación integral, de forma que se obtienen estimas de la abundancia de la población que permiten proyectar las capturas y simular el comportamiento de las pesquerías a medio y largo plazo, pudiendo así analizar el riesgo asociado a diversas medidas de gestión.

Evaluación de recursos pesqueros. Participación activa en los organismos internacionales para el establecimiento del consejo científico de gestión de los recursos marinos vivos (ICCAT, IOTC, ICES/CIEM, NAFO, etc).

Tecnología pesquera. Desarrollo de tecnologías para mejorar la eficiencia de la pesca, tales como mejora de diseño de artes y aparejos de pesca, métodos de detección y atracción, estudio del comportamiento de las especies ante las artes y los aparejos de pesca y aplicación de tecnologías de la información y comunicaciones a la pesca. Todo ello desde el punto de vista de gestión pesquera responsable, por medio de la evaluación del esfuerzo pesquero, el estudio de la selectividad y el impacto de las artes de pesca, y la minimización de los descartes.

Ecología. Estudio del impacto de la variabilidad natural y el cambio climático en el ecosistema marino y las pesquerías. Entre ellos, la productividad biológica, los procesos que afectan al reclutamiento de los stocks pesqueros, el acoplamiento de modelos biológicos y físicos y la relación entre medio marino y pesquerías a través de la teledetección.

Socio-economía de la pesca. Estudio de los componentes económicos y sociales de las pesquerías, con el fin de aportar información para un desarrollo sostenible las mismas.

En todas estas áreas las matemáticas, y principalmente la estadística, juegan un papel importante, y son muchos y muy diversos los campos de aplicación que intervienen. Estos van desde los métodos numéricos utilizados en oceanografía física, hasta los modelos socio-económicos, pasando por teoría del muestreo para la obtención de índices indirectos, geostadística en el estudio de las distribuciones espaciales de la población, análisis multivariante para la caracterización de hábitats, y cuestiones de aplicación más general, como diseño de experimentos, inferencia estadística (tanto frecuentista como bayesiana) o métodos computacionales de simulación.

Tanto es así, que en los últimos años la demanda de matemáticos y estadísticos en el ámbito de las pesquerías a nivel general, y en AZTI en particular, ha aumentado de forma considerable. Su labor básica normalmente consiste en formar parte, junto con biólogos, biólogos marinos, físicos, ecólogos,

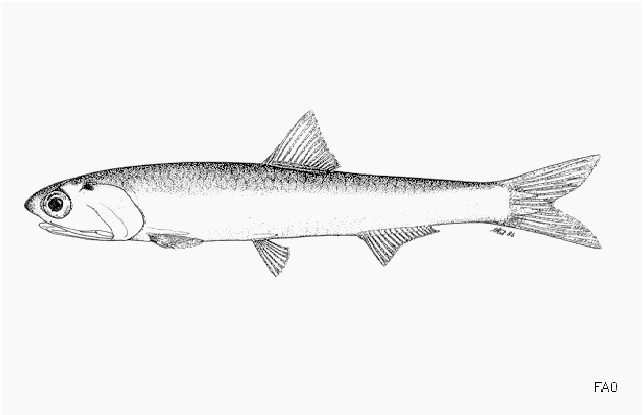


Figura 1: Engraulis encrasicholus, L. tomada de FAO [18]

químicos y/o ingenieros, de equipos de trabajo multidisciplinares que hacen frente a problemas de diversa índole. Hay que resaltar que en AZTI existe un interés genuino por la investigación propiamente dicha, que se traduce en la realización por los investigadores de la Tesis Doctoral.

El presente artículo describe el Método de Producción Diaria de Huevos (MPDH), uno de los métodos directos de evaluación que lleva a cabo AZTI con el fin de obtener índices directos de abundancia para la población de anchoa del Golfo de Bizkaia, *Engraulis encrasicholus, L.*, ver Figura (1). Se trata aquí de ilustrar la incorporación de los últimos avances en modelado estadístico, describiendo el trabajo realizado en el proyecto europeo n^o 99/080 titulado "Using environmental variables with improved DEPM methods to consolidate the series of sardine and anchovy estimates". En este proyecto, además de AZTI han colaborado el IEO (Instituto Español de Oceanografía), IPIMAR (Instituto de Investigação das Pescas e do Mar) y el grupo de investigación RUWPA (Research Unit of Wildlife Population Assessment) de la Universidad de St Andrews (Escocia).

En la sección 2 se describen las bases del MPDH. En la sección 3 se presentan los modelos aditivos generalizados (GAMs -generalized additive models), se enumeran brevemente las mejoras metodológicas desarrolladas en el transcurso del proyecto y se muestran de forma resumida los resultados principales obtenidos en la aplicación a la serie de datos desde 1996 hasta 1999 para la anchoa del Golfo de Bizkaia. Finalmente, en la sección 4 se discuten algunos de los resultados y se establecen futuras líneas de trabajo.

2 El Método de Producción Diaria de Huevos

El Método de Producción Diaria de Huevos desarrollado inicialmente por Parker ([16]) y modificado posteriormente por Stauffer y Piquelle ([21]), viene

aplicándose con éxito en la evaluación de la población de anchoa en el Golfo de Bizkaia desde 1987 ([11]). Este método consiste en estimar la biomasa desovante como el cociente entre las estimas de producción total diaria de huevos y de fecundidad diaria (número de huevos puestos por día y unidad de peso) de la población. Más concretamente, incorporando el hecho de que la anchoa es ponedora múltiple que realiza repetidas puestas a lo largo de la temporada de freza:

$$SSB = \frac{P_0}{DF} = \frac{D_0 A}{RS\frac{F}{W}}\,, \tag{1}$$

donde SSB (spawning stock biomass) es la biomasa desovante en toneladas, P_0 la producción total diaria de huevos, DF (daily fecundity) la fecundidad diaria de la población, D_0 la producción diaria de huevos por unidad de superficie, A el área de puesta, R la proporción en peso de las hembras en la población, S la fracción diaria de las hembras maduras en puesta, F la fecundidad parcial (número de huevos puestos por día y unidad de peso de cada hembra), y W el peso medio de las hembras maduras.

La varianza asociada a la estima de SSB se calcula a partir de las estimas de las varianzas y covarianzas asociadas a las estimas de cada uno de los parámetros por medio del método delta descrito en [19].

A continuación, siguiendo a [17], se detalla cómo se obtienen las estimas para cada uno de los parámetros de (1). Buenas referencias a nivel general pueden encontrarse en [12] y [14].

2.1 Producción diaria de huevos

Para la estimación de la producción total diaria de huevos, P_0, todos los años se realiza una campaña de ictioplancton, (ver [6]), en el Golfo de Bizkaia entre los meses de mayo y junio, que es el período en que sucede la puesta. Esta campaña sigue un diseño de muestreo de tipo adaptivo sistemático con transectos perpendiculares a costa, de forma que los límites geográficos y la intensidad del muestreo se van modificando de acuerdo a la abundancia de huevos hallada en cada estación.

El muestreo se realiza utilizando principalmente una red llamada CalVET (California Vertical Egg Tow), red de plancton vertical especialmente diseñada para muestrear huevos de anchoa y sardina. Dado que se quiere estimar la densidad media de huevos en la columna vertical de agua, se coloca un flujómetro en la boca de la red que permite calcular la distancia efectiva recorrida por la red y detectar así los posibles comportamientos anómalos ocurridos durante el lance. Además, se calcula el volumen filtrado, que permitirá más adelante transformar el número de huevos por m^3 observado en cada estación en densidades de huevos:

$$D_i = \frac{N_i R_i}{\pi r^2 l_i}\,, \tag{2}$$

donde D_i es la densidad de huevos, N_i el número de huevos muestreado, R_i la profundidad de la columna de agua muestreada, r el radio de la boca de la red,

l_i la longitud efectiva dada por el flujómetro, y R_i la profundidad de la columna de agua muestreada, todos ellos en la estación i.

Una vez que los huevos han sido recogidos y debidamente triados, se clasifican en estadios de acuerdo a sus características morfológicas. A continuación, se transforman en cohortes diarias (clases de huevos que han sido puestos el mismo día) usando información sobre las tasas de desarrollo de los estadios en función de la temperatura, obtenida mediante experimentos de incubación. Para cada estación, las frecuencias de huevos por cohorte se transforman en densidades mediante la fórmula (2), y la edad media de los huevos en cada cohorte se calcula como la edad media ponderada por el número de huevos en cada edad. Entonces, la producción diaria de huevos, D_0, y la tasa de mortalidad diaria, Z, se estiman al ajustar el modelo de mortalidad exponencial

$$E[D_{ij}] = D_0 e^{-Z a_{ij}} , \tag{3}$$

donde D_{ij} y a_{ij} representan respectivamente la densidad de huevos (número de huevos por m^2) y la edad media (en días) en la cohorte j de la estación i.

A su vez, el área total muestreada se post-estratifica en un estrato positivo que contiene el área de puesta, y un estrato negativo que abarca las estaciones situadas fuera de los límites del área de puesta. El área de puesta se estima como la suma de las áreas que representan las estaciones del estrato positivo.

Así pues, la producción total diaria de huevos se estima como el producto de las estimas de la producción diaria de huevos por unidad de superficie y del área de puesta.

2.2 Fecundidad parcial

Con en fin de estimar los parámetros reproductores de la población que aparecen en (1), coincidiendo en el espacio y en el tiempo con la campaña de ictioplancton, se tiene que realizar una campaña de muestreo de adultos. Tradicionalmente, en el caso de la anchoa del Golfo de Bizkaia, estas muestras provienen de la campaña acústica de evaluación realizada anualmente por IFREMER (Institut français de recherche pour l'exploitation de la mer) a bordo del buque de investigación Thalassa. Estas muestras se suelen completar con algunas recogidas de forma oportunista por los barcos de cerco de la flota comercial vasca que se encuentran faenando.

De las pescas realizadas en el Thalassa se selecciona primeramente una muestra aleatoria de 2 kg. El muestreo finaliza bien cuando se ha asignado el sexo a un mínimo de 1 kg de anchoas (o 60 individuos), cuando se han obtenido 25 hembras no hidratadas, o cuando se han superado los 120 ejemplares sin haber logrado el objetivo de tener 25 hembras no hidratadas. Las muestras obtenidas en los barcos pesqueros se preservan y se envían al laboratorio para su posterior análisis.

Las estimas de los parámetros reproductores de los adultos se basan en el tipo de muestreo realizado (ver [5]). En este caso, el muestreo se corresponde con un muestreo en dos etapas: una primera etapa en que se seleccionan las

estaciones en que se van a efectuar los lances, que se aproxima a un muestreo de probabilidad proporcional a la abundancia, y una segunda etapa en que de las capturas obtenidas en cada lance se selecciona una muestra aleatoria de individuos.

Bajo estas condiciones, un estimador insesgado de la media poblacional μ de cualquiera de las características de los adultos Y que se quiere determinar es

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \bar{y}_i}{n},$$

donde $\hat{\mu}$ denota la estima de la media de la población, n es el número de estaciones, m_i es el número de individuos submuestreados en el lance i, y_{ij} es el valor observado correspondiente al individuo j del lance i, e

$$\bar{y}_i = \frac{\sum_{j=1}^{m_i} y_{ij}}{m_i}$$

es la media observada en la submuestra del lance i.

Sin embargo, a menudo el tamaño de las submuestras difiere entre los diferentes lances y la expresión anterior se tiene que modificar a

$$\hat{\mu} = \frac{\sum_{i=1}^{n} m_i \bar{y}_i}{\sum_{i=1}^{n} m_i}, \tag{4}$$

donde el estimador de la varianza asociada a este estimador es

$$\widehat{Var}(\bar{y}) = \frac{\sum_{i=1}^{n} m_i^2 (\bar{y}_i - \bar{y})^2}{n(n-1)\left(\sum_{i=1}^{n} \frac{m_i}{n}\right)^2}. \tag{5}$$

Estas ecuaciones (4) y (5) se utilizan para obtener las estimas y las correspondientes estimas de varianza de los parámetros de adultos de la población, tomando m_i y $\bar{y}_i$ como corresponda en cada caso.

Para poder estimar el peso de las hembras maduras de la población, se corrige el peso de las hembras de la submuestra debido al efecto de su conservación en formaldehído. Además, con el fin de corregir el sobrepeso de las hembras hidratadas debido a la retención de líquidos, su peso se estima a partir de un modelo de regresión lineal de peso total frente a peso sin gónada ajustado utilizando las observaciones recogidas de las hembras no hidratadas.

La fecundidad parcial solo puede ser observada en las hembras en estado de hidratación. Así pues, siguiendo el método de los ovocitos hidratados descrito en [8], se calcula la fecundidad parcial de las hembras hidratadas como

$$F_{ij} = O_{ij} \frac{1}{3} \sum_{k=1}^{3} \frac{u_{ijk}}{\nu_{ijk}} ,$$

donde O_{ij} es el peso de la gónada de la hembra j del lance i, y ν_{ijk} y u_{ijk} son respectivamente el peso y el número de ovocitos hidratados hallados en la muestra k del tejido ovárico de la hembra j del lance i. Ahora bien, hay una alta correlación entre el número de huevos por puesta y el peso sin gónada, de forma que la fecundidad parcial para las hembras maduras no hidratadas se puede estimar a partir de un modelo de regresión lineal de la fecundidad parcial frente al peso sin gónada ajustado a las observaciones obtenidas de las hembras hidratadas. En este caso, la expresión (5) para el cálculo de la varianza asociada a la estima de F se corrige con el fin de incluir la variabilidad asociada al hecho de que la fecundidad parcial deducida para las hembras hidratadas es, en sí misma, una estimación.

La frecuencia de puesta, S, se define como la proporción de hembras en puesta diariamente, y se estima como la incidencia de folículos post-ovulatorios de día 1 hallados en las gónadas de las hembras maduras ([9]):

$$S_i = \frac{N1_i}{N0_i + N1_i + N2_i + N3_i} ,$$

donde NJ_i es el número de hembras con folículos post-oculatorios de día J halladas en la muestra i. Ahora bien, se sabe que las hembras con ovocitos con núcleos en migración o hidratación son sobremuestreadas, lo cual se corrige calculando la frecuencia de puesta de las hembras de la muestra de cada lance como:

$$S_i = \frac{N1_i}{2N1_i + N2_i} .$$

Finalmente, la razón de sexo en número de la población se asume que es 1:1, y la razón de sexo en peso de la muestra de cada lance se deriva como la razón de los pesos medios entre hembras y machos.

3 Aplicación de GAM's para estimar P_0

Hasta ahora se han descrito las bases del MPDH, tal y como se ha venido implementando a lo largo de los años. Sin embargo, hoy en día existen modelos matemáticos, como los modelos lineales generalizados (GLM) y los modelos aditivos generalizados (GAM), que se adaptan mejor a la naturaleza de algunos de los problemas que plantea el MPDH.

Los GLM, ([13]), son una generalización de los modelos de regresión lineal, pero permiten asumir un amplio rango de distribuciones de probabilidad, además de un cierto grado de no linealidad. En general, un modelo lineal generalizado se representa como

$$E[y_i] \equiv \mu_i = g^{-1}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\right),$$

donde y sigue una distribución de la familia exponencial, $x_1, \ldots, x_p$ son p variables explicativas, $\beta_0, \beta_1, \ldots, \beta_p$ son los p+1 parámetros a estimar del modelo y $g(\cdot)$ es llamada función enlace (link function), que es una función monótona y diferenciable.

En particular, siguiendo el trabajo de [2], el modelo de mortalidad exponencial (3) que tradicionalmente se ajusta por mínimos cuadrados no lineal ponderando cada observación por el área que representa, se puede reescribir de forma más natural como un GLM:

$$\log(E[N_{ij}]) = \log(R_i) + \log(D_0) - Za_{ij}, \tag{6}$$

donde N_{ij} es el número de huevos de la cohorte j de la estación i, que se asume sigue bien una distribución de Poisson o bien una binomial negativa, con función enlace logarítmica, $log(D_0)$ y Z son los parámetros a estimar, y $\log(R_i)$, que representa el logaritmo del área efectiva muestreada en cada estación i, no es más que un offset.

Los GAM, ([7]), tienen forma aún más general que los GLM, ya que el predictor pasa de ser lineal a convertirse en curvas o superficies más flexibles representadas bien mediante funciones univariables, $s(x_i)$, como los splines cúbicos o bien mediante funciones multivariables, $s(x_1, \ldots, x_p)$, como los thin plate splines.

De forma general, un modelo aditivo generalizado se puede expresar como

$$E[y_i]) = g^{-1}(\beta_0 + s(x_{1i}, \ldots, x_{pi})).$$

En este caso, el modelo de mortalidad exponencial (3) planteado como un GAM pasa a tener la forma:

$$\log(E[N_{ij}]) = \log(R_i) + \log(D_0) - Za_{ij}, \tag{7}$$

donde, a diferencia de (6), $\log(D_0)$ y Z son smooths de variables explicativas, como pueden ser las variables geográficas o medioambientales recogidas a lo largo de la campaña de ictioplancton.

Las primeras aplicaciones de los GAM en métodos de producción de huevos se pueden ver en [1], [3] y [4]. Ya en ellas se mostró el gran potencial de los GAMs para modelar la distribución espacial de la producción de huevos y estudiar su relación con el medioambiente, disminuyendo además la varianza asociada a la estima de la produción total. Desde entonces han sido varios los trabajos presentados aplicando GAMs a las campañas de ictioplancton, como por ejemplo [22].

Siguiendo esta línea de trabajo, AZTI ha participado en el proyecto de financiación europea n^o 99/080. "Using environmental variables with improved DEPM methods to consolidate the series of sardine and anchovy estimates",

que como su propio título índica, trataba de resolver algunos de los problemas metodológicos detectados en [3], con el fin de mejorar la eficiencia y la precisión de las estimas de biomasa obtenidas para la anchoa y la sardina ibero-atlánticas, y obtener a su vez mapas de distribución de ambas especies, que permitan estudiar su relación con variables espaciales y medioambientales y su evolución temporal.

Los avances más importantes realizados en lo que a los GAM se refiere son:

- Selección del modelo de forma automática basada en el estadístico GCV (generalised cross validation), cuyas bases se describen en [23] y [25].
- Posibilidad de modelar las interacciones entre las covariables por medio de thin plate splines, ([24]).
- Desarrollo de un marco teórico sólido para la estimación de intervalos de confianza GAM, extendiendo el modelo bayesiano de Silverman ([20]) para splines cúbicos de una dimensión.
- Implementación de la familia binomial negativa dentro de los GAM, permitiendo así modelar la sobredispersión.
- Posibilidad de modelar en (7) no sólo la tasa de producción de huevos sino también la tasa de mortalidad.

Todo esto ha sido implementado en una librería de R (http://www.r-project.org/). Además, también se han implementado:

- Cálculo automatizado del área y los límites de la campaña.
- Nuevo método de asignación de edades bayesiano, basado en un modelo multinomial de la clasificación por estadios observada en los experimentos de incubación.
- Estimación de la varianza por medio de un bootstrap no paramétrico.

En el caso de la anchoa del golfo de Bizkaia, esta metodología se utilizó para la serie de datos 1996-1999. Las variables explicativas disponibles para ajustar las superficies de tasa de producción diaria por unidad de área y mortalidad de (7) fueron latitud, longitud, distancia a lo largo de la costa a un punto dado, distancia perpendicular a costa, temperatura de superficie, salinidad de superficie y profundidad. Tras un análisis exploratorio y teniendo en cuenta el conocimiento previo, para la producción diaria de huevos se consideraron un smooth bivariable de longitud y latitud para describir la situación espacial de la puesta, un smooth bivariable de salinidad y temperatura para caracterizar la situación medioambiental y un smooth bivariable de profundidad y distancia a un punto fijo a lo largo de la costa para estudiar si el efecto de la profundidad variaba en el espacio. En todos los años la tasa de mortalidad se asumió constante en el espacio, para así poder comparar los resultados con los obtenidos mediante el método tradicional.

Así, se ajustaron una serie de modelos con distribución Poisson con sobre-dispersión y función enlace logarítmica. El modelo final se seleccionó por ser aquel con estadístico GCV de menor valor ([23] y [25]). La producción total de huevos se calculó como la suma de los valores predichos por el modelo sobre una rejilla regular que cubría el área de la campaña. La varianza se estimó por medio de un bootstrap no parámetrico que incluye la varianza asociada al proceso de asignación de edades.

La Figura (2) muestra las estimas y los intervalos de confianza de la producción diaria de huevos de 1996 a 1999 obtenidas mediante el método tradicional y la nueva metodología GAM. Las estimas puntuales de ambos métodos son consistentes, con diferencias entre el 1 y 20 %. Contrariamente a lo esperado, la nueva metodología no redujo la varianza asociada a P_0. Sin embargo, ha de destacarse que las varianzas obtenidas mediante el método tradicional son pequeñas y resultan poco realistas, por ejemplo el CV es de 5 % para 1997, mientras que las estimas obtenidas mediante la nueva metodología GAM son verosímiles, CV entre 13 y 21 %. Esto indica que las estimas de varianza del método tradicional podrían estar negativamente sesgadas.

La Figura (3) muestra las superficies de producción diaria de huevos de 1996 a 1999 obtenidas a partir de la nueva metodología GAM.

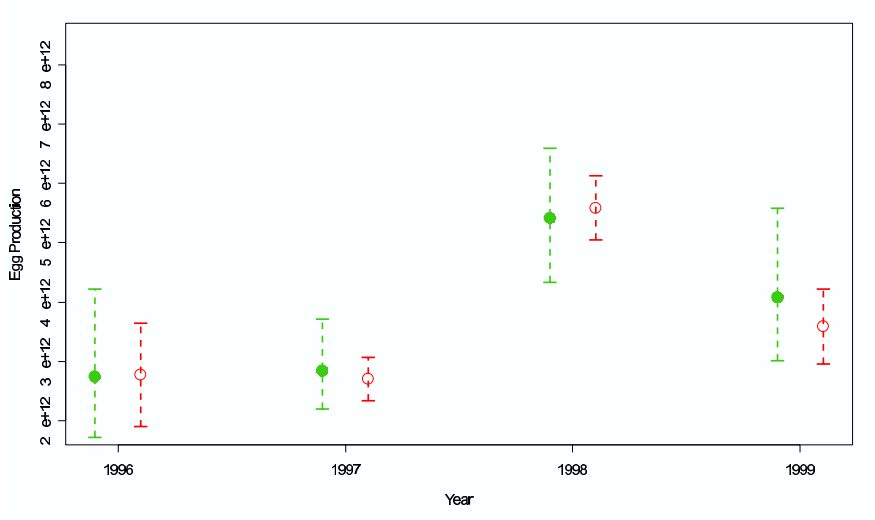


Figura 2: Serie de las estimas de la producción diaria de huevos junto con los intervalos de confianza al 95 % (líneas verticales) calculados por el método tradicional (en rojo asumiendo normalidad) y por el nuevo método GAM (en verde, asumiendo lognormalidad).

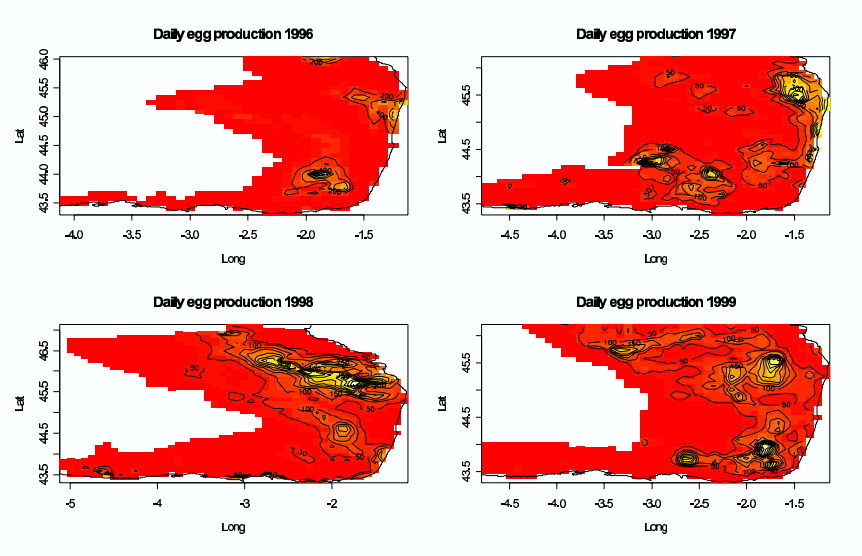


Figura 3: Producción diaria de huevos de anchoa en el Golfo de Bizkaia ajustadas por medio de GAMs.

4 Conclusiones

En general, la aplicación de los GAMs al MPDH ha supuesto un avance considerable. Por un lado, permite obtener la distribución espacial de la producción diaria de huevos y estudiar así cambios espacio-temporales de la puesta y su relación con la situación medio-ambiental, y por otro lado permite disminuir la varianza asociada a las estimas. Aunque éste no ha sido el caso de la anchoa, debido probablemente a un sesgo negativo en las estimas de varianza tradicionales. Y todo ello sobre unas bases metodológicas bien fundadas estadísticamente.

Hay que indicar que los GAMs se pueden aplicar no sólo a la producción de huevos, sino que también se pueden utilizar para estudiar la estructura espacial de la fecundidad diaria. Como se ha descrito en la sección 2.2 los parámetros de adultos clave para la anchoa del Golfo de Bizkaia son el peso medio de las hembras, W, y la fracción de puesta, S, ya que la fecundidad parcial, F depende directamente de W por medio de un modelo de regresión lineal y el porcentaje en peso de hembras con respecto al total, R, se deriva de W bajo la hipótesis de que el porcentaje en número de hembras con respecto al total es de 1:1. Así pues, una vez modeladas las superficies de W y S se pueden deducir las superficies de F y R, y luego la superficie de DF. Por último, el mapa de la biomasa desovante, SSB, se puede obtener dividiendo las superficies de la producción diaria de huevos y de la fecundidad diaria predichas por los modelos GAM. A este respecto ya se han hecho las primeras pruebas en [10] considerando para W

un GAM con distribución normal, función enlace identidad y smooth bivariable de longitud y latitud, y para S un GAM con distribución binomial, función enlace logit y smooth bivariable formado por longitud y latitud. Sin embargo, se está en una fase preliminar. Uno de los objetivos a medio plazo es poder implementar la estimación de la varianza asociada a los parámetros de adultos.

Este artículo ha tratado simplemente de mostrar una de las múltiples aplicaciones de las matemáticas al mundo de los recursos pesqueros, esperando así el haber contribuido a acercar los lectores a este campo.

Referencias

[1] N.H. Augustin, D.L. Borchers, E.D. Clarke, S.T. Buckland, and M. Walsh. Spatio-temporal modelling for the annual egg production method of stock assessment using generalised additive models. Can. J. Fish. Aquat. Sci. 55, 2608–2621, 1998.

[2] M. Bernal. A likelihood model and a new ageing procedure for improving the daily egg production estimates in species with fast developing eggs. MSc Thesis, University of St. Andrews, St. Andrews, Scotland, 1985.

[3] D.L. Borchers, S.T. Buckland, I.G. Priede and S. Ahmadi. Improving the precision of the daily egg production method using generalized additive models. Can. J. Fish. Aquat. Sci. 54: 2727-2742, 1997.

[4] D.L. Borchers, A. Richardson and L.Motos. Modelling the spatial distribution of fish eggs using generalized additive models. Ozeanografika 2, 1997.

[5] W.G. Cochran. Sampling techniques, John Wiley & Sons, Inc., New York, 3rd edition, 1977.

[6] D.R. Gunderson. Survey of fisheries resources, John Wiley & Sons, Inc., 1993.

[7] T.J. Hastie and R.J. Tibshirani. Generalized Additive Models. Chapman and Hall, London, 1990.

[8] J.R. Hunter, N.C.H. Lo and R.J.H. Leong. Batch fecundity in multiple spawning fishes. In R. Lasker (ed.) An egg production method for estimating spawning biomass of pelagic fish: application to the northern anchovy, *Engraulis mordax, p.*. US Dep. Commer., NOAA Tech. Rep. NMFS 36, 1985.

[9] J.R. Hunter and B.J. Macewicz. Measurement of spawning frequency in multiple spawning fishes. In R. Lasker (ed.) An egg production method for estimating spawning biomass of pelagic fish: application to the northern anchovy, *Engraulis mordax, p.*. US Dep. Commer., NOAA Tech. Rep. NMFS 36, 1985.

[10] ICES. Report of the study group on the estimation of the spawning biomass of sardine and anchovy. ICES CM 2002/G:01, 2002.

[11] ICES. Report of the working group on the assessment of mackerel, horse mackerel, sardine and anchovy. ICES CM 2003/ACFM:07, 2003.

[12] R. Lasker (ed.) An egg production method for estimating spawning biomass of pelagic fish: application to the northern anchovy, Engraulis mordax. U.S. Dep. Commer. NOAA Tech. Rep. NMFS 36, 1985.

[13] P. McCullagh and J.A. Nelder. Generalized Linear Models. Chapman and Hall, London, 2nd edition, 1989.

[14] L. Motos. Estimación de la biomasa desovante de la anchoa en el Golfo de Vizcaya, Engraulis encrasicolus, a partir de su producción de huevos. Bases metodológicas y aplicación, Tesis doctoral, Universidad del País Vasco-Euskal Herriko Unibertsitatea, Leioa, 1994.

[15] I.Palomera and P.Rubiés. The European Anchovy and its Environment. Sci. Mar. 60 (Supl.2), 1996.

[16] K. Parker. A direct method for estimating northern anchovy, *Engraulis mordax*, spawning biomass. Fish. Bull. 78: 541-544, 1980.

[17] S. Piquelle and G. Stauffer. Parameter estimation for an egg production method of anchovy biomass assessment. In R. Lasker (ed.) An egg production method for estimating spawning biomass of pelagic fish: application to the northern anchovy, *Engraulis mordax, p.*. US Dep. Commer., NOAA Tech. Rep. NMFS 36, 1985.

[18] W. Schneider. FAO species identification sheets for fishery purposes. Field guide to the commercial marine resources of the Gulf of Guinea. Prepared and published with the support of the FAO Regional Office for Africa. FAO, Roma, 1990.

[19] G.A.F. Seber. The estimation of animal abundance and related parameters. Charles Griffin and Co., London, 2nd edition, 1982.

[20] B.W. Silverman. Some aspects of the spline smoothing approach to nonparametric regression curve fitting. J.R. Statist. Soc. (B) 47: 1-52, 1985.

[21] G.D. Stauffer and S.J. Piquelle. Estimation of the 1980 spawning biomass of the subpopulation of northern anchovy. Natl. Mar. Fish. Serv. Southwest Fish. Cent., La Jolla, CA, Admin. Rep. LJ-80-09, 1980.

[22] Y. Stratoudakis, M. Bernal, D.L. Borchers and M.F. Borges. Changes in the distribution of sardine eggs and larvae off Portugal, 1985-2000. Fisheries Oceanography 12: 49-60, 2003.

[23] S.N. Wood. Modelling and smoothing parameter estimation with multiple quadratic penalties. J. Roy. Statist. Soc. (B) 62: 413-428, 2000.

[24] S.N. Wood. Thin plate regression splines. J. Roy. Statist. Soc. (B) 65: 95-114, 2003.

[25] S.N. Wood and N.H. Augustin. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling 157: 157-177, 2002.

Bol. Soc. Esp. Mat. Apl. nº 28(2004), 169–184

Enseñanza de las Matemáticas basada en los estilos de aprendizaje

A. Nevot

Departamento de Matemática Aplicada. E. U. de Arquitectura Técnica de la Universidad Politécnica de Madrid

nevot.luna@euatm.upm.es

Resumen

La intención de estas propuestas pedagógicas para la enseñanza de las matemáticas, basada en los diferentes Estilos de Aprendizaje, es proponer un plan de acción para aquellos aspectos que bloquean el aprendizaje, aportando propuestas y sugerencias de mejora, en el ámbito de la enseñanza de las matemáticas.

Durante los últimos años se han llevado a cabo diversos trabajos de investigación sobre los estilos de aprendizaje. Así, se ha comprobado que los estilos de aprendizaje del profesor son muy importantes porque repercuten en su manera de enseñar. Es frecuente que el profesor tienda a enseñar como le gustaría que le enseñaran a él, es decir, como le gustaría aprender. Diversas investigaciones prueban que los estudiantes aprenden con más efectividad cuando se les enseña con sus estilos de aprendizaje preferidos.

1 Introducción

Contrariamente a ideas y prácticas muy extendidas, es la enseñanza la que debe adaptarse al enseñado. Es el alumno el que debe ocupar el centro de todo acto educativo y, a medida que adquiere madurez, el alumno debe sentirse cada vez más libre de decidir por sí mismo lo que quiere aprender y en lo que desea formarse. Por tanto, la docencia es cada día más un arte, además de una profesión, en la que se impone la calidad en todas sus actividades profesionales y humanas (R. Díez [4]).

Las inquietudes que surgen diariamente en el desarrollo de la labor docente son fruto de la experiencia, del día a día en contacto con los alumnos. En los comienzos de la carrera profesional se actúa, en la mayoría de los casos, de forma impulsiva y sin detenerse a analizar las consecuencias de dichas actuaciones.

Con el paso de los años, de forma intuitiva y sin ninguna base teórica, se van fabricando una batería de recetas técnicas o trucos que posibilitan mejorar, o al menos así se cree, nuestra labor docente Sin embargo, esa batería va creciendo infinitamente año tras año. De tal forma que, cuando se cree conocer toda la casuística del aula, surge algo novedoso y difuso que trastoca toda nuestra buena disposición para solucionar los nuevos problemas y acontecimientos.

La siguiente reflexión, expresada por R.J. Sternberg [10], refleja un cambio radical en la interpretación de la práctica docente y nos puede servir de motivación. "Un día me di cuenta de que, en todos los años anteriores, muchos de los estudiantes que había considerado tontos no lo eran en absoluto: simplemente no aprendían de una manera compatible con mi forma de enseñar; y además me di cuenta de que el hecho de que yo enseñara el material de una sola manera no les había dado ni una oportunidad".

"En el ámbito más concreto de las matemáticas —afirman R. Dunn y k. Dunn [5]— es muy posible que los alumnos que obtienen notas más altas en matemáticas la consigan porque se les está enseñando en la forma que mejor va con su estilo peculiar. Y si los profesores de matemáticas cambiaran sus estrategias instructivas para acomodarlas a los estilos de los alumnos con calificaciones más bajas, es muy probable que disminuyera el número de éstos".

Estas reflexiones iniciales, seleccionadas de forma intencionada, pretenden centrar y al mismo tiempo justificar el origen y motivación de este trabajo y que hay que buscarlo en el tema apasionante de los Estilos de Aprendizaje y los Estilos de Enseñanza, porque permiten enfocar la enseñanza y el aprendizaje desde una perspectiva novedosa, práctica y con innumerables aportaciones al quehacer docente en todos los ámbitos y edades.

2 ¿Qué son los Estilos de Aprendizaje?

Quién no se ha preguntado en alguna ocasión: ¿por qué aprendo mejor en determinadas circunstancias y peor en otras? En general se atribuye a las aptitudes, y en muchos casos, en parte, son los estilos de aprendizaje la causa.

El concepto de Estilo de Aprendizaje es definido de forma muy variada por diversos autores, si bien la mayoría coinciden en que se trata de cómo la mente procesa la información o cómo es influida por las percepciones de cada individuo. Una de las definiciones, que proponen diversos autores (C. Alonso [1], C. Alonso, D. Gallego y P. Honey [2]) y que asumimos, es la siguiente: "Los Estilos de Aprendizaje son los rasgos cognitivos, afectivos y fisiológicos, que sirven como indicadores relativamente estables, de cómo los discentes perciben, interaccionan y responden a sus ambientes de aprendizaje."

P. Honey y A. Mumford [8] prescinden parcialmente de la insistencia en el factor inteligencia, que no es fácilmente modificable, insistiendo en otras facetas más accesibles y mejorables. Clasifican los Estilos de Aprendizaje en cuatro tipos: Activo, Reflexivo, Teórico y Pragmático. Y los describen así:

> *Estilo Activo.* Las personas que tienen predominancia en este estilo se implican plenamente y sin prejuicios en nuevas experiencias. Son

de mente abierta, nada escépticos y acometen con entusiasmo las tareas nuevas. Sus días están llenos de actividad. Se crecen ante los desafíos de nuevas experiencias, y se aburren con los largos plazos. Piensan que por lo menos una vez hay que intentarlo todo. Son personas muy de grupo que se involucran en los asuntos de los demás y centran a su alrededor todas las actividades.

Estilo Reflexivo. A los reflexivos les gusta considerar experiencias y observarlas desde diferentes perspectivas. Reúnen datos, analizándolos con detenimiento antes de llegar a alguna conclusión. Su filosofía consiste en ser prudente. Disfrutan observando la actuación de los demás, escuchan a los demás y no intervienen hasta que se han adueñado de la situación. Crean a su alrededor un aire ligeramente distante y condescendiente.

Estilo Teórico. Los teóricos enfocan los problemas de forma vertical escalonada, por etapas lógicas. Tienden a ser perfeccionistas. Integran los hechos en teoría coherentes. Son profundos en su sistema de pensamiento, a la hora de establecer teorías, principios y modelos. Les gusta analizar y sintetizar. Buscan la racionalidad y la objetividad huyendo de los subjetivo y de lo ambiguo. Par ellos si es lógico son bueno.

Estilo Pragmático. El punto fuerte de las personas con predominancia en estilo pragmático es la aplicación práctica de las ideas. Descubren el aspecto positivo de las nuevas ideas y aprovechan la primera oportunidad para experimentarlas. Les gusta actuar rápidamente y con seguridad con aquellas ideas y proyectos que les atraen. Tienden a ser impacientes cuando hay personas que teorizan. Pisan la tierra cuando hay que tomar una decisión o resolver un problema. Su filosofía es siempre se puede hacer mejor, si funciona es bueno.

C. Alonso [1] añade una serie de características a los cuatro estilos de aprendizaje definidos por P. Honey y A. Mumford [8] anteriormente. Así, divide estas características en dos grupos: características principales (más significativas) y otras características.

Las personas con predominio claro de Estilo Activo poseerán algunas de las siguientes características principales: animador, improvisador, descubridor, arriesgado y espontáneo. Otras características son: creativo, novedoso, aventurero, renovador, inventor, vital, vividor de la experiencia, generador de ideas, lanzado, protagonista, chocante, innovador, conversador, líder, voluntarioso, divertido, participativo, competitivo, deseoso de aprender, solucionador de problemas y cambiante.

Las personas en las que predomine el Estilo Reflexivo tendrán alguna de las siguientes características: ponderado, concienzudo, receptivo,

analítico y exhaustivo. Otras características son: observador, recopilador, paciente, cuidadoso, detallista, elaborador de argumentos, previsor de alternativas, estudioso de comportamientos, registrador de datos, investigador, asimilador, escritor de informes y/o declaraciones, lento, distante, prudente, inquisidor y sondeador.

Entre las características de las personas con un alto grado de Estilo Teórico destacan: metódico, lógico, objetivo, crítico y estructurado. Otras características son: disciplinado, planificado, sistemático, ordenado, sintético, razonador, pensador, relacionador, perfeccionista, generalizador, explorador, inventor de procedimientos y buscador de hipótesis, modelos, preguntas, supuestos subyacentes, conceptos, finalidad clara, racionalidad, "por qué", sistemas de valores...

Mientras que las personas que tengan un predominio en Estilo Pragmático presentan algunas de las siguientes características: experimentador, práctico, directo, eficaz y realista. Otras características son: técnico, útil, rápido, decidido, planificador, positivo, concreto, objetivo, claro, seguro de sí, organizador, actual, solucionador de problemas, aplicador de lo aprendido y planificador de acciones.

3 Propuestas pedagógicas

Cuando un alumno tiene preferencia alta por un determinado Estilo de Aprendizaje conviene reconocer cuándo aprenderá mejor y qué posibles dificultades o inconvenientes presenta. Y, por otra parte, aquellos alumnos con preferencia baja en un determinado Estilo de Aprendizaje, conviene saber cómo reconocerlo, desarrollarlo y fortalecerlo.

El aprendizaje de las matemáticas tiene su "propia" pedagogía. La visión que los estudiantes y profesores tienen acerca de las matemáticas en las situaciones de aprendizaje es muy compleja y diversa. Lo que no admite duda es que los profesores estarán mejor equipados para su tarea si pueden comprender cómo se ven las Matemáticas desde la perspectiva del que aprende.

Se comenzará haciendo un diagnóstico de las ventajas y desventajas que se dan en el Aprendizaje de los estudiantes que muestran una alta predominancia en cada uno de los Estilos. A continuación, se analizarán los posibles bloqueos de tipo cognitivo, afectivo o cultural. El primer paso esencial para el tratamiento de los bloqueos que, en mayor o menor grado, afectan a nuestra personalidad, consiste en conocerlos. Si logramos librarnos de unos cuantos bloqueos en un grado razonable, el progreso de nuestra actividad global mejorará sensiblemente. Y, finalmente, se propondrán una serie de sugerencias pedagógicas para lograr desbloquearlos. Todo ello tomando como referencia permanente la enseñanza de la matemática.

"El trabajo de docente —indica H. Gardner [6]— se parece al de un compositor, que teniendo presente toda la partitura se puede centrar en unos pasajes o unos instrumentos concretos. El docente debe plantear unas preguntas, unidades y ejercicios de comprensión que encajen bien entre sí, debe hacer que

los estudiantes se interesen por el tema y, en última instancia, debe procurar que la inmensa mayoría pueda comprender el tema con profundidad".

3.1 Estilo activo

3.1.1 Predominancia alta

Los estudiantes con predominancia alta en Estilo Activo poseen una serie de preferencias y dificultades (Tabla 1), que indican las situaciones en las que aprenden mejor o se sienten más cómodos y, aquellas otras, en las que se encuentran con dificultades y se muestran más incómodos.

Preferencias	Dificultades
• Intentar cosas nuevas • Resolver problemas • Competir en equipo • Dirigir debates • Hacer presentaciones • No tener que escuchar sentado mucho tiempo • Realizar actividades diversas	• Exponer temas con mucha carga teórica • Prestar atención a los detalles • Trabajar en solitario • Repetir la misma actividad • Limitarse a cumplir instrucciones precisas • Estar pasivo: oír conferencias, explicaciones,... • No poder participar

Tabla 1: Estilo Activo

3.1.2 Bloqueos

Los bloqueos más frecuentes que impiden el desarrollo del Estilo Activo son:

- *Miedos.* Miedo al fracaso, a la equivocación. Experimentar el fracaso y la equivocación en algunas tareas, nos permite aprender también cómo hacer las cosas mejor. Sin embargo, —afirman R.J. Sternberg y L. Spear-Swerling [11]—, unos, que obtienen generalmente resultados bajos, tienen miedo al fracaso porque lo han experimentado demasiadas veces; otros, por el contrario, no han sido capaces de aceptar los fracasos ocasionales como parte normal de su aprendizaje. Existen ocasiones en las que no conviene correr riesgos, pero hay otras en las que hay que hacerlo y la indolencia puede acarrear la pérdida de oportunidades

- *Ansiedades.* La ansiedad ante cosas nuevas preocupa e inquieta.

- *Sentirnos obligados a hacer algo que no queremos.* Puede ser debido al esfuerzo que comporta o porque no vemos qué valor puede tener. Necesitamos experimentar para sentirnos a gusto, además es motivante y favorece el aprendizaje con cierta autonomía y control.
- *Falta de confianza en sí mismo.* Una tendencia excesiva al juicio crítico es un defecto que nos hace desconfiar de nuestras propias capacidades. Muchas veces no nos deja avanzar.
- *Pensar las cosas muy detenidamente.* Un cierto grado de reflexión es necesario. Ahora bien, darle vueltas y más vueltas a las cosas no permite avanzar e impide tomar decisiones.

3.1.3 Sugerencias de propuestas didácticas

Las posibles propuestas didácticas para mejorar el Estilo Activo son:

- *Hacer algo nuevo, algo que nunca se ha hecho antes, al menos de vez en cuando.* Por ejemplo, como señala M. de Guzmán [7], hay que intentar aproximarse a problemas desconocidos, aunque sea con cierto recelo. No sabemos si es fácil o difícil, si estará a nuestro alcance o no. Jugamos con él, cada vez se hace menos hostil. Lo manipulamos, y se hace más amigo, nos da pistas y nos anima a explorarlo.
- *Activar la curiosidad.* La curiosidad —afirma J. Alonso [3]— es un proceso activado por características de la información como su novedad, su complejidad, su carácter inesperado, su ambigüedad y su variabilidad. Es evidente que el profesor capta la atención de los alumnos de esta manera.
- *Practicar la resolución de problemas en grupo.* Este tipo de trabajo requiere de cooperación y diálogo con los compañeros.
- *Cambiar de actividad en la hora de clase.* Hacer el cambio lo más diverso posible. Por ejemplo, después de una exposición breve por parte del profesor o de un alumno, cambiar a una actividad de experimentación (individual o en grupo) como la resolución de ejercicios o problemas, comprobar o verificar propiedades, etc. Es necesario proponer a los alumnos una gran variedad de tareas.
- *Forzarse a uno mismo a ocupar el primer plano.* Ofrecerse voluntario para resolver un ejercicio o para exponer un tema en clase. Cuando se trabaja en grupo, obligarse a hacer de moderador o secretario.
- *Discusión de ideas.* Los alumnos preguntan y responden cuestiones entre ellos, explican sus respuestas o estrategias, sugieren ideas y discuten sobre las mismas.
- *Puesta en común.* Se trata de exponer las conjeturas, los resultados parciales, las ideas más significativas, ofreciendo las explicaciones adecuadas para facilitar la comprensión.

- *Pedir a un estudiante que describa oralmente su proceso de resolución de un problema*, que comunique sus ideas, con ayuda del protocolo realizado.
- *Resolver ejercicios que consistan en la repetición de una determinada técnica* previamente expuesta por el profesor. Es decir, aquellos ejercicios que tienen por finalidad la consolidación y automatización de técnicas.
- *Permitir cometer errores.* Cuando se exploran cosas nuevas es inevitable cometer errores. Pero se debe aprender de ellos. Sin embargo, en los centros se tiende a no perdonarlos y, como consecuencia, se acaba teniendo miedo a errar y, por tanto, a pensar de forma independiente y creativa. La insistencia en respuestas correctas fomenta el conformismo, no la creatividad.
- *Estimular el razonamiento crítico.* El profesor plantea preguntas para estimular el razonamiento y el debate. Fomenta el diálogo entre el profesor y el alumno y de los alumnos entre sí.

3.2 Estilo reflexivo

3.2.1 Predominancia alta

Las preferencias y dificultades de los estudiantes con predominancia alta en Estilo Reflexivo se indican en la Tabla 2, mostrando las situaciones en las que aprenden mejor y, aquellas otras, en las que se encuentran con dificultades.

Preferencias	Dificultades
• Observar y reflexionar • Llevar su propio ritmo de trabajo • Tener tiempo para asimilar, escuchar, preparar • Trabajar concienzudamente • Oír los puntos de vista de otros • Hacer análisis detallados y pormenorizados	• Ocupar el primer plano • Actuar de líder • Presidir reuniones o debates • Participar en reuniones sin planificación • Expresar ideas espontáneamente • Estar presionado de tiempo

Tabla 2: Estilo Reflexivo

3.2.2 Bloqueos

Los bloqueos más frecuentes que impiden el desarrollo del Estilo Reflexivo son:

- *Carecer de tiempo suficiente para planificar y pensar.* Dejar tiempo para la reflexión es fundamental. Pero si no tenemos la oportunidad de pensar en lo que estamos haciendo y de reflexionar en lo que ha ido bien, lo que ha ido mal y por qué, las oportunidades de mejorar a largo plazo serán escasas.

- *Obligación de cambiar rápidamente de actividad.* Cambiar de actividad exige un gran esfuerzo de voluntad, de decisión. Pero en este mundo que nos ha tocado vivir las personas que aprenden a enfrentarse al cambio están más preparadas para sobrevivir y prosperar.

- *Impaciencia.* La impaciencia es falta de paz, de tranquilidad, ir con prisas. Quien asiduamente se enfrenta a problemas semejantes a los que le proponen, a su ritmo, con tranquilidad, será capaz de enfrentarse a problemas a plazo fijo, a tomar decisiones con inmediatez. En cualquier caso la prisa siempre es mala consejera.

- *La falta de control.* Algunos estudiantes son capaces de realizar trabajos académicos excelentes, pero sus aptitudes no están desarrolladas debido a la tendencia que tienen a trabajar de manera impulsiva e irreflexiva. Las mejores soluciones suelen obtenerse después de un período de reflexión

- *La falta de orientación hacia el producto.* Algunos están muy preocupados por el proceso mediante el que se hacen las cosas, pero no tanto por el resultado. En general y desgraciadamente, nos juzgarán fundamentalmente por el resultado.

3.2.3 Sugerencias de propuestas didácticas

Las posibles sugerencias de propuestas didácticas para mejorar el Estilo Reflexivo son:

- *Practicar la manera de escribir con sumo cuidado.* Escribir un enunciado de un teorema, una demostración, el desarrollo de un ejercicio o problema.

- *Salir a la pizarra a resolver un problema o a realizar una tarea.* Hay alumnos que nunca se ofrecen voluntarios para esta actuación, sobre todo por miedo a equivocarse. Debe, pues, fomentarse la participación en el aula como una actividad regular y procurar que genere satisfacción personal.

- *Elaborar protocolos.* Se trata de registrar de forma ordenada todo lo que ha sucedido a lo largo del proceso de resolución de un ejercicio o problema, una demostración de un teorema.

- *Recoger información mediante la observación.* Por ejemplo, escribiendo toda la información posible que se extraiga de una presentación de modo gráfico (tablas, diagramas, gráficos en general,...) realizada por parte del profesor o de otro alumno.

- *Comunicar información mediante expresión oral.* Por ejemplo, explicación oral y justificada del proceso seguido en la resolución de problemas.
- *Investigar, añadir información nueva a la ya existente.* Se trataría de todos aquellos procedimientos relacionados con la búsqueda, recogida y selección de información necesaria para definir y plantear un determinado problema y, después, resolverlo. A modo de ejemplo, la búsqueda en textos, revistas o en bases de datos, de información estadística.
- *Dejar tiempo para pensar de forma creativa.* Somos una sociedad con prisas. Necesitamos tiempo para pensar un problema, desmenuzarlo y producir una solución creativa. Por tanto, se debe dejar suficiente tiempo en los deberes y en los exámenes. Desgraciadamente, en muchas ocasiones, tanto los profesores como los estudiantes no tenemos tiempo para pensar, y mucho menos para pensar de forma creativa. Hay que dar tiempo para que se haga.
- *Observar como imitación interior.* El alumno que observa a su profesor mientras éste explica una lección o realiza un ejercicio, le imita interiormente. La observación de una actividad suele ser útil para su posterior realización independiente.
- *Captación matemática de un proceso.* La captación de un desarrollo matemático por parte del profesor requiere la actividad del intérprete (alumno). Esto es, no basta la explicación del profesor, es necesaria la participación activa del alumno.
- *A toda acción práctica debe seguir una fase de reflexión.* Los alumnos razonan sus propuestas de solución, formulan sus reflexiones. El profesor procura que se escuchen mutuamente y entiendan lo que sus compañeros dicen. Oye sus reflexiones, ayuda a interpretarlas y las hace comprensibles para los alumnos; destaca las ideas importantes; expresa de nuevo lo que los estudiantes han expuesto con vaguedad; repite varias veces lo importante.
- *La alegría de conocer.* Experimentar la alegría solucionando problemas, reconociendo su claridad y belleza, es fundamental para el trabajo en matemáticas.
- *El principio de la ayuda mínima.* El profesor observará lo que el grupo de clase es capaz de hacer por sí mismo, de una forma autónoma. Paulatinamente irá tomando la dirección, guiará hacia los conocimientos que considere esenciales. Hasta el final no mostrará a los alumnos la respuesta.
- *Activar y mantener el interés.* Para mantener la atención del alumno centrada en el desarrollo de una explicación o en la realización de una tarea, se debe conectar lo que el alumno sabe y lo que el profesor va diciendo. Para ello, J. Alonso [3] señala las siguientes estrategias:

a) Activar los conocimientos previos al comenzar la clase (objetivos planteados, razones por las que se tratan de conseguir y principales puntos a tratar) que conducirán a una curiosidad, estimularán el recuerdo de lo que se sabe, e incluso, a la búsqueda de nueva información sobre el tema.

b) Utilizar ilustraciones y ejemplos. El uso frecuente de ilustraciones y ejemplos son recursos importantes para mantener el interés.

- *Exposición oral del profesor*. El profesor se encarga de presentar la materia que hay que aprender. Su utilización óptima es para presentar información nueva.

3.3 Estilo téorico

3.3.1 Predominancia alta

Se indican en la Tabla 3 las situaciones en las que aprenden mejor y en las que se encuentran con dificultades, los estudiantes con predominancia alta en Estilo Teórico.

Preferencias	Dificultades
• Sentirse en situaciones claras y estructuradas • Participar en sesiones de preguntas y respuestas • Entender conocimientos complicados • Leer u oír hablar sobre ideas y conceptos bien presentados • Leer u oír hablar sobre ideas y conceptos que insistan en la racionalidad y la lógica • Tener que analizar una situación completa	• Verse obligado a hacer algo sin un contexto o finalidad clara • Tener que participar en situaciones donde predominen las emociones y los sentimientos • Participar en actividades no estructuradas • Participar en problemas abiertos • Verse, por la improvisación, ante la confusión de métodos o técnicas alternativas

Tabla 3: Estilo Teórico

3.3.2 Bloqueos

Los bloqueos más frecuentes que impiden el desarrollo del Estilo Teórico son:

- *Dejarse llevar por las primeras impresiones.* Visión estereotipada que consiste en ver, ante una situación determinada, solamente lo que esperamos ver. Es necesario permanecer abierto a lo extraño, a las desviaciones de lo que aparentemente se espera ver.

- *Preferir la intuición y la subjetividad.* La rigidez en la utilización de diversos procesos de pensamiento constituye un tipo importante de bloqueo. La rigidez mental impide la flexibilidad de pensamiento necesaria para cambiar estrategias o modificarlas.

- *Desagrado ante enfoques estructurados y organizados.* Todos sentimos en alguna ocasión en nuestro trabajo intelectual un cierto rechazo hacia algunas de las tareas que nos vemos obligados a llevar a cabo. En unos casos sentimos rechazo porque encontramos la tarea aburrida, rutinaria, opaca. En otros casos nos resulta la actividad antipática porque nos resulta extraña, no familiar, no connatural a nuestra forma espontánea de proceder (M. de Guzmán [7]).

- *La dependencia excesiva de los demás* (profesor y compañeros). Muchos estudiantes confían en que, o bien los demás les solucionen los problemas, o bien les expliquen de forma permanente cómo afrontarlos, ya que, sin esa ayuda, se encuentran totalmente perdidos.

- *Preferencia por la espontaneidad y el riesgo.* Asumir riesgos sensatos y estimular a los otros a asumirlos es beneficioso. Señala R.J. Sternberg [9], que se debe valorar la creatividad de los estudiantes a la hora de llevar a cabo una práctica o un proyecto.

- *Incapacidad de convertir el pensamiento en acción.* No basta con tener buenas ideas, sino también la capacidad de ponerlas en práctica, trasladar el pensamiento a la acción. Esto es, hacer Matemáticas.

- *Incapacidad para terminar y llevar a cabo los trabajos.* Algunas personas son incapaces de llegar hasta el final, cualquier cosa que empiezan no son capaces de finalizarla. Se enredan en cualquier paso intermedio.

3.3.3 Sugerencias de propuestas didácticas

Las posibles propuestas de sugerencias didácticas para mejorar el Estilo Teórico son:

- *Leer atentamente y de forma pausada un teorema, una proposición, una propiedad o el enunciado de un problema.* Después tratar de resumir lo que se ha leído, diciéndolo con palabras propias.

- *Tomar una situación compleja y analizarla.* Por ejemplo, dado un problema novedoso buscar las posibles relaciones con otros que se tengan almacenados en la memoria de tal forma que la información inicial se transforme en otra información que permita obtener su solución. O de

otra manera, decodificar la información, es decir, traducir la información inicial a un nuevo código o lenguaje con el que el alumno esté familiarizado y le permita conectar la información nueva con las ya existentes.

- *Prever contratiempos y preparase para resolverlos.* Debemos ser optimistas siempre que sea posible; el pesimismo agota la energía, mina el empuje. Deberíamos aprender a ver los contratiempos como oportunidades de aprendizaje y no como causas de desesperación.

- *Resumir teorías e hipótesis, formular y comprobar conjeturas.* El profesor debe recompensar explícitamente los esfuerzos creativos de los estudiantes, además del conocimiento, habilidades analíticas y la redacción.

- *Practicar la manera de hacer preguntas.* M. de Guzmán [7] considera la pregunta como una actitud y señala: "la pregunta es como un anzuelo para extraer ideas originales. El esfuerzo consciente por preguntarse y preguntar genera una actitud inquisitiva, que es la base de todo progreso en el conocimiento".

- *Cuestionar los supuestos.* R.J. Sternberg [9] afirma que todo pensamiento creativo comienza con una pregunta: "¿por qué?" Los profesores debemos estimular a los alumnos a que cuestionen los supuestos.

- *Adquirir experiencia.* En el caso de la aplicación rígida de algoritmos matemáticos, suele ser útil crear situaciones donde los estudiantes deban pensar como el matemático que ideo el algoritmo e intenten por su cuenta desarrollarlo de nuevo.

- *La codificación selectiva.* Supone separar la información relevante de la irrelevante.

- *La perseverancia.* Algunos estudiantes se dan por vencidos con demasiada facilidad. Si en los primeros intentos no tienen éxito, abandonan. La perseverancia es imprescindible en la realización de un ejercicio de Matemáticas

- *Formulación algebraica.* El alumno debe dotar a las fórmulas y a las frases de sentido. Tiene que poder explicarla, justificarla en su lenguaje. Con ello demuestra que los signos son para él portadores de significado.

- *Aprender de memoria y automatizar.* En Matemáticas hay que hacer ejercicios y aprender frases de memoria. La finalidad es su automatización. Algunas fórmulas, determinados enunciados y reglas, hay que aprenderlas de memoria

- *Aplicar los conceptos.* Hay que dar ocasión a los alumnos de emplear los instrumentos que han adquiridos. Por ejemplo, si se trata de un problema que hay que resolver de forma autónoma, debe preguntarse dónde cree que existen las aplicaciones prácticas y teóricas de los conceptos estudiados.

3.4 Estilo Pragmático

3.4.1 Predominancia alta

Las preferencias y desventajas que presentan los estudiantes con predominancia alta en Estilo Pragmático figuran en la Tabla 4.

Preferencias	Dificultades
• Aprender técnicas inmediatamente aplicables • Percibir muchos ejemplos y anécdotas • Experimentar y practicar técnicas con asesoramiento de un experto • Recibir indicaciones prácticas y técnicas	• Aprender cosas que no tengan una aplicabilidad inmediata • Trabajar sin instrucciones claras sobre cómo hacerlo • Considerar que las personas no avanzan con suficiente rapidez

Tabla 4: Estilo Pragmático

3.4.2 Bloqueos

Los bloqueos más frecuentes que impiden el desarrollo del Estilo Pragmático son:

- *Considerar las técnicas útiles exageradas.* Contemplación, abstracción, especulación, por ejemplo, no son actividades mentales muy de moda para los prácticos. Sin embargo, de ellas han dependido fundamentalmente los grandes avances del pensamiento humano, incluso en las ciencias.

- *No saber para qué sirve lo que se estudia* puede resultar desmotivante. Los estudiantes, en general, prefieren trabajar en algo que resulte útil, que no en algo que no se sabe para qué sirve. Sin embargo, en innumerables ocasiones la aplicabilidad no es inmediata, hay que ir subiendo peldaños paso a paso hasta ver el horizonte práctico.

- *Dejar los temas abiertos.* En la fase inicial de un determinado problema concédete la oportunidad de volar libremente, déjate llevar por conjeturas imaginativas, por tu fantasía, todo ello por encima de planteamientos lógicos. Ya vendrá el rigor (M. de Guzmán [7]).

- *La distracción y la falta de concentración.* Hay personas que se distraen con mucha facilidad y suelen tener breves lapsos de atención y, como

consecuencia de ello, no suele cundirles mucho. El profesor debe proporcionar a sus alumnos un ambiente adecuado para trabajar y animarles a lograr sus objetivos (R.J. Sternberg y L. Spear-Swerling [11]).

3.4.3 Sugerencias de propuestas didácticas

Las posibles propuestas de sugerencias didácticas para mejorar el Estilo Pragmático son:

- *Llevar a cabo la corrección de ejercicios y la posterior autoevaluación.*
- *Recabar ayuda de personas que tienen experiencia.* M. de Guzmán [7] indica que el experto y el aprendiz se manifiestan ante un problema difícil de forma muy distinta; el experto manifiesta una mayor intuición y flexibilidad para abandonar un camino equivocado, mientras que el aprendiza suele presentar cierta inmovilidad de pensamiento.
- *Aprender del maestro.* En la relación entre el maestro y el aprendiz, el maestro aborda y plantea un problema nuevo y hace que el principiante intervenga en su resolución. De esta manera, el aprendiz presencia muchos ejemplos de la aplicación adecuada, y dispone de numerosas ocasiones para poner en práctica su propia comprensión (H. Gardner [6]).
- *Experimentar y observar.* La experimentación es una de las técnicas más fructíferas para el descubrimiento y la resolución de problemas. De la observación surge una conjetura, se sigue experimentando y se contrasta.
- *Estudiar las técnicas que utilizan otras personas.* Cuando se descubra que algo hacen bien, imitarlos. El profesor debe actuar de "entrenador" en el sentido de que, al principio y en multitud de ocasiones, mostrará las habilidades y las técnicas que, posteriormente, el alumno utilizará de forma estratégica en la resolución de ejercicios y problemas.
- *Recibir información de una actuación en clase.* Después de una intervención en clase, una presentación o en la realización de un ejercicio, recibir información de cómo se ha hecho.
- *Ejercitar.* Plantear problemas que tengan como finalidad la utilización de las distintas técnicas, algoritmos y destrezas matemáticas en contextos distintos de los que se han aprendido y enseñado.
- *Utilizar imágenes.* Muchos ejercicios y problemas se hacen más asequibles cuando se utiliza una representación adecuada de los elementos que en ellos intervienen. Se piensa generalmente mejor con el apoyo de imágenes que con palabras, números, símbolos, y fórmulas.
- *Crear "entornos de aprendizaje asistidos por ordenador".* Los estudiantes pueden investigar cualquier tema de interés por su cuenta o en colaboración con otros compañeros. Intercambian información, se

comunican con estudiantes de otros lugares y también pueden consultar con expertos a través de Internet.

4 Reflexión conlusiva

A lo largo de este trabajo se ha pretendido aportar una serie de propuestas didácticas en el quehacer diario de la clase de matemáticas en bachillerato, pero trasladable en muchas de sus cuestiones a la universidad o, al menos, a la universidad que yo deseo. Evidentemente y como se puede comprobar, no figuran ejemplos matemáticos concretos para lograr los objetivos deseados, y eso es así porque los ejemplos concretos que sirven a cada profesor depende del momento, del lugar, de los alumnos, del centro, del curso y de muchos otros factores específicos de la didáctica de la matemática. Pero eso no ha sido lo que se ha pretendido aquí. Lo que verdaderamente se ha perseguido es "vivir mejor el aprendizaje, vivir mejor la enseñanza", aspectos mucho más profundos y más formativos que acompañan a la personalidad del docente.

Como ya se ha señalado anteriormente, tanto el aprendizaje como la enseñanza son por fortuna para los que nos dedicamos a esta profesión procesos dinámicos, si bien los cambios se van produciendo muy lentamente. Nuestra labor docente acumulada, la experiencia, hace que cualquier cambio o transformación de nuestro papel y actuación en el aula, por ínfima que parezca, sea vivido con una cierta convulsión. Pero debemos obligarnos a estar permanentemente evolucionando y aprendiendo, con el fin de incorporar a nuestra tarea todo aquello que de positivo nos puedan aportar otros compañeros o la lectura e investigación de diversos textos. Debemos estar "vivos" en el aula.

La enseñanza es un arte para el que hay que poseer unas cualidades innatas que, adornadas de técnicas, entusiasmo y alegría, permiten disfrutarlo y transmitirlo. Convertir lo oscuro en claro y lo complejo en simple. Más que un programa es una filosofía. En resumidas cuentas nuestra actitud en el aula refleja con cierta precisión nuestra actitud ante la vida. No transmitimos lo que queremos ser sino lo que somos. No enseñamos con nuestras palabras sino con nuestros hechos.

En la enseñanza se debe aprender día a día. Aprendemos de nuestros alumnos y ellos de nosotros. No podemos enseñar de forma idéntica a como nos enseñaron, ni podemos ni debemos actuar en el aula como actuaban los que nos enseñaron hace veinte o treinta años. Es tan importante formarse en la materia específica como en pedagogía y psicología. Es tan importante conocer la asignatura que se explica como interesarse por los acontecimientos de la vida actual en sus múltiples facetas.

Referencias

[1] C. Alonso. *Estilos de aprendizaje: análisis y diagnóstico en estudiantes universitarios.* Universidad Complutense, 1992.

[2] C. Alonso, D. Gallego y P. Honey. *Los estilos de aprendizaje.* Mensajero, 1995.

[3] J. Alonso. *Motivar para el aprendizaje.* Edebé, 1997.

[4] R. Díez. *Aprender para el futuro. Nuevo marco de la tarea docente.* Fundación Santillana, 1998.

[5] R. Dunn y K. Dunn. *La enseñanza y el estilo individual de aprendizaje.* Anaya, 1984.

[6] H. Gardner. *La educación de la mente y el conocimiento de las disciplinas.* Paidós, 2000.

[7] M. de Guzmán. *Para pensar mejor.* Labor, 1991.

[8] P. Honey y A. Mumford. *The Manual of Learning Styles.* Maindehead, Berkshire: P. Honey, Ardingly House, 1986.

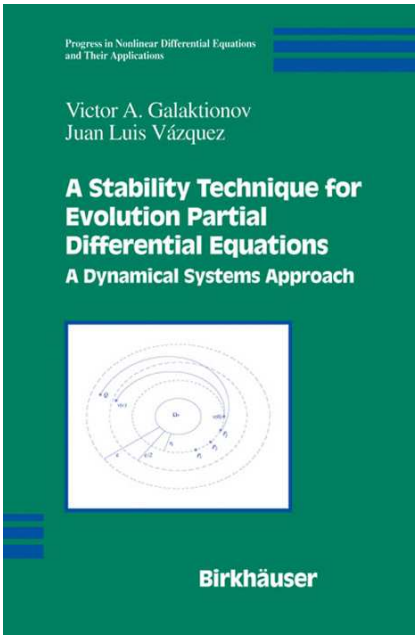
[9] R.J. Sternberg. *Inteligencia exitosa.* Paidós, 1997.

[10] R.J. Sternberg. *Estilos de pensamiento.* Paidós, 1999.

[11] R.J. Sternberg y L. Spear-Swerling. *Enseñar a pensar.* Santillana, 2000.

A Stability Technique for Evolution Partial Differential Equations. A Dynamical System Approach
Viktor A. Galaktionov y Juan Luis Vázquez
Birkhäuser, Boston, 2004.
ISBN: 0-8176-4146-7 (377 páginas)

Por Shergei Chmarev



El libro presenta un nuevo método de análisis del comportamiento asintótico de las soluciones de ecuaciones en derivadas parciales de evolución. El método es el "estado-del-arte"de la teoría contemporánea de las ecuaciones no lineales en derivadas parciales y es fruto de más de diez años de colaboración de los autores, expertos reconocidos a nivel mundial en esta rama de matemáticas. Este método es aplicable a distintos problemas de evolución que se formulan en términos de ecuaciones en derivadas parciales del tipo parabólico o hiperbólico y que pueden incluir términos de primer o segundo orden u operadores de órdenes superiores. Además, también es aplicable a los problemas con fronteras libres y a otros tipos de ecuaciones y sistemas.

La teoría desarrollada esta basada en novedosos resultados sobre la estabilidad de soluciones de sistemas dinámicos de dimensión infinita. Se considera una ecuación diferencial no autónoma de evolución del tipo

$$u_t = \mathbf{A}(u) + \mathbf{C}(u,t), \quad t > 0,$$

donde u pertenece a un espacio de Banach, $\mathbf{A}$ es un operador autónomo (independiente del tiempo) y $\mathbf{C}$ es una perturbación asintóticamente pequeña, $\mathbf{C}(u(t),t) \to 0$ cuando $t \to \infty$ (en un sentido débil) a lo largo de las órbitas $\{u(t)\}$. Se supone que el comportamiento asintótico de la ecuación autónoma,

$u_t = \mathbf{A}(u)$, es conocido. El resultado principal es el Teorema de Estabilidad que afirma que para largos tiempos las órbitas de la ecuación no autónoma convergen a una clase límite de la ecuación autónoma. Lo típico para los métodos estándar es que la condición de estabilidad se imponga sobre la ecuación original. En cambio, en el método presentado en el libro solo se exige que sea estable la ecuación limite no perturbada.

El primer capítulo del libro esta dedicado a la demostración del Teorema de Estabilidad, que sirve luego como herramienta principal en el estudio de las distintas ecuaciones no lineales de evolución. Este capítulo contiene abundante información sobre los métodos y resultados conocidos en la actualidad, muchos de ellos obtenidos por los autores del libro.

El resto del libro está dedicado al estudio de fenómenos asintóticos intrínsecos de distintos operadores no lineales. Se estudian ecuaciones que surgen en la teoría de difusión-convección-reacción, ecuaciones cuyas soluciones muestran *blow-up*, ecuaciones de Navier-Stokes, ecuaciones de Hamilton-Jacobi y ecuaciones completamente no lineales. En esta colección de aplicaciones del método, se analizan con una especial atención los casos de estudio en los que aparecen situaciones críticas o se revela la formación de singularidades, y donde fallan otros métodos de análisis no lineal. Aparte de los estudios del comportamiento límite de las soluciones, se presentan diferentes técnicas de estimación a priori para las soluciones de ecuaciones parabólicas no lineales. Cada capítulo finaliza con una sección dedicada a las posibles generalizaciones de los resultados incluidos, y una revisión de la bibliografía existente.

El trabajo es un brillante ejemplo del análisis de una amplia gama de sofisticados fenómenos no lineales basado en una aproximación unificada. Es una fusión de un delicado análisis y de una variedad de aplicaciones cuidadosamente seleccionadas. El libro es recomendable a los investigadores que trabajan en el campo de ecuaciones en derivadas parciales y física matemática y puede ser útil para los alumnos de tercer ciclo. La lista de referencias bibliográficas contiene 329 entradas que, aparte de otros méritos, lo hace también un excelente libro de referencia.