Universidad Rey Juan Carlos

APUNTES DE TEORÍA

MÉTODOS MATEMÁTICOS APLICADOS A LA INGENIERÍA

Métodos matemáticos aplicados a la Ingeniería de Materiales

Métodos matemáticos aplicados a la Ingeniería de la Energía

Métodos numéricos (módulo I) en el Máster en Ingeniería Industrial

E. Schiavi, A. I. Muñoz Montalvo, C. Conde

Septiembre 2022

Publicado en: https://burjcdigital.urjc.es

Índice general

Capítulo 1

Introducción a las ecuaciones en derivadas parciales

1.1. Generalidades.

Todas las operaciones básicas, físicas y químicas de la Ingeniería Química, implican el transporte de una o varias de las tres magnitudes fundamentales siguientes: cantidad de movimiento, energía y materia. Estos fenómenos de transporte se producen en el seno de los fluidos o entre sólidos y fluidos, bien como consecuencia de las diferencias de concentraciones de dichas magnitudes en aquellos (representando la tendencia de todos los sistemas a alcanzar el equilibrio) bien como consecuencia del movimiento de los propios fluidos. Para poder diseñar adecuadamente los equipos e instalaciones donde hayan de desarrollarse las operaciones indicadas, se requiere una información precisa sobre los caudales de transporte de las citadas magnitudes físicas[1]. El análisis de los fenómenos de transporte nos conduce *naturalmente* a la consideración de las ecuaciones de conservación de la masa (ecuación de continuidad), de la cantidad de movimiento (ecuación del equilibrio) y de la energía. Tales ecuaciones suelen ser de un tipo determinado, llamado ecuaciones en derivadas parciales o, más brevemente, EDP[2].

En este curso desarrollaremos las técnicas matemáticas básicas que permiten abordar tales problemas desde los puntos de vista analítico (exacto)

[1] Los tres fenómenos de transporte se producen casi siempre simultáneamente en todas las operaciones básicas, tanto físicas como químicas, pero la importancia relativa de los mismos varía en cada caso, siendo frecuente que sólo uno de ellos, por su mayor lentitud, determine el tamaño del equipo necesario para el desarrollo de dichas operaciones. Véase el libro de Costa Novella, Vol 2 (Fenómenos de transporte).

[2] Usaremos indistintamente la sigla EDP para indicar una ecuación en derivadas parciales o un conjunto de ecuaciones en derivadas parciales.

y numérico (aproximado). Tras haber introducido la notación y definiciones básicas para los sucesivos tratamientos, y puesto que en el análisis del flujo (transporte) de fluidos (líquidos o gases) las ecuaciones del movimiento (vectorial), continuidad (escalar) y energía (escalar) se pueden combinar en una única ecuación vectorial de conservación (véase la sección dedicada a los sistemas hiperbólicos de leyes de conservación) empezaremos este tema con el análisis de las ecuaciones hiperbólicas cuasilineales de primer orden lo que nos permitirá, mediante la interpretación dinámica (en términos de la mecánica de fluidos) del proceso de resolución, entrar rápidamente en materia relacionando los fenómenos físicos y su formulación matemática rigurosa, adquirir una cierta soltura en el manejo de los operadores de derivación parcial, introducir el concepto de características, ver su aplicación e interpretación geométrica, familiarizarnos con los problemas de valor inicial (PVI) y los problemas de valor inicial y de contorno (PVIC) entendidos como procedimientos para seleccionar entre las infinitas soluciones posibles de una EDP aquella con sentido físico[3]. Digamos que relacionaremos rápidamente las técnicas matemáticas con la problemática físico-química a estudiar creando un vínculo entre nuestra asignatura y las asignaturas del área físico-química donde el formalismo matemático es conveniente y, a veces, necesario.

1.1.1. Notación y Conceptos fundamentales. Ejemplos

Siguiendo la notación introducida en el curso de Elementos de Matemáticas, denotaremos las derivadas parciales de una función de varias variables, digamos $u(x,t)$, en las formas

$$u_x = \frac{\partial u}{\partial x}, \qquad u_{xx} = \frac{\partial^2 u}{\partial x^2}, \qquad u_{xt} = \frac{\partial^2 u}{\partial x \partial t}, \qquad u_t = \frac{\partial u}{\partial t}.$$

Utilizaremos indistintamente ambos tipos de notación aunque observamos que al trabajar con campos de velocidades (que aparecen en los problemas de fluidodinámica) la notación con el subíndice puede generar algo de confusión y no es recomendada. En los libros de texto se suele, en efecto, denotar $\boldsymbol{v} = (v_x, v_y, v_z)$ donde v_x indica la componente horizontal de la velocidad y no su derivada parcial. En este caso es conveniente utilizar la notación extendida de operadores $\partial/\partial x$ pues, de lo contrario, al calcular la aceleración de una partícula en un campo de fuerzas tendríamos $a = v_{x_x}$ lo que resultaría bastante incomprensible. Una notación alternativa, suficientemente clara, sería $a = (v_x)_x$. Pasamos ahora a describir las regiones del espacio donde ha de satisfacerse una EDP. Denotaremos por $\Omega \subset I\!R^n$, $n = 1, 2, 3$ (ocasionalmente

[3]No nos referimos aquí a las soluciones de entropía, para las cuales se necesitan fundamentos matemáticos muy avanzados sino a la búsqueda de soluciones que cumplan unas condiciones de contorno y/o iniciales obtenidas mediante experimentos y mediciones.

se utilizará la letra N para indicar la dimensión espacial) a una región (dominio) abierta del espacio n-dimensional, que eventualmente podrá ser infinita (no acotada) y coincidente con todo el espacio: $\Omega \equiv I\!R^n$. El símbolo $\equiv$ denota igualdad entre conjuntos. También se usa para indicar que una función es, por ejemplo, idénticamente nula en una región, en la forma $f \equiv 0$. El símbolo $\doteq$ denotará igualdad *por definición* y es el análogo del operador de asignación $:=$ que aparece en muchos programas informáticos (por ejemplo en Maple).

Si Ω es una región acotada[4] entonces su frontera, denotada por $\partial\Omega$, se supondrá suficientemente regular para que el análisis posterior tenga validez [5]. En general la condición de regularidad que pediremos será la existencia de un vector normal en todo punto de $\partial\Omega$. Denotaremos además por $\bar{\Omega} = \Omega \cup \partial\Omega$ al cierre del dominio Ω. Cuando una de las variables independientes representa el tiempo se suele introducir otra notación (y nomenclatura) que utilizaremos al trabajar con problemas de evolución (es decir donde el estado del sistema evoluciona con el tiempo).

Otras notaciones típicas, del tipo $(x_1, x_2) = (x, y)$ y $(x_1, x_2, x_3) = (x, y, z)$ (para denotar puntos del plano o del espacio en coordenadas cartesianas) o $\boldsymbol{F}, \vec{v}$ (para denotar vectores) y $[A]$ para denotar matrices, serán también utilizadas.

Conceptos y definiciones

Entramos ahora en materia considerando la definición general de una ecuación en derivadas parciales.

Definición 1.1.1 *Se llama ecuación diferencial en derivadas parciales a una ecuación de la forma:*

$$F\left(x_1, x_2, ..., x_n, u, \frac{\partial u}{\partial x_1}, ..., \frac{\partial u}{\partial x_n}, ..., \frac{\partial^m u}{\partial x_1^{k_1} \partial x_2^{k_2} ... \partial x_n^{k_n}}\right) = 0, \qquad (1.1)$$

que relaciona las variables independientes $x_1, x_2, ..x_n$ ($n > 1$), la variable dependiente $u = u(x_1, x_2, ..., x_n)$ y sus derivadas parciales hasta el orden m, siendo m un número entero tal que $m \geq 1$. Los superíndices $k_1, k_2, .., k_n$ son números enteros no negativos tales que $k_1 + k_2 + ... + k_n = m$.

[4] Recuérdese que la condición necesaria y suficiente para que un conjunto esté acotado es que exista una bola cerrada que lo contenga. Véase el tema 7 de los guiones del primer curso.

[5] Puede parecer excesivo todo este tipo de precisiones sobre el dominio y su frontera pero es conveniente observar que muchos resultados y técnicas de análisis matemático están condicionados a la geometría del dominio en consideración. Su frontera tiene también un papel fundamental y la teoría de trazas que aparece en el cálculo variacional lo confirma. Para la definición exacta y una clasificación muy rigurosa de los dominios dependiendo de su frontera se puede consultar el libro de Adams (véase la referencia en la sección final de este capítulo dedicada a la bibliografía avanzada).

Ejemplo 1.1.1 *Típicos ejemplos de ecuaciones diferenciales en derivadas parciales pueden ser:*

$$1)\quad u_{xx}+u_{yy}=0,\quad 2)\quad u_t=\alpha^2 u_{xx},\quad 3)\quad u_{tt}=-\beta^2 u_{xxxx},\quad 4)\quad u_t=i\beta u_{xx},$$

(siendo α, β constantes reales e i la unidad imaginaria). También:

$$5)\quad u_{xx}+u_{yy}+\alpha^2 u=0,\qquad 6)\quad u_t+\alpha u u_x+\beta u_{xxx}=0,\qquad 7)\quad u_{xx}+x u_{yy}=0$$

En todos los casos se trata de ecuaciones en derivadas parciales *famosas* que suelen llevar el nombre de su descubridor. En concreto se trata de la ecuación de Laplace (1), de la ecuación de Fourier (2), de la ecuación de Euler-Bernoulli (3), de la ecuación de Schrodinger (4), de la ecuación de Helmholtz (5), de la ecuación de Korteweg-de Vries (6) y de la ecuación de Tricomi (7). Aparecen en numerosos problemas de la física matemática. Algunas de ellas las aprenderemos a conocer a lo largo del curso; otras, distintas, las iremos descubriendo al entrar más en materia.

Precisaremos ahora lo que se entiende por orden de una EDP (puesto que en los ejemplos anteriores han aparecido diferentes órdenes de derivación parcial en la misma ecuación). Nótese que se trata de una extensión directa de la definición análoga de orden de una EDO.

Definición 1.1.2 *Se llama orden de una ecuación diferencial del tipo (1.1) al mayor de los órdenes de las derivadas parciales que aparecen en la ecuación.*

Las 7 ecuaciones propuestas en el ejemplo son (respect.) de orden 2,2,4,2 y 2,3,2. Una EDP típica de primer orden es la ecuación de advección[6] (o convección):

$$u_t + a u_x = 0,$$

que será analizada más adelante.

Cuando se consideran varias EDP simultáneamente se generan sistemas de ecuaciones en derivadas parciales. Excepto unos pocos casos concretos es extremadamente difícil el estudio de tales sistemas, especialmente si se consideran ecuaciones de distinta naturaleza[7]. Un caso especial lo constituyen los sistemas de EDP de primer orden. Bajo ciertas hipótesis existe una teoría analítica y un

[6] Originariamente el término advección se utilizaba para los fenómenos de transporte de masas de aire y se extendió después su uso al transporte genérico de materia reservándose el término de convección para el transporte de energía. Actualmente se suele abarcar todos estos fenómenos de transporte con el término de convección.

[7] Véase la sección referente a la clasificación de las ecuaciones en derivadas parciales en elípticas, parabólicas e hiperbólicas.

tratamiento numérico de estos sistemas que aparecen, a menudo, trabajando con fenómenos de transporte. Veamos por ejemplo el sistema que modeliza la propagación de una pequeña perturbación en un gas (una onda sonora). Se trata de un modelo matemático para la propagación del sonido.

Ejemplo 1.1.2 *El movimiento unidimensional de un gas, cuya viscosidad es inapreciable, se describe mediante:*

$$\begin{cases} \rho\dfrac{\partial u}{\partial x} + u\dfrac{\partial \rho}{\partial x} + \dfrac{\partial \rho}{\partial t} &= 0, \\ \rho\dfrac{\partial u}{\partial t} + \rho u\dfrac{\partial u}{\partial x} + \dfrac{\partial p}{\partial x} &= \rho F, \end{cases}$$

donde $u(x,t)$ es la velocidad en el punto x y en el tiempo t, $\rho(x,t)$ es la densidad de masa, $p(x,t)$ es la presión y $F(x,t)$ es una fuerza dada por unidad de masa. La primera de estas ecuaciones expresa la conservación de la masa; la otra es la segunda ley de Newton del movimiento.

Consideremos nuevamente la ecuación diferencial en derivadas parciales (1.1) de orden m. La ecuación estará planteada (en el sentido de que se tiene que satisfacer) en una región abierta (eventualmente infinita) de $\mathbb{R}^n$, digamos $\Omega \subset \mathbb{R}^n$, $n \geq 2$. Denotamos además por $C^m(\Omega)$ al conjunto de funciones continuas con derivadas continuas hasta el orden m en la región Ω. Para $m = 1$ se tiene el espacio $C^1(\Omega)$ de funciones continuamente diferenciables (o diferenciables con continuidad).

Introducida la definición de ecuación en derivadas parciales y caracterizado el orden de la misma, veamos lo que se entiende por solución de una EDP.

Definición 1.1.3 *Se llama solución de una ecuación diferencial de orden m del tipo (1.1) en cierta región abierta $\Omega \subset \mathbb{R}^n$ de variación de las variables independientes $x_1, x_2, .., x_n$ a una función*

$$u = u(x_1, x_2, ..., x_n) \in C^m(\Omega),$$

tal que sustituyendo esta función y sus derivadas en la ecuación (1.1) se obtiene una identidad.

Observación 1.1.1 *En la definición anterior no se especifica la regularidad de la solución en la frontera $\partial\Omega$ de la región (abierta) Ω. Lo que generalmente se pide es que la función sea continua o diferenciable en $\partial\Omega$, que admita todas las derivadas parciales que aparecen en la ecuación en el interior de Ω (que es lo mismo que Ω pues es una región abierta por hipótesis) y que se satisfaga la ecuación en el interior de Ω.*

Observación 1.1.2 *La definición anterior caracteriza las soluciones de una EDP de orden m como aquellas funciones de clase $C^m(\Omega)$ que satisfacen la ecuación en todo punto de $\Omega \subset \mathbb{R}^n$, $n = 1, 2, 3$. A este tipo de soluciones se les suele llamar* ***soluciones clásicas*** *pues son soluciones (en el sentido de que es posible realizar las operaciones de derivación parcial que aparecen en la ecuación alcanzando una identidad) y son clásicas pues esta identidad se tiene en todo punto de la región Ω. Este tipo de soluciones son las que han sido objeto de estudio a partir de los primeros trabajos de Fourier, Laplace, Liouville, etc... En realidad es posible demostrar la existencia de un tipo de soluciones, llamadas* ***soluciones débiles****, que satisfacen la ecuación (o una versión modificada de la misma [8]) no en todo punto del dominio sino en una parte digamos* significativa *(en casi[9] todo punto). No entraremos en detalles sobre la teoría (la Teoría de Distribuciones de L. Schwartz) que justifica este tipo de análisis (conocido como análisis funcional) pero avisamos al lector de la existencia de este tipo de soluciones que han sido y son actualmente objeto de atento estudio e investigación. Varios ejemplos de soluciones débiles se encontrarán al final de la parte de este tema dedicada a las ecuaciones en derivadas parciales de primer orden donde consideraremos unos problemas de Cauchy con ecuaciones de tipo lineal o cuasilineal[10] cuya solución no es de clase $C^1(\Omega)$ (y no es, por tanto, una solución clásica en el sentido de la definición (1.1.3) sino débil). Veremos cómo la prescripción de datos iniciales discontinuos o cuya derivada es discontinua puede generar soluciones discontinuas.*

Como ya se ha señalado en la introducción se llaman *ecuaciones diferenciales en derivadas parciales* a aquéllas ecuaciones en las que las funciones desconocidas dependen de más de una variable independiente y aparecen operadores de derivación parcial. El proceso de resolución de una EDP se llama también **integración** de la EDP. En algunos casos la resolución de la ecuación en derivadas parciales es directa entendiendo por ello que se realiza mediante una integración directa. Recordemos, para ello, las fórmulas de integración para derivadas parciales que vienen dadas por

$$\int \frac{\partial f(x,y)}{\partial x} dx = f(x,y) + \phi(y), \qquad \int \frac{\partial f(x,y)}{\partial y} dx = f(x,y) + \psi(x),$$

[8] Se está hablando aquí de las ecuaciones en forma de divergencia a las cuales son aplicables las técnicas del cálculo variacional. Para ello es básica la teoría de distribuciones de L. Schwartz en la que no profundizaremos en este curso.

[9] Es fundamental aquí el concepto de conjunto de medida cero que aparece en la teoría de la medida de Lebesgue.

[10] Más adelante daremos definiciones precisas de todo este tipo de terminología.

siendo $\phi(y)$, $\psi(x)$ funciones derivables arbitrarias. Las fórmulas anteriores se aplican en la integración indefinida. Por ejemplo, si conocemos la derivada parcial primera de una función, digamos $f_x(x,y) = 2xy$, entonces considerando y como constante e integrando en x se tiene:

$$\int 2xydx = 2y \int xdx = yx^2 + \phi(y),$$

luego $f(x,y) = yx^2 + \phi(y)$. Nótese que cualquier función del tipo $f(x,y) = yx^2 + \phi(y)$ satisface $f_x = 2xy$, independientemente de la función ϕ elegida. Es una generalización del concepto de primitivas para funciones de una variable real.

Para la integración definida (es decir especificando los extremos de integración) se tiene:

$$\int_{h_1(y)}^{h_2(y)} \frac{\partial f(x,y)}{\partial x} dx = [f(x,y) + \phi(y)]_{x=h_1(y)}^{x=h_2(y)} = f(h_2(y), y) - f(h_1(y), y).$$

Por ejemplo, si $h_1(y) = 1$, $h_2(y) = 2y$ la integración definida de $f_x(x,y) = 2xy$ nos dará:

$$\int_1^{2y} 2xydx = y \int_1^{2y} 2xdx = [yx^2 + \phi(y)]_{x=1}^{x=2y} = 4y^3 - y = f(2y, y) - f(1, y),$$

siendo $f(x,y) = yx^2 + \phi(y)$. Es una generalización de la Regla de Barrow para funciones de una variable. De forma análoga se tiene la fórmula:

$$\int_{g_1(x)}^{g_2(x)} \frac{\partial f(x,y)}{\partial y} dy = [f(x,y) + \psi(x)]_{g_1(x)}^{g_2(x)} = f(x, g_2(x)) - f(x, g_1(x)).$$

Nótese que la variable de integración no puede aparecer en los límites de integración. Por ejemplo, no tiene sentido escribir $\int_0^x ydx$. Simplemente se introduce una variable *muda* en la forma $\int_0^x yds$ lo que indica claramente que el resultado de la integración es una función de x (el extremo superior del intervalo de integración)

Podemos ahora considerar el siguiente ejemplo:

Ejemplo 1.1.3 *Resolver la ecuación:*

$$\frac{\partial u}{\partial x}(x,y) = y + x.$$

Se trata evidentemente de una EDP de primer orden que es del tipo (1.1) para los valores paramétricos $m = 1$, $n = 2$, $k_1 = 1$, $k_2 = 0$ y la notación $(x_1, x_2) = (x, y)$. Buscamos una solución del tipo $z = u(x, y)$ (dependiente de dos variables) que sea de clase C^1. Integrando directamente (y de manera indefinida) respecto a x, obtenemos:

$$u(x, y) = xy + \frac{x^2}{2} + \phi(y),$$

donde $\phi(y)$ es una función derivable arbitraria de y. Para cada elección concreta de $\phi(y)$ tendremos una única solución, es decir una función $u = z(x, y)$ tal que $u_x = y + x$.

Por ser $\phi(y)$ arbitraria hemos obtenido infinitas soluciones de la EDP que dependen de *una* función arbitraria. La expresión anterior es por tanto la **solución general** de la EDP en cuestión. Nótese que la gráfica de una solución del tipo $z = u(x, y)$ es una superficie. En el caso bidimensional, las superficies que son soluciones de una EDP de primer orden se llamarán **superficies integrales** de la EDP. Matizaremos este concepto en la sección dedicada a la interpretación geométrica del proceso de resolución de una EDP de primer orden.

Ejemplo 1.1.4 *Resolver la ecuación:*

$$\frac{\partial^2 u}{\partial x \partial y}(x, y) = 0.$$

Se trata de una EDP de segundo orden que es del tipo (1.1) para los valores paramétricos $m = 2$, $n = 2$, $k_1 = 1$, $k_2 = 1$ y la notación $(x_1, x_2) = (x, y)$. Buscamos una solución del tipo $z = u(x, y)$ que sea de clase C^2. Integrando respecto a x, obtenemos:

$$\frac{\partial u}{\partial y}(x, y) = \phi(y),$$

donde $\phi(y)$ es una función (continua) arbitraria de y. Integrando ahora respecto a y se obtiene:

$$u(x, y) = \int \phi(y) dy + \phi_1(x),$$

donde $\phi_1(x)$ es una función (derivable) arbitraria de x. O bien, designando

$$\int \phi(y) dy = \phi_2(y)$$

a una primitiva de $\phi(y)$, tendremos finalmente

$$u(x,y) = \phi_1(x) + \phi_2(y),$$

donde $\phi_2(y)$, en virtud de la arbitrariedad (y continuidad) de $\phi(y)$, es también una función arbitraria (y derivable) de y.

Los ejemplos expuestos sugieren por tanto que la solución general de una ecuación en derivadas parciales de primer orden depende de una función arbitraria; la solución general de una ecuación en derivada parciales de segundo orden depende de dos funciones arbitrarias, y la solución general de una ecuación en derivada parciales de orden p probablemente dependerá de p funciones arbitrarias. Estas consideraciones son ciertas, pero deben ser precisadas. De ello se ocupa el teorema de S.V. Kovaléskaia (1850-1891) sobre la existencia y unicidad de la solución del problema de Cauchy asociado a una ecuación en derivadas parciales. Las hipótesis y el enunciado del teorema de Sofya Kovaléskaia [11] desbordan los objetivos de este curso (pues exigen la aplicación del concepto de derivada para funciones de variable compleja y de la teoría de las funciones analíticas [12]) y no serán considerados. Alternativamente desarrollaremos un análisis local del problema de Cauchy asociado a ecuaciones en derivadas parciales de primer orden que nos proporcionará los resultados teóricos de existencia y unicidad de soluciones del problema considerado.

En este tema estudiaremos brevemente sólo los métodos de integración (resolución) de las ecuaciones en derivadas parciales de primer orden, cuya teoría

[11] Véase, por ejemplo el libro de Fritz John, Partial Differential Equations, pag 61, donde se presenta el teorema bajo el nombre de teorema de Cauchy-Kovaléskaia.

[12] La condición de analiticidad de una función en un punto introduce una restricción muy fuerte a una función. Implica en efecto la existencia de todas las derivadas de orden superior en un entorno del punto y esto garantiza la existencia de una serie de potencias convergente que representa la función en un entorno del punto. Esto está en marcado contraste con el comportamiento de las funciones reales, para las que es posible que exista la derivada primera y que sea continua sin que por ello se pueda deducir la existencia de la derivada segunda. Las funciones analíticas tienen un papel muy importante en la resolución de problemas de flujos bidimensionales a través de la función de corriente. Una introducción bastante sencilla a la teoría de las funciones analíticas se puede encontrar en el libro de Apostol, T.M., (1982), Análisis matemático. Segunda ed. Editorial Reverté. capítulo 16. Resultados matemáticos avanzados se encuentran en el capítulo 3 del libro de F. John. Una aplicación a los problemas de fluidodinámica se encuentra en el capítulo 2, Vol 5 del libro de Costa Novella dedicado al flujo potencial de un fluido perfecto (o ideal). Nosotros introduciremos paulatinamente este tipo de problemas mediante un ejemplo concreto de flujo potencial que analizaremos según se vayan desarrollando las técnicas matemáticas necesarias para su tratamiento (y comprensión). Una introducción muy interesante al uso de la función de corriente en la resolución de problemas de flujos incompresibles (líquidos) que aparecen en mecánica de fluidos en el caso bidimensional se encuentra en el libro de W. Deen, Analysis of Transport Phenomena, capítulo 5, sección 5.9, pag 239. Aplicaciones avanzadas de la teoría se encuentran en los capítulos 7 y 8 del mismo libro.

está estrechamente ligada a la integración de ciertos sistemas de ecuaciones ordinarias. Las ecuaciones en derivadas parciales de orden mayor se integran por métodos completamente distintos[13] y serán tratadas (en el caso de las ecuaciones de segundo orden) en los temas siguientes (temas 2 y 3). La resolución numérica de las ecuaciones en derivadas parciales de primer y segundo orden se considerará en el tema 6.

Finalizamos esta sección resolviendo un problema muy básico y común de la teoría de campos que aparece en fluidodinámica: la determinación de la función de corriente asociada a un campo de velocidades bidimensional irrotacional.

Calcularemos también el potencial del campo resolviendo así el problema de la determinación de las líneas equipotenciales de un campo escalar de $I\!R^2$ planteado en el tema 12, sección 12.1.4 del guión de la asignatura del primer curso dedicado a la Teoría de Campos. Mediante el conocimiento de las líneas de corriente y de las líneas equipotenciales analizaremos algunas propiedades geométricas interesantes del flujo.

Una aplicación de la teoría de campos

Empezaremos introduciendo el concepto y la definición de la función de corriente asociada a un campo de velocidades.

La función de corriente es un objeto matemático extremadamente útil para resolver problemas de flujo de fluidos incompresibles (líquidos) donde sólo hay dos componentes del campo no nulas y sólo dos coordenadas espaciales. Esto incluye los fluidos con un movimiento plano o los procesos de transporte (flujo) simétricos respecto a un eje. Se trata de una herramienta muy útil para deducir los campos de velocidad y presiones en tales problemas y proporciona además información para la visualización de los patrones de flujo. Detalles sobre la terminología utilizada en esta aplicación se encuentran en los temas de Teoría de Campos, EDO y Sistemas de EDO de los guiones de la asignatura de Elementos de Matemáticas del primer curso. Mayores detalles se pueden encontrar en el libro de Deen[14], pag 239, donde se deducen las ecuaciones que gobiernan (determinan) la función de corriente en los distintos sistemas de coordenadas, cartesianas, cilíndricas y esféricas típicamente utilizados en la

[13] En realidad en los temas 2 y 3 presentaremos algunos métodos de resolución de EDP de segundo orden que se pueden aplicar también a ecuaciones de primer orden. Nos referimos por ejemplo al método de separación de variables o al método de la transformada de Laplace.

[14] W.M. Deen, (1998), Analysis of Transport Phenomena. Oxford University Press.

Ingeniería Química.

Función de corriente

Introducimos la definición de la función de corriente como un campo escalar $\psi : \mathbb{R}^2 \to \mathbb{R}$ cuya gráfica (dada por la superficie de ecuación $z = \psi(x, y)$) verifica un cierto problema geométrico. En este caso se busca la superficie $\psi(x, y)$ tal que sus líneas de nivel (también llamadas líneas de corriente) sean tangentes a un campo de velocidades $\boldsymbol{v}$ dado (es decir conocido). Puesto que el gradiente de esta superficie, $\nabla\psi$, es ortogonal a las líneas de nivel de ψ, para que éstas sean tangentes al campo de velocidades $\boldsymbol{v}$ se tiene que $\nabla\psi$ deberá ser ortogonal al mismo campo luego tendrá que ser de la forma $\nabla\psi = (-v_y, v_x)$ ya que

$$(v_x, v_y) \cdot (-v_y, v_x) = -v_x v_y + v_y v_x \equiv 0,$$

es decir,

$$\nabla\psi \cdot \boldsymbol{v} = \left(\frac{\partial\psi}{\partial x}, \frac{\partial\psi}{\partial y}\right) \cdot (v_x, v_y) = \frac{\partial\psi}{\partial x} v_x + \frac{\partial\psi}{\partial y} v_y = 0, \tag{1.2}$$

siendo

$$\begin{cases} \dfrac{\partial\psi}{\partial x} = -v_y, & \dfrac{\partial\psi}{\partial y} = v_x. \end{cases} \tag{1.3}$$

Esta es la definición típicamente utilizada para la función de corriente en un sistema de coordenadas cartesianas rectangulares $(x, y) \in \mathbb{R}^2$ y un campo de velocidades $\boldsymbol{v} : \mathbb{R}^2 \to \mathbb{R}^2$ definido por:

$$\boldsymbol{v}(x, y) = (v_x(x, y), v_y(x, y)).$$

Nótese que sería perfectamente equivalente definir $\nabla\psi = (v_y, -v_x)$ (es decir en la dirección perpendicular pero en el sentido opuesto) puesto que la condición de ortogonalidad $\boldsymbol{v} \cdot \nabla\psi = 0$, sería igualmente cierta.

Conociendo el campo de velocidades $\boldsymbol{v} = (v_x, v_y)$, podremos determinar la función de corriente mediante integración directa de las EDP que aparecen en el sistema (1.3).

La utilidad de la función de corriente en la visualización del flujo de un fluido nace del hecho que ψ (la función de corriente) es *constante a lo largo de una línea de corriente*. Nótese que el operador

$$\boldsymbol{v} \cdot \nabla\psi,$$

mide la tasa de cambio de ψ a lo largo de una línea de corriente y la ecuación $\boldsymbol{v} \cdot \nabla\psi = 0$ nos dice simplemente que la tasa es nula (pues ψ es constante a lo largo de cada una de ellas por definición).

Veremos más adelante que las líneas sobre las que $\psi(x, y)$ es constante deben ser perpendiculares a las líneas equipotenciales (que son las líneas a lo largo de las cuales el potencial $\phi(x, y)$ es constante).

Como aplicación de lo anterior obtendremos un posible método para el cálculo de las *líneas de campo* de un campo vectorial, por ejemplo de un campo de velocidades $\boldsymbol{v}$.

Recordamos que las líneas de campo se interpretan geométricamente como las curvas que en todos sus puntos son tangentes al vector de campo $\boldsymbol{v}$ (en todo instante t) y tienen una interpretación física muy clara. Se trata en efecto de un problema típico que aparece en mecánica de fluidos y fenómenos de transporte (fluidodinámica) donde las líneas de campo se llaman también **líneas de corriente** o **trayectorias**. Si el campo de velocidades es estacionario la trayectoria que seguiría una partícula puntual de fluido depositada en un punto coincide con la línea trazada por la corriente del fluido y esto justifica la terminología adoptada. Establecida la analogía entre los conceptos de líneas de campo y líneas de corriente (para campos estacionarios éstas últimas no son otra cosa que las primeras consideradas en el contexto de la mecánica de fluidos), podemos ahora considerar el siguiente ejemplo:

Ejemplo 1.1.5 *Se considera el campo de velocidad bidimensional estacionario* $\boldsymbol{v} = (v_x, v_y)$ *dado por*

$$v_x = x, \qquad v_y = -y.$$

Determinar las líneas de corriente del campo. Determinar, si existe, el potencial del campo.

Abordaremos el problema planteado en el ejemplo 1.1.5 resolviendo un sistema de ecuaciones diferenciales en derivadas parciales de primer orden para la función de corriente asociada al campo. Estas ecuaciones nacen a partir de la definición de la función de corriente. De manera similar, es decir, resolviendo un sistema de ecuaciones diferenciales en derivadas parciales de primer orden para la función potencial asociada al campo dado (que es irrotacional) determinaremos el potencial del campo.

Cálculo de las líneas de corriente mediante la función de corriente.

Utilizando la definición de la función de corriente dada en (1.3) (válida para

coordenadas rectángulares con $v_z = 0$ y ninguna dependencia en la variable z) se tiene:

$$\begin{cases} \dfrac{\partial \psi}{\partial y} = v_x = x, & \dfrac{\partial \psi}{\partial x} = -v_y = y. \end{cases}$$

Integrando directamente estas ecuaciones diferenciales en derivadas parciales se obtienen las siguientes expresiones para la función de corriente:

$$\psi(x, y) = xy + f(x), \qquad \psi(x, y) = xy + g(y),$$

siendo $f(x)$, $g(y)$, funciones arbitrarias diferenciables. Las dos expresiones corresponden a la misma función, de donde se obtiene que $f(x) = g(y) = C$. El valor de C es arbitrario. La función de corriente es por tanto:

$$\psi(x, y) = xy + C.$$

Si ponemos $C = 0$, tenemos

$$\psi(x, y) = xy,$$

y la ecuación para las líneas de corriente es $\psi(x, y) = xy = K$, siendo $K \in I\!R$. La elección hecha para la constante tiene el efecto de asignar el valor $\psi = 0$ a las líneas de corriente que corresponden a $x = 0$ o $y = 0$ (es decir: los ejes coordenados).

Función potencial

La **función potencial** (o **potencial escalar**) es un concepto que ya hemos introducido en el tema de teoría de campos del guión del primer curso.

Recordamos aquí brevemente la definición de potencial escalar de un campo vectorial. Consideremos nuevamente el campo de velocidades del ejemplo $\boldsymbol{v} = (v_x, v_y)$ dado por:

$$v_x = x, \qquad v_y = -y.$$

En este contexto, la función potencial se llama **velocidad potencial**.

Puesto que

$$\frac{\partial v_x}{\partial y} = 0 = \frac{\partial v_y}{\partial x},$$

se tiene,

$$\mathrm{rot}\boldsymbol{v}(x, y) = \nabla \wedge \boldsymbol{v}(x, y) = \left(\frac{\partial v_y}{\partial x}(x, y) - \frac{\partial v_x}{\partial y}(x, y) \right) \boldsymbol{k} = \boldsymbol{0},$$

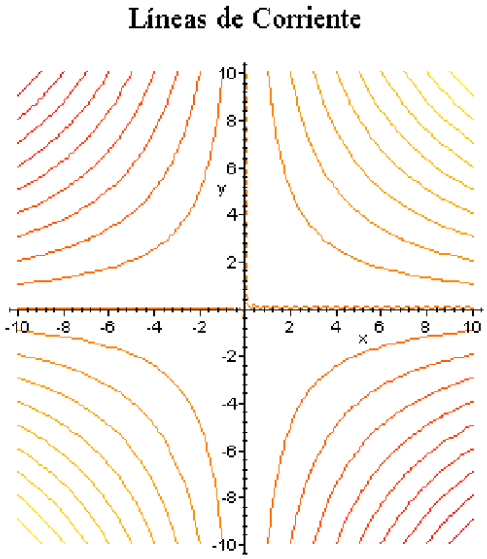


Figura 1.1: Líneas de corriente del fluido: son las curvas de nivel de la función corriente.

luego el campo de velocidades es un campo vectorial plano irrotacional pues el **rotacional escalar**

$$\left(\frac{\partial v_y}{\partial x}(x,y) - \frac{\partial v_x}{\partial y}(x,y)\right),$$

es nulo.

El campo vectorial considerado se puede por tanto identificar con un campo conservativo luego deriva de un campo escalar $\phi(x,y)$ llamado **potencial escalar** de forma que

$$\nabla\phi = \boldsymbol{v}.$$

Cálculo de las líneas equipotenciales mediante el potencial escalar.

La determinación del potencial escalar de un campo conservativo se realiza integrando las igualdades

$$\left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}\right) = (v_x, v_y),$$

En nuestro caso vemos que la función escalar $\phi(x,y)$ verifica el sistema de EDP:

$$\left\{ \quad \frac{\partial \phi}{\partial x} = v_x = x, \qquad \frac{\partial \phi}{\partial y} = v_y = -y. \right.$$

Integrando se tiene:

$$\left\{ \quad \phi(x,y) = \frac{1}{2}x^2 + f(y), \qquad \phi(x,y) = -\frac{1}{2}y^2 + g(x). \right.$$

Derivando parcialmente con respecto a y en la primera expresión de ϕ y derivando parcialmente con respecto a x en la segunda expresión de ϕ se deduce:

$$\frac{\partial \phi}{\partial y} = v_y = -y = f'(y), \qquad \frac{\partial \phi}{\partial x} = v_x = x = g'(x),$$

luego para determinar las expresiones de las funciones arbitrarias f y g es suficiente resolver las EDO:

$$\{ \quad f'(y) = -y, \qquad g'(x) = x,$$

y obtener:

$$\left\{ \quad f(y) = -\frac{1}{2}y^2 + C_1, \qquad g(x) = \frac{1}{2}x^2 + C_2, \right.$$

a partir de las cuales se tiene la siguiente expresión del potencial escalar del campo de velocidades:

$$\phi(x,y) = \frac{1}{2}(x^2 - y^2) + C,$$

siendo $C = C_1 = C_2$ una constante arbitraria de integración. Las líneas equipotenciales de este campo escalar (véase su definición en el tema de teoría de campos del guión de primer curso), se definen por $\phi(x,y) = \alpha$, $\alpha \in I\!R$ es decir,

$$\phi(x,y) = \frac{1}{2}(x^2 - y^2) + C = \alpha,$$

luego tienen la ecuación:

$$x^2 - y^2 = K, \qquad K = 2(\alpha - C),$$

y son por tanto,

$$\left\{ \begin{array}{rcll} x & = & \pm\sqrt{K + y^2} & K \geq 0 \\ y & = & \pm\sqrt{x^2 - K} & K \leq 0, \end{array} \right.$$

y el gradiente de ϕ en cada punto (es decir los vectores de **máximo ascenso**) marca la velocidad en ese punto: $\nabla\phi = \boldsymbol{v}$.

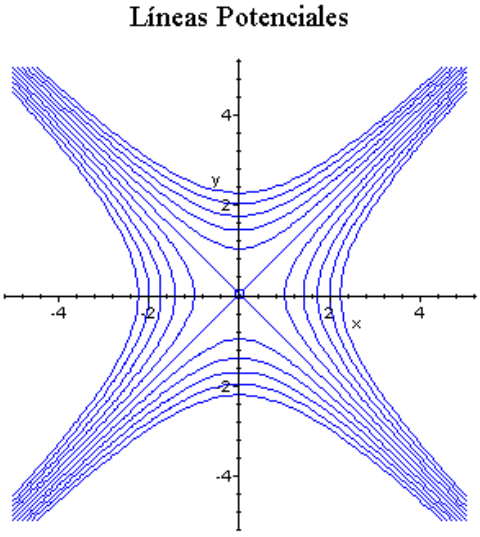


Figura 1.2: Líneas potenciales: son las curvas de nivel de la función potencial.

Ortogonalidad de las líneas de corriente con las líneas equipotenciales de un flujo plano irrotacional

Es posible demostrar que las líneas equipotenciales son ortogonales a las líneas de campo (o líneas de corriente). Lo veremos primero en nuestro caso concreto donde:

$$\phi(x,y) = x^2 - y^2, \qquad \psi(x,y) = xy,$$

son el potencial de velocidad y la función de corriente determinados antes (se han elegido constantes arbitrarias tales que $\phi(0,0) = \psi(0,0) = 0$); posteriormente veremos que la propiedad de ortogonalidad se tiene en general (si el flujo es plano e irrotacional).

Sea (x_0, y_0) un punto genérico del plano $\mathbb{R}^2$. Por comodidad supondremos $y_0 > 0$ (se razonaría de forma análoga si $y_0 < 0$). Si $y_0 = 0$ la línea de corriente que pasa por $(x_0, 0)$ es $\psi(x,y) = xy = 0$, es decir, el eje $y = 0$. Si $(x_0, y_0) = (0,0)$ entonces la línea de corriente viene dada por los ejes coordenados de ecuación $y = 0$ o $x = 0$.

La línea equipotencial y la línea de corriente que pasa por (x_0, y_0), $y_0 > 0$ serán, respectivamente,

$$\phi(x,y) = x^2 - y^2 = x_0^2 - y_0^2, \qquad \psi(x,y) = xy = x_0 y_0,$$

es decir,

$$y = y(x) = \pm\sqrt{x^2 - x_0^2 + y_0^2}, \qquad y = y(x) = \frac{x_0 y_0}{x}.$$

Calculando sus derivadas (que nos dan el coeficiente angular de sus rectas tangentes) se tiene:

$$y' = y'(x) = \pm\frac{x}{\sqrt{x^2 - x_0^2 + y_0^2}}, \qquad y' = y'(x) = -\frac{x_0 y_0}{x^2},$$

y evaluando en x_0 (recuérdese que, por hipótesis, $y_0 > 0$):

$$y'(x_0) = m_1 \doteq \pm\frac{x_0}{|y_0|} = \pm\frac{x_0}{y_0}, \qquad y' = y'(x_0) = m_2 \doteq -\frac{x_0 y_0}{x_0^2} = -\frac{y_0}{x_0}.$$

Un conocido resultado de la geometría analítica nos dice que dos rectas son ortogonales cuando el producto de los coeficientes angulares es -1. Calculamos entonces el producto $m_1 m_2$ para obtener

$$m_1 m_2 = \left(\pm\frac{x_0}{|y_0|}\right)\left(-\frac{y_0}{x_0}\right) = \mp\frac{y_0}{|y_0|} = \mp 1,$$

luego si $y_0 > 0$, entonces la línea equipotencial

$$y = y(x) = \sqrt{x^2 - x_0^2 + y_0^2},$$

cuya recta tangente tiene coeficiente angular $m_1 = x_0/y_0$, es ortogonal a la línea de corriente

$$y(x) = \frac{x_0 y_0}{x},$$

cuya recta tangente tiene coeficiente angular $m_2 = -y_0/x_0$. Si $y_0 < 0$ entonces la línea equipotencial

$$y = y(x) = -\sqrt{x^2 - x_0^2 + y_0^2},$$

cuya recta tangente tiene coeficiente angular $m_1 = x_0/y_0$, es ortogonal a la línea de corriente

$$y(x) = \frac{x_0 y_0}{x},$$

cuya recta tangente tiene coeficiente angular $m_2 = -y_0/x_0$.

En lugar de desarrollar los cálculos anteriores, es posible demostrar la ortogonalidad de las líneas de corriente con las líneas equipotenciales en el caso de flujos planos irrotacionales mediante la siguiente observación. Por la definición (1.3) de la función de corriente se tiene que:

$$\boldsymbol{v} \cdot \nabla\psi = (v_x, v_y) \cdot \left(\frac{\partial\psi}{\partial x}, \frac{\partial\psi}{\partial y}\right) = v_x\frac{\partial\psi}{\partial x} + v_y\frac{\partial\psi}{\partial y} = -v_x v_y + v_y v_x \equiv 0.$$

Por otra parte, al ser $\boldsymbol{v}$ un campo potencial, se tiene que $\nabla\phi = \boldsymbol{v}$. Juntando las dos ecuaciones, $\boldsymbol{v} \cdot \nabla\psi = 0$ y $\nabla\phi = \boldsymbol{v}$, se deduce,

$$\nabla\psi \cdot \nabla\phi = 0,$$

lo que expresa la ortogonalidad de los respectivos gradientes. Ahora bien, $\nabla\psi(x_0, y_0)$ es un vector ortogonal a la línea de corriente que pasa por (x_0, y_0) mientras que $\nabla\phi(x_0, y_0)$ es un vector ortogonal a la línea equipotencial que pasa por (x_0, y_0) luego las líneas mencionadas se cruzan ortogonalmente.

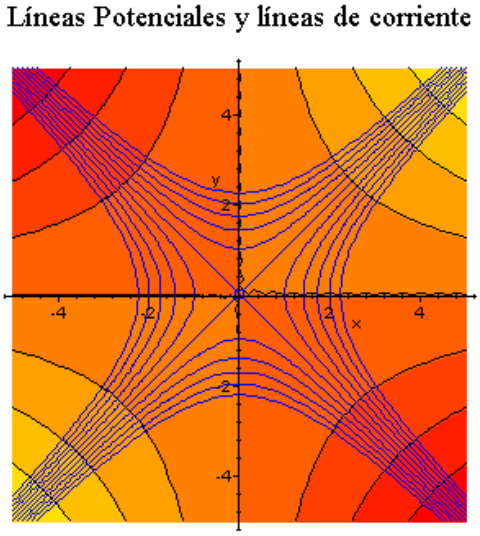


Figura 1.3: Líneas potenciales y de corriente: obsérvese la ortogonalidad entre ellas.

Volveremos a este ejemplo tras haber introducido otros conceptos, definiciones y técnicas de resolución necesarios para ulteriores desarrollos. Puesto que los fundamentos de la interpretación geométrica del proceso de resolución de una EDP se apoyan fuertemente en los de curvas y superficies hemos resumido algunos resultados básicos en los anexos al primer tema.

1.2. Ecuaciones cuasilineales de primer orden

La ecuación (1.1) en el caso en que $m = 1$ nos proporciona la denominada ecuación cuasilineal de primer orden:

$$F\left(x_1, x_2, ..., x_n, u, \frac{\partial u}{\partial x_1}, ..., \frac{\partial u}{\partial x_n}\right) = 0, \tag{1.4}$$

que tendrá que verificarse en una región $\Omega \subset I\!R^n$, siendo $n \geq 2$. Si $n = 2$ entonces se pone $(x_1, x_2) = (x, y) \in \Omega \subset I\!R^2$ y se tiene la forma general de la EDP de primer orden cuasilineal bidimensional dada por:

$$F\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right) = 0 \tag{1.5}$$

La ecuación en la forma (1.5) admite una interpretación geométrica en términos de la teoría de campos (tema 12 de la asignatura de primer curso).

Dos problemas que surgen de modo *natural* al estudiar un campo vectorial continuo (aunque supondremos algo más de regularidad, digamos campos de clase C^1 para poder aplicar las técnicas del cálculo diferencial vectorial) son los problemas sobre la determinación de las líneas vectoriales y de las superficie vectoriales. Casi con la misma frecuencia surge el problema sobre la determinación de la familia de superficies ortogonales a las líneas vectoriales. Analizaremos estos problemas en las siguientes secciones haciendo hincapié en su interpretación geométrica. Previamente, es necesario considerar algunos aspectos, conceptos y definiciones de la teoría matemática general de las EDP de primer orden.

1.2.1. Teoría matemática general

Seguiremos en esta exposición el libro de Elsgoltz[15] capítulo 5. También se ha utilizado el capítulo 1 del libro de F. John[16].

Se denomina ecuación **cuasilineal de primer orden** en derivadas parciales en forma normal (o explícita) a una ecuación de la forma:

$$P_1(x_1, x_2, ..., x_n, u)\frac{\partial u}{\partial x_1} + ... + P_n(x_1, x_2, ..., x_n, u)\frac{\partial u}{\partial x_n} = R(x_1, x_2, ..., x_n, u). \tag{1.6}$$

Nótese que el tipo de ecuación (1.6) es un caso particular de (1.4) donde la incógnita del problema es $u(x_1, ..., x_n)$. Esta ecuación es lineal con respecto a las derivadas, pero puede no ser lineal con respecto a la función desconocida u (en cuyo caso es cuasilineal). Concretamente, se tiene la siguiente clasificación:

CLASIFICACIÓN

[15] Elsgoltz, L., (1983). Ecuaciones diferenciales y cálculo variacional. Tercera ed. Editorial Mir.

[16] Fritz John, (1982). Partial differential equations. IV Ed. Applied Mathematical Sciences. Springer Verlag.

1. Si $R \equiv 0$ y $P_i = P_i(x_1, x_2, ..., x_n)$ (no dependen de u) la ecuación (1.6) se llama **lineal homogénea**:

$$P_1(x_1, x_2, ..., x_n)\frac{\partial u}{\partial x_1} + ... + P_n(x_1, x_2, ..., x_n)\frac{\partial u}{\partial x_n} = 0.$$

Si definimos $\vec{x} = (x_1, ..., x_n)$, $\vec{P} = (P_1, ..., P_n)$ la ecuación anterior se puede escribir en la forma equivalente:

$$\vec{P}(\vec{x}) \cdot \nabla u(\vec{x}) = 0,$$

siendo $\vec{x} \in \Omega \subset I\!R^n$. Si además los coeficientes P_i son constantes, $P_i \equiv c_i$, $c_i \in I\!R$, $i = 1, .., n$, la ecuación (1.6) se dice **lineal homogénea con coeficientes constantes**:

$$c_1\frac{\partial u}{\partial x_1} + ... + c_n\frac{\partial u}{\partial x_n} = 0.$$

Si definimos $\vec{c} = (c_1, ..., c_n)$ y utilizamos la notación anterior la ecuación se puede escribir en la forma equivalente:

$$\vec{c} \cdot \nabla u(\vec{x}) = 0,$$

siendo $\vec{x} \in \Omega \subset I\!R^n$. Ejemplos concretos pueden ser:

$$\frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} = 0, \qquad e^{yx}\frac{\partial u}{\partial x} - (x + y)\frac{\partial u}{\partial y} = 0,$$

que son dos ecuaciones en derivadas parciales de primer orden lineales homogéneas. La primera con coeficientes constantes, $\vec{c} = (c_1, c_2) = (1, -1)$ y la segunda con coeficientes variables, $\vec{P} = (P_1(x, y), P_2(x, y)) = (e^{xy}, x + y)$.

2. Si $R \neq 0$ y $R(x_1, x_2, ..., x_n, u)$ es lineal en u y además se tiene que los coeficientes P_i son funciones del tipo (como antes) $P_i = P_i(x_1, x_2, ..., x_n)$ (es decir no dependen de u) entonces la ecuación (1.6) se llama **lineal no homogénea**:

$$P_1(x_1, x_2, ..., x_n)\frac{\partial u}{\partial x_1} + ...+ P_n(x_1, x_2, ..., x_n)\frac{\partial u}{\partial x_n} = a(x_1, ..., x_n)u + b(x_1, ..x_n)$$

siendo a, b funciones coeficientes que no dependen de u:

$$R(x_1, ..., x_n) = a(x_1, ..., x_n)u + b(x_1, ..x_n).$$

Esta ecuación se puede escribir en la forma equivalente

$$\vec{P}(\vec{x}) \cdot \nabla u(\vec{x}) =$$

$$a(\vec{x})u + b(\vec{x}),$$

siendo $\vec{x} \in \Omega \subset I\!R^n$. Ejemplos pueden ser

$$\frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} = 2u - (x^2 + y^2), \qquad e^{yx}\frac{\partial u}{\partial x} - (x+y)\frac{\partial u}{\partial y} = x + y + u,$$

que son dos ecuaciones en derivadas parciales de primer orden lineales no homogéneas. La primera con coeficientes constantes y la segunda con coeficientes variables.

3. Si algún coeficiente P_i depende de u, en la forma $P_i = P_i(x_1, x_2, ..., x_n, u)$ y además $R \equiv 0$ la ecuación (1.6) se llama **cuasilineal homogénea** y es de la forma:

$$P_1(x_1, x_2, ..., x_n, u)\frac{\partial u}{\partial x_1} + ... + P_n(x_1, x_2, ..., x_n, u)\frac{\partial u}{\partial x_n} = 0,$$

donde la homogeneidad se deduce de la condición $R = 0$. Esta ecuación se puede escribir en la forma equivalente

$$\vec{P}(\vec{x}, u) \cdot \nabla u(\vec{x}) = 0,$$

siendo $\vec{x} \in \Omega \subset I\!R^n$.

Nótese que si los coeficientes P_i dependen de u en forma lineal la ecuación sigue siendo cuasilineal. Por ejemplo, el término uu_{x_1} es no lineal aunque la dependencia $P_1(x_1, x_2, ..., x_n, u) = u$ es lineal. De hecho, es un término cuadrático pues se puede escribir en la forma:

$$u\frac{\partial u}{\partial x_1} = \frac{1}{2}\frac{\partial}{\partial x_1}(u^2).$$

Ejemplo:

$$u\frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} = 0, \qquad e^{yx}\frac{\partial u}{\partial x} - (u + x + y)\frac{\partial u}{\partial y} = 0,$$

son dos ecuaciones en derivadas parciales de primer orden cuasilineales homogéneas.

4. Sea $R \neq 0$ (no idénticamente nula; puede depender de u). Si algún coeficiente P_i depende de u, $P_i = P_i(x_1, x_2, ..., x_n, u)$, entonces la ecuación (1.6) se llama **cuasilineal no homogénea**. Esta ecuación se puede escribir en la forma equivalente:

$$\vec{P}(\vec{x}, u) \cdot \nabla u(\vec{x}) = R(\vec{x}, u),$$

siendo $\vec{x} \in \Omega \subset I\!R^n$. Por ejemplo:

$$u\frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} = x, \qquad e^{yx}\frac{\partial u}{\partial x} - (u + x + y)\frac{\partial u}{\partial y} = -u^2,$$

son dos ecuaciones en derivadas parciales de primer orden cuasilineales no homogéneas.

En el caso bidimensional la notación anterior se simplifica. Si $n = 2$, entonces $(x_1, x_2) = (x, y) \in \Omega \subset I\!R^2$ y la ecuación cuasilineal de primer orden (1.6) se tranforma en:

$$P(x, y, u)\frac{\partial u}{\partial x} + Q(x, y, u)\frac{\partial u}{\partial y} = R(x, y, u), \tag{1.7}$$

siendo $P_1(x, y, u) = P(x, y, u)$, $P_2(x, y, u) = Q(x, y, u)$. Obviamente la clasificación correspondiente se obtiene particularizando lo anterior. Típicamente se denota la solución $u(x, y)$ en la forma $z = u(x, y)$, que es la ecuación (explícita y en coordenadas cartesianas) de una superfice en el espacio tridimensional. Las superficies $z = u(x, y)$, que son soluciones de la EDP se llaman **superficies integrales** de la EDP.

Introducida la forma general de las ecuaciones que trataremos en las secciones siguientes pasaremos ahora a dar una interpretación geométrica del proceso de resolución. El resultado principal que veremos es que la solución general de una EDP del tipo (1.4) (o en la forma explícita (1.6)) se puede obtener resolviendo sistemas (eventualmente acoplados) de ecuaciones diferenciales ordinarias que nacen al considerar ciertos problemas geométricos típicos de la teoría de campos.

Interpretación geométrica

Daremos ahora una interpretación geométrica del problema de la resolución de una EDP lineal no homogénea (o cuasilineal) de primer orden.

Veremos que resolver una EDP de tal tipo implicará buscar las líneas de campo (o líneas vectoriales) de un campo vectorial *naturalmente* asociado a la EDP. El problema se reconducirá por tanto a resolver un sistema de EDO cuyas soluciones (denominadas las características de la EDP) serán las ecuaciones de las líneas de campo del campo vectorial. La solución general de la EDP original nos dará las ecuaciones de las superficies vectoriales del campo. Seleccionar una superficie vectorial concreta equivaldrá a resolver un problema de Cauchy prefijando una curva en el espacio por donde tiene que pasar la superficie

seleccionada. Analizaremos, además, los problemas que surgen al prefijar una curva que sea característica para la EDP.

Para mayor claridad en la interpretación geométrica, estudiemos la ecuación cuasilineal con dos variables independientes $(x, y) \in \Omega \subset I\!R^2$ dada en (1.7), es decir,

$$P(x, y, u)\frac{\partial u}{\partial x} + Q(x, y, u)\frac{\partial u}{\partial y} = R(x, y, u),$$

siendo $u = u(x, y)$ la incógnita del problema. La ecuación (1.7) se deduce de la ecuación (1.6) para $N = 2$, $(x_1, x_2) = (x, y)$, $(P_1, P_2) = (P, Q)$, $R = R$. Supondremos que las funciones coeficientes P, Q, R son diferenciables con continuidad (de clase $C^1(\Omega)$) en la región considerada de variación de las variables, Ω, y que no se anulan simultáneamente en un punto de Ω pues de lo contrario obtendríamos en aquel punto la identidad $0 = 0$ que sería cierta para cualquier función y la ecuación *desaparecería*. Otras situaciones *delicadas* se podrían presentar al anularse uno de los dos coeficientes P o Q. En tal caso la ecuación degeneraría, en el sentido de que cambiaría de tipo, pasando de EDP a EDO pues nos quedaría sólo una derivada parcial. En estos casos sería todavía posible encontrar las soluciones de la EDP pero en un sentido débil. También podría ocurrir que ambos coeficientes P y Q se anularan simultáneamente en un punto (sin que lo hiciera R). En tal caso pasaríamos de una EDP a una ecuación algebraica siendo aún más *dramática* la degeneración. En estos casos sigue siendo posible desarrollar un análisis local de la ecuación pero procuraremos evitar estos casos digamos *patológicos*.

La interpretación geométrica

La idea básica de la interpretación geométrica consiste en asociar a la ecuación (1.7) el campo vectorial de clase $C^1(\Omega)$, siendo Ω una región abierta de $I\!R^3$,

$$\boldsymbol{F} : \Omega \subset I\!R^3 \to I\!R^3,$$

dado por:

$$\boldsymbol{F}(x, y, u) = P(x, y, u)\boldsymbol{i} + Q(x, y, u)\boldsymbol{j} + R(x, y, u)\boldsymbol{k}, \tag{1.8}$$

siendo $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$, vectores unitarios dirigidos por los ejes de coordenadas.

Antes de entrar en detalles con la interpretación geométrica recordaremos algunos conceptos y definiciones básicos para sucesivos tratamientos.

Las **líneas vectoriales** de este campo (es decir, las líneas cuyas tangentes tienen en cada punto la dirección del vector $\boldsymbol{F}$ en dicho punto) son las **curvas**

características de la EDP y se determinan por la condición de paralelismo entre el vector $\boldsymbol{T} = \boldsymbol{i}dx + \boldsymbol{j}dy + \boldsymbol{k}du$ (vector director de la tangente a las líneas buscadas) y el vector $\boldsymbol{F}$. Utilizando el producto vectorial esta condición se escribe en la forma:

$$\boldsymbol{T} \wedge \boldsymbol{F} = (Rdy - Qdu)\boldsymbol{i} + (Pdu - Rdx)\boldsymbol{j} + (Qdx - Pdy)\boldsymbol{k} = \boldsymbol{0}.$$

A lo largo de una curva característica se tiene por tanto:

$$\frac{dx}{P(x,y,u)} = \frac{dy}{Q(x,y,u)} = \frac{du}{R(x,y,u)}. \tag{1.9}$$

Las ecuaciones (1.9) se conocen con el nombre de ecuaciones características de la EDP y su resolución nos proporciona las líneas vectoriales del campo.

Introduciendo un parámetro θ podemos escribir la condición (1.9) (que define las curvas características de la EDP) en la forma

$$\frac{dx}{P(x,y,u)} = \frac{dy}{Q(x,y,u)} = \frac{du}{R(x,y,u)} = d\theta,$$

para obtener el sistema de ecuaciones diferenciales:

$$\begin{cases} \dfrac{dx}{d\theta} = P(x,y,u), & \dfrac{dy}{d\theta} = Q(x,y,u), & \dfrac{du}{d\theta} = R(x,y,u). \end{cases}$$

Este sistema se dice autónomo ya que la variable independiente θ no aparece explícitamente. Suponiendo que P, Q, R son de clase $C^1(\Omega)$ sabemos por la teoría de las ecuaciones diferenciales ordinarias (véanse los temas correspondientes de EDO y sistemas de EDO del primer curso) que a través de cualquier punto de Ω pasa exactamente una curva característica de la EDP o (lo que es lo mismo) una línea vectorial del campo.

Nótese que aunque la introducción del parámetro θ pueda parecer bastante *artificial*, será en realidad entendida como *natural* al considerar la expresión paramétrica de las curvas buscadas. Consideraremos esta situación en la sección dedicada al caso paramétrico.

Las superficies formadas por las **líneas vectoriales** del campo se llaman **superficies vectoriales**. Si una superficie S definida por $u = u(x,y)$ es la unión de curvas características, entonces S es una superficie integral (es decir una solución de la EDP). Por otra parte, cualquier superficie integral es unión de curvas características (lo que equivale a afirmar que a través de cualquier punto de la superficie S pasa una curva característica contenida en S). Estos resultados son una consecuencia del siguiente teorema (cuya demostración se puede encontrar en el libro de F. John, pag 10)

Teorema 1.2.1 *Sea dado un punto $P_0 = (x_0, y_0, z_0) \in \mathbb{R}^3$ que pertenece a una superficie integral S de la EDP definida por $z = u(x, y)$. Sea Γ la curva característica que pasa por el punto P_0. Entonces Γ está completamente contenida en S.*

Como consecuencia directa de este resultado, dos superficie integrales que tienen un punto P_0 en común se intersecan a lo largo de toda la característica que pasa por P_0 (es decir que ambas contienen completamente a la característica). Tenemos ahora una simple descripción de la solución general de (1.7): las superficies integrales $z = u(x, y)$ que son unión de curvas características.

Siendo $\boldsymbol{N}$ un vector normal a la superficie vectorial, esta superficie se caracteriza por la ecuación $\boldsymbol{N} \cdot \boldsymbol{F} = 0$ pues el vector $\boldsymbol{N}$ que tiene la dirección de la normal a la superficie, es ortogonal al vector $\boldsymbol{F}$ del campo en todo punto de ésta. Esta ecuación tiene dos expresiones distintas dependiendo del tipo de representación que se utilize para definir la superficie. En concreto, se tiene:

1. Si la superficie vectorial se determina explícitamente por la ecuación $z = u(x, y)$, entonces el vector normal es

$$\boldsymbol{N} = \frac{\partial u}{\partial x}\boldsymbol{i} + \frac{\partial u}{\partial y}\boldsymbol{j} - \boldsymbol{k} = (u_x, u_y, -1),$$

   y la condición de ortogonalidad $\boldsymbol{N} \cdot \boldsymbol{F} = 0$ toma la forma de la ecuación (1.7). El vector normal unitario es:

$$\boldsymbol{n} = \frac{\boldsymbol{N}}{\|\boldsymbol{N}\|} = \frac{(u_x, u_y, -1)}{\sqrt{1 + u_x^2 + u_y^2}}.$$

2. Si la superficie vectorial se determina implícitamente por la ecuación $S(x, y, z) = 0$, siendo:

$$\nabla S = \left(\frac{\partial S}{\partial x}, \frac{\partial S}{\partial y}, \frac{\partial S}{\partial z}\right),$$

   entonces el vector normal es:

$$\boldsymbol{N} = \frac{\partial S}{\partial x}\boldsymbol{i} + \frac{\partial S}{\partial y}\boldsymbol{j} + \frac{\partial S}{\partial z}\boldsymbol{k},$$

   y la condición de ortogonalidad $\boldsymbol{N} \cdot \boldsymbol{F} = 0$ toma la forma de la ecuación:

$$P(x, y, u)\frac{\partial S}{\partial x} + Q(x, y, u)\frac{\partial S}{\partial y} + R(x, y, u)\frac{\partial S}{\partial u} = 0, \tag{1.10}$$

siendo $S(x, y, u)$ la solución buscada. El vector normal unitario es, en este caso:

$$\boldsymbol{n} = \frac{\nabla S}{\|\nabla S\|}.$$

Por consiguiente, para hallar las superficies vectoriales hay que integrar la ecuación cuasilineal (1.7), o la ecuación lineal homogénea (1.10), según se busque la ecuación de las superficies vectoriales en forma explícita o implícita.

El proceso de resolución

Veamos ahora como actuar cuando la integración de una EDP no es directa. Puesto que las superficies vectoriales pueden formarse por líneas vectoriales, la integración de la ecuación cuasilineal no homogénea (1.7) (que nos proporciona las ecuaciones explícitas de las superficies integrales soluciones de la EDP) se reduce a la integración del sistema de ecuaciones diferenciales ordinarias de las líneas vectoriales que aparece en (1.9) y que aquí recordamos:

$$\frac{dx}{P(x, y, u)} = \frac{dy}{Q(x, y, u)} = \frac{du}{R(x, y, u)}.$$

Sean $\psi_1(x, y, u) = c_1$, $\psi_2(x, y, u) = c_2$ dos soluciones (se las llama también **integrales primeras**) independientes del sistema (1.9). Obsérvese que la independencia se obtiene al considerar sólo dos de las tres ecuaciones características. Esta selección se realiza siguiendo los siguientes principios: siempre se considera la primera de ellas,

$$\frac{dx}{P(x, y, u)} = \frac{dy}{Q(x, y, u)},$$

que nos proporciona las características de la EDP y después se considera una cualquiera (se suele elegir la más sencilla) entre

$$\frac{dx}{P(x, y, u)} = \frac{du}{R(x, y, u)}, \qquad \frac{dy}{Q(x, y, u)} = \frac{du}{R(x, y, u)},$$

lo que nos proporciona las superficies integrales de la EDP. En general, las curvas ψ_1, ψ_2 (obtenidas al resolver dos de las tres ecuaciones diferenciales ordinarias) son las **características** de la ecuación en derivadas parciales. Su expresión define, como ya vimos en la sección anterior, las líneas vectoriales del campo asociado a la EDP. Puesto que las superficies vectoriales (soluciones de la EDP) deben contener todas las características podemos encontrar la

ecuación de las superficies vectoriales estableciendo una dependencia continua cualquiera $\Phi(c_1, c_2) = 0$ entre los parámetros c_1, c_2 que nos permita *pasar con continuidad* de una línea vectorial a otra generando así una superfice. Eliminando los parámetros c_i del sistema:

$$\psi_1(x, y, u) = c_1, \qquad \psi_2(x, y, u) = c_2, \qquad \Phi(c_1, c_2) = 0,$$

obtenemos la ecuación buscada:

$$\Phi(\psi_1(x, y, u), \psi_2(x, y, u)) = 0,$$

siendo Φ una función arbitraria. De este modo, la integral de la ecuación cuasilineal (1.7) (dependiente de una función arbitraria)

$$P(x, y, u)\frac{\partial u}{\partial x} + Q(x, y, u)\frac{\partial u}{\partial y} = R(x, y, u),$$

puede obtenerse por el método siguiente:

1. Se integra el sistema auxiliar de ecuaciones (1.9):

$$\frac{dx}{P(x, y, u)} = \frac{dy}{Q(x, y, u)} = \frac{du}{R(x, y, u)}.$$

2. Se hallan dos integrales primeras de éste:

$$\psi_1(x, y, u) = c_1, \qquad \psi_2(x, y, u) = c_2.$$

3. Se obtiene la integral buscada en la forma:

$$\Phi(\psi_1(x, y, u), \psi_2(x, y, u)) = 0,$$

siendo Φ una función arbitraria.

Ejemplo 1.2.1 *Determinar la integral de la ecuación*

$$\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = 1$$

que depende de una función arbitraria.

Se trata de una EDP de primer orden, lineal, con coeficientes constantes no homogénea. El sistema auxiliar de ecuaciones características es

$$dx = dy = du.$$

Considerando las ecuaciones $dx = dy$ y $dx = du$, sus primeras integrales tienen la forma

$$1) \quad \psi_1(x, y, u) = x - y = c_1, \qquad 2) \quad \psi_2(x, y, z) = u - x = c_2.$$

La integral (o solución general) de la EDP original es

$$3) \qquad \Phi(c_1, c_2) = \Phi(x - y, u - x) = 0,$$

siendo Φ una función arbitraria. De 2) se tiene que $u = x + c_2$. Como por 3) y 1)

$$\Phi(c_1, c_2) = \Phi(x - y, c_2) = 0,$$

se tendrá que $c_2 = f(c_1) = f(x - y)$, siendo f es una función de clase $C^1(\mathbb{R}^2)$ arbitraria. Por tanto,

$$u = x + f(x - y).$$

Nótese en efecto que, al sustituir en la ecuación se tiene:

$$\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = \frac{\partial}{\partial x}\left(x + f(x - y)\right) + \frac{\partial}{\partial y}\left(x + f(x - y)\right) = 1 + f' - f' = 1, \quad \forall\, f,$$

es decir, una identidad, luego $u = x + f(x - y)$ es una familia (infinita) de superficies que representa la solución general de la EDP. Las características de la EDP son las rectas del plano de ecuación $y = x + C$, $C \in \mathbb{R}$.

Interpretación geométrica del proceso resolutivo de la EDP del ejemplo (1.2.1)

El proceso de resolución seguido en el ejemplo (1.2.1) ha consistido en considerar el campo vectorial de clase $(C^1(\mathbb{R}^3))^3 \doteq C^1(\mathbb{R}^3) \times C^1(\mathbb{R}^3) \times C^1(\mathbb{R}^3)$ (es decir, cada componente del campo es de clase $C^1(\Omega)$) definido por:

$$\boldsymbol{F}(x, y, u) = P(x, y, u)\boldsymbol{i} + Q(x, y, u)\boldsymbol{j} + R(x, y, u)\boldsymbol{k} = \boldsymbol{i} + \boldsymbol{j} + \boldsymbol{k}, \qquad (1.11)$$

es decir, $\boldsymbol{F}(x, y, u) = (1, 1, 1)$. Las **líneas vectoriales** de este campo (es decir, las líneas cuyas tangentes tienen en cada punto la dirección del vector $\boldsymbol{F}$ en dicho punto) se determinan por la condición de paralelismo entre el vector $\boldsymbol{T} = dx\boldsymbol{i} + dy\boldsymbol{j} + du\boldsymbol{k}$, dirigido por la tangente a las líneas buscadas (véase la sección en los anexos dedicada a la representación de curvas de este mismo guión) y el vector $\boldsymbol{F} = \boldsymbol{i} + \boldsymbol{j} + \boldsymbol{k}$. Tal condición se expresa en forma de producto vectorial de la siguiente forma:

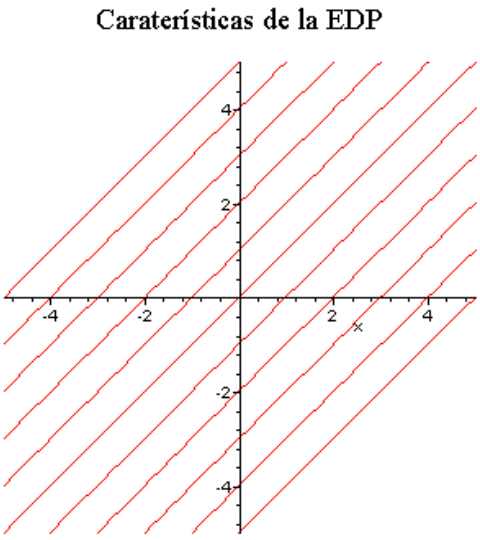


Figura 1.4: Características de la EDP.

$$\boldsymbol{T} \wedge \boldsymbol{F} = (dy - du)\boldsymbol{i} + (du - dx)\boldsymbol{j} + (dx - dy)\boldsymbol{k} = \boldsymbol{0} = (0, 0, 0),$$

que es la ecuación auxiliar $dx = dy = du$ de las características de la EDP. Las líneas vectoriales de este campo son:

$$\psi_1(x, y, u) = x - y = c_1, \qquad \psi_2(x, y, u) = u - x = c_2,$$

y son una familia bi-paramétrica (pues dependen de (c_1, c_2)) de líneas vectoriales (las características de la EDP). De esta familia se extrajo (arbitrariamente) una familia uniparamétrica estableciendo una dependencia continua cualquiera $\Phi(c_1, c_2) = 0$ entre los parámetros (c_1, c_2). Esta operación de extracción (arbitraria pues Φ es una función arbitraria) de una familia uniparamétrica ha correspondido a la determinación de la integral de la ecuación original:

$$\Phi(x - y, u - x) = \Phi(c_1, c_2) = 0,$$

que nos proporcionó la ecuación buscada de las superficies vectoriales:

$$u = u(x, y) = x + \phi(x - y).$$

Interpretación geométrica del problema del flujo potencial: deducción del método de las trayectorias

En esta sección veremos cómo la interpretación geométrica de las EDP permite interpretar los problemas de flujo asociados a un campo de velocidades mediante los conceptos de curvas características y superficie integrales. Lo que haremos será identificar las curvas características con las líneas de corriente y las superficie integrales con la gráfica de la función de corriente. Deduciremos así una forma alternativa de determinar las líneas de corriente de un campo de velocidades mediante la integración de una EDO. El cálculo de las líneas de corriente mediante este método se conoce como **método de las trayectorias**. Nótese que este método se puede usar para cualquier flujo donde el campo de velocidades es conocido. No es necesario que el flujo sea bidimensional o incompresible. También se aplica al tratamiento de problemas de flujo no estacionario.

Otra forma de deducir las líneas de corriente será mediante la resolución de un sistema de EDO lo que nos proporcionará una representación paramétrica de estas curvas. Recuérdese que en el ejemplo (1.1.5) las líneas de corriente se determinaron sólo tras la determinación de la función de corriente mediante integración directa de las EDP que la definen.

En la sección anterior vimos que, dada una EDP de primer orden, si una superficie S definida por $z = \psi(x, y)$, es la unión de curvas características, entonces S es una superficie integral (es decir una solución de la EDP). Por otra parte, cualquier superficie integral es unión de curvas características. Al trabajar con problemas de flujo las curvas características se denominan **líneas de corriente**[17]. Las superficies formadas por las líneas de corriente del campo de velocidades se llaman **superficies vectoriales** y la gráfica de la función de corriente es una de ellas.

Siendo $\boldsymbol{N}$ un vector normal a la superficie vectorial, esta superficie se caracteriza por la ecuación $\boldsymbol{N} \cdot \boldsymbol{v} = 0$, pues el vector $\boldsymbol{N}$, que tiene la dirección de la normal a la superficie, es ortogonal al vector $\boldsymbol{v}$ del campo en todo punto de ésta.

Si la superficie vectorial se determina explícitamente por la ecuación $z =$

[17] En mecánica de medios continuos se distingue claramente entre líneas de campo (trayectorias) y líneas de corriente. Sin embargo si el campo de velocidades es estacionario, las líneas de campo y líneas de corriente son iguales. Trataremos, en este tema, tan sólo este caso (es decir, estudiaremos sólo regímenes de flujo estacionarios).

$\psi(x, y)$, entonces el vector normal es:

$$\boldsymbol{N} = \frac{\partial \psi}{\partial x}\boldsymbol{i} + \frac{\partial \psi}{\partial y}\boldsymbol{j} - \boldsymbol{k} = (\psi_x, \psi_y, -1),$$

y la condición de ortogonalidad $\boldsymbol{N} \cdot \boldsymbol{v} = 0$, toma la forma de la ecuación:

$$v_x \frac{\partial \psi}{\partial x} + v_y \frac{\partial \psi}{\partial y} = 0.$$

La idea básica de la interpretación geométrica del problema propuesto en el ejemplo (1.1.5) (determinación de las líneas de corriente de un campo de velocidades plano dado) consiste por tanto, en asociar a la ecuación de primer orden, lineal homogénea,

$$P(x, y)\frac{\partial \psi}{\partial x} + Q(x, y)\frac{\partial \psi}{\partial y} = 0,$$

el campo vectorial de velocidades,

$$\boldsymbol{v}(x, y) = v_x(x, y)\boldsymbol{i} + v_y(x, y)\boldsymbol{j}.$$

La ecuación que caracteriza la función de corriente ψ asociada a un campo de velocidades $\boldsymbol{v}$ es:

$$\boldsymbol{v} \cdot \nabla\psi = v_x \frac{\partial \psi}{\partial x} + v_y \frac{\partial \psi}{\partial y} = 0.$$

Puesto que el campo dado era $(v_x, v_y) = (x, -y)$, se obtiene la EDP (de primer orden lineal homogénea):

$$x\frac{\partial \psi}{\partial x} - y\frac{\partial \psi}{\partial y} = 0.$$

Recordando la definición de ψ dada en (1.3):

$$\frac{\partial \psi}{\partial x} = -v_y, \qquad \frac{\partial \psi}{\partial y} = v_x,$$

se deduce que ψ es una superficie integral (solución) de la EDP de primer orden.

Las **líneas vectoriales** de este campo (es decir, las líneas cuyas tangentes tienen en cada punto la dirección del vector $\boldsymbol{F}$ en dicho punto) se interpretan como las **líneas de corriente** del campo de velocidades y representan las **curvas características** de la EDP. Se determinan por la condición de

paralelismo entre el vector $\boldsymbol{T} = \boldsymbol{i}dx + \boldsymbol{j}dy + \boldsymbol{k}d\psi$ (vector director de la tangente a las líneas buscadas) y el vector $\boldsymbol{v}$. Utilizando el producto vectorial esta condición se escribe en la forma:

$$\boldsymbol{T} \wedge \boldsymbol{v} = (Rdy - Qdu)\boldsymbol{i} + (Pdu - Rdx)\boldsymbol{j} + (Qdx - Pdy)\boldsymbol{k}$$

Puesto que $P = v_x$, $Q = v_y$ y $R = 0$, se tiene:

$$\boldsymbol{T} \wedge \boldsymbol{v} = v_y d\psi \boldsymbol{i} + v_x d\psi \boldsymbol{j} + (v_y dx - v_x dy)\boldsymbol{k} = \boldsymbol{0}.$$

A lo largo de una curva característica se tiene por tanto:

$$\frac{dx}{v_x} = \frac{dy}{v_y} \tag{1.12}$$

y la solución $\psi(x, y)$ es constante, $\psi \equiv C$, a lo largo de una característica pues su tasa de cambio (dada por $\boldsymbol{v} \cdot \nabla\psi$) es nula por la definición y construcción de ψ. El hecho de que ψ es constante se denota en la forma:

$$\frac{dx}{v_x} = \frac{dy}{v_y} = \frac{d\psi}{0}.$$

Las ecuaciones (1.12) se conocen con el nombre de ecuaciones características de la EDP y su resolución nos proporciona las líneas de corriente del campo.

Ecuaciones paramétricas de las líneas de corriente

El método de las trayectorias para la determinación de las líneas de corriente consiste en la resolución de la ecuación (1.12) previa introducción de un parámetro. Este procedimiento permite obtener propiamente las ecuaciones paramétricas de las líneas de corriente en términos de trayectorias.

En efecto, introduciendo un parámetro θ podemos escribir la condición (1.12) (que define las curvas características de la EDP) en la forma:

$$\frac{dx}{v_x} = \frac{dy}{v_y} = \frac{d\psi}{0} = d\theta,$$

para obtener el sistema de ecuaciones diferenciales ordinarias:

$$\begin{cases} \dfrac{dx}{d\theta} = P(x, y) = v_x, & \dfrac{dy}{d\theta} = Q(x, y) = v_y, & \dfrac{d\psi}{d\theta} = R(x, y) = 0, \end{cases}$$

es decir,

$$\begin{cases} \dfrac{dx}{d\theta} = x, & \dfrac{dy}{d\theta} = -y, & \dfrac{d\psi}{d\theta} = 0. \end{cases}$$

El sistema es desacoplado (es decir, podemos resolver las ecuaciones que lo componen de manera independiente, de una en una). Puesto que se trata de EDO lineales de primer orden separables se obtiene (separando variables) la representación paramétrica de las líneas de corriente:

$$x(\theta) = c_1 e^{\theta}, \qquad y(\theta) = c_2 e^{-\theta}, \qquad \psi(\theta) = c_3.$$

Los valores de c_1, c_2 y c_3 se determinan imponiendo unas condiciones ulteriores que complementan la EDP. Analizaremos tales situaciones en las secciones siguientes, al introducir la definición de Problema de Cauchy asociado a una EDP de primer orden.

El problema de Cauchy

Si entre las infinitas superficies vectoriales del campo $\boldsymbol{F}$ se deseara determinar la superficie que pasa por una línea dada, definida a su vez por las ecuaciones $\Phi_1(x, y, u) = 0$, $\Phi_2(x, y, u) = 0$, entonces la función Φ ya no será arbitraria sino que se determina $\Phi(c_1, c_2)$ por eliminación de x, y, u de las ecuaciones:

$$\Phi_1(x, y, u) = 0, \quad \Phi_2(x, y, u) = 0, \quad \psi_1(x, y, u) = c_1, \quad \psi_2(x, y, u) = c_2,$$

a través del cual se tiene la ecuación $\Phi(c_1, c_2) = 0$ y la integral buscada será

$$\Phi(\psi_1(x, y, u), \psi_2(x, y, u)) = 0,$$

Para poder explicar con rigor este proceso de selección de soluciones precisamos antes la definición de Problema de Cauchy para EDP cuasilineales de primer orden.

El problema de Cauchy para ecuaciones cuasilineales

Para seleccionar una entre las infinitas superficies integrales de una EDP cuasilineal de primer orden seguiremos el siguiente camino que nos llevará a la definición del problema de Cauchy para EDP cuasilineales de primer orden, que no es otra cosa que un método para generar soluciones a partir de la prescripción de un conjunto de funciones que representan los datos del problema (digamos la información física o experimental) sobre el fenómeno a estudiar. Es decir, buscamos asociar soluciones a datos mediante una aplicación entre espacios vectoriales (el espacio de los datos como espacio inicial y el espacio de

las soluciones como espacio de llegada). El espacio de las soluciones se describe entonces mediante el espacio (usualmente más simple) de los datos[18].

Una manera simple de seleccionar una función $u(x, y)$ en el conjunto infinito de soluciones de (1.7)

$$P(x, y, u)\frac{\partial u}{\partial x} + Q(x, y, u)\frac{\partial u}{\partial y} = R(x, y, u),$$

consiste en prescribir una curva Γ en el espacio tridimensional que debe estar contenida en la superficie integral $z = u(x, y)$. Sea Γ una curva dada paramétricamente mediante:

$$x = f(s), \qquad y = g(s), \qquad u = h(s) \tag{1.13}$$

Estamos buscando una solución $u(x, y)$ de (1.7), tal que la relación

$$h(s) = u(f(s), g(s)) \tag{1.14}$$

sea una identidad.

Definición 1.2.1 (Problema de Cauchy) *Sea dada la ecuación cuasilineal de primer orden (1.7). Un problema de Cauchy asociado a la ecuación (1.7), consiste en encontrar la función $u(x, y)$ asociada a los datos $f(s)$, $g(s)$ y $h(s)$ mediante (1.13).*

Obviamente se entiende por función asociada a los datos a una función que satisface la EDP y la condición (1.14). Nótese que es posible dar muchas parametrizaciones distintas de la misma curva Γ eligiendo distintos parámetros s. Sin embargo la introducción de un parámetro diferente σ tal que $s = \phi(\sigma)$ no cambia la solución del problema de Cauchy. En cuanto al tipo de solución, nos contentaremos con **soluciones locales**[19] definidas para (x, y) en un entorno del punto $x_0 = f(s_0)$, $y_0 = g(s_0)$. En muchos problemas físicos la variable y se identifica con el tiempo y x con la posición en el espacio. Además en dichos problemas es frecuente que y_0 se tome como inicio del intervalo temporal en el que se plantea el problema. Es entonces natural proponerse resolver el problema de determinar una solución $u(x, y)$ a partir del conocimiento de su *valor inicial* en el tiempo $y = 0$:

$$u(x, 0) = u_0(x). \tag{1.15}$$

[18] No queremos (ni podemos) entrar en detalles sobre este tipo de problemas propio de la teoría de la regularidad de las EDP. Sólo señalamos que la caracterización del espacio de soluciones de una EDP mediante la información sobre los datos del problema representa uno de los problemas más interesantes de la teoría de las ecuaciones en derivadas parciales.

[19] Es una de las características más notables de las EDP de primer orden hiperbólicas: la existencia local, es decir hasta un tiempo finito.

Definición 1.2.2 (Problema de Valor Inicial) *Sea dada la ecuación cuasilineal de primer orden (1.7). Un problema de valor inicial para la ecuación consiste en encontrar la función $u(x,y)$ que satisface la ecuación (1.7) y la condición inicial (1.15).*

Nótese que un problema de valor inicial (PVI) es un caso especial de un problema de Cauchy en el cual la curva Γ tiene la representación

$$x = s, \qquad y = 0, \qquad u = u_0(s). \tag{1.16}$$

Sea $\Omega \subset I\!R^3$ y sea $\boldsymbol{F} = (P, Q, R)$ un campo vectorial de clase $(C^1(\Omega))^3$. Sea dada además una curva inicial Γ_0 en la forma paramétrica (1.13). Es posible entonces demostrar que el problema de Cauchy admite solución si se verifica la siguiente condición suficiente:

$$\Delta = \begin{vmatrix} f'(s_0) & g'(s_0) \\ P(x_0, y_0, u_0) & Q(x_0, y_0, u_0) \end{vmatrix} = Q_0 f_0^{'} - P_0 g_0^{'} \neq 0. \tag{1.17}$$

En el caso del problema de valor inicial se tiene (aplicando (1.16) con $f(s) = s$, $g(s) = 0$):

$$\Delta = \begin{vmatrix} 1 & 0 \\ P(x_0, y_0, u_0) & Q(x_0, y_0, u_0) \end{vmatrix} = Q(x_0, y_0, u_0) \neq 0. \tag{1.18}$$

La condición de no anulación del determinante jacobiano Δ garantiza la existencia (local pues se debe a una aplicación del teorema de la función implícita válida en un entorno del punto $(x_0, y_0, u_0) \in I\!R^3$) de una superficie integral $u = u(x, y)$. La unicidad se deduce del teorema (1.2.1). Nótese que la condición (1.17) es fundamental para la existencia de una única solución $u(x, y)$ de clase C^1 pues se puede demostrar que $\Delta = 0$ es incompatible con la existencia de tales soluciones a menos que la curva prefijada Γ_0 sea característica (en cuyo caso se tendrán infinitas soluciones).

Acabamos la parte teórica con el siguiente resultado de existencia y unicidad para el problema de Cauchy asociado a una curva inicial Γ_0 dada en la forma paramétrica:

$$x = f(s), \qquad y = g(s), \qquad u = u_0(s), \tag{1.19}$$

siendo f, g funciones derivables con continuidad. Se tiene entonces el siguiente teorema:

Teorema 1.2.2 *Sea dada la ecuación en derivadas parciales*

$$P(x,y,u)\frac{\partial u}{\partial x} + Q(x,y,u)\frac{\partial u}{\partial y} = R(x,y,u),$$

y sea Γ_0 *una curva inicial dada por (1.19) con la condición que* $(f')^2+(g')^2 \neq 0$ *(es decir* f' *y* g' *no ambas nulas simultáneamente). Sea además,*

$$\Delta = Q_0 x_0^{'} - P_0 y_0^{'} = Q_0 f_0^{'} - P_0 g_0^{'},$$

el determinante jacobiano. Se tiene entonces:

1. *Si* $\Delta \neq 0$ *en toda* Γ_0 *entonces existe una única solución del problema de Cauchy.*
2. *Si* $\Delta = 0$ *en toda* Γ_0 *y* Γ_0 *es una curva característica entonces existen infinitas soluciones del problema de Cauchy.*
3. *Si* $\Delta = 0$ *en toda* Γ_0 *y* Γ_0 ***no*** *es una curva característica entonces no existe solución del problema de Cauchy.*

Lo anterior se puede interpretar de la siguiente forma: si el problema de Cauchy (o el PVI) tienen solución y $\Delta = 0$ a lo largo de Γ_0 entonces la curva Γ_0 es ella misma una curva característica de la EDP. Pero si Γ_0 es una curva característica entonces infinitas superficies integrales pasan a través de (contienen) Γ_0. Observemos que si la función u no es de clase C^1 (en un entorno de Γ_0 y en la misma Γ_0) entonces no es posible deducir, a partir de $\Delta = 0$, que la curva Γ_0 es característica. En efecto, pueden existir soluciones de la EDP que pasen por una curva no característica Γ_0 y para las cuales $\Delta = 0$.

Nótese finalmente que este teorema no cubre el caso en el cual $\Delta = 0$ sólo en algún punto de Γ_0 puesto que en este caso las soluciones no serían de clase C^1. Aparecerían soluciones débiles. Existe una teoría al respecto que no será aquí considerada. Una referencia a un posible tratamiento general de este caso se encuentra en el libro de Courant-Hilbert, pag 65. En general es suficiente considerar por separado las regiones de degeneración de los coeficientes de la EDP (es decir donde se anulan) y utilizar el análisis local mediante el estudio del determinante jacobiano Δ a lo largo de Γ_0 en las regiones de no degeneración.

Ejemplo 1.2.2 (Problema de Cauchy) *Hallar la superficie integral de la ecuación:*

$$x\frac{\partial u}{\partial y} - y\frac{\partial u}{\partial x} = 0,$$

que pasa por la curva inicial Γ_0 *dada por* $x = 0$, $u = y^2$.

Se trata de una EDP de primer orden, lineal, con coeficientes variables, homogénea. Para que la ecuación no pueda degenerar imponemos la condición $x^2 + y^2 \neq 0$. El problema de Cauchy está bien planteado (existe una única solución). En efecto, notamos que, parametrizando la curva dada en la forma:

$$x = f(s) = 0, \qquad y = g(s) = s, \qquad u = u_0(s) = h(s) = s^2,$$

se tiene $f'(s) = 0$, $g'(s) = 1$. Considerando que $P = -y$, $Q = x$ se deduce $\Delta = Qf' - Pg' = y \neq 0$ si $y \neq 0$ y esto es cierto a lo largo de la curva Γ_0 para todo $s \neq 0$. Si $s = 0$ entonces $x = y = 0$ pero este punto ya había sido excluido con la condición de no degeneración. Aplicando el teorema (1.2.2) se tiene asegurada la existencia de una única superficie integral que contiene la curva inicial Γ_0.

El campo vectorial a estudiar es:

$$\boldsymbol{F}(x, y, u) = (P(x, y, u), Q(x, y, u), R(x, y, u)) = (-y, x, 0).$$

La ecuación auxiliar $\boldsymbol{T} \wedge \boldsymbol{F} = \boldsymbol{0}$ nos conduce en este caso a:

$$\frac{dx}{-y} = \frac{dy}{x} = \frac{du}{0}.$$

La notación $du/0$ significa que $du = 0$, es decir, que u es constante a lo largo de las características. Esta notación (dividir por cero) se aplica también en el caso $P \equiv 0$ o $Q \equiv 0$. A partir de la ecuación auxiliar se pueden obtener las siguientes integrales primeras (características o líneas vectoriales del campo o líneas de corriente si el campo $\boldsymbol{F}$ se considera un campo de velocidades estacionario plano $\boldsymbol{F} = \boldsymbol{v} = (v_x, v_y)$):

$$x dx = -y dy, \qquad \Longrightarrow \qquad \frac{1}{2}x^2 = -\frac{1}{2}y^2 + K \qquad \Longrightarrow \qquad x^2 + y^2 = 2K = C_2,$$

es decir, $\psi_2(x, y, u) = x^2 + y^2 = c_2$ y

$$du = 0 \qquad \Longrightarrow \qquad u = C_1,$$

o sea, $\psi_1(x, y, u) = u = c_1$. Hemos obtenido:

$$\psi_1(x, y, u) = u = c_1, \qquad \psi_2(x, y, u) = x^2 + y^2 = c_2.$$

Nótese que la curva Γ_0 no es (una curva) característica pues no satisface ninguna de las ecuaciones $\psi = c_1$, $\psi = c_2$. Considerando las condiciones $x = 0$, $u = y^2$ se tiene el siguiente sistema de ecuaciones:

$$u = c_1 \quad x^2 + y^2 = c_2 \quad x = 0 \quad u = y^2,$$

y se deduce que (eliminando las variables x, y, u) $c_1 = c_2$. Por ello la superficie integral buscada será

$$0 = c_1 - c_2 = \Phi(c_1, c_2) = \Phi(u, x^2 + y^2) = u - (x^2 + y^2),$$

es decir,

$$u = x^2 + y^2.$$

Nótese que evidentemente al seccionar la superficie con el plano $x = 0$ se obtiene la curva prefijada $u = y^2$. En términos de la mecánica de fluidos, la superficie integral obtenida corresponde a la gráfica de la función de corriente asociada al campo de velocidades.

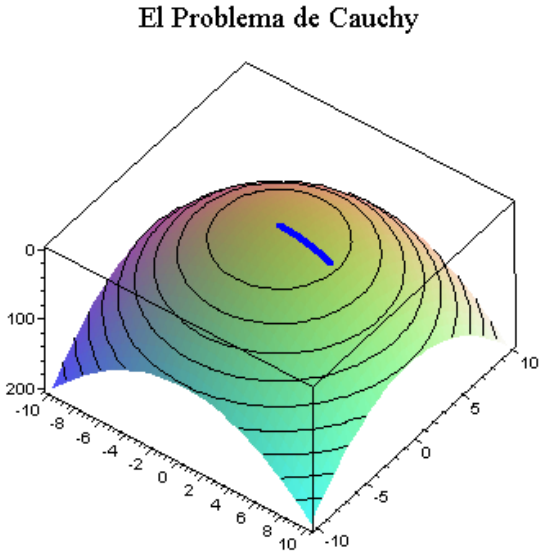


Figura 1.5: El problema de Cauchy.

Obsérvese además que el problema quedaría indeterminado si la curva dada por $\Phi_1 = \Phi_2 = 0$ fuese característica ya que en este caso diferentes superficies integrales pasarían por dicha curva. Es decir: existirían infinitas soluciones. Encontraremos concretamente esta situación en el siguiente ejemplo:

Ejemplo 1.2.3 (Problema de Cauchy) *Hallar la superficie integral de la ecuación:*

$$x\frac{\partial u}{\partial y} - y\frac{\partial u}{\partial x} = 0,$$

que pasa por la circunferencia $u = 1$, $x^2 + y^2 = 4$.

Puesto que la línea dada es característica, el problema es indeterminado, es decir, existen infinitas soluciones (superficies) que pasan por la circunferencia dada. Por ejemplo los paraboloides de revolución:

$$u = x^2 + y^2 - 3, \qquad 4u = x^2 + y^2, \qquad u = -x^2 - y^2 + 5,$$

son soluciones pues corresponden a las superficies:

$$\Phi(c_1, c_2) = c_2 - c_1 - 3 = 0, \qquad \Phi(c_1, c_2) = c_2 - 4c_1 = 0, \qquad \Phi(c_1, c_2) = c_1 + c_2 - 5 = 0.$$

En efecto, cualquier superficie de revolución $u = \Phi(x^2 + y^2)$, cuyo eje de rotación coincida con el eje Oz, es superficie integral de la EDP. También la esfera $x^2 + y^2 + u^2 = 5$ es solución pues corresponde a la superficie:

$$\Phi(c_1, c_2) = c_1^2 + c_2 - 5 = 0.$$

El problema de Cauchy anterior estaba mal planteado (pues existen infinitas soluciones). Nótese que, parametrizando la curva dada en la forma:

$$x = f(\theta) = 2\cos\theta, \qquad y = g(\theta) = 2\sin\theta, \qquad u = u_0(\theta) = 1,$$

se tiene, $f'(\theta) = -2\sin\theta$, $g'(\theta) = 2\cos\theta$. Considerando que $P = -y$, $Q = x$ se deduce $\Delta = Qf' - Pg' = -4\sin\theta\cos\theta + 4\sin\theta\cos\theta \equiv 0$, $\forall\,\theta$.

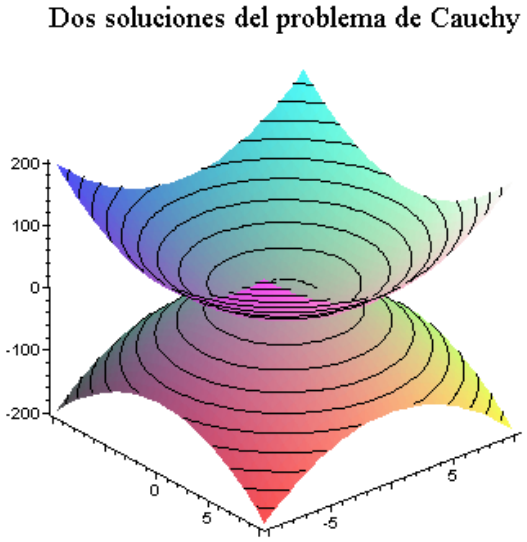


Figura 1.6: Problema mal planteado: existen dos soluciones del mismo problema de Cauchy.

Hasta ahora hemos considerado el caso en que la expresión de la curva inicial venía dada en coordenadas cartesianas mediante un sistema de dos ecuaciones, cuya solución (la curva) es la intersección de dos superficies (las ecuaciones). Consideraremos ahora el caso en el que la curva inicial se expresa en forma paramétrica.

El caso paramétrico

Si la ecuación de la curva Γ_0, por la cual se exige trazar la superficie integral de la ecuación:

$$P(x,y,u)\frac{\partial u}{\partial x} + Q(x,y,u)\frac{\partial u}{\partial y} = R(x,y,u),$$

se proporciona en forma paramétrica:

$$\Gamma_0 = \{ \quad x_0 = x_0(s), \qquad y_0 = y_0(s), \qquad u_0 = u_0(s),$$

entonces, es también conveniente buscar la solución en forma paramétrica[20]

$$x = x(t,s), \qquad y = y(t,s), \qquad u = u(t,s).$$

Para ello se introduce un parámetro t en el sistema (1.9) que determina las características, haciendo

$$\frac{dx}{P(x,y,u)} = \frac{dy}{Q(x,y,u)} = \frac{du}{R(x,y,u)} = dt. \tag{1.20}$$

Para que las características pasen por la curva dada, se busca la solución del sistema (1.20) que satisface (para $t = 0$ o $t = T_0$) las condiciones iniciales:

$$x_0 = x_0(s) = x(T_0,s), \qquad y_0 = y_0(s) = y(T_0,s), \qquad u_0 = u_0(s) = u(T_0,s).$$

Para estas condiciones iniciales (y para s fija) obtenemos una característica que pasa por un punto (fijado) de la curva. Cuando s es variable se obtiene la familia biparamétrica de características:

$$x = x(t,s), \qquad y = y(t,s), \qquad u = u(t,s),$$

que pasan por los puntos de la curva dada (en este caso se considera que la curva dada no es característica). El conjunto de puntos que pertenece a esta familia de características forma precisamente la integral buscada.

[20] Véase el tema 12 de la asignatura de primer curso (y también los anexos a este guión) para la definición de una superficie paramétrica.

Ejemplo 1.2.4 *Sea dada la ecuación:*

$$\frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} = 1.$$

Clasificar la EDP, determinar el campo asociado y hallar la superficie integral que pasa por la curva:

$$x_0 = s, \qquad y_0 = s^2, \qquad u_0 = s^3.$$

Se trata evidentemente de una EDP de primer orden, lineal, no homogénea con coeficientes constantes. Está asociada al campo $F = (1, -1, 1)$. El sistema de ecuaciones que determina las características tiene la forma:

$$dx = -dy = du = dt.$$

Su solución general es:

$$x = t + c_1, \qquad y = -t + c_2, \qquad u = t + c_3.$$

Las constantes arbitrarias se determinan mediante las condiciones iniciales:

$$c_1 = s, \qquad c_2 = s^2, \qquad c_3 = s^3,$$

y sustituyendo en la solución general se tiene:

$$x = t + s, \qquad y = -t + s^2, \qquad u = t + s^3.$$

* Parametrización mediante la longitud de arco

Consideraremos EDP de primer orden en dos variables independientes. La familia más general se puede escribir en la forma:

$$A\frac{\partial u}{\partial x} + B\frac{\partial u}{\partial y} = C.$$

Si $A = A(x, y)$, $B = B(x, y)$ y $C = C(x, y, u)$ es lineal en u la ecuación es **lineal**. En los demás casos es **no lineal**[21] En cada punto del plano x, y podemos definir un vector unitario:

$$\boldsymbol{s} \equiv \left(\frac{A}{\sqrt{A^2 + B^2}}, \frac{B}{\sqrt{A^2 + B^2}} \right),$$

[21] Una clasificación más precisa, clasifica las ecuaciones no lineales en semi-lineales, cuasi-lineales, completamente no lineales y doblemente no lineales dependiendo de las dependencias funcionales de las funciones coeficientes

y escribir la PDE de primer orden en la forma:

$$\boldsymbol{s} \cdot \nabla u = \frac{C}{\sqrt{A^2 + B^2}} \equiv D.$$

Fijada una curva inicial (que no coincida con una característica) podemos dibujar en el plano, para cada punto de la curva inicial, una trayectoria que es tangente en cada punto al vector $\boldsymbol{s}$. La pendiente de la tangente (en cada punto) es:

$$\frac{dy}{dx} = \frac{B}{A} \tag{1.21}$$

y es la EDO verificada por la trayectoria. Estas trayectorias se llaman **características**. Denotamos con σ a la longitud de arco[22] a lo largo de una característica. Supongamos ahora que los coeficientes A, B y C son constantes reales. Se tiene entonces:

$$\boldsymbol{s} = \left(\frac{dx}{d\sigma}, \frac{dy}{d\sigma} \right).$$

Este resultado se puede comprobar fácilmente definiendo una parametrización de las curvas características en la forma:

$$\boldsymbol{r}(t) = x(t)\boldsymbol{i} + y(t)\boldsymbol{j}, \quad t \geq t_0 = 0,$$

luego,

$$\boldsymbol{r}'(t) = x'(t)\boldsymbol{i} + y'(t)\boldsymbol{j}, \qquad \|\boldsymbol{r}'(t)\| = \sqrt{(x')^2 + (y')^2},$$

y por tanto, puesto que las ecuaciones características vienen dadas por (paramétricamente):

$$\begin{cases} \dfrac{dx}{dt} = A, & \dfrac{dy}{dt} = B, \end{cases}$$

se tiene que la relación entre el parámetro t y el parámetro longitud de arco es:

$$\sigma \doteq \int_0^t \|\boldsymbol{r}'(t)\| dt = \int_0^t (\sqrt{A^2 + B^2}) dt = (\sqrt{A^2 + B^2})t,$$

luego la diferencial de longitud de arco es:

$$d\sigma = \|\boldsymbol{r}'(t)\| dt = (\sqrt{A^2 + B^2}) dt.$$

Por lo anterior, escribimos entonces $\boldsymbol{s} \cdot \nabla u = D$ en la forma:

[22] Véase el tema 12, sección 10 del guión de la asignatura del primer curso para la definición de este concepto.

$$\frac{du}{d\sigma} = D. \tag{1.22}$$

Cuando $A(x,y)$, $B(x,y)$ son funciones conocidas de las variables independientes, podemos integrar primero (1.21) para encontrar las características y después (1.22) para obtener u. Si $A(x,y,u)$, $B(x,y,u)$ dependen de u (pero no de sus derivadas) entonces las características no son conocidas *a priori* y hay que resolver el problema acoplado (1.21), (1.22) (generalmente con un método numérico en diferencias finitas, ver Mei, pag 21) para encontrar las características y la incógnita u.

Concluimos esta sección observando que el esquema de resolución indicado para el caso bidimensional se puede generalizar al caso $(n+1)-$dimensional. Detalles se pueden encontrar en el libro de Elsgoltz, capítulo 5 pag 252 o en el libro de Courant-Hilbert, pag 69.

* Un ejemplo de formación de ondas de choque en una ecuación cuasilineal. Ecuación de Bürgers

Un problema de valor inicial que a menudo se toma como ejemplo para ilustrar ciertos fenómenos que pueden aparecer al trabajar con ecuaciones cuasilineales de primer orden es: *Encontrar una función* $u = u(x,y)$ *que satisfaga las igualdades*:

$$\left\{ \begin{array}{rcll} \dfrac{\partial u}{\partial y} + u\dfrac{\partial u}{\partial x} & = & 0, & x \in I\!R, \quad y > 0, \\ u(x,0) & = & u_0(x), & x \in I\!R. \end{array} \right. \tag{1.23}$$

Este problema, que modeliza entre otros fenómenos el rompimiento de la barrera del sonido por aviones supersónicos fue propuesto por J.M. Bürgers y, en su honor, se conoce como el problema de Bürgers[23]. La solución u de la ecuación cuasilineal se puede interpretar como un campo de velocidad longitudinal (en la dirección x) que varía con el tiempo y. La ecuación (1.23) afirma que cualquier partícula del fluido tiene aceleración nula (no hay efectos inerciales), luego tiene velocidad constante a lo largo de las trayectorias del fluido. Esto es evidente si ponemos $u = v_x = v_x(x,y)$ siendo $\boldsymbol{v} = (v_x(x,y), v_y(x,y)) = (v_x(x,y), 0)$ (nótese que la variable y que aparece en las componentes del campo de velocidades denota una variable espacial) y calculamos su derivada material definida

[23] En realidad la ecuación de Bürgers propiamente dicha es una ecuación parabólica de segundo orden a partir de la cual se puede deducir (mediante un límite asintótico que representa la hipótesis de difusión despreciable) la ecuación hiperbólica cuasilineal de primer orden que analizaremos.

mediante el operador de derivación total (donde t representa ahora el tiempo) [24]

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \boldsymbol{v} \cdot \nabla = \frac{\partial}{\partial t} + \sum_{i=1}^{n} v_i \frac{\partial}{\partial x_i},$$

siendo $\boldsymbol{v} = (v_1, .., v_n)$, $\boldsymbol{x} = (x_1, .., x_n)$. En nuestro caso $n = 2$, $(x_1, x_2) = (x, y)$, $v_1 = v_x$, $v_2 = v_y$. La afirmación anterior equivale por tanto a verificar (se deja por ejercicio al lector) que $D\boldsymbol{v}/Dt = 0$.

Considérese nuevamente el problema (1.23) siendo y la variable temporal. La condición inicial $u(x, 0) = u_0(x)$ describe la distribución inicial de la velocidad, que corresponde a la curva en el espacio $\mathbb{R}^3$ (véase la ecuación (1.16)):

$$x = s, \qquad y = 0, \qquad u = u_0(s). \tag{1.24}$$

Introduciendo un parámetro θ las ecuaciones diferenciales ordinarias asociadas son:

$$\frac{dx}{d\theta} = P(x, y, u) = u, \qquad \frac{dy}{d\theta} = Q(x, y, u) = 1, \qquad \frac{du}{d\theta} = R(x, y, z) = 0.$$

que combinadas con la condición inicial (1.24) para $\theta = 0$ nos dan la representación paramétrica de la solución $z = u(x, y)$ del PVI:

$$x = s + u\theta, \qquad y = \theta, \qquad u = u_0(s). \tag{1.25}$$

Eliminando los parámetros s, θ, se tiene la fórmula (implícita) de la solución:

$$u = u_0(s) = u_0(x - u\theta) = u_0(x - uy).$$

* Discontinuidades y soluciones débiles

En la primera parte del ejercicio anterior hemos encontrado una fórmula implícita de la solución. Queremos analizar más en detalle su naturaleza (su comportamiento cualitativo). Si proyectamos la característica (curva en el espacio) en el plano xy (es decir eliminamos u en (1.25)) obtenemos la curva C_s (proyección de la característica):

$$x = s + u_0(s)y,$$

a lo largo de la cual la solución tiene el valor constante,

$$u(x, 0) = u(s, 0) = u_0(s).$$

[24] Algunos libros de texto denominan a la derivada material, derivada substancial. Véase, por ejemplo, el libro de Costa-Novella, Vol 2, dedicado a los fenómenos de transporte.

Físicamente, $x = s + u_0(s)y$ define el camino de una partícula localizada en $x = s$ en el tiempo $y = 0$. Veamos que ocurre si dos características tienen un punto en común. Sea $(x^*, y^*) \in I\!R^2$ el punto común a las dos características. Entonces, puesto que la ecuación de las características es $x = s + u_0(s)y$ se tendrá:

$$x^* = s_1 + u_0(s_1)y^*, \qquad x^* = s_2 + u_0(s_2)y^*,$$

y se deduce, $s_1 + u_0(s_1)y^* = s_2 + u_0(s_2)y^*$, es decir,

$$y^* = -\frac{s_2 - s_1}{u_0(s_2) - u_0(s_1)}.$$

Si $s_2 \neq s_1$ y $u_0(s_2) \neq u_0(s_1)$ la función toma valores distintos $u_0(s_1)$, $u_0(s_2)$, en el punto (x^*, y^*), luego la función es multivaluada y *salta*. Hay una discontinuidad. En particular se puede demostrar que u está condenada a *saltar* y a presentar una singularidad en su derivada parcial primera si u_0 tiene soporte compacto (se entiende por ello que la función es nula fuera de un intervalo compacto), excepto en el caso trivial donde $u_0(s) \equiv 0$.

Pensando en y como en el tiempo, para las funciones $u_0(s)$ (dato inicial del problema) tales que $y^* > 0$, esta relación define (y predice) la existencia de un instante de explosión (singularidad) de la solución. Es fácil comprobar en efecto que, si $u_0'(s) < 0$, entonces $y^* > 0$. Para todo este tipo de datos iniciales la solución $u(x, y)$ será discontinua en algún tiempo $y^* > 0$. La naturaleza de esta singularidad será más clara si consideramos los valores de la derivada $u_x(x, y)$ a lo largo de la característica $x = s + u_0(s)y$. A partir de la expresión (implícita) de la solución:

$$u = u_0(x - uy),$$

y aplicando derivación implícita se tiene:

$$u_x = [u_0'(x - uy)][(1 - u_x y)] = u_0' - u_x u_0' y,$$

y se deduce,

$$u_x = \frac{u_0'(s)}{1 + u_0'(s)y},$$

luego si $u_0'(s) < 0$, entonces $u_x \to \infty$ para

$$y = -\frac{1}{u_0'(s)}.$$

El tiempo y más pequeño para el cual esta relación se cumple corresponde al valor $s = s_0$ en el cual $u_0'(s)$ tiene un mínimo (negativo). En el instante

$T = y = -1/u_0'(s)$ la solución tiene un crecimiento incontrolable y su derivada explota (se pone vertical). No puede existir una solución de clase C^1 más allá del instante T. Nótese que este tipo de comportamiento es típico de las ecuaciones no lineales.

Es posible, sin embargo, definir unas soluciones débiles del PVI que existen más allá del tiempo T. Para ello, debemos de dar un sentido a la EDP cuasi-lineal para una clase de funciones más amplia que C^1 (o incluso continuas). Para ello se escribe la EDP

$$u_y + uu_x = 0,$$

en forma de divergencia (es decir mediante el operador de divergencia aplicado a funciones adecuadas),

$$\frac{\partial R(u)}{\partial y} + \frac{\partial S(u)}{\partial x} = 0, \tag{1.26}$$

siendo $R(u)$ y $S(u)$ funciones tales que $S'(u) = uR'(u)$. Por ejemplo, podríamos elegir $R(u) = u$ y $S(u) = (1/2)u^2$. Nótese que se trata de una elección natural pues:

$$u_y + uu_x = u_y + \frac{1}{2}(u^2)_x = \operatorname{div}\left(u, \frac{1}{2}u^2\right) = \operatorname{div}(R(u), S(u),) = \frac{\partial R(u)}{\partial y} + \frac{\partial S(u)}{\partial x} = 0.$$

Integrando la relación (1.26) en $x \in (a, b)$, se tiene la **ley de conservación**:

$$0 = \frac{d}{dy}\int_a^b R(u(x,y))dx + S(u(b,y)) - S(u(a,y)). \tag{1.27}$$

Por otra parte cualquier función de clase C^1 que verifica (1.27) es solución de (1.26). Ahora bien, la ecuación (1.27) tiene sentido para funciones mucho más generales y puede ser utilizada para definir las **soluciones débiles** de (1.26). En particular consideramos el caso donde u es una solución C^1 de (1.26) en cada una de dos regiones del plano separadas por una curva $x = \xi(y)$ al cruzar la cual la solución experimentará un salto (choque). Denotando los valores límites de u a la izquierda y la derecha de la curva por u^- y u^+, respectivamente, se tiene (a partir de (1.27) que:

$$\begin{aligned}
0 &= S(u(b,y)) - S(u(a,y)) + \frac{d}{dy}\left(\int_a^\xi R(u(x,y))dx + \int_\xi^b R(u(x,y))dx\right) \\
&= S(u(b,y)) - S(u(a,y)) + \xi' R(u^-) - \xi' R(u^+) - \int_a^\xi \frac{\partial S(u)}{\partial x}dx - \int_\xi^b \frac{\partial S(u)}{\partial x}dx \\
&= -[R(u^+) - R(u^-)]\xi' - S(u^-) + S(u^+).
\end{aligned}$$

Hemos encontrado la **condición de salto** (*shock conditions*):

$$\frac{\partial \xi}{\partial y} = \frac{S(u^+) - S(u^-)}{R(u^+) - R(u^-)}, \tag{1.28}$$

que relaciona la velocidad de propagación $d\xi/dy$ de la discontinuidad con el tamaño del salto de R y S. Nótese que (1.28) depende no sólo de la EDP sino también de la elección hecha ($R(u) = u$ y $S(u) = (1/2)u^2$) entre las funciones que satisfacen $S'(u) = uR'(u)$.

Con esta elección de las funciones S y R y para el dato inicial:

$$u(x,0) = u_0(x) = \begin{cases} -\dfrac{2}{3}\sqrt{3x} & x > 0, \\ 0 & x < 0, \end{cases}$$

es posible verificar que la función

$$u(x,y) = \begin{cases} -\dfrac{2}{3}\left(y + \sqrt{3x + y^2}\right) & 4x + y^2 > 0, \\ 0 & 4x + y^2 < 0, \end{cases}$$

es una solución débil de la ecuación en forma de divergencia.

* Propagación de discontinuidades en EDP de primer orden

Se considera la ecuación de primer orden, lineal, no homogénea de coeficientes constantes:

$$\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = 1, \qquad y \geq 0, \quad -\infty < x < \infty,$$

donde u es conocida en los puntos $(x_0, 0)$ del eje x: $u(x_0, 0) = u_0(x)$. La dirección característica viene dada por $dx = dy$ luego la característica que pasa por el punto $(x_0, 0)$ es (por integración directa):

$$y = x - x_0.$$

A lo largo de la característica se tiene que $du = dy$ luego por integración directa deducimos que la solución a lo largo de esta recta es:

$$u(x,y) = u_0(x - y) + y.$$

Consideraremos dos posibles situaciones: que los datos iniciales sean discontinuos o que lo sean sus derivadas. Seguiremos la exposición que aparece en el libro de Smith, en el capítulo dedicado a las ecuaciones hiperbólicas y sus características.

* **Valores iniciales discontinuos**

Supongamos tener un dato inicial del tipo

$$u(x,0) = u_0(x) = \begin{cases} f_1(x) & -\infty < x < x_A, \\ \\ f_2(x) & x_A < x < \infty, \end{cases}$$

y sean $x_p < x_A < x_q$, números reales. A la izquierda de la característica $y = x - x_A$ la solución u_I es $u_I = f_1(x_p) + y$ a lo largo de $y = x - x_p$. A la derecha de la línea recta $y = x - x_A$ la solución u_D es $u_D = f_2(x_q) + y$ a lo largo de $y = x - x_q$. Luego para el mismo valor de y en las dos soluciones se tiene:

$$u_I - u_D = f_1(x_p) - f_2(x_q).$$

Si nos acercamos por ambos lados a x_A, vemos que la cantidad $u_I - u_D$ es discontinua a lo largo de $y = x - x_A$ cuando

$$\lim_{x_p \to x_A} f_1(x_p) \neq \lim_{x_q \to x_A} f_2(x_q).$$

Esto muestra que cuando los valores iniciales son discontinuos en un punto x_A entonces la solución es discontinua a lo largo de la característica que pasa por x_A. Además el efecto de esta discontinuidad no disminuye según nos vayamos alejando de x_A a lo largo de la característica. Con las EDP parabólicas y elípticas el efecto de prescribir una discontinuidad inicial es muy diferente pues tiende a localizarse y a disminuir rapidamente al aumentar la distancia al punto de discontinuidad.

* **Derivadas iniciales discontinuas**

Supongamos tener ahora un dato inicial del tipo

$$u(x,0) = u_0(x) = \begin{cases} 0, & -\infty < x \leq 0, \\ \\ x, & 0 < x < \infty. \end{cases}$$

Evidentemente la función (derivada del dato inicial):

$$\frac{\partial u}{\partial x}(x,0) = \begin{cases} 0, & -\infty < x < 0, \\ \\ 1, & 0 < x < \infty, \end{cases}$$

es discontinua en $(0,0)$. Utilizando la ecuación vemos que también

$$\frac{\partial u}{\partial y}(x,0),$$

es discontinua en $(0,0)$. Por otra parte $u(x,0)$ es continua en $(0,0)$. Como antes se deduce que la solución u a lo largo de la característica $y = x - x_p$ que pasa por el punto $(x_p, 0)$ es

$$u - u_p = y = x - x_p.$$

La solución u_I a la izquierda de $y = x$ es

$$u_I = u_0(x-y) + y = 0 + y = y, \qquad -\infty < x \leq 0, \quad y \geq 0,$$

y la solución u_D a la derecha de $y = x$ es

$$u_D = u_0(x-y) + y = x - y + y = x, \qquad 0 < x < \infty, \quad y \geq 0.$$

Se deduce que $u_I = u_D$ a lo largo de $y = x$, pero que

$$\frac{\partial u_I}{\partial x} = 0, \qquad \frac{\partial u_D}{\partial x} = 1,$$

es decir, la solución es continua a lo largo de la característica que pasa por el punto de discontinuidad pero las discontinuidades iniciales de las derivadas parciales se propagan con velocidad constante a lo largo de esta característica.

* Las ecuaciones de Pfaff

Si en las secciones anteriores hemos visto la relación existente entre determinar las líneas vectoriales (y las superficies vectoriales de un campo) y resolver una EDP de primer orden lineal no homogénea veremos ahora las denominadas **ecuaciones de Pfaff** cuya resolución permite determinar la familia de superficies ortogonales a las líneas vectoriales. En efecto, la ecuación de éstas superficies tiene la forma $\boldsymbol{F} \cdot \boldsymbol{T} = 0$, siendo $\boldsymbol{T}$ un vector contenido en el plano tangente a las superficies buscadas:

$$\boldsymbol{T} = dx\boldsymbol{i} + dy\boldsymbol{j} + du\boldsymbol{k} = (dx, dy, du).$$

Si $\boldsymbol{F} = P\boldsymbol{i} + Q\boldsymbol{j} + R\boldsymbol{k} = (P, Q, R)$, entonces $\boldsymbol{F} \cdot \boldsymbol{T} = (P, Q, R) \cdot (dx, dy, du)$ nos conduce a

$$P(x,y,u)dx + Q(x,y,u)dy + R(x,y,u)du = 0. \tag{1.29}$$

Existen dos casos posibles dependiendo de si el campo vectorial $\boldsymbol{F} = P\boldsymbol{i} + Q\boldsymbol{j} + R\boldsymbol{k}$ es potencial o no. En el primer caso (si el campo es potencial) entonces $\boldsymbol{F} = \nabla S$, es decir,

$$P = \frac{\partial S}{\partial x}, \qquad Q = \frac{\partial S}{\partial y}, \qquad P = \frac{\partial S}{\partial z},$$

y las superficies buscadas son superficies de nivel $S(x,y,z) = c$ de la función potencial S. En este caso, la determinación de éstas no representa dificultad, puesto que

$$S = \oint_{(x_0,y_0,z_0)}^{(x,y,z)} Pdx + Qdy + Rdz,$$

donde la integral curvilínea se toma por cualquier camino entre el punto escogido (x_0, y_0, z_0) y el punto con coordenadas variables (x, y, z), por ejemplo, por la línea quebrada compuesta por segmentos de recta paralelos a los ejes coordenados.

Ejemplo 1.2.5 *Sea dada la ecuación de Pfaff*

$$(6x + yz)dx + (xz - 2y)dy + (xy + 2z)dz = 0.$$

Determinar la familia de superficies ortogonales a las líneas vectoriales del campo asociado.

El campo vectorial asociado es:

$$\boldsymbol{F} = (6x + yz)\boldsymbol{i} + (xz - 2y)\boldsymbol{j} + (xy + 2z)\boldsymbol{k},$$

luego rot$\boldsymbol{F} \equiv \boldsymbol{0}$. Entonces, $\boldsymbol{F} = \nabla S$, siendo:

$$S = \oint_{(x_0,y_0,z_0)}^{(x,y,z)} Pdx+Qdy+Rdz = \oint_{(0,0,0)}^{(x,y,z)} (6x+yz)dx+(xz-2y)dy+(xy+2z)dz.$$

Tomemos como camino de integración una línea quebrada con segmentos paralelos a los ejes de coordenadas. Integrando obtenemos:

$$S = 3x^2 - y^2 + z^2 + xyz.$$

Por tanto, la integral buscada será $S = c$, es decir,

$$3x^2 - y^2 + z^2 + xyz = c.$$

Si el campo no es potencial, en ciertos casos se puede escoger un factor escalar $\mu(x,y,z)$ tal que $\mu\boldsymbol{F} = \nabla S$ es potencial.

Ejemplo 1.2.6 *Sea dada la ecuación de Pfaff:*

$$dx + e^{-x}dy = 0.$$

Determinar la familia de superficies ortogonales a las líneas vectoriales del campo asociado.

El campo vectorial asociado es:

$$\boldsymbol{F} = \boldsymbol{i} + e^{-x}\boldsymbol{j} = (1, e^{-x}, 0).$$

Calculando su rotacional se tiene:

$$\text{rot}\boldsymbol{F} = (0, 0, -e^{-x}) \neq \boldsymbol{0},$$

luego el campo $\boldsymbol{F}$ no deriva de un potencial. Sin embargo si multiplicamos el campo $\boldsymbol{F}$ por la función escalar $\mu(x, y, z) = e^x$ se tiene que el campo:

$$\boldsymbol{G} = \mu\boldsymbol{F} = e^x\boldsymbol{i} + \boldsymbol{j} = (e^x, 1, 0).$$

es irrotacional (verificarlo por ejercicio), luego

$$\boldsymbol{G} = \mu\boldsymbol{F} = \nabla S$$

es potencial, siendo

$$S = \oint_{(x_0,y_0,z_0)}^{(x,y,z)} Pdx + Qdy + Rdz = \oint_{(0,0,0)}^{(x,y,z)} e^x dx + dy.$$

Integrando se obtiene:

$$S = e^x + y.$$

Por tanto, la integral buscada será $S = c$, es decir,

$$e^x + y = c.$$

Observación 1.2.1 *Observamos que las líneas vectoriales del campo* $\boldsymbol{F} = (1, e^{-x}, 0)$, *son exactamente iguales a las líneas vectoriales del campo* $\boldsymbol{G} = (e^x, 1, 0)$, *puesto que verifican la misma ecuación característica:*

$$\frac{dx}{1} = \frac{dy}{e^{-x}}.$$

Nótese aquí la relación entre las ecuaciones de Pfaff y las ecuaciones diferenciales exactas (véase las secciones 2.5, 2.6 del capítulo 13 del guión de primero). Las ecuaciones de Pfaff se pueden integrar sólo si el campo vectorial de coeficientes asociado a la ecuación de Pfaff es potencial o si existe una función (factor integrante de la ecuación) tal que el producto de esta función escalar por el campo vectorial es un campo potencial. En estos casos la ecuación de Pfaff es una EDO exacta (o reducible a una EDO exacta) y es directamente resoluble.

Se puede demostrar que la **condición necesaria y suficiente** para que exista un factor integrante μ es que se cumpla la ecuación:

$$\boldsymbol{F} \cdot \text{rot}\boldsymbol{F} = 0,$$

es decir, que los vectores $\boldsymbol{F}$ y rot$\boldsymbol{F}$ sean ortogonales. Si esta condición (llamada condición de **integración total**) no se cumple, entonces no existe ninguna familia de superficies $S(x, y, z) = c$ ortogonales a las líneas vectoriales del campo $\boldsymbol{F}(x, y, z)$. Nótese que en el ejemplo (1.2.6) la condición $\boldsymbol{F} \cdot \text{rot}\boldsymbol{F} = 0$ se cumple.

Ejemplo 1.2.7 *Dada la ecuación de Pfaff*

$$zdx + (x - y)dy + zydz = 0,$$

se verifica que la condición $\boldsymbol{F} \cdot rot\boldsymbol{F} = 0$*, donde*

$$\boldsymbol{F} = z\boldsymbol{i} + (x - y)\boldsymbol{j} + zy\boldsymbol{k}$$

no se cumple.

En este caso la ecuación de Pfaff no es integrable directamente y hace falta utilizar unas técnicas que desbordan los objetivos del curso. Más detalles se pueden encontrar en el libro de Elsgoltz [25].

1.2.2. Introducción progresiva de las EDP hiperbólicas de primer orden

Se entiende por introducción progresiva la presentación de distintos modelos de dificultad creciente para el entendimiento del proceso de resolución de EDP de tipo hiperbólico. Un libro de texto interesante en este sentido es el libro de Strikwerda[26].

Empezaremos considerando la estructura de un problema de valor inicial para una ecuación en derivada parciales de primer orden lineal o cuasilineal, homogénea o no homogénea. Extenderemos el análisis al caso de sistemas de EDP de primer orden y discutiremos el importante papel de las características en el proceso de resolución de tales problemas. Tras considerar, en las secciones siguientes, los problemas de contorno (en dominios espacialmente acotados), pasaremos al estudio de los problemas de valor inicial y de contorno, definiendo los conceptos de problemas bien puestos y problemas mal puestos.

El caso lineal homogéneo con coeficientes constantes

El prototipo de todas las EDP hiperbólicas de primer orden (en particular de las homogéneas con coeficientes constantes) es la ecuación de onda uni-direccional (conocida también con el nombre de **ecuación de advección** o de convección):

[25] Elsgoltz, L. (1983). Ecuaciones diferenciales y cálculo variacional. III Ed. Editorial Mir.

[26] Strikwerda J. C., Finite Difference Schemes and Partial Differential Equations. Chapman & Hall. International Thomson Publishing. 1989

$$u_t + V u_x = 0, \tag{1.30}$$

siendo V una constante: $V \in \mathbb{R}$. Los subíndices denotan, como siempre, los operadores de derivación parcial $u_x = \partial u/\partial x$ y $u_t = \partial u/\partial t$. Un **problema de valores iniciales** consiste en prescribir el comportamiento de u en un instante inicial, $t = 0$, digamos $u(x, 0) = u_0(x)$, $\forall\, x \in \mathbb{R}$ y en determinar los valores de $u(x, t)$, $\forall\, x \in \mathbb{R}$, $\forall\, t > 0$. Es decir, determinar $u(x, t)$ tal que:

$$\begin{cases} \dfrac{\partial u}{\partial t} + V\dfrac{\partial u}{\partial x} &= 0 \qquad 0 < x < \infty, \quad t > 0, \\ u(x,0) &= u_0(x), \quad 0 < x < \infty, \quad t = 0. \end{cases} \tag{1.31}$$

Es fácil comprobar que la solución del problema de valor inicial (1.31) es:

$$u(x, t) = u_0(x - Vt). \tag{1.32}$$

Se puede además demostrar (utilizando los resultados de la sección (1.2.1), concretamente el teorema (1.2.2)) que la solución dada es la única solución del problema de valor inicial. En efecto, parametrizamos la curva inicial Γ_0:

$$x = f(s) = s, \qquad y = g(s) = 0, \qquad u = u_0(s),$$

e identificamos $\boldsymbol{F} = (P, Q, R) = (V, 1, 0)$. Se tiene entonces que

$$\Delta = Qf' - Pg' = 1 \neq 0.$$

La fórmula de la solución (1.32) puede decirnos varias cosas: en primer lugar nos dice que en cualquier instante la solución es una copia del dato inicial obtenida por simple traslación hacia la derecha si $V > 0$ y la izquierda si $V < 0$. Otra forma de verlo consiste en observar que la solución depende sólo de $\xi = x - Vt$. Las rectas del plano (x, t) donde ξ es constante se llaman características. El parámetro V tiene las dimensiones de una distancia dividida por el tiempo y se llama la velocidad de propagación a lo largo de las características. Es decir que se puede interpretar la solución como una onda que se propaga con velocidad constante V sin cambiar de forma a lo largo de la dirección x (de ahí el nombre de ecuación de onda unidimensional). Un segundo aspecto relevante de la fórmula $u(x, t) = u_0(x - Vt)$ es que tiene sentido aunque u_0 no sea diferenciable mientras la ecuación parece tener sentido sólo si u es diferenciable. En general, se admitirán soluciones discontinuas para las EDP hiperbólicas[27].

[27] Un ejemplo de solución discontinua es una onda de choque (shock wave), típica de las ecuaciones no lineales. Véase la sección dedicada a las soluciones discontinuas.

El caso lineal no homogéneo con coeficientes constantes

Para ilustrar con más detalle el concepto de característica consideraremos ahora la EDP hiperbólica más general:

$$u_t + Vu_x + bu = f(x,t), \tag{1.33}$$

siendo V y b constantes, complementada con la *condición inicial* $u(0,x) = u_0(x)$. Se trata por tanto de resolver el problema de valor inicial

$$\begin{cases} \dfrac{\partial u}{\partial t} + V\dfrac{\partial u}{\partial x} + bu &= f(x,t) \quad 0 < x < \infty, \quad t > 0, \\ u(x,0) &= u_0(x), \quad 0 < x < \infty, \quad t = 0. \end{cases} \tag{1.34}$$

Si $b = 0$ y $f \equiv 0$ recuperamos (1.31). Un cambio de variables lineal del tipo

$$\tau = t, \quad \xi = x - Vt \qquad \Longrightarrow \qquad t = \tau, \quad x = \xi + V\tau,$$

junto a

$$u(x,t) = u(\xi + V\tau, \tau) = \tilde{u}(\xi, \tau),$$

(las variables $\tilde{u}$ y u representan la misma función pero la tilda es necesaria para distinguir entre los dos sistemas ((x,t) y (ξ,τ)) de coordenadas para las variables independientes) permite escribir (1.33) en la forma:

$$\begin{cases} \dfrac{\partial \tilde{u}}{\partial \tau} + b\tilde{u} &= f(\xi + V\tau, \tau) \quad -V\tau < \xi < \infty, \quad \tau > 0, \\ \tilde{u}(\xi, 0) &= u_0(\xi), \qquad\qquad 0 < \xi < \infty, \quad \tau = 0, \end{cases} \tag{1.35}$$

que es una EDO en τ (aunque ξ sea sólo un parámetro, la función $\tilde{u}$ depende formalmente de las dos variables (τ, ξ); de ahí el significado del símbolo de derivación parcial aún cuando se puede considerar la ecuación en el sentido de una EDO) que se complementa con la condición inicial $\tilde{u}(\xi, 0) = u_0(\xi)$. El problema de valor inicial asociado se puede resolver por la fórmula de variación de las constantes (véase la sección correspondiente de los guiones de primer curso para la deducción de la fórmula mediante el método de Lagrange):

$$\tilde{u}(\xi,\tau) = u_0(\xi)e^{-b\tau} + \int_0^\tau f(\xi + V\sigma, \sigma)e^{-b(\tau-\sigma)}d\sigma.$$

Volviendo a las variables originales tenemos una representación para la solución dada por:

$$u(x,t) = u_0(x - Vt)e^{-bt} + \int_0^t f(x - V(t-s), s)e^{-b(t-s)}ds. \tag{1.36}$$

A partir de (1.36) se deduce que $u(x,t)$ depende sólo del valor de $u_0(x)$ en el punto x^* tal que $x^* = x - Vt$ y del valor de $f(x,t)$ en todos los puntos de la característica que pasa por $(x^*, 0)$ para $0 \leq t' \leq t$.

Este método de resolución se puede fácilmente extender a ecuaciones del tipo:

$$u_t + Vu_x = f(x,t,u), \qquad u(x,0) = u_0(x), \tag{1.37}$$

complementadas con $u(x,0) = u_0(x)$. Se puede mostrar (se deja como ejercicio propuesto) que (1.37) es equivalente a la familia de P.V.I.:

$$\frac{\partial \tilde{u}}{\partial \tau} = f(\xi + V\tau, \tau, \tilde{u}), \tag{1.38}$$

con $\tilde{u}(\xi, 0) = u_0(\xi)$. La relación entre la solución del problema original (1.37) y la solución del problema (1.38) viene dada por:

$$u(x,t) = \tilde{u}(x - Vt, t).$$

No siempre es posible resolver explícitamente el problema (1.38). En tales casos es necesario acudir a métodos numéricos.

1.2.3. Sistemas hiperbólicos con coeficientes constantes

Consideraremos sistemas hiperbólicos espacialmente unidimensionales, $\vec{x} = x \in I\!R$, con coeficientes constantes. La variable $\vec{U}$ es un vector de dimensión d:

$$\vec{U} = \vec{U}(x,t) = (u_1(x,t), ..., u_d(x,t)) \in I\!R^d, \quad d \geq 2.$$

Definición 1.2.3 *Un sistema de la forma:*

$$\vec{U}_t + [A]\vec{U}_x + [B]\vec{U} = \vec{F}(x,t), \tag{1.39}$$

se dice que es **hiperbólico** *si la matriz* $[A]$ *es diagonalizable con autovalores reales. Los autovalores* a_i *de* $[A]$ *se denominan* **velocidades características** *del sistema.*

Recuérdese que si $[A]$ es diagonalizable, entonces existe una matriz no singular $[P]$ tal que $[D] = [P^{-1}][A][P]$, siendo $[D]$ una matriz diagonal.

En el caso particular de que $[B] \equiv 0$, definiendo entonces el cambio de variables $\vec{W} = [P^{-1}]\vec{U}$, se tiene:

$$\vec{W}_t + [D]\vec{W}_x = [P^{-1}]\vec{F}(x,t) = \vec{G}(t,x),$$

es decir,

$$w_t^i + \lambda_i w_x^i = g^i(t,x), \qquad i = 1...d,$$

que es de la forma (1.33) para las componentes de los vectores $\vec{W} = (w^1, ...w^d)$, $\vec{G} = (g^1, ...g^d)$ y los autovalores λ_i, $i = 1, .., d$ de la matriz del sistema $[A]$.

Por tanto, cuando $[B] \equiv 0$ el sistema hiperbólico se reduce a un sistema desacoplado de ecuaciones hiperbólicas escalares independientes que se resuelven por separado. Consideraremos esta situación en el siguiente ejemplo, donde supondremos $[B] \equiv 0$ y $\vec{F}(x,t) \equiv 0$ (caso homogéneo).

Ejemplo 1.2.8 *Se considera el PVI dado por el sistema hiperbólico:*

$$\begin{cases} u_t + 2u_x + v_x &= 0, \\ \\ v_t + u_x + 2v_x &= 0, \end{cases}$$

y las condiciones iniciales:

$$u(x,0) = u_0(x) = \begin{cases} 1 & |x| \leq 1 \\ \\ 0 & |x| > 1 \end{cases}, \qquad v(x,0) = v_0(x) = 0.$$

Encontrar su solución.

Se utiliza la teoría matricial, concretamente la técnica de diagonalización de matrices cuadradas.

El sistema se puede escribir en la forma:

$$\begin{pmatrix} u \\ \\ v \end{pmatrix}_t + \begin{pmatrix} 2 & 1 \\ \\ 1 & 2 \end{pmatrix} \begin{pmatrix} u \\ \\ v \end{pmatrix}_x = 0,$$

siendo $\vec{U} = (u, v)$. Es decir,

$$\vec{U}_t + [A]\vec{U}_x = 0, \qquad [A] = \begin{pmatrix} 2 & 1 \\ \\ 1 & 2 \end{pmatrix}.$$

Es fácil ver que la matriz $[A]$ (que es simétrica) tiene autovalores

$$\lambda_1 = 3, \qquad \lambda_2 = 1.$$

Los autovectores son $\vec{\mu}_1 = (1,1)$ y $\vec{\mu}_2 = (1,-1)$. La matriz es diagonalizable y el sistema se puede desacoplar mediante el cambio $\vec{W} = [P^{-1}]\vec{U}$. Nótese que

la matriz $[P]$ viene dada por (escribiendo en columna las componentes de los autovectores $\vec{\mu}_1$ y $\vec{\mu}_2$)

$$[P] = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad [P^{-1}] = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix},$$

luego,

$$\begin{pmatrix} w^1 \\ w^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(u+v) \\ \frac{1}{2}(u-v) \end{pmatrix}.$$

Un simple razonamiento nos confirma que el cambio de coordenadas a efectuar es el anterior. En efecto, sumando y restando las dos ecuaciones:

$$u_t + 2u_x + v_x = 0, \qquad v_t + u_x + 2v_x = 0,$$

el sistema se puede reescribir en la forma:

$$\begin{aligned} \frac{1}{2}\left[(u+v)_t + 3(u+v)_x\right] &= 0, \qquad 0 < x < \infty, \\ \frac{1}{2}\left[(u-v)_t + (u-v)_x\right] &= 0, \qquad 0 < x < \infty, \end{aligned}$$

y definiendo $w^1 = u+v$, $w^2 = u-v$, se tienen los dos PVI para las componentes (w^1, w^2).

$$\left\{ \begin{array}{rcl} w^1_t + 3w^1_x &=& 0 \\ w^1(0,x) &=& \frac{1}{2}u_0(x) \end{array} \right., \qquad \left\{ \begin{array}{rcl} w^2_t + w^2_x &=& 0 \\ w^2(0,x) &=& \frac{1}{2}u_0(x) \end{array} \right., \tag{1.40}$$

donde las condiciones de contorno se deducen observando que $v_0(x) = 0$. La solución es por tanto,

$$w^1(t,x) = w^1_0(x-3t), \qquad w^2(t,x) = w^2_0(x-t).$$

En términos de las componentes originales $(u(x,t), v(x,t))$ se tiene:

$$\begin{aligned} u(x,t) &= w^1 + w^2 = \frac{1}{2}[u_0(x-3t) + u_0(x-t)], \\ v(x,t) &= w^1 - w^2 = \frac{1}{2}[u_0(x-3t) - u_0(x-t)]. \end{aligned}$$

En las hipótesis $[B] \equiv 0$, $F(x,t) \equiv 0$, el proceso de resolución de un sistema hiperbólico del tipo (1.39) se puede resumir (utilizando una notación matricial) en los siguientes pasos:

MÉTODO DE RESOLUCIÓN DE SISTEMAS HOMOGÉNEOS

1. Resolver el sistema hiperbólico $\vec{U}_t + [A]\vec{U}_x = 0$, siendo $\vec{U} = (u, v)$ un vector y $[A]$ una matriz cuadrada.

2. Diagonalizar la matriz $[A]$ y obtener la matriz $[P]$ tal que $[D] = [P^{-1}][A][P]$. Recuérdese que siempre es posible por definición de sistema hiperbólico.

3. Observar que $[A] = [P][D][P^{-1}]$, luego debemos resolver

$$\vec{U}_t + [A]\vec{U}_x = \vec{U}_t + [P][D][P^{-1}]\vec{U}_x = 0,$$

   de donde se deduce la ecuación matricial

$$[P]\vec{U}_t + [D][P]\vec{U}_x = 0.$$

4. Definir $\vec{W} = [P^{-1}]\vec{U}$ (es decir se utiliza la matriz de paso $[P]$ para obtener las coordenadas en las cuales $[A]$ es diagonalizable).

5. Resolver el sistema hiperbólico desacoplado $\vec{W}_t + [D]\vec{W}_x = 0$ resolviendo una por una las ecuaciones hiperbólicas escalares y los PVI asociados.

Si $[B] \neq 0$ el sistema es acoplado. El efecto de los términos $[B]\vec{U}$ puede ser aquel de generar crecimiento, decaimiento u oscilaciones en la solución pero no modifica la propiedad cualitativa de propagación a lo largo de las características típica de los problemas hiperbólicos. La definición de sistema hiperbólico para una dimensión genérica espacial, $\vec{x} = (x_1, .., x_n) \in I\!R^n$, $n > 1$, se encuentra en el cap. 9 del libro de Strikwerda.

1.2.4. Ecuaciones hiperbólicas lineales homogéneas con coeficientes variables

Consideraremos unas ecuaciones hiperbólicas con coeficientes variables para las cuales la velocidad característica es una función del tiempo y del espacio. En concreto, sea dada la ecuación

$$u_t + a(x,t)u_x = 0, \tag{1.41}$$

complementada con la condición inicial $u(x,0) = u_0(x)$, cuya velocidad de propagación es $a(x,t)$. Tal y como hicimos con la ecuación (1.33) (coeficientes constantes) introducimos las variables τ y ξ, siendo $t = \tau$ y dejando ξ todavía indeterminada. Se tiene:

$$\frac{\partial \tilde{u}}{\partial \tau} = \frac{\partial t}{\partial \tau} u_t + \frac{\partial x}{\partial \tau} u_x = u_t + \frac{\partial x}{\partial \tau} u_x.$$

Siguiendo la analogía con el caso de coeficientes constantes, ponemos:

$$\frac{dx}{d\tau} = a(x,t) = a(x,\tau),$$

que es una EDO para x que nos da la velocidad a lo largo de la característica a través del punto (x,τ) como $a(x,\tau)$. Prescribimos además el valor inicial para la curva característica que pasa por el punto (x,τ) como ξ (de esta forma queda determinado el cambio de variables). Luego la ecuación (1.41) es equivalente al PVI constituido por el sistema de EDO y las condiciones iniciales:

$$\left\{ \begin{array}{rcll} \dfrac{d\tilde{u}}{d\tau} & = & 0, & \tilde{u}(\xi,0) = u_0(\xi) \\ \dfrac{dx}{d\tau} & = & a(x,\tau), & x(0) = x_0 = \xi, \end{array} \right. \tag{1.42}$$

donde ξ es un parámetro que permite *recorrer* con continuidad las características de la EDP (es constante a lo largo de cada característica pero varía pasando de una a otra).

Tal y como se deduce a partir de la primera ecuación del sistema, u es constante a lo largo de cada curva característica, pero la característica no tiene por qué ser una recta (lo sería si a fuese constante).

Ejemplo 1.2.9 (Problema de valor inicial) *Se considera el PVI:*

$$\left\{ \begin{array}{rcl} u_t + x u_x & = & 0, \\ u(x,0) & = & u_0(x), \end{array} \right.$$

siendo la condición inicial:

$$u(x,0) = u_0(x) = \left\{ \begin{array}{ll} 1 & 0 \leq x \leq 1, \\ 0 & x \notin [0,1], \end{array} \right.$$

Encontrar su solución.

Utilizando (1.42) se tienen las ecuaciones

$$\frac{d\tilde{u}}{d\tau} = 0, \qquad \frac{dx}{d\tau} = x.$$

La solución general de la EDO para $x(\tau)$ es $x(\tau) = ce^{\tau}$. Puesto que $x(0) = \xi$ se tiene $x(\tau) = \xi e^{\tau}$ es decir,

$$\xi = xe^{-\tau}.$$

La ecuación para la $\tilde{u}$ muestra que $\tilde{u}$ es independiente de τ y utilizando la condición inicial se tiene:

$$\tilde{u}(\tau, \xi) = u_0(\xi),$$

luego,

$$u(x,t) = \tilde{u}(\tau, \xi) = u_0(\xi) = u_0(xe^{-t}),$$

y se tiene, para $t > 0$,

$$u(x,t) = \begin{cases} 1, & 0 \leq x \leq e^t, \\ 0, & x \notin [0, e^t]. \end{cases}$$

Observación 1.2.2 *Nótese que el dato inicial del problema, $u_0(x)$ es una función de soporte compacto*[28]*, siendo el soporte el intervalo $[0,1]$. Para cada instante de tiempo fijado, digamos t^*, también la solución tiene soporte compacto $[0, e^{t^*}]$. La solución representa por tanto una propagación del dato inicial a lo largo de las características. El soporte se dilata pero sigue siendo compacto para cada instante. Esta es una propiedad cualitativa típica de las ecuaciones de tipo hiperbólico.*

Ejemplo 1.2.10 (Problema de valor inicial) *Se considera el PVI:*

$$\begin{cases} u_t + xu_x &= u, \\ u(x,0) &= u_0(x), \end{cases}$$

siendo la condición inicial:

$$u(x,0) = u_0(x) = \begin{cases} 1 & 0 \leq x \leq 1, \\ 0 & x \notin [0,1]. \end{cases}$$

Encontrar su solución.

[28] Se entiende por ello una función que es idénticamente nula por fuera de un intervalo cerrado y acotado de la recta real.

El problema es ahora no homogéneo. Se define $u(x,t) = \tilde{u}(\xi,\tau)$ y se aplica la regla de la cadena para obtener las ecuaciones:

$$\frac{d\tilde{u}}{d\tau} = \tilde{u}, \qquad \frac{dx}{d\tau} = x.$$

La solución general de la EDO para $x(\tau)$ es $x(\tau) = ce^{\tau}$. Puesto que $x(0) = \xi$ se tiene $x(\tau) = \xi e^{\tau}$, es decir,

$$\xi = xe^{-\tau}.$$

La ecuación para la $\tilde{u}$ muestra que:

$$\tilde{u}(\tau,\xi) = Ce^{\tau}.$$

Utilizando la condición inicial:

$$\tilde{u}(0,\xi) = u_0(\xi) = C,$$

luego $\tilde{u}(\tau,\xi) = u_0(\xi)e^{\tau}$ y por tanto,

$$u(x,t) = \tilde{u}(\tau,\xi) = u_0(\xi)e^{\tau} = u_0(xe^{-t})e^{t},$$

y se tiene, para $t > 0$,

$$u(x,t) = \begin{cases} e^t & 0 \le x \le e^t, \\ 0 & x \notin [0, e^t]. \end{cases}$$

Nótese finalmente que estos métodos se pueden extender también a ecuaciones no lineales (cuasilineales) de la forma:

$$u_t + a(x,t)u_x = f(x,t,u).$$

Las ecuaciones para las cuales las velocidades características dependen de u, es decir con velocidad característica $a(t,x,u)$, por ejemplo las ecuaciones del tipo:

$$u_t + a(x,t,u)u_x = f(x,t,u),$$

requieren un cuidado especial pues las curvas características se pueden intersecar. En este caso la solución toma valores distintos a lo largo de cada una de las características luego en el punto de intersección debe haber un salto. Hay una discontinuidad. Los gradientes de la solución se ponen verticales y tienden al infinito. Se genera un frente y hay singularidad en las derivadas parciales.

1.2.5. * Sistemas hiperbólicos con coeficientes variables

Consideraremos en esta sección los sistemas unidimensionales (espacialmente) de ecuaciones hiperbólicas de primer orden con coeficientes variables. Utilizaremos la misma notación adoptada en el caso de coeficientes constantes. Las matrices se entienden ahora como funciones matriciales.

Definición 1.2.4 *El sistema*

$$\vec{U}_t + [A(x,t)]\vec{U}_x + [B(x,t)]\vec{U} = \vec{F}(x,t),$$

complementado con la condición inicial $\vec{U}(x,0) = \vec{U}_0(x)$ *es* **hiperbólico** *si existe una función matricial* $[P(x,t)]$ *tal que*

$$[P^{-1}(x,t)][A(x,t)][P(x,t)] = [D(x,t)] = \begin{pmatrix} a_1(x,t) & & & 0 \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & a_d(x,t) \end{pmatrix}$$

es diagonal, con autovalores reales y las normas matriciales de $[P(x,t)]$, $[P^{-1}(x,t)]$ *están acotadas en* x *y* t *para* $x \in \mathbb{R}$, $t \geq 0$.

Las curvas características del sistema vienen dadas por las soluciones de las EDO:

$$\frac{dx^i}{dt} = a_i(x,t), \quad x^i(0) = \xi^i, \qquad i = 1,..,d.$$

Definiendo $\vec{W} = [P^{-1}(x,t)]\vec{U}$, obtenemos el sistema para la $\vec{W}$:

$$\vec{W}_t + [D(x,t)]\vec{W}_x = [P^{-1}(x,t)]\vec{F}(x,t) + [G(x,t)]\vec{W}$$

siendo $[G] = -[P]^{-1}([P]_t + [A][P]_x - [B][P])$.

1.2.6. * Problemas de valor inicial y de contorno para una EDP de primer orden con coeficientes constantes

Consideraremos ahora EDP hiperbólicas de primer orden en un intervalo finito en lugar de en toda la recta real. La mayoría de las aplicaciones de las EDP se plantean en dominios con una frontera y es muy importante prescribir correctamente los datos del problema a lo largo de la frontera, es decir las **condiciones de contorno**. El problema de determinar una solución de una ecuación diferencial cuando se prescriben datos iniciales y datos en el contorno se llama un **problema de valor inicial y de contorno**. El análisis de

este tipo de problemas mostrará nuevamente la importancia del concepto de característica.

En esta sección consideraremos los problemas de valor inicial y de contorno para EDP de primer orden (hiperbólicas) en una dimensión espacial.

Se considera la ecuación de onda unidimensional

$$u_t + V u_x = 0, \qquad 0 \leq x \leq 1, \quad t \geq 0. \tag{1.43}$$

Si $V > 0$ las características en esta región son líneas rectas que se propagan de izquierda a derecha. Podemos prescribir la solución en la frontera $x = 0$, junto a una condición inicial, para que la solución esté definida $\forall t \geq 0$. Por otra parte no podemos prefijar arbitrariamente la función en la otra frontera ($x = 1$) pues la solución podría estar sobre determinada (es decir podría no existir una solución que cumpla todas las condiciones impuestas).

Si especificamos un dato inicial

$$u(x, 0) = u_0(x),$$

y un dato de contorno

$$u(0, t) = g(t),$$

entonces la solución del problema de valor inicial y de contorno constituido por la ecuación (1.43) y las condiciones límite anteriores es:

$$u(x,t) = \begin{cases} u_0(x - Vt) & x - Vt > 0, \\ g(t - V^{-1}x) & x - Vt < 0. \end{cases}$$

A lo largo de la característica dada por $x - Vt = 0$ tendremos una discontinuidad de salto en u si $u_0(x = 0) \neq g(t = 0)$. Si $V < 0$ el papel de las dos fronteras se invierte y el análisis es similar.

Ejemplo 1.2.11 *Calcular la solución de la ecuación*

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \qquad 0 < x < \infty, \quad t > 0,$$

complementada con la condición de contorno:

$$u(0, t) = g(t) = 2t, \quad t > 0,$$

y la condición inicial:

$$u(x, 0) = u_0(x) = \begin{cases} x(x-2) & 0 \leq x \leq 2, \\ 2(x-2) & x \geq 2. \end{cases}$$

Puesto que $V \equiv 1$, se tiene que la solución es:

$$u(x,t) = \begin{cases} u_0(x-t) & x-t > 0, \\ g(t-x) & x-t < 0. \end{cases}$$

Nótese que en el origen $(x,t) = (0,0)$ se tiene $u_0(0) = g(0)$ y no hay discontinuidad de salto. Explícitamente (utilizando los datos del problema) la expresión de la solución es:

$$u(x,t) = \begin{cases} \left\{ \begin{array}{ll} (x-t)(x-t-2) & 0 < x-t \leq 2, \\ 2(x-t-2) & x-t \geq 2 \end{array} \right\} & x-t > 0, \\ 2(t-x) & x-t < 0. \end{cases}$$

Veamos el porqué. El sistema auxiliar es:

$$dt = dx = \frac{du}{0},$$

luego u es constante a lo largo de la línea recta característica $t = x + C$. Por tanto la ecuación de la característica que pasa por el punto $(x_0, 0)$ del eje x es $t = x - x_0$ y si $u(x,0) = u_0(x)$ entonces la solución a lo largo de esta característica es:

$$u(x,t) = u_0(x_0) = u_0(x-t).$$

Además, puesto que el dato inicial viene dado en $0 < x < \infty$, se tiene que $u_0(s)$ está definida para $s > 0$ luego la expresión anterior de la solución es válida sólo en la región tal que $s = x - t > 0$. De manera análoga, si $u(0,t) = g(t)$, la solución a lo largo de la característica $t - t_0 = x$ que pasa por el punto $(0, t_0)$ es:

$$u(x,t) = g(t_0) = g(t-x).$$

Puesto que el dato de contorno viene dado para $t > 0$, se tiene que $g(s)$ está definida para $s > 0$ luego la expresión anterior de la solución es válida sólo en la región tal que $s = t - x > 0$. Juntando las dos expresiones se tiene:

$$u(x,t) = \begin{cases} u_0(x-t) & x-t > 0, \\ g(t-x) & x-t < 0. \end{cases}$$

* Problemas de valor inicial y de contorno para sistemas de primer orden con coeficientes constantes

Consideremos ahora el sistema hiperbólico:

$$\begin{pmatrix} u^1 \\ u^2 \end{pmatrix}_t + \begin{pmatrix} a & b \\ b & a \end{pmatrix} \begin{pmatrix} u^1 \\ u^2 \end{pmatrix}_x = 0, \qquad 0 \le x \le 1, \quad t \ge 0. \tag{1.44}$$

Los autovalores (o velocidades características) del sistema son:

$$\lambda_1 = a + b, \qquad \lambda_2 = a - b.$$

Considereremos sólo el caso donde a y b son positivos. Si $0 < b < a$, entonces los autovalores son ambos positivos y las dos familias de características se propagan de izquierda a derecha. Esto significa que la solución (vector) $\vec{U} = (u^1, u^2)^T$ se debe prefijar en $x = 0$ y ningún dato se debe especificar en $x = 1$. Nótese que la pendiente de las características es el inverso de la velocidad (a^{-1}), luego las características más lentas serán aquellas con mayor pendiente. El caso más interesante se da cuando $0 < a < b$. Los autovalores son de signo contrario y las familias de características se propagan en direcciones opuestas. Escribimos el sistema en la forma:

$$\begin{pmatrix} u^1 + u^2 \\ u^1 - u^2 \end{pmatrix}_t + \begin{pmatrix} a+b & 0 \\ 0 & a-b \end{pmatrix} \begin{pmatrix} u^1 + u^2 \\ u^1 - u^2 \end{pmatrix}_x = 0,$$

donde las ecuaciones se tienen que verificar para $0 \le x \le 1$ y $t \ge 0$. Ciertamente una forma de determinar una única solución consiste en prescribir $u^1 + u^2$ en $x = 0$ y prescribir $u^1 - u^2$ en $x = 1$. Sin embargo existen otras posibilidades. Por ejemplo, cualquier condición de la forma

$$\begin{aligned} u^1 + u^2 &= \alpha_0(u^1 - u^2) + \beta_0(t) \qquad \text{en } x = 0, \\ u^1 - u^2 &= \alpha_1(u^1 + u^2) + \beta_1(t) \qquad \text{en } x = 1, \end{aligned} \tag{1.45}$$

determinará (únicamente) la solución del problema. Los coeficientes α_0, α_1 pueden ser funciones de t o constantes. Las condiciones de contorno que determinan una única solución se dicen **bien puestas** y el problema se dice **bien planteado**. Las condiciones (1.45) están bien puestas para el sistema (1.44). Además son las únicas condiciones bien puestas posibles para el sistema (1.44). Las condiciones de contorno (1.45) expresan el valor de las variables características en la característica *entrante*[29] en términos de la variable característica

[29] Por característica entrante se entiende una característica que entra en el dominio en la frontera considerada. Una característica saliente es una que deja el dominio.

saliente. Por tanto, si especificamos u^1 o u^2 en $x = 0$ el problema está bien planteado y las condiciones bien puestas. Especificar u^1 o u^2 en $x = 1$ también genera un problema bien planteado con condiciones bien puestas. Sin embargo, especificar $u^1 - u^2$ en $x = 0$ o especificar $u^1 + u^2$ en $x = 1$ origina un problema mal planteado pues las condiciones están mal puestas.

Para que un problema hiperbólico de valores iniciales y de contorno esté bien puesto, el número de condiciones de contorno debe ser igual al número de características *entrantes* en el dominio.

Ejemplo 1.2.12 *Considérese el sistema hiperbólico (1.44) en el intervalo* $[0, 1]$*, con* $a = 0$ *y* $b = 1$ *y las condiciones de contorno* $u^1(0, t) = 0$ *(en la frontera lateral izquierda) y* $u^1(1, t) = 1$ *(en la frontera lateral derecha). Determinar su solución siendo* $u^1(x, 0) = x$ *y* $u^2(x, 0) = 1$ *los datos iniciales.*

Las condiciones de contorno están bien puestas pues son del tipo (1.45) para los valores paramétricos $\alpha_0 = -1$, $\alpha_1 = -1$ y las funciones $\beta_0(t) \equiv 0$, $\beta_1(t) \equiv 2$.

Los autovalores (o velocidades características) del sistema son:

$$\lambda_1 = a + b = 1, \qquad \lambda_2 = a - b = -1.$$

Escribimos el sistema en la forma:

$$\begin{pmatrix} u^1 + u^2 \\ u^1 - u^2 \end{pmatrix}_t + \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} u^1 + u^2 \\ u^1 - u^2 \end{pmatrix}_x = 0,$$

donde las ecuaciones se tienen que verificar para $0 \leq x \leq 1$ y $t \geq 0$. Introduciendo, como antes, las componentes $w^1 = \frac{1}{2}(u^1 + u^2)$ y $w^2 = \frac{1}{2}(u^1 - u^2)$ se tiene:

$$\begin{pmatrix} w^1 \\ w^2 \end{pmatrix}_t + \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} w^1 \\ w^2 \end{pmatrix}_x = 0,$$

y nos hemos reconducido a resolver los dos PVI para las componentes (w^1, w^2):

$$\begin{cases} w_t^1 + w_x^1 &= 0, \\ w^1(x, 0) &= \frac{1}{2}(u_0^1(x) + u_0^2(x)) = \frac{1}{2}(x + 1), \end{cases}$$

$$\begin{cases} w_t^2 - w_x^2 &= 0, \\ w^2(x,0) &= \dfrac{1}{2}(u_0^1(x) - u_0^2(x)) = \dfrac{1}{2}(x-1), \end{cases} \tag{1.46}$$

donde las condiciones de contorno se deducen a partir de los datos del problema. La solución es por tanto:

$$w^1(x,t) = w_0^1(x-t) = \frac{1}{2}(x-t+1), \qquad w^2(x,t) = w_0^2(x+t) = \frac{1}{2}(x+t-1).$$

En términos de las componentes originales $(u(x,t), v(x,t))$ se tiene:

$$u^1(x,t) = w^1 + w^2 = w_0^1(x-t) + w_0^2(x+t) = x,$$

$$u^2(x,t) = w^1 - w^2 = w_0^1(x-t) - w_0^2(x+t)] = 1-t.$$

Se comprueba directamente que las condiciones de contorno para la componente u^1 en las fronteras laterales están satisfechas.

* Problemas periódicos

Es posible considerar también **problemas periódicos**. Por ejemplo, supongamos que se quiera resolver la ecuación de onda unidimensional en el intervalo $[0,1]$, donde la solución satisface:

$$u(0,t) = u(1,t), \qquad \forall\, t \geq 0.$$

Esta condición se llama condición de contorno periódica. Un problema periódico para una función $u(x,t)$ con $x \in [0,1]$ es equivalente a un problema en la recta real sustituyendo la condición anterior por:

$$u(x,t) = u(x+p,t), \qquad \forall\, t \geq 0, \quad \forall\, p \in \mathbb{Z}.$$

Nótese que, hablando con rigor, una condición de tipo periódico no es una condición de contorno puesto que para las soluciones de los problemas periódicos no existe frontera (o contorno). No se aplican por tanto los comentarios hechos sobre las condiciones de contorno bien (y mal) puestas.

1.2.7. * Sistemas de leyes de conservación: la notación de operadores

En esta sección presentamos la forma general de un sistema de leyes de conservación en varias variables espaciales y daremos un ejemplo importante que aparece en la física matemática al describir la dinámica de gases.

Sea Ω una región abierta de $\mathbb{R}^n$ y sean f_j, $1 \leq j \leq d$ funciones de clase C^1 tales que $f_j : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$, es decir: sean dadas d funciones vectoriales de n

variables suficientemente regulares para que esté bién definido y sea continuo el gradiente de cada una de ellas. Consideraremos el sistema de n ecuaciones de conservación dado por:

$$\frac{\partial \boldsymbol{u}}{\partial t} + \sum_{j=1}^{d} \frac{\partial f_j}{\partial x_j}(\boldsymbol{u}) = 0, \tag{1.47}$$

para $\boldsymbol{x} = (x_1, ..., x_d) \in I\!R^d$ y $t > 0$, y donde $\boldsymbol{u} = (u_1, ..., u_n)$ es un campo vectorial $\boldsymbol{u} : I\!R^d \times (0, +\infty) \to \Omega$. El sistema (1.47) se puede escribir en la forma de divergencia:

$$\boldsymbol{u}_t + \operatorname{div} \boldsymbol{f}(\boldsymbol{u}) = 0,$$

siendo $\boldsymbol{f}$ una función matricial con valores en M_{nd}. El conjunto Ω se llama el **conjunto de estados** (del sistema). Las funciones $f_j = (f_{1j}, ..., f_{nj})$ (es decir las componentes vectoriales de la función matricial) se llaman las **funciones de flujo**. Formalmente el sistema (1.47) expresa la conservación de las cantidades $u_1, ..., u_n$. En efecto, sea D un dominio arbitrario de $I\!R^d$ y sea $\boldsymbol{n} = (n_1, ..., n_d)$ el vector normal unitario exterior a la frontera ∂D de D. Entonces integrando la ecuación (1.47) en $D \subset I\!R^d$ y aplicando el teorema de la divergencia (en el caso d-dimensional) se tiene:

$$\frac{d}{dt} \int_D \boldsymbol{u} d\boldsymbol{x} + \sum_{j=1}^{d} \int_{\partial D} f_j(\boldsymbol{u}) n_j dS = 0.$$

Esta ecuación de equilibrio (balance) tiene un significado muy natural: la variación de $\displaystyle\int_D \boldsymbol{u} d\boldsymbol{x}$, representada por:

$$\frac{d}{dt} \int_D \boldsymbol{u} d\boldsymbol{x},$$

es igual a las pérdidas a través de la frontera:

$$\sum_{j=1}^{d} \int_{\partial D} f_j(\boldsymbol{u}) n_j dS.$$

Caracterizaremos ahora los sistemas estríctamente hiperbólicos[30]. Para ello se calcula la matriz jacobiana de cada función de flujo $f_j(\boldsymbol{u})$ dada por:

$$[A_j(\boldsymbol{u})] = \left(\frac{\partial f_{ij}}{\partial u_k}(\boldsymbol{u}) \right)_{1 \le i,k \le n}.$$

[30]Nótese que la mayoría de los sistemas de leyes de conservación que aparecen en las aplicaciones son hiperbólicos. Más exactamente, son simetrizables y esto implica que son hiperbólicos. La simetrizabilidad se debe a la existencia de una función de entropía. Más detalles se pueden encontrar en el libro de Godlewski y Raviart que aparece en la bibliografía avanzada.

El sistema (1.47) se dice **hiperbólico** si, $\forall\, \boldsymbol{u} \in \Omega$ y cada $\boldsymbol{w} = (w_1, ..., w_d)$ la matriz

$$[A(\boldsymbol{u}, \boldsymbol{w})] = \sum_{j=1}^{d} w_j A_j(\boldsymbol{u})$$

tiene n autovalores reales:

$$\lambda_1(\boldsymbol{u}, \boldsymbol{w}) \leq \lambda_2(\boldsymbol{u}, \boldsymbol{w}) \leq ..... \leq \lambda_n(\boldsymbol{u}, \boldsymbol{w}) \leq ...$$

y n correspondientes autovectores linealmente independientes

$$r_1(\boldsymbol{u}, \boldsymbol{w}),\ r_2(\boldsymbol{u}, \boldsymbol{w}),\ ....., r_n(\boldsymbol{u}, \boldsymbol{w}).$$

Si además los autovalores son todos distintos entonces el sistema (1.47) se denomina **estríctamente hiperbólico**. Definimos finalmente el problema de Cauchy para sistemas de leyes de conservación.

Definición 1.2.5 (Problema de Cauchy) *Determinar una función*

$$\boldsymbol{u} :\ I\!R^d \times [0, +\infty) \to \Omega, \qquad \boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}, t) \in \Omega,$$

que satisfaga la ecuación (1.47) y la condición inicial:

$$\boldsymbol{u}(\boldsymbol{x}, 0) = \boldsymbol{u}_0(\boldsymbol{x}), \quad \boldsymbol{x} \in I\!R^d,$$

siendo $\boldsymbol{u}_0 : I\!R^d \to \Omega$ *una función dada.*

Como aplicación de lo anterior consideraremos ahora las ecuaciones de las dinámicas de los gases en coordenadas eulerianas[31] $(\vec{x}, t) = (x_1, x_2, x_3, t)$. Lo que haremos será escribir estas ecuaciones en la forma de una única ecuación de conservación del tipo (1.47).

Ejemplo 1.2.13 *Un ejemplo muy importante en las aplicaciones es el sistema de ecuaciones que gobierna la dinámica de los gases. Despreciando la conducción de calor, las ecuaciones de Euler para un fluido compresible no viscoso (un gas) son*

[31] Utilizar coordenadas eulerianas consiste en dar una descripción espacial de la región considerada. Se centra así la atención en los puntos del espacio en lugar de seguir la evolución temporal de las patículas que componen la distribución de masas inicial del sistema (lo que equivaldría a una descripción *lagrangiana* (o material)).

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \displaystyle\sum_{i=1}^{3} \frac{\partial}{\partial x_j}(\rho v_j) = 0, \\ \dfrac{\partial}{\partial t}(\rho v_i) + \displaystyle\sum_{i=1}^{3} \frac{\partial}{\partial x_j}(\rho v_i v_j + p\delta_{ij}) = 0, \qquad 1 \leq i \leq 3, \\ \dfrac{\partial}{\partial t}(\rho e) + \displaystyle\sum_{i=1}^{3} \frac{\partial}{\partial x_j}((\rho e + p) v_j) = 0. \end{cases} \tag{1.48}$$

siendo ρ la densidad del fluido, $\boldsymbol{v} = (v_1, v_2, v_3)$ la velocidad, p la presión, ϵ la energía interna específica (por unidad de masa), $e = \epsilon + \|\boldsymbol{u}\|/2$ la energía total específica y δ_{ij} el símbolo de Kronecker: $\delta_{ij} = 1$ si $i = j$, $\delta_{ij} = 0$ si $i \neq j$.

Las ecuaciones (1.48) expresan las leyes de conservación de la masa, del momento y de la energía total del fluido. Se tienen que complementar con una ecuación de estado (que caracteriza el gas estudiado) que suele ser del tipo

$$p = p(\rho, \epsilon). \tag{1.49}$$

Las ecuaciones (1.48) y (1.49) constituyen un sistema de $5 + 1 = 6$ ecuaciones en las 6 variables $(\rho, v_1, v_2, v_3, p, \epsilon)$. En el caso de un gas ideal politrópico la ecuación de estado sería por ejemplo $p = (\gamma - 1)\rho\epsilon$, siendo $\gamma > 1$. Si definimos $m_i = \rho v_i$, $i = 1.,3$, $E = \rho e$ el sistema (1.48) se puede escribir en la forma de una única ecuación vectorial del tipo (1.47):

$$\frac{\partial \boldsymbol{\Phi}}{\partial t} + \sum_{j=1}^{d} \frac{\partial f_j}{\partial x_j}(\boldsymbol{\Phi}) = 0,$$

siendo la incógnita dada por el campo vectorial

$$\Phi = \begin{pmatrix} \rho \\ m_1 \\ m_2 \\ m_3 \\ E \end{pmatrix},$$

las funciones de flujo dadas por

$$f_1(\Phi) = \begin{pmatrix} m_1 \\ p + m_1^2/2 \\ m_1 m_2/\rho \\ m_1 m_3/\rho \\ (E+p)m_1/\rho \end{pmatrix}, \; f_2(\Phi) = \begin{pmatrix} m_2 \\ m_1 m_2/\rho \\ p + m_2^2/\rho \\ m_2 m_3/\rho \\ (E+p)m_2/\rho \end{pmatrix}, \; f_3(\Phi) = \begin{pmatrix} m_3 \\ m_1 m_3/\rho \\ m_2 m_3/\rho \\ p + m_3^2/\rho \\ (E+p)m_3/\rho \end{pmatrix},$$

la ecuación de estado

$$p = p(\rho, (E - |m|^2/2\rho)/\rho) = (\gamma - 1)(E - |m|^2/2\rho)$$

y el conjunto de estados

$$\Omega = \{(\rho, \boldsymbol{m} = (m_1, m_2, m_3), E) \,/\, \rho > 0, \boldsymbol{m} \in I\!R^3, E - |m|^2/2\rho > 0\},$$

donde aparecen las condiciones de admisibilidad física de las soluciones del problema.

1.3. Ecuaciones lineales de segundo orden

La forma general de una ecuación diferencial en derivadas parciales **lineal** de segundo orden con dos variables independientes x, y es

$$A(x,y)\frac{\partial^2 u}{\partial x^2} + B(x,y)\frac{\partial^2 u}{\partial x \partial y} + C(x,y)\frac{\partial^2 u}{\partial y^2} + a(x,y)\frac{\partial u}{\partial x} + b(x,y)\frac{\partial u}{\partial y} + c(x,y)u = f(x,y), \tag{1.50}$$

siendo A, B, C, a, b, c, f funciones reales. Si las funciones A, B, C son idénticamente nulas, es decir $A(x,y) = B(x,y) = C(x,y) \equiv 0$ en un dominio $\Omega \subset I\!R^2$ entonces la ecuación es de primer orden y se estudia con las técnicas expuestas en las dos primeras secciones de este tema. Si $f(x,y) \equiv 0$ en Ω entonces la ecuación se dice que es **homogénea**, y en caso contrario, **no homogénea**. Al designar el primer miembro de (1.50) por $L[u]$ (notación de operadores) se puede escribir de forma más compacta la ecuación (1.50) como:

$$L[u] = f(x,y),$$

siendo la ecuación homogénea correspondiente $L[u] = 0$. Aquí L es el operador diferencial lineal definido en el espacio $C^2(D)$ mediante la función $u(x,y)$. Con más precisión, se define el operador $L[\cdot] : C^2(D) \to C(D)$ por:

$$L[u] = \left(A(x,y)\frac{\partial^2}{\partial x^2} + B(x,y)\frac{\partial^2}{\partial x \partial y} + C(x,y)\frac{\partial^2}{\partial y^2} + a(x,y)\frac{\partial}{\partial x} + \right.$$

$$\left. b(x,y)\frac{\partial}{\partial y} + c(x,y) \right)[u].$$

El operador L es, evidentemente, lineal pues verifica

$$L[\alpha u + \beta v] = \alpha L[u] + \beta L[v], \qquad \forall\, u, v \in C^2(\Omega), \quad \forall\, \alpha, \beta \in I\!R.$$

Observación 1.3.1 *Nótese que no todos los operadores son lineales. Por ejemplo, el operador L tal que:*

$$L[u] \doteq \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2,$$

no lo es. En realidad, muchos problemas físicos implican operadores no lineales. Las hipótesis de simplificación que se suelen hacer al considerar un modelo matemático de la realidad física están dirigidas a sustituir un operador no lineal por uno lineal. Esto es típico, en el sentido de que es posible aproximar las soluciones de muchos problemas sustituyendo operadores no lineales por otros lineales. En el tratamiento analítico exacto desarrollado en este curso nos referiremos tan sólo a tales casos. El mundo *no lineal será considerado en el capítulo dedicado a la aproximación numérica.*

Ya sabemos (véase el tema de EDO del primer curso) que si una EDO es lineal y homogénea entonces a partir de soluciones conocidas es posible generar otras soluciones por superposición. Para una EDP lineal homogénea la situación es parecida. En efecto, utilizando la linealidad del operador L, se tienen las siguientes propiedades de las soluciones de ecuaciones diferenciales en derivadas parciales:

PROPIEDADES

1. **Caso homogéneo**.

   Teorema 1.3.1 *Si $u(x,y)$ es una solución de la ecuación en derivadas parciales lineal homogénea*

   $$L[u] = 0,$$

   entonces $cu(x,y)$ (donde c es una constante cualquiera) es también solución de $L[u] = 0$.

Teorema 1.3.2 *Si $u_1(x,y)$ y $u_2(x,y)$ son soluciones de la ecuación en derivadas parciales lineal homogénea*

$$L[u] = 0,$$

entonces la suma $u_1(x,y) + u_2(x,y)$ es también solución de $L[u] = 0$.

Los teoremas anteriores se resumen diciendo que el conjunto de soluciones de una EDP lineal homogénea es un espacio vectorial. De otra forma, si cada una de las funciones $u_1(x,y)$, $u_2(x,y)$, ... $u_k(x,y)$ es solución de la ecuación lineal homogénea $L[u] = 0$ entonces la combinación lineal

$$c_1u_1(x,y) + c_2u_2(x,y) + ... + c_ku_k(x,y),$$

donde c_i, $i = 1, ...k$ son constantes reales arbitrarias, también es solución de esta ecuación. Nótese que esta propiedad tiene lugar también para las EDO lineales homogéneas (véase la definición de sistema fundamental de soluciones para una EDO lineal homogénea de orden n de la sección 3 del capítulo 13 del guión del primer curso). Sin embargo una EDO lineal homogénea de orden n tiene *exactamente* n familias de soluciones (un número finito por tanto) linealmente independientes (cuya combinación lineal genera la solución general de la EDO). Una EDP lineal homogénea puede tener un conjunto infinito de soluciones linealmente independientes. Consecuentemente para EDP lineales homogéneas tendremos que operar no sólo con combinaciones lineales de un número finito de soluciones, sino también con las series infinitas[32]

$$\sum_{n=1}^{\infty} c_n u_n(x,y).$$

Es decir, el conjunto de soluciones de una EDP del tipo (1.50) es un espacio lineal (vectorial) de dimensión infinita contrariamente a lo que ocurría con las EDO de orden n cuyo espacio de soluciones tiene dimensión n.

2. **Caso no homogéneo**.

Teorema 1.3.3 *Si $u(x,y)$ es solución de la EDP lineal no homogénea*

$$L[u] = f$$

y $v(x,y)$ es solución de la EDP homogénea correspondiente

$$L[u] = 0,$$

entonces la suma $u + v$ es solución de la EDP no homogénea $L[u] = f$.

[32]Nos encontraremos con este tipo de soluciones, en forma de series infinitas, al resolver las EDP con el método de separación de variables (tema 2 de este guión).

Se tiene además el siguiente resultado conocido con el nombre de **Principio de superposición.**

Teorema 1.3.4 *Si $u_1(x,y)$ es una solución de la ecuación*

$$L[u] = f_1$$

y $u_2(x,y)$ es una solución de

$$L[u] = f_2,$$

entonces $u_1 + u_2$ es solución de la ecuación $L[u] = f_1 + f_2$.

En el próximo tema veremos que el principio de superposición tiene una gran aplicación en el proceso de resolución de los problemas de contorno y/o iniciales para EDP lineales de segundo orden.

1.3.1. Ecuaciones de segundo orden: curvas características y clasificación

Sea dada la ecuación diferencial lineal general de segundo orden (1.50):

$$A(x,y)\frac{\partial^2 u}{\partial x^2}+B(x,y)\frac{\partial^2 u}{\partial x \partial y}+C(x,y)\frac{\partial^2 u}{\partial y^2}+a(x,y)\frac{\partial u}{\partial x}+b(x,y)\frac{\partial u}{\partial y}+c(x,y)u = f(x,y)$$

en cierta región $\Omega \subset I\!R^2$. Esta ecuación, en un punto (x,y), se denomina hiperbólica, parabólica o elíptica según el siguiente criterio:

CRITERIO DE CLASIFICACIÓN

1. La ecuación (1.50) es **hiperbólica** en Ω si:
$$\Delta = B^2 - 4AC > 0 \quad \text{en } \Omega.$$
2. La ecuación (1.50) es **parabólica** en Ω si:
$$\Delta = B^2 - 4AC \equiv 0 \quad \text{en } \Omega.$$
3. La ecuación (1.50) es **elíptica** en Ω si:
$$\Delta = B^2 - 4AC < 0 \quad \text{en } \Omega.$$

Nótese que las funciones coeficientes a, b y c de los términos de primer orden no aparecen en absoluto en el criterio de clasificación. Esto se interpreta diciendo que los términos de primer orden no pueden modificar la *naturaleza* de una ecuación lineal de segundo orden. Es decir, los términos de primer orden influyen cuantitativamente pero no cualitativamente en las soluciones de la ecuación.

A partir del tipo de ecuación considerado se pueden deducir importantes propiedades que sugieren no sólo qué métodos de resolución son adecuados sino también criterios para determinar si un problemas está bien planteado o no.

Los conceptos de hiperbolicidad, parabolicidad y elipticidad son locales, es decir, se tienen en un punto concreto (x_0, y_0). El conjunto de los puntos donde se tienen estas propiedades son las regiones donde la ecuación (o el operador diferencial que la define) es hiperbólica, parabólica y elíptica.

Ejemplo 1.3.1 *Clasificar la ecuación de Tricomi:*
$$u_{xx} + xu_{yy} = 0.$$
Se trata evidentemente de una EDP de segundo orden lineal y homogénea. Identificando coeficientes se tiene
$$A(x,y) \equiv 1, \qquad B(x,y) \equiv 0, \qquad C(x,y) = x.$$
Por tanto, $B^2 - 4AC = -4x$, luego la ecuación es hiperbólica si $x < 0$ y es elíptica si $x > 0$. Si $x = 0$ (en el eje y) la ecuación es degenerada, pues cambia de tipo pasando de EDP a EDO (paramétrica).

Observación 1.3.2 *Los términos hiperbólico, parabólico y elíptico derivan de la clasificación típica (de la geometría analítica) de las curvas cuadráticas (cónicas).*

Formas canónicas

Haciendo unos cambios adecuados en las variables independientes,

$$\xi = f(x,y), \quad \eta = g(x,y), \qquad f, g \in C^2,$$

y aplicando la regla de la cadena obtenemos las siguientes formas canónicas:

1. La ecuación (1.50) es **hiperbólica** en Ω si se puede reconducir a una de las dos siguientes expresiones:

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = F\left(\xi, \eta, u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}\right), \qquad \frac{\partial^2 u}{\partial \xi^2} - \frac{\partial^2 u}{\partial \eta^2} = \Phi\left(\xi, \eta, u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}\right)$$

(son dos formas canónicas de las ecuaciones de tipo hiperbólico de segundo orden). Las curvas (coordenadas) (ξ, η) se llaman **características**. En el caso hiperbólico lineal uni-dimensional con coeficientes constantes hay exactamente dos curvas (rectas) características dadas por:

$$\xi = y - \alpha_- x, \qquad \eta = y - \alpha_+ x$$

donde las pendientes $\alpha_\pm$ vienen dadas por:

$$\alpha_\pm = \frac{1}{2A}\left[-B \pm \sqrt{B^2 - 4AC}\right]$$

Las formas canónicas son:

$$u_{\xi\eta} + a u_\xi + b u_\eta + cu + d = 0,$$

y (definiendo $\sigma = (\xi + \eta)/2$, $\tau = (\xi - \eta)/2$)

$$u_{\sigma\sigma} - u_{\tau\tau} + (a+b)u_\sigma + (a-b)u_\tau + cu + d = 0$$

El prototipo de EDP de segundo orden hiperbólica es la ecuación de ondas $u_{tt} = c^2 \Delta u$. En el caso unidimensional se expresa en la forma $u_{tt} = c^2 u_{xx}$. La ecuación de ondas se puede resolver fácilmente introduciendo las nuevas coordenadas (las características)

$$\xi = x + ct, \qquad \eta = x - ct,$$

y aplicando la regla de la cadena a través de la cual se deduce la simple ecuación:

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0.$$

Esta ecuación se satisface si y sólo si

$$u(\xi, \eta) = p(\xi) + q(\eta),$$

es decir,

$$u(x,t) = p(x+ct) + q(x-ct),$$

donde p y q son funciones derivables cualesquiera de una variable. En consecuencia si u es solución de la ecuación de ondas tiene que ser del tipo anterior. La fórmula anterior es la integral general de la ecuación de onda unidimensional. D' Alembert la encontró en 1747. Esta ecuación nos dice que toda solución es la suma de una onda que se desplaza hacia la izquierda con velocidad $-c$ y de otra que lo hace hacia la derecha con velocidad $+c$. Puesto que las ondas se desplazan en direcciones opuestas, la forma de $u(x,t)$ en general cambiará con el tiempo. Esta interpretación originó el nombre de ecuación de ondas. Resolver un problema de valores iniciales y de contorno concreto corresponderá a determinar las funciones p y q adecuadas. Una referencia clásica muy clara es el libro de Weimberger, capítulo 2, dedicado a la ecuación de ondas unidimensional. Como en el caso de las EDP de primer orden se admiten soluciones discontinuas (es decir débiles o generalizadas) y las discontinuidades se propagan a lo largo de las características.

2. La ecuación (1.50) es **parabólica** en Ω si:

$$\frac{\partial^2 u}{\partial \eta^2} = \Phi\left(\xi, \eta, u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}\right)$$

En este caso existe sólo una familia de características reales con pendiente

$$\alpha = \alpha_1 = \alpha_2 = \frac{B}{2A}.$$

En el caso unidimensional lineal de coeficientes constantes la forma canónica es:

$$u_{\eta\eta} + au_\xi + bu_\eta + cu + d = 0,$$

El prototipo de EDP de segundo orden parabólica es la ecuación del calor $u_t = k\Delta u$. En el caso unidimensional se expresa en la forma $u_t = ku_{xx}$. Nos ocuparemos de ella más tarde.

3. La ecuación (1.50) es **elíptica** en Ω si:

$$\frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2} = \Phi\left(\xi, \eta, u, \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \eta}\right).$$

En este caso $B^2 - 4AC < 0$ luego las raíces α_1, α_2 son complejas y no existen características reales. En el caso bi-dimensional con coeficientes

constantes si definimos $\sigma = (\xi + \eta)/2$, $\tau = (\xi - \eta)/2i$ se tiene la forma canónica para ecuaciones elípticas:

$$u_{\sigma\sigma} + u_{\tau\tau} + (a+b)u_\sigma + (a-b)u_\tau + cu + d = 0.$$

El prototipo de EDP de segundo orden elíptica es la ecuación de Laplace $\Delta u = 0$. En el caso bidimensional se expresa en la forma $u_{xx} + u_{yy} = 0$. Este tipo de ecuación será abordada más adelante.

Los razonamientos anteriores puden ser generalizados al caso de coeficientes variables. A diferencia del caso con coeficientes constantes, pueden existir características reales sólo en una parte del dominio de interés. Las ecuaciones pueden cambiar de tipo de una región a otra. Esto ocurre, por ejemplo, en dinámica de gases (*transonic gas dynamics*). En el libro de Mei, pag 28, puden encontrarse más detalles.

Resumimos lo anterior considerando el operador:

$$L[u] \doteq A\frac{\partial^2 u}{\partial t^2} + B\frac{\partial^2 u}{\partial x \partial t} + C\frac{\partial^2 u}{\partial x^2},$$

siendo A, B, C constantes dadas. Nos concentramos por tanto en la parte de segundo orden del operador lineal general de segundo orden (pues es la parte que gobierna el proceso de clasificación). Utilizaremos además variables (x, t) típicas de los problemas de evolución (parabólicos e hiperbólicos). Podemos entonces transformar $L[u]$ en un múltiplo de

$$\frac{\partial^2 u}{\partial \xi \partial \eta}$$

si y sólo si

$$B^2 - 4AC > 0.$$

Siempre que esto se cumple, L es **hiperbólico**. La transformación en este caso viene dada por:

$$\xi = 2Ax + [-B + \sqrt{B^2 - 4AC}]t, \qquad \eta = 2Ax + [-B - \sqrt{B^2 - 4AC}]t$$

y el operador se transforma en

$$L[u] = -4A(B^2 - 4AC)\frac{\partial^2 u}{\partial \xi \partial \eta}$$

El caso $A = 0$ (y $C \neq 0$) puede tratarse del mismo modo, con $\xi = t$, $\eta = x - (C/B)t$. El caso $A = C = 0$ es trivial puesto que el operador ya es de la forma buscada.

Si $B^2 - 4AC = 0$ se dice que L es **parabólico**. En este caso la transformación:

$$\xi = 2Ax - Bt, \qquad \eta = t,$$

transforma $L[u]$ en

$$L[u] = A\frac{\partial^2 u}{\partial \eta^2},$$

que es la forma corriente para un operador parabólico. La solución general de $L[u] = 0$ es ahora:

$$p(\xi) + \eta q(\xi).$$

Esta solución puede interpretarse como una onda de forma fija que se mueve con velocidad $B/2A$ junto con otra que crece linealmente con el tiempo y se mueve con la misma velocidad. Un operador parabólico tiene sólo una familia de características $\xi = C$ (constantes). Las discontinuidades de las derivadas se propagan a lo largo de estas características.

Finalmente, si $B^2 - 4AC < 0$, el operador L es **elíptico**. El prototipo de operador elíptico es:

$$\frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2}.$$

No tiene características. Sin embargo, la transformación

$$\xi = \frac{2Ax - Bt}{\sqrt{4AC - B^2}}, \qquad \eta = t,$$

hace

$$L[u] = A\left[\frac{\partial^2 u}{\partial \xi^2} + \frac{\partial^2 u}{\partial \eta^2}\right].$$

En la forma $L[u] = 0$, se tiene la ecuación de Laplace. Las derivadas parciales de una solución de la ecuación de Laplace no presentan discontinuidades.

Un gran número de problemas físicos puede reducirse a una o varias de las ecuaciones anteriores.

Observación 1.3.3 *Cuando el número n de variables independientes es superior a dos, también se diferencian las ecuaciones de tipo hiperbólico, parabólico y elíptico. Finalmente observamos que no existe relación entre la clasificación*

de las EDP de segundo orden y las de primer orden ya que éstas últimas son todas hiperbólicas y el criterio las clasificaría (poniendo $A = B = C \equiv 0$), erróneamente, como parabólicas. Sin embargo sí existe relación entre las ecuaciones lineales hiperbólicas de primer y segundo orden. Esta relación atañe las propiedades cualitativas de las soluciones. Por ejemplo, la propiedad de propagación con velocidad finita de las perturbaciones (el dato inicial). Este fenómeno (véase la observación referente la existencia de soluciones de soporte compacto para EDP de primer orden) no puede ocurrir en las ecuaciones lineales parabólicas en las cuales la velocidad de propagación es infinita, entendiendo por ello que si el dato inicial tiene soporte compacto la solución se propaga instantáneamente a todo el dominio (es decir se difunde en todo el espacio).

1.3.2. Problemas de contorno

Para describir completamente uno u otro proceso físico no basta con tener sólo la ecuación diferencial que rige el proceso, hace falta plantear el estado inicial de este proceso (condiciones iniciales para problemas de evolución) y el régimen en la frontera $\partial\Omega$ de la región Ω donde tiene lugar el proceso. Se distinguen dos tipos principales de problemas de contorno para EDP:

1. Problemas de contorno para ecuaciones de tipo elíptico. Se trata de problemas estacionarios donde se plantean las condiciones en la frontera $\partial\Omega$ y no hay condiciones iniciales.

2. Problemas de contorno y de valores iniciales para ecuaciones en derivadas parciales en dominios acotados. Se trata de problemas de evolución para ecuaciones de tipo hiperbólico o parabólico donde se plantean condiciones iniciales y de contorno en dominios acotados $\Omega \subset \mathbb{R}^n$, $|\Omega| \doteq \operatorname{diam}\Omega < \infty$.

Empezando por el caso más sencillo consideramos la ecuación parabólica de conductibilidad térmica unidimensional, conocida con el nombre de ecuación del calor:

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1, \quad t > 0, \tag{1.51}$$

siendo $k > 0$ la conductividad térmica del medio conductor (supuesto homogéneo) considerado. Complementamos la ecuación (1.51) con una condición inicial (nótese que el orden de derivación temporal de la EDP es uno)

$$u(x, 0) = f(x), \qquad 0 < x < 1, \tag{1.52}$$

y con condiciones de contorno en las *fronteras laterales* del dominio. Por ejemplo (detallaremos en la siguiente sección los distintos tipos de condiciones de contorno que se suelen utilizar) podríamos escribir

$$u(0,t)=\phi_0(t), \quad t>0, \qquad u(1,t)=\phi_1(t), \quad t>0, \tag{1.53}$$

en $x=0$, $\forall t>0$ y $x=1$, $\forall t>0$, siendo $\phi_0(t)$, $\phi_1(t)$ funciones dadas (conocidas).

Nótese que por fronteras laterales se entienden los conjuntos de puntos Γ_0, $\Gamma_1 \subset I\!R^2$ definidos por:

$$\Gamma_0=\{(x,t)\in I\!R^2/\,(x,t)=(0,t),\, t>0\}=\{0\}\times(0,\infty)$$

y

$$\Gamma_1=\{(x,t)\in I\!R^2/\,(x,t)=(1,t),\, t>0\}=\{1\}\times(0,\infty).$$

Más en general, si $\Omega \subset I\!R^N$ es un dominio N-dimensional de frontera $\partial\Omega$ entonces la ecuación se tiene en el cilindro (temporal) $Q=\Omega\times(0,T)$ de frontera lateral $\Sigma=\partial\Omega\times(0,T)$ donde $T\in I\!R^+$ (soluciones locales en tiempo) o $T=+\infty$ (soluciones globales). En este caso el problema de valor inicial y de contorno asociado a la ecuación del calor con los valores de la solución prefijados en el contorno (frontera lateral) se formula: *Hallar una función*

$$u(x,t):\bar{\Omega}\times[0,+\infty)\to I\!R$$

tal que,

$$\left\{\begin{array}{rcll} \dfrac{\partial u}{\partial t} & = & \Delta u, & \text{en } Q,\\ u(x,t) & = & g(x,t), & \text{en } \Sigma,\\ u(x,0) & = & u_0(x), & \text{en } \Omega. \end{array}\right.$$

donde $\Delta=\displaystyle\sum_{i=1}^{N}\frac{\partial^2}{\partial {x_i}^2}$, designa el operador laplaciano respecto de las variables espaciales, t es la variable tiempo y $u_0(x)$ es una función dada.

A veces es útil considerar (pues ahí se suele[33] localizar el máximo de la solución) la frontera parabólica ∂Q del cilindro $Q=\Omega\times(0,T)$ que se define por:

$$\partial Q=(\bar{\Omega}\times\{0\})\bigcup(\partial\Omega\times[0,T]).$$

[33]La naturaleza de las hipótesis que garantizan la veracidad de esta afirmación va mucho más allá de los objetivos de este curso. Una exposición clara pero de nivel avanzado se puede encontrar en el capítulo 10 sobre problemas de evolución del libro de H. Brezis, (1983). Análisis Funcional. Teoría y aplicaciones. Alianza Editorial.

Cuando Ω es infinito se suelen distinguir los casos $\Omega = I\!R^N$, $N = 1, 2, 3$ y, en el caso unidimensional, $\Omega = (-\infty, 0)$, $\Omega = (0, \infty)$ que corresponden al caso de dominio semi-infinito.

1.3.3. Condiciones iniciales y de contorno

Para presentar los distintos tipos de condiciones iniciales y de contorno que se suelen imponer en las aplicaciones consideraremos la ecuación del calor unidimensional. Existen tres tipos principales de condiciones iniciales y de contorno:

1. Especificar la función en la frontera del dominio. Se conoce como condición del tipo Dirichlet. Por ejemplo, si nuestra incógnita es la función $u(x,t)$ y queremos resolver la EDP del calor

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2} = \nabla \cdot (k\nabla u),$$

en un dominio acotado (un intervalo abierto de la recta real) $\Omega = (a, b)$ se pueden prefijar las siguientes condiciones límite:

$$u(x,0) = f(x)\,(t=0),\ u(a,t) = g_1(t)\,(x=a),\ u(b,t) = g_2(t)\,(x=b)$$

Este tipo de condiciones, junto con la EDP, genera lo que se llama un **problema de Dirichlet** o de primer tipo. Nótese que se ha impuesto *una* condición inicial de tipo Dirichlet y *dos* condiciones de contorno (siempre de tipo Dirichlet) en las fronteras laterales Γ_a y Γ_b.

2. Especificar la *derivada* de la función en la frontera del dominio. Se conoce como condición del tipo Neumann y el problema asociado se llama **problema de Neumann** (o de segundo tipo). Por ejemplo,

$$\frac{\partial u}{\partial x}(a,t) = 0 \quad (x=a), \qquad -k\frac{\partial u}{\partial x}(b,t) = q \quad (x=b),$$

siendo q una constante. Para dominios $\Omega \subset I\!R^n$, $n > 1$, se prescribe la componente normal de vector gradiente en la forma:

$$\nabla u \cdot \boldsymbol{n} = \frac{\partial u}{\partial \boldsymbol{n}},$$

siendo $\boldsymbol{n}$ el vector normal a la curva (o hipersuperficie) que representa la frontera del dominio. Físicamente este tipo de condición define el flujo de

calor que atraviesa la frontera, que por ley de Fourier de la conducción del calor, es proporcional al gradiente de temperaturas. Por ejemplo, en un problema estacionario de conducción de calor en una geometría cilíndrica (coordenadas (r, z)) podríamos escribir

$$\boldsymbol{q} = -k\nabla u,$$

siendo las componentes del flujo

$$q_r = -k\frac{\partial u}{\partial r}, \qquad q_z = -k\frac{\partial u}{\partial z}.$$

La componente $q_r(r, z)$ del vector de flujo $\boldsymbol{q} = (q_r, q_z)$ denota flujo de calor molecular en la dirección radial mientras $q_z(r, z)$ representa el flujo de calor molecular en la dirección axial.

Se trata en este caso de un flujo puramente difusivo. Si consideramos también el flujo convectivo (que es un fenómeno de transporte del calor que tiene lugar debido al transporte de materia) se tienen condiciones del tipo:

$$q_r = -k\frac{\partial u}{\partial r} + V_r u, \qquad q_z = -k\frac{\partial u}{\partial z} + V_z u,$$

siendo (V_r, V_z) las componentes radial y normal (respectivamente) de un campo de velocidades conocido.

3. Especificar la función y su derivada en la frontera del dominio.

Se conoce como condición de tipo **Robin** (o mixto) y el problema asociado es un **problema mixto** (o de tercer tipo). Si consideramos por ejemplo el flujo de calor a través de las paredes de un conducto por el cual fluye un fluido

$$-k\frac{\partial u}{\partial r} = h(u - u_\infty) \quad r = R, \quad \forall\, z,$$

siendo h un coeficiente de transporte de calor característico del material sólido del conducto que controla el flujo (de calor) en la frontera y u_∞ un valor conocido (dato del problema) de la temperatura ambiente alrededor del sistema. Hay que controlar cuidadosamente los signos que aparecen en este tipo de condición. En procesos de enfriamiento la cantidad $u - u_\infty$ es positiva (pues obviamente la temperatura externa que controla el sistema es menor que la temperatura del fluido que se quiere enfriar) luego

$$\frac{\partial u}{\partial r} < 0,$$

pues se está perdiendo calor por las fronteras laterales. En efecto el flujo de calor q_r es positivo,

$$q_r = -\frac{\partial u}{\partial r} > 0,$$

y esto quiere decir que el calor se está difundiendo de donde hay más (en el interior del conducto) hacia donde hay menos (en el exterior). Comentarios parecidos se pueden hacer en los procesos de transporte (difusión molecular) de materia donde se aplica la ley de Fick. La componente radial del flujo en coordenadas cilíndricas sería por ejemplo:

$$J_r = -D\frac{\partial c}{\partial r},$$

siendo $c(r, z)$ la concentración de una especie química y D su coeficiente de difusión.

Este tipo de condiciones es muy realista pues expresa que el flujo de calor conductivo es proporcional a la diferencia de temperaturas entre las paredes del conducto y el exterior.

Las condiciones anteriores se dicen **homogéneas** si se satisfacen (sin cambiar de expresión) para un múltiplo cualquiera de la variable incógnita.

El primer tipo de condiciones (condiciones de tipo Dirichlet) incluye las condiciones iniciales, que para una función $u(x, t)$ se pueden escribir en la forma:

$$u(x, 0) = f(x).$$

Esta condición significa que en el instante $t = 0$ la distribución de temperatura viene dada por la función $f(x)$. En alguna frontera podría haber variaciones temporales así que podríamos tener, por ejemplo en $x = 0$:

$$u(0, t) = g(t).$$

Ninguna de las dos condiciones del tipo 1 planteadas es homogénea (a menos que las funciones f y g sean idénticamente nulas). Sin embargo, si el valor en el contorno es una constante fijada, del tipo:

$$u(0, t) = u_0,$$

entonces podemos definir otra variable dependiente $\theta = u - u_0$ y obtener una condición

$$\theta(0, t) = 0,$$

que es homogénea en la frontera.

En ocasiones las condiciones del tipo 2, en sistemas de coordenadas cilíndricas y esféricas, es posible utilizarlas como condición de simetría:

$$\frac{\partial u}{\partial r} = 0, \quad r = 0.$$

Nótese que para tales sistemas de coordenadas $r \geq 0$, luego para asegurar perfiles simétricos de u tenemos que imponer la condición anterior. En el caso de una pared de un conducto aislada térmicamente (si se considera un problema de transporte de calor) o impermeable (si se considera un problema de flujo de materia en un medio poroso saturado), la condición es

$$-k\frac{\partial u}{\partial r} = 0, \quad r = R.$$

En el caso de las paredes de un conducto calentadas eléctricamente, la entrada de calor puede ser uniforme y constante (controlando por ejemplo la intensidad de corriente) luego podríamos imponer en un contorno del conducto:

$$-k\frac{\partial u}{\partial r} = q, \quad r = R.$$

El tercer tipo de condiciones es mixto e incluye la función y su derivada (o su integral). Por ejemplo, la condición:

$$-k\frac{\partial u}{\partial r} = h(u - u_\infty) \quad r = R, \quad \forall\, z,$$

que simplemente modeliza el equilibrio entre el flujo conductivo y la transferencia de calor en la pared de un conducto. Puede ser reconducida a una condición homogénea definiendo $\theta = u - u_\infty$ (si u_∞ es constante):

$$-k\frac{\partial \theta}{\partial r} = h\theta \quad r = R, \quad \forall\, z.$$

En los límites $h \to \infty$ y $h \to 0$ se recuperan las condiciones homogéneas del tipo 1 y 2. Este tipo de condiciones de contorno se conoce también con el nombre de **condición de tipo convectivo**[34]. Ocasionalmente, una condición de tipo mixto puede nacer como un equilibrio integro-diferencial. Un ejemplo se encuentra en Rice y Do, pag 408. Tal tipo de condiciones se suele abarcar

[34] Véase por ejemplo pag 42 del libro de W.M. Deen, (1998), Analysis of Transport Phenomena. Oxford University Press.

con la técnica de la transformada de Laplace que veremos en el tercer tema de esta asignatura.

Otros tipos de condiciones de contorno son igualmente posibles. En los procesos de transporte de calor por radiación (por ejemplo en combustión o en procesos que se desarrollan a altas temperaturas) se aplica la ley de Stefan-Boltzmann para describir el mecanismo de transporte de calor entre superficies sólidas separadas por gases (que se asumen transparentes a la radiación). Otras condiciones se imponen en procesos de fusión o evaporación donde tienen lugar cambios de fases[35].

El número de condiciones de contorno o iniciales necesario para resolver una EDO corresponde al número de constantes arbitrarias generadas en el curso del análisis. Una ecuación de orden n genera n constantes arbitrarias. Esto implica que el número total de condiciones (de contorno o iniciales) a imponer es n. En las ecuaciones en derivadas parciales, existen siempre al menos dos variables independientes así, por ejemplo, una ecuación que describe el régimen transitorio de distribución de temperatura en una barra cilíndrica de metal:

$$\rho c_p \frac{\partial T}{\partial t} = k \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial T}{\partial r} \right) = k \left(\frac{\partial^2 T}{\partial r^2} + \frac{1}{r} \frac{\partial T}{\partial r} \right),$$

necesitará (normalmente) una condición para el tiempo (una condición inicial) y dos para las fronteras fijas espaciales (digamos $r = 0$ y $r = R$). En principio, por tanto, se necesita fijar generalmente una condición para cada orden de cada derivada parcial. Sin embargo, esto no siempre es el caso. Por ejemplo, una condición inicial no es necesaria si buscamos una solución periódica en el tiempo. Un caso particular sería:

$$T(r, t) = f(r) e^{iwt}.$$

En tales casos, sólo es necesario imponer condiciones en las fronteras ($r = 0$ y $r = R$). Comentarios parecidos se aplican cuando se habla de la posición angular en coordenadas cilíndricas o esféricas. Concretamente cuando la solución debe ser periódica en el ángulo:

$$T(r, \theta) = T(r, \theta + 2\pi).$$

[35] Muy interesante en este sentido es el capítulo 2 sección 2.5, pag 41 del libro de Deen dedicado a la transferencia de calor en las interfases.

1.4. Ecuaciones elípticas.

El estudio de los procesos estacionarios (que no varían con el tiempo) de diferente naturaleza física conduce a la consideración de EDP de tipo elíptico.

1.4.1. Ecuaciones de Laplace y de Poisson

El operador lineal:

$$\Delta u = \nabla^2 u = (\nabla \cdot \nabla)u = \nabla \cdot (\nabla u) = \mathrm{div}(\nabla u) = \sum_{i=1}^{N} \frac{\partial^2 u}{\partial {x_i}^2},$$

se llama **laplaciano N-dimensional** de u. Al operador

$$\Delta = \sum_{i=1}^{N} \frac{\partial^2}{\partial {x_i}^2} = \frac{\partial^2}{\partial {x_1}^2} + \frac{\partial^2}{\partial {x_2}^2} + ... + \frac{\partial^2}{\partial {x_N}^2},$$

se le llama **operador laplaciano** en honor del matemático francés Pierre Laplace (1749-1827). A la ecuación

$$\Delta u = 0, \qquad \Omega \subset I\!R^N,$$

se la conoce como ecuación de Laplace y es la más simple de las ecuaciones de tipo elíptico. La ecuación de Laplace siempre es homogénea. Si se considera la ecuación no homogénea

$$\Delta u = f(x_1, ..x_N), \qquad \Omega \subset I\!R^N,$$

entonces tenemos la ecuación de Poisson (EDP elíptica lineal no homogénea que lleva este nombre en honor del matemático S. D. Poisson (1781-1840)) que corresponde a un estado de equilibrio originado por una fuerza exterior. Si f dependiera también de u en la forma:

$$\Delta u + \lambda u = 0,$$

entonces tendríamos la ecuación (lineal) de Helmholtz que es un caso particular de las ecuaciones elípticas semilineales,

$$-\Delta u + f(x, u) = 0, \qquad \Omega \subset I\!R^N.$$

Otra clase de EDP elípticas muy importante en las aplicaciones son las cuasilineales (ecuaciones no lineales donde las no linealidades aparecen en las derivadas de orden inmediatamente más bajo del orden de la ecuación). Su expresión general es:

$$-\Delta u + f(x, u, \nabla u) = 0, \qquad \Omega \subset I\!R^N.$$

Por último, definimos el operador bi-laplaciano:

$$\nabla^4 u = \nabla^2(\nabla^2 u) = \Delta(\Delta u) = \Delta^2 u,$$

que es un operador elíptico del cuarto orden. Se trata en este caso de resolver la ecuación elíptica bi-armónica de cuarto orden

$$\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = 0.$$

Se puede demostrar (véase el libro de Deen, capítulo 5, pag 239) que la función de corriente de un fluido incompresible estacionario bidimensional satisface la ecuación bi-armónica lo que expresa la conservación de la cantidad de movimiento para flujos (lentos) de Stokes de fluidos newtonianos. Los flujos de Stokes (*creeping flows* en la literatura anglosajona) tienen muchas aplicaciones tecnológicas (microcirculación, reología de suspensiones, dispersiones coloidales o el procesado de polímeros) y aparecen relacionados con varios fenómenos naturales asociados a fluidos muy viscosos.

Dependiendo de los valores de n (es decir del número de variables independientes consideradas) se tienen las siguientes expresiones del operador laplaciano en coordenadas cartesianas:

$$N = 1 \quad \Delta u \equiv \frac{d^2u}{dx^2} = 0, \qquad N = 2 \quad \Delta u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0,$$

$$N = 3 \quad \Delta u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$$

donde el caso $N = 1$ corresponde a una EDO y los casos $n = 2, 3$ corresponden a una EDP.

Definición 1.4.1 *Una función $u \in C^2(\Omega)$ tal que $\Delta u = 0$ en una región $\Omega \subset I\!R^N$ se dice* **armónica** *en Ω.*

Ejemplo 1.4.1 *Demostrar que la función:*

$$f(x, y) = e^x \cos y,$$

es armónica en el plano $I\!R^2$.

La función dada es evidentemente de clase $C^2(I\!R^2)$ (de hecho es de clase C^∞). Es inmediato calcular

$$f_x(x, y) = e^x \cos y, \qquad f_{xx}(x, y) = e^x \cos y,$$

y

$$f_y(x,y) = -e^x \sin y, \qquad f_{yy}(x,y) = -e^x \cos y,$$

por tanto, la laplaciana de f es

$$\Delta f(x,y) = f_{xx}(x,y) + f_{yy}(x,y) = e^x \cos y - e^x \cos y \equiv 0,$$

y así f es armónica.

Existen muchas aplicaciones físicas de las funciones armónicas pues describen con alto grado de precisión diferentes fenómenos naturales. Por ejemplo, los procesos de conducción del calor en estado estacionario (es decir, una vez que la temperatura del material conductor se ha estabilizado y que no varía en el tiempo) donde se considera el vector densidad del flujo de calor cuyo potencial $\phi(x,y)$ (la temperatura en el medio conductor) satisface la ecuación de Laplace en todo punto del espacio donde no haya fuentes o sumideros de calor. Las funciones armónicas aparecen también al describir el flujo de un fluido *ideal* (no viscoso pues no hay pérdidas de energía por fricción interna) incompresible donde se considera el campo de velocidades de un fluido irrotacional (ausencia de vórtices o remolinos) cuyo potencial es armónico o en la teoría de la electrostática donde la concentración del flujo eléctrico en un punto del espacio se describe por medio del vector densidad de flujo eléctrico cuyo potencial electrostático verifica, en una región sin cargas, la ecuación de Laplace.

Laplaciana bidimensional en coordenadas polares

La introducción de coordenadas polares:

$$x = r\cos\theta, \quad y = r\sin\theta,$$

transforma $u(x,y)$ en $v(r,\theta)$ y considerando la laplaciana bidimensional:

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \tag{1.54}$$

se tiene (aplicando la regla de la cadena estudiada en el primer curso) la siguiente fórmula:

$$\Delta u \equiv \frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\frac{\partial v}{\partial r} + \frac{1}{r^2}\frac{\partial^2 v}{\partial \theta^2}.$$

Ejemplo 1.4.2 *Probar que la función $u(r,\theta) = r^2\cos 2\theta$ es una función armónica.*

Es análogo al ejemplo hecho en coordenadas cartesianas. Escribiendo la ecuación de Laplace en coordenadas polares se tiene que u es una función armónica si

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} \equiv 0.$$

Efectuando las derivaciones indicadas se tiene:

$$\frac{\partial^2 u}{\partial r^2} = 2\cos 2\theta, \quad \frac{1}{r}\frac{\partial u}{\partial r} = 2\cos 2\theta, \quad \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = -4\cos 2\theta.$$

luego u es armónica.

Laplaciana tri-dimensional en coordenadas cilíndricas

La introducción de coordenadas cilíndricas:

$$x = r\cos\theta, \quad y = r\sin\theta, \quad z = z,$$

transforma $u(x, y, z)$ en $v(r, \theta, z)$ y considerando la laplaciana tri-dimensional:

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}, \tag{1.55}$$

se tiene (mediante aplicación de la regla de la cadena) la siguiente fórmula:

$$\Delta u \equiv \frac{\partial^2 v}{\partial r^2} + \frac{2}{r}\frac{\partial v}{\partial r} + \frac{1}{r^2}\frac{\partial^2 v}{\partial \theta^2} + \frac{\partial^2 v}{\partial z^2}.$$

Laplaciana tri-dimensional en coordenadas esféricas

La introducción de coordenadas esféricas:

$$x = r\cos\theta\sin\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\phi,$$

transforma $u(x, y, z)$ en $v(r, \theta, \phi)$ y considerando la laplaciana tri-dimensional (1.55) se tiene (por aplicación de la regla de la cadena) la siguiente fórmula:

$$\Delta u \equiv \frac{\partial^2 v}{\partial r^2} + \frac{2}{r}\frac{\partial v}{\partial r} + \frac{1}{r^2}\frac{\partial^2 v}{\partial \phi^2} + \frac{1}{r^2}\frac{\cos\phi}{\sin\phi}\frac{\partial v}{\partial \phi} + \frac{1}{r^2}\frac{1}{\sin\phi}\frac{\partial^2 v}{\partial \theta^2}.$$

Laplaciana N-dimensional en coordenadas radiales

Un tipo de coordenadas extremadamente útil en el estudio de problemas de transporte en los cuales existan ciertas condiciones de simetría son las coordenadas radiales. Sea $\boldsymbol{x} \in \Omega \subset I\!R^N$, $\boldsymbol{x} = (x_1, x_2, ..., x_N)$. Definimos el cambio de variables:

$$r = \sqrt{x_1^2 + x_2^2 + ... + x_N^2}, \qquad u(\boldsymbol{x}) = v(r).$$

Entonces se tiene que:

$$\Delta u \equiv \frac{1}{r^{N-1}} \frac{d}{dr} \left(r^{N-1} \frac{dv}{dr} \right) = \frac{d^2 v}{dr^2} + \frac{N-1}{r} \frac{dv}{dr}.$$

1.4.2. Algunos fenómenos físico-técnicos que modelizan

Tal y como vimos en la sección anterior las ecuaciones de Laplace y Poisson en dos o tres dimensiones aparecen en problemas que conciernen campos potenciales como el electrostático, el gravitacional o el campo de velocidades (en mecánica de fluidos al trabajar con flujos potenciales incompresibles). Por ejemplo, la velocidad potencial para el flujo estacionario de un fluido incompresible y no viscoso satisface la ecuación de Laplace. En estas hipótesis también la función de corriente verifica la ecuación de Laplace. Es la expresión matemática de la idea de que en ausencia de fuentes o sumideros la tasa con la cual el fluido incompresible entra en una región es la misma con la cual la deja. Una solución de la ecuación de Laplace se puede interpretar también como la distribución de temperatura en un estado de equilibrio estable. Aparece por tanto en la descripción de un régimen (de conducción) estacionario. También es indicativa para la descripción del comportamiento de algunos sistemas transitorios (que evoluciona con el tiempo) que se estabilizan para tiempos grandes (a largo plazo).

De forma parecida, el potencial eléctrico V asociado a una distribución bidimensional de electrones con densidad de carga ρ satisface la ecuación

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\rho}{\epsilon} = 0,$$

siendo ϵ la constante dieléctrica. Esta ecuación no es otra cosa que la expresión del teorema de Gauss que afirma que el flujo eléctrico total a través de cualquier superficie cerrada es igual a la carga total encerrada por la superficie.

Otra ecuación destacada de la Matemática Aplicada es la ecuación elíptica

bidimensional de difusión-convección estacionaria

$$\underbrace{\boldsymbol{v}\cdot\nabla T}_{\text{convección}} = \underbrace{\epsilon\nabla^2 T}_{\text{difusión}},$$

que se verifica en un dominio $\Omega \subset I\!R^2$ donde el campo de velocidades $\boldsymbol{v}$ es irrotacional, es decir,

$$\boldsymbol{v} = (\psi_y, -\psi_x),$$

siendo $\psi(x, y)$ la función de corriente y donde ϵ es el coeficiente de dispersión del medio. Las líneas de corriente del flujo son las curvas donde ψ es constantes (véase el ejemplo examinado al comienzo de este tema para el cálculo de las líneas de corriente en el flujo de un fluido).

1.4.3. * Fórmulas de Green

Introduciremos ahora las **fórmulas de Green**[36], una herramienta muy importante en la teoría del potencial. Su aplicación a las funciones armónicas permitirá deducir sus propiedades fundamentales.

Recordemos en primer lugar el teorema de la divergencia (primer curso, tema 13) a partir del cual será posible deducir las fórmulas de Green.

Teorema 1.4.1 (Gauss-Ostrogradski) *Sea $\Omega \subset I\!R^3$ un dominio acotado limitado por la superficie $\partial\Omega$, orientable y cerrada. Sea $\boldsymbol{F} : D \subset I\!R^3 \to I\!R^3$, $\Omega \subset D$, un campo vectorial*

$$\boldsymbol{F}(x, y, z) = P(x, y, z)\boldsymbol{i} + Q(x, y, z)\boldsymbol{j} + R(x, y, z)\boldsymbol{k},$$

tal que sus componentes P, Q y R son campos escalares continuos con derivadas parciales continuas en Ω: $\boldsymbol{F} \in (C^1(\Omega))^3$. Entonces el flujo del campo $\boldsymbol{F}$ a través de la superficie (cerrada) $\partial\Omega$ es igual a la integral triple (integral de volumen) de la divergencia del campo:

$$\int_{\partial\Omega} \boldsymbol{F}\cdot\boldsymbol{n}ds = \int_{\Omega} div(\boldsymbol{F})dxdydz,$$

siendo $\boldsymbol{n}$ el vector normal exterior a la superficie, y

$$div(\boldsymbol{F}) = \nabla\cdot\boldsymbol{F} = \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z},$$

el operador de divergencia.

[36] Recuérdese que las fórmulas de Green ya han sido consideradas en la sección 12.4 del tema 12 del primer curso. Un ejercicio interesante se encuentra propuesto (y resuelto) en el guión de ejercicios.

Utilizando el teorema de Gauss podemos ahora deducir las fórmulas de Green.

Sean ϕ, ψ dos funciones escalares de clase C^2 en Ω.

Observación 1.4.1 *Hipótesis de regularidad suficientes para la aplicación de las fórmulas de Green son: $\psi \in C^2(\Omega) \cap C^1(\bar{\Omega})$, $\phi \in C^1(\Omega) \cap C^0(\bar{\Omega})$ para la primera y $\psi, \phi \in C^2(\Omega) \cap C^1(\bar{\Omega})$ para la segunda fórmula de Green (ver Courant-Hilbert, pag 252). Tales hipótesis no son en realidad necesarias y es posible suponer una regularidad menor. Para ello es necesario conocer la teoría de distribuciones que no se contempla en este curso.*

Entonces, utilizando las propiedades del operador de divergencia:

$$\nabla \cdot (\phi \nabla \psi) = \nabla \phi \cdot \nabla \psi + \phi \nabla \cdot \nabla \psi = \nabla \phi \cdot \nabla \psi + \phi \Delta \psi$$

y considerando $\boldsymbol{F} = \phi \nabla \psi$ en el teorema de la divergencia se tiene:

Primera fórmula de Green

$$\int_\Omega (\nabla \phi \cdot \nabla \psi + \phi \nabla^2 \psi) d\Omega = \int_{\partial\Omega} \phi \frac{\partial \psi}{\partial \boldsymbol{n}} ds, \tag{1.56}$$

donde $\dfrac{\partial \psi}{\partial \boldsymbol{n}} = \boldsymbol{n} \cdot \nabla \psi$ es la componente normal exterior del vector gradiente.

La relación (1.56) se conoce con el nombre de primera fórmula de Green. Intercambiando ϕ y ψ y restando se tiene la segunda fórmula de Green:

Segunda fórmula de Green

$$\int_\Omega (\phi \nabla^2 \psi - \psi \nabla^2 \phi) d\Omega = \int_{\partial\Omega} \left(\phi \frac{\partial \psi}{\partial \boldsymbol{n}} - \psi \frac{\partial \phi}{\partial \boldsymbol{n}} \right) ds. \tag{1.57}$$

Nótese finalmente que fórmulas análogas se tienen en el plano y en dimensiones arbitrarias (ver Courant-Hilbert, pag 257).

Utilizando las fórmulas anteriores es posible deducir las siguientes propiedades de las funciones armónicas:

Teorema 1.4.2 *Si $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$ es una función armónica y si $\partial\Omega$ es una superfice orientable y cerrada, entonces la integral de superficie de su derivada normal vale cero:*

$$\int_{\partial\Omega} \frac{\partial u}{\partial \boldsymbol{n}} = 0 \tag{1.58}$$

La demostración de lo anterior es inmediata considerando $\phi = 1$, $\psi = u$ armónica ($\Delta u = 0$) y aplicando la primera fórmula de Green. La integral (de superficie) (1.58) se denomina Integral de Gauss.

Si $\phi \in C^2(\Omega) \cap C^1(\bar{\Omega})$ es una función armónica y consideramos $\phi = \psi$ en la primera fórmula de Green se obtiene la identidad:

$$\int_{\Omega} |\nabla \phi|^2 d\Omega = \int_{\partial\Omega} \phi \frac{\partial \phi}{\partial \boldsymbol{n}}.$$

La integral de volumen que aparece en la fórmula anterior se conoce con el nombre de Integral de Dirichlet y juega un papel muy importante en la teoría del potencial. Una consecuencia inmediata de la identidad anterior es el siguiente teorema para funciones armónicas:

Teorema 1.4.3 *Siendo $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$ una función armónica, $\Omega \subset I\!R^3$ un dominio acotado limitado por la superficie $\partial\Omega$ orientable y cerrada se verifica:*

1. *Si u se anula en la superficie $\partial\Omega$ entonces $u \equiv 0$ en Ω.*
2. *Si $\dfrac{\partial u}{\partial \boldsymbol{n}} = 0$ en $\partial\Omega$ entonces u es constante en Ω.*

La demostración es inmediata observando que en ambos casos $\displaystyle\int_{\partial\Omega} \phi \frac{\partial \phi}{\partial \boldsymbol{n}} = 0$ luego la integral de Dirichlet es nula y por tanto u es constante. En el primer caso además la constante tiene que coincidir con el valor en la frontera (que es cero).

Si Ω es una esfera de centro (x_0, y_0, z_0), radio R, superficie $\partial\Omega$ y u es una función armónica que satisface el teorema (1.4.2) se tiene:

$$u(x_0, y_0, z_0) = \frac{1}{4\pi R^2} \int_{\partial\Omega} u ds,$$

es decir:

Teorema 1.4.4 *El valor de una función armónica en un punto (x_0, y_0, z_0) es igual a la media aritmética de sus valores en cada esfera centrada en (x_0, y_0, z_0).*

El teorema (1.4.4) tiene importante consecuencias:

Principio del máximo para funciones armónicas

Teorema 1.4.5 *Sea u una función regular en una región conexa Ω y continua en la superficie $\partial\Omega$ que la limita. Entonces el valor máximo y mínimo de u se alcanzan en la frontera $\partial\Omega$. La función u alcanza sus valores máximo y mínimo en el interior Ω si y sólo si u es constante.*

Corolario 1.4.1 *Si una función armónica u, regular en Ω y continua en la frontera $\partial\Omega$ es constante en $\partial\Omega$ entonces es constante en todo Ω.*

Unicidad para la ecuación de Laplace

Una simple aplicación de la primera fórmula de Green permite demostrar la unicidad de soluciones de la ecuación de Laplace en un volumen $\Omega \subset I\!R^3$ con condiciones Dirichlet no homogéneas en la superficie del volumen $\partial\Omega$. [37]

Sean ψ_1 y ψ_2 dos soluciones de la ecuación de Laplace en un volumen $\Omega \subset I\!R^3$ satisfaciendo la condición de contorno $\psi_i = f(x)$, $i = 1, 2$ en la frontera $\partial\Omega$. Entonces, por la linealidad del operador, se tiene que la función diferencia, $\Theta = \psi_1 - \psi_2$ verifica $\Delta\Theta = 0$ y $\Theta = 0$ en la frontera y por la I fórmula de Green se tiene:

$$\int_\Omega |\nabla\Theta|^2 d\Omega = \int_{\partial\Omega} \Theta \frac{\partial\Theta}{\partial n} d\partial\Omega;$$

pero $\Theta = 0$ en $\partial\Omega$, luego

$$\int_\Omega |\nabla\Theta|^2 d\Omega = 0.$$

Por ser el integrando no negativo lo anterior implica $|\nabla\Theta|^2 = 0$ en cualquier punto de Ω luego Θ es constante en Ω y como vale cero en $\partial\Omega$ se tiene: $\Theta \equiv 0$ y $\psi_1 \equiv \psi_2$.

Otra forma de expresar este resultado (utilizando los teoremas anteriores) es la siguiente:

Teorema 1.4.6 *Dos funciones armónicas en Ω que sean continuas en $\Omega \cup \partial\Omega$ y coincidan en $\partial\Omega$ son idénticas en todo Ω.*

La demostración es inmediata observando que la diferencia entre dos funciones armónicas que satisfagan el teorema (1.4.6) es también una función armónica que se anula en la frontera luego, por el corolario (1.4.1), es idénticamente nula en Ω.

[37] Este tipo de razonamiento se puede extender para obtener la unicidad de la solución de la ecuación de Laplace con condiciones de contorno más generales.

1.5. Ecuaciones parabólicas.

Las ecuaciones en derivadas parciales de segundo orden de tipo parabólico se presentan al estudiar los procesos de conducción térmica, difusión de materia y en general procesos termodinámicos. Modelizan fenómenos transitorios, de evolución (en contraposición con los estados de equilibrio de las ecuaciones elípticas, estacionarias), donde el estado del sistema varía con el tiempo. En general, tales fenómenos de difusión pueden completarse con procesos de convección y dispersión que aparecen, por ejemplo, en varios modelos de difusión de contaminantes o en filtración de aguas subterráneas.

1.5.1. La ecuación de difusión y algunas variantes

Empezaremos por el caso más sencillo: la ecuación parabólica de conducción térmica uni-dimensional, conocida con el nombre de ecuación del calor:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 \leq x \leq 1, \quad t > 0. \tag{1.59}$$

Nótese la ausencia de fuentes de calor internas al sistema (la ecuación es, en efecto, homogénea). Las ecuaciones del tipo (1.59) se resolverán analíticamente en el tema 2 (mediante el método de separación de variables) y en el tema 3 (mediante la técnica de la transformada de Laplace) donde se considerarán situaciones más generales. La ecuación (1.59) puede ser generalizada de distintas formas:

1. Introduciendo un término de fuente (o sumidero) estacionario $q(x)$ en el interior del dominio:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + q(x), \qquad 0 \leq x \leq 1, \quad t > 0, \tag{1.60}$$

   donde $q(x) > 0$ si es una fuente y $q(x) < 0$ si es un sumidero. La ecuación (1.60) aparece en problemas de conducción de calor transitorios (evolutivos) con generación de energía. El término de generación de energía (o término de fuente) $q(x)$ se conoce en la literatura como *steady forcing* (fuente estacionaria). Este tipo de ecuaciones se puede resolver analíticamente con algunas variantes de los métodos que se utilizan para las ecuaciones del tipo (1.59). Un ejemplo de aplicación a un problema concreto se puede encontrar en el libro de Mei, pag 75.

2. Introduciendo un término de fuente transitorio (o fuente de calor) en el interior del dominio:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \qquad 0 \leq x \leq 1, \quad t > 0, \tag{1.61}$$

siendo $f(x,t) > 0$. La ecuación (1.61) es la clásica ecuación de reacción-difusión lineal. El término de generación de energía $f(x,t)$ se conoce en la literatura como *transient forcing* (fuente transitoria). También este tipo de ecuaciones se puede resolver análiticamente con algunas variantes de los métodos que se utilizan para las ecuaciones del tipo (1.59) y (1.60). Un ejemplo de aplicación a un problema concreto se puede encontrar en el libro de Mei, pag 75.

3. Considerando los efectos de disipación viscosa (producción de energía térmica y disipación de la energía mecánica) en el interior del dominio debidos a procesos termodinámicos:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t,u), \qquad 0 \leq x \leq 1, \quad t > 0, \tag{1.62}$$

siendo $f(x,t,u) > 0$. Debido al signo (positivo) de $f(x,t,u)$ este término tiene el efecto cualitativo de generar crecimiento en las soluciones, pudiendo aparecer efectos de *blow up* (singularidades o *explosiones*) de las mismas soluciones que tienen así un carácter local. La ecuación (1.62) es **lineal** si $f(x,t,u)$ es lineal en u. Si $f(x,t,u)$ no es lineal en u entonces se dice **semilineal**. Si f fuese del tipo $f(x,t,u,u_x)$, es decir si dependiera también de la derivada parcial primera u_x y además lo hiciera en forma no lineal entonces la ecuación sería de tipo **cuasilineal**.

4. Incluyendo un término de absorción (o sumidero) en el interior del dominio:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - f(x,t,u), \qquad 0 \leq x \leq 1, \quad t > 0, \tag{1.63}$$

siendo $f(x,t,u) > 0$. Valen las mismas consideraciones sobre el carácter de linealidad de (1.63) hechas en el caso anterior. Cualitativamente el término de absorción causa un decaimiento en los perfiles de las soluciones. En la forma $f(x,t,u) = u$ y en problemas de conducción de calor, se le conoce como el término de enfriamiento de Newton.

5. Permitiendo un coeficiente de difusión variable espacialmente

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(x)\frac{\partial u}{\partial x}\right], \qquad 0 \leq x \leq 1, \quad t > 0, \tag{1.64}$$

siendo $a = a(x)$ una función positiva y acotada en $[0,1]$, es decir $0 < a(x) < \infty$. Este tipo de ecuaciones nace, por ejemplo, al considerar procesos de conducción de calor siendo la conductividad calorífica k una función no constante en todo el medio: $k = k(x)$.

6. Introduciendo efectos no lineales de difusión mediante un coeficiente del tipo $a(x,u)$:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(x,u)\frac{\partial u}{\partial x}\right], \qquad 0 \leq x \leq 1, \quad t > 0. \tag{1.65}$$

La ecuación (1.65) describe fenómenos de filtración en medios porosos donde la expresión típica del coeficiente es $a(x,u) = u^m$ o procesos de conducción calorífica en un medio cuya conductibilidad térmica depende de la propia temperatura.

Son ecuaciones de tipo **no lineal** y no suelen tener soluciones analíticas exactas, especialmente cuando se complementan con un término de fuente $f(x,t)$. En estos casos se suelen calcular numéricamente las soluciones de (1.65).

7. Modelizando el flujo de fluidos no newtonianos mediante un coeficiente del tipo $a(x,u,u_x)$:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left[a(x,u,u_x)\frac{\partial u}{\partial x}\right], \qquad 0 \leq x \leq 1, \quad t > 0. \tag{1.66}$$

Los comentarios hechos para (1.65) valen, obviamente, en este caso más general. No suele ser posible resolver[38] este tipo de ecuaciones en forma exacta y hay que acudir a métodos numéricos.

8. Considerando fenómenos de convección (por ejemplo en problemas de fluidodinámica)

$$\frac{\partial u}{\partial t} + V(x,t)\frac{\partial u}{\partial x} = \mu\frac{\partial^2 u}{\partial x^2}, \qquad 0 \leq x \leq 1, \quad t > 0, \tag{1.67}$$

siendo $V(x,t)$ una distribución de velocidades en la dirección longitudinal x conocida. La ecuación (1.67) es una ecuación de difusión-convección. Es lineal pues $V(x,t)$ es un dato del problema independiente de u. La incógnita u suele representar la concentración de un contaminante (o más en general de una especie química) o la temperatura de un fluido y μ es el coeficiente de difusión del medio.

9. Si $V(x,t) = u(x,t)$, siendo u la incógnita del problema representando un campo de velocidades en la dirección x se tiene

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \mu\frac{\partial^2 u}{\partial x^2}, \quad t > 0. \tag{1.68}$$

[38] En realidad, como ya observamos, la posibilidad de resolución analítica no está descartada en presencia de ciertas simetrías en la ecuación y en las condiciones de contorno. En tal caso se suele usar el método de combinación de variables que es un método aplicable al caso lineal en los supuestos anteriores. Más detalles sobre este método se verán en el tema 3.

Si $\mu = \epsilon$, siendo $0 < \epsilon \ll 1$ un parámetro pequeño, la ecuación anterior modeliza flujos poco viscosos y se conoce en la literatura como la ecuación de Bürgers. La ecuación (1.68) es una ecuación de tipo no lineal debido al término no lineal uu_x de convección. El término de *viscosidad* ϵu_{xx} tiene un papel *regularizante* sobre las soluciones. En el límite $\epsilon \to 0$ se tiene la ecuación hiperbólica de primer orden considerada en la primera parte de este tema. Se pierden las propiedades regularizantes propias de la ecuación del calor y se pueden desarrollar singularidades en forma de ondas de choque.

10. Simulando fenómenos de dispersión:

$$c(x,t)\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 \leq x \leq 1, \quad t > 0, \tag{1.69}$$

siendo $c(x,t) > 0$. En problemas térmicos $c(x,t)$ está relacionada con la capacidad calorífica del medio. En problemas de flujo con el coeficiente de almacenamiento.

11. Dejando variar x en todo $\mathbb{R}$ se pueden considerar problemas en dominios infinitos:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad -\infty < x < \infty, \quad t > 0. \tag{1.70}$$

Dos aplicaciones concretas de este tipo de modelos se encuentran en el libro de Mei, pag 137 y 139. Este tipo de problemas de difusión (o conducción) unidimensional en dominios no acotados se resolverán analíticamente mediante el método de la Transformada de Fourier en el tema 3. Se suelen completar con condiciones de decaimiento en el infinito ($u \to \pm\infty$ cuando $|x| \to \infty$) y con condiciones iniciales especiales que simulan impulsos (fuerzas puntuales, localizadas) aplicadas al sistema. También se modelizan problemas con cargas puntuales o temperaturas iniciales discontinuas.

12. Dejando variar x en todo $\mathbb{R}^+ = (0, \infty)$ (dominio semi-infinito):

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < \infty, \quad t > 0. \tag{1.71}$$

Nuevamente se trata de procesos de difusión o conducción unidimensional. Se suelen resolver con el método de las Transformadas senos y cosenos de Fourier. Analizaremos algunos casos modelo en el segundo tema.

13. Considerando más variables espaciales:

$$\frac{\partial u}{\partial t} = \Delta u, \qquad x \in \Omega \subset \mathbb{R}^n, \quad t > 0, \tag{1.72}$$

siendo Ω un dominio acotado. Para $n = 2$ se puede encontrar un ejemplo de proceso de conducción de calor bidimensional en un rectángulo en el libro de Mei, pag 78. Varios ejemplos aparecen también en el libro de Zill y en general en la bibliografía comentada al final del capítulo. En el tema 2 desarrollaremos un ejemplo de difusión de la concentración de una especie química en un dominio circular.

14. Incluyendo varios de los fenómenos y casos anteriores en la clase de ecuaciones parabólicas cuasilineales:

$$c\frac{\partial u}{\partial t} = \Delta u + f(x, t, u, \nabla u), \qquad x \in \Omega \subset I\!R^n, \quad t > 0, \tag{1.73}$$

de reacción-difusión-dispersión-convección.

Las ecuaciones anteriores y otras similares se tienen que complementar con una condición inicial y adecuadas condiciones en el contorno del dominio para obtener problemas de valor inicial y de contorno o problemas (elípticos) sólo de contorno que estén bien planteados. En el caso de dominios no acotados (sin contorno, del tipo que aparece en (1.71)), complementando la ecuación con una condición inicial se obtiene el problema de Cauchy o de valor inicial para ecuaciones en derivadas parciales.

1.5.2. * Algunos fenómenos físico-técnicos que modelizan

Ilustramos aquí, muy brevemente, otros modelos matemáticos (algunos básicos y otros avanzados) que aparecen en las ciencias aplicadas[39].

El proceso de difusión (de materia o calor por ejemplo) se vuelve más interesante cuando otros procesos tienen lugar. Por ejemplo, si durante una reacción química se libera calor internamente con una tasa, digamos $f(T)$, por unidad de volumen entonces la ecuación del calor toma la forma:

$$T_t = \nabla^2 T + f(T).$$

Aunque las soluciones existan y sean únicas para las elecciones típicas que aparecen en las aplicaciones, éstas no tienen porque existir para todo tiempo t. Así, podemos ver que T se hace singular en un tiempo finito si $f'' > 0$ (funciones convexas), caso, por ejemplo en el que $f(T) = T^2$ y este fenómeno se conoce como *blow up*. El ejemplo más conocido es cuando $f(T) = \lambda e^T$ que

[39] Véase el libro de A.C. Fowler, Mathematical Models in the Applied Sciences. (1997).

aparece en teoría de la combustión. Este fenómeno es propio de la difusión no lineal.

Otra ecuación de difusión no lineal viene dada por ejemplo por la ecuación de calor con conductibilidad térmica dependiente de la temperatura:

$$\rho c_p T_t = \nabla \left(k(T) \nabla T \right).$$

Información sobre este tipo de dependencia en los modelos químicos se puede encontrar en el vol 2 del libro de Costa Novella sobre Fenómenos de Transporte donde hay una aplicación concreta pero al caso unidimensional (EDO).

El flujo de un gas en un medio poroso es otro fenómeno típico de difusión no lineal. Para verlo, se considera la ley de conservación (de la masa)

$$\rho_t + \nabla \cdot \boldsymbol{q} = 0,$$

donde $\boldsymbol{q}$ es el flujo (de materia) y ρ la densidad del gas. En analogía con la ley de Fourier y la ley de Fick, este flujo viene dado (en hipótesis de flujo lento) por la ley de Darcy:

$$\boldsymbol{q} = -\frac{k}{\mu} \rho \nabla p,$$

donde k es la *permeabilidad* y μ la viscosidad del gas. Aquí p representa la presión (responsable del flujo). Finalmente, utilizando la ley constitutiva para los gases perfectos (o ideales) $p = \rho RT$ se tiene que (en hipótesis de flujo isotermo, es decir, T constante):

$$\rho_t = \nabla \cdot \left[\left\{ \frac{kRT}{\mu} \right\} \rho \nabla \rho \right],$$

que es una ecuación del tipo (1.65). Esta ecuación, escrita en variables adimensionales adecuadas se puede resolver mediante el método de combinación de variables que estudiaremos en el tercer tema de esta asignatura. Nótese que el coeficiente de difusión $a(x, \rho) = c\rho$ ($c = kRT/\mu$ es constante) tiende a cero cuando ρ tiende a cero, un comportamiento muy interesante conocido con el nombre de *degeneración* y que consideraremos más adelante. Sólo observamos que la señal evidente de difusión no lineal degenerada es la existencia de un frente que se propaga con velocidad finita (contrariamente al caso de la ecuación del calor propiamente dicha donde hay velocidad inifinita de propagación de las perturbaciones). Digamos que las ecuaciones parabólicas degeneradas tienen propiedades cualitativas más propias de las ecuaciones hiperbólicas que de las parabólicas.

Otro contexto donde se generan ecuaciones parabólicas no lineales degeneradas es el de la modelización del flujo de aguas subterráneas en un medio

poroso no saturado. Complementando la Ley de Darcy con una ecuación para la conservación de la fase (o las fases[40]) y siendo ϕ la porosidad del suelo, ρ la densidad material (masa por unidad de volumen *del fluido*) la ecuación que describe el modelo para el flujo en un medio poroso rígido es:

$$\frac{\partial(\rho\phi)}{\partial t} = \nabla \cdot \left[\frac{k}{\mu}\rho\nabla\rho\right].$$

Otro ejemplo es el flujo de agua en un terreno saturado que se describe (en el caso unidimensional y en la dirección transversal z) mediante la ecuación:

$$\phi_t - V\phi_x = (D\phi_z)_z,$$

siendo V y D constantes positivas y ϕ la porosidad del medio.

Ecuaciones del tipo (1.65) (complementadas con un término de reacción del tipo $f(x,t)$) surgen también al describir la evolución (en el caso isotermo) del espesor $h(x,t)$ de una capa de hielo, en la forma:

$$h_t = \frac{\partial}{\partial x}\left[\frac{1}{3}h^3 h_x\right] + a(x,t).$$

En este caso $m = 3$, y $a(x,t)$ es una función que representa la tasa de acumulación/ablación de hielo debida a las condiciones atmosféricas y la ecuación es una ecuación de difusión no lineal degenerada[41]. Excepto cuando $a(x,t) \equiv 0$ esta ecuación no puede ser resuelta analíticamente. Se puede sin embargo analizar cuantitativamente (es decir, a nivel numérico) o cualitativamente (mediante métodos de energía). También es posible realizar un detallado análisis asintótico (una técnica propia de la teoría de la perturbación). Una ulterior dificultad (especialmente a nivel numérico aunque existen técnicas avanzadas para su tratamiento) consiste en el hecho que el dominio donde se tiene que verificar la ecuación no es conocido *a priori* y tiene que ser determinado junto con la solución. El marco correcto en el cual considerar esta ecuación se construye mediante la teoría de operadores multívocos y el problema es unilateral[42] (del

[40] Por ejemplo en un yacimiento petrolífero las fases son petróleo y agua.

[41] Intuitivamente hay degeneración cuando el coeficiente de difusión se anula en alguna subregión del dominio. Se trata de un fenómeno propio de las ecuaciones no lineales (es decir no puede ocurrir en las ecuaciones lineales) pero no ocurre en todas las ecuaciones no lineales siendo caracterizado por un delicado balance entre la velocidad de difusión y el tamaño del dominio. Un estudio detallado se encuentra en el libro de J.I. Díaz sobre fronteras libres citado en la bibliografía avanzada. Nótese que el fenómeno puede darse tanto en ecuaciones elípticas como en ecuaciones parabólicas siendo necesario pero no suficiente el carácter no lineal del término de difusión.

[42] Se trata de un tipo de formulación matemática muy utilizado en mecánica de fluidos que permite incorporar al problema matemático distintas fenomenologías que satisfacen unas condiciones de admisibilidad física.

tipo problema de obstáculo). La teoría y el comportamiento cualitativo de las soluciones de esta ecuación no lineal depende, críticamente, de los valores de m según que se tenga $0 < m < 1$ (caso singular) o $m > 1$ (caso degenerado).

En general se emplea la ecuación (1.66) (de tipo no lineal parabólico eventualmente degenerado) para modelizar las dinámicas no lineales de los fluidos geofísicos. Por ejemplo, la expresión $a(x, u, u_x) = |u_x|^{p-2}$ permite describir el flujo lento, viscoso, no newtoniano de las grandes masas de hielo polar que recubren la Antártida y Groenlandia. Otro fluido geofísico no newtoniano es el magma.

Finalizamos este tema con unos comentarios sobre los sistemas de ecuaciones en derivadas parciales. Es evidente (véanse los comentarios en la primera sección sobre los fenómenos de transporte) que una descripción detallada del estado de un sistema se puede obtener a partir de la consideranción simultánea de las ecuaciones en derivadas parciales que nacen al aplicar las leyes de conservación. Surgen así unos sistemas **modelo** que aparecerán muy a menudo en la carrera de un ingeniero químico. El modelo fundamental o básico, a partir del cual se obtienen otros casos digamos *habituales*, es el sistema de ecuaciones que se genera al considerar la ecuación de conservación de la masa (ecuación de continuidad) con la de conservación de la cantidad de movimiento (ecuación del equilibrio) para la descripción del flujo de un fluido newtoniano. Nos referimos al sistema de **Navier Stokes**:

$$\left\{ \begin{array}{rcl} \nabla \cdot \boldsymbol{u} & = & 0, \\ \\ \boldsymbol{u}_t + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} & = & -\dfrac{1}{\rho}\nabla p + \nu \nabla^2 \boldsymbol{u}, \end{array} \right. \tag{1.74}$$

siendo $\nu = \mu/\rho$ la *viscosidad cinemática*, μ la viscosidad dinámica, ρ la densidad, p la presión y $\boldsymbol{u}$ la velocidad. Si U es una velocidad típica del fluido y l es una dimensión típica de la geometría del flujo entonces el sistema (1.74) se puede escribir en variables adimensionales en la forma

$$\left\{ \begin{array}{rcl} \nabla \cdot \boldsymbol{u} & = & 0, \\ \\ \boldsymbol{u}_t + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} & = & -\nabla p + \dfrac{1}{\mathrm{Re}}\nabla^2 \boldsymbol{u}, \end{array} \right. \tag{1.75}$$

siendo el parámetro adimensional $\mathrm{Re} = Ul/\nu$ llamado **número de Reynolds**. Si $\mathrm{Re} \ll 1$ (es decir, para valores del número de Reynolds muy pequeños) el movimiento del flujo es muy lento (se llama **flujo de Stokes**) y la ecuación del momento se puede aproximar por:

$$\nabla p = \nabla^2 \boldsymbol{u},$$

donde la presión ha sido reescalada mediante 1/Re. Se trata de una perturbación regular, al menos en dominios finitos. Si Re$\gg 1$ (es decir, si el número de Reynolds es muy grande) entonces la aproximación de la ecuación del equilibrio es la **ecuación de Euler**:

$$\boldsymbol{u}_t + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u} = -\nabla p,$$

que se obtiene en el límite:

$$\lim_{\mathrm{Re}\to\infty}\left(\frac{1}{\mathrm{Re}}\nabla^2\boldsymbol{u}\right)\to 0.$$

Se trata de una perturbación singular pues al eliminar los efectos viscosos se reduce el orden de la ecuación y pueden aparecer muchas complicaciones.

Tras el abanico de posibilidades evidenciado en esta sección (y en las anteriores) podemos afirmar que las ecuaciones diferenciales en derivadas parciales gobiernan numerosos fenómenos que aparecen en física-matemática. Esto no quiere decir que las ecuaciones ordinarias (que suelen aparecer en modelos más básicos) no puedan ser válidas para modelizar adecuadamente un fenómeno. Tampoco es cierto (en general) que las ecuaciones ordinarias son simples y más facilmente resolubles pero es evidente que las simplificaciones (mediante análisis dimensional) de los modelos suelen empobrecer el grado de aproximación a la realidad del fenómeno físico considerado. Digamos que se suele simplificar el modelo[43] hasta poder alcanzar una solución analítica exacta o analítica aproximada del mismo que permita validar los resultados numéricos. Tales simplificaciones suelen venir sugeridas por el análisis dimensional (previo) de las ecuaciones lo que proporciona una aproximación *sensible* del modelo originario. Las ecuaciones en derivadas parciales representan, por tanto, un grado superior de aproximación a la realidad del fenómeno modelizado sin que ello haga inútil o redundante el manejo de las ecuaciones ordinarias. Nótese en efecto que utilizaremos EDO para resolver analíticamente EDP.

Para los ingenieros químicos el campo de aplicación dominante de las ecuaciones en derivadas parciales es, sin duda, el de los fenómenos de transporte entendiéndose por ello el transporte difusivo-dispersivo-convectivo-reactivo de materia o transporte conductivo-convectivo de calor. Además las técnicas que veremos para su resolución tendrán múltiples aplicaciones, por ejemplo la Transformada de Laplace al trabajar con control y diseño de experimentos

[43] Una técnica consolidada en la modelización consiste en efecto en desmontar el modelo (deducido a partir de las leyes de conservación y de las leyes constitutivas) en sus piezas fundamentales, analizarlo y, tras ello, volver a considerar, uno a uno, los términos inicialmente despreciados para tener así una idea de los efectos producidos y de su importancia relativa.

o las series de Fourier en el análisis de la estabilidad de los esquemas numéricos en diferencias finitas utilizados para aproximar las soluciones de las EDP. Digamos que proporcionaremos métodos para resolver problemas con la idea de que el alcance de los métodos vaya mucho más allá de la resolución del problema inicialmente planteado. Entraremos en detalles en este tipo de problemas en los temas 2 y 3, al considerar algunos ejemplos concretos que se pueden encontrar, por ejemplo, en los libros de Costa Novella que aparecen en la bibliografía.

1.6. Anexo

En esta sección recordaremos rápidamente varios conceptos necesarios para el entendimiento de las otras secciones. Se trata de un material que básicamente se vió en el curso de primero pero que ha sido complementado con una interpretación directa y práctica del lenguaje matemático asociado a los conceptos de trayectorias y superficies en términos de la mecánica de fluidos. Sirven de enlace con la terminología adoptada en cursos paralelos a éste.

1.6.1. Introducción a las trayectorias

Matemáticamente es útil pensar en una curva C como un conjunto de valores de una función que manda un intervalo de números reales al plano o al espacio. A dicha aplicación le llamaremos **trayectoria**. Por lo común se denota una trayectoria mediante $\boldsymbol{c}$. Entonces la imagen C de una trayectoria corresponde a la curva. Frecuentemente escribimos t como la variable independiente y la imaginamos como el tiempo, de manera que $\boldsymbol{c}(t)$ es la posición en el tiempo t de una partícula en movimiento, la cual describe una curva conforme t varía. También decimos que $\boldsymbol{c}$ (la trayectoria) parametriza C (la curva).

Definición 1.6.1 *Una* **trayectoria** *en* $\mathbb{R}^n$ *es una aplicación* $\boldsymbol{c} : [a, b] \to \mathbb{R}^n$*; Si* $n = 2$*, es una* **trayectoria en el plano** *y si* $n = 3$*, es una* **trayectoria en el espacio**. *La colección* C *de puntos* $\boldsymbol{c}(t)$*, conforme* t *varía en* $[a, b]$ *se denomina* **curva**, *y* $\boldsymbol{c}(a)$, $\boldsymbol{c}(b)$ *son sus* **puntos extremos**. *Se dice que la trayectoria* $\boldsymbol{c}$ **parametriza** *la curva* C.

Si $\boldsymbol{c}$ *es una trayectoria en* $\mathbb{R}^3$*, podemos escribir:*

$$\boldsymbol{c}(t) = (x(t), y(t), z(t)),$$

y llamamos a $x(t), y(t), z(t)$ **funciones componentes** *de* $\boldsymbol{c}$. *Formamos de manera análoga las funciones componentes en* $\mathbb{R}^2$ *o, en general, en* $\mathbb{R}^n$.

Si imaginamos los puntos $\boldsymbol{c}(t)$ como la curva descrita por una partícula y t como el tiempo, es razonable definir el vector velocidad como sigue:

Definición 1.6.2 *Si* $\boldsymbol{c}$ *es una trayectoria y es diferenciable, decimos que* $\boldsymbol{c}$ *es una* **trayectoria diferenciable**. *La* **velocidad** *de* $\boldsymbol{c}$ *en el tiempo* t *se define mediante:*

$$\boldsymbol{c}'(t) = \lim_{h \to 0} \frac{\boldsymbol{c}(t+h) - \boldsymbol{c}(t)}{h}.$$

La **rapidez** *de la trayectoria* $\boldsymbol{c}(t)$ *es la longitud del vector velocidad:* $||\boldsymbol{c}'(t)||$.

Si $\boldsymbol{c}(t) = (x(t), y(t))$ en $\mathbb{R}^2$ entonces:

$$\boldsymbol{c}'(t) = (x'(t), y'(t)) = x'(t)\boldsymbol{i} + y'(t)\boldsymbol{j},$$

y si $\boldsymbol{c}(t) = (x(t), y(t), z(t))$ en $\mathbb{R}^3$ entonces:

$$\boldsymbol{c}'(t) = (x'(t), y'(t), z'(t)) = x'(t)\boldsymbol{i} + y'(t)\boldsymbol{j} + z'(t)\boldsymbol{k}.$$

Nótese que la velocidad $\boldsymbol{c}'(t)$ es un vector **tangente** a la trayectoria $\boldsymbol{c}(t)$ en el tiempo t. Si C es una curva descrita por $\boldsymbol{c}$ y si $\boldsymbol{c}'(t) \neq \boldsymbol{0}$, entonces $\boldsymbol{c}'(t)$ es un vector tangente a la curva C en el punto $\boldsymbol{c}(t)$.

Si $\boldsymbol{c}$ representa la trayectoria de una partícula que se mueve, entonces su **vector velocidad** es una función vectorial $\boldsymbol{v} = \boldsymbol{c}'(t)$ (que depende de t) y su **rapidez** es $||\boldsymbol{v}||$. La derivada $\boldsymbol{a} = d\boldsymbol{v}/dt = \boldsymbol{c}''(t)$ se llama **aceleración** de la curva. Si la curva es $(x(t), y(t), z(t))$, entonces la aceleración en el tiempo t está dada por:

$$\boldsymbol{a}(t) = x''(t)\boldsymbol{i} + y''(t)\boldsymbol{j} + z''(t)\boldsymbol{k}.$$

Muy a menudo se tratará de evaluar campos escalares a lo largo de curvas en el espacio. Particularmente útil es el siguiente resultado (que es un caso particular de la regla de la cadena):

Teorema 1.6.1 *Sea $\boldsymbol{c} : \mathbb{R} \to \mathbb{R}^3$, $\boldsymbol{c}(t) = (x(t), y(t), z(t))$, una trayectoria diferenciable y $f : \mathbb{R}^3 \to \mathbb{R}$ un campo escalar. Sea $h : \mathbb{R} \to \mathbb{R}$ definida por:*

$$h(t) = f(\boldsymbol{c}(t)) = f(x(t), y(t), z(t)).$$

Entonces,

$$\frac{dh}{dt} = \nabla f(\boldsymbol{c}(t)) \cdot \boldsymbol{c}'(t),$$

donde $\boldsymbol{c}'(t) = (x'(t), y'(t), z'(t))$.

Nos preguntamos cual es la longitud de una trayectoria. Puesto que la rapidez $||\boldsymbol{c}'(t)||$ es la tasa de cambio de la distancia recorrida respecto al tiempo, la distancia recorrida por un punto que se mueve a lo largo de la curva debe ser la integral de la rapidez respecto al tiempo sobre el intervalo de tiempo $[t_0, t_1]$.

Definición 1.6.3 *La longitud de una trayectoria, llamada* **longitud de arco** *s (también se suele denotar por σ), es:*

$$s \doteq \int_{t_0}^{t_1} ||\boldsymbol{c}'(t)|| dt.$$

Se deduce que la longitud de arco de la trayectoria $\boldsymbol{c}(t) = (x(t), y(t), z(t))$, para $t_0 \leq t \leq t_1$ es:

$$s \doteq \int_{t_0}^{t_1} \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2} dt.$$

La **función de longitud de arco** $s(t)$ para una trayectoria dada $\boldsymbol{c}(t)$ se define por:

$$s(t) \doteq \int_{t_0}^{t} ||\boldsymbol{c}'(\tau)|| d\tau,$$

y representa la distancia que una partícula, viajando por la trayectoria $\boldsymbol{c}$, habrá recorrido en el tiempo t si comienza en el tiempo t_0; esto es, proporciona la longitud de $\boldsymbol{c}$ entre $\boldsymbol{c}(t_0)$ y $\boldsymbol{c}(t)$.

Pasamos ahora a la definición de la diferencial de la longitud de arco.

Definición 1.6.4 *Un* **desplazamiento infinitesimal** *de una partícula que sigue una trayectoria:*

$$\boldsymbol{c}(t) = x(t)\boldsymbol{i} + y(t)\boldsymbol{j} + z(t)\boldsymbol{k},$$

es

$$d\boldsymbol{s} = dx\boldsymbol{i} + dy\boldsymbol{j} + dz\boldsymbol{k} = \left(\frac{dx}{dt}\boldsymbol{i} + \frac{dy}{dt}\boldsymbol{j} + \frac{dz}{dt}\boldsymbol{k}\right) dt,$$

y su longitud:

$$ds = \sqrt{dx^2 + dy^2 + dz^2} = \sqrt{\left(\frac{dx}{dt}\boldsymbol{i}\right)^2 + \left(\frac{dy}{dt}\boldsymbol{j}\right)^2 + \left(\frac{dz}{dt}\boldsymbol{k}\right)^2} dt,$$

es la **diferencial de longitud de arco**.

La fórmulas anteriores nos ayudan a recordar la fórmula de la longitud de arco:

$$s = \int_{t_0}^{t_1} ds,$$

y, en general, la definición de **integral de trayectoria** (o integral curvilínea o integral del campo escalar $f(x, y, z)$ a lo largo de la trayectoria).

Definición 1.6.5 *Sea $\boldsymbol{c} : [a, b] \to \mathbb{R}^3$ una trayectoria de clase C^1 y $f : \boldsymbol{c}(t) \subset \mathbb{R}^3 \to \mathbb{R}$ continua para cada $t \in [a, b]$. Se define entonces la integral de trayectoria del campo escalar f como:*

$$\int_{\boldsymbol{c}} f ds \doteq \int_{t_0}^{t_1} f(x(t), y(t), z(t)) ||\boldsymbol{c}'(t)|| dt.$$

Los elementos diferenciales ds con respecto de los cuales se integra vienen dados por la longitud de la diferencial de longitud de arco. Nótese que, a veces, la integral de trayectoria se denota con:

$$\int_{\boldsymbol{c}} f ds = \int_{\boldsymbol{c}} f(x,y,z) ds = \int_{t_0}^{t_1} f(\boldsymbol{c}(t)) ||\boldsymbol{c}'(t)|| dt.$$

Un caso importante de la integral de trayectoria se presenta cuando la trayectoria $\boldsymbol{c}$ describe una curva plana. Suponiendo que f es una función real de dos variables la integral de trayectoria a lo largo de $\boldsymbol{c}$ es:

$$\int_{\boldsymbol{c}} f(x,y) ds = \int_{t_0}^{t_1} f(x(t),y(t)) \sqrt{x'(t)^2 + y'(t)^2} dt.$$

Otro tipo de integral de trayectoria, esta vez de un campo vectorial $\boldsymbol{F} : I\!R^3 \to I\!R^3$ continuo sobre una trayectoria diferenciable con continuidad es la **integral de línea** (o integral de circulación). El elemento diferencial considerado es ahora $d\boldsymbol{s}$ es decir la diferencial de longitud de arco. Con más precisión se tiene:

Definición 1.6.6 *Sea $\boldsymbol{F}$ un campo vectorial continuo sobre la trayectoria $\boldsymbol{c}$ de clase C^1 en $[t_0, t_1]$. Definimos la* **integral de línea** *de $\boldsymbol{F}$ a lo largo de $\boldsymbol{c}$ como*

$$\int_{\boldsymbol{c}} \boldsymbol{F} \cdot d\boldsymbol{s} \doteq \int_{t_0}^{t_1} \boldsymbol{F}(\boldsymbol{c}(t)) \cdot \boldsymbol{c}'(t) dt.$$

Para trayectorias que satisfagan $\boldsymbol{c}'(t) \neq 0$, hay otra fórmula útil para la integral de línea. Siendo $\boldsymbol{T}(t) = \boldsymbol{c}'(t)$ el vector tangente a la trayectoria ($\boldsymbol{t} = \boldsymbol{c}'(t)/||\boldsymbol{c}'(t)||$ es el vector tangente unitario), tenemos:

$$\int_{\boldsymbol{c}} \boldsymbol{F} \cdot d\boldsymbol{s} \doteq \int_{t_0}^{t_1} \boldsymbol{F}(\boldsymbol{c}(t)) \cdot \boldsymbol{c}'(t) dt = \int_{t_0}^{t_1} \left[\boldsymbol{F}(\boldsymbol{c}(t)) \cdot \boldsymbol{T}(t)\right] ||\boldsymbol{c}'(t)|| dt.$$

Esta fórmula nos dice que la integral de línea del campo vectorial es igual a la integral de trayectoria de su componente tangencial.

Para campos vectoriales gradientes (o conservativos o irrotacionales) existe una fórmula muy fácil de cálculo de sus integrales de línea.

Definición 1.6.7 *Sea $f : I\!R^3 \to I\!R$ un campo escalar de clase C^1 sobre la trayectoria $\boldsymbol{c}$ de clase C^1 en $[t_0, t_1]$. Entonces:*

$$\int_{\boldsymbol{c}} \nabla f \cdot d\boldsymbol{s} = f(\boldsymbol{c}(t_1)) - f(\boldsymbol{c}(t_0)).$$

Volviendo al concepto de longitud de arco notamos que se puede extender a trayectorias en el espacio n-dimensional. Véase a este respecto la sección 4.2, pag 260, del libro de Marsden y Tromba.

Construimos ahora una **reparametrización** de $\boldsymbol{c}$ mediante la longitud de arco. Para ello caracterizamos primero el concepto de reparametrización. Sea $\boldsymbol{c}(t)$ una trayectoria dada para $t \in [a,b]$. Sea $s = \alpha(t)$ una nueva variable donde $\alpha(t)$ una función de clase C^1 estríctamente creciente definida en $[a,b]$. Las hipótesis efectuadas aseguran que para cada $s \in [\alpha(a), \alpha(b)]$ existe un t único tal que $\alpha(t) = s$. Definimos entonces la trayectoria $\boldsymbol{d} : [\alpha(a), \alpha(b)] \to I\!R^3$ mediante $\boldsymbol{d}(s) = \boldsymbol{c}(t)$. Las curvas imagenes de $\boldsymbol{c}$ y $\boldsymbol{d}$ son las mismas (es decir que $\boldsymbol{d}$ es una reparametrización de la curva imagen de la trayectoria $\boldsymbol{c}$) y $\boldsymbol{c}$ y $\boldsymbol{d}$ tienen la misma longitud de arco. Elegimos ahora la función $\alpha(t)$, definiendo

$$s = \alpha(t) = \int_a^t ||\boldsymbol{c}'(\tau)|| d\tau.$$

Si definimos $\boldsymbol{d}$ como se hizo antes, mediante $\boldsymbol{d}(s) = \boldsymbol{c}(t)$ entonces se verifica que:

$$||\boldsymbol{d}'(s)|| = 1.$$

Se dice entonces que $\boldsymbol{d}(s)$ es una **reparametrización** de $\boldsymbol{c}$ dada por la **longitud de arco**. Una trayectoria parametrizada mediante la longitud de arco tiene rapidez unitaria.

Pasamos a la definición de los conceptos geométricos de tangente unitaria, rapidez unitaria, vector normal principal y vector binormal.

Definición 1.6.8 *Dada una trayectoria* $\boldsymbol{c} : [a,b] \to I\!R^3$ *tal que* $\boldsymbol{c}'(t) \neq 0$ *para todo* t *el vector*

$$\boldsymbol{t}(t) \doteq \frac{\boldsymbol{c}'(t)}{||\boldsymbol{c}'(t)||}$$

es tangente a $\boldsymbol{c}$ *en* $\boldsymbol{c}(t)$ *y, como* $||\boldsymbol{t}|| = 1$, *el vector* $\boldsymbol{t}$ *se llama* **tangente unitaria** *a* $\boldsymbol{c}$.

Dada una trayectoria parametrizada mediante la longitud de arco, digamos por $\boldsymbol{c}(s)$, la **curvatura** en un punto $\boldsymbol{c}(s)$ sobre una trayectoria se define por:

$$k = ||\boldsymbol{t}'(s)|| = ||\boldsymbol{c}''(s)||.$$

Pasamos ahora a la definición de vector normal principal y vector binormal.

Definición 1.6.9 *Dada una trayectoria* $\boldsymbol{c}(t)$*, si* $\boldsymbol{t}'(t) \neq \boldsymbol{0}$ *se tiene que el vector:*

$$\boldsymbol{n}(t) \doteq \frac{\boldsymbol{t}'(t)}{||\boldsymbol{t}'(t)||},$$

es normal a $\boldsymbol{t}(t)$ *y es unitario. El vector* $\boldsymbol{n}$ *se llama vector normal principal. Un tercer vector unitario que es perpendicular tanto a* $\boldsymbol{t}$ *como a* $\boldsymbol{n}$ *se define por* $\boldsymbol{b} = \boldsymbol{t} \wedge \boldsymbol{n}$ *y se llama* **vector binormal**. *Los vectores* $\boldsymbol{t}$*,* $\boldsymbol{n}$ *y* $\boldsymbol{b}$ *forman un sistema de vectores ortogonales entre sí que siguen la regla de la* mano derecha *y que se va moviendo a lo largo de la trayectoria.*

Una aplicación muy importante del concepto de trayectoria consiste en la caracterización de las **líneas de flujo** de un campo vectorial. En mecánica de fluidos el campo vectorial a considerar es el campo de velocidades de un fluido. Un campo de velocidades se dice **estacionario** en una región del espacio (o del plano para campos bidimensionales) si la velocidad del fluido que pasa por los puntos de la región considerada no cambia con el tiempo (nótese que esto no quiere decir que el fluido no se está moviendo). Por ejemplo se dice que el flujo de agua por una tubería es estacionario si en cada punto de la tubería la velocidad del fluido que pasa por ese punto no cambia con el tiempo.

Definición 1.6.10 *Si* $\boldsymbol{F}$ *es un campo vectorial, una* **línea de flujo** *para* $\boldsymbol{F}$ *es una trayectoria* $\boldsymbol{c}(t)$ *tal que verifica el sistema de EDO:*

$$\boldsymbol{c}'(t) = \boldsymbol{F}(\boldsymbol{c}(t)).$$

Esto es, el campo de velocidad de la trayectoria viene generado *(o producido) por el campo vectorial* $\boldsymbol{F}$*.*

En el contexto del flujo de agua por una tubería una línea de flujo se puede asemejar a la trayectoria seguida por una partícula *pequeña* (con densidad igual a la del fluido y rozamiento con él nulo) suspendida en el fluido. Por ello, las líneas de flujo se llaman, apropiadamente, **líneas de corriente** o **curvas integrales**. El vector velocidad $\boldsymbol{v}$ de un fluido es tangente a una línea de flujo y la expresión de esta propiedad es $\boldsymbol{v} \wedge \boldsymbol{t} = \boldsymbol{0}$, que es la ecuación vectorial del sistema de EDO que determina las líneas de flujo del campo dado por $\boldsymbol{c}'(t) = \boldsymbol{F}(\boldsymbol{c}(t))$. Si

$$\boldsymbol{c}(t) = (x(t), y(t), z(t)) = x(t)\boldsymbol{i} + y(t)\boldsymbol{j} + z(t)\boldsymbol{k}, \qquad \boldsymbol{F} = \boldsymbol{v} = P\boldsymbol{i} + Q\boldsymbol{j} + R\boldsymbol{k},$$

siendo

$$P = P(x, y, z), \qquad Q = Q(x, y, z), \qquad R = R(x, y, z),$$

se tiene que el sistema de EDO es:

$$\begin{cases} x'(t) &= P(x(t), y(t), z(t)), \\ y'(t) &= Q(x(t), y(t), z(t)), \\ z'(t) &= Z(x(t), y(t), z(t)). \end{cases}$$

Ejemplo 1.6.1 *Probar que la trayectoria* $\boldsymbol{c}(t) = (\cos t, \sin t)$ *es una línea de flujo para el campo* $\boldsymbol{F}(x, y) = (-y, x)$. *Determinar las otras líneas de flujo.*

En primer lugar, interpretamos el campo vectorial como un campo de velocidades de un fluido:

$$\boldsymbol{F}(x, y) = \boldsymbol{v}(x, y) = (v_x(x, y), v_y(x, y)) = (-y, x) = -y\boldsymbol{i} + x\boldsymbol{j}.$$

La imagen de la trayectoria dada es un círculo. Para que sea una línea de flujo debe verificar las ecuaciones $\boldsymbol{c}'(t) = \boldsymbol{F}(\boldsymbol{c}(t))$, lo que es cierto pues:

$$\boldsymbol{c}'(t) = -\sin(t)\boldsymbol{i} + \cos(t)\boldsymbol{j} = F(\cos(t), \sin(t)) = -\sin(t)\boldsymbol{i} + \cos(t)\boldsymbol{j},$$

así que tenemos una línea de flujo. Las otras líneas de flujo también son círculos. En efecto, puesto que $P(x, y) = -y$, $Q(x, y) = x$ y no hay dependencia en la variable z (pues se trata de un flujo plano) se tiene que resolver:

$$\begin{cases} x'(t) &= -y(t), \\ y'(t) &= x(t), \end{cases}$$

sujeto a las condiciones iniciales $x(t_0) = x_0$, $y(t_0) = y_0$.

Utilizando las técnicas expuestas en el tema 14 de la asignatura de Elementos de Matemáticas y recordando la operación de exponenciación de matrices se deduce que las soluciones del sistema son:

$$\begin{aligned} x(t) &= R\cos(t - t_0), \\ y(t) &= R\sin(t - t_0), \end{aligned}$$

es decir, las trayectorias:

$$\boldsymbol{c}(t) = (R\cos(t - t_0), R\sin(t - t_0)),$$

cuyas imágenes son las curvas dadas por todos los círculos centrados en el origen, de radio R y que *arrancan* en los puntos del semieje positivo de las x dados por $(x_0, y_0) = (R, 0)$. Por cada punto del plano xy y en cada instante de tiempo pasa una única línea de flujo.

Flujo de un campo

Es conveniente utilizar una notación especial para la única solución que pasa por un punto dado en el tiempo $t = 0$. Es posible razonar en una genérica dimensión n.

Se fija una condición inicial $\boldsymbol{x}_0 = \boldsymbol{x}(0) \in I\!R^n$ y se sigue a lo largo de la línea de flujo durante un tiempo t hasta alcanzar la nueva posición $\phi(\boldsymbol{x}, t)$. Es decir, que $\phi(\boldsymbol{x}, t) \in I\!R^n$, se define como la posición del punto en la línea de flujo que pasa por $\boldsymbol{x}_0$ después de haber transcurrido un tiempo t. Matemáticamente esto se traduce diciendo que $\phi(\boldsymbol{x}, t)$ está definida como la solución del PVI asociado al sistema:

$$\left\{ \begin{array}{rcl} \dfrac{\partial}{\partial t}\phi(\boldsymbol{x}, t) & = & \boldsymbol{F}(\phi(\boldsymbol{x}, t)), \\ \phi(\boldsymbol{x}, 0) & = & \boldsymbol{x}_0. \end{array} \right.$$

La función ϕ, que se considera como función de las variables $(\boldsymbol{x}, t)$, $\phi : I\!R^n \times [0, \infty) \to I\!R^n$ se llama **flujo** de $\boldsymbol{F}$. Se demuestra también que ϕ es una función diferenciable.

El concepto de flujo nos permite introducir el concepto del operador diferencial **derivada material** de un campo escalar f respecto a un campo vectorial $\boldsymbol{F}$. En concreto, sea $f(\boldsymbol{x}, t)$ una función con valores reales, $f : I\!R^n \times [0, \infty)$ y sea $\boldsymbol{F} : I\!R^n \to I\!R^n$ un campo vectorial.

Definición 1.6.11 *Se define la* **derivada material** *de un campo escalar f con respecto a un campo vectorial $\boldsymbol{F}$ como:*

$$\frac{Df}{Dt} \doteq \frac{\partial f}{\partial t} + \nabla f(\boldsymbol{x}) \cdot \boldsymbol{F}.$$

La definición anterior se interpreta observando que la derivada material coincide con la derivada con respecto de t de f transportada por el flujo de $\boldsymbol{F}$, es decir, la derivada con respecto de t de $f(\phi(\boldsymbol{x}, t), t)$.

Flujos potenciales

Recordaremos en esta sección las definiciones de la función potencial para flujos irrotacionales (potenciales). Este tipo de flujo es propio de los fluidos no viscosos (que más adelante caracterizaremos como ideales o perfectos). En efecto, en un fluido sin viscosidad no pueden existir tensiones (fuerzas) tangenciales o rasantes (de cizalla) sobre sus elementos, y las fuerzas de presión o campo que actúen sobre ellos, podrán provocar deformaciones pero nunca

rotaciones de los mismos. Finalmente observamos que en los supuestos de flujo plano y estacionario las líneas de corriente de un flujo potencial siempre empiezan (y acaban) en el infinito.

Sea $\boldsymbol{v} = (v_x, v_y, v_z)$ un campo vectorial de velocidades estacionario[44] tal que $\boldsymbol{w} = \text{rot}\boldsymbol{v} = 0$ (es decir con vorticidad $\boldsymbol{w}$ nula). Entonces existe una función potencial $\phi(x, y, z)$ tal que $\boldsymbol{v} = \nabla\phi$. Nos limitaremos aquí al caso bidimensional. Dependiendo del sistema de coordenadas se tiene:

Sistema de Coord.	**Función potencial ϕ**
Rectangulares	$\boldsymbol{v} = (v_x, v_y) = \left(\dfrac{\partial\phi}{\partial x}, \dfrac{\partial\phi}{\partial y}\right)$
Polares	$\boldsymbol{v} = (v_r, v_\theta) = \left(\dfrac{\partial\phi}{\partial r}, \dfrac{1}{r}\dfrac{\partial\phi}{\partial\theta}\right)$

Flujos incompresibles

Recordaremos brevemente las definiciones de la función de corriente para flujos incompresibles en los sistemas de coordenadas rectangulares y cilíndricas.

Obsérvese que los fluidos incompresibles se caracterizan por la condición de divergencia nula. Si la densidad del fluido es constante entonces el fluido es, por supuesto, incompresible. Por otra parte existen fluidos incompresibles cuya densidad no es constante. En efecto, en condiciones isotermas, la compresibilidad de un fluido, c, se define mediante:

$$c = \frac{1}{\rho}\frac{d\rho}{dp},$$

siendo ρ la densidad y p la presión del fluido. Si el fluido es incompresible:

$$\frac{d\rho}{dp} = 0.$$

Esto se cumple obviamente si ρ (la densidad) es constante en todo el fluido. También se puede cumplir si ρ no es constante *pero es independiente de la presión*, $\rho = \rho(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, como por ejemplo en el caso de flujos multifásicos

[44] Esta hipótesis no es absolutamente necesaria. Sólo simplifica el tratamiento matemático.

incompresibles (agua-petróleo, por ejemplo). En este sentido la incompresibilidad se puede traducir como *constancia de la densidad respecto a la presión* (aunque pueda depender de otras variables que no sean a su vez función de la posición.

El flujo de un líquido siempre es incompresible (digamos que se puede despreciar el efecto de la compresibilidad) pero el flujo de un gas podrá también ser considerado como tal (es decir puede fluir sin importantes variaciones de su densidad) si el flujo es subsónico[45] pero no deja de ser más que una idealización.

Sea $\boldsymbol{v} = (v_x, v_y, v_z)$ un campo vectorial de velocidades estacionario tal que $\text{div}\boldsymbol{v} = 0$ (es decir con divergencia nula). Entonces existe una función de corriente $\psi(x, y, z)$ tal que $\boldsymbol{v} \cdot \nabla\psi = 0$. Nos limitaremos aquí al caso bidimensional luego tendremos, respectivamente, cuatro casos:

I Coordenadas rectangulares con $v_z = 0$ y ninguna dependencia en la variable z.

II Coordenadas cilíndricas con $v_z = 0$ y ninguna dependencia en la variable z (son equivalentes a las coordenadas polares).

III Coordenadas cilíndricas con $v_\theta = 0$ y ninguna dependencia en la variable θ.

IV Coordenadas Esféricas con $v_\phi = 0$ y ninguna dependencia en la variable ϕ.

Dependiendo del sistema de coordenadas se tiene:

[45] La clasificación del flujo de un gas en subsónico (flujo incompresible) y transónico, supersónico o hipersónico (flujos compresibles) se realiza mediante el cálculo del número de Mach que expresa la relación entre la velocidad media del gas y la del sonido en su seno.

Sistema de Coord.	Función de corriente
I) Rectangulares con $v_z = 0$	$\boldsymbol{v} = (v_x, v_y) = \left(\dfrac{\partial \psi}{\partial y}, -\dfrac{\partial \psi}{\partial x}\right)$
II) Cilíndricas con $v_z = 0$	$\boldsymbol{v} = (v_r, v_\theta) = \left(\dfrac{1}{r}\dfrac{\partial \psi}{\partial \theta}, -\dfrac{\partial \psi}{\partial r},\right)$
III) Cilíndricas con $v_\theta = 0$	$\boldsymbol{v} = (v_r, v_z) = \left(\dfrac{1}{r}\dfrac{\partial \psi}{\partial z}, -\dfrac{1}{r}\dfrac{\partial \psi}{\partial r},\right)$
IV) Esféricas con $v_\phi = 0$	$\boldsymbol{v} = (v_r, v_\theta) = \left(\dfrac{1}{r^2 \sin\theta}\dfrac{\partial \psi}{\partial \theta}, -\dfrac{1}{r \sin\theta}\dfrac{\partial \psi}{\partial r}\right)$

Fluidos perfectos

Deduciremos ahora la **ecuación de Euler** para un **fluido perfecto** (o ideal) entendiendo por ello un fluido que fluye sin viscosidad[46] (el análogo del rozamiento entre los sólidos). El flujo *real* es el flujo de los fluidos reales con viscosidad apreciable, siempre rotacional.

Consideremos un fluido no viscoso que se mueve en un campo de velocidad $\boldsymbol{v}$. Cuando decimos que el fluido es perfecto queremos decir que en cualquier parte del fluido Ω, actúan fuerzas de presión sobre la frontera $\partial\Omega$ a lo largo de su normal. Suponemos que la fuerza por unidad de área que actúa sobre $\partial\Omega$ es $-p\boldsymbol{n}$, siendo $p(x, y, z, t)$ la función de **presión** (es un campo escalar). Así la fuerza total de presión que actúa sobre $\partial\Omega$ es:

$$\boldsymbol{F}_{\partial\Omega} = \text{Fuerza } = -\iint_{\partial\Omega} p\boldsymbol{n} dS.$$

Ésta es una cantidad vectorial. Si aplicamos el teorema de la divergencia a cada componente del campo vectorial de fuerzas se tiene que:

$$\boldsymbol{F}_{\partial\Omega} = -\iiint_{\Omega} \nabla p dx dy dz.$$

[46] En muchos textos (véase por ejemplo el vol. 2 sobre fenómenos de Transporte, capítulo 1, sección 7.4, pag 50 de los libros de Costa Novella) se identifican los fluidos perfectos como los fluidos cuyo flujo es no viscoso, incompresible e irrotacional. Otros textos añaden la hipótesis de flujo estacionario.

Sea $\phi_t(\boldsymbol{x}) = \phi(\boldsymbol{x}, t)$ el flujo del campo $\boldsymbol{v}$ y sea $\Omega_t = \phi_t(\Omega)$ una región en movimiento. Aplicando la segunda ley de Newton a Ω_t, la razón de cambio (la variación) de la cantidad de movimiento de Ω_t es igual a la fuerza que actúa sobre ella:

$$\frac{d}{dt}\iiint_{\Omega_t} \rho \boldsymbol{v} dxdydz = \boldsymbol{F}_{\partial\Omega_t} = -\int\int\int_{\Omega_t} \nabla p \, dxdydz.$$

Utilizando la **ecuación del transporte** que afirma que en cualquier región Ω_t que se mueve con el fluido se tiene:

$$\frac{d}{dt}\iiint_{\Omega_t} f(x,y,z,t)dxdydz = \iiint_{\Omega_t} \left(\frac{Df}{Dt} + f\text{div}\boldsymbol{F}\right) dxdydz,$$

(siendo Df/Dt el operador de derivación material) y aplicando el **teorema del transporte** se obtiene:

$$\iiint_{\Omega_t} \left[\frac{\partial}{\partial t}(\rho \boldsymbol{v}) + \boldsymbol{v} \cdot \nabla(\rho \boldsymbol{v}) + \rho \boldsymbol{v} \text{div}\boldsymbol{v}\right] dxdydz = -\iiint_{\Omega_t} \nabla p \, dxdydz.$$

Puesto que Ω_t es arbitraria (es decir que lo anterior es cierto para cualquier región que se mueve con el fluido), esto es equivalente a

$$\frac{\partial}{\partial t}(\rho \boldsymbol{v}) + \boldsymbol{v} \cdot \nabla(\rho \boldsymbol{v}) + \rho \boldsymbol{v} \text{div}\boldsymbol{v} = -\nabla p.$$

Utilizando la ecuación de continuidad $\rho_t + \text{div}\boldsymbol{J} = 0$ siendo $\boldsymbol{J} = \rho \boldsymbol{v}$ escrita en la forma:

$$\frac{\partial \rho}{\partial t} + \boldsymbol{v} \cdot \nabla \rho + \rho \text{div}\boldsymbol{v} = 0,$$

se deduce:

$$\rho\left(\frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v}\right) = -\nabla p, \tag{1.76}$$

que es la ecuación de Euler para un fluido perfecto. Para fluidos compresibles, p (la presión) es una función dada (conocida) de ρ (por ejemplo, para muchos gases $p = A\rho^\gamma$ para constantes A y γ dadas). Si, por otra parte el fluido es incompresible, p se determina a partir de la condición $\text{div}\boldsymbol{v} = 0$. En este caso, la condición de divergencia nula complementada con la ecuación de Euler (1.76) gobiernan el movimiento del fluido.

1.6.2. Representaciones de curvas y superficies

Existen básicamente tres métodos de representación de superficies que permiten expresar analíticamente tal lugar geométrico[47] . Uno es la *representación implícita* en el que se considera una superficie como un conjunto de puntos (x, y, z) que satisfacen una ecuación de la forma:

$$F(x, y, z) = 0.$$

Por ejemplo, la esfera de radio 1 y centro en el origen tiene la representación implícita:

$$x^2 + y^2 + z^2 - 1 = 0.$$

Algunas veces (véase la sección 9.13 del guión del primer curso dedicada a las funciones definidas implícitamente) podemos despejar en la ecuación una de las coordenadas en función de las otras dos, por ejemplo, z en función de x e y. Cuando esto es posible obtenemos una *representación explícita* dada por una o varias ecuaciones de la forma:

$$z = f(x, y).$$

Por ejemplo, al despejar z en la representación implícita anterior de la esfera se tiene:

$$z = \sqrt{1 - x^2 - y^2}, \qquad z = -\sqrt{1 - x^2 - y^2}.$$

La primera es la representación explícita de la semiesfera superior y la segunda de la inferior.

Existe un tercer método de representación de superficies que es más útil en el estudio de las mismas; es la *representación paramétrica* o *vectorial* por medio de tres ecuaciones que expresan x, y, z en función de dos parámetros u y v:

$$x = X(u, v), \qquad y = Y(u, v), \qquad z = Z(u, v). \tag{1.77}$$

Aquí el punto (u, v) puede variar en un conjunto conexo bidimensional Ω en el plano uv, y los puntos (x, y, z) correspondientes constituyen una porción de superficie en el espacio $\mathbb{R}^3$. Este método es análogo al de la representación de una curva en $\mathbb{R}^3$ mediante tres ecuaciones con un solo parámetro. La presencia de los dos parámetros en (1.77) permite transmitir dos grados de libertad al punto (x, y, z). Si introducimos el radio vector que une el origen con un

[47] Una superficie, a groso modo, puede identificarse con el lugar geométrico de un punto que se mueve en el espacio con dos grados de libertad. Véase el libro de Apostol, Vol 2, capítulo 12.

punto genérico (x, y, z) de la superficie, podemos combinar las tres ecuaciones paramétricas (1.77) en una ecuación vectorial de la forma:

$$\boldsymbol{r}(u, v) = X(u, v)\boldsymbol{i} + Y(u, v)\boldsymbol{j} + Z(u, v)\boldsymbol{k}, \tag{1.78}$$

donde $(u, v) \in \Omega$. Ésta es la llamada *ecuación vectorial* de la superficie. Existen muchas representaciones paramétricas de la misma superficie. Una de ellas puede obtenerse a partir de la forma explícita $z = f(x, y)$, tomando:

$$X(u, v) = u, \qquad Y(u, v) = v, \qquad Z(u, v) = f(u, v).$$

Ejemplo 1.6.2 *La representación paramétrica de una esfera de radio a y centro en el origen es:*

$$x = a\cos u\cos v, \qquad y = a\sin u\cos v, \qquad z = a\sin v.$$

Si elevamos al cuadrado las tres ecuaciones y sumamos resulta

$$x^2 + y^2 + z^2 = a^2$$

y vemos que todo punto (x, y, z) que satisface las ecuaciones paramétricas está en la esfera.

Considérese ahora una superficie representada por la ecuación vectorial (1.78). Si X, Y y Z son derivables en Ω podemos considerar los dos vectores:

$$\frac{\partial \boldsymbol{r}}{\partial u} = \frac{\partial X}{\partial u}\boldsymbol{i} + \frac{\partial Y}{\partial u}\boldsymbol{j} + \frac{\partial Z}{\partial u}\boldsymbol{k}, \qquad \frac{\partial \boldsymbol{r}}{\partial v} = \frac{\partial X}{\partial v}\boldsymbol{i} + \frac{\partial Y}{\partial v}\boldsymbol{j} + \frac{\partial Z}{\partial v}\boldsymbol{k}.$$

Ambos vectores son tangentes a la superficie. Más concretamente, el vector $\partial \boldsymbol{r}/\partial u$ es tangente a una u-curva en la superficie (con la expresión u-curva se entiende una curva en la superficie que se obtiene manteniendo v constante) y el vector $\partial \boldsymbol{r}/\partial v$ es tangente a una v-curva en la superficie (una curva en la superficie que se obtiene manteniendo u constante). En general estos vectores no son ortogonales entre ellos y tampoco son unitarios. Sin embargo su producto vectorial es ortogonal a ambos luego es normal a la superficie. El producto vectorial:

$$\boldsymbol{N} \doteq \frac{\partial \boldsymbol{r}}{\partial u} \wedge \frac{\partial \boldsymbol{r}}{\partial v} = \boldsymbol{r}_u \wedge \boldsymbol{r}_v,$$

se denomina **producto vectorial fundamental** de la representación $\boldsymbol{r}$. Si (u, v) es un punto en Ω en el cual $\partial \boldsymbol{r}/\partial u$ y $\partial \boldsymbol{r}/\partial v$ son continuas y el producto vectorial fundamental no es nulo, el punto imagen $\boldsymbol{r}(u, v)$ se llama *punto regular* de $\boldsymbol{r}$. Los puntos donde al menos una de estas condiciones no se verifica se llaman puntos singulares. Una superfice se llama **regular** si todos sus puntos son regulares. En cada punto regular los vectores $\partial \boldsymbol{r}/\partial u$ y $\partial \boldsymbol{r}/\partial v$ determinan

un plano que tiene el vector $\partial \boldsymbol{r}/\partial u \wedge \partial \boldsymbol{r}/\partial v$ como normal. Por esta razón el plano determinado por $\partial \boldsymbol{r}/\partial u$ y $\partial \boldsymbol{r}/\partial v$ se llama **plano tangente**. En cada punto regular de una superficie se designa como $\boldsymbol{n}$ al vector unitario normal que tenga el mismo sentido que el producto vectorial fundamental:

$$\boldsymbol{n} = \frac{\dfrac{\partial \boldsymbol{r}}{\partial u} \wedge \dfrac{\partial \boldsymbol{r}}{\partial v}}{\left\| \dfrac{\partial \boldsymbol{r}}{\partial u} \wedge \dfrac{\partial \boldsymbol{r}}{\partial v} \right\|} = \frac{\boldsymbol{r}_u \wedge \boldsymbol{r}_v}{\|\boldsymbol{r}_u \wedge \boldsymbol{r}_v\|}.$$

La continuidad de las derivadas parciales implica la continuidad de su producto vectorial y esto, a su vez, significa que el plano tangente se mueve con continuidad en una superfice regular.

Observación 1.6.1 *Una manera alternativa de calcular un vector normal unitario consiste en representar la superficie como $S(x, y, z) = 0$ y utilizar la fórmula:*

$$\boldsymbol{n} = \frac{\nabla S}{\|\nabla S\|}.$$

Aplicando lo anterior veremos ahora cómo se determina el vector normal a una superficie con representación explícita $z = f(x, y)$.

Para ello podemos usar x e y como parámetros lo que nos proporciona la ecuación vectorial:

$$\boldsymbol{r}(x, y) = x\boldsymbol{i} + y\boldsymbol{j} + f(x, y)\boldsymbol{k}.$$

Para calcular el producto vectorial fundamental observemos que, si f es diferenciable,

$$\frac{\partial \boldsymbol{r}}{\partial x} = \boldsymbol{i} + \frac{\partial f}{\partial x}\boldsymbol{k}, \qquad \frac{\partial \boldsymbol{r}}{\partial y} = \boldsymbol{j} + \frac{\partial f}{\partial y}\boldsymbol{k}.$$

Esto nos conduce a:

$$\frac{\partial \boldsymbol{r}}{\partial u} \wedge \frac{\partial \boldsymbol{r}}{\partial v} = -\frac{\partial f}{\partial x}\boldsymbol{i} - \frac{\partial f}{\partial y}\boldsymbol{j} + \boldsymbol{k}.$$

Nótese que el producto vectorial fundamental nunca es nulo para este tipo de representación puesto que la componente z del producto vale 1. Los únicos puntos singulares que pueden presentarse son los puntos en los que al menos una de las derivadas parciales f_x o f_y no es continua.

Ejemplo 1.6.3 *Calcular el vector normal a la superficie dada por la ecuación:*

$$z = f(x, y) = \sqrt{1 - x^2 - y^2},$$

que representa un hemisferio de radio 1 y centro en el origen si $x^2 + y^2 \leq 1$.

En primer lugar parametrizamos la superficie usando x e y como parámetros. La ecuación vectorial es:

$$\boldsymbol{r}(x,y) = x\boldsymbol{i} + y\boldsymbol{j} + \left(\sqrt{1-x^2-y^2}\right)\boldsymbol{k}.$$

Nótese que $\boldsymbol{r}$ aplica el disco unidad $D = \{(x,y) \,:\, x^2+y^2 \leq 1\}$ sobre el hemisferio y dicha aplicación es uno a uno[48]. Las derivadas parciales $\partial\boldsymbol{r}/\partial x$, $\partial\boldsymbol{r}/\partial y$ existen y son continuas en el interior del disco pero no existen en su frontera. Por consiguiente todo punto del ecuador es un punto singular de esta representación. En el interior del disco (donde f es diferenciable) se tiene:

$$\frac{\partial \boldsymbol{r}}{\partial x} = \boldsymbol{i} + \frac{\partial f}{\partial x}\boldsymbol{k} = \boldsymbol{i} - \frac{x}{\sqrt{1-x^2-y^2}}\boldsymbol{k},$$

$$\frac{\partial \boldsymbol{r}}{\partial y} = \boldsymbol{j} + \frac{\partial f}{\partial y}\boldsymbol{k} = \boldsymbol{j} - \frac{y}{\sqrt{1-x^2-y^2}}\boldsymbol{k}.$$

El producto vectorial fundamental es:

$$\frac{\partial \boldsymbol{r}}{\partial x} \wedge \frac{\partial \boldsymbol{r}}{\partial y} = -\frac{\partial f}{\partial x}\boldsymbol{i} - \frac{\partial f}{\partial y}\boldsymbol{j} + \boldsymbol{k} = \frac{x}{\sqrt{1-x^2-y^2}}\boldsymbol{i} + \frac{y}{\sqrt{1-x^2-y^2}}\boldsymbol{j} + \boldsymbol{k},$$

luego su norma es

$$\left\| \frac{\partial \boldsymbol{r}}{\partial u} \wedge \frac{\partial \boldsymbol{r}}{\partial v} \right\| = \sqrt{\frac{x^2}{1-x^2-y^2} + \frac{y^2}{1-x^2-y^2} + 1} = \sqrt{\frac{1}{1-x^2-y^2}} = \frac{1}{\sqrt{1-x^2-y^2}}.$$

El vector unitario normal a la superficie será por tanto,

$$\boldsymbol{n} = \frac{\dfrac{\partial \boldsymbol{r}}{\partial x} \wedge \dfrac{\partial \boldsymbol{r}}{\partial x}}{\left\| \dfrac{\partial \boldsymbol{r}}{\partial x} \wedge \dfrac{\partial \boldsymbol{r}}{\partial x} \right\|} = \left(x, y, \sqrt{1-x^2-y^2}\right).$$

Nótese que es unitario pues $\|\boldsymbol{n}\| = x^2+y^2+1-x^2-y^2 = 1$.

A continuación se presentan unas tablas para los distintos sistemas de coordenadas más utilizados, la expresión de los operadores vectoriales de derivación en cada uno de los sistemas y la forma final de las ecuaciones de conservación utilizadas en numerosos problemas.

[48] Recuérdese que esto quiere decir que puntos distintos en el disco unidad tienen imagen distinta en el hemisferio.

1.6.3. Tablas de operadores diferenciales

Es habitual trabajar en alguno de los tres sistemas de coordenadas siguientes: coordenadas cartesianas (también llamadas rectangulares), coordenadas cilíndricas y coordenadas esféricas.

Las coordenadas polares, que se usan en $\mathbb{R}^2$, se pueden obtener fácilmente al proyectar en el plano $z \equiv 0$ las coordenadas cilíndricas y no serán descritas como caso autónomo sino que se considerarán como un caso particular de las cilíndricas (que representan por tanto su generalización a $\mathbb{R}^3$).

La notación utilizada será: (x, y, z) para la localización de un punto en coordenadas cartesianas, (r, θ, z) para la localización de un punto en coordenadas cilíndricas y (ρ, θ, ϕ) para la localización de un punto en coordenadas esféricas.

Coordenadas	**Fórmulas de conversión**
Cilíndricas $\to$ Cartesianas	$x = r\cos\theta, \qquad y = r\sin\theta, \qquad z = z$
Cartesianas $\to$ Cilíndricas	$r = \sqrt{x^2 + y^2}, \qquad \tan\theta = (y/x), \qquad z = z$
Esféricas $\to$ Cartesianas	$x = \rho\sin\phi\cos\theta, \qquad y = \rho\sin\phi\sin\theta, \qquad z = \rho\cos\phi$
Cartesianas $\to$ Esféricas	$\rho = \sqrt{x^2 + y^2 + z^2}, \qquad \tan\theta = (y/x),$ $\phi = \arccos\left(\dfrac{z}{\sqrt{x^2 + y^2 + z^2}}\right)$
Esféricas $\to$ Cilíndricas	$r = \rho\sin\phi, \qquad \theta = \theta, \qquad z = \rho\cos\phi$
Cilíndricas $\to$ Esféricas	$\rho = \sqrt{r^2 + z^2}, \qquad \theta = \theta, \qquad \phi = \arccos\left(\dfrac{z}{\sqrt{r^2 + z^2}}\right)$

Operador Divergencia

Dependiendo del sistema de coordenadas utilizado se expresa de distinta manera. Si el campo vectorial $\boldsymbol{v}$ viene dado en coordenadas rectangulares en-

tonces $\boldsymbol{v} = (v_x, v_y, v_z)$. Si viene dado en coordenadas cilíndricas entonces $\boldsymbol{v} = (v_r, v_\theta, v_z)$ y si viene dado en coordenadas esféricas entonces $\boldsymbol{v} = (v_r, v_\theta, v_\phi)$.

Coordenadas	**Operador de Divergencia**
Rectangulares	$\nabla \cdot \boldsymbol{v} = \dfrac{\partial v_x}{\partial x} + \dfrac{\partial v_y}{\partial y} + \dfrac{\partial v_z}{\partial z}$
Cilíndricas	$\nabla \cdot \boldsymbol{v} = \dfrac{1}{r}\dfrac{\partial}{\partial r}(r v_r) + \dfrac{1}{r}\dfrac{\partial v_\theta}{\partial \theta} + \dfrac{\partial v_z}{\partial z} = \dfrac{\partial v_r}{\partial r} + \dfrac{1}{r} v_r + \dfrac{1}{r}\dfrac{\partial v_\theta}{\partial \theta} + \dfrac{\partial v_z}{\partial z}$
Esféricas	$\nabla \cdot \boldsymbol{v} = \dfrac{1}{r^2}\dfrac{\partial}{\partial r}(r^2 v_r) + \dfrac{1}{r \sin\theta}\dfrac{\partial}{\partial \theta}(v_\theta \sin\theta) + \dfrac{1}{r \sin\theta}\dfrac{\partial v_\phi}{\partial \phi}$

Operador Derivada Material

Las leyes generales de conservación en mecánica de fluidos se pueden escribir más fácilmente en términos del operador diferencial:

$$\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + \boldsymbol{v} \cdot \nabla, \tag{1.79}$$

que se conoce con el nombre de **derivada material**. La derivada material tiene un significado físico concreto: es la tasa de cambio de un campo escalar vista por un observador que se mueve con el fluido. Por ejemplo, aparece en la ecuación de conservación de la energía aplicado al campo escalar de temperaturas. También se aplica a un campo vectorial (por ejemplo, el campo de velocidades) actuando componente a componente.

Coordenadas	**Operador de Derivada Material aplicado a T**
Rectangulares $T(x,y,z,t)$	$\dfrac{DT}{Dt} = \dfrac{\partial T}{\partial t} + v_x \dfrac{\partial T}{\partial x} + v_y \dfrac{\partial T}{\partial y} + v_z \dfrac{\partial T}{\partial z}$
Cilíndricas $T(r,\theta,z,t)$	$\dfrac{DT}{Dt} = \dfrac{\partial T}{\partial t} + v_r \dfrac{\partial T}{\partial r} + \dfrac{v_\theta}{r} \dfrac{\partial T}{\partial \theta} + v_z \dfrac{\partial T}{\partial z}$
Esféricas $T(r,\theta,\phi,t)$	$\dfrac{DT}{Dt} = \dfrac{\partial T}{\partial t} + v_r \dfrac{\partial T}{\partial r} + \dfrac{v_\theta}{r} \dfrac{\partial T}{\partial \theta} + \dfrac{v_\phi}{r \sin\theta} \dfrac{\partial T}{\partial \phi}$

Consideremos el mismo operador pero aplicado al campo vectorial de velocidades:

$$\frac{D\boldsymbol{v}}{Dt} \equiv \frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v}. \tag{1.80}$$

Se tiene entonces:

Coord. Rectangulares	**Operador de Derivada Material aplicado a $\boldsymbol{v}$**
$v_x(x,y,z)$	$\dfrac{Dv_x}{Dt} = \dfrac{\partial v_x}{\partial t} + v_x \dfrac{\partial v_x}{\partial x} + v_y \dfrac{\partial v_x}{\partial y} + v_z \dfrac{\partial v_x}{\partial z}$
$v_y(x,y,z)$	$\dfrac{Dv_y}{Dt} = \dfrac{\partial v_y}{\partial t} + v_x \dfrac{\partial v_y}{\partial x} + v_y \dfrac{\partial v_y}{\partial y} + v_z \dfrac{\partial v_y}{\partial z}$
$v_z(x,y,z)$	$\dfrac{Dv_z}{Dt} = \dfrac{\partial v_z}{\partial t} + v_x \dfrac{\partial v_z}{\partial x} + v_y \dfrac{\partial v_z}{\partial y} + v_z \dfrac{\partial v_z}{\partial z}$

Operador Laplaciano

El operador laplaciano puede aparecer[49] en las ecuaciones de conservación de la cantidad de movimiento y de la energía. En el primer caso (es decir considerando la ecuación de conservación de la cantidad de movimiento) es sabido que para un fluido newtoniano con densidad y viscosidad constantes se tiene:

$$\nabla \cdot [\tau] = \mu \nabla^2 \boldsymbol{v},$$

siendo $[\tau]$ el tensor de tensiones (que denotamos como una matriz cuadrada)[50] (o fuerzas) viscosas, siendo μ la viscosidad del fluido. Típicamente se denota $\nabla^2 \boldsymbol{v} = \Delta \boldsymbol{v}$. En el segundo caso, al aplicar la ley de conducción del calor de Fourier se tiene:

$$-\nabla \cdot \boldsymbol{q} = -\nabla \cdot (-k \nabla T) = k \nabla^2 T,$$

siendo $\boldsymbol{q}$ el vector de flujo (de calor) y k la conductibilidad calorífica. Nuevamente se suele denotar $\nabla^2 T = \Delta T$. En general dado un campo escalar $u(x, y, z, t)$ (que puede ser T o una componente del campo de velocidades) se tiene:

[49] Por ejemplo no aparece al trabajar con fluidos perfectos que satisfacen la ecuación de Euler.

[50] No consideraremos en este curso la definición de tensor. Sólo observamos que un escalar se puede considerar un tensor de orden cero, un vector se considera un tensor de orden uno y que que un tensor de segundo orden en el espacio tridimensional es un objeto matemático que se puede representar como un conjunto ordenado de nueve números. Cada uno de ello se asocia con dos direcciones. Habitualmente se utilizan las matrices para denotar los tensores.

Coordenadas	Operador Laplaciano
Rectangulares	$\Delta u = \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} + \dfrac{\partial^2 u}{\partial z^2}$
Cilíndricas	$\Delta u = \dfrac{1}{r}\dfrac{\partial}{\partial r}\left(r\dfrac{\partial u}{\partial r}\right) + \dfrac{1}{r^2}\dfrac{\partial^2 u}{\partial \theta^2} + \dfrac{\partial^2 u}{\partial z^2} = \dfrac{\partial^2 u}{\partial r^2} + \dfrac{1}{r}\dfrac{\partial u}{\partial r} + \dfrac{1}{r^2}\dfrac{\partial^2 u}{\partial \theta^2} + \dfrac{\partial^2 u}{\partial z^2}$
Esféricas	$\Delta u = \dfrac{1}{r^2}\dfrac{\partial u}{\partial r}\left(r^2\dfrac{\partial u}{\partial r}\right) + \dfrac{1}{r^2\sin\phi}\dfrac{\partial}{\partial \phi}\left(\sin\phi\dfrac{\partial u}{\partial \phi}\right) + \dfrac{1}{r^2\sin^2\phi}\left(\dfrac{\partial^2 u}{\partial \theta^2}\right)$
Radiales	$\Delta u = \dfrac{1}{r^{N-1}}\dfrac{\partial}{\partial r}\left(r^{N-1}\dfrac{\partial u}{\partial r}\right) = \dfrac{\partial^2 u}{\partial r^2} + \dfrac{(N-1)}{r}\dfrac{\partial u}{\partial r}, \quad N = 1, 2, 3$

1.6.4. Tablas de Ecuaciones

Recordaremos aquí las expresiones en distintos sistemas de coordenadas de las ecuaciones básicas que aparecen en los problemas de transporte de la mecánica de fluidos.

Ecuación de continuidad

En general se corresponde a la ley de conservación de la materia total (véase por ejemplo la ecuación 3.75 del vol. 2 del libro de Costa Novella). En forma de balance local de masa:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0, \tag{1.81}$$

siendo ρ la densidad de masa por unidad de volumen y $\boldsymbol{v}$ el campo de velocidades[51]. Desarrollando las operaciones de derivación que aparecen en (1.81) se tiene:

$$\frac{\partial \rho}{\partial t} + \boldsymbol{v} \cdot \nabla \rho + \rho \nabla \cdot \boldsymbol{v} = 0. \tag{1.82}$$

[51] Recuérdese que en Mecánica de Medios Continuos se entiende por fluido un cuerpo cuyas moléculas tienen poca coherencia entre si, de modo que pueden deslizarse libremente una sobre otras (líquidos), o separase y desplazarse con completa independencia (gases), tomando siempre la forma del recipiente que lo contiene.

Esta ecuación (o su forma más compacta (1.81)) expresa la conservación de materia en fenómenos en que esta se transporta sólo convectivamente (por ejemplo en el movimiento de un fluido). Pero no es el caso más general pues no incluye otras formas de transporte. La identificación de la ecuación de conservación de la masa con el nombre de **ecuación de continuidad** es propio de la mecánica de fluidos[52] y es una consecuencia de aplicar la ley de conservación de la materia al fluido. Volviendo al ejemplo (1.1.2) (que describe el movimiento unidimensional de un gas) ponemos $\boldsymbol{v} = u(x)$, $\mathrm{div}\boldsymbol{v} = u_x$, $\nabla\rho = \rho_x$ y vemos que la ecuación de conservación de la masa propuesta en (1.1.2):

$$\rho\frac{\partial u}{\partial x} + u\frac{\partial \rho}{\partial x} + \frac{\partial \rho}{\partial t} = 0,$$

es del tipo (1.82) siendo $\boldsymbol{v}\cdot\nabla\rho = u\rho_x$ y $\rho(\nabla\cdot\boldsymbol{v}) = \rho u_x$.

Si la densidad del fluido es constante (fluidos incompresibles) la ecuación de continuidad se simplifica a la condición de divergencia nula:

$$\nabla\cdot\boldsymbol{v} = 0, \tag{1.83}$$

que es la expresión matemática de la conservación de la masa para materiales incompresibles (representados por campos cuya divergencia es nula).

Dependiendo del sistema de coordenadas utilizado se expresa de distinta manera.

Sistema de Coord.	**Ecuación de continuidad para fluidos incomp.**
Rectangulares	$\nabla\cdot\boldsymbol{v} = \dfrac{\partial v_x}{\partial x} + \dfrac{\partial v_y}{\partial y} + \dfrac{\partial v_z}{\partial z} = 0$
Cilíndricas	$\nabla\cdot\boldsymbol{v} = \dfrac{1}{r}\dfrac{\partial}{\partial r}(rv_r) + \dfrac{1}{r}\dfrac{\partial v_\theta}{\partial \theta} + \dfrac{\partial v_z}{\partial z} = 0$
Esféricas	$\nabla\cdot\boldsymbol{v} = \dfrac{1}{r^2}\dfrac{\partial}{\partial r}(r^2v_r) + \dfrac{1}{r\sin\theta}\dfrac{\partial}{\partial \theta}(v_\theta\sin\theta) + \dfrac{1}{r\sin\theta}\dfrac{\partial v_\phi}{\partial \phi} = 0$

[52] Nótese que la mecánica de fluidos abarca la hidrostática, la hidrodinámica (para fluidos incompresibles) y la aerodinámica (para fluidos compresibles) subsónica o supersónica.

Ecuación de conservación de la cantidad de movimiento

Se trata de una ecuación vectorial. En mecánica de fluidos se la suele llamar con el nombre de ecuación del equilibrio o ecuación de Navier-Stokes y modeliza los fluidos newtonianos con densidad y viscosidad constantes.

Obsérvese que la nomenclatura utilizada para las ecuaciones de conservación de la masa (ecuación de continuidad) y la ecuación de conservación de la cantidad de movimiento (ecuación de Navier-Stokes) surge de su aplicación a la mecánica de fluidos[53].

En realidad hay tres *grandes* leyes de conservación: ley de conservación de la masa, ley de conservación de la energía y ley de conservación de la cantidad de movimiento. Su aplicación al caso de los fluidos proporciona la ecuación de continuidad, las ecuaciones de Navier-Stokes y la ecuación de la energía. La aplicación a otras parcelas, como la mecánica de sólidos rígidos, la mecánica cuántica (distancias microscópicas), la mecánica relativista (distancias macroscópicas) o el electromagnetismo, se manifiesta por otras ecuaciones que se conocen con otros nombres pero que responden a las mismas leyes de conservación. La ecuación (vectorial) de Navier-Stokes es, por ejemplo, una consecuencia de la ecuación de conservación de la cantidad de movimiento:

$$\rho \frac{D\boldsymbol{v}}{Dt} = \rho \boldsymbol{g} - \nabla P + \mu \nabla^2 \boldsymbol{v}. \tag{1.84}$$

Dependiendo del sistema de coordenadas utilizado se expresa de distinta manera. Por ejemplo, las ecuaciones de Navier-Stokes en coordenadas rectangulares son:

[53] La mecánica de fluidos es sólo una rama de la mecánica de medios continuos que es la ciencia que estudia en general los medios deformables constituidos por *infinitos* puntos. Otra rama nace por ejemplo al estudiar los cuerpos elásticos (membranas, vigas, cuerdas vibrantes).

	Ecuaciones de Navier-Stokes en Coord. Rectangulares
x	$\rho\left[\frac{\partial v_x}{\partial t}+v_x\frac{\partial v_x}{\partial x}+v_y\frac{\partial v_x}{\partial y}+v_z\frac{\partial v_x}{\partial z}\right]=\rho g_x-\frac{\partial P}{\partial x}+\mu\left[\frac{\partial^2 v_x}{\partial x^2}+\frac{\partial^2 v_x}{\partial y^2}+\frac{\partial^2 v_x}{\partial z^2}\right]$
y	$\rho\left[\frac{\partial v_y}{\partial t}+v_x\frac{\partial v_y}{\partial x}+v_y\frac{\partial v_y}{\partial y}+v_z\frac{\partial v_y}{\partial z}\right]=\rho g_y-\frac{\partial P}{\partial y}+\mu\left[\frac{\partial^2 v_y}{\partial x^2}+\frac{\partial^2 v_y}{\partial y^2}+\frac{\partial^2 v_y}{\partial z^2}\right]$
z	$\rho\left[\frac{\partial v_z}{\partial t}+v_x\frac{\partial v_z}{\partial x}+v_y\frac{\partial v_z}{\partial y}+v_z\frac{\partial v_z}{\partial z}\right]=\rho g_z-\frac{\partial P}{\partial z}+\mu\left[\frac{\partial^2 v_z}{\partial x^2}+\frac{\partial^2 v_z}{\partial y^2}+\frac{\partial^2 v_z}{\partial z^2}\right]$

Las ecuaciones de Navier-Stokes en Coordenadas Cilíndricas son:

Componente	Ecuaciones de Navier-Stokes
componente r	$\rho\left[\frac{\partial v_r}{\partial t}+v_r\frac{\partial v_r}{\partial r}+\frac{v_\theta}{r}\frac{\partial v_r}{\partial \theta}-\frac{v_\theta^2}{r}+v_z\frac{\partial v_r}{\partial z}\right]=$ $\rho g_r-\frac{\partial P}{\partial r}+\mu\left[\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}(rv_r)\right)+\frac{1}{r^2}\frac{\partial^2 v_r}{\partial \theta^2}-\frac{2}{r^2}\frac{\partial v_\theta}{\partial \theta}+\frac{\partial^2 v_r}{\partial z^2{}^2}\right]$
componente θ	$\rho\left[\frac{\partial v_\theta}{\partial t}+v_r\frac{\partial v_\theta}{\partial r}+\frac{v_\theta}{r}\frac{\partial v_\theta}{\partial \theta}+\frac{v_r v_\theta}{r}+v_z\frac{\partial v_\theta}{\partial z}\right]=$ $\rho g_\theta-\frac{\partial P}{\partial \theta}+\mu\left[\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}(rv_\theta)\right)+\frac{1}{r^2}\frac{\partial^2 v_\theta}{\partial \theta^2}+\frac{2}{r^2}\frac{\partial v_\theta}{\partial \theta}+\frac{\partial^2 v_\theta}{\partial z^2{}^2}\right]$
componente z	$\rho\left[\frac{\partial v_z}{\partial t}+v_r\frac{\partial v_z}{\partial r}+\frac{v_\theta}{r}\frac{\partial v_z}{\partial \theta}+v_z\frac{\partial v_z}{\partial z}\right]=$ $\rho g_z-\frac{\partial P}{\partial z}+\mu\left[\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v}{\partial z}\right)\right)+\frac{1}{r^2}\frac{\partial^2 v_z}{\partial \theta^2}+\frac{\partial^2 v_z}{\partial z^2{}^2}\right]$

Ecuación de conservación de la energía

En muchos problemas que requieren una ecuación de la energía los efectos térmicos son mucho más importantes que los efectos mecánicos. Esto ocurre

por ejemplo cuando existen en el sistemas grandes variaciones de temperatura o cuando se produce calor por reacción. Si suponemos que la densidad ρ y el calor específico c_p son constantes entonces la ecuación de la energía toma la forma:

$$\rho c_p \frac{DT}{Dt} = -\nabla \cdot \boldsymbol{q} + H_V, \tag{1.85}$$

donde el término de fuente H_V representa la tasa de producción de energía debida a fuentes de energía externas por unidad de volumen. El ejemplo más común es el calor que se produce por la resistencia al flujo de una corriente eléctrica. El flujo de calor en un sólido o en un fluido puro se calcula utilizando la ley de Fourier:

$$\boldsymbol{q} = -k\nabla T.$$

Si la conductividad calorífica k es constante entonces la forma usual de la ecuación de la energía interna es:

$$\rho c_p \frac{DT}{Dt} = k\nabla^2 T + H_V. \tag{1.86}$$

En esta ecuación no está representado el fenómeno de la disipación viscosa que es la conversión de la energía cinética en calor debida a la fricción interna al fluido. Normalmente es despreciable exceptuando algunos flujos a alta velocidad (con gradientes de velocidades elevados) o el flujo de fluidos extremadamente viscosos (el hielo o el magma de los volcanes por ejemplo) Dependiendo del sistema de coordenadas considerado se tienen las siguientes expresiones de la ecuación (1.86), siendo $T(x, y, z, t)$ en coordenadas rectangulares, $T(r, \theta, z, t)$ en coordenadas cilíndricas y $T(r, \theta, \phi, t)$ en coordenadas esféricas:

Conservación de la Energía

Coordenadas Rectangulares (x, y, z)

$$\frac{\partial T}{\partial t} + v_x \frac{\partial T}{\partial x} + v_y \frac{\partial T}{\partial y} + v_z \frac{\partial T}{\partial z} = \alpha \left[\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right] + \frac{H_V}{\rho c_p}$$

Coordenadas Cilíndricas (r, θ, z)

$$\frac{\partial T}{\partial t} + v_r \frac{\partial T}{\partial r} + \frac{v_\theta}{r} \frac{\partial T}{\partial \theta} + v_z \frac{\partial T}{\partial z} = \alpha \left[\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial T}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 T}{\partial \theta^2} + \frac{\partial^2 T}{\partial z^2} \right] + \frac{H_V}{\rho c_p}$$

Coordenadas Esféricas (r, θ, ϕ)

$$\frac{\partial T}{\partial t} + v_r \frac{\partial T}{\partial r} + \frac{v_\theta}{r} \frac{\partial T}{\partial \theta} + \frac{v_\phi}{r \sin\theta} \frac{\partial T}{\partial \phi} =$$

$$= \alpha \left[\frac{1}{r^2} \frac{\partial T}{\partial r} \left(r^2 \frac{\partial T}{\partial r} \right) + \frac{1}{r^2 \sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial T}{\partial \theta} \right) + \frac{1}{r^2 \sin^2\theta} \left(\frac{\partial^2 T}{\partial \theta^2} \right) \right] + \frac{H_V}{\rho c_p}$$

1.7. Bibliografía

- Apostol, T.M., (1982), Análisis matemático. II ed. Editorial Reverté.
- Apostol, T.M., (1996), Calculus. II ed. Editorial Reverté.
- Costa Novella, E. (1986). Ingeniería Química. Vol. 2, Fenómenos de Transporte. Vol. 3, Flujo de Fluidos. Vol. 4, Transmisión de Calor. Vol. 5, Transferencia de Materia. Alhambra Universidad.
- Courant, R y Hilbert, D. (1989). Partial Differential Equations. Vol 2. John Wiley & Sons, Inc.
- Deen, W.M., (1998). Analysis of Transport Phenomena. Oxford University Press.
- Elsgoltz, L. (1983). Ecuaciones diferenciales y cálculo variacional. III Ed. Editorial Mir.
- Kreyszig, E. (1993). Advanced Engineering Mathematics. VII Ed. John Wiley & Sons, Inc.
- Marsden, J. E. y Tromba, A. J. (1991). Cálculo vectorial. (III edición). Ed. Addison-Wesley Iberoamericana
- Mei, C.C. (1997). Mathematical analysis in Engineering. Cambridge University Press.
- Rice, R.G, Do, D.D., (1995). Applied Mathematics and Modelling for Chemical Engineers. John Wiley & Sons, Inc.
- Smith, G.D. (1999). Numerical Solutions of Partial Differential Equations. Finite Difference Methods. III Ed. Oxford University Press.
- Strikwerda J. C., (1989). Finite Difference Schemes and Partial Differential Equations. Chapman & Hall. International Thomson Publishing.
- Weinberger, H., (1965). A first course in partial differential equations. Blaisdell.
- Zill, D.G., (1977). Ecuaciones Diferenciales con aplicaciones de modelado. International Thomson Ed. 6 ed.

1.7.1. Bibliografía avanzada

- Adams, R., (1975), Sobolev spaces. Academic Press.
- Brezis, H., (1983) Analyse Fonctionelle. Masson. Paris.
- Díaz, J.I., (1985). Nonlinear Partial Differential Equations and Free Boundaries. Ed. Pitman. Londres.
- Fowler, A.C, (1997). Mathematical Models in the Applied Sciences. Cambridge texts in Applied Mathematics. Cambridge University Press.
- Froment, G.F y Bischoff, K.B., (1990). Chemical Reactor Analysis and Design. II Ed. John Wiley & Sons, Inc.
- Godlewski, E., Raviart, P.A., (1991). Hyperbolic Systems of Conservation Laws. Ed. Ellipses.
- John, F., (1982). Partial differential equations. IV Ed. Applied Mathematical Sciences. Springer Verlag.

Capítulo 2

Métodos numéricos para la resolución de ecuaciones no lineales

2.1. Motivación y generalidades

Ejemplo 2.1.1 *(Hanna et al.[9]): La ecuación de Peng-Robinson es una ecuación de estado que proporciona la presión P de un gas mediante:*

$$P = \frac{RT}{V-b} - \frac{a}{V(V+b)+b(V-b)}$$

donde a y b son constantes, T es la temperatura absoluta a la que se encuentra el gas, V es el volumen específico y R es la constante de los gases perfectos (8.31441J/(mol ^oK)). Para el CO_2 las constantes a y b toman los valores a = 364.61 $m^6.kPa$ / $(kg.mol)^2$ y b = 0.02664 $m^3/kg.mol$. Supongamos que se desea encontrar la densidad (es decir 1/V) del CO_2 a una presión de $1{\cdot}10^4$ kPa y a una temperatura de 340^oK usando la ecuación de Peng-Robinson. Ello implicaría tener que encontrar el valor de V para el que: $10^4 = \frac{8{,}31441340}{V-0{,}02664} - \frac{364{,}61}{V(V+0{,}02664)+0{,}02664(V-0{,}02664)}$, *lo cual no es en modo alguno evidente. La solución ..... más adelante.*

Ejemplo 2.1.2 *(Hanna et al. [9]): Una mezcla de un mol de CO y 3 moles de H_2 se deja reaccionar a 500^oC y 1 atmósfera de presión hasta alcanzar su equilibrio químico y se desea estimar la composición de la mezcla de equilibrio. Las reacciones significativas (y las constantes de equilibrio) para este sistema son:*

$$
\begin{aligned}
3H_2 + CO &\rightleftharpoons CH_4 + H_2O \qquad (K = 69{,}18)\\
CO + H_2O &\rightleftharpoons CO_2 + H_2 \qquad (K = 4{,}68)\\
CO_2 + C &\rightleftharpoons 2CO \qquad (K = 0{,}0056)\\
5H_2 + 2CO &\rightleftharpoons C_2H_6 + 2H_2O \qquad (K = 0{,}141)
\end{aligned}
$$

Puesto que todos los componentes de las reacciones anteriores, salvo el C, aparecen en fase gaseosa, las relaciones de equilibrio, asumiendo que los gases se comportan como gases perfectos, se pueden expresar mediante:

$$\frac{y_{CH_4} y_{H_2O}}{y_{H_2}^3 y_{CO}} = 69{,}18 \qquad \text{(a)}$$

$$\frac{y_{CO_2} y_{H_2}}{y_{CO} y_{H_2O}} = 4{,}68 \qquad \text{(b)}$$

$$\frac{y_{CO}^2}{y_{CO_2}} = 0{,}0056 \qquad \text{(c)}$$

$$\frac{y_{C_2H_6} y_{H_2O}^2}{y_{H_2}^5 y_{CO}^2} = 0{,}141 \qquad \text{(d)}$$

donde y_α es la fracción molar de la especie "α". Usando las "coordenadas de reacción" x_i (i = 1, 2, 3, 4) de cada una de las cuatro reacciones anteriores y denotando por especie 1 al H_2, por especie 2 al CO, por especie 3 al CH_4, por especie 4 al H_2O, por especie 5 al CO_2 y por especie 6 al C_2H_6, las fracciones molares de cada especie y_i (i = 1, 2, 3, 4, 5, 6) satisfacen el sistema de ecuaciones siguiente:

$$
\begin{aligned}
y_1 &= (3 - 3x_1 + x_2 - 5x_4)/D\\
y_2 &= (1 - x_1 - x_2 + 2x_3 - 2x_4)/D\\
y_3 &= x_1/D\\
y_4 &= (x_1 - x_2 + 2x_4)/D\\
y_5 &= (x_2 - x_3)/D\\
y_6 &= x_4/D
\end{aligned}
$$

donde $D = (4 - 2x_1 + x_3 - 4x_4)$ representa el número total de moles presentes en el equilibrio. Entrando con estas expresiones en las ecuaciones (a), (b), (c)

y (d) se obtiene finalmente el sistema de cuatro ecuaciones no lineales:

$$\begin{aligned}
\frac{x_1(x_1 - x_2 + 2x_4)D^2}{(3 - 3x_1 + x_2 - 5x_4)^3(1 - x_1 - x_2 + 2x_3 - 2x_4)} &= 69{,}18 \\
\frac{(x_2 - x_3)(3 - 3x_1 + x_2 - 5x_4)}{(1 - x_1 - x_2 + 2x_3 - 2x_4)(x_1 - x_2 + 2x_4)} &= 4{,}68 \\
\frac{(1 - x_1 - x_2 + 2x_3 - 2x_4)^2}{(x_2 - x_3)D} &= 0{,}0056 \\
\frac{x_4(x_1 - x_2 + 2x_4)^2D^4}{(3 - 3x_1 + x_2 - 5x_4)^5(1 - x_1 - x_2 + 2x_3 - 2x_4)^2} &= 0{,}141
\end{aligned}$$

La resolución del sistema de ecuaciones no lineales anterior nos conduce a los valores de las coordenadas de reacción y, a partir de ellas, a la determinación de las fracciones molares de cada especie. Una solución de este sistema, con sentido químico, es

$$x_1 = 0{,}681604, \quad x_2 = 1{,}58962 \cdot 10^{-2}, \quad x_3 = -0{,}128703, \quad x_4 = 1{,}40960{,}10^{-5}$$

o

$$y_1 = 0{,}387162, \quad y_2 = 0{,}00179676, \quad y_3 = 0{,}271769,$$
$$y_4 = 0{,}265442, \quad y_5 = 0{,}0576549, \quad y_6 = 5{,}62034 \cdot 10^{-6}.$$

La cuestión es ¿cómo se han calculado estas soluciones?

Los dos ejemplos anteriores pretenden ilustrar el hecho de que en numerosas aplicaciones físicas y técnicas, y en concreto en muchos problemas propios de ingeniería, aparece la necesidad de tener que resolver ecuaciones o sistemas de ecuaciones no lineales.

Este tipo de sistemas tiene peculiaridades que los diferencian notablemente de los sistemas lineales. Así por ejemplo, los sistemas lineales de n ecuaciones con n incógnitas en los que la matriz del sistema es regular sólo admiten una solución. A diferencia de este caso, los sistemas no lineales, aunque tengan el mismo número de incógnitas que de ecuaciones, desde un punto de vista matemático, pueden admitir una, ninguna o varias soluciones. El elegir entre ellas las que sirven a la aplicación concreta que motivó el sistema de ecuaciones debe hacerse en función de los criterios físicos, químicos y técnicos que regulen el problema en cuestión (por ejemplo, aunque matemáticamente puedan tener sentido, químicamente serían inadmisibles fracciones molares 1 de una especie química negativas o superiores a 1).

Una segunda diferencia es la debida al hecho de que un sistema lineal que admita solución única puede ser resuelto de forma exacta mediante un número finito de operaciones (recuérdense los métodos directos de resolución

de sistemas lineales de ecuaciones: Gauss, LU, Choleski, Crout, QR, etc...). En el caso de los sistemas no lineales, en general, la solución no podrá ser encontrada mediante un número finito de operaciones. En este sentido, los métodos de resolución de sistemas de ecuaciones no lineales serán métodos de tipo iterativo mediante los cuales se construirá una sucesión de vectores que, en los casos en que el método funcione, se irán aproximando hacia uno de los vectores solución del sistema no lineal.

Observación 2.1.1 *Interprétese correctamente lo que se acaba de leer. No quiere ello decir que ninguna ecuación no lineal pueda resolverse de forma directa. Ahí están las ecuaciones de segundo grado que, siendo no lineales, pueden resolverse de forma exacta mediante un número finito de operaciones. O si se buscan las soluciones de la ecuación:* $(x-1)^{10} = 0$ *también es obvio que estas son* $x = 1$ *(con multiplicidad* 10*). Lo que se está diciendo es que no hay, por ejemplo, ningún método directo que nos permita calcular las raíces de cualquier polinomio de grado* 10*.*

El hecho de que los métodos de resolución de ecuaciones y sistemas no lineales sean de tipo iterativo nos plantea muchas cuestiones. Entre ellas cabe citar las siguientes:

a) ¿Cómo se genera la sucesión de vectores que puedan aproximarse a la solución?

b) Dado que es imposible evaluar los infinitos vectores de la sucesión anterior, ¿cómo se sabe que ya se está "suficientemente" cerca de una solución?.

c) Si la solución encontrada mediante un método no es la que pueda interesarnos ¿cómo buscar otras posibles soluciones?.

d) En el caso de tener diferentes métodos que nos proporcionen las soluciones de un sistema ¿cómo elegir el mejor entre ellos?.

A estas y otras cuestiones intentaremos dar respuesta en el presente tema. La descripción general de los principales métodos de resolución puede hacerse de una forma muy intuitiva sin necesidad de recurrir a artificios matemáticos complicados. No obstante la justificación rigurosa de las técnicas de resolución y el análisis de las condiciones que pueden garantizar su convergencia así como el estudio de la velocidad con que convergen exigirá acudir a conceptos matemáticos previos.

Conviene por último que el lector tome conciencia desde el primer momento de un hecho relativo a los métodos de resolución de sistemas de ecuaciones no

lineales: **no existe un método universal de resolución de sistemas de ecuaciones no lineales**. Algunos de ellos funcionarán sobre ciertos sistemas y no servirán para resolver otros. Los métodos que presenten un buen comportamiento sobre algunos sistemas pueden no ser los mejores para resolver otros sistemas diferentes. Más bien cabría decir que **cada sistema no lineal requerirá su método de resolución idóneo**.

2.2. Conceptos previos

Puesto que, como se acaba de señalar, los métodos que abordaremos serán de tipo iterativo y en ellos se generará una sucesión de vectores que, en el mejor de los casos, se vayan aproximando hacia un vector solución, conviene comenzar recordando algunos conceptos sobre sucesiones. En este sentido, en primer lugar, nos ubicaremos en conjuntos sobre los que se haya definido una forma de medir la distancia entre sus elementos (esto es, en un espacio métrico (E, d)). En este espacio métrico comenzamos recordando la siguiente definición:

Definición 2.2.1 *Dada una sucesión infinita de elementos* $\{x_i\}_{i=1}^{\infty}$ *del espacio métrico* (E, d) *se dice que la sucesión es* ***convergente*** *hacia el elemento* $x^* \in E$, *si para cualquier valor* $\varepsilon > 0$ *siempre se puede encontrar un número natural* N *tal que para todo índice* $n > N$ *se verifica que* $d(x_n, x^*) < \varepsilon$. *Al elemento* x^* *anterior, si existe, se le denomina* ***límite*** *de la sucesión* $\{x_i\}_{i=1}^{\infty}$.

Una sucesión de un espacio métrico podrá tener límite o no tenerlo pero en el caso de que exista este límite siempre será el único límite de la sucesión.

Se llama la atención del lector sobre el hecho de que el límite de una sucesión, si existe, tiene que ser un elemento del propio espacio métrico al que pertenecen los elementos de la sucesión. Así por ejemplo, si nos ubicamos en el espacio de los números racionales (Q) con la distancia fundamental ($d_f(x, y) = |x - y|$) la sucesión $\{x_n = \sum_{i=1}^{n}(1/i!)\}_{n=1}^{\infty}$ tiene elementos tan cercanos al número e como se desee. En efecto, recuérdese que el número e, entre otras definiciones, se podía obtener mediante $e = \sum_{i=1}^{\infty}(1/i!)$ por lo que dado un valor de ε bastará escoger N suficientemente elevado para que todos los elementos x_n de la sucesión anterior con $n > N$ disten de e una cantidad inferior a ε. Parecería pues que el número e es el límite de la sucesión anterior. Sin embargo e no es un número racional por lo que la sucesión anterior no tendrá límite en el espacio (Q, d_f) considerado. Sí lo tendría sin embargo, en el espacio $(\mathbb{R}, d_f)$.

Las sucesiones que "parece" que convergen (a falta de saber si hacia lo que convergen es un elemento del espacio en el que se está trabajando) tienen un nombre concreto: sucesiones de Cauchy. Más rigurosamente se puede dar la siguiente definición de este tipo de sucesiones:

Definición 2.2.2 *Dada una sucesión infinita de elementos* $\{x_i\}_{i=1}^{\infty}$ *del espacio métrico* (E,d) *se dice que la sucesión es una* ***sucesión de Cauchy****, si para cualquier valor* $\varepsilon > 0$ *siempre se puede encontrar un número natural* N *tal que para todo par de índices* $n > N$ *y* $m > N$ *se verifica que* $d(x_n, x_m) < \varepsilon$.

Ejemplo 2.2.1 *Trabajando con la distancia fundamental* d_f *(en el espacio métrico Q) la sucesión* $\left\{x_n = \sum_{i=1}^{n} \frac{1}{i!}\right\}_{n=1}^{\infty}$ *es una sucesión de Cauchy. En efecto se tiene que, dados tres números naturales* N, n *y* m *tales que* $N < n < m$ *:*

$$d(x_n, x_m) = \left|\sum_{i=1}^{n} \frac{1}{i!} - \sum_{i=1}^{m} \frac{1}{i!}\right| = \sum_{n+1}^{m} \frac{1}{i!} \leq \sum_{N+1}^{\infty} \frac{1}{i!} = e - \sum_{i=1}^{n} \frac{1}{i!}$$

por lo que para cualquier valor de ε *(positivo) bastará escoger* N *suficientemente elevado para que* $d_f(x_n, x_m) < \varepsilon$ *para todo par de números* n *y* m *mayores que* N.

Ejemplo 2.2.2 *En* $(\mathbb{R}, d_f)$ *la sucesión* $\left\{x_n = \frac{1}{n+1}\right\}_{n=1}^{\infty}$ *es una sucesión de Cauchy pues, siendo* $N < n$ *y* $N < m$ *se verificará:*

$$d_f(x_n, x_m) = \left|\frac{1}{n+1} - \frac{1}{m+1}\right| \leq \left|\frac{1}{n+1}\right| + \left|\frac{1}{m+1}\right| \leq$$

$$\left|\frac{1}{N+1}\right| + \left|\frac{1}{N+1}\right| = \frac{2}{N+1}$$

por lo que dado cualquier valor de ε *bastará con tomar* $N > \left(\frac{2}{\varepsilon} - 1\right)$ *para que* $d_f(x_n, x_m)$ *sea inferior a* ε.

Según la definición que se ha dado anteriormente las sucesiones de Cauchy son tales que las distancias entre todos sus elementos pueden hacerse inferiores a cualquier valor ε, a partir de un cierto índice N (que dependerá del valor ε elegido), es decir, que cuanto mayor sea el índice n, menos distará x_n de "x_∞". En otros términos, parece que estas sucesiones convergen hacia "algo". Lo único que faltaría para que ese "algo" fuese el límite de la sucesión es que perteneciese al espacio métrico en el que se considere la sucesión. En este sentido, para evitar el problema de que el supuesto límite tuviera la descortesía de no pertenecer al espacio, cuando sea posible se trabajará con espacios que tengan la sana costumbre de incluir entre sus elementos a todos los posibles límites de sus sucesiones. Estos espacios también tienen un nombre concreto: espacios métricos completos. Más rigurosamente:

Definición 2.2.3 *Se dice que el espacio métrico* (E,d) *es un* ***espacio métrico completo*** *si toda sucesión de Cauchy de elementos de* E *es una sucesión convergente en* (E,d).

Ejemplo 2.2.3 *En $\mathbb{R}^n$ se pueden considerar las distancias*

$$d_1(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|\,, \qquad d_2(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2},$$

$$d_\infty(\mathbf{x},\mathbf{y}) = \max_{1\le i\le n} \{|x_i - y_i|\}$$

Los espacios métricos $(\mathbb{R}^n, d_1)$, $(\mathbb{R}^n, d_2)$ y $(\mathbb{R}^n, d_\infty)$ son espacios métricos completos. Y siendo C un conjunto cerrado de $\mathbb{R}^n$ los espacios métricos (C, d_1), (C, d_2) y (C, d_∞) son también espacios métricos completos.

Ejemplo 2.2.4 *En el conjunto de los números reales $\mathbb{R}$ se define la distancia fundamental mediante $d_f(x,y) = |x-y|$. El espacio métrico $(\mathbb{R}, d_f)$ es un espacio métrico completo. Y siendo $[a,b]$ un intervalo cerrado el espacio $([a,b], d_f)$ también es un espacio métrico completo.*

Lo hasta aquí dicho es aplicable a espacios métricos en general. No obstante será habitual trabajar en espacios que tengan la estructura de espacios vectoriales (por ejemplo para buscar en ellos el vector solución de un sistema no lineal). En ellos la forma de medir distancias se asocia al concepto de norma de un vector (que, a su vez, generaliza el concepto de módulo de un vector). De forma más concreta puede darse la siguiente definición:

Definición 2.2.4 *Siendo E un espacio vectorial definido sobre un cuerpo K (habitualmente $K = \mathbb{R}$ o $K = C$) se denomina* ***norma*** *sobre E, y se representa por $\|\cdot\|$, a toda aplicación definida en E, que toma valores reales no negativos y verifica las condiciones siguientes:*

$$i) \qquad ||\mathbf{x}|| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$$

$$ii) \qquad ||\lambda\mathbf{x}|| = |\lambda|||\mathbf{x}||, \qquad \forall\lambda \in K, \quad \forall\mathbf{x} \in E$$

$$iii) \qquad ||\mathbf{x}+\mathbf{y}|| \le ||\mathbf{x}|| + ||\mathbf{y}|| \qquad \forall\,\mathbf{x},\mathbf{y} \in E$$

A todo espacio vectorial E sobre el que se defina una norma $\|\cdot\|$ se le denomina **espacio vectorial normado** *y se representará por $(E, \|\cdot\|)$.*

Observación 2.2.1 *En la definición anterior $|\lambda|$ representa el valor absoluto de λ si se trabaja en el cuerpo $K = \mathbb{R}$ de los números reales y el módulo de λ si se trabajase sobre el cuerpo $K = C$ de los números complejos.*

Ejemplo 2.2.5 *En $\mathbb{R}^n$ son normas vectoriales las siguientes aplicaciones:*

$$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|\,, \qquad ||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2},$$

$$||\mathbf{x}||_\infty = \max_{1\le i\le n} \{|x_i|\}$$

Definición 2.2.5 *Siendo $\|\cdot\|$ y $\|\cdot\|'$ dos normas definidas sobre un mismo espacio vectorial E, se dice que ambas normas son* ***equivalentes*** *si existen dos constantes k_1 y k_2 tales que se verifica:*

$$k_1||\mathbf{x}|| \leq ||\mathbf{x}||' \leq k_2||\mathbf{x}||, \qquad \forall\, \mathbf{x} \in E.$$

Ejemplo 2.2.6 *Si n es finito las normas sobre $\mathbb{R}^n$ introducidas en el ejemplo anterior son equivalentes. Es más, si la dimensión del espacio vectorial E es finita, todas las normas vectoriales sobre él definidas son equivalentes. Aun podría decirse más: si la dimensión del espacio vectorial normado es finita el espacio es completo.*

En un espacio vectorial normado $(E, \|\cdot\|)$ podrían definirse muy diferentes distancias. Entre todas ellas es habitual trabajar con la distancia que induce la norma vectorial. De forma más precisa, se puede definir como sigue:

Proposición 2.2.1 *Siendo $(E, \|\cdot\|)$ un espacio vectorial normado se verifica que la aplicación $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$ es una distancia denominada* ***distancia asociada a la norma*** *$\|\cdot\|$.*

Demostración:
Se verifica que:

$$d(\mathbf{x}, \mathbf{x}) = \|\mathbf{x} - \mathbf{x}\| = 0$$

$$\forall \mathbf{x}, \mathbf{y} \in E \,/\, \mathbf{x} \neq \mathbf{y}: \;\; d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| > 0$$

$$\forall \mathbf{x}, \mathbf{y} \in E: \; d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \|\mathbf{y} - \mathbf{x}\| = d(\mathbf{y}, \mathbf{x})$$

$$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in E: \; d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{x} - \mathbf{z}\| + \|\mathbf{z} - \mathbf{y}\| = d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$$

por lo que la aplicación definida es una distancia.

c.q.d.

Ejemplo 2.2.7 *A las normas antes definidas se les asocian respectivamente las distancias d_1, d_2 y d_∞ consideradas en ejemplos precedentes.*

Una conclusión de lo anterior es que los espacios vectoriales normados: $(\mathbb{R}^n, \|\cdot\|_1)$, $(\mathbb{R}^n, \|\cdot\|_2)$ y $(\mathbb{R}^n, \|\cdot\|_\infty)$ son espacios métricos completos.

Observación 2.2.2 *El que dos normas sean equivalentes no quiere decir que al aplicarlas a un mismo vector tomen el mismo valor pero sí que nos indica que si una de ellas toma un valor "elevado" al aplicarla a un cierto vector de E, la otra también tomará un valor "elevado" al aplicarla al mismo vector. Y si el valor es "pequeño" para una de ellas también lo será para la otra. En ese sentido las distancias que a ellas están asociadas también serán equivalentes.*

Y por ello si una sucesión es convergente con la distancia asociada a una de las normas también lo será con la otra. Como en el caso de trabajar en $\mathbb{R}^n$ todas las normas son equivalentes a efectos de analizar la convergencia de una sucesión de vectores será equivalente hacerlo con una u otra norma que se considere en $\mathbb{R}^n$.

En algunas ocasiones trabajaremos con el espacio formado por las matrices cuadradas de orden n. Estos conjuntos tienen estructura de espacio vectorial y por tanto sobre ellos sería posible definir también normas y distancias como se acaba de describir. No obstante sobre el espacio de las matrices cuadradas de orden n a las normas que en él se utilizan se las exige algo más. De forma concreta para estos espacios se tiene la siguiente definición:

Definición 2.2.6 *Siendo M_n el espacio de las matrices cuadradas de orden n definidas sobre un cuerpo K (con $K = \mathbb{R}$ o $K = C$) se denomina* ***norma matricial*** *definida sobre M_n a toda aplicación definida en M_n que toma valores reales no negativos y que verifica las propiedades siguientes:*

$$\begin{aligned}
&i) && \|\mathbf{A}\| = 0 \Leftrightarrow \mathbf{A} = \mathbf{0} \\
&ii) && \|\lambda\mathbf{A}\| = |\lambda|\|\mathbf{A}\|, \qquad \forall\lambda \in K, \quad \forall\mathbf{A} \in M_n \\
&iii) && \|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\| \qquad \forall\mathbf{A},\mathbf{B} \in M_n \\
&iv) && \|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|, \qquad \forall\mathbf{A},\mathbf{B} \in M_n
\end{aligned}$$

Ejemplo 2.2.8 *En M_n son normas matriciales las siguientes:*

$$\|\mathbf{A}\|_F = \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{n}|a_{ij}|^2},\ \|\mathbf{A}\|_1 = \max_{1\leq j\leq n}\left\{\sum_{i=1}^{n}|a_{ij}|\right\},$$

$$\|\mathbf{A}\|_\infty = \max_{1\leq i\leq n}\left\{\sum_{j=1}^{n}|a_{ij}|\right\}$$

donde $\|\mathbf{A}\|_F$ se conoce como la norma de Fröbenius.

Asimismo, siendo $\rho(\mathbf{A})$ el radio espectral de la matriz $\mathbf{A}$ (es decir, el módulo del valor propio de $\mathbf{A}$ que tenga mayor módulo) y designando por $\mathbf{A^*}$ a la matriz adjunta de $\mathbf{A}$ (la traspuesta si $\mathbf{A}$ es una matriz real) la aplicación $\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^*\cdot\mathbf{A})}$ también es una norma matricial sobre M_n.

Entre todas las normas matriciales que se pueden definir en M_n es útil considerar un grupo de ellas que tiene una peculiaridad interesante: estar definidas a partir de una norma vectorial y, por tanto, estar vinculadas a normas vectoriales. Este grupo de normas matriciales se conocen con el nombre de normas

matriciales subordinadas a una norma vectorial y pasamos a definirlo a continuación (mediante una propiedad que puede consultarse por ejemplo en Ciarlet & Lions [4])

Proposición 2.2.2 *Sea $\|\cdot\|$ una norma vectorial definida sobre K^n (con $K = \mathbb{R}$ o $K = C$) y sea M_n el espacio de las matrices cuadradas de orden n definidas sobre K. La aplicación $\|\cdot\|$ definida de cualquiera de las formas (equivalentes entre sí) siguientes:*

$$\|\mathbf{A}\| = Sup_{\{\mathbf{v}\in K^n\setminus\mathbf{0}\}}\left(\frac{\|\mathbf{Av}\|}{\|\mathbf{v}\|}\right) = Sup_{\{\mathbf{v}\in K^n\setminus\mathbf{0}/\|\mathbf{v}\|\leq 1\}}\left(\|\mathbf{Av}\|\right) =$$

$$Sup_{\{\mathbf{v}\in K^n/\,\|\mathbf{v}\|=1\}}\left(\|\mathbf{Av}\|\right)$$

es una norma matricial que se denomina ***norma matricial subordinada*** *a la norma vectorial $\|\cdot\|$.*

Observación 2.2.3 *En la definición anterior se ha utilizado el mismo símbolo ($\|\cdot\|$) para referirse a la norma matricial y a la norma vectorial. Fácilmente distinguirá el lector entre una y otra por el tipo de elemento al que se aplica.*

Las normas matriciales subordinadas permiten trabajar con formas de medir "coherentes" entre los vectores y las matrices cuando estos aparecen mezclados en los problemas que deban abordarse. Es importante en este sentido tener siempre presente la siguiente propiedad que relaciona el valor de una norma vectorial con la norma matricial subordinada a ella:

Proposición 2.2.3 *Siendo $\|\cdot\|$ una norma matricial (subordinada a la norma vectorial $\|\cdot\|$) se verifica que:*

$$\|\mathbf{Av}\| \leq \|\mathbf{A}\|\|\mathbf{v}\| \qquad \forall \mathbf{A} \in M_n, \qquad \forall \mathbf{v} \in K^n$$

Demostración:

Si $\mathbf{v} = \mathbf{0}$ entonces $\|\mathbf{Av}\| = 0$ y $\|\mathbf{A}\|\|\mathbf{v}\| = 0$, $\forall\, \mathbf{A} \in M_n$ por lo que se verifica el teorema (con el signo "=").

Si $\mathbf{v} \neq \mathbf{0}$ se tiene que $\|\mathbf{v}\| \neq 0$ y por tanto:

$$\|\mathbf{Av}\| = \left\|\mathbf{A}\frac{\mathbf{v}}{\|\mathbf{v}\|}\|\mathbf{v}\|\right\| = \left\|\mathbf{A}\frac{\mathbf{v}}{\|\mathbf{v}\|}\right\|\|\mathbf{v}\| = \frac{\|\mathbf{Av}\|}{\|\mathbf{v}\|}\|\mathbf{v}\| \leq$$

$$\leq Sup_{\mathbf{u}\in K^n\setminus\mathbf{0}}\left(\frac{\|\mathbf{Au}\|}{\|\mathbf{u}\|}\right)\|\mathbf{v}\| = \|\mathbf{A}\|\|\mathbf{v}\|$$

y esto se tiene $\forall\, \mathbf{A} \in M_n$ y $\forall\, \mathbf{v} \in K^n \setminus \mathbf{0}$.

c.q.d.

Observación 2.2.4 *Las normas matriciales $\|\cdot\|_1$, $\|\cdot\|_2$, y $\|\cdot\|_\infty$ antes definidas son normas matriciales subordinadas a las normas vectoriales de $\mathbb{R}^n$ definidas con los mismos subíndices. Sin embargo, la norma de Fröbenius no es una norma matricial subordinada.*

Ejemplo 2.2.9 *Sea* $\mathbf{A}$ *la matriz:*

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}$$

Se verifica que:

$$\sum_{j=1}^{3} |a_{1,j}| = |1| + |1| + |0| = 2$$

$$\sum_{j=1}^{3} |a_{2,j}| = |1| + |2| + |1| = 4$$

$$\sum_{j=1}^{3} |a_{3,j}| = |-1| + |1| + |2| = 4$$

por lo que $\|\mathbf{A}\|_\infty = Sup(2,4,4) = 4$. *Por otra parte:*

$$\sum_{i=1}^{3} |a_{i,1}| = |1| + |1| + |-1| = 3$$

$$\sum_{i=1}^{3} |a_{i,2}| = |1| + |2| + |1| = 4$$

$$\sum_{i=1}^{3} |a_{i,3}| = |0| + |1| + |2| = 3$$

por lo que $\|\mathbf{A}\|_1 = Sup(3,4,3) = 4$. *Asimismo:*

$$\mathbf{A}^T.\mathbf{A} = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} . \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 2 & -1 \\ 2 & 6 & 4 \\ -1 & 4 & 5 \end{pmatrix}$$

El polinomio característico de $\boldsymbol{A^T A}$ *es:*

$$p(\lambda) = \begin{vmatrix} (3-\lambda) & 2 & -1 \\ 2 & (6-\lambda) & 4 \\ -1 & 4 & (5-\lambda) \end{vmatrix} = -\lambda(\lambda^2 - 14\lambda + 12)$$

cuyas raíces son los valores propios $\lambda_1 = 0$, $\lambda_2 = 7+\sqrt{7}$ *y* $\lambda_3 = 7-\sqrt{7}$. *Por tanto el radio espectral de* $\boldsymbol{A^T A}$ *es:* $\rho(\mathbf{A}^T.\mathbf{A}) = Sup(0, 7+\sqrt{7}, 7-\sqrt{7}) = 7+\sqrt{7}$. *Y la norma-2 de* $\boldsymbol{A}$ *será:*

$$\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^T\mathbf{A})} = \sqrt{7+\sqrt{7}} \approx 3{,}106$$

En los métodos iterativos que plantearemos en este capítulo, la sucesión de vectores que, en su caso, vayan aproximándose hacia la solución, se generará a partir de un vector inicial $\mathbf{x}^{(0)}$ mediante un esquema de cálculo que se traducirá en determinar una función $\mathbf{g}(\mathbf{x})$ con la que el proceso iterativo se reducirá a:

$$\mathbf{x}^{(0)} \quad dado$$

$$\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)}) \qquad (i = 0, 1, 2, ...)$$

Más adelante estudiaremos la forma de definir esta función $\mathbf{g}(\mathbf{x})$ (que obviamente tendrá que ver con el sistema que se quiera resolver). No obstante ya puede apreciarse con la consideración que acaba de hacerse que el buen funcionamiento de un método dependerá de cómo sea la función $\mathbf{g}(\mathbf{x})$ escogida. Por tanto, nos interesará trabajar, cuando ello sea posible con aplicaciones $\mathbf{g}(\mathbf{x})$ para las que se pueda asegurar que la sucesión que con ellas se genere es convergente hacia alguna solución del sistema que se quiere resolver.

Observación 2.2.5 *En este sentido, dado un sistema de ecuaciones* $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ *y considerado un método de resolución del mismo en la forma: "Dado* $\mathbf{x}^{(0)} \in D$, $\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)})$, $(i = 0, 1, ...)$*", se dice que el método es consistente con el sistema de ecuaciones cuando para toda solución* $\mathbf{x}^*$ *del sistema se verifica que* $\mathbf{x}^* = \mathbf{g}(\mathbf{x}^*)$ *y, viceversa, todo vector que verifique que* $\mathbf{x}^* = \mathbf{g}(\mathbf{x}^*)$ *es solución del sistema. Asimismo se dirá que el método es estable cuando para cualquier vector* $\mathbf{x}^{(0)}$ *de partida tomado en* D *se verifique que la sucesión* $\left\{\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)})\right\}_{i=0}^{\infty}$ *es una sucesión convergente. Y por último diremos que el método es convergente en el dominio* D*, para el sistema considerado, cuando para cualquier* $\mathbf{x}^{(0)} \in D$ *se verifique que la sucesión* $\left\{\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)})\right\}_{i=0}^{\infty}$ *converge hacia alguna raíz* $\mathbf{x}^*$ *de la ecuación. Obsérvese que con estas definiciones, todo método que sea a la vez consistente y estable es convergente y que todo método que sea convergente es consistente y estable. Esta relación de consistencia, estabilidad y convergencia volveremos a plantearlas al estudiar métodos numéricos para la resolución de ecuaciones diferenciales y, allí, será de enorme utilidad para el análisis de la convergencia de los esquemas numéricos. Nosotros, en este capítulo, nos limitaremos a analizar la convergencia de los métodos numéricos por otros procedimientos.*

Si $\mathbf{x}^*$ fuese el límite de la sucesión generada mediante el esquema anterior y además la aplicación $\mathbf{g}(\mathbf{x})$ fuese continua se verificará que:

$$\lim_{i\to\infty}(\mathbf{x}^{(i+1)}) = \lim_{i\to\infty}\mathbf{g}(\mathbf{x}^{(i)}) \quad \Leftrightarrow \quad \mathbf{x}^* = \mathbf{g}(\mathbf{x}^*)$$

Los puntos $\mathbf{x}^*$, del dominio sobre el que está definida una aplicación $\mathbf{g}(\mathbf{x})$, que verifican que $\mathbf{x}^* = \mathbf{g}(\mathbf{x}^*)$ reciben el nombre de puntos fijos de la aplicación. Más concretamente:

Definición 2.2.7 *Siendo* $\mathbf{g}$ *una aplicación definida en un espacio métrico (E, d) y con valores en el mismo espacio métrico, se denomina* ***punto fijo*** *de la aplicación* $\mathbf{g}$ *a todo elemento* $\mathbf{x}^*$ *de* E *tal que* $\mathbf{x}^* = \mathbf{g}(\mathbf{x}^*)$.

Interesará por tanto trabajar con funciones $\mathbf{g}$ que posean un punto fijo. Un tipo de tales funciones son las que se denominan contracciones y que pasamos a definir a continuación:

Definición 2.2.8 *Sean (E, d) y (V, d') dos espacios métricos y sea* $g : E \to V$ *una aplicación definida en* E *y con valores en* V. *Se dice que* g *es una* ***aplicación lipschitciana*** *cuando existe una constante real* $k > 0$ *tal que:*

$$d'(g(x), g(y)) \leq k d(x, y) \qquad \forall\, x, y \in E$$

A la menor constante k *que verifica la condición anterior se la denomina* ***constante de Lipschitz (o razón)*** *de la aplicación.*

Observación 2.2.6 *En el caso de que se esté trabajando sobre los espacios vectoriales normados* $(E, \|\cdot\|)$ *y* $(V, \|\cdot\|')$ *toda aplicación* $\mathbf{g} : E \to V$ *que sea lipschitciana de razón* k *verificará:*

$$\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\|' \leq k\|\mathbf{x} - \mathbf{y}\|, \qquad \forall\, \mathbf{x}, \mathbf{y} \in E$$

Proposición 2.2.4 *Toda aplicación lipschitciana definida en* (E, d) *y con valores en* (V, d') *es una aplicación continua en todo* E.

Demostración:

Si $\mathbf{g} : E \to V$ es una aplicación lipschitciana con constante de Lipschitz k se verificará que $d'(\mathbf{g}(\mathbf{x}), \mathbf{g}(\mathbf{y})) \leq k \cdot d(\mathbf{x}, \mathbf{y})$. Por tanto para cualquier valor de ε estrictamente positivo y para cualquier punto $\mathbf{x} \in E$ se tiene que:

$$\forall\, \varepsilon > 0, \quad \exists\, \delta = \frac{\varepsilon}{k} \;\; / d(\mathbf{x}, \mathbf{y}) < \delta \Rightarrow d'(\mathbf{g}(\mathbf{x}), \mathbf{g}(\mathbf{y})) \leq k \cdot d(\mathbf{x}, \mathbf{y}) < \varepsilon$$

Por tanto g es continua en todo punto $x \in E$.

c.q.d.

Definición 2.2.9 *A toda aplicación lipschitciana que verifique las dos condiciones siguientes:*

1^a) *Estar definida en un espacio métrico* (E, d) *sobre sí mismo:* $g : E \to E$,

2^a) *Tener una constante de Lipschitz estrictamente inferior a* 1,

se la denomina ***contracción*** *sobre E.*

El hecho de que para las contracciones se garantice la convergencia de la sucesión generada mediante el esquema de cálculo:

$$\mathbf{x}^{(0)} \quad dado$$
$$\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)}) \qquad (i = 0, 1, 2, ...)$$

se debe al teorema que se expone a continuación. Se recomienda prestar atención a la demostración del mismo, realizada mediante la técnica de **aproximaciones sucesivas**, pues en ella se recogen las bases de los métodos de resolución de sistemas de ecuaciones.

Teorema 2.2.1 (del punto fijo) *Toda contracción definida sobre un espacio métrico completo admite un único punto fijo.*

Demostración:

a) **Existencia**. Comencemos demostrando la existencia del punto fijo. Sea $\mathbf{g} : E \to E$ una contracción, de constante de Lipschitz $k < 1$, definida en el espacio métrico (E, d) y sea $\mathbf{x}^{(0)}$ un elemento cualquiera de E. Considérese entonces la sucesión formada a partir de $\mathbf{x}^{(0)}$ mediante:

$$\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)}), \qquad i = 0, 1, 2, , ...$$

Para la sucesión $\left\{\mathbf{x}^{(i)}\right\}_{i=0}^{\infty}$ se verificará:

$$d(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = d(\mathbf{g}(\mathbf{x}^{(0)}), \mathbf{g}(\mathbf{x}^{(1)})) \leq kd(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})$$

$$d(\mathbf{x}^{(2)}, \mathbf{x}^{(3)}) = d(\mathbf{g}(\mathbf{x}^{(1)}), \mathbf{g}(\mathbf{x}^{(2)})) \leq kd(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \leq k^2 d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})$$

.....

$$d(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}) \leq k^n d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})$$

.....

De estas desigualdades, aplicando la desigualdad triangular de las distancias, resultará que:

$$d(\mathbf{x}^{(n)}, \mathbf{x}^{(n+p)}) \leq d(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}) + d(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+p)}) \leq$$

$$\leq d(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}) + d(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+2)}) + d(\mathbf{x}^{(n+2)}, \mathbf{x}^{(n+p)}) \leq$$

$$\leq d(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}) + d(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+2)}) + ... + d(\mathbf{x}^{(n+p-1)}, \mathbf{x}^{(n+p)}) \leq$$

$$\leq k^n d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) + k^{(n+1)} d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) + ... + k^{(n+p-1)} d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) =$$

$$= k^n d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})[1 + k + ... + k^{(p-1)}] \leq k^n d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) \left(\sum_{i=0}^{\infty} k^i \right)$$

En la expresión anterior, el sumatorio que aparece representa la suma de una progresión geométrica de razón $k(< 1)$ y cuyo primer término toma el valor 1. Por tanto:

$$\sum_{i=0}^{\infty} k^i = \frac{1}{1-k}$$

lo que nos conduce a que:

$$d(\mathbf{x}^{(n)}, \mathbf{x}^{(n+p)}) \leq \frac{k^n}{1-k} d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})$$

y puesto que, al ser $\mathbf{g}(\mathbf{x})$ una contracción, k es estríctamente inferior a 1, para cualquier valor de ε positivo bastará con considerar el índice natural N de forma que:

$$N \geq \frac{\log\left(\frac{\varepsilon(1-k)}{d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})} \right)}{\log(k)}$$

para que se verifique que $d(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) < \varepsilon$ para todo par de índices n y m mayores que N. En definitiva, la sucesión $\left\{ \mathbf{x}^{(i)} \right\}_{i=0}^{\infty}$ es una sucesión de Cauchy. Y como, por las hipótesis del teorema, se está trabajando en un espacio métrico completo, admitirá límite $\mathbf{x}^*$. Y puesto que al ser $\mathbf{g}$ una contracción es continua se verificará que:

$$\mathbf{g}(\mathbf{x}^*) = \lim_{i \to \infty} \mathbf{g}(\mathbf{x}^{(i)}) = \lim_{i \to \infty} \mathbf{x}^{(i+1)} = \mathbf{x}^*$$

Luego $\mathbf{g}(\mathbf{x})$ admite un punto fijo que es el límite de la sucesión generada mediante:

$$\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)}), \qquad i = 0, 1, 2, ...$$

a partir de cualquier elemento $\mathbf{x}^{(0)}$ perteneciente a E.

b) **Unicidad**. Demostremos ahora la unicidad del punto fijo. Esta cuestión es cómoda demostrarla mediante reducción al absurdo. En efecto, consideremos por un momento que en las condiciones del teorema hubiera dos elementos distintos de E, que denotaremos por $\mathbf{a}$ y $\mathbf{b}$, que fuesen puntos fijos. Al ser distintos se tendrá que $d(\mathbf{a},\mathbf{b}) > 0$. Pero por otra parte se debe verificar que:

$$d(\mathbf{a},\mathbf{b}) = d(\mathbf{g}(\mathbf{a}),\mathbf{g}(\mathbf{b})) \leq kd(\mathbf{a},\mathbf{b}) < d(\mathbf{a},\mathbf{b})$$

Y que un número real sea estríctamente menor que sí mismo obviamente es absurdo. Obsérvese que si por el contrario se supusiera que $\mathbf{a} = \mathbf{b}$, las desigualdades anteriores se transformarían en:

$$0 = d(\mathbf{a},\mathbf{b}) = d(\mathbf{g}(\mathbf{a}),\mathbf{g}(\mathbf{b})) \leq kd(\mathbf{a},\mathbf{b}) = 0$$

que sí tiene sentido. Por tanto, es absurdo suponer que existen puntos fijos distintos.

c.q.d.

Observación 2.2.7 *Entiéndase bien el teorema anterior. En él sólo se afirma lo que se afirma. Ello no imposibilita que otras aplicaciones que no sean contracciones, o que estén definidas en espacios que no sean completos, puedan tener uno o varios puntos fijos. Simplemente nos asegura que si nuestra aplicación es una contracción y está definida sobre un espacio métrico completo siempre existirá un único punto fijo de la aplicación. La demostración de la existencia del punto fijo nos indica además cómo puede encontrarse: como límite de la sucesión* $\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)})$ *generada a a partir de cualquier* $\mathbf{x}^{(0)}$ *perteneciente al espacio* E.

Ejemplo 2.2.10 *A continuación se consideran algunas situaciones concretas que ilustran las caracteríticas principales del teorema del punto fijo.*

1. *La aplicación* $g : (\mathbb{R}, d_f) \to (\mathbb{R}, d_f)$ *definida mediante* $g(x) = \frac{x}{2}$ *es una contracción ya que:*

$$d_f(g(x), g(y)) = \left|\frac{x}{2} - \frac{y}{2}\right| = \frac{1}{2}|x-y| = \frac{1}{2}d_f(x,y) \leq \frac{1}{2}d_f(x,y).$$

   Esta aplicación, al estar definida en un espacio métrico completo sólo admite un punto fijo: $x = 0$.

2. *La misma aplicación anterior pero definida en el espacio métrico* $(]0,1[, d_f)$ *no admite punto fijo pues* 0 *no pertenece al espacio. Obsérvese que* $(]0,1[, d_f)$ *no es completo.*

3. *La misma aplicación pero definida en el espacio* $([1,2],d_f)$ *tampoco admite punto fijo. Obsérvese que aunque* $([1,2],d_f)$ *sí es completo* g *no es una contracción pues no toma valores en* $([1,2],d_f)$ *(pues por ejemplo* $g(1{,}5)=3/4\notin[1,2]$*).*

4. *La misma aplicación definida en* $(]-1,1[,\ d_f)$ *tiene por punto fijo* $x=0$*. Nótese que no está definida sobre un completo y sin embargo admite un (único) punto fijo.*

5. *La aplicación* $g(x)=\frac{2x+6}{3x+2}$ *definida en* (E,d_f) *siendo* $E=\{x\in\mathbb{R}/x\geq 2/3\}$ *es una contracción ya que:*

$$|g(x)-g(y)|=\left|\frac{2x+6}{3x+2}-\frac{2y+6}{3y+2}\right|=\left|\frac{14y-14x}{9xy+6x+6y+4}\right|=$$

$$=\frac{14\,|x-y|}{|9xy+6x+6y+4|}\leq\frac{14\,|x-y|}{\left|9\frac{2}{3}\frac{2}{3}+6\frac{2}{3}+6\frac{2}{3}+4\right|}=\frac{7}{8}\,|x-y|$$

Al estar definida en un espacio métrico completo y tomar valores en él admitirá un único fijo que está dado por:

$$x^*=g(x^*)\Leftrightarrow x^*=\frac{2x^*+6}{3x^*+2}\Leftrightarrow 3(x^*)^2+2x^*=2x^*+6\Rightarrow x^*=\sqrt{2}$$

6. *La misma aplicación del ejemplo anterior sobre* $(\mathbb{R}-\{2/3\},\ d_f)$ *admite dos puntos fijos dados por* $x^*=\pm\sqrt{2}$*. Obsérvese que el espacio sobre el que está definida la aplicación no es un espacio métrico completo pues al haber quitado de* $\mathbb{R}$ *el punto* $2/3$ *habrá sucesiones de Cauchy (como por ejemplo* $\left\{x_i=\frac{2}{3}+\frac{1}{i+1}\right\}_{i=0}^{\infty}$*) que no tengan límite.*

Observación 2.2.8 *La demostración de la existencia del punto fijo que se ha realizado en el teorema precedente ya pone de manifiesto muchos aspectos importantes sobre los métodos iterativos de resolución de ecuaciones no lineales. En efecto, si logramos definir una contracción* $\mathbf{g}$ *con la que generar una sucesión que converja hacia la solución de las ecuaciones no lineales a resolver ya habremos dado un primer paso. La distancia que separe* $\mathbf{x}^{(n)}$ *de la solución* $\mathbf{x}^*(=\text{“}\mathbf{x}^{(\infty)}\text{”})$ *podrá estimarse mediante:*

$$d(\mathbf{x}^{(n)},\mathbf{x}^*)=d(\mathbf{x}^{(n)},\mathbf{x}^{(\infty)})\leq d(\mathbf{x}^{(n)},\mathbf{x}^{(n+1)})+d(\mathbf{x}^{(n+1)},\mathbf{x}^{(\infty)})\leq$$

$$\leq d(\mathbf{x}^{(n)},\mathbf{x}^{(n+1)})+d(\mathbf{x}^{(n+1)},\mathbf{x}^{(n+2)})+d(\mathbf{x}^{(n+2)},\mathbf{x}^{(\infty)})\leq$$

$$\leq d(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}) + d(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+2)}) + ... + d(\mathbf{x}^{(n+j)}, \mathbf{x}^{(n+j+1)}) + ... \leq$$

$$\leq k^n d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) + k^{(n+1)} d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) + ... + k^{(n+j)} d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) + ... =$$

$$= k^n d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})[1 + k + ... + k^{(j)} + ...] = k^n d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) \left(\sum_{i=0}^{\infty} k^i \right)$$

luego

$$d(\mathbf{x}^{(n)}, \mathbf{x}^*) \leq \frac{k^n}{1-k} d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})$$

Ello pone de manifiesto diferentes hechos (para el caso en que ya se haya resuelto el problema de que la sucesión converja). Entre ellos podemos citar:

a) No es indiferente el elemento $\mathbf{x}^{(0)}$ *con el que se inicialice la sucesión pues el "ahorrarnos" iteraciones en el proceso depende de* $d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})$.

b) Cuanto más próxima a 1 *sea la constante de Lipschitz de la contracción más pequeño será* $(1-k)$ *y por tanto mayor será el número* (n) *de iteraciones que se necesitarán para estar "razonablemente" cerca de la solución. En otros términos cuanto más próxima a* 0 *sea la constante de Lipschitz de las contracciones con las que se trabaje, menor será el esfuerzo de cálculo necesario para obtener "buenas" aproximaciones de las soluciones.*

c) Si en lugar de acotar la distancia a la solución con $d(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})$ *se acotara con* $d(\mathbf{x}^{(n)}, \mathbf{x}^{(n-1)})$ *se tendría que:*

$$d(\mathbf{x}^{(n)}, \mathbf{x}^*) = d(\mathbf{x}^{(n)}, \mathbf{x}^{(\infty)}) \leq d(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}) + d(\mathbf{x}^{(n+1)}, \mathbf{x}^{(\infty)}) \leq$$

$$\leq d(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}) + d(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+2)}) + d(\mathbf{x}^{(n+2)}, \mathbf{x}^{(\infty)}) \leq$$

$$\leq d(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}) + d(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+2)}) + ... + d(\mathbf{x}^{(n+j)}, \mathbf{x}^{(n+j+1)}) + ... \leq$$

$$\leq k d(\mathbf{x}^{(n-1)}, \mathbf{x}^{(n)}) + k^2 d(\mathbf{x}^{(n-1)}, \mathbf{x}^{(n)}) + ... + k^{(j)} d(\mathbf{x}^{(n-1)}, \mathbf{x}^{(n)}) + ... =$$

$$= kd(\mathbf{x}^{(n-1)}, \mathbf{x}^{(n)})[1 + k + ... + k^{(j)} + ...] = kd(\mathbf{x}^{(n-1)}, \mathbf{x}^{(n)}) \left(\sum_{i=0}^{\infty} k^i \right)$$

luego

$$d(\mathbf{x}^{(n)}, \mathbf{x}^*) \leq \frac{k}{1-k} d(\mathbf{x}^{(n)}, \mathbf{x}^{(n-1)})$$

Ello nos indica que la distancia entre dos aproximaciones consecutivas de la solución es una forma de medir la distancia de la última de ellas a la solución... ponderada por el factor $\frac{k}{1-k}$ *(lo que nos vuelve a llevar a la consideración de que interesan valores de la constante de Lipschitz lo más pequeños posible).*

2.3. Métodos generales para la resolución de una única ecuación no lineal

En este apartado se expondrán métodos generales que nos permiten encontrar soluciones de una ecuación no lineal. Como se señala en el último subapartado, para ecuaciones no lineales concretas (por ejemplo polinómicas) es posible adaptar algunos de estos métodos y, en algunos casos, existen métodos de aplicación específica a ellas.

Una ecuación no lineal la representaremos genéricamente en la forma $f(x) = 0$. Obsérvese que lo anterior no quiere decir que la función $f(x)$ sea la función idénticamente nula. Simplemente es una forma de representar la ecuación a la que nos enfrentemos. Más concretamente, tras la expresión $f(x) = 0$ se ocultará el siguiente problema:

Dada una función $f(x)$ *determínese, si es posible, algún valor* x^* *para el que se verifique que* $f(x^*) = 0$.

A los valores x^* para los que la función $f(x)$ se anula habitualmente se les denomina raíces (o ceros) de la función. En general si $f(x)$ admite el valor x^* como raíz, se podrá encontrar un número positivo m y una función $\varphi(x)$ tales que $\varphi(x^*) \neq 0$ y, en un entorno de x^*,

$$f(x) = (x - x^*)^m \varphi(x).$$

En esa situación se dirá que x^* es una raíz de multiplicidad m. Así por ejemplo, la función $f(x) = (x-1)^2(x+2)\sin(x)$ admite, entre otras, las raíces $x_1^* = 1$ (de multiplicidad 2) y $x_2^* = -2$ (de multiplicidad 1). O la función $f(x) = \sqrt[3]{x}(x-4)^2 e^x$ admite la raíz $x_1^* = 0$ (de multiplicidad 1/3) y la raíz $x_2^* = 4$ (de multiplicidad 2).

No siempre está tan claro cuál es la multiplicidad de una raíz. Por ejemplo, la función $f(x) = \sin(x)$ tiene una raíz en $x = 0$. Para expresarla en la forma

$f(x) = xg(x)$ la función $g(x)$ debería ser $g(x) = \sin(x)/x$ pero esta función no está definida en $x = 0$. Por ello, teniendo en cuenta que $\lim_{x\to 0}[\sin(x)/x] = 1$ la función $g(x)$ se define en este caso como:

$$g(x) = \begin{cases} \dfrac{\sin(x)}{x} & \text{si} \quad x \neq 0 \\ 1 & \text{si} \quad x = 0 \end{cases}$$

De forma análoga podríamos operar con las demás raíces de la función ($x = n\pi$, siendo n un número entero no negativo).

A las raíces de multiplicidad 1 se las llama raíces simples. A las de multiplicidad 2 se las designa como raíces dobles, a las de multiplicidad 3 como raíces triples, etc... Pero según como se ha definido anteriormente las multiplicidad de una raíz, éstas pueden ser en general números reales no necesariamente enteros.

No obstante, el concepto de multiplicidad de una raíz que nosotros consideraremos se referirá sólo a valores enteros positivos. En este sentido, siendo m un entero positivo y $f(x)$ una función de clase $C^m(I)$ (donde I es el intervalo en el que está definida), diremos que una raíz x^* de $f(x)$ es de multiplicidad m si se verifica que:

$$f(x^*) = f'(x^*) = f''(x^*) = ... = f^{(m-1}(x^*) = 0, \quad f^{(m}(x^*) \neq 0$$

En efecto, si se verifican las condiciones anteriores, al ser f suficientemente regular, un desarrollo en serie de Taylor nos permitirá escribir que para todo $x \in I$ existe $\xi_x \in I$ tal que,

$$\begin{aligned} f(x) = f(x^*) + (x - x^*)f'(x^*) + \frac{(x - x^*)^2}{2}f''(x^*) + ...+ \\ + \frac{(x - x^*)^{(m-1)}}{(m-1)!}f^{(m-1}(x^*) + \frac{(x - x^*)^m}{m!}f^{(m}(\xi_x) = \\ = (x - x^*)^m \frac{1}{m!}f^{(m}(\xi_x) = (x - x^*)^m\varphi(x) \end{aligned}$$

donde se ha denotado por $\varphi(x)$ a la función $\varphi(x) = \frac{1}{m!}f^{(m}(\xi_x)$. Ello demuestra que las condiciones impuestas efectivamente conducen a que la raíz sea de multiplicidad m en el sentido de la definición dada inicialmente.

Observación 2.3.1 *Es importante hacer las siguientes observaciones:*

1. *Los métodos numéricos que se abordan a continuación, en general, perderán velocidad de convergencia cuando las raíces a determinar tengan multiplicidad superior a* 1. *De aquí la importancia de poder distinguir entre las raíces simples y las raíces múltiples.*

2. *Si se ha determinado una raíz x^* de multiplicidad m de la función $f(x)$ y se desean determinar otras raíces, puede calcularse la función:*

$$\varphi(x) = \begin{cases} \dfrac{1}{(x-x^*)^m} f(x) & si \quad x \neq x^* \\ \lim_{x \to x^*} \dfrac{1}{(x-x^*)^m} f(x) & si \quad x = x^* \end{cases}$$

y determinar las raíces de $\varphi(x)$ que ya no admitirá a x^* como raíz y sin embargo sí que admitirá a las demás raíces de $f(x)$ como raíces suyas.

Como ya se señaló, los métodos iterativos que analizaremos en este apartado consistirán en, a partir de un valor x_0 dado, generar una sucesión $\{x_i\}_{i=0}^{\infty}$ que converja hacia alguna solución de la ecuación.

Comencemos presentando un método muy intuitivo que nos permitirá ir asentando ideas.

2.3.1. El método de bipartición

Considérese una ecuación no lineal de la forma $f(x) = 0$ y supongamos que se conocen dos puntos a y b del dominio en el que está definida $f(x)$ tales que: $a < b$ y que en ellos $f(a)$ tiene signo contrario a $f(b)$, es decir que $f(a)f(b) < 0$. Obviamente estamos suponiendo que $f(a)$ y $f(b)$ son no nulos pues si alguno de ellos fuese nulo ya se tendría una solución de la ecuación. En estas condiciones si $f(x)$ es una función continua en $[a, b]$, por aplicación del teorema de los valores intermedios, existirá al menos un punto x^* de este intervalo en el que $f(x)$ se anule. Por ello, junto a la hipótesis de que $f(a)f(b) < 0$ supondremos también que $f \in C([a, b])$.

Observación 2.3.2 *El que exista "al menos" un punto en el que se anule $f(x)$ no quiere decir que sólo haya uno. Contando cada raíz de $f(x)$ tantas veces como sea su multiplicidad, si $f(a)f(b) < 0$, habrá en general un número impar de raíces de $f(x)$ en el intervalo $[a, b]$. Y si $f(a)f(b)$ fuese positivo o no hay ninguna raíz o habrá un número par de ellas.*

Una primera aproximación de este punto x^* puede ser el punto medio:

$$x_1 = \frac{a+b}{2}$$

Si $f(x_1) = 0$ ya se tendría calculada una raíz. Pero por lo general, salvo que se tenga mucha suerte, se tendrá que $f(x_1) \neq 0$. Pero, al haber supuesto que la función es continua, si $f(a)f(x_1) < 0$ se podrá afirmar que en el intervalo

$[a, x_1]$ habrá al menos una solución de la ecuación. Y si $f(a).f(x_1) > 0$ se verificará que $f(x_1)f(b) < 0$ lo que nos indicaría que en el intervalo $[x_1, b]$ existirá al menos una raíz. Por tanto se habrá definido así un nuevo intervalo $[a_1, b_1]$ en el que existirá una solución. A él puede aplicársele nuevamente el proceso anterior.

En general, partiendo de un intervalo $[a_j, b_j]$ en el que $f(a_j)f(b_j) < 0$ se denotará por x_{j+1} al punto medio del intervalo:

$$x_{j+1} = \frac{a_j + b_j}{2}$$

procediendo a continuación de la forma siguiente:

a) Si $f(x_{j+1}) = 0$ se habrá obtenido una solución de la ecuación: el punto x_{j+1}.

b) Si $f(a_j)f(x_{j+1}) < 0$ se denotará por: $a_{j+1} = a_j$ y por $b_{j+1} = x_{j+1}$.

c) Si $f(a_j)f(x_{j+1}) > 0$ se denotará por: $a_{j+1} = x_{j+1}$ y por $b_{j+1} = b_j$.

Al nuevo intervalo $[a_{j+1}, b_{j+1}]$ se le vuelve a aplicar el mismo proceso.

El problema que se nos puede plantear es: y si ningún valor $f(x_j)$ (j = 1, 2, ...) tiene la gracia de anularse ¿cuándo se detiene el proceso iterativo?. La respuesta a esta pregunta dependerá de la precisión con la que se desee obtener la aproximación de la solución buscada. En efecto, si se parte de un intervalo $[a, b]$ la longitud del mismo es $|b - a|$. En la primera iteración se obtendrá un intervalo $[a_1, b_1]$ cuya longitud será la mitad del anterior, es decir $\frac{|b-a|}{2}$. A su vez, en la segunda iteración se obtendrá un intervalo $[a_2, b_2]$ de longitud mitad que el anterior, es decir $\frac{|b-a|}{2^2}$. Siguiendo este proceso, en la j-ésima iteración se obtendrá un intervalo $[a_j, b_j]$ cuya longitud será $\frac{|b-a|}{2^j}$. Si se tomara como aproximación de la solución x^* existente en dicho intervalo el punto medio x_{j+1} es evidente que

$$|x_{j+1} - x^*| \leq \frac{|b_j - a_j|}{2} = \frac{|b - a|}{2^{(j+1)}}.$$

Por tanto, si se desease estar seguro de que la distancia de la aproximación x_{j+1} a la raíz x^* fuese inferior a un determinado valor ε deberían realizarse un número j de iteraciones tal que:

$$\frac{|b-a|}{2^{(j+1)}} < \varepsilon \quad \Rightarrow \quad (j+1)\log(2) > \log\left(\frac{|b-a|}{\varepsilon}\right) \quad \Rightarrow \quad j > \frac{\log\left(\frac{|b-a|}{\varepsilon}\right)}{\log(2)} - 1$$

Observación 2.3.3 *Obsérvese que el número de iteraciones anterior aseguraría la precisión deseada pero que ello no impide el que algún valor x_i calculado en alguna iteración anterior pueda estar igual de cerca (o más) de la solución buscada.*

Todo el proceso que hasta aquí se acaba de describir podría sintetizarse en el siguiente algoritmo:

Algoritmo del método de bipartición:

Dada la ecuación $f(x) = 0$, el indicador de precisión ε y dos puntos a y b en los que $f(a)f(b) < 0$,

1. Estimar el menor número natural N tal que:

$$N > \frac{\log\left(\frac{|b-a|}{\varepsilon}\right)}{\log(2)} - 1$$

2. **Para** $j = 1$, **hasta** $j = N$, **con paso** 1, **hacer:**

$$x_j \leftarrow \frac{a+b}{2}$$

Si $(f(x_j) = 0)$ **entonces:**

tomar x_j como raíz x^* y finalizar el proceso

si no:

Si $(f(x_j)f(a) < 0)$ **entonces**:

$b \leftarrow x_j$

si no:

$a \leftarrow x_j$

fin condición.

fin condición.

Fin bucle en j.

$x^* \approx \frac{a+b}{2}$

Fin del algoritmo.

El algoritmo anterior nos permitirá encontrar, en las condiciones que garantizan el éxito del proceso (esto es que $f(x)$ sea continua), una de las raíces

existentes en $[a,b]$. ¿'Pero qué condiciones nos podrían garantizar que dicha raíz es la única existente en $[a,b]$?. Una posibilidad para ello podría ser el que la función $f(x)$ fuese estrictamente monótona en el intervalo $[a,b]$. En otros términos, el proceso anterior nos demuestra el siguiente resultado:

Proposición 2.3.1 *Si la función $f(x)$ es continua y estríctamente monótona en el intervalo $[a,b]$ y además es tal que $f(a)f(b) < 0$, dado un valor real positivo ε y denotando por N al menor número natural tal que:*

$$N > \frac{\log\left(\frac{|b-a|}{\varepsilon}\right)}{\log(2)} - 1$$

se verifica que N iteraciones del proceso de bipartición antes descrito conducen a un valor x_{N+1} que dista de la única solución existente en el intervalo $[a,b]$ de la ecuación $f(x) = 0$ una magnitud inferior a ε.

Ilustremos el método descrito con un ejemplo.

Ejemplo 2.3.1 *(Cortesía del Pr. J. Aguado): La presión de vapor del n-hexano y del n-octano se pueden relacionar con la temperatura a la que se encuentren mediante las siguientes expresiones:*

$$\log(P^0_{C_6}) = 15{,}8737 - \frac{2697{,}55}{T - 48{,}784}$$

$$\log(P^0_{C_8}) = 15{,}9798 - \frac{3127{,}60}{T - 63{,}633}$$

donde la presión P^0_i está dada en milímetros de mercurio y la temperatura T en grados Kelvin. Ello nos permite estimar la temperatura de ebullición del n-hexano a 2 atmósferas (1520 mm Hg) en 364.39$^o K$ y la del n-octano a la misma presión en 425.07$^o K$. Se desea conocer, también a la presión de 2 atmósferas, la temperatura de ebullición de una mezcla líquida que contenga un 50 % en moles de ambos componentes.

Para ello, denotando por x_1 a la fracción molar en la fase líquida de n-hexano y por x_2 a la fracción molar del n-octano se tendrá que $x_1 = x_2 = 0{,}5$. Puesto que el vapor estará en equilibrio, siendo P su presión total (1520 mm Hg) y designando por y_1 e y_2 a las fracciones de cada componente en el vapor se tendrá que:

$$y_1 = \frac{P^0_1}{P} x_1 = \frac{P^0_1}{2P}$$

$$y_2 = \frac{P^0_2}{P} x_2 = \frac{P^0_2}{2P}$$

debiendo verificarse que:

$$y_1 + y_2 = 1 \quad \Leftrightarrow \quad \frac{P_1^0}{2P} + \frac{P_2^0}{2P} = 1$$

lo que, reemplazando P_1^0 y P_2^0 por sus expresiones en función de la temperatura, nos conduce a la ecuación no lineal:

$$f(T) = \frac{e^{15{,}8737 - \frac{2697{,}55}{T - 48{,}784}}}{3040} + \frac{e^{15{,}9798 - \frac{3127{,}60}{T - 63{,}633}}}{3040} - 1 = 0$$

La temperatura de ebullición T^* de la mezcla será superior a la temperatura de ebullición del n-hexano puro (364.39^{o}K) e inferior a la del n-octano puro (425.07^{o}K). Por ello un intervalo natural en el que buscar la solución puede ser [364, 425]. En este intervalo se verifica que $f(T)$ es una función continua (es suma de exponenciales cuyos exponentes están bien definidos en el intervalo de trabajo) y además es estríctamente monótona creciente (pues es la suma de funciones estríctamente monótonas crecientes).

Si en la ecuación anterior a T se le da el valor $T = 364^oK$ se tendrá que $f(364) < 0$. Análogamente si a T se le da el valor $T = 425^oK$ se tendrá que $f(425) > 0$. Por todo ello existirá una única solución de la ecuación en dicho intervalo. Si se desea encontrar esta solución con una precisión de $\varepsilon = 10^{-6}$ deberán realizarse al menos un número N de iteraciones del método de bipartición tal que:

$$N > \frac{\log\left(\frac{61}{10^{-6}}\right)}{\log(2)} - 1 \approx 24{,}862$$

es decir 25 iteraciones. En la primera iteración se tomará:

$$x_1 = \frac{364 + 425}{2} = 394{,}5$$

resultando que $f(394{,}5) = 0{,}277432 > 0$. Por tanto, la raíz buscada estará en el intervalo [364, 394.5]. En la segunda iteración se considerará:

$$x_2 = \frac{364 + 394{,}5}{2} = 379{,}25$$

resultando que $f(379{,}25) = -0{,}123283 < 0$ por lo que la solución se buscará en [379.25, 394.5]. En la tercera iteración:

$$x_3 = \frac{379{,}25 + 394{,}5}{2} = 386{,}875$$

valor para el que $f(386{,}875) = 0{,}0626451 > 0$ por lo que la solución se buscará en el intervalo [379.25, 386.875]. Posteriores iteraciones del método de bipartición nos van conduciendo a los valores: $x_4 = 383{,}0625$, $x_5 = 384{,}96...$, ..., $x_{26} = 384{,}4294930547....$ verificándose en este punto que

$$f(384{,}4294930547..) = -0{,}857630E - 08$$

Observación 2.3.4 *Las primeras referencias sobre este método de resolución de ecuaciones, aplicado a la determinación de las raíces de un polinomio de grado 3, se deben al polifacético científico belga Simon Stevin que vivió entre 1548 y 1620.*

2.3.2. El método de aproximaciones sucesivas

El método de aproximaciones sucesivas (o del punto fijo) para determinar una solución de la ecuación no lineal $f(x) = 0$ se basa en el teorema del punto fijo demostrado en la sección anterior (teorema (2.2.1)). Para ello el primer paso que se realiza en este método consiste en reescribir la ecuación $f(x) = 0$ en la forma $x = g(x)$.

Adviértase que existen múltiples posibilidades para transformar la ecuación $f(x) = 0$ en otra del tipo $x = g(x)$. Por ejemplo podría despejarse (de la forma que sea) x de la expresión de la ecuación $f(x) = 0$. O podrá sumarse la variable x en ambos lados de la ecuación y designar por $g(x)$ a $(f(x) + x)$:

$$0 = f(x) \Leftrightarrow x = f(x) + x = g(x)$$

O, siendo $\alpha \neq 0$ podrá realizarse el proceso:

$$0 = f(x) \Leftrightarrow x = \alpha f(x) + x = g(x)$$

O bien, siendo $\alpha \neq 0$ y $\beta \neq 0$:

$$0 = f(x) \Leftrightarrow \alpha x = \alpha x + \beta f(x) \Leftrightarrow x = \frac{\alpha x + \beta f(x)}{\alpha} = g(x)$$

O bien:

$$0 = f(x) \Leftrightarrow x^k = \alpha f(x) + x^k \Leftrightarrow x = \sqrt[k]{\alpha f(x) + x^k} = g(x)$$

O por ejemplo:

$$0 = f(x) \Leftrightarrow \cos(x) = f(x) + \cos(x) \Leftrightarrow x = \arccos(f(x) + \cos(x))$$

Y muchas otras opciones serían posibles. No obstante, no debe confundirse el hecho de que sea posible considerar múltiples formas de rescribir la ecuación en la forma $x = g(x)$ con el que sea indiferente la forma de hacerlo. En efecto la elección de la función $g(x)$ no es independiente de la eficacia del método pudiéndose formar funciones $g(x)$ que no estén definidas en parte del intervalo en el que se vaya a trabajar, o que no sean aplicaciones, o que no sean continuas, o que .....no sean contracciones. Desde luego no tendrán el mismo comportamiento unas que otras. Pero dejemos para un poco más adelante el análisis sobre cuales son las "buenas" funciones $g(x)$ que nos interesan. En todo

caso, una vez rescrita la ecuación $f(x) = 0$ en la forma $x = g(x)$ el método de aproximaciones sucesivas busca un punto fijo de la aplicación $g(x)$ mediante el **esquema de cálculo** siguiente:

Dado un valor x_0 se genera la sucesión $\{x_{i+1} = g(x_i)\}_{i=0}^{\infty}$.

Según lo visto en la sección anterior se tiene el siguiente resultado:

Teorema 2.3.1 *Si $g(x)$ es una contracción definida sobre un intervalo $[a, b]$ entonces el método de aproximaciones sucesivas que se acaba de describir genera, a partir de cualquier valor inicial $x_0 \in [a, b]$, una sucesión $\{x_i\}_{i=0}^{\infty}$ que converge hacia la única solución de la ecuación $x = g(x)$ en el intervalo $[a, b]$.*

Demostración: En virtud del teorema del punto fijo, por ser $g(x)$ una contracción definida sobre el espacio métrico completo $([a, b], d_f)$ admitirá un único punto fijo x^* que será el límite de la sucesión $\{x_i\}_{i=0}^{\infty}$.

c.q.d.

Observación 2.3.5 *Es necesario hacer las siguientes observaciones:*

1. *Puesto que la ecuación $f(x) = 0$ es equivalente a $x = g(x)$, en las condiciones del teorema anterior, x^* será solución en $[a, b]$ de la ecuación equivalente $f(x) = 0$.*

2. *En otros términos las buenas funciones $g(x)$ que nos interesan son aquellas que sean contracciones sobre un determinado intervalo $[a, b]$ en el que se buscará la única solución en él existente. Además, como se justificó en las notas realizadas al teorema del punto fijo, cuanto menor sea la constante de Lipschitz de la contracción $g(x)$ más rápidamente convergerá el método hacia la solución.*

3. *Interprétese bien el teorema anterior. En él se asegura que bajo ciertas hipótesis (el ser $g(x)$ una contracción en $([a, b], d_f)$) el método de aproximaciones sucesivas nos conduce a la única solución existente en $[a, b]$ de la ecuación $f(x) = 0$. Pero no se impide que el método funcione si no se verifican las hipótesis. Simplemente no se asegura su buen funcionamiento.*

El demostrar que una aplicación $g(x)$ es una contracción mediante la determinación de su constante de Lipschitz puede, en ciertas ocasiones, resultar algo laborioso. Por ello pueden contemplarse variantes más restrictivas (pero más fácilmente aplicables en la práctica) del teorema anterior. Un ejemplo de ello es el siguiente teorema:

Teorema 2.3.2 *Si $g(x)$ es una aplicación de clase $C^1([a,b])$, que toma valores en $[a,b]$ y verificando la condición:*

$$\exists k < 1 \ / \ |g'(x)| \leq k, \qquad \forall x \in [a,b]$$

entonces la sucesión $\{x_i\}_{i=0}^{\infty}$ generada, a partir de cualquier $x_0 \in [a,b]$, converge hacia la única solución de la ecuación $x = g(x)$ en $[a,b]$.

Demostración: Por aplicación del teorema del valor medio se verificará que:

$$\forall x, y \in [a,b] \quad \exists\, z \in]a,b[\,/ \quad g(x) - g(y) = g'(z)(x-y)$$

y por haber supuesto que la primera derivada estaba acotada en valor absoluto se tendrá que:

$$\forall x, y \in [a,b] : \quad |g(x) - g(y)| \leq k\,|x-y| < |x-y|$$

por lo que, teniendo en cuenta que $g : [a,b] \to [a,b]$, resulta que $g(x)$ es una contracción. Aplicando el teorema precedente quedará totalmente demostrado este.

c.q.d.

Observación 2.3.6 *De hecho una aplicación g de clase $C^1([a,b], I\!R)$ sólo puede contracción si $|g'(x)| < 1$. Pero puede haber aplicaciones de clase C^0 que sean contracciones y que en algún punto no admitan derivadas. De aquí el hecho de que el teorema anterior (2.3.2) sea algo más restrictivo que el teorema (2.3.1).*

Observación 2.3.7 *Cuando en las aplicaciones se utilice este teorema para comprobar que la aplicación considerada es una contracción se tomará como aproximación de la constante de Lipschitz el valor $k = \max\limits_{x \in [a,b]} \left\{ |g'(x)| \right\}$.*

Los dos teoremas precedentes establecen condiciones suficientes de convergencia global del método sobre un intervalo $[a,b]$ (esto es independientemente del punto $x_0 \in [a,b]$ con el que se arranque el proceso iterativo). Cuando se conozca un cierto entorno de la solución buscada pueden establecerse resultados de convergencia local (esto es para valores de x_0 suficientemente cercanos a la solución). Así por ejemplo se tiene el siguiente teorema:

Teorema 2.3.3 *Si existe una solución x^* de la ecuación $x = g(x)$ en un intervalo $[a,b]$ en el que $g(x)$ es de clase $C^1([a,b])$ y $|g'(x^*)| < 1$ entonces existe un valor $\delta > 0$ tal que si $|x^* - x_0| < \delta$ la sucesión $\{x_{i+1} = g(x_i)\}_{i=0}^{\infty}$ verifica que:*

a) $|x^* - x_i| < \delta$, $\forall\, x_i$,

b) $\lim_{i\to\infty} x_i = x^*$.

Demostración: Por ser $g'(x)$ continua en todo $x \in [a,b]$ existirá un intervalo abierto de centro x^* y radio δ' tal que en él se verifique:

$$|g'(x)| \leq k < 1 \qquad \forall x \in]x^* - \delta', x^* + \delta'[$$

Considerando un valor $\delta < \delta'$ se tendrá por tanto que:

$$|g'(x)| \leq k < 1 \qquad \forall x \in [x^* - \delta, x^* + \delta]$$

Además, al ser g de clase $C^1([x^* - \delta, x^* + \delta])$, un desarrollo en serie de Taylor nos conduce a que:

$$\forall x \in [x^* - \delta, x^* + \delta] \quad \exists \xi_x \in [x^* - \delta, x^* + \delta] \text{ tal que}$$

$$g(x) = g(x^*) + (x - x^*)g'(\xi_x)$$

luego $\forall x \in [x^* - \delta, x^* + \delta]$ se tiene que

$$|x^* - g(x)| \leq |x - x^*| \leq \delta$$

lo que demuestra que $\forall x \in [x^* - \delta, x^* + \delta]$ la imagen $g(x)$ también pertenece al intervalo $[x^* - \delta, x^* + \delta]$. Consecuentemente $g(x)$ es una contracción en $[x^* - \delta, x^* + \delta]$. Ello conduce a que: $\forall x_i \in \{x_i\}_{i=1}^{\infty}$ se tiene que

$$|x_i - x^*| = |g(x_{i-1}) - g(x^*)| \leq k|x_{i-1} - x^*| \leq$$

$$\leq k^2|x_{i-2} - x^*| \leq ..... \leq k^i|x_0 - x^*| \leq k^i\delta$$

Al ser $k < 1$ bastará con escoger el índice i suficientemente elevado para que todos los elementos de la sucesión con índice mayor que i sean tan cercanos a x^* como se desee. En otros términos

$$x^* = \lim_{x\to\infty} x_i.$$

c.q.d.

Observación 2.3.8 *Cuanto menor sea el valor de $|g'(x^*)|$ menor será la cota de $|x_i - x^*|$ obtenida en la demostración anterior y por ello mejor (en el sentido de más rápida) será la convergencia del método si se parte de un punto suficientemente cercano a la solución.*

Los teoremas precedentes establecen condiciones suficientes para que el método de aproximaciones sucesivas converja. De esta forma, si se verifican las hipótesis de cualquiera de los teoremas anteriores, seleccionado el punto inicial x_0, todo consistirá en generar a partir de él $x_1 = g(x_0)$, y a partir de este $x_2 = g(x_1)$, y así sucesivamente. Tras hacer infinitas iteraciones alcanzaríamos la solución buscada. Pero, evidentemente, no pueden realizarse "infinitas" iteraciones. Por ello la cuestión que nos planteamos ahora es ¿cuántas iteraciones nos garantizarían una precisión determinada?. La respuesta a este dilema nos la proporciona el siguiente teorema:

Teorema 2.3.4 *Siendo $g(x)$ una contracción definida en el intervalo $[a,b]$ la distancia entre la única solución x^* de la ecuación $x = g(x)$ y cualquier elemento de la sucesión $\{x_n = g(x_{n-1})\}_{n=0}^{\infty}$, generada a partir de cualquier valor $x_0 \in [a,b]$, está acotada mediante la expresión:*

$$|x^* - x_n| \leq \frac{k^n}{1-k} |x_1 - x_0|$$

donde k es la constante de Lipschitz de la contracción.

Demostración: Véase la segunda observación realizada tras la demostración del teorema del punto fijo (2.2.1).

c.q.d.

Observación 2.3.9 *Es importante notar que:*

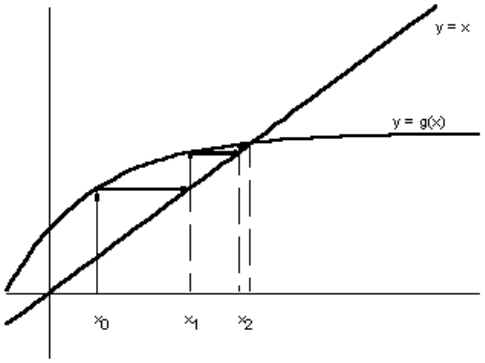
1. *Bajo las hipótesis del teorema precedente, si se desea asegurar que el error cometido es menor que un cierto valor ε la expresión anterior nos conduce que deben realizarse un número N de iteraciones tal que:*

$$\frac{k^N}{1-k} |x_1 - x_0| < \varepsilon \Rightarrow N > \left| \frac{\log\left(\frac{\varepsilon(1-k)}{|x_1 - x_0|}\right)}{\log(k)} \right|$$

2. *Si no se conoce el valor exacto de la constante de Lipschitz de la aplicación puede estimarse de forma aproximada de diferentes formas. Por ejemplo, tras cada iteración del método podría obtenerse una aproximación de dicha constante mediante:*

$$k \approx |g'(x_i)| \approx \frac{|g(x_{i-1}) - g(x_{i-2})|}{|x_{i-1} - x_{i-2}|} = \frac{|x_i - x_{i-1}|}{|x_{i-1} - x_{i-2}|}$$

3. *Una interpretación gráfica del método consiste simplemente en buscar la intersección entre la bisectriz del primer cuadrante y la contracción $g(x)$ mediante sucesivos "escalones" comprendidos entre la gráfica de $g(x)$ y la bisectriz del primer cuadrante. Es decir:*



En la práctica, en lugar de calcular a priori el número de iteraciones a realizar se va estimando en cada iteración la distancia del valor en ella hallado a la solución exacta. Esta estimación se realiza simplemente evaluando la diferencia entre las dos últimas aproximaciones halladas que, cuando $g(x)$ es una contracción, son un indicador de la cercanía a la solución exacta en virtud del siguiente teorema:

Theorem 1 *Siendo $g(x)$ una contracción definida en el intervalo $[a,b]$ la distancia entre la única solución x^* de la ecuación $x = g(x)$ y cualquier elemento de la sucesión $\{x_n = g(x_{n-1})\}_{n=0}^{\infty}$, generada a partir de cualquier valor $x_0 \in [a,b]$, está acotada mediante la expresión:*

$$|x^* - x_n| \leq \frac{k}{1-k} \left| x_n - x_{(n-1)} \right|$$

donde k es la constante de Lipschitz de la contracción.

Demostración: Véase la observación (2.2.8), apartado c), realizada tras la demostración del teorema del punto fijo (2.2.1).

c.q.d.

Con ello, cuando $g(x)$ sea una contracción, al ser $k < 1$, bastará con hacer un número de iteraciones tal que $\left| x_n - x_{(n-1)} \right|$ sea suficientemente pequeño para asegurar que $|x^* - x_n|$ también es pequeño. Este control de la convergencia debe acompañarse con la limitación del número de iteraciones a realizar en previsión de los casos en los que, no siendo $g(x)$ una contracción, el método no converja. Más concretamente un algoritmo del método de aproximaciones sucesivas, en el que se parte de la ecuación equivalente $x = g(x)$ es el siguiente:

Algoritmo del método de aproximaciones sucesivas:

Dada la ecuación $x = g(x)$, el indicador de precisión ε , un valor máximo del número de iteraciones que se permiten realizar ($maxiter$) y un punto x_0 con el que inicializar el proceso,

$tol \leftarrow 2\varepsilon$

$iteración \leftarrow 0$

Mientras ($(iteración < maxiter)$ **y** $(tol > \varepsilon)$ **), hacer:**

$$x_1 \leftarrow g(x_0)$$

$$tol \leftarrow |x_1 - x_0|$$

$$iteración \leftarrow iteración + 1$$

$$x_0 \leftarrow x_1$$

Fin bucle condicional.

Si $(tol < \varepsilon)$ **entonces:**

tomar x_1 como solución

si no:

Escribir un mensaje de error en el proceso de cálculo

fin condición.

Fin del algoritmo.

Ilustremos el método que se acaba de describir mediante un ejemplo.

Ejemplo 2.3.2 *(Cortesía del Pr. J. Arsuaga): La expresión de Plank proporciona la densidad de energía radiada u (energía por unidad de volumen) por un emisor perfecto (un cuerpo negro) que se encuentre a la temperatura absoluta T en el intervalo de frecuencias desde un valor ϑ hasta $\vartheta + \delta\vartheta$ mediante:*

$$u(\vartheta, T) = \frac{8\pi h}{c^3} \frac{\vartheta^3}{e^{\frac{h\vartheta}{kT}} - 1}$$

En la ecuación anterior π puede tomarse como 3.1416, h es la constante de Plank $(6{,}62610^{-34} J\,s)$*, k es la constante de Boltzmann* $(1{,}3806610^{-23} J/K)$ *y c es la velocidad de la luz en el vacío* $(c = 310^8 m/s)$*. Se desea saber la frecuencia $\vartheta > 0$ para la que, a una determinada temperatura fija $T > 0$, se hace máxima la densidad de energía emitida.*

Para ello, siendo $u(\vartheta, T) = u_T(\vartheta)$ la energía radiada a temperatura prefijada $T > 0$, es decir

$$u_T(\vartheta) = \frac{8\pi h}{c^3} \frac{\vartheta^3}{e^{\frac{h\vartheta}{kT}} - 1}$$

denotemos por M a la constante:

$$M = \frac{8\pi h}{c^3}$$

y por N a la constante:

$$N = \frac{h}{k}$$

Con esta notación resultará que:

$$u_T(\vartheta) = M \frac{\vartheta^3}{e^{N\frac{\vartheta}{T}} - 1}.$$

El extremo de esta función se alcanzará en algún punto en el que se anule su primera derivada, es decir, que para calcularlo debe resolverse la ecuación no lineal:

$$\frac{du_T}{d\vartheta}(\vartheta) = M. \frac{3\vartheta^2 \left(e^{N\frac{\vartheta}{T}} - 1\right) - \vartheta^3 \frac{N}{T} e^{N\frac{\vartheta}{T}}}{\left(e^{N\frac{\vartheta}{T}} - 1\right)^2} = 0.$$

Puesto que el denominador de la expresión anterior nunca se anula (los valores de ϑ considerados son estrictamente positivos) y la constante M tampoco es nula, las soluciones de la ecuación anterior son las mismas que las de la ecuación:

$$3\vartheta^2 \left(e^{N\frac{\vartheta}{T}} - 1\right) - \vartheta^3 \frac{N}{T} e^{N\frac{\vartheta}{T}} = 0$$

o, dividiendo esta expresión por ϑ^2, (lo cual nos eliminaría dos veces la frecuencia $\vartheta = 0$ que está descartada del conjunto de frecuencias con interés práctico) se obtiene otra ecuación con las mismas soluciones no nulas que las de la ecuación anterior:

$$3\left(e^{N\frac{\vartheta}{T}} - 1\right) - N\frac{\vartheta}{T} e^{N\frac{\vartheta}{T}} = 0$$

Llamando α a la relación: $\alpha = N\frac{\vartheta}{T}$ la ecuación anterior se transforma en:

$$3\left(e^{\alpha} - 1\right) - \alpha e^{\alpha} = 0 \Longleftrightarrow \left(3 - \alpha\right) e^{\alpha} = 3 \Leftrightarrow$$

$$\Leftrightarrow 3 - \alpha = 3e^{-\alpha} \Leftrightarrow \alpha = 3(1 - e^{-\alpha})$$

Una solución de esta ecuación, obviamente, es la solución trivial $\alpha = 0$. Esta solución nos conduciría a $\vartheta = 0$, es decir otra vez a la frecuencia nula que está descartada del conjunto de frecuencias con interés práctico. Por tanto

intentaremos buscar otras soluciones de la ecuación $\alpha = g(\alpha) = 3\,(1 - e^{-\alpha})$. Ello puede hacerse mediante el **método de aproximaciones sucesivas**. Para ello, teniendo en cuenta que N, T y ϑ son positivas podríamos pensar en ubicarnos, en principio, en el espacio métrico $([0, \infty[, d_f)$ que es un espacio métrico completo. Sin embargo en él $g(\alpha)$ no es una contracción (basta con comprobar que $g'(1) = 3/e \approx 1{,}104$. Busquemos pues un intervalo en el que $g(\alpha)$ sí sea una contracción. Puesto que:

$$g'(\alpha) = 3e^{-\alpha}$$

se verificará que:

$$0 < g'(\alpha) < 1, \qquad \forall\, \alpha > \log(3) \approx 1{,}0986$$

Por este motivo buscaremos la solución de la ecuación en $[\ln(3), \infty[$. Nótese que al ser $g(\alpha)$ una función continua monótona creciente y verificarse que $g(0) = 3(1-1) = 0$ y que

$$g(\ln(3)) = 3(1 - e^{-\ln(3)}) \approx 2 > \ln(3)$$

sólo se ha perdido la solución (inútil en la práctica) $\alpha = 0$ al descartar el intervalo $[0, \ln(3)[$ del espacio de búsqueda de las soluciones, y que además:

a) sólo habrá una solución de la ecuación $\alpha = g(\alpha)$ distinta de la solución nula,

b) la única solución existente pertenece a $[\ln(3), \infty[$,

c) el método de aproximaciones sucesivas nos va a conducir a dicha solución no nula.

Apliquemos pues el método partiendo de $\alpha_0 = 1{,}1$. Se irán obteniendo sucesivamente los valores siguientes:

$$\alpha_1 = g(\alpha_0) = 3(1 - e^{-1{,}1}) = 2{,}001386749$$

$$\begin{aligned}
\alpha_2 &= g(\alpha_1) = 3(1 - e^{-2{,}001386749}) = 2{,}594556788 \\
\alpha_3 &= g(\alpha_2) = 3(1 - e^{-2{,}594556788}) = 2{,}775963098 \\
\alpha_4 &= g(\alpha_3) = 3(1 - e^{-2{,}775963098}) = 2{,}813131625 \\
\alpha_5 &= g(\alpha_4) = 3(1 - e^{-2{,}813131625}) = 2{,}819949757 \\
\alpha_6 &= g(\alpha_5) = 3(1 - e^{-2{,}819949757}) = 2{,}821173187 \\
\alpha_7 &= g(\alpha_6) = 3(1 - e^{-2{,}821173187}) = 2{,}821391836
\end{aligned}$$

$$\alpha_8 = g(\alpha_7) = 3(1 - e^{-2,821391836}) = 2,821430884$$
$$\alpha_9 = g(\alpha_8) = 3(1 - e^{-2,821430884}) = 2,821437856$$
$$\alpha_{10} = g(\alpha_9) = 3(1 - e^{-2,821437856}) = 2,821439101$$
$$\alpha_{11} = g(\alpha_{10}) = 3(1 - e^{-2,821439101}) = 2,821439324$$
$$\alpha_{12} = g(\alpha_{11}) = 3(1 - e^{-2,821439324}) = 2,821439364$$
$$\alpha_{13} = g(\alpha_{12}) = 3(1 - e^{-2,821439364}) = 2,821439371$$
$$\alpha_{14} = g(\alpha_{13}) = 3(1 - e^{-2,821439371}) = 2,821439372$$
$$\alpha_{15} = g(\alpha_{14}) = 3(1 - e^{-2,821439372}) = 2,821439372$$

no obteniéndose diferencias de valor, con los 9 decimales que hemos utilizado en el cálculo, para posteriores iteraciones. Por tanto la solución buscada será $\alpha^* \approx 2,821439372$. A partir de este valor, puesto que habíamos denotado por $\alpha = N\frac{\vartheta}{T}$, se tendrá que la frecuencia a la que se maximiza la energía está dada por:

$$\vartheta^* \approx 2,821439372\frac{T}{N} = 2,821439372\frac{hT}{k}.$$

Observación 2.3.10 *Nótese que:*

1. *El resultado del ejercicio anterior muestra que la relación entre la frecuencia a la que se emite la máxima energía y la temperatura siempre es:*

$$\frac{\vartheta^*}{T} \approx \frac{2,821439372}{N} = \frac{2,821439372}{\frac{h}{k}} = \frac{2,8214393721,3806610^{-23}}{6,62610^{-34}} \approx$$

$$\approx 5,87910^{10} s^{-1} K^{-1}$$

2. *La anterior es una forma de expresar la llamada "ley del desplazamiento" de Wien que dice que la frecuencia a la que se produce la emisión máxima es directamente proporcional a la temperatura absoluta del cuerpo emisor.*

3. *A partir de la fórmula de Plank:*

$$u(\vartheta, T) = \frac{8\pi h}{c^3}\frac{\vartheta^3}{e^{\frac{h\vartheta}{kT}} - 1}$$

se puede obtener la ecuación de Stefan-Boltzmann (que históricamente es anterior) según la cual la potencia total radiada por unidad de superficie (a todas las frecuencias) a una determinada temperatura absoluta es directamente proporcional a la 4^a potencia de la misma, es decir:

$$S = \sigma T^4$$

donde σ es la constante de Stefan-Boltzmann. Basta para obtener esta expresión efectuar el proceso de integración: $S = \int_0^\infty u(\vartheta, T)d\vartheta$..... *pero eso es objeto de otra disciplina.*

La técnica de sobreiteración

En ocasiones la aplicación del método de aproximaciones sucesivas a la ecuación $x = g(x)$ conducirá a un proceso que converge muy lentamente (por tener su constante de Lipschitz próxima a 1) o que no converge. En esas ocasiones será conveniente modificar la ecuación equivalente convirtiéndola en otra de la forma $x = h(x)$ en la que $h(x)$ tenga mejores propiedades de cara a la convergencia del método hacia la solución x^*. Una estrategia que en ocasiones nos puede ayudar en este proceso consiste en modificar la ecuación de la forma:

$$x = g(x) \Leftrightarrow x + \rho x = g(x) + \rho x \Leftrightarrow x = \frac{g(x) + \rho x}{1 + \rho} = h(x)$$

Se dispone así de un parámetro ρ con el que intentar mejorar la velocidad de convergencia del algoritmo de aproximaciones sucesivas. Ello se podría lograr, en virtud del teorema de convergencia local antes presentado, dando a ρ el valor de $-g'(x^*)$ (cuando este sea no nulo) pues en ese caso:

$$h'(x^*) = \frac{g'(x^*) + (-g'(x^*))}{1 - g'(x^*)} = 0$$

La dificultad de este proceso, conocido con el nombre de técnica de sobreiteración, radica en estimar el valor de $g'(x^*)$ .... sin conocer x^*. No obstante, aunque parezca increíble, en ocasiones esto podrá hacerse. Como botón de muestra sirva el siguiente ejemplo.

Ejemplo 2.3.3 *Determinar la solución de la ecuación no lineal* $x^2 - a = 0$ *siendo* $a \in \mathbb{R}$, $a > 0$.

Si se desea conocer la solución de la ecuación no lineal $x^2 - a = 0$ siendo a un número estrictamente positivo puede procederse, en primer lugar de la siguiente forma

$$x^2 - a = 0 \Leftrightarrow xx = a \Leftrightarrow x = \frac{a}{x} = g(x)$$

No obstante, esta función $g(x)$ nos sirve de poco para calcular la raíz pues si se parte de un valor $x_0 \neq \sqrt{a}$ se tendrá que:

$$x_1 = g(x_0) = \frac{a}{x_0}; \quad x_2 = g(x_1) = \frac{a}{\frac{a}{x_0}} = x_0; \quad x_3 = g(x_2) = \frac{a}{x_0}; ....$$

es decir, que la sucesión que nos proporciona el método de aproximaciones sucesivas será:

$$\left\{x_0, \frac{a}{x_0}, x_0, \frac{a}{x_0}, x_0, \frac{a}{x_0}, ...\right\}$$

que como se ve no converge hacia nada.

No obstante en este caso se tendrá que:

$$g'(x^*) = -\frac{a}{(x^*)^2}$$

y como en la solución buscada se verificará que: $(x^*)^2 = a$ resultará que:

$$g'(x^*) = -\frac{a}{a} = -1$$

Por tanto puede intentarse el método de sobreiteración tomando como valor

$$\rho = -g'(x^*) = 1.$$

Con ello la ecuación se transformará en:

$$x = h(x) = \frac{\frac{a}{x} + x}{1+1} = \frac{a + x^2}{2x} = \frac{1}{2}(x + \frac{a}{x})$$

Así si por ejemplo, se considera que $a = 16$ el esquema anterior, partiendo de $x_0 = 1$ nos conduce a que:

$$x_1 = h(x_0) = \frac{17}{2} = 8{,}5$$

$$x_2 = h(x_1) = \frac{88{,}25}{17} = 5{,}191176470$$

$$x_3 = 4{,}136664722$$

$$x_4 = 4{,}002257525$$

$$x_5 = 4{,}000000637$$

$$x_6 = 4{,}000000001$$

$$x_6 = 4{,}000000000$$

Observación 2.3.11 *Este procedimiento de cálculo de raíces cuadradas es atribuido a Herón de Alejandría (arquitecto e ingeniero que vivió en la segunda mitad del siglo I) y se conoce con el nombre de regla de Herón o regla mecánica para el cálculo de raíces cuadradas. A pesar de llevar el nombre de Herón, pues la regla aparece recogida por primera vez en su obra "Métrica", se cree que este método es debido en realidad a Arquímedes de Siracusa (ingeniero y matemático del siglo III antes de Cristo). A pesar de su antigüedad, este método es empleado actualmente en numerosas calculadoras científicas debido a su gran velocidad de convergencia hacia el valor de la raíz cuadrada de cualquier número real positivo a.*

2.3.3. El método de Newton-Raphson

Considérese la ecuación $f(x) = 0$ en la que supongamos que $f(x)$ es una función de clase $C^2([a,b])$. Supongamos además que la ecuación anterior admite una solución x^* en el intervalo $[a,b]$. Para cualquier otro valor $x_0 \in [a,b]$, denotando por h al valor tal que $x^* = x_0 + h$, la expresión del desarrollo en serie de Taylor nos permitiría escribir que:

$$0 = f(x^*) = f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0 + \theta h), \qquad \theta \in [0,1].$$

Si conocido x_0 se fuese capaz de determinar h resolviendo la ecuación:

$$f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0 + \theta h) = 0$$

podría evaluarse x^* como $x^* = x_0 + h$. Pero para resolver esta ecuación primero deberíamos conocer el valor de θ (lo cual no es obvio) y una vez conocido resolver una ecuación, en general, no lineal pues obsérvese que h interviene en la expresión $f''(x_0 + \theta h)$. Por tanto, salvo en situaciones muy particulares, no se ganaría gran cosa reemplazando el problema de resolver $f(x) = 0$ por el de resolver

$$F(h) = f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0 + \theta h) = 0.$$

El **método de Newton-Raphson** (o método de linealización de Newton) se sustenta en simplificar la expresión anterior linealizándola. Para ello considera que si se está suficientemente cerca de la solución (es decir, si h es suficientemente pequeño) el término $\frac{h^2}{2} f''(x_0 + \theta h)$ podrá despreciarse frente a los otros términos de la ecuación. Por ello resuelve la ecuación lineal:

$$f(x_0) + Hf'(x_0) = 0$$

de la que se obtiene que:

$$H = -\frac{f(x_0)}{f'(x_0)}$$

Obviamente, al ser diferente la ecuación linealizada que la proporcionada por el desarrollo de Taylor, se tendrá que $H \neq h$ y por tanto

$$x^* = x_0 + h \neq x_1 = x_0 + H.$$

De una forma intuitiva (que más adelante deberemos precisar cuándo es correcta) puede pensarse que aunque x_1 sea diferente de x^* será un valor más próximo a x^* que x_0 pues lo hemos obtenido "aproximando" el valor h que nos llevaba de x_0 a x^*. Ello, al menos, será así cuando h sea suficientemente pequeño, es decir cuando x_0 sea suficientemente próximo a x^*. Con ello el método

de Newton-Raphson propone repetir este proceso de forma recursiva hasta estar lo suficientemente cercanos a la solución buscada. Más concretamente el **método de Newton-Raphson** consiste en:

$$\textit{Dado un valor } x_0\textit{, generar la sucesión } \left\{ x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \right\}_{i=0}^{\infty}.$$

Observación 2.3.12 *Un poco de historia sobre el método. Esta idea para aproximar las raíces de las ecuaciones tiene sus antecedentes en trabajos anteriores a Newton, debidos al matemático francés François Viète que vivió entre 1540 y 1603. El método se recogió por primera vez en la obra "Algebra" del matemático inglés John Wallis aparecida en 1685. No obstante el prolífico matemático y físico inglés Sir Isaac Newton la había aplicado ya a la resolución de algunas ecuaciones no lineales en escritos suyos anteriores partiendo de un valor suficientemente próximo de la raíz y calculando, en una primera aproximación, el incremento que le acercaba más a la raíz, en una segunda aproximación, el incremento del incremento que mejoraba la aproximación anterior, en una tercera aproximación el incremento del incremento del incremento, y así sucesivamente. Así, Newton aplica el método a la ecuación*

$$x^3 - 2x - 5 = 0 \qquad (1)$$

y considera como valor $x_0 = 2$, que supone que dista menos de la unidad de una solución exacta, suponiendo entonces que una raíz de esta ecuación es $x^* = (2 + \delta x)$. Sustituyendo esta expresión en la propia ecuación obtiene:

$$(\delta x)^3 + 6\,(\delta x)^2 + 10(\delta x) - 1 = 0 \qquad (2)$$

Al haber supuesto que $\delta x < 1$, Newton desprecia los términos en $(\delta x)^3$ y en $(\delta x)^2$ aproximando $\delta x \approx 0{,}1$. Tras ello considera que $\delta x = 0{,}1 + \delta^2 x$ con lo que inyectada esta expresión en (2) se obtiene que:

$$\left(\delta^2 x\right)^3 + 6{,}3\left(\delta^2 x\right)^2 + 11{,}23(\delta^2 x) + 0{,}061 = 0 \qquad (3)$$

de donde despreciando los términos en $\left(\delta^2 x\right)^3$ y en $\left(\delta^2 x\right)^2$ obtiene que $\delta^2 x \approx -0{,}0054$ por lo que $\delta x \approx 0{,}1 - 0{,}0054$. Considera entonces que $\delta^2 x = -0{,}054 + \delta^3 x$ y vuelve a repetir el proceso anterior, sustituyendo esta expresión en (3) y obteniendo una aproximación de $\delta^3 x$. Procediendo de forma iterativa Newton obtiene una aproximación aceptable de la raíz.

En 1690 en una publicación del matemático inglés Joseph Raphson, en la que no menciona a Newton, se describe el método de una forma más próxima a cómo se utiliza hoy en día: actualizando el valor de la aproximación de la raíz y calculando un nuevo incremento para esta actualización. Concretamente, sobre la misma ecuación no lineal Raphson realizaría la primera iteración considerando que $x^* = (2 + \Delta_1 x)$ y obteniendo, al igual que Newton, $\Delta_1 x = 0{,}1$

por lo que llama $x_1 = 2{,}1$ *Tras ello supone que* $x^* = 2{,}1 + \Delta_2 x$ *y procede sustituyendo* $(2{,}1 + \Delta_2 x)$ *en la ecuación (1), despreciando los términos cuadráticos y cúbicos y estimando que* $\Delta_2 x = -0{,}0054$ *y continuando así el proceso.*

Como puede apreciarse ambas formas de proceder son equivalentes, pero, operacionalmente, es más sencilla la forma en que Raphson modifica la técnica propuesta por Newton. Pero también puede observarse que en ambas formas de proceder no aparece explícitamente la primera derivada de la función que define la ecuación. Téngase en cuenta que es por esta misma época cuando el propio Newton en Inglaterra y el gran matemático Gottfried Wilhelm Leibniz en el continente europeo están asentando las bases del cálculo infinitesimal. Además, por supuesto que la forma de obtener el método tanto por Newton como por Raphson fue más heurística (pues hasta el año 1715 en la publicación del matemático inglés Brook Taylor "Methodus incrementorum directa e inversa" no se presentó el desarrollo en serie que lleva su nombre). Hubo que esperar más de un siglo, hasta 1818, para que el matemático francés Joseph Fourier comenzase el análisis de las condiciones de convergencia del método y acabase de formularlo en la forma en que es utilizado hoy en día.

Sobre este método, en primer lugar, puede observarse que si denotamos por:

$$g(x) = x - \frac{f(x)}{f'(x)}$$

estamos en presencia de un caso particular del método de aproximaciones sucesivas antes contemplado. En otros términos, se tienen las siguientes propiedades:

Proposición 2.3.2 *Si la función* $g(x) = x - \dfrac{f(x)}{f'(x)}$ *es una contracción definida en* $[a, b]$ *la sucesión dada por*

$$\left\{ x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \right\}_{i=0}^{\infty}$$

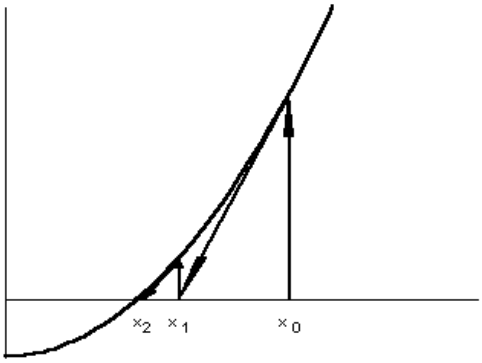
obtenida a partir de cualquier punto $x_0 \in [a, b]$ *converge hacia la única solución de la ecuación* $f(x) = 0$ *en* $[a, b]$.

Demostración: Es un caso particular de los teoremas de convergencia del método de aproximaciones sucesivas.

c.q.d.

Proposición 2.3.3 *Si la función* $g(x) = x - \frac{f(x)}{f'(x)}$ *definida en* $[a, b]$ *toma valores en* $[a, b]$, *es de clase* $C^1([a, b])$ *y además:*

$$|g'(x)| = \left| \frac{f''(x) f(x)}{(f'(x))^2} \right| < 1 \qquad \forall x \in [a, b]$$



entonces la sucesión dada por

$$\left\{ x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \right\}_{i=0}^{\infty}$$

obtenida a partir de cualquier punto $x_0 \in [a,b]$ *converge hacia la única solución de la ecuación* $f(x) = 0$ *en* $[a,b]$.

Demostración: Es un caso particular del teorema 1.8. de convergencia del método de aproximaciones sucesivas.

c.q.d.

Observación 2.3.13 *Obsérvese lo siguiente:*

1. *Que la aplicación del método de Newton-Raphson exige que los valores de* $f'(x_i)$ *no se anulen.*

2. *Una interpretación gráfica del método puede obtenerse teniendo en cuenta que* $f'(x_i)$ *geométricamente representa la tangente trigonométrica de la recta tangente a la gráfica de* $f(x)$ *en el punto* $(x_i, f(x_i))$ *por lo que* $\left|f(x_i)/f'(x_i)\right|$ *será la distancia existente entre* x_i *y el punto de corte de la recta tangente a* $f(x)$ *en* $(x_i, f(x_i))$ *con el eje de abscisas. Es decir que las primeras iteraciones del proceso se pueden representar de la forma en que se recoge en la figura siguiente:*

Al igual que se hizo con el método de aproximaciones sucesivas, las condiciones que garantizan la convergencia global del método pueden ser sustituidas por otras que garantizan su convergencia local (esto es si el punto x_0 con el que se inicializa el método es suficientemente cercano a la solución buscada). Con ello se pueden rebajar las "exigencias" sobre la función que garanticen el correcto funcionamiento del método. En concreto es de aplicación a este método el siguiente teorema:

Teorema 2.3.5 *Si $f \in C^2[a,b]$ y x^* es una solución de la ecuación $f(x) = 0$ en la que $f'(x^*) \neq 0$ entonces existe un valor $\delta > 0$ tal que la sucesión $\left\{x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}\right\}_{i=0}^{\infty}$ generada a partir de cualquier punto $x_0 \in [x^* - \delta, x^* + \delta]$ converge hacia x^*.*

Demostración: Por ser $f'(x)$ continua en x^* existirá un intervalo $[x^* - \delta_1, x^* + \delta_1]$ en el que $f(x) \neq 0, \forall\, x \in [x^* - \delta_1, x^* + \delta_1]$. Por tanto la aplicación:

$$g(x) = x - \frac{f(x)}{f'(x)}$$

también estará definida y será continua en $[x^* - \delta_1, x^* + \delta_1]$ y verificará que $x^* = g(x^*)$. Además como:

$$g'(x) = \frac{f(x)f''(x)}{(f'(x))^2}$$

y se ha supuesto que $f(x) \in C^2([a,b])$ resultará que $g(x) \in C^1([x^* - \delta_1, x^* + \delta_1])$. Y como $f(x^*) = 0$ se tendrá que:

$$g'(x^*) = \frac{f(x^*)f''(x^*)}{(f'(x^*))^2} = 0$$

Luego se tiene una aplicación $g(x)$ de clase $C^1([x^* - \delta_1, x^* + \delta_1])$ y tal que $|g'(x^*)| = 0 < 1$. Por tanto, al ser $g'(x)$ continua en x^* para cualquier valor $k < 1$ existirá un valor $0 < \delta \leq \delta_1$ tal que:

$$|g'(x)| \leq k < 1, \qquad \forall\, x \in [x^* - \delta, x^* + \delta]$$

Además se verificará que $\forall x \in [x^* - \delta, x^* + \delta]$ tal que $g(x) \in [x^* - \delta, x^* + \delta]$. En efecto, si $x \in [x^* - \delta, x^* + \delta]$ se tendrá que $|x^* - x| \leq \delta$ y por el teorema de los incrementos finitos se tendrá que $\exists z \in [x^* - \delta, x^* + \delta]$:

$$|g(x) - x^*| = |g(x) - g(x^*)| = |g(z)|\,|x - x^*| \leq k\delta < \delta$$

En resumen $g(x)$ es una contracción definida sobre el espacio métrico completo $([x^* - \delta, x^* + \delta], d_f)$ y por tanto admitirá un único punto fijo x^* que es el límite de la sucesión $\{x_{i+1} = g(x_i)\}_{i=0}^{\infty}$ sea cual sea el punto $x_0 \in [x^* - \delta, x^* + \delta]$ con el que se inicialice.

c.q.d.

El teorema anterior puede ser completado estimando cotas de la distancia a la solución de la aproximación obtenida en una iteración respecto a la obtenida en la iteración precedente. En efecto, esto se hace en el siguiente teorema:

Teorema 2.3.6 *Dada la ecuación $f(x) = 0$ y suponiendo que $f(x)$ es una función que verifica las siguientes hipótesis:*

a) está definida en un intervalo $[a, b]$ en el que existe una solución x^* de la ecuación

b) $f'(x)$ es una aplicación lipschitciana de razón k en $[a, b]$

c) $\exists \beta > 0$ tal que $|f'(x)| > \beta \qquad \forall x \in [a, b]$

entonces existe algún valor δ tal que si se considera $x_0 \in [x^* - \delta, x^* + \delta,]$ la sucesión

$$\left\{ x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \right\}_{i=0}^{\infty}$$

converge hacia x^* y además se verifica que:

$$|x_{i+1} - x^*| \leq \frac{k}{2\beta} |x_i - x^*|$$

Demostración: Demostremos en primer lugar la última desigualdad del teorema. Para ello se sabe que:

$$x_{i+1} - x^* = x_i - \frac{f(x_i)}{f'(x_{i)}} - x^* = x_i - x^* - \frac{f(x_i) - f(x^*)}{f'(x_{i)}} =$$

$$= \frac{1}{f'(x_i)} . \left[f(x^*) - f(x_i) - f'(x_i)(x^* - x_i) \right]$$

Como por otra parte se tendrá que:

$$f(x^*) - f(x_i) = \int_{x_i}^{x^*} f'(x) dx$$

y por tanto:

$$f(x^*) - f(x_i) - f'(x_i)(x^* - x_i) = \int_{x_i}^{x^*} (f'(x) - f'(x_i)) dx$$

se obtiene que:

$$\left| f(x^*) - f(x_i) - f'(x_i)(x^* - x_i) \right| = \left| \int_{x_i}^{x^*} (f'(x) - f'(x_i)) dx \right| \leq$$

$$\leq \int_{x_i}^{x^*} \left| f'(x) - f'(x_i) \right| .dx \leq k \int_{x_i}^{x^*} |x - x_i| \, dx \leq \frac{k}{2} |x - x_i|^2$$

por lo que:

$$|x_{i+1} - x^*| \leq \frac{k}{2\beta} |x - x_i|^2$$

Una vez demostrada la desigualdad anterior la convergencia de la sucesión se garantizaría logrando que fuese una sucesión de Cauchy y que estuviéramos trabajando en un espacio métrico completo. Para ello basta con tomar $\delta_1 = \text{mín}\{|x^* - a|, |x^* - b|\}$ y considerar

$$\delta = \text{mín}\left\{\delta_1, \theta\left(\frac{2\beta}{k}\right)\right\}$$

donde θ es un valor de $]0,1[$ elegido de tal forma que $\theta\left(\frac{2\beta}{k}\right) < 1$. Con ello la distancia entre $x_0 \in [x^* - \delta, x^* + \delta]$ y x^* será inferior o igual a $\theta\left(\frac{2\beta}{k}\right) < 1$, la distancia entre x_1 y x^* verificará que:

$$|x_1 - x^*| \leq \frac{k}{2\beta} |x - x_0|^2 \leq \theta^2 \frac{2\beta}{k} \leq \theta = \theta$$

y por recurrencia se comprueba que la distancia a la sucesión va decreciendo de forma tal que bastará con considerar un índice lo suficientemente grande para hacerla, a partir de él, tan pequeña como se desee. Por tanto, al ser $([x^* - \delta, x^* + \delta], d_f)$ un completo si x_0 pertenece a este intervalo quedará garantizada la convergencia del método de Newton-Raphson.

c.q.d.

Observación 2.3.14 *Obsérvese lo siguiente:*

1. *Que la desigualdad del teorema anterior*

$$|x_{i+1} - x^*| \leq \frac{k}{2\beta} |x_i - x^*|$$

no garantiza por sí sola la convergencia del método pues simplemente establece una cota del valor de $|x_{i+1} - x^*|$. Es la pertenencia del valor inicial x_0 a un intervalo suficientemente pequeño en torno a la raíz la que garantiza el éxito del proceso.

2. *En este sentido es interesante reflexionar sobre el significado de la relación $k/2\beta$. En ella k es la constante de Lipschitz de $f'(x)$. Pero multiplicando la ecuación $f(x) = 0$ por un parámetro $\alpha \neq 0$ se obtiene una ecuación equivalente $\alpha f(x) = F(x) = 0$ en la que la constante de Lipschitz de la función $F'(x)$se ve afectada por el parámetro α. Obviamente también se vería afectado por este parámetro el valor de β (cota inferior del valor absoluto que toma $f'(x)$ en el intervalo de trabajo). Por ello la relación k/β*

es un límite superior de la "no linealidad" de $f(x)$ y la desigualdad del teorema anterior nos indica que cuanto menor sea este índice de no linealidad más rápida será la convergencia hacia la solución de la ecuación. En el caso extremo de que $f(x)$ sea lineal (una recta de ecuación $kx+c$) la constante de Lipschitz de $f'(x)$ será 0 y la cota inferior de la derivada será k por lo que la convergencia se alcanzará en una iteración del método.

Ilustremos el proceso anterior con un ejemplo que si bien no es propio de la ingeniería química es real como la vida misma:

Ejemplo 2.3.4 *El dinero necesario para pagar la cuota correspondiente a un crédito hipotecario a interés fijo se suele estimar mediante la denominada "ecuación de la anualidad ordinaria":*

$$Q = \frac{A}{i}[1-(1+i)^{-n}]$$

en donde Q es la cantidad (en euros) pedida en préstamo, A es la cuota (en euros) que debe pagar el beneficiario del préstamo, i es la tasa de interés (en tanto por 1) fijado por la entidad bancaria que concede el préstamo y n es el número de periodos durante los cuales se realizan pagos de la cuota (meses si se paga mensualmente, trimestres si se paga trimestralmente, semestres si se paga semestralmente o años si se paga anualmente).

Una pareja que desea comenzar una vida en común se plantea adquirir una vivienda y para ello saben que necesitan pedir un préstamo de 150000 euros a pagar semestralmente durante un plazo que ellos desean que sea de 10 años. Sabiendo que para atender este pago pueden destinar una cantidad máxima de 600 euros mensuales, calcúlese cual es el tipo máximo de interés al que pueden negociar su préstamo con las entidades bancarias.

Puesto que el pago es semestral, en 10 años realizarán un total de 20 cuotas. Además dado que pueden pagar 600 euros al mes, cada semestre podrán afrontar el pago de 3600 euros. Ello hace que la ecuación de la anualidad ordinaria quede:

$$150000 = \frac{3600}{i}[1-(1+i)^{-20}]$$

o bien

$$f(i) = 150000 - \frac{3600}{i}[1-(1+i)^{-20}] = 0$$

Se tiene entonces que:

$$f'(i) = \frac{3600}{i}\left[\frac{1}{i}\left(1-(1+i)^{-20}\right) - 20(1+i)^{-21}\right]$$

por lo que el método de Newton nos conduciría al esquema de cálculo:

$$i_{j+1} = i_j - \frac{150000 - \frac{3600}{i_j}[1-(1+i_j)^{-20}]}{\frac{3600}{i_j}\left[\frac{1}{i_j}\left(1-(1+i_j)^{-20}\right) - 20(1+i_j)^{-21}\right]}$$

que, partiendo de $i_0 = 0{,}03$ *nos proporcionará la siguiente sucesión de valores:*

$$i_1 = -0{,}1647.., \qquad i_2 = -0{,}1212..., \qquad i_3 = -0{,}0852..., \qquad i_4 = -0{,}0659..,$$

$$i_5 = -0{,}0617..., \qquad i_6 = -0{,}0616..., \qquad i_7 = -0{,}0616...$$

Como resultado de lo anterior se dan cuenta que difícilmente podrán encontrar la vivienda que desean pues parece razonable pensar que ningún banco o caja de ahorros les concederá un préstamo a un interés semestral negativo del $-6{,}16\,\%$*. Por ello tras planear dejar de ver a sus respectivas amistades y reinvertir el dinero que gastan en copas para adquirir la casa de sus sueños aumentan la cantidad que mensualmente pueden dedicar a amortizar el crédito hasta 5400 euros semestrales y asumen endeudarse durante 15 años en lugar de 10. Con ello el método de Newton-Raphson se convierte ahora en el esquema iterativo:*

$$i_{j+1} = i_j - \frac{150000 - \frac{5400}{i_j}[1-(1+i_j)^{-30}]}{\frac{5400}{i_j}\left[\frac{1}{i_j}\left(1-(1+i_j)^{-30}\right) - 30(1+i_j)^{-31}\right]}$$

y les proporciona la siguiente sucesión:

$$i_0 = 0{,}03, \quad i_1 = -0{,}0022.., \quad i_2 = 0{,}0044..., \quad i_3 = 0{,}050..., \quad i_4 = 0{,}0050...$$

Tras innumerables gestiones, la pareja en cuestión no encuentra ninguna entidad bancaria que les conceda el préstamo de 150000 euros al 0.5 % de interés semestral. Por ello, haciendo de tripas corazón, deciden endeudarse durante 20 años en lugar de 15 pero pagando la misma cantidad de 5400 euros semestrales pues (al ser un miembro de la pareja profesor asociado de tipo 1 en la Universidad Rey Juan Carlos y el otro administrativo de la escala C en un Organismo Oficial) les es imposible pagar más. Con ello el esquema iterativo se convierte en:

$$i_{j+1} = i_j - \frac{150000 - \frac{5400}{i_j}[1-(1+i_j)^{-40}]}{\frac{5400}{i_j}\left[\frac{1}{i_j}\left(1-(1+i_j)^{-40}\right) - 40(1+i_j)^{-41}\right]}$$

y les conduce a que: $i_0 = 0{,}03$, $i_1 = 0{,}0175...$, $i_2 = 0{,}0190...$, $i_3 = 0{,}0191...$, $i_4 = 0{,}0191...$

Desmoralizados al seguir sin encontrar entidad bancaria alguna que les conceda el préstamo al interés del $1'91\,\%$ *semestral, la pareja toma la decisión de*

renunciar a su casa ideal y busca otra (más alejada de la zona que les gusta, sin buenas comunicaciones, más antigua, más pequeña y construída con materiales de peor calidad) para la que sólo necesitan un préstamo de 100000 euros y mantienen las demás condiciones anteriores: pago de 5400 euros semestrales y 20 años de "condena". El esquema de Newton-Raphson en este caso les lleva a:

$$i_{j+1} = i_j - \frac{100000 - \frac{5400}{i_j}[1-(1+i_j)^{-40}]}{\frac{5400}{i_j}\left[\frac{1}{i_j}(1-(1+i_j)^{-40}) - 40(1+i_j)^{-41}\right]}$$

luego $i_0 = 0{,}03$, $i_1 = 0{,}0423..$, $i_2 = 0{,}0444...$, $i_3 = 0{,}0445...$, $i_4 = 0{,}0445..$,

Como finalmente ya encuentran una entidad que (tras duras negociaciones y previo avalamiento de fiadores solventes) les otorga el préstamo al interés del $4'45\,\%$ *la pareja puede comenzar una feliz vida en pareja en la que durante 20 años renunciarán a sus amigos (tranquilos que ellos también se pringarán) sin dinero para nada que no sea la supervivencia más elemental y, por supuesto, pagar la vivienda, y residiendo en una casa que no es la que les gusta.*

Observación 2.3.15 *Téngase en cuenta lo siguiente:*

1. *Esperamos que el lector que haya seguido el ejemplo anterior no eche la culpa de la situación a Sir Isaac Newton que aunque algo tuvo que ver con la banca no es el que fija los sueldos ni los tipos de interés bancario, ni al matemático inglés contemporáneo de Newton, Joseph Raphson. Además, para tranquilidad del lector, hemos de informarle que en la pareja del ejemplo uno de ellos, el administrativo del Organismo Oficial (pues el otro, aunque cobraba poco, se encontraba a gusto en una Universidad de calidad), al poco tiempo, salió elegido concejal del ayuntamiento en el pueblo al que fueron a vivir y de allí saltó al Consejo de Administración de Telefónica en el que, aparte de olvidar el método de Newton-Raphson, obtuvo pingües beneficios comprando lo que se llama algo así como "stock options".*

2. *Hablando más en serio, obsérvese que en el primer caso del ejemplo anterior, cuando la amortización del crédito era de 3600 euros y el plazo de pago 20 semestralidades el dinero total que se pagaba es de 72000 euros que no cubre el préstamo solicitado (150000 euros). Por ello no es extraño que el interés resultante sea negativo aun cuando ello no tenga sentido en la realidad. Pero es que tampoco lo tiene que en un préstamo se devuelva menos dinero del recibido.*

3. *También se puede observar utilizando el ejemplo anterior que la convergencia del método depende de la "cercanía" del punto de partida a la solución. En efecto si en el último de los supuestos considerados (préstamo de 100000 euros, amortización semestral de 5400 euros y 40 pagos*

semestrales) se hubiese partido de un interés inicial del 300 % (es decir $i_0 = 3$) la sucesión obtenida (y los valores de $f(i_j)$) resultan ser: $i_1 = -160{,}66666..$ ($f(i_1) = 1{,}0003410^6$, $f'(i_1) = 0{,}20919..$), $i_1 = -478{,}35425..$ ($f(i_1) = 1{,}0000010^6$, $f'(i_1) = 0{,}235910^{-7}$), ...valores que no convergen hacia nada.

Observación 2.3.16 *Si se quiere encontrar el valor de la raíz cuadrada de un número $a > 0$ puede buscarse como la solución de la ecuación no lineal $f(x) = x^2 - a = 0$. Para ello el método de Newton-Raphson nos conduce al esquema:*

$$x_{i+1} = x_i - \frac{x_i^2 - a}{2x_i} = \frac{1}{2}\left(x_i + \frac{a}{x_i}\right)$$

recuperándose así el método de Herón (o regla mecánica) con el que ilustrábamos la técnica de sobreiteración.

En ocasiones, para funciones $f(x)$ que satisfagan alguna hipótesis adicional, pueden rebajarse las condiciones que aseguran la convergencia del método de Newton Raphson. Un ejemplo de ello es el siguiente teorema:

Teorema 2.3.7 *Si $f(x) \in C^2([a,b])$, es creciente y convexa en $[a,b]$ y admite alguna raíz en $[a,b]$, entonces la sucesión $\left\{x_{i+1} = x_i - \dfrac{f(x_i)}{f'(x_i)}\right\}_{i=0}^{\infty}$ generada a partir de cualquier valor $x_0 \in [a,b]$, converge hacia la única solución de la ecuación $f(x) = 0$ en $[a,b]$.*

Demostración: La unicidad de la solución de $f(x) = 0$ es evidente por ser $f(x)$ continua y creciente en $[a,b]$. Además por ser $f(x)$ convexa se verificará que $f''(x) > 0$, $\forall\, x \in [a,b]$. Y por ser creciente se verificará que $f'(x) > 0$, $\forall x \in [a,b]$. Denotemos por $h_i = x_i - x^*$. Se tendrá que:

$$h_{i+1} = x_{i+1} - x^* = x_i - x^* - \frac{f(x_i)}{f'(x_i)} = \frac{h_i f'(x_i) - f(x_i)}{f'(x_i)}$$

Como por otra parte, desarrollando en serie de Taylor se tiene que existe un punto $\xi_i \in [a,b]$ para el que:

$$0 = f(x^*) = f(x_i - h_i) = f(x_i) - h_i f'(x_i) + \frac{1}{2}h_i^2 f''(\xi_i)$$

resultará que:

$$h_i f'(x_i) - f(x_i) = \frac{1}{2}h_i^2 f''(\xi_i)$$

Entrando con esta expresión en la que nos proporcionaba h_{i+1} se obtiene:

$$h_{i+1} = \frac{h_i f'(x_i) - f(x_i)}{f'(x_i)} = \frac{1}{2}\frac{f''(\xi_i)}{f'(x_i)} h_i^2 > 0$$

Lo anterior nos indica que x_{i+1} siempre será mayor que x^*. Además, por ser $f(x)$ creciente, $f(x_{i+1}) > f(x^*) = 0$. Luego las sucesiones $\{x_i\}_{i=1}^{\infty}$ y $\{f(x_i)\}_{i=1}^{\infty}$ son sucesiones acotadas inferiormente por x^* y por 0 respectivamente. Por otra parte se tiene que:

$$h_{i+1} = \frac{h_i f'(x_i) - f(x_i)}{f'(x_i)} = h_i - \frac{f(x_i)}{f'(x_i)} < h_i$$

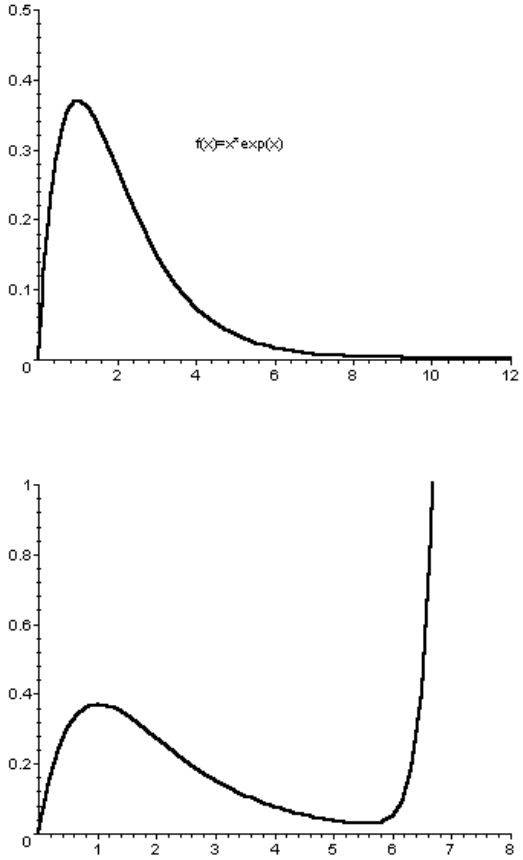
lo que nos indica que la sucesión $\{h_i\}_{i=1}^{\infty}$ es una sucesión decreciente y siempre positiva (pues recuérdese que $x_i > x^*$ para todo $i > 0$). Ello quiere decir que su límite será 0 y por tanto el límite de $\{x_i\}_{i=1}^{\infty}$ será x^*.

c.q.d.

Observación 2.3.17 *Resultados similares al anterior podrían obtenerse si $f(x)$, además de ser de clase $C^2([a,b])$ fuese convexa decreciente, o cóncava creciente o cóncava decreciente. Dejamos al lector la demostración de los mismos.*

En cuanto a la forma de detener el proceso iterativo, cuando $g(x) = x - \dfrac{f(x)}{f'(x)}$ sea una contracción puede seguirse la misma estrategia que en el método de aproximaciones sucesivas, es decir que cuando $|x_i - x_{i-1}|$ sea inferior a un cierto ε puede considerarse que x_i es una buena aproximación de la solución x^*. Pero un código informático que recoja el método de Newton-Raphson debería ser aplicable a situaciones en las que $g(x)$ no es una contracción y detectar por sí solo si se encuentra una solución de la ecuación o no. ¿Cómo saber en esos casos que se está cerca de la solución buscada?. Obsérvese por ejemplo que si $f'(x_i)$ toma un valor elevado el valor de $|x_{i+1} - x_i|$ puede hacerse muy pequeño sin necesidad de que $f(x_i)$ sea próximo a 0. Eso nos llevaría a que un criterio de detención del proceso iterativo sería que $|f(x_{i+1})| < \delta$ donde δ es un parámetro fijado de antemano y suficientemente pequeño. Lamentablemente este criterio tampoco es fiable pues puede darse el caso de funciones en las que $f(x_{i+1})$ sea muy pequeño estándose muy alejados de la solución. Por ejemplo, si se considera la función $f(x) = xe^{-x}$ y se quiere encontrar una solución no negativa de la ecuación $f(x) = 0$, la única solución es $x^* = 0$ pero para valores de $x_i = 10^i$ el valor de $f(x_i)$ se hace tan pequeño como se desee con tal de tomar i lo suficientemente elevado.

Puede entonces pensarse en que cuando las derivadas de la función tengan valor absoluto elevado los valores de $|x_{i+1} - x_i|$ serán pequeños pero el criterio de que $|f(x_{i+1})| < \delta$ nos servirá para saber si se está cerca o lejos de la solución



buscada en tanto que cuando $|f(x_{i+1})| < \delta$ será el analizar si $|x_{i+1} - x_i| < \varepsilon$ lo que nos permitirá discernir si estamos cerca o no de la solución buscada. Lamentablemente tampoco este criterio cubre todas las situaciones que puedan darse pues puede haber situaciones en las que la sucesión se acumule en torno a un mínimo suficientemente próximo a 0 pero lejano de la solución. Por ejemplo, si la gráfica de una función fuera como la de la figura, en torno a $x = 6$ se puede producir una acumulación de valores que nos conducirán hacia el mínimo de la función en lugar de hacia la solución $x^* = 0$.

Nótese que en este caso la derivada de $f(x)$ se anula en algún punto por lo que no se verifican las hipótesis que aseguran la convergencia del proceso. En tal caso procedería cambiar de punto de arranque del proceso para ubicarnos en una zona en la que sí esté garantizada la convergencia.

Aunque existen test de control de la cercanía a la solución basados en la consideración de los valores de $|x_{i+1} - x_i|$ y de los de f y sus derivadas en x_{i+1}, cambiando de punto de arranque del proceso iterativo cuando el método

se "atasca" en torno a algún punto que no es raíz, nosotros nos limitaremos en estos apuntes a considerar como control de cercanía a la solución el que $|x_{i+1} - x_i| < \varepsilon$ y que $|f(x_{i+1})| < \delta$. Esta estrategia será suficiente para los casos en que esté asegurada la convergencia del método y será acompañada con la limitación del número máximo de iteraciones que se permite realizar para asegurar la finalización del algoritmo en los casos en que no haya convergencia.

Ello nos permite escribir un algoritmo recogiendo el método de Newton-Raphson como el que sigue:

Algoritmo del método de Newton-Raphson:

Dada la ecuación $f(x) = 0$, los indicadores de precisión ε y δ, un valor máximo del número de iteraciones que se permiten realizar ($maxiter$) y un punto x_0 con el que inicializar el proceso,

$tolx \leftarrow 2\varepsilon$

$tolf \leftarrow 2\delta$

$iteración \leftarrow 0$

Mientras ($(iteración < maxiter)$ **y** $((tolx > \varepsilon)$ **o** $(tolf > \delta)$ **), hacer:**

Si $(f'(x_0) = 0)$ **entonces:**

Escribir mensaje de error (derivada nula) y finalizar el proceso

si no:

$$x_1 \leftarrow x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$tolx \leftarrow |x_1 - x_0|$$

$$tolf \leftarrow |f(x_1)|$$

$$iteración \leftarrow iteración + 1$$

$$x_0 \leftarrow x_1$$

fin condición.

Fin bucle condicional.

Si $((tolx < \varepsilon)$ **y** $(tolf < \delta)$ **) entonces:**

tomar x_1 como solución

si no:

Escribir un mensaje de error en el proceso de cálculo

fin condición.

Fin del algoritmo.

Observación 2.3.18 *En muchas ocasiones la diferencia entre dos valores consecutivos de las aproximaciones obtenidas se relativizan para expresarlas porcentualmente. Para ello en el algoritmo anterior puede sustituirse la línea:*

$$tolx \leftarrow |x_1 - x_0|$$

por otras que sean de la forma:

Si $(x_1 \neq 0\ )$ **entonces:**
$\quad tolx \leftarrow \frac{|x_1 - x_0|}{|x_1|} 100$
si no:
$\quad tolx \leftarrow |x_1 - x_0|$
fin condición.

Variantes del método de Newton-Raphson: métodos de la secante y de "regula falsi"

El método de Newton que se acaba de exponer es un método que, generalmente, tiene un buen comportamiento en la resolución de muy diferentes ecuaciones no lineales. Su principal inconveniente práctico consiste en la necesidad de calcular los valores de $f'(x_i)$ en cada iteración. Por ello existen variantes del método de Newton que tratan de obviar este cálculo aproximando el valor de $f'(x_i)$. Entre ellos, los más populares son los denominados **método de la secante** y **método de regula falsi**.

a) Método de la secante.

Este método aproxima el valor de $f'(x_i)$ mediante:

$$f'(x_i) \approx \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}$$

con lo que el esquema iterativo del método de Newton-Raphson se ve modificado a:

$$x_{i+1} = x_i - \frac{f(x_i)}{\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}} = \frac{x_{i-1} f(x_i) - x_i f(x_{i-1})}{f(x_i) - f(x_{i-1})}$$

Obsérvese que para aplicar el método se necesitan dos valores x_0 y x_1 con los que inicializar el proceso. Por ello en el método de la secante la primera iteración se realiza mediante el método de Newton siguiéndose el siguiente proceso:

$$\textit{Dado } x_0$$

$$x_1 \quad \leftarrow \quad x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$x_{i+1} \leftarrow \frac{x_{i-1}f(x_i) - x_i f(x_{i-1})}{f(x_i) - f(x_{i-1})} \qquad i = 1, 2, ...$$

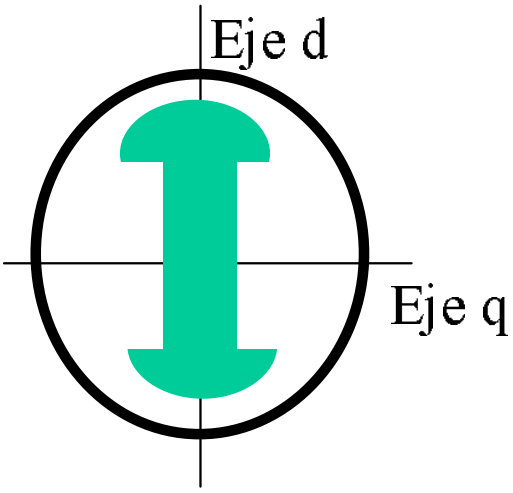
Observación 2.3.19 *El método de la secante toma su nombre del hecho de que gráficamente se puede interpretar el método de forma similar al de Newton pero sustituyendo la recta tangente a la curva por el punto $(x_i, f(x_i))$ por la recta secante que pasa por los puntos $(x_{i-1}, f(x_{i-1}))$ y $(x_i, f(x_i))$.*

Ilustremos el funcionamiento del método con un ejemplo:

Ejemplo 2.3.5 *(Cortesía del Pr. E. Conde): La energía eléctrica que habitualmente consumimos se genera y transmite por las redes eléctricas en forma de corriente alterna, es decir, en forma de onda sinusoidal, a una determinada frecuencia (en Europa 50 Hz). La existencia de campos magnéticos hace aparecer efectos inductivos y capacitivos en la red que introducen desfases entre las ondas de corriente y las ondas de tensión. Una forma de representar estas ondas sinusoidales es mediante variables del campo complejo. En este sentido puede hablarse de la denominada "potencia activa" (P) que es la potencia útil (capaz de transformarse en trabajo) que se transmite por la red. Pero también existe la denominada "potencia reactiva" (Q) que es aquella que no realiza trabajo útil pero que es necesaria para la creación de los campos magnéticos y eléctricos de los componentes del sistema. Es a esta potencia reactiva a la que se le asigna la componente imaginaria en la forma de analizar los sistemas eléctricos. También en estos sistemas se define la "potencia aparente" (S) que no es más que el módulo de la potencia transmitida, es decir $|S| = \sqrt{P^2 + Q^2}$. De esta forma la potencia transmitida S (que es demandada por las cargas conectadas a la red) se puede descomponer en una potencia activa (P, parte real de S) y una potencia reactiva (Q, parte imaginaria de S). El argumento principal de la potencia transmitida ($\omega = arctg(P/Q)$) representa el desfase existente entre la onda de tensión y la de intensidad.*

La sección de los generadores de polos salientes, instalados en diferentes centrales eléctricas, es del estilo a la que se recoge en la figura siguiente.

En dicha sección se observa la existencia de una parte móvil giratoria en la que existe un campo electromagnético, llamada rotor (y que es accionada normalmente por una turbina a la que arrastra el vapor a presión generado en las calderas de las centrales térmicas, o la corriente de agua en las centrales hidráulicas) y una parte fija, el estátor, en la que existe un bobinado en el que se genera el campo eléctrico (por el movimiento del rotor) que es el que entrega, previa transformación, la potencia a trasmitir por la red eléctrica. Al no existir una distancia fija entre todos los puntos del contorno del rotor y del estátor, la resistencia en cada punto al paso del flujo magnético (reluctancia) es diferente en unas direcciones u otras. Por ello para su estudio se recurre a una aproximación que descompone los fenómenos de la máquina según el denominado eje



directo "d" (menor entrehierro) y el denominado eje de cuadratura "q" (máximo entrehierro).

Al girar el rotor se genera en el estátor una fuerza electromotriz E (que no es más que una tensión). Sin embargo, como el generador no es ideal, existe una diferencia entre esta fuerza electromotriz y la tensión V que se tiene en los bornes del generador. Como ambas son ondas sinusoidales, entre una y otra onda aparece un desfase que se denomina ángulo de carga.

Es este ángulo de carga el que determina la relación entre la potencia activa, la reactiva y la potencia aparente que puede suministrar el generador para una fuerza electromotriz E y una tensión en bornes V dadas. Para asegurar un funcionamiento adecuado del generador el ángulo de carga se debe mantener dentro de unos límites. Por ejemplo, en una central eléctrica, cuando se quiere aumentar la potencia suministrada manteniendo los valores de V y de E, se aumenta el flujo motriz de la turbina, lo cual a su vez hace aumentar el ángulo de carga. Pero esto no se puede hacer de forma descontrolada puesto que se deben respetar los límites de estabilidad (desde luego nunca se deben sobrepasar los 90º). En la práctica este ángulo se mantiene entre 30º y 40 º.

Un generador de energía eléctrica de polos salientes, similar a los habitualmente instalados en las centrales hidroeléctricas, tiene como características asignadas:

S_N = *Potencia aparente nominal* = 15 *MVA*

V_N = *Tensión nominal* = *13.6 kV*

f_N = *frecuencia nominal* = *50 Hz*

X_d = *reactancia síncrona según eje directo* = *0.91 por unidad*

X_q = *reactancia síncrona según el eje de cuadratura* = *0.76 por unidad*

En un momento determinado el generador está entregando a la red eléctrica una potencia activa de 10 MW, teniendo en bornes su tensión nominal (13.6 kV) y una fuerza electromotriz interna $E = 16{,}592$ kV. Se desea encontrar el ángulo de carga del generador (es decir, el ángulo formado entre la tensión en bornes V y la fuerza electromotriz E).

La fuerza electromotriz de un generador de polos salientes (despreciando la resistencia interna) se puede expresar por:

$$E = V + iX_d I_d + iX_q I_q$$

Las potencias activa y reactiva por unidad suministradas a su vez se relacionan con la tensión en bornes V y con la fuerza electromotriz E mediante las expresiones:

$$P = \frac{|V|\,|E|}{X_d}\sin(\delta) + \frac{|V|^2}{2}\left(\frac{1}{X_q} - \frac{1}{X_d}\right)\sin(2\delta)$$

$$Q = \frac{|V|\,|E|}{X_d}\cos(\delta) - |V|^2\left(\frac{\cos(\delta)}{X_d} + \frac{\sin^2(\delta)}{X_q}\right)$$

considerando los valores nominales de la máquina resultará:

$$P = \frac{10}{15} = \frac{2}{3}$$

$$|V| = \frac{13{,}6}{13{,}6} = 1$$

$$|E| = \frac{16{,}592}{13{,}6} = 1{,}22$$

valores para los que la expresión de la potencia activa por unidad nos conduce a

$$\frac{2}{3} = C_1\sin(\delta) + C_2\sin(2\delta)$$

donde

$$C_1 = \frac{16{,}592}{13{,}60{,}91}$$

$$C_2 = \frac{1}{2}\left(\frac{1}{0{,}76} - \frac{1}{0{,}91}\right)$$

En otros términos se trata de encontrar una solución de la ecuación:

$$f(\delta) = C_1\sin(\delta) + C_2\sin(2\delta) - \frac{2}{3} = 0$$

Para ello, puesto que

$$f'(\delta) = C_1 \cos(\delta) + 2C_2 \cos(2\delta)$$

utilizando el método de Newton, se tendrá que:

$$\delta_{i+1} = \delta_i - \frac{C_1 \sin(\delta_i) + C_2 \sin(2\delta_i) - \frac{2}{3}}{C_1 \cos(\delta_i) + 2C_2 \cos(2\delta_i)}$$

Partiendo de $\delta_0 = 0$ se tiene la siguiente tabla de valores:

Iteración	δ_i	$f(\delta_i)$
1	0,428023270207	$-0{,}0282916312890$
2	0,448797366525	$-0{,}19342675203210^{-3}$
3	0,448941379375	$-0{,}95497175591010^{-8}$
4	0,448941386486	$0{,}29381878874410^{-16}$

por lo que el ángulo buscado será $\delta^* \approx 0{,}448941386486\ rad$ ($\approx 25{,}72246^o$)

Si en lugar del método de Newton se hubiese utilizado el método de la secante, la primera iteración se realizaría exactamente igual que en el método de Newton y, a partir de ella, se emplearía el esquema:

$$\delta_{i+1} \leftarrow \frac{\delta_{i-1} f(\delta_i) - \delta_i f(\delta_{i-1})}{f(\delta_i) - f(\delta_{i-1})} \qquad i = 1, 2, ...$$

es decir, que en este caso:

$$\delta_{i+1} = \frac{\delta_{i-1}(C_1 \sin(\delta_i) + C_2 \sin(2\delta_i) - \frac{2}{3}) -}{C_1(\sin(\delta_i) - \sin(\delta_{i-1})) + C_2(\sin(2\delta_i) - \sin(2\delta_{i-1}))} -$$

$$- \frac{\delta_i(C_1 \sin(\delta_{i-1}) + C_2 \sin(2\delta_{i-1}) - \frac{2}{3})}{C_1(\sin(\delta_i) - \sin(\delta_{i-1})) + C_2(\sin(2\delta_i) - \sin(2\delta_{i-1}))} \quad (i = 1, 2, ...)$$

lo que, partiendo también de $\delta_0 = 0.$ nos conduce a la tabla:

Iteración	δ_i	$f(\delta_i)$
1	0,428023270207	$-0{,}0282916312890$
2	0,446992490293	$0{,}18969220086510^{-1}$
3	0,448927715744	$0{,}19352254510310^{-2}$
4	0,448941377367	$0{,}13661622824610^{-4}$
5	0,448941386486	$0{,}91188854560110^{-8}$

En esta ocasión hemos necesitado una iteración más para obtener una solución satisfactoria. Como luego se justificará, el método de la secante tiene una velocidad de convergencia inferior al método de Newton aunque superior al

método de aproximaciones sucesivas. Obsérvese también que el valor del residuo es, en nuestra solución aproximada por el método de la secante, superior al residuo que obtuvimos con el método de Newton. Ello es debido a que, aunque los 12 primeros decimales utilizados en el desarrollo del ejemplo coinciden en ambos casos, hay diferencias en decimales posteriores que el programa con el que se han realizado los cálculo sí tuvo en cuenta.

b) El método de "Regula Falsi".

Este método es una combinación del método de bipartición y el método de la secante. En él se considera una ecuación $f(x) = 0$ y un intervalo $[a, b]$ en el que $f(x)$ sea continua y además se verifique que $f(a)f(b) < 0$. Con ello, según se indicó al analizar el método de bipartición se puede estar seguro de que en $[a, b]$ existe al menos una raíz. Tras ello se denomina x_1 al punto de corte con el eje de abscisas de la recta secante que pasa por los puntos $(a, f(a))$, $(b, f(b))$, es decir, que será el punto:

$$x_1 = \frac{af(b) - bf(a)}{f(b) - f(a)}$$

Si $f(x_1)f(a) < 0$ se puede asegurar que en el intervalo (a, x_1) existirá una solución de la ecuación. En el caso de que $f(x_1)f(a) > 0$ se puede afirmar lo mismo para el intervalo (x_1, b). Y en el caso de que $f(x_1) = 0$ se habrá determinado ya la solución. En todo caso o se tiene la solución de la ecuación o se dispone de un intervalo más pequeño en el que volver a repetir el proceso. Esta forma de proceder es repetida las veces que sea necesario hasta encontrar un intervalo en el que exista una solución y con una longitud inferior a la precisión deseada. Más concretamente el algoritmo del método será:

Algoritmo del método de Regula Falsi:

Dada la ecuación $f(x) = 0$, el indicador de precisión ε y dos puntos a y b en los que $f(a)f(b) < 0$,

Mientras $|b - a| > \varepsilon$, **hacer:**

$$x \leftarrow \frac{af(b) - bf(a)}{f(b) - f(a)}$$

Si $(f(x) = 0)$ **entonces:**
tomar x como raíz x^* y finalizar el proceso
si no:
Si $(f(x)f(a) > 0)$ **entonces**:
$b \leftarrow x$

si no:
$a \leftarrow x$
fin condición.
fin condición.
Fin bucle condicional.
$x^* \leftarrow x$
Fin del algoritmo.

Observación 2.3.20 *Este método de "regula falsi", aquí presentado como variante del método de Newton-Raphson, históricamente es anterior. En efecto, las primeras referencias de este método (aunque probablemente sea anterior), consistentes en la aplicación a la resolución de algunos ejemplos de ecuaciones lineales, se deben al gran matemático árabe Muhamad ibn Musà al-Kwarizmi que trabajó en el califato de Bagdad a finales del siglo VIII y comienzos del siglo IX. No obstante no es esta la más importante contribución de este matemático. En efecto, Al-Kwarizmi contribuyó notablemente a la introducción del sistema de numeración decimal (arábigo) tomado del sistema de numeración hindú y en el que entre otras cosas se introduce el número* 0 *(no existente en los sistemas de numeración anteriores). Con ello además divulga cómo operar con los números decimales frente a los métodos, entonces tradicionales, basados en ábacos. En su obra "Aritmética" Al-Kwarizmi proporciona las reglas para realizar las operaciones aritméticas en este sistema decimal. Las traducciones que de esta obra se hicieron posteriormente al latín y en las que su nombre era latinizado dieron origen a la palabra "algoritmo" que hoy, en general, denota los procesos que conducen a la consecución de objetivos. Asimismo, Al-Kwarizmi publicó una obra ("Hibab al-jabar wa-ad-muquabala") sobre la resolución de ecuaciones en las que "restaura el orden" (en árabe al-jabar) de las ecuaciones que pretende resolver. Esta publicación es considerada como el inicio de la rama de las matemáticas que hoy en día se llama Álgebra (que debe su nombre a la latinización del término al-jabar).*

Volviendo al método de regula falsi, otro matemático del califato de Bagdad, Abu Kamil, en torno al año 900, usa este método, en su obra "Sobre los aumentos y las disminuciones", para la resolución de ecuaciones lineales de una incógnita mediante 1 ó 2 ensayos.

c) Otras variantes del método de Newton.

Existen otras variantes del método de Newton que son relativamente utilizadas en algunas aplicaciones. Así por ejemplo está el denominado **método de Newton modificado** en el que el valor de f' se estima sólo cada k iteraciones actuándose entre dichas iteraciones con el último valor de f' calculado.

Otra variante, conocida con el nombre de **método de Newton mejorado** se basa en en el siguiente razonamiento: al justificar los orígenes del método de Newton escribíamos:

$$0 = f(x^*) = f(x_i) + (x^* - x_i)f'(x_i) + \frac{(x^* - x_i)^2}{2}f''(x_i) + ...$$

desarrollo que una vez linealizado nos conducía a que: $(x^* - x_i) \approx -\frac{f(x_i)}{f'(x_i)}$ de donde se obtenía una aproximación x_{i+1} de la solución como $x_i - \frac{f(x_i)}{f'(x_i)}$. En el método de Newton mejorado se usa el hecho de que $(x^* - x_i) \approx -\frac{f(x_i)}{f'(x_i)}$ para sustituir esta expresión en uno de los dos factores $(x^* - x_i)$ que intervienen en el término de segundo grado del desarrollo de Taylor, despreciando los de mayor orden, con lo que:

$$0 = f(x^*) \approx f(x_i) + (x^* - x_i)f'(x_i) - (x^* - x_i)\frac{f(x_i)}{2f'(x_i)}f''(x_i)$$

de donde

$$(x^* - x_i) \approx -\frac{f(x_i)}{f'(x_i) - \frac{f(x_i)f''(x_i)}{2f'(x_i)}} = -\frac{2f(x_i)f'(x_i)}{2(f'(x_i))^2 - f(x_i)f''(x_i)}$$

y generándose, a partir de un x_0, la sucesión:

$$x_{i+1} = x_i - \frac{2f(x_i)f'(x_i)}{2(f'(x_i))^2 - f(x_i)f''(x_i)}$$

Observación 2.3.21 *Este método también es conocido con el nombre de método de Halley en honor al astrónomo inglés Edmund Halley, contemporáneo y amigo de Newton, que observó y estudió el famoso cometa que lleva su nombre y del que, en 1707, calculó su periodo prediciendo su aparición cada 76 años e identificándolo por tanto como el mismo cometa que había aparecido en las cercanías de la tierra en numerosas ocasiones anteriores y de las que se tenían numerosas referencias. Entre ellas la que le sirvió a Giotto para dibujarlo en su cuadro sobre la Natividad como la famosa estrella de los Reyes Magos.*

2.3.4. Velocidad de convergencia de los métodos iterativos

Una misma ecuación no lineal podrá ser resuelta en ocasiones por diferentes métodos iterativos. Para poder optar entre uno u otro interesará conocer cual de ellos nos acerca más rápidamente a la solución de la ecuación. Ello se hace a través del concepto denominado "orden de convergencia" que pasamos a definir a continuación:

Definición 2.3.1 *Siendo* $\{x_i\}_{i=0}^{\infty}$ *una sucesión convergente hacia* x^* *en la que* $x_i \neq x^*$ *para todo valor del índice* i*, se dice que* ***la sucesión converge*** *hacia* x^* ***con orden*** p ***y con una constante de error asintótico*** β *cuando existen dos números reales positivos* p *y* β *tales que:*

$$\lim_{i\to\infty} \frac{|x_{i+1} - x^*|}{|x_i - x^*|^p} = \beta$$

En este sentido se dice que un ***método iterativo*** *de la forma* $x_{i+1} = g(x_i)$ *es* ***de orden*** p *cuando la sucesión* $\{x_i\}_{i=0}^{\infty}$ *converja hacia una solución de* $x = g(x)$ *con orden* p*. En el caso de que* p *sea igual a 1 se dice que el método* ***converge linealmente****. Y si* $p = 2$ *se dice que el método* ***converge cuadráticamente****. Cuando* $p > 1$ *se dirá que la* ***convergencia*** *es* ***superlineal****.*

En general, dada una sucesión $\{x_i\}$ que converja hacia x^* con orden de convergencia p y otra sucesión $\left\{x_i^{'}\right\}$ que converja hacia x^* con orden de convergencia $q < p$ se verifica que los elementos de $\{x_i\}$ se acercan más rápidamente hacia x^* por lo menos a partir de un cierto índice de los elementos de la sucesión.

Una forma cómoda de conocer el orden de un método iterativo, cuando éste está dado por un número natural, nos la proporciona el siguiente teorema:

Theorem 2 *Siendo* $g(x)$ *una función de clase* $C^{p+1}([a,b])$ *y tal que en* $[a,b]$ *admite un único punto fijo* x^* *en el que se cumplen las hipótesis:*

$$g^{(k}(x^*) = 0 \qquad para\ k = 1, 2, ..., (p-1)$$

$$g^{(p}(x^*) \neq 0$$

entonces se verifica que la sucesión generada a partir de un $x_0 \in [a,b]$*, mediante* $\{x_{i+1} = g(x_i)\}_{i=0}^{\infty}$ *si converge hacia* x^* *lo hace con un orden de convergencia* p*.*

Demostración: Considerando el desarrollo en serie de Taylor:

$$x_{i+1} = g(x_i) = g(x^* + (x_i - x^*)) = g(x^*) + (x_i - x^*)g'(x^*) + ...+$$

$$+..... + \frac{(x_i - x^*)^p}{p!} g^{(p}(x^*) + O((x_i - x^*)^{(p+1)})$$

y dado que $x^* = g(x^*)$ y que las (p-1) primeras derivadas de g son nulas en x^* se tendrá que

$$\frac{(x_{i+1} - x^*)}{(x_i - x^*)^p} = \frac{1}{p!} g^{(p}(x^*) + O((x_i - x^*))$$

por lo que si la sucesión converge hacia x^* se verificará

$$\lim_{i\to\infty}\frac{(x_{i+1}-x^*)}{(x_i-x^*)^p}=\frac{1}{p!}g^{(p}(x^*)$$

c.q.d.

El teorema anterior nos permite analizar fácilmente el orden de convergencia de los métodos iterativos cuando este orden es un número entero. Así para el método de aproximaciones sucesivas y para el método de Newton-Raphson se tiene:

Proposición 2.3.4 *Si $g(x)$ es una contracción en $[a,b]$, el método de aproximaciones sucesivas es, al menos, de convergencia lineal.*

Demostración: Basta con comprobar que $g'(x^*)$ no tiene por qué ser nula.

c.q.d.

Proposición 2.3.5 *En las condiciones de convergencia del método de Newton-Raphson, si x^* es una solución simple de la ecuación $f(x)=0$, y $f(x)$ es de clase $C^2([a,b])$, este método es, al menos, de convergencia cuadrática.*

Demostración: En el método de Newton-Raphson:

$$g(x)=x-\frac{f(x)}{f'(x)}$$

por lo que si x^* es una raíz simple de $f(x)$ se verificará que

$$g'(x^*)=\frac{f(x^*)f''(x^*)}{(f'(x^*))^2}=0$$

y

$$g''(x^*)=\frac{f''(x^*)}{f'(x^*)}$$

que en general no tiene por qué anularse.

c.q.d.

Observación 2.3.22 *Adviértase que en la proposición anterior se especifica que la raíz buscada debe ser simple. Si no lo fuese $f'(x^*)$ sería nulo y el razonamiento anterior no sería correcto. En general si x^* fuese una raíz de multiplicidad m y el método de Newton-Raphson convergiese hacia ella, se podrá escribir la ecuación en la forma:*

$$f(x)=(x-x^*)^m h(x)=0$$

en la que x^* no es raíz de $h(x)$, por lo que

$$f'(x) = m(x-x^*)^{(m-1)}h(x) + (x-x^*)^m h'(x)$$

de donde

$$g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{(x-x^*)h(x)}{mh(x) + (x-x^*)h'(x)}$$

y

$$g'(x^*) = 1 - \frac{h(x^*)mh(x^*)}{m^2h^2(x^*)} = 1 - \frac{1}{m}$$

En resumen, en este caso sólo se puede asegurar la convergencia lineal.

Otros métodos en los que el orden de convergencia no es entero deben analizarse a partir de la definición dada para el orden de convergencia de la sucesión que generan. Así por ejemplo se tiene la siguiente propiedad:

Proposición 2.3.6 *Siendo f una función de clase $C^3([a,b])$, el método de la secante para la búsqueda de raíces simples de la ecuación $f(x) = 0$, cuando converge, presenta una convergencia de orden $\left(\frac{1+\sqrt{5}}{2}\right)$.*

Demostración:
Denotemos por $h_i = x_i - x^*$. Se tiene entonces que

$$h_{i+1} = x_{i+1} - x^* = \frac{x_{i-1}f(x_i) - x_i f(x_{i-1})}{f(x_i) - f(x_{i-1})} - x^* =$$

$$= \frac{f(x_i)(x_{i-1} - x^*) - f(x_{i-1})(x_i - x^*)}{f(x_i) - f(x_{i-1})} =$$

$$= \frac{f(x_i)h_{i-1} - f(x_{i-1})h_i}{f(x_i) - f(x_{i-1})} =$$

$$= \frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})} \frac{\frac{f(x_i)}{h_i} - \frac{f(x_{i-1})}{h_{i-1}}}{x_i - x_{i-1}} h_i h_{i-1}$$

Examinemos las fracciones que intervienen en la expresión anterior. En primer lugar puesto que mediante desarrollos en serie de Taylor:

$$f(x_i) = f(x^* + h_i) = h_i f'(x^*) + \frac{1}{2}h_i^2 f''(x^*) + O(h_i^3) \quad \Rightarrow$$

$$\Rightarrow \quad \frac{f(x_i)}{h_i} = \frac{1}{2}h_i f''(x^*) + O(h_i^2)$$

y análogamente,

$$\frac{f(x_{i-1})}{h_{i-1}} = \frac{1}{2}h_{i-1} f''(x^*) + O(h_{i-1}^2)$$

resultará que

$$\frac{f(x_i)}{h_i} - \frac{f(x_{i-1})}{h_{i-1}} = \frac{1}{2}(h_i - h_{i-1})f''(x^*) + O(h_{i-1}^2)$$

de donde, dado que $x_i - x_{i-1} = (x_i - x^*) - (x_{i-1} - x^*) = h_i - h_{i-1}$, se tiene que

$$\frac{\frac{f(x_i)}{h_i} - \frac{f(x_{i-1})}{h_{i-1}}}{x_i - x_{i-1}} = \frac{1}{2}f''(x^*) + O(h_{i-1}^2)$$

Análogamente, combinando los desarrollos en serie anteriores se obtiene que

$$\frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})} = \frac{1}{f'(x^*)} - O(h_{i-1}^2)$$

En resumen,

$$h_{i+1} = \frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})} \frac{\frac{f(x_i)}{h_i} - \frac{f(x_{i-1})}{h_{i-1}}}{x_i - x_{i-1}} h_i h_{i-1} \approx \frac{f''(x^*)}{2f'(x^*)} h_i h_{i-1} =$$

$$= C h_i h_{i-1}$$

Si el método converge con un orden p y una constante de error asintótico β se verificará que

$$\lim_{i\to\infty} \frac{h_{i+1}}{h_i^p} = \beta$$

lo que nos llevaría a que

$$\lim_{i\to\infty} h_i = \left(\frac{1}{\beta}\lim_{i\to\infty} h_{i+1}\right)^{1/p}$$

y análogamente

$$\lim_{i\to\infty} h_{i-1} = \left(\frac{1}{\beta}\lim_{i\to\infty} h_i\right)^{1/p}$$

por lo que tomando límites en la expresión antes obtenida,

$$h_{i+1} \approx C h_i h_{i-1} \Rightarrow \lim_{i\to\infty} h_{i+1} = C\lim_{i\to\infty} h_i (\frac{1}{\beta}\lim_{i\to\infty} h_i)^{1/p} = \frac{C}{\beta^{\frac{1}{p}}}(\lim_{i\to\infty} h_i)^{1+\frac{1}{p}}$$

$$\Rightarrow \beta\lim_{i\to\infty} h_i^p = \frac{C}{\beta^{\frac{1}{p}}}(\lim_{i\to\infty} h_i)^{1+\frac{1}{p}}$$

de donde finalmente,

$$\frac{\beta^{1-\frac{1}{p}}}{C} = (\lim_{i\to\infty} h_i)^{1-p+\frac{1}{p}}$$

Puesto que el lado izquierdo de la igualdad anterior es una constante distinta de cero, mientras que al haber supuesto la convergencia del método se verifica que $\lim_{i\to\infty} h_i = 0$, se debe verificar que

$$1 - p + \frac{1}{p} = 0$$

o lo que es lo mismo que: $1+p-p^2 = 0$ de donde $p = \frac{1\pm\sqrt{5}}{2}$. En esta expresión de p el signo negativo no tiene sentido pues conduciría a órdenes de convergencia negativos por lo que finalmente $p = \frac{1+\sqrt{5}}{2}$.

c.q.d.

Observación 2.3.23 *Se observa que:*

1. *En resumen el método de la secante (para la búsqueda de raíces simples) tiene una convergencia del orden* 1,62.. *es decir menor que el método de Newton pero mayor que (en general) el método de aproximaciones sucesivas. No obstante en el método de Newton el esfuerzo computacional en cada iteración puede ser mayor ya que debe estimarse en cada iteración el valor de $f(x_i)$ y el de $f'(x_i)$ lo que nos conduce a que debe optarse entre un menor número de iteraciones más costosas o un mayor número de iteraciones menos costosas.*

2. *El número $\left(\frac{1+\sqrt{5}}{2}\right)$ se conoce con el nombre de número aúreo y aparece en otras parcelas de las matemáticas tales como la optimización. Este número, por ejemplo, es el límite de la relación entre cada uno de los números y su precursor de la conocida sucesión de Fibonacci (generada por el matemático italiano, de finales del siglo XII y comienzos del XIII, Leonardo de Pisa (o Pisano) más conocido como Fibonacci al ser "filo de Bonaccio"). Pero esto, como ya se ha dicho, es objeto de otra parcela de las matemáticas.*

Ejercicio propuesto: Demuéstrese que el método de Newton mejorado (o de Halley) es de orden 3.

2.3.5. Aceleración de la convergencia de los métodos iterativos: método Δ^2 de Aitken

Cuando el método de resolución de ecuaciones no lineales que se esté empleando para resolver una ecuación no lineal no posea convergencia, al menos, cuadrática, puede utilizarse la estrategia conocida con el nombre de método delta-dos (Δ^2) de A. C. Aitken (matemático del siglo XX) para mejorar su velocidad de convergencia. Antes de examinar en qué consiste esta estrategia presentemos los fundamentos teóricos de la misma.

Definition 1 *Dada una sucesión* $\{x_i\}_i^\infty$ *se denomina* ***diferencia progresiva de primer orden*** *en el punto* x_i*, y se representará por* Δx_i*, al valor:*

$$\Delta x_i = x_{i+1} - x_i, \qquad i \geq 0$$

Análogamente se define la ***diferencia progresiva de orden m*** *en el punto* x_i*, y se representará por* $\Delta^m x_i$*, mediante:*

$$\Delta^m x_i = \Delta\left(\Delta^{(m-1)} x_i\right), \qquad i \geq 0, \quad m \geq 2$$

En concreto la **diferencia progresiva de orden 2** será, según la definición anterior:

$$\Delta^2 x_i = \Delta\left(\Delta x_i\right) = \Delta x_{i+1} - \Delta x_i = x_{i+2} - 2x_{i+1} + x_i \qquad i \geq 0$$

Teorema 2.3.8 *Sea* $\{x_i\}_{i=0}^\infty$ *una sucesión convergente hacia* x^**, y sea* $\{y_i\}_{i=0}^\infty$ *una nueva sucesión generada a partir de la primera mediante:*

$$y_i = x_i - \frac{(\Delta x_i)^2}{\Delta^2 x_i} = \frac{x_i x_{i+2} - x_{i+1}^2}{x_{i+2} - 2x_{i+1} + x_i} \qquad i \geq 0$$

Bajo la hipótesis de que exista una constante c *tal que* $|c| < 1$ *y una sucesión* $\{\delta_i\}_{i=0}^\infty$ *tal que* $\lim_{i\to\infty}\delta_i = 0$ *y tales que:*

$$x_{i+1} - x^* = (c + \delta_i)(x_i - x^*) \qquad i \geq 0$$

se verifica entonces que la sucesión $\{y_i\}_{i=0}^\infty$ *converge hacia* x^* *y además:*

$$\lim_{i\to\infty} \frac{y_i - x^*}{x_i - x^*} = 0$$

Demostración: Denotemos por $h_i = x_i - x^*$. Se tiene entonces que para todo valor del índice i:

$$y_i = \frac{x_i x_{i+2} - x_{i+1}^2}{x_{i+2} - 2x_{i+1} + x_i} = \frac{(x^* + h_i)(x^* + h_{i+2}) - (x^* + h_{i+1})^2}{(x^* + h_{i+2}) - 2(x^* + h_{i+2}) + (x^* + h_i)} =$$

$$= x^* + \frac{h_i h_{i+2} - h_{i+1}^2}{h_{i+2} - 2h_{i+1} + h_i}$$

Utilizando el hecho de que $h_{i+1} = (c + \delta_i)h_i$ se tiene que

$$h_{i+2} = (c + \delta_{i+1})(c + \delta_i)h_i$$

y

$$y_i - x^* = \frac{(c + \delta_{i+1})(c + \delta_i)h_i^2 - (c + \delta_i)^2 h_i^2}{(c + \delta_{i+1})(c + \delta_i)h_i - 2(c + \delta_i)h_i + h_i} =$$

$$= \frac{(c+\delta_{i+1})(c+\delta_i)-(c+\delta_i)^2}{(c+\delta_{i+1})(c+\delta_i)-2(c+\delta_i)+1}h_i$$

Tomando límites en la expresión anterior, dado que $c \neq 1$ (única raíz de $c^2 - 2 \cdot c + 1 = 0$) que $\{x_i\}_{i=1}^{\infty}$ converge hacia x^* y que $\{\delta_i\}_{i=0}^{\infty}$ converge hacia 0, es evidente que $\{y_i\}_{i=0}^{\infty}$ convergerá hacia x^*. Además de dicha expresión resultará que:

$$\lim_{i\to\infty}\frac{y_i - x^*}{h_i} = \lim_{i\to\infty}\frac{(c+\delta_{i+1})(c+\delta_i)-(c+\delta_i)^2}{(c+\delta_{i+1})(c+\delta_i)-2(c+\delta_i)+1} = 0$$

c.q.d.

El teorema anterior nos muestra que la sucesión $\{y_i\}_{i=0}^{\infty}$ converge más rápidamente hacia x^* que la solución dada.

Considérese ahora una sucesión $\{x_{i+1} = g(x_i)\}_{i=0}^{\infty}$, generada por un método de orden menor que 2, que converja hacia x^* y tal que $|g'(x^*)| < 1$. Por ser el orden de convergencia superior o igual a 1 pero inferior a 2 se verificará, a partir de un desarrollo de Taylor y teniendo en cuenta que $g'(x^*) \neq 0$, que:

$$\frac{x_{i+1} - x^*}{x_i - x^*} = g'(x^*) + O((x_i - x^*)^2)$$

por lo que llamando $c = g'(x^*)$ y $\delta_i = O((x_i - x^*)^2)$ se verificarán las condiciones del teorema anterior.

Todo lo anterior puede utilizarse para elaborar un algoritmo que mejore la velocidad de convergencia de los métodos de orden de convergencia inferior a 2. Este algoritmo, conocido también con el nombre de método de Steffensen, es el que se recoge a continuación:

Algoritmo del método de Steffensen:

Dada la ecuación $x = g(x)$, el indicador de precisión ε, un valor máximo del número de iteraciones que se permiten realizar ($maxiter$) y un punto x_0 con el que inicializar el proceso,

$tol \leftarrow 2\varepsilon$

$iteración \leftarrow 0$

Mientras ($(iteración < maxiter)$ **y** $(tol > \varepsilon)$), **hacer:**

$$\begin{aligned} x_1 &\leftarrow g(x_0) \\ x_2 &\leftarrow g(x_1) \end{aligned}$$

Si $((x_2 - 2x_1 + x_0) \neq 0)$ **entonces:**

$$x_3 \leftarrow x_0 - \frac{(x_1 - x_0)^2}{x_2 - 2x_1 + x_0}$$

si no:

hacer $tol \leftarrow \frac{\varepsilon}{2}$

fin condición.

$$tol \leftarrow |x_3 - x_0|$$

$$iteración \leftarrow iteración + 1$$

$$x_0 \leftarrow x_3$$

Fin bucle condicional.

Si $(tol < \varepsilon)$ **entonces:**

tomar x_3 como solución

si no:

Escribir un mensaje de error en el proceso de cálculo

fin condición.

Fin del algoritmo.

Observación 2.3.24 *En el algoritmo anterior se calcula x_3 en cada iteración (valor de la sucesión corregida) tras asegurarse de que el denominador ($\Delta^2 x_0$) es no nulo. Si lo fuese se da como solución aproximada la obtenida en la iteración anterior.*

Ejemplo 2.3.6 *Puesto que estamos llegando al final de la materia que se va a incluir en estos guiones sobre los métodos de resolución de una ecuación no lineal, ilustremos el funcionamiento del método de Steffensen sobre el primero de los ejemplos (tomados de O.T. Hanna & O.C. Sandall [9]) con los que abríamos este tema. Recordemos que su enunciado era:*

La ecuación de Peng-Robinson es una ecuación de estado que proporciona la presión P de un gas mediante:

$$P = \frac{RT}{V-b} - \frac{a}{V(V+b)+b(V-b)}$$

donde a y b son constantes, T es la temperatura absoluta a la que se encuentra el gas, V es el volumen específico y R es la constante de los gases perfectos ($8{,}31441\, J/(mol^o K)$). Para el CO_2 las constantes a y b toman los valores $a = 364{,}61\, m^6 kPa/(kg\, mol)^2$ y $b = 0{,}02664\, m^3/kg\, mol$. Supongamos que se desea encontrar la densidad (es decir $1/V$) del CO_2 a una presión de 10^4 kPa y a una temperatura de 340^o K usando la ecuación de Peng-Robinson. Ello implicaría tener que encontrar el valor de V para el que:

$$10^4 = \frac{340R}{V-0{,}02664} - \frac{364{,}61}{V(V+0{,}02664)+0{,}02664(V-0{,}02664)}$$

Para aplicar el método de aproximaciones sucesivas a esta ecuación no lineal puede despejarse una de las incógnitas de la forma que sigue:

$$10^4(V-0{,}02664) = 340R - \frac{364{,}61(V-0{,}02664)}{V(V+0{,}02664)+0{,}02664(V-0{,}02664)} \Rightarrow$$

$$V = 0{,}02664 + 340\cdot 10^{-4}R - \frac{364{,}61\cdot 10^{-4}(V-0{,}02664)}{V(V+0{,}02664)+0{,}02664(V-0{,}02664)} = g(V)$$

que nos conduce al esquema iterativo:

$$V_{i+1} = 0{,}02664 + 340\cdot 10^{-4}R - \frac{364{,}61\cdot 10^{-4}(V_i-0{,}02664)}{V_i(V_i+0{,}02664)+0{,}02664(V_i-0{,}02664)}$$

Como punto de partida puede considerarse que si el gas fuese perfecto la ecuación de los gases perfectos

$$P = \frac{RT}{V} \Rightarrow V = \frac{RT}{P}$$

nos conduciría en este caso a

$$V_0 = \frac{8{,}31441\cdot 340}{10^4} \approx 2866\cdot 10^{-4}$$

En los resultados que siguen, además del valor de V_i y de $g(V_i)$ se proporciona el valor del residuo r_i estimado como:

$$r_i = \frac{340R}{V_i-0{,}02664} - \frac{364{,}61}{V_i(V_i+0{,}02664)+0{,}02664(V_i-0{,}02664)} - 10^4$$

A partir del valor inicial considerado, el método de aproximaciones sucesivas nos proporciona la siguiente tabla de valores (realizando iteraciones hasta que la distancia entre dos valores consecutivos se hace inferior a 10^{-8} y $|r_i| < 10^{-5}$):

Iteración	V_i	$g(V_i)$	r_i
1	0,211311226884	0,187353020426	−1297,34376394
2	0,187353020426	0,177275001886	−627,081646107
3	0,177275001886	0,172576048103	−311,943022813
4	0,172576048103	0,170283680111	−157,080311681
5	0,170283680111	0,169141118216	−79,5413967845
6	0,169141118216	0,168565615384	−40,3858467251
...	.....	.....	.....
27	0,167973123629	0,167973123232	−0,0000280358854345
28	0,167973123232	0,167973123031	−0,0000142727608292
29	0,167973123031	0,167973123232	−0,00000726610596669

por lo que se puede tomar $V^*_{a.s.} \approx 0{,}167973123031$ (y por tanto la densidad buscada sería su inversa 5,9533333..). La determinación de este valor ha costado 29 iteraciones del método de aproximaciones sucesivas. Si en su lugar se hubiera utilizado el algoritmo de Steffensen se obtendría la siguiente tabla de valores (con los mismos controles de tolerancia):

Iteración	V_i	$g(V_i)$	r_i
1	0,176170684169	0,172043245471	$-276{,}026202967$
2	0,168072867021	0,168023886001	$-3{,}46319927617$
3	0,167973138878	0,167973130996	$-0{,}000557724441576$
4	0,167973122821	0,167973122821	$-0{,}14049206242810^{-10}$
5	0,167973122821	0,167973122821	$0{,}37303493627410^{-12}$

Obsérvese que en la última iteración se repiten los valores de x_i y $g(x_i)$ de la cuarta iteración aunque el valor del residuo difiere. Ello es debido a que x_4 y x_5 no son iguales pero se diferencian a partir de valores decimales que están más allá de los considerados en este ejemplo.

Nótese asimismo que en sólo 5 iteraciones del método de Steffensen se ha logrado una solución $V^*_{St} \approx 0{,}167973122821$ muy similar a la proporcionada por el método de aproximaciones sucesivas ($V^*_{a.s.} \approx 0{,}167973123031$) habiendo diferencias entre ellas del orden de 10^{-9} pero que hacen mucho menor el residuo. Es decir, con menos iteraciones se ha logrado una solución más precisa. Eso sí, siendo honestos, debemos reconocer que en cada iteración del método de Steffensen se han realizado 2 evaluaciones de la función $g(x)$ (frente a una en el método de aproximaciones sucesivas) y seis operaciones elementales más (las que nos proporcionaban el valor "corregido" x_3 en cada iteración). No obstante, este mayor esfuerzo computacional en cada iteración, al menos en este caso, merece la pena.

2.3.6. Algunos comentarios finales sobre los métodos de resolución de una ecuación no lineal

1^o) Los métodos presentados en este apartado constituyen los métodos más básicos de resolución de ecuaciones no lineales. Existen muchos otros aplicables a ecuaciones no lineales de determinado tipo. Entre ellos merece la pena destacar los métodos de resolución de ecuaciones polinómicas tales como los de Bairstow, Bernoulli, Dandelin-Graeffe, etc... que pueden encontrarse en la bibliografía sobre este tema (por ejemplo en C. Conde & G. Winter [5] o en E. Durand [6] o en D. Kincaid y W. Cheney [10]).

2^o) En el esquema iterativo del método de la secante, de forma gráfica, se sigue la recta secante al grafo de la función que pasa por los puntos $(x_i, f(x_i))$ y $(x_{i-1}, f(x_{i-1}))$. Algunas variantes de este método consisten en ajustar en las tres últimas aproximaciones halladas una parábola que pase por

$(x_{i-2}, f(x_{i-2}))$, $(x_{i-1}, f(x_{i-1}))$ y $(x_i, f(x_i))$ determinando con ella un nuevo punto, el x_{i+1}, como uno de los puntos en que dicha parábola corta al eje de abscisas y continuando con él el proceso. Esta es la idea en que se basa el denominado método de Müller cuya descripción puede encontrarse, por ejemplo en R. Burden & J.D. Faires [3].

3^o) Asimismo existen variantes de los métodos antes expuestos para el caso de trabajar con funciones de variable compleja. En D. Kincaid & W. Cheney [10] o en C. Conde & G. Winter [5] podrán encontrarse, por ejemplo, las adaptaciones del método de Newton al caso complejo.

4^o) Otra familia de métodos para la búsqueda de soluciones de una ecuación no lineal se sustenta en los métodos de optimización. Para ello se puede construir la función $r(x) = f^2(x)$. Esta función $r(x)$ siempre tomará valores positivos o nulos por lo que, si $f(x) = 0$ admite solución, los mínimos de $r(x)$ tendrán valor 0. Pueden entonces emplearse técnicas de minimización (que desbordan los objetivos de este tema) para determinar los puntos mínimos de la función $r(x)$. Algoritmos tales como el de Marquardt-Levenberg o la familia de métodos de optimización global (algoritmos genéticos, métodos de "recocido simulado", etc...) pueden ser aplicados a $r(x)$. Remitimos al lector a la bibliografía de este tema (por ejemplo J.L. de la Fuente O'Connor [7]) para un estudio de métodos de optimización.

5^o) Otros métodos muy en boca hoy en día son los métodos de continuación (o de homotopía). En ellos, dada la ecuación $f(x) = 0$, se considera que la variable x depende a su vez (mediante una función desconocida "a priori") de un parámetro $\lambda \in [0, 1]$ de forma tal que cuando λ tome el valor 1 se verifique que $x(1) = x^*$, siendo x^* una solución de la ecuación planteada. Por ejemplo, es habitual, dado un valor inicial x_0, considerar que $f(x(\lambda)) = (1 - \lambda)f(x_0)$. De esta forma cuando λ tome el valor 1 se deberá verificar que $f(x(1)) = 0$ con lo que $x^* = x(1)$. Con esta elección se tendrá que:

$$\left\{ \begin{array}{c} \frac{df}{dx}(x(\lambda))\frac{dx}{d\lambda}(\lambda) = -f(x_0) \\ x(0) = x_0 \end{array} \right\} \qquad \lambda \in [0, 1]$$

En las expresiones anteriores $\frac{df}{dx}(x(\lambda))$ puede calcularse (pues $f(x)$) es conocida) y será una expresión que dependerá de $x(\lambda)$. Por tanto las ecuaciones anteriores representan un problema de valor inicial cuya resolución nos determinará la expresión de la función $x(\lambda)$ buscada.

A su vez la resolución del problema de valor inicial se realiza en la práctica mediante métodos numéricos como los que se abordarán en el capítulo siguiente de estos guiones. Por ello retomaremos este método en el próximo capítulo como

una aplicación de los métodos de resolución numérica de problemas de valor inicial.

6^o) Una vez determinada una solución de la ecuación $f(x) = 0$, otras raíces pueden buscarse utilizando la denominada técnica de deflacción en la que $f(x)$ se expresa como:

$$f(x) = (x - x^*)h(x) \Leftrightarrow h(x) = \frac{f(x)}{(x - x^*)}$$

y se buscan otras posibles raíces de la ecuación $h(x) = 0$. En rigor la función $h(x)$ no estaría definida en el punto $x = x^*$ (pues en él aparecería una división por 0). Por ello, si se tiene necesidad de trabajar en dicho punto, se definirá:

$$h(x^*) = \lim_{x \to x^*} \frac{f(x)}{(x - x^*)}$$

7^o) Habitualmente, si una función $f(x)$ tiene diferentes raíces, dependiendo del punto de partida x_0 con el que se arranque el proceso iterativo se podrá determinar una u otra de ellas. Al conjunto de puntos que tomados como valor de arranque del método iterativo conducen a la solución x^* se le llama **dominio** (o cuenca) **de atracción** de la raíz. Por ejemplo, si $f(x) = x^2 - \frac{1}{4}$ es obvio que sus raíces son $\pm\frac{1}{2}$. Si se buscan las raíces de esta ecuación mediante el método de Newton puede comprobarse que:

a) Si $x_0 \leq -\frac{1}{2}$ la solución que proporciona el método es $x^* = -\frac{1}{2}$

b) Si $-\frac{1}{2} < x_0$ el método conduce a la solución $x^* = \frac{1}{2}$.

Por ello para la ecuación considerada y para el método de Newton-Raphson el dominio de atracción de la raíz $\frac{-1}{2}$ es $]-\infty, -\frac{1}{2}]$ y el de la raíz $\frac{1}{2}$ es $]\frac{-1}{2}, \infty[$.

Pero si la misma ecuación se intenta resolver mediante el método de aproximaciones sucesivas utilizando el esquema iterativo:

$$x_{i+1} = x_i^2 + x_i - \frac{1}{4}$$

se tiene que:

a) Si $x_0 < \frac{-3}{2}$ o $x_0 > \frac{1}{2}$ el método diverge

b) So $x_0 = \frac{-3}{2}$ o $x_0 = \frac{1}{2}$ el método converge (en una iteración) a $\frac{1}{2}$

c) Para otros valores de x_0 el método converge hacia $\frac{-1}{2}$

Por ello para la ecuación considerada y para el método de aproximaciones sucesivas el dominio de atracción de la raíz $\frac{-1}{2}$ es $]\frac{-3}{2}, \frac{1}{2}]$ y el de la raíz $\frac{1}{2}$ es $\{\frac{-3}{2}, \frac{1}{2}\}$.

8^o) Debido al comentario anterior es aconsejable buscar intervalos en los que la función sólo tenga una raíz y tome valores con signos alternos en los

extremos del intervalo, combinando los métodos de resolución con el método de bipartición. En Press et al. [12] pueden encontrarse métodos de "separación de raíces" (el proceso de *bracketing*) que permiten buscar tales intervalos mediante la exploración de los valores de la función en diferentes puntos. También en Press [12] pueden encontrarse programas en FORTRAN 90 que recogen los métodos que hemos presentado anteriormente combinándolos con el método de bipartición.

9^o) La existencia de raíces múltiples, como ya se señaló anteriormente, puede ralentizar la aproximación hacia las raíces de una función. Pero no es este el único inconveniente que tales raíces presentan. En efecto, si la multiplicidad de una raíz es elevada también puede tenerse la sensación de haber encontrado dicha raíz estando relativamente alejados de ella. Así si en el proceso de búsqueda de una raíz x^* de multiplicidad m de la función $f(x)$ se tiene en un momento un valor aproximado α, el valor de la función en α será: $f(\alpha) = (\alpha - x^*)^m \varphi(\alpha)$. Si α es "próximo" a x^*, aunque no lo suficientemente próximo como para que sea su raíz, y m es elevado el valor de $f(\alpha)$ puede ser, computacionalmente hablando, nulo sin que esta sea la raíz buscada. Un ejemplo para aclarar lo anterior (tomado de Shampine & Allen & Pruess [13]) es aquel en el que x^* es una raíz de multiplicidad 10 y $\alpha - x^* = 10^{-4}$. En ese caso, la precisión puede no ser la deseada y sin embargo: $f(\alpha) = 10^{-40} g(\alpha)$ lo que implicaría que si se está trabajando en un ordenador con precisión simple y $|g(\alpha)| < 1$, el ordenador toma $f(\alpha)$ como 0 por lo que "detecta" una raíz en α.

10^o) Para evitar problemas como los anteriores, en ocasiones es útil "escalar" la ecuación a resolver. Ello consiste simplemente en reemplazar la ecuación $f(x) = 0$ por otra de la forma $F(x) = s(x)f(x) = 0$ que admitirá, entre otras, las soluciones de la ecuación inicial. La función $s(x)$ debe ser escogida adecuadamente y se denomina función de escala (o factor de escala cuando es una constante). Ello puede contribuir a aumentar los valores de la función en las supuestas raíces. Así, por poner un ejemplo obvio, si se desean encontrar raíces de la ecuación $10^{-38}\cos(x) = 0$, y se trabaja en un ordenador con precisión simple, la estimación en cualquier punto x de la recta real del valor $f(x) = 10^{-38}\cos(x)$ es siempre nulo (pues en simple precisión no se pueden almacenar números tan pequeños). En otros términos, en simple precisión cualquier valor real sería para el ordenador solución de la ecuación dada. Si la ecuación se escala multiplicándola por 10^{38} se tiene ahora $F(x) = \cos(x) = 0$ y sobre esta ecuación el ordenador ya puede distinguir si se está en las cercanías de una raíz o no.

No siempre es tan sencillo el proceso de "escalar" ecuaciones. En Shampine & Allen & Pruess [13] pueden encontrarse algunos ejemplos muy instructivos

y simples a este respecto.

11^o) En algunos códigos, la búsqueda de raíces múltiples de $f(x)$ se realiza buscando las raíces simples de $h(x) = f(x)/f'(x)$.

2.4. Métodos de resolución de sistemas de ecuaciones no lineales.

Los métodos de aproximaciones sucesivas, Newton-Raphson y sus variantes, presentados en el apartado anterior para el caso de una única ecuación pueden extenderse fácilmente al caso de sistemas de n ecuaciones no lineales con n incógnitas.

Este tipo de sistemas los escribiremos en la forma:

$$\begin{cases} f_1(x_1, x_2, ..., x_n) = 0 \\ f_2(x_1, x_2, ..., x_n) = 0 \\ ..... \\ f_n(x_1, x_2, ..., x_n) = 0 \end{cases}$$

o más brevemente como $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ donde $\mathbf{0}$ es el vector nulo de n componentes, $\mathbf{x}$ es un vector de $I\!R^n$ y $\mathbf{f}$ es la función vectorial dependiente de n variables reales dada por:

$$\mathbf{f} : I\!R^n \to I\!R^n$$

$$\mathbf{x} = (x_1, x_2, ..., x_n)^T \to \mathbf{f}(\mathbf{x}) = \begin{cases} f_1(x_1, x_2, ..., x_n) \\ f_2(x_1, x_2, ..., x_n) \\ ..... \\ f_n(x_1, x_2, ..., x_n) \end{cases}$$

Al igual que se indicó para el caso de un única ecuación, no debe confundirse esta forma de representar el sistema de ecuaciones con el que la función $\mathbf{f}(\mathbf{x})$ sea idénticamente nula. En efecto con la notación $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ estaremos representando de forma breve en este apartado el problema siguiente: "Encontrar, si es posible, algún vector $\mathbf{x}^*$ de $I\!R^n$ para el que se verifique que $\mathbf{f}(\mathbf{x}^*) = \mathbf{0}$"

2.4.1. El método de aproximaciones sucesivas para sistemas de n ecuaciones no lineales

Este método se basa en el teorema del punto fijo. Para su aplicación, al igual que en el caso de una ecuación, se transforma el sistema $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ en otro equivalente (es decir con las mismas soluciones, de la forma $\mathbf{x} = \mathbf{g}(\mathbf{x})$).

La forma de realizar esta transformación no es única. Por ejemplo, si se suma el vector $\mathbf{x}$ en ambos términos del sistema se tendrá que:

$$\mathbf{x} = \mathbf{x} + \mathbf{f}(\mathbf{x})$$

por lo que podría tomarse $\mathbf{g}(\mathbf{x}) = \mathbf{x} + \mathbf{f}(\mathbf{x})$.

Pero también podría realizarse el proceso:

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \Longleftrightarrow \mathbf{x} = \mathbf{x} - \mathbf{f}(\mathbf{x})$$

por lo que $\mathbf{g}(\mathbf{x}) = \mathbf{x} - \mathbf{f}(\mathbf{x})$ también sería otra forma de realizar la transformación antes aludida.

O podría despejarse (total o parcialmente) de la primera ecuación x_1, de la segunda x_2, ... y de la n-ésima ecuación x_n con lo que también escribiríamos el sistema en la forma deseada.

También ahora debe ponerse atención en el hecho de que a pesar de que existan muy diferentes formas de rescribir $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ en la forma $\mathbf{x} = \mathbf{g}(\mathbf{x})$ no todas son equivalentes para aplicar sobre ellas el método de aproximaciones sucesivas y debe seleccionarse con cuidado la forma en la que se da este paso para que, a la luz de lo que nos indiquen los teoremas y propiedades que a continuación presentaremos, se garantice la convergencia (lo más rápidamente posible) del método.

Una vez escrito el sistema de ecuaciones en la forma $\mathbf{x} = \mathbf{g}(\mathbf{x})$ el método de aproximaciones sucesivas consiste en seleccionar "arbitrariamente" (aunque mejor cuanto más cercano esté a la solución buscada) un vector $\mathbf{x}^{(0)}$ con el que inicializar el esquema iterativo siguiente:

$$\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)}) \qquad (i = 0, 1, 2, ...)$$

De esta forma se genera una sucesión de vectores $\{\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)})\}_{i=0}^{\infty}$. Y según el teorema del punto fijo se tiene el resultado siguiente:

Teorema 2.4.1 *Si para alguna norma definida sobre $\mathbb{R}^n$ se verifica que $\mathbf{g}(\mathbf{x})$ es una contracción sobre un dominio D cerrado de $\mathbb{R}^n$ y $\mathbf{x}^{(0)}$ es un punto de D, entonces el método de aproximaciones sucesivas antes planteado converge hacia la única solución en D de la ecuación $\mathbf{x} = \mathbf{g}(\mathbf{x})$.*

Demostración: Por ser D un cerrado de $\mathbb{R}^n$ será completo. Y por aplicación directa del teorema de punto fijo, al ser $\mathbf{g}(\mathbf{x})$ una contracción se verificará que:

1^o) Sólo existirá un punto $\mathbf{x}^*$ de D para el que $\mathbf{x}^* = \mathbf{g}(\mathbf{x}^*)$

2^o) La sucesión $\{\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)})\}_{i=0}^{\infty}$ converge hacia $\mathbf{x}^*$ en el sentido de la norma utilizada.

Y si la sucesión converge para la norma con la que se trabaja en $\mathbb{R}^n$ también lo hará para cualquier norma definida sobre $\mathbb{R}^n$ pues al ser $\mathbb{R}^n$ de dimensión finita todas las normas sobre él definidas serán equivalentes.

c.q.d.

Observación 2.4.1 *Es conveniente observar que*

1. *Puesto que el sistema de ecuaciones* $\mathbf{x} = \mathbf{g}(\mathbf{x})$ *es equivalente al sistema* $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ *en las condiciones del teorema anterior el método de aproximaciones sucesivas converge hacia una solución* $\mathbf{x}^*$ *del sistema* $\mathbf{f}(\mathbf{x}) = \mathbf{0}$.

2. *En otros términos las funciones* $\mathbf{g}(\mathbf{x})$ *con las que nos interesa trabajar son aquellas que, para alguna norma vectorial* $\|\|$, *sean una contracción sobre algún dominio cerrado* D *de* $\mathbb{R}^n$.

3. *El teorema anterior nos proporciona unas condiciones que aseguran la convergencia del método. Son pues condiciones suficientes para la convergencia del método. Pero el teorema no dice nada sobre su necesidad. Y en efecto puede haber situaciones particulares en las que no verificándose las condiciones del teorema (que* $\mathbf{g}(\mathbf{x})$ *sea una contracción sobre el dominio* D *en el que se busca el vector solución) el método también converja. A este respecto el teorema anterior se limita a no asegurar el buen funcionamiento del método en el caso de que no se satisfagan las hipótesis en él hechas pero sin impedir su buen funcionamiento en dichos casos.*

El demostrar que una aplicación $\mathbf{g}(\mathbf{x})$ es una contracción para alguna norma, mediante la determinación de su constante de Lipschitz puede, en ciertas ocasiones, resultar algo laborioso. Por ello pueden contemplarse variantes más restrictivas (pero más fácilmente aplicables en la práctica) del teorema anterior. En ellas, asumiendo que todas las componentes de la función $\mathbf{g}(\mathbf{x})$ son derivables en todos los puntos del dominio de trabajo, se utiliza la matriz Jacobiana de la aplicación $\mathbf{g}(\mathbf{x})$, que denotaremos por $[\mathbf{J}_\mathbf{g}(\mathbf{x})]$ y que recordamos que se definía mediante:

$$[\mathbf{J}_\mathbf{g}(\mathbf{x})] = \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{x}) & \frac{\partial g_1}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial g_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial g_2}{\partial x_1}(\mathbf{x}) & \frac{\partial g_2}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial g_2}{\partial x_n}(\mathbf{x}) \\ \dots & \dots & \dots & \dots \\ \frac{\partial g_n}{\partial x_1}(\mathbf{x}) & \frac{\partial g_n}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial g_n}{\partial x_n}(\mathbf{x}) \end{bmatrix}$$

Con esta notación se tiene el siguiente teorema:

Teorema 2.4.2 *Si* $\mathbf{g}(\mathbf{x})$ *es una aplicación de clase* $\left(C^1(D)\right)^n$ *y que toma valores en el cerrado* D *verificando para alguna norma matricial subordinada la condición:*

$$\exists k < 1 \ / \qquad \|[\mathbf{J}_\mathbf{g}(\mathbf{x})]\| \leq k < 1 \qquad \forall \mathbf{x} \in D$$

entonces la sucesión $\left\{\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)})\right\}_{i=0}^{\infty}$ *generada, a partir de cualquier* $\mathbf{x}^{(0)} \in D$, *converge hacia la única solución de la ecuación* $\mathbf{x} = \mathbf{g}(\mathbf{x})$ *en* D.

Demostración: Por aplicación del teorema del valor medio se verificará que:

$$\forall \mathbf{x}, \mathbf{y} \in D \quad \exists \mathbf{z} \in \overset{\circ}{D} \ / \quad \mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y}) = [\mathbf{J_g}(\mathbf{z})]\,(\mathbf{x} - \mathbf{y})$$

y por haber supuesto que para alguna norma matricial subordinada, el valor de $\|[\mathbf{J_g}(\mathbf{x})]\|$ estaba acotado por k, trabajando con la norma vectorial a la que está subordinada la anterior norma matricial se tendrá que:

$$\begin{aligned} \forall \mathbf{x}, \mathbf{y} \quad \in \quad & D: \quad ||\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})|| = \|[\mathbf{J_g}(\mathbf{z})]\,(\mathbf{x} - \mathbf{y})\| \leq \\ \leq \quad & \|[\mathbf{J_g}(\mathbf{z})]\|\,\|\mathbf{x} - \mathbf{y}\| \leq k||\mathbf{x} - \mathbf{y}|| < ||\mathbf{x} - \mathbf{y}|| \end{aligned}$$

por lo que, teniendo en cuenta que $\mathbf{g} : D \to D$, resulta que $\mathbf{g}(\mathbf{x})$ es una contracción. Aplicando el teorema precedente quedará totalmente demostrado.

c.q.d.

Observación 2.4.2 *Cuando en las aplicaciones se utilice este teorema para comprobar que la aplicación considerada es una contracción se tomará como aproximación de la constante de Lipschitz el valor* $k = Máx_{\mathbf{x} \in D} \{\|[\mathbf{J_g}(\mathbf{x})]\|\}$.

Los dos teoremas precedentes establecen condiciones suficientes de convergencia *global* del método sobre un dominio cerrado D (esto es independientemente del punto $\mathbf{x}^{(0)} \in D$ con el que se inicialice el proceso iterativo). Cuando se conozca un cierto entorno de la solución buscada pueden establecerse resultados de convergencia *local* (es decir, para puntos $\mathbf{x}^{(0)}$ suficientemente próximos a la solución). Así por ejemplo, se tiene el siguiente teorema:

Teorema 2.4.3 *Si existe una solución* $\mathbf{x}^*$ *de la ecuación* $\mathbf{x} = \mathbf{g}(\mathbf{x})$ *en un dominio cerrado* D *en el que* $\mathbf{g}(\mathbf{x})$ *es de clase* $\left(C^1(D)\right)^n$ *y para alguna norma matricial subordinada* $\|[\mathbf{J_g}(\mathbf{x}^*)]\| < 1$ *entonces existe un valor* $\delta > 0$ *tal que si, trabajando con la norma vectorial asociada a la norma matricial anterior,* $||\mathbf{x}^* - \mathbf{x}^{(0)}|| < \delta$ *la sucesión* $\left\{\mathbf{x}^{(i+1)} = \mathbf{g}(\mathbf{x}^{(i)})\right\}_{i=0}^{\infty}$ *verifica que:*

a) $||\mathbf{x}^* - \mathbf{x}^{(i)}|| < \delta \quad \forall \mathbf{x}^{(i)}$

b) $\lim_{i \to \infty} \mathbf{x}^{(i)} = \mathbf{x}^*$.

Demostración: Por ser $\dfrac{\partial g_i}{\partial x_j}(\mathbf{x})$ (i, j = 1, ..., n) continuas en todo $\mathbf{x} \in D$ existirá una bola abierta de centro $\mathbf{x}^*$ y radio δ' , $B(\mathbf{x}^*, \delta')$ tal que en ella se verifique:

$$\|[\mathbf{J_g}(\mathbf{x})]\| \leq k < 1 \qquad \forall \mathbf{x} \in B(\mathbf{x}^*, \delta')$$

Considerando un valor $\delta < \delta'$ se tendrá por tanto que,

$$\|[\mathbf{J_g}(\mathbf{x})]\| \leq k < 1 \qquad \forall \mathbf{x} \in B'(\mathbf{x}^*, \delta)$$

donde $B'(\mathbf{x}^*, \delta)$ es una bola cerrada de centro $\mathbf{x}^*$ y radio δ.

Consecuentemente $\mathbf{g}(\mathbf{x})$ es una contracción en $B'(\mathbf{x}^*, \delta)$. Ello conduce a que

$$\forall \mathbf{x}^{(i)} \in \left\{\mathbf{x}^{(i)}\right\}_{i=1}^{\infty} : \quad \left\|\mathbf{x}^{(i)} - \mathbf{x}^*\right\| = \left\|\mathbf{g}(\mathbf{x}^{(i-1)}) - \mathbf{g}(\mathbf{x}^*)\right\| < k \left\|\mathbf{x}^{(i-1)} - \mathbf{x}^*\right\| <$$

$$< k^2 \left\|\mathbf{x}^{(i-2)} - \mathbf{x}^*\right\| < ..... < k^i \left\|\mathbf{x}^{(0)} - \mathbf{x}^*\right\| < k^i \delta < \delta$$

Por otra parte, al ser $k < 1$ bastará con escoger el índice i suficientemente elevado para que todos los elementos de la sucesión con índice mayor que i sean tan cercanos a x^* como se desee. En otros términos $\mathbf{x}^* = \lim_{i \to \infty} \mathbf{x}^{(i)}$.

c.q.d.

Observación 2.4.3 *Cuanto menor sea el valor de* $\|[\mathbf{J}_{\mathbf{g}}(\mathbf{x}^*)]\|$ *menor será la cota de* $\left\|\mathbf{x}^{(i)} - \mathbf{x}^*\right\|$ *obtenida en la demostración anterior y por ello mejor será la convergencia del método si se parte de un punto suficientemente cercano a la solución.*

Los teoremas precedentes establecen condiciones suficientes para que el método de aproximaciones sucesivas converja. De esta forma, si se verifican las hipótesis de cualquiera de los teoremas anteriores, seleccionado el punto inicial $\mathbf{x}^{(0)}$, todo consistirá en generar a partir de él $\mathbf{x}^{(1)} = \mathbf{g}(\mathbf{x}^{(0)})$, y a partir de este $\mathbf{x}^{(2)} = \mathbf{g}(\mathbf{x}^{(1)})$, y así sucesivamente. Tras hacer infinitas iteraciones alcanzaríamos la solución buscada. Pero, evidentemente, no pueden realizarse "infinitas" iteraciones. Por ello la cuestión que nos planteamos ahora es ¿cuántas iteraciones nos garantizarían una precisión determinada?. La respuesta a este dilema nos la proporciona el siguiente teorema:

Teorema 2.4.4 *Siendo* $\mathbf{g}(\mathbf{x})$ *una contracción definida en el dominio cerrado* D *la distancia (en el sentido de la norma vectorial con la que se demuestre que* $\mathbf{g}(\mathbf{x})$ *es una contracción) entre la única solución* $\mathbf{x}^*$ *de la ecuación* $\mathbf{x} = \mathbf{g}(\mathbf{x})$ *y cualquier elemento de la sucesión* $\left\{\mathbf{x}^{(n)} = \mathbf{g}(\mathbf{x}^{(n-1)})\right\}_{n=0}^{\infty}$, *generada a partir de cualquier valor* $\mathbf{x}^{(0)} \in D$, *está acotada mediante la expresión:*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(n)}\right\| \leq \frac{k^n}{1-k} \left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|$$

donde k *es la constante de Lipschitz de la contracción.*

Demostración: Véase la "Nota 2^a" realizada tras la demostración del teorema del punto fijo.

c.q.d.

Observación 2.4.4 *Obsérvese lo siguiente:*

1. *Bajo las hipótesis del teorema precedente, si se desea asegurar que la norma (siempre en el sentido de la norma vectorial para la que* $\mathbf{g}(\mathbf{x})$ *es una contracción) del error cometido es menor que un cierto valor* ε *la expresión anterior nos dice que deben realizarse un número* N *de iteraciones tal que:*

$$\frac{k^N}{1-k}\left\|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right\| < \varepsilon \Rightarrow N > \frac{\log\left(\frac{\varepsilon(1-k)}{\left\|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right\|}\right)}{\log(k)}$$

2. *Si no se conoce el valor exacto de la constante de Lipschitz de la aplicación puede estimarse de forma aproximada de diferentes maneras. Por ejemplo, tras cada iteración del método podría obtenerse una aproximación de dicha constante estimando la norma de la matriz jacobiana en el último punto hallado.*

En la práctica, en lugar de calcular a priori el número de iteraciones a realizar se va estimando en cada iteración la distancia del punto en ella hallado a la solución exacta. Esta estimación se realiza simplemente evaluando la diferencia entre las dos últimas aproximaciones halladas que, cuando $\mathbf{g}(\mathbf{x})$ es una contracción, son un indicador de la cercanía a la solución exacta en virtud del siguiente teorema:

Teorema 2.4.5 *Siendo* $\mathbf{g}(\mathbf{x})$ *una contracción definida en el cerrado* D *la distancia entre la única solución* $\mathbf{x}^*$ *en* D *de la ecuación* $\mathbf{x}=\mathbf{g}(\mathbf{x})$ *y cualquier elemento de la sucesión*

$$\left\{\mathbf{x}^{(n)}=\mathbf{g}(\mathbf{x}^{(n-1)})\right\}_{n=0}^{\infty},$$

generada a partir de cualquier punto $\mathbf{x}^{(0)}\in D$*, está acotada mediante la expresión:*

$$\left\|\mathbf{x}^*-\mathbf{x}^{(n)}\right\| \leq \frac{k}{1-k}\left\|\mathbf{x}^{(n)}-\mathbf{x}^{(n-1)}\right\|$$

donde k *es la constante de Lipschitz de la contracción.*

Demostración: Véase la segunda observación realizada tras la demostración del teorema del punto fijo.

c.q.d.

Con ello, cuando $\mathbf{g}(\mathbf{x})$ sea una contracción, al ser $k<1$, bastará con hacer un número de iteraciones tal que $\left\|\mathbf{x}^{(n)}-\mathbf{x}^{(n-1)}\right\|$ sea suficientemente pequeño para asegurar que $\left\|\mathbf{x}^*-\mathbf{x}^{(n)}\right\|$ también es pequeño. Este control de la convergencia debe acompañarse con la limitación del número de iteraciones a realizar, en previsión de los casos en los que, no siendo $\mathbf{g}(\mathbf{x})$ una contracción, el método

no converja, y con el control del valor de $\left\|\mathbf{f}(\mathbf{x}^{(n)})\right\| = \left\|\mathbf{g}(\mathbf{x}^{(n)}) - \mathbf{x}\right\|$ en cada iteración. Más concretamente un algoritmo del método de aproximaciones sucesivas, en el que se parte de la ecuación equivalente $\mathbf{x} = \mathbf{g}(\mathbf{x})$ es el siguiente:

Algoritmo del método de aproximaciones sucesivas:

Dada la ecuación $\mathbf{x} = \mathbf{g}(\mathbf{x})$, los indicadores de precisión ε y δ, un valor máximo del número de iteraciones que se permiten realizar ($maxiter$) y un punto $\mathbf{x}^{(0)}$ con el que inicializar el proceso,

$tolx \leftarrow 2\varepsilon$

$tolf \leftarrow 2\delta$

$iteración \leftarrow 0$

Mientras ($(iteración < maxiter)$ **y** $((tolx > \varepsilon)\ o\ (tolf > \delta))$**, hacer:**

$$\mathbf{x}^{(1)} \leftarrow \mathbf{g}(\mathbf{x}^{(0)})$$

$$\begin{aligned} tolx &\leftarrow \left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\| \\ tolf &\leftarrow \left\|\mathbf{g}(\mathbf{x}^{(1)}) - \mathbf{x}^{(1)}\right\| \end{aligned}$$

$$iteración \leftarrow iteración + 1$$

$$\mathbf{x}^{(0)} \leftarrow \mathbf{x}^{(1)}$$

Fin bucle condicional.

Si $((tolx < \varepsilon)\ y\ (tolf < \delta))$ **entonces:**

tomar $\mathbf{x}^{(1)}$ como solución

si no:

Escribir un mensaje de error en el proceso de cálculo

fin condición.

Fin del algoritmo.

Ilustremos el método que se acaba de describir mediante un ejemplo.

Ejemplo 2.4.1 *(Cortesía del Pr. B. Coto): Según el modelo de Wilson las expresiones de los coeficientes de actividad a dilución infinita (γ_i^∞) de una mezcla binaria están dadas por las expresiones:*

$$\begin{aligned} \ln(\gamma_1^\infty) &= 1 - \ln(\Lambda_{12}) - \Lambda_{21} \\ \ln(\gamma_2^\infty) &= 1 - \ln(\Lambda_{21}) - \Lambda_{12} \end{aligned}$$

donde Λ_{12} y Λ_{21} son los parámetros binarios de la mezcla. Se desea saber con el modelo de Wilson el valor de los parámetros binarios en los dos casos siguientes:

a) En el caso de una mezcla binaria ideal en la que los coeficientes de actividad a dilución infinita son $\gamma_1^\infty = \gamma_2^\infty = 1{,}091$.

b) En una mezcla de agua y etanol para la cual $\gamma_1^\infty = 7{,}20$ *y* $\gamma_2^\infty = 2{,}74$.

Solución:

Por simplicidad denotaremos como $x_1 = \Lambda_{12}$ y como $x_2 = \Lambda_{21}$. Con ello las ecuaciones del modelo de Wilson se escriben como:

$$\begin{aligned} \ln(\gamma_1^\infty) &= 1 - \ln(x_1) - x_2 \\ \ln(\gamma_2^\infty) &= 1 - \ln(x_2) - x_1 \end{aligned}$$

Caso a): Una primera opción para intentar resolver estas ecuaciones mediante el método de aproximaciones sucesivas consiste en despejar x_1 de la segunda ecuación y x_2 de la primera de ellas con lo que:

$$\begin{aligned} x_1 &= 1 - \ln(\gamma_2^\infty) - \ln(x_2) \\ x_2 &= 1 - \ln(\gamma_1^\infty) - \ln(x_1) \end{aligned}$$

por lo que llamando

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} 1 - \ln(\gamma_2^\infty) - \ln(x_2) \\ 1 - \ln(\gamma_1^\infty) - \ln(x_1) \end{pmatrix}$$

puede intentarse el esquema iterativo:

$$\mathbf{x}^{(i+1)} = \mathbf{g}\left(\mathbf{x}^{(i)}\right) = \begin{pmatrix} g_1\left(x_1^{(i)}, x_2^{(i)}\right) \\ g_2\left(x_1^{(i)}, x_2^{(i)}\right) \end{pmatrix} = \begin{pmatrix} 1 - \ln(\gamma_2^\infty) - \ln\left(x_2^{(i)}\right) \\ 1 - \ln(\gamma_1^\infty) - \ln\left(x_1^{(i)}\right) \end{pmatrix}$$

Obsérvese que

$$[\mathbf{J}_\mathbf{g}(\mathbf{x})] = \begin{bmatrix} 0 & -\frac{1}{x_2} \\ -\frac{1}{x_1} & 0 \end{bmatrix}$$

por lo que sus valores propios serán:

$$\sqrt{\frac{1}{x_1 x_2}}$$

lo que nos indica que si $\left|\frac{1}{x_1 x_2}\right| \geq 1$ la convergencia del método no estará asegurada (puesto que el valor de toda norma matricial siempre es superior al radio espectral de la matriz). En efecto, si con estas funciones inicializamos

el método de aproximaciones sucesivas se van obteniendo, a partir de $x_1^{(0)} = x_2^{(0)} = 1$, los valores siguientes:

$$x_1^{(1)} = x_2^{(1)} = 0{,}912905..., \quad x_1^{(2)} = x_2^{(2)} = 1{,}00402..., \quad x_1^{(3)} = x_2^{(3)} = 0{,}90888...,$$

$$x_1^{(4)} = x_2^{(4)} = 1{,}00844..., \quad x_1^{(5)} = x_2^{(5)} = 0{,}90449..., \quad x_1^{(6)} = x_2^{(6)} = 1{,}013279...,$$

$$....., \quad x_1^{(59)} = x_2^{(59)} = 0{,}1758157..., \quad x_1^{(60)} = x_2^{(60)} = 2{,}65122...,$$

sucesión de vectores en la que sus componentes forman sucesiones oscilantes que divergen. Por tanto esta elección de la función $\mathbf{g}(\mathbf{x})$ no es adecuada para la resolución del sistema de ecuaciones. Puede observarse que para los valores que se van obteniendo, el radio espectral de la matriz jacobiana va tomando valores mayores que 1 (para los vectores en que sus componentes son menores que 1) y menores que 1 (para los vectores en que sus componentes son mayores que 1). Por tanto ninguno de los teoremas anteriores nos garantiza la convergencia.

En su lugar puede procederse de la forma siguiente:

$$\left(\begin{array}{l} \ln(\gamma_1^\infty) = 1 - \ln(x_1) - x_2 \\ \ln(\gamma_2^\infty) = 1 - \ln(x_2) - x_1 \end{array} \right) \Longleftrightarrow$$

$$\Longleftrightarrow \left(\begin{array}{l} \alpha x_1 = 1 + \alpha x_1 - \ln(\gamma_1^\infty) - \ln(x_1) - x_2 \\ \alpha x_2 = 1 + \alpha x_2 - \ln(\gamma_2^\infty) - \ln(x_2) - x_1 \end{array} \right) \Longleftrightarrow$$

$$\Longleftrightarrow \left(\begin{array}{l} x_1 = g_1(x_1, x_2) = \frac{1}{\alpha}(1 + \alpha x_1 - \ln(\gamma_1^\infty) - \ln(x_1) - x_2) \\ x_2 = g_2(x_1, x_2) = \frac{1}{\alpha}(1 + \alpha x_2 - \ln(\gamma_2^\infty) - \ln(x_2) - x_1) \end{array} \right)$$

Con esta nueva forma de proceder la matriz Jacobiana será:

$$[\mathbf{J_g}(\mathbf{x})] = \frac{1}{\alpha} \left[\begin{array}{cc} \left(\alpha - \frac{1}{x_1}\right) & -1 \\ -1 & \left(\alpha - \frac{1}{x_2}\right) \end{array} \right]$$

La norma-1 de la matriz Jacobiana será por tanto,

$$\|[\mathbf{J_g}(\mathbf{x})]\|_1 = \frac{1}{\alpha} Máx \left\{ \left|\alpha - \frac{1}{x_1}\right| + 1, \left|\alpha - \frac{1}{x_2}\right| + 1 \right\}$$

Restringiéndonos a valores de α positivos, para que $\frac{1}{\alpha}(\left|\alpha - \frac{1}{x}\right| + 1)$ sea inferior a 1 se debe verificar que

$$\left|\alpha - \frac{1}{x}\right| + 1 < \alpha \Longrightarrow x < 1$$

lo que asegura que $\|[\mathbf{J_g}(\mathbf{x})]\|_1$ es menor que 1 siempre que

$$x_1 < 1 \quad \text{y} \quad x_2 < 1$$

Con ello puede darse a α un valor que nos proporcione dominios suficientemente amplios para obtener una buena convergencia. Por ejemplo, si a α le asignamos el valor 2, admitiendo en el caso a) la igualdad de valores para las soluciones x_1^* y para x_2^* puesto que responden a la misma ecuación, se tiene que si $x \in [0{,}639, 1]$ las imágenes de $g_i(x, x)$ (i=1, 2) pertenecen al intervalo $[0{,}95, 1] \subset [0{,}639, 1]$. Por tanto si ambas variables se toman con el mismo valor en el intervalo $[0{,}639, 1-\varepsilon]$ (siendo ε tan pequeño como se desee) la convergencia queda asegurada. Así partiendo ahora de $x_1^{(0)} = x_2^{(0)} = 1$ se tiene que:

$$x_1^{(1)} = x_2^{(1)} = 0{,}9564526645...., \qquad x_1^{(2)} = x_2^{(2)} = 0{,}9569409865.....,$$
$$x_1^{(3)} = x_2^{(3)} = 0{,}956929935..., \qquad x_1^{(4)} = x_2^{(4)} = 0{,}9569301837...,$$
$$x_1^{(5)} = x_2^{(5)} = 0{,}9569301781..., x_1^{(6)} = x_2^{(6)} = 0{,}9569301782...$$

por lo que $x_1^* = x_2^* \approx 0{,}956930178$.

Caso b): En este caso el sistema a resolver será:

$$\begin{aligned} \ln(7{,}20) &= 1 - \ln(x_1) - x_2 \\ \ln(2{,}74) &= 1 - \ln(x_2) - x_1 \end{aligned}$$

de donde

$$\left(\begin{array}{l} x_1 = g_1(x_1, x_2) = \frac{1}{\alpha}(1 + \alpha x_1 - \ln(7{,}20) - \ln(x_1) - x_2) \\ x_2 = g_2(x_1, x_2) = \frac{1}{\alpha}(1 + \alpha x_2 - \ln(2{,}74) - \ln(x_2) - x_1) \end{array} \right)$$

Tomando para α el valor 4 (dejamos al lector la justificación de qué valores serían admisibles para este parámetro) el esquema iterativo a seguir se reduce a:

$$x_1^{(i+1)} = \frac{1}{4}(1 + 4x_1^{(i)} - \ln(7{,}20) - \ln(x_1^{(i)}) - x_2^{(i)})$$
$$x_2^{(i+1)} = \frac{1}{4}(1 + 4x_2 - \ln(2{,}74) - \ln(x_2^{(i)}) - x_1^{(i)})$$

por lo que comenzando con los valores $x_1^{(0)} = x_2^{(0)} = 0{,}956$ (aproximación de los hallados en el caso a)) se tienen los vectores:

$$\mathbf{x}^{(1)} = \begin{pmatrix} 0{,}484729 \\ 0{,}726260 \end{pmatrix} \rightarrow \mathbf{x}^{(2)} = \begin{pmatrix} 0{,}240685 \\ 0{,}683050 \end{pmatrix} \rightarrow \mathbf{x}^{(3)} = \begin{pmatrix} 0{,}182469 \\ 0{,}716186 \end{pmatrix} \rightarrow$$

$$\rightarrow \mathbf{x}^{(4)} = \begin{pmatrix} 0{,}185196 \\ 0{,}752033 \end{pmatrix} \rightarrow \mathbf{x}^{(5)} = \begin{pmatrix} 0{,}175253 \\ 0{,}774988 \end{pmatrix} \rightarrow ..... \rightarrow$$

$$\rightarrow \mathbf{x}^{(45)} = \begin{pmatrix} 0{,}162447 \\ 0{,}843320 \end{pmatrix} \rightarrow \mathbf{x}^{(46)} = \begin{pmatrix} 0{,}162447 \\ 0{,}843321 \end{pmatrix}$$

momento en el que detenemos el proceso iterativo al verificarse que la norma del vector diferencia entre las dos últimas aproximaciones es inferior a 10^{-6}.

Como puede apreciarse, con la función vectorial:

$$\mathbf{g}(x_1, x_2) = \begin{pmatrix} \frac{1}{4}(1 + 4x_1 - \ln(7{,}20) - \ln(x_1) - x_2) \\ \frac{1}{4}(1 + 4x_2 - \ln(2{,}74) - \ln(x_2) - x_1) \end{pmatrix}$$

el método de aproximaciones sucesivas nos ha conducido a la solución buscada pero en un número relativamente elevado de iteraciones. Conocida la solución, podemos comprobar (a posteriori) que en ella la norma-1 de la matriz Jacobiana toma un valor de 0,9535 es decir relativamente próximo a 1. Ello nos puede hacer pensar que quizás otras formas de determinar la función $\mathbf{g}(\mathbf{x})$ podrían ser más ventajosas. En efecto, las ecuaciones dadas pueden manipularse de la forma siguiente:

$$\begin{pmatrix} \ln(\gamma_1^{\infty}) = 1 - \ln(x_1) - x_2 \\ \ln(\gamma_2^{\infty}) = 1 - \ln(x_2) - x_1 \end{pmatrix} \Leftrightarrow \begin{pmatrix} 0 = \ln(e) - (\ln(\gamma_1^{\infty}) + \ln(x_1)) - x_2 \\ 0 = \ln(e) - (\ln(\gamma_2^{\infty}) + \ln(x_2)) - x_1 \end{pmatrix} \Leftrightarrow$$

$$\Leftrightarrow \begin{pmatrix} 0 = \ln\left(\frac{e}{\gamma_1^{\infty} x_1}\right) - x_2 \\ 0 = \ln\left(\frac{e}{\gamma_2^{\infty} x_2}\right) - x_1 \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 = \frac{e}{\gamma_1^{\infty} x_1}.e^{-x_2} \\ 1 = \frac{e}{\gamma_2^{\infty} x_2}.e^{-x_1} \end{pmatrix} \Leftrightarrow$$

$$\Leftrightarrow \begin{pmatrix} x_1 = g_1(x_1, x_2) = \frac{1}{\gamma_1^{\infty} e^{(x_2 - 1)}} \\ x_2 = g_2(x_1, x_2) = \frac{1}{\gamma_2^{\infty} e^{(x_1 - 1)}} \end{pmatrix}$$

Con ello la matriz Jacobiana ahora será:

$$[\mathbf{J}_{\mathbf{g}}(\mathbf{x})] = \begin{bmatrix} 0 & \frac{-1}{\gamma_2^{\infty} e^{(x_1 - 1)}} \\ \frac{-1}{\gamma_1^{\infty} e^{(x_2 - 1)}} & 0 \end{bmatrix}$$

por lo que en el punto fijo de la aplicación $\mathbf{g}(\mathbf{x})$ tomará la expresión:

$$[\mathbf{J}_{\mathbf{g}}(\mathbf{x}^*)] = \begin{bmatrix} 0 & \frac{-1}{\gamma_2^{\infty} e^{(x_1^* - 1)}} \\ \frac{-1}{\gamma_1^{\infty} e^{(x_2^* - 1)}} & 0 \end{bmatrix}$$

que sustiyuyendo los valores correspondientes a este caso se convierte en:

$$[\mathbf{J_g}(\mathbf{x}^*)] = \begin{bmatrix} 0 & -0{,}427 \\ -0{,}321 & 0 \end{bmatrix}$$

por lo que la norma-1 de la matriz jacobiana en la solución es 0,427 valor inferior al que obtuvimos antes y que demuestra que se tendrá una mayor velocidad de convergencia (al menos en un entorno de la raíz). En efecto, aplicando el esquema iterativo:

$$\left(\begin{array}{l} x_1^{(i+1)} = \dfrac{1}{\gamma_1^{\infty} e^{(x_2^{(i)}-1)}} \\ x_2^{(i+1)} = \dfrac{1}{\gamma_2^{\infty} e^{(x_1^{(i)}-1)}} \end{array} \right)$$

a partir de los valores: $x_1^{(0)} = x_2^{(0)} = 0{,}956$ (aproximación de los hallados en el caso a)) se tienen los vectores:

$$\mathbf{x}^{(1)} = \begin{pmatrix} 0{,}145136 \\ 0{,}381380 \end{pmatrix} \rightarrow \mathbf{x}^{(2)} = \begin{pmatrix} 0{,}257828 \\ 0{,}858049 \end{pmatrix} \rightarrow \mathbf{x}^{(3)} = \begin{Bmatrix} 0{,}160072 \\ 0{,}766603 \end{Bmatrix} \rightarrow$$

$$\rightarrow \mathbf{x}^{(4)} = \begin{pmatrix} 0{,}175400 \\ 0{,}845328 \end{pmatrix} \rightarrow \mathbf{x}^{(5)} = \begin{pmatrix} 0{,}162121 \\ 0{,}832470 \end{pmatrix} \rightarrow ..... \rightarrow$$

$$\rightarrow \mathbf{x}^{(15)} = \begin{pmatrix} 0{,}162447 \\ 0{,}843323 \end{pmatrix} \rightarrow \mathbf{x}^{(16)} = \begin{pmatrix} 0{,}162447 \\ 0{,}843323 \end{pmatrix}$$

deteniéndose el proceso iterativo en la 16^a iteración al ser el valor de la norma-1 del vector diferencia entre las aproximaciones de las dos últimas iteraciones inferior a 10^{-6} y ser el valor de

$$\begin{pmatrix} 1 - \ln(7{,}2) - \ln(0{,}162447) - 0{,}843323 \\ 1 - \ln(2{,}74) - \ln(0{,}843323) - 0{,}162447 \end{pmatrix} = \begin{pmatrix} -5.42 \times 10^{-7} \\ 3.19 \times 10^{-7} \end{pmatrix}$$

por tanto, también inferior (en norma-1) a 10^{-6}.

Aún podría mejorarse la velocidad de convergencia en este caso modificando el esquema iterativo según la variante del método que se presenta a continuación.

Una variante del método de aproximaciones sucesivas

En el método de aproximaciones sucesivas que se ha descrito anteriormente se sigue el esquema iterativo:

$$\left.\begin{array}{l} x_1^{(i+1)} = g_1(x_1^{(i)}, x_2^{(i)}, ..., x_{j-1}^{(i)}, x_j^{(i)}, x_{j+1}^{(i)}, ..., x_n^{(i)}) \\ x_2^{(i+1)} = g_2(x_1^{(i)}, x_2^{(i)}, ..., x_{j-1}^{(i)}, x_j^{(i)}, x_{j+1}^{(i)}, ..., x_n^{(i)}) \\ \qquad\qquad ..... \\ x_j^{(i+1)} = g_j(x_1^{(i)}, x_2^{(i)}, ..., x_{j-1}^{(i)}, x_j^{(i)}, x_{j+1}^{(i)}, ..., x_n^{(i)}) \\ \qquad\qquad ..... \\ x_n^{(i+1)} = g_n(x_1^{(i)}, x_2^{(i)}, ..., x_{j-1}^{(i)}, x_j^{(i)}, x_{j+1}^{(i)}, ..., x_n^{(i)}) \end{array}\right\} \qquad (i = 0, 1, ...)$$

es decir, que para la estimación del valor de la variable $x_j^{(i+1)}$ se utilizan los valores obtenidos en la iteración anterior para todas las variables que se están aproximando. No obstante, en ese momento ya se dispone de los valores de $x_1^{(i+1)}, x_2^{(i+1)}, ..., x_{j-1}^{(i+1)}$ que, si el método posee buenas propiedades de convergencia puede pensarse que estén más próximos a los de la solución que los de la iteración anterior. En este sentido, en ocasiones, se sustituye el esquema iterativo antes considerado por el siguiente:

$$\left.\begin{array}{c} x_1^{(i+1)} = g_1(x_1^{(i)}, x_2^{(i)}, ..., x_{j-1}^{(i)}, x_j^{(i)}, x_{j+1}^{(i)}, ..., x_n^{(i)}) \\ x_2^{(i+1)} = g_2(x_1^{(i+1)}, x_2^{(i)}, ..., x_{j-1}^{(i)}, x_j^{(i)}, x_{j+1}^{(i)}, ..., x_n^{(i)}) \\ \qquad\qquad ..... \\ x_j^{(i+1)} = g_j(x_1^{(i+1)}, x_2^{(i+1)}, ..., x_{j-1}^{(i+1)}, x_j^{(i)}, x_{j+1}^{(i)}, ..., x_n^{(i)}) \\ \qquad\qquad ..... \\ x_n^{(i+1)} = g_n(x_1^{(i+1)}, x_2^{(i+1)}, ..., x_{j-1}^{(i+1)}, x_j^{(i+1)}, x_{j+1}^{(i+1)}, ..., x_n^{(i)}) \end{array}\right\} \qquad (i = 0, 1, ...)$$

Por la analogía entre esta forma de proceder y la que se sigue en el método de Gauss-Seidel para la resolución de sistemas algebraicos de ecuaciones lineales este método se conoce, en algunos textos, con el nombre de método de Gauss-Seidel, aunque más adelante, como variante del método de Newton, presentaremos otro método también designado con este nombre.

Hemos de advertir no obstante que, pese a lo "razonable" que parece el razonamiento seguido para derivarlo, no siempre funciona mejor esta variante del método que el método clásico de aproximaciones sucesivas pues con él se puede estar modificando la función $\mathbf{g}(\mathbf{x})$ de la iteración empeorando su constante de Lipschitz.

Un ejemplo, en el que se acelera la convergencia actuando con la variante que acabamos de presentar, es el del problema propuesto por el Pr. B. Coto que conduce al sistema utilizado en el ejemplo antes resuelto. En efecto en ese caso el método de aproximaciones sucesivas nos conducía al esquema:

$$\left(\begin{array}{l} x_1^{(i+1)} = \dfrac{1}{\gamma_1^{\infty} e^{(x_2^{(i)}-1)}} \\ x_2^{(i+1)} = \dfrac{1}{\gamma_2^{\infty} e^{(x_1^{(i)}-1)}} \end{array} \right) \Leftrightarrow \left(\begin{array}{l} x_1^{(i+1)} = \dfrac{1}{\gamma_1^{\infty} e^{\left(\frac{1}{\gamma_2^{\infty} e^{(x_1^{(i-1)}-1)}} - 1 \right)}} \\ x_2^{(i+1)} = \dfrac{1}{\gamma_2^{\infty} e^{\left(\frac{1}{\gamma_1^{\infty} e^{(x_2^{(i-1)}-1)}} - 1 \right)}} \end{array} \right)$$

en el que se puede apreciar que el valor de cada incógnita se va aproximando con el que se obtuvo para la aproximación de dicha incógnita dos iteraciones antes. Parece que de esta forma se están generando, para cada variable, dos sucesiones independientes que convergen ambas en el mismo punto. Por ello puede procederse según el esquema:

$$\left(\begin{array}{l} x_1^{(i+1)} = \dfrac{1}{\gamma_1^{\infty} e^{(x_2^{(i)}-1)}} \\ x_2^{(i+1)} = \dfrac{1}{\gamma_2^{\infty} e^{(x_1^{(i+1)}-1)}} \end{array} \right) \Leftrightarrow \left(\begin{array}{l} x_1^{(i+1)} = \dfrac{1}{\gamma_1^{\infty} e^{\left(\frac{1}{\gamma_2^{\infty} e^{(x_1^{(i)}-1)}} - 1 \right)}} \\ x_2^{(i+1)} = \dfrac{1}{\gamma_2^{\infty} e^{\left(\frac{1}{\gamma_1^{\infty} e^{(x_2^{(i)}-1)}} - 1 \right)}} \end{array} \right)$$

Aplicado este esquema iterativo a nuestro problema en cuestión, inicializado con los valores $x_1^{(0)} = x_2^{(0)} = 0{,}956$ se tienen los vectores:

$$x^{(1)} = \begin{pmatrix} 0{,}145136 \\ 0{,}858049 \end{pmatrix} \rightarrow x^{(2)} = \begin{pmatrix} 0{,}160072 \\ 0{,}845328 \end{pmatrix} \rightarrow x^{(3)} = \begin{pmatrix} 0{,}162121 \\ 0{,}843598 \end{pmatrix} \rightarrow$$

$$\rightarrow x^{(4)} = \begin{pmatrix} 0{,}162402 \\ 0{,}843361 \end{pmatrix} \rightarrow x^{(5)} = \begin{pmatrix} 0{,}162441 \\ 0{,}843325 \end{pmatrix} \rightarrow x^{(6)} = \begin{pmatrix} 0{,}162446 \\ 0{,}843324 \end{pmatrix} \rightarrow$$

$$\rightarrow x^{(7)} = \begin{pmatrix} 0{,}162447 \\ 0{,}843323 \end{pmatrix} \rightarrow x^{(8)} = \begin{pmatrix} 0{,}162447 \\ 0{,}843323 \end{pmatrix}$$

habiéndose obtenido la misma solución en la mitad de iteraciones que antes.

2.4.2. El método de Newton-Raphson para sistemas de n ecuaciones no lineales.

Considérese nuevamente el sistema de n ecuaciones no lineales con n incógnitas representado por

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \Longleftrightarrow \left(\begin{array}{c} f_1(x_1, x_2, ..., x_n) = 0 \\ f_2(x_1, x_2, ..., x_n) = 0 \\ ..... \\ f_n(x_1, x_2, ..., x_n) = 0 \end{array} \right)$$

Al igual que se hizo en el caso de una variable, supongamos que en un dominio cerrado $D \subset I\!R^n$ $\mathbf{f}(\mathbf{x})$ es una función de clase $(C^2(D))^n$. Y supongamos además que la ecuación anterior admite una solución $\mathbf{x}^*$ en el dominio D. Para cualquier otro vector $\mathbf{x}^{(0)} \in D$, denotando por $\boldsymbol{\delta}\mathbf{x}$ al vector tal que $\mathbf{x}^* = \mathbf{x}^{(0)} + \boldsymbol{\delta}\mathbf{x}$, la expresión del desarrollo en serie de Taylor nos permitiría afirmar, para cada una de las ecuaciones del sistema, que existen los valores $\theta_j \in [0, 1]$ $(j = 1, 2, .., n)$ tales que:

$$\begin{aligned} 0 &= f_j(\mathbf{x}^*) = f(\mathbf{x}^{(0)} + \boldsymbol{\delta}\mathbf{x}) = \\ &= f_j(\mathbf{x}^{(0)}) + \left\{ \nabla f_j(\mathbf{x}^{(0)}) \right\}^T \boldsymbol{\delta}\mathbf{x} + \frac{1}{2} \{\boldsymbol{\delta}\mathbf{x}\}^T \left[\mathbf{H}_{f_j}(\mathbf{x}^{(0)} + \theta_j \boldsymbol{\delta}\mathbf{x}) \right] \boldsymbol{\delta}\mathbf{x} \end{aligned}$$

donde $\left[H_{f_j}(\mathbf{x}) \right]$ es la matriz hessiana de la función $f_j(\mathbf{x})$.

Si conocido $\mathbf{x}^{(0)}$ se fuese capaz de determinar $\boldsymbol{\delta}\mathbf{x}$ resolviendo el sistema formado para $j = 1, 2, .., n$ por las ecuaciones:

$$f_j(\mathbf{x}^{(0)}) + \left\{ \nabla f_j(\mathbf{x}^{(0)}) \right\}^T \boldsymbol{\delta}\mathbf{x} + \frac{1}{2} \{\boldsymbol{\delta}\mathbf{x}\}^T \left[\mathbf{H}_{f_j}(\mathbf{x}^{(0)} + \theta_j \boldsymbol{\delta}\mathbf{x}) \right] \boldsymbol{\delta}\mathbf{x} = 0$$

podría determinarse $\mathbf{x}^*$ como $\mathbf{x}^* = \mathbf{x}^{(0)} + \boldsymbol{\delta}\mathbf{x}$. Pero para resolver este sistema primero deberíamos conocer los valores de θ_j (lo cual no es obvio) y, una vez conocidos, resolver un sistema, en general, no lineal pues obsérvese que $\boldsymbol{\delta}\mathbf{x}$ interviene en la expresión de las matrices hessianas $\left[H_{f_j}(\mathbf{x}^{(0)} + \theta_j \boldsymbol{\delta}\mathbf{x}) \right]$. Por tanto, salvo en situaciones muy particulares, no se ganaría gran cosa reemplazando el problema de resolver $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ por el de resolver el sistema anterior.

El método de Newton-Raphson (o método de linealización de Newton) se sustenta en simplificar las expresiones anteriores linealizándolas. Para ello considera que si se está suficientemente cerca de la solución (es decir, si $||\boldsymbol{\delta}\mathbf{x}||$ es suficientemente pequeño) los términos

$$\left(\frac{1}{2} \{\boldsymbol{\delta}\mathbf{x}\}^T \left[\mathbf{H}_{f_j}(\mathbf{x}^{(0)} + \theta_j \boldsymbol{\delta}\mathbf{x}) \right] \boldsymbol{\delta}\mathbf{x} \right)$$

podrán despreciarse frente a los otros términos de cada ecuación del sistema. Por ello en este método se resuelve el sistema **lineal**:

$$\mathbf{f}(\mathbf{x}^{(0)}) + \left[\mathbf{J_f}(\mathbf{x}^{(0)})\right] \boldsymbol{\Delta}\mathbf{x}^{(0)} = 0$$

del que se obtiene que

$$\boldsymbol{\Delta}\mathbf{x}^{(0)} = -\left[\mathbf{J_f}(\mathbf{x}^{(0)})\right]^{-1} \mathbf{f}(\mathbf{x}^{(0)})$$

Obviamente, al ser diferente el sistema linealizado que el proporcionado por el desarrollo de Taylor, se tendrá que $\boldsymbol{\Delta}\mathbf{x}^{(0)} \neq \boldsymbol{\delta}\mathbf{x}$ y por tanto

$$\mathbf{x}^* = \mathbf{x}^{(0)} + \boldsymbol{\delta}\mathbf{x} \neq \mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \boldsymbol{\Delta}\mathbf{x}^{(0)}$$

De una forma intuitiva (que después deberemos precisar cuándo es correcta) puede pensarse que aunque $\mathbf{x}^{(1)}$ sea diferente de $\mathbf{x}^*$ será un vector más próximo a $\mathbf{x}^*$ que $\mathbf{x}^{(0)}$ pues lo hemos obtenido "aproximando" el valor $\boldsymbol{\delta}\mathbf{x}$ que nos llevaba de $\mathbf{x}^{(0)}$ a $\mathbf{x}^*$. Con ello el método de Newton-Raphson propone repetir este proceso de forma recursiva hasta estar lo suficientemente cercanos a la solución buscada. Más concretamente el método de Newton-Raphson consiste en:

Dado un vector $\mathbf{x}^{(0)}$, generar la sucesión:

$$\left\{\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]^{-1} \mathbf{f}(\mathbf{x}^{(i)})\right\}_{i=0}^{\infty}.$$

Sobre este método, en primer lugar, puede observarse que si denotamos por:

$$\mathbf{g}(\mathbf{x}) = \mathbf{x} - [\mathbf{J_f}(\mathbf{x})]^{-1} \mathbf{f}(\mathbf{x})$$

estamos en presencia de un caso particular del método de aproximaciones sucesivas antes contemplado en el apartado 1.4.1.. En otros términos, se tiene la siguiente propiedad:

Proposición 2.4.1 *Si la función* $\mathbf{g}(\mathbf{x}) = \mathbf{x} - [\mathbf{J_f}(\mathbf{x})]^{-1} \mathbf{f}(\mathbf{x})$ *es, para alguna norma matricial, una contracción definida en* D *la sucesión dada por*

$$\left\{\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]^{-1} \mathbf{f}(\mathbf{x}^{(i)})\right\}_{i=0}^{\infty}$$

obtenida a partir de cualquier vector $\mathbf{x}^{(0)} \in D$ *converge hacia la única solución de la ecuación* $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ *en* D.

Demostración: Es un caso particular de los teoremas de convergencia del método de aproximaciones sucesivas.

c.q.d.

Del teorema anterior, por analogía a lo realizado en el caso de una única ecuación no lineal, podrían derivarse teoremas de convergencia que actuaran sobre las primeras y segundas derivadas parciales de las componentes de la aplicación vectorial $\mathbf{f}(\mathbf{x})$. Dejamos al lector el desarrollo de tales teoremas y pasamos a enunciar algunos otros en los que las hipótesis se realizan directamente sobre la propia función vectorial $\mathbf{f}(\mathbf{x})$ y su matriz jacobiana y que pueden ser de más fácil aplicación al análisis de la convergencia del método. Previamente a la demostración de dichos teoremas necesitaremos introducir el siguiente lema:

Lema 2.4.1 *Siendo* $\mathbf{f} : D \to D$ *una aplicación de clase* $(C^1(D))^n$ *y siendo* D *un cerrado de* $\mathbb{R}^n$, *si existe una constante estrictamente positiva* α, *tal que para alguna norma vectorial y para la norma matricial a ella subordinada se verifique:*

$$\|[\mathbf{J_f}(\mathbf{x})] - [\mathbf{J_f}(\mathbf{y})]\| \leq \alpha \|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y} \in D$$

entonces se verifica también que:

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y}) - \mathbf{J_f}(\mathbf{y})(\mathbf{x} - \mathbf{y})\| \leq \frac{\alpha}{2} \|\mathbf{x} - \mathbf{y}\|^2 \qquad \forall \mathbf{x}, \mathbf{y} \in D$$

Demostración:

Siendo $\mathbf{x}$ e $\mathbf{y}$ dos vectores genéricos de D denotemos por $\mathbf{q}(t)$ a la función vectorial dependiente de un único parámetro real definida por:

$$\mathbf{q}(t) = \mathbf{f}(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))$$

Esta función, habida cuenta de las hipótesis realizadas sobre $\mathbf{f}$ es derivable $\forall t \in [0, 1]$. Así, denotando por $\mathbf{z} = \mathbf{y} + t(\mathbf{x} - \mathbf{y})$ se tiene que:

$$\mathbf{q}'(t) = \frac{d\mathbf{q}}{dt}(t) = \lim_{\Delta t \to 0} \frac{\mathbf{f}(\mathbf{y} + (t + \Delta t)(\mathbf{x} - \mathbf{y})) - \mathbf{f}(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))}{\Delta t} =$$

$$= \lim_{\Delta t \to 0} \frac{\mathbf{f}(\mathbf{z} + \Delta t(\mathbf{x} - \mathbf{y})) - \mathbf{f}(\mathbf{z})}{\Delta t} = [\mathbf{J_f}(\mathbf{z})]\,(\mathbf{x} - \mathbf{y})$$

de donde

$$\left\|\mathbf{q}'(t) - \mathbf{q}'(0)\right\| = \|[[\mathbf{J_f}(\mathbf{z})] - [\mathbf{J_f}(\mathbf{y})]]\,(\mathbf{x} - \mathbf{y})\| \leq$$

$$\leq \|[[\mathbf{J_f}(\mathbf{z})] - [\mathbf{J_f}(\mathbf{y})]]\| \,\|\mathbf{x} - \mathbf{y}\| =$$

$$= \|[[\mathbf{J_f}(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))] - [\mathbf{J_f}(\mathbf{y})]]\| \,\|\mathbf{x} - \mathbf{y}\| \leq$$

$$\leq \alpha t \|\mathbf{x} - \mathbf{y}\|^2$$

Esta desigualdad, a su vez, puede utilizarse en el proceso siguiente:

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y}) - \mathbf{J_f}(\mathbf{y})(\mathbf{x}-\mathbf{y})\| = \left\|\mathbf{q}(1) - \mathbf{q}(0) - \mathbf{q}'(0)\right\| =$$

$$= \left\|\int_0^1 (\mathbf{q}'(t) - \mathbf{q}'(0))dt\right\| \leq \int_0^1 \left\|\mathbf{q}'(t) - \mathbf{q}'(0)\right\| dt \leq$$

$$\leq \int_0^1 \alpha t \|\mathbf{x}-\mathbf{y}\|^2 .t = \frac{\alpha}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

c.q.d.

Observación 2.4.5 *El que se verifique la hipótesis del lema precedente:*

$$\exists \alpha \in I\!R_+ \quad / \quad \|[\mathbf{J_f}(\mathbf{x})] - [\mathbf{J_f}(\mathbf{y})]\| \leq \alpha \|\mathbf{x}-\mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y} \in D$$

se expresa diciendo que la matriz Jacobiana es lipschitciana de razón α en D para la norma $\|\|$.

Con ayuda de este lema puede procederse a presentar y demostrar el siguiente teorema:

Teorema 2.4.6 *Siendo D un cerrado de $I\!R^n$ y siendo $\mathbf{f} : D \to D$ una aplicación de clase $(C^1(D))^n$ para la que, utilizando alguna norma vectorial y para la norma matricial a ella subordinada, se verifican las dos hipótesis siguientes:*

a) $\exists \alpha \in I\!R_+ \quad / \quad \|[\mathbf{J_f}(\mathbf{x})] - [\mathbf{J_f}(\mathbf{y})]\| \leq \alpha \|\mathbf{x}-\mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in D$

b) $\exists \beta \in I\!R_+ \quad / \quad \left\|[\mathbf{J_f}(\mathbf{x})]^{-1}\right\| < \beta \qquad \forall \mathbf{x} \in D$

entonces para la sucesión $\left\{\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]^{-1} \mathbf{f}(\mathbf{x}^{(i)})\right\}_{i=0}^{\infty}$ *obtenida a partir de cualquier vector* $\mathbf{x}^{(0)} \in D$ *se verifica que:*

$$\left\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right\| \leq \frac{\alpha\beta}{2}\left\|\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}\right\|^2$$

Demostración: Se tiene que:

$$\left\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right\| = \left\|-\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]^{-1}\mathbf{f}(\mathbf{x}^{(i)})\right\| \leq \left\|\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]^{-1}\right\| \left\|\mathbf{f}(\mathbf{x}^{(i)})\right\| \leq$$

$$\leq \beta \left\|\mathbf{f}(\mathbf{x}^{(i)})\right\|$$

y como de:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]^{-1}\mathbf{f}(\mathbf{x}^{(i)})$$

se deduce que:

$$\mathbf{f}(\mathbf{x}^{(i)}) = -\left[\mathbf{J}_{\mathbf{f}}(\mathbf{x}^{(i)})\right]\left(\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right)$$

se tiene, utilizando el lema precedente, que:

$$\left\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right\| \leq \beta \left\|\mathbf{f}(\mathbf{x}^{(i)})\right\| =$$

$$= \beta \left\|\mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i)}) - \left[\mathbf{J}_{\mathbf{f}}(\mathbf{x}^{(i)})\right]\left(\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right)\right\| \leq$$

$$\leq \frac{\alpha\beta}{2}\left\|\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}\right\|^2$$

c.q.d.

El teorema anterior nos muestra que la relación entre la norma del vector diferencia entre las aproximaciones halladas en las iteraciones $(i+1)$ e i es proporcional (con factor $C = \alpha\beta/2$) al cuadrado de la norma del vector diferencia entre las aproximaciones halladas en las iteraciones i e $(i-1)$. Pero por sí solo este teorema no nos justifica que el método converja. Simplemente nos indica que si en algún momento $\left\|\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}\right\|^2 < (1/C)$ entonces se habrá logrado una sucesión de Cauchy y, al estar en un completo, por ello una sucesión convergente. Para acabar de obtener un resultado que garantice la convergencia es necesario imponer más condiciones en el método. Como por ejemplo las que se recogen en el teorema siguiente que, junto a las hipótesis a) y b) del teorema anterior añade una nueva:

Teorema 2.4.7 *Siendo D un cerrado de $\mathbb{R}^n$ y siendo $\mathbf{f} : D \to D$ una aplicación de clase $(C^1(D))^n$ para la que, utilizando alguna norma vectorial y para la norma matricial a ella subordinada, se verifican las dos hipótesis siguientes:*

a) $\exists \alpha \in \mathbb{R}_+ \ / \ \|[\mathbf{J}_{\mathbf{f}}(\mathbf{x})] - [\mathbf{J}_{\mathbf{f}}(\mathbf{y})]\| \leq \alpha \|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in D$

b) $\exists \beta \in \mathbb{R}_+ \ / \ \left\|[\mathbf{J}_{\mathbf{f}}(\mathbf{x})]^{-1}\right\| < \beta \qquad \forall \mathbf{x} \in D$

entonces para la sucesión $\left\{\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{J}_{\mathbf{f}}(\mathbf{x}^{(i)})\right]^{-1}\mathbf{f}(\mathbf{x}^{(i)})\right\}_{i=0}^{\infty}$ *obtenida a partir de cualquier vector* $\mathbf{x}^{(0)} \in D$ *para el que se verifique la condición*

c) $\exists \mu < \frac{2}{\alpha\beta} \in \mathbb{R}_+ \ / \qquad \left\|\left[\mathbf{J}_{\mathbf{f}}(\mathbf{x}^{(0)})\right]^{-1}\mathbf{f}(\mathbf{x}^{(0)})\right\| \leq \mu$

existe el límite $\mathbf{x}^*$ *de la sucesión* $\left\{\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{J}_{\mathbf{f}}(\mathbf{x}^{(i)})\right]^{-1}\mathbf{f}(\mathbf{x}^{(i)})\right\}_{i=0}^{\infty}$ *que es una raíz del sistema* $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ *en D y se verifica que:*

$$\left\|\mathbf{x}^{(i+1)} - \mathbf{x}^*\right\| \leq \frac{r^{2^i - 1}}{1 - r^{2^i}}\mu$$

donde $r = \alpha\beta\mu/2 < 1$.

Demostración: Por aplicación directa del teorema anterior:

$$\left\|\mathbf{x}^{(i+1)}-\mathbf{x}^{(i)}\right\| \leq \frac{\alpha\beta}{2}\left\|\mathbf{x}^{(i)}-\mathbf{x}^{(i-1)}\right\|^2$$

Por recursión, llamando $C=\frac{\alpha\beta}{2}$ se tiene:

$$\left\|\mathbf{x}^{(i+1)}-\mathbf{x}^{(i)}\right\| \leq C\left\|\mathbf{x}^{(i)}-\mathbf{x}^{(i-1)}\right\|^2 \leq C^3\left\|\mathbf{x}^{(i-1)}-\mathbf{x}^{(i-2)}\right\|^4 \leq$$

$$\leq C^7\left\|\mathbf{x}^{(i-2)}-\mathbf{x}^{(i-3)}\right\|^8 \leq ..... \leq C^{2^i-1}\left\|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right\|^{2^i}$$

y como

$$\mathbf{x}^{(1)}-\mathbf{x}^{(0)}=\left[\mathbf{J}_{\mathbf{f}}(\mathbf{x}^{(0)})\right]^{-1}\mathbf{f}(\mathbf{x}^{(0)})$$

utilizando la hipótesis c) resultará que

$$\left\|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right\|^{2^i} \leq \mu^{2^i}$$

de donde

$$\left\|\mathbf{x}^{(i+1)}-\mathbf{x}^{(i)}\right\| \leq C^{(2^i-1)}\mu^{2^i}=r^{(2^i-1)}\mu$$

donde se ha denotado por r a: $r=\frac{\alpha\beta\mu}{2}<1$. Ello demuestra que, bajo las hipótesis del teorema, en la sucesión formada mediante el método de Newton-Raphson la distancia entre dos vectores consecutivos puede hacerse tan pequeña como se desee.

Por otra parte es claro que se verificará que siendo j e i dos índices tales que $j>i$:

$$\left\|\mathbf{x}^{(j+1)}-\mathbf{x}^{(i)}\right\| \leq \left\|\mathbf{x}^{(j+1)}-\mathbf{x}^{(j)}\right\|+\left\|\mathbf{x}^{(j)}-\mathbf{x}^{(i)}\right\| \leq$$

$$\leq \left\|\mathbf{x}^{(j+1)}-\mathbf{x}^{(j)}\right\|+\left\|\mathbf{x}^{(j)}-\mathbf{x}^{(j-1)}\right\|+\left\|\mathbf{x}^{(j-1)}-\mathbf{x}^{(i)}\right\| \leq$$

$$\leq ... \leq \left\|\mathbf{x}^{(j+1)}-\mathbf{x}^{(j)}\right\|+\left\|\mathbf{x}^{(j)}-\mathbf{x}^{(j-1)}\right\|+....+\left\|\mathbf{x}^{(i+1)}-\mathbf{x}^{(i)}\right\| \leq$$

$$\left\|\mathbf{x}^{(j+1)}-\mathbf{x}^{(i)}\right\|\mu\left(r^{2^j-1}+r^{2^{(j-1)}-1}+.....+r^{2^{(i+1)}-1}+r^{2^i-1}\right)=$$

$$=\mu r^{-1}\left(r^{2^j}+r^{2^{(j-1)}}+.....+r^{2^{(i+1)}}+r^{2^i}\right)$$

y como para cualquier entero k se tiene que

$$r^{2^k} = r^{2^i 2^{(k-i)}} = \left(r^{2^i}\right)^{2^{(k-i)}}$$

resultará

$$\left\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(i)}\right\| \leq \mu r^{-1}\left(r^{2^j} + r^{2^{(j-1)}} + ... + r^{2^{(i+1)}} + r^{2^i}\right) =$$

$$= \mu r^{-1}\left(\left(r^{2^i}\right)^{2^{(j-i)}} + \left(r^{2^i}\right)^{2^{(j-i-1)}} + ... + \left(r^{2^i}\right)^2 + r^{2^i}\right) =$$

$$= \mu r^{(2^i-1)}\left(1 + r^{2^i} + \left(r^{2^i}\right)^3 + \left(r^{2^i}\right)^7 + ... + \left(r^{2^i}\right)^{2^{(j-i)}-1}\right).$$

Puesto que $r < 1$ se tiene que:

$$\left\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(i)}\right\| \leq \mu r^{(2^i-1)}\left(1 + r^{2^i} + \left(r^{2^i}\right)^2 + \left(r^{2^i}\right)^3 + ... + \left(r^{2^i}\right)^{(j-i)}\right) =$$

$$= \mu r^{(2^i-1)}\left(\frac{1}{1-r^{2^i}} - \frac{\left(r^{2^i}\right)^{(j-i+1)}}{1-r^{2^i}}\right)$$

La desigualdad anterior muestra que, siendo $j > i$, el valor de $\left\|\mathbf{x}^{(j)} - \mathbf{x}^{(i)}\right\|$ puede hacerse tan pequeño como se desee con tal de tomar el índice i suficientemente elevado. Puesto que la sucesión $\left\{\mathbf{x}^{(i)}\right\}_{j=0}^{\infty}$ es una sucesión de Cauchy y D es un cerrado, ésta convergerá hacia un vector $\mathbf{x}^*$. Tomando límites en la desigualdad anterior resulta además que:

$$\begin{aligned}\left\|\mathbf{x}^* - \mathbf{x}^{(i)}\right\| &= \lim_{j\to\infty}\left\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(i)}\right\| \leq \\ &\leq \lim_{j\to\infty}\left(\mu r^{(2^i-1)}\left(\frac{1}{1-r^{2^i}} - \frac{\left(r^{2^i}\right)^{(j-i)}}{1-r^{2^i}}\right)\right) = \frac{r^{(2^i-1)}}{1-r^{2^i}}\mu\end{aligned}$$

lo que acaba de demostrar el teorema.

c.q.d.

El teorema precedente demuestra que, bajo las hipótesis en él impuestas, el método de Newton-Raphson converge. El teorema anterior a éste, demuestra además que la convergencia del método es cuadrática.

El método de Newton, en cada iteración, exige evaluar la matriz $\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]^{-1}$. Ello, en el caso de que el número de ecuaciones sea elevado, requiere un gran

esfuerzo computacional. Por ello se han desarrollado diferentes variantes del método de Newton en las que, perdiendo algo de su velocidad de convergencia, se aproxima dicha matriz inversa de la jacobiana. Algunas de estas variantes las presentaremos un poco más adelante.

Asimismo, el método de Newton-Raphson suele programarse de forma algo diferente a como lo hemos expuesto hasta ahora. En efecto, en el método de Newton-Raphson en cada iteración se suele determinar el vector incremento a través de la resolución de un sistema de n ecuaciones lineales (consúltese la bibliografía para el estudio de métodos numéricos de resolución de tales tipos de sistemas) y tras ello se suma el vector de incrementos al vector con el que se inicializó la iteración. Más concretamente un algoritmo del método puede ser el siguiente:

Algoritmo del método de Newton-Raphson para sistemas

Dado el sistema de ecuaciones no lineales $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, los indicadores de precisión ε y δ, un valor máximo del número de iteraciones que se permiten realizar ($maxiter$) y un vector $\mathbf{x}$ con el que inicializar el proceso,

$tolx \leftarrow 2\varepsilon$
$tolf \leftarrow 2\delta$
$iteración \leftarrow 0$
Mientras ($(iteración < maxiter)$ **y** $((tolx > \varepsilon)$ **o** $(tolf > \delta)$ **), hacer:**
Evaluar la matriz $[\mathbf{J_f}(\mathbf{x})]$
Resolver el sistema de ecuaciones lineales: $[\mathbf{J_f}(\mathbf{x})]\,\boldsymbol{\delta}\mathbf{x} = \mathbf{f}(\mathbf{x})$
Si (el sistema no puede resolverse) **entonces:**
Escribir mensaje de error (jacobiana singular) y finalizar el proceso
si no:

$$\mathbf{x} \leftarrow \mathbf{x} - \boldsymbol{\delta}\mathbf{x}$$

$$tolx \leftarrow \|\boldsymbol{\delta}\mathbf{x}\|$$

$$tolf \leftarrow \|\mathbf{f}(\mathbf{x})\|$$

$$iteración \leftarrow iteración + 1$$

fin condición.
Fin bucle condicional.
Si $((tolx < \varepsilon)$ **y** $(tolf < \delta)$ **) entonces:**
tomar $\mathbf{x}$ como solución
si no:
Escribir un mensaje de error en el proceso de cálculo

fin condición.
Fin del algoritmo.

Observación 2.4.6 *A la matriz* $[\mathbf{J_f(x)}]$, *por analogía con lo que representa la derivada de una función, se la denomina en algunos textos en lugar de matriz jacobiana, matriz tangente.*

Ilustremos el funcionamiento del método con un ejemplo.

Ejemplo 2.4.2 *(Propuesto en Hanna & Sandall [9]): La caída de presión en la circulación de un flujo turbulento en una tubería recta de sección circular constante puede estimarse mediante la expresión:*

$$\Delta p = \frac{\rho f L u^2}{2D}$$

donde ρ *es la densidad del fluido,* L *es la longitud de la tubería,* D *es el diámetro de la sección,* u *es la velocidad del fluido y* f *es el coeficiente de fricción de la tubería. Este coeficiente es a su vez proporcional al número de Reynolds* $Re = \frac{Du\rho}{\mu}$ *donde* μ *es la viscosidad del fluido) según la relación* $f = Re^{-0,25}$. *Asimismo, en los problemas de tuberías es usual trabajar no ya en términos de velocidad del fluido si no de caudal del fluido, siendo el caudal* $Q = \frac{\pi D^2}{4}.u$. *Con ello, la caída de presión en la tubería puede expresarse mediante una ley del tipo:*

$$\Delta p(x) = K(x).\,(Q(x))^{1,75}$$

donde $\Delta p(x)$ *es la pérdida de presión en el punto que dista* x *unidades de longitud de aquél respecto al que se mide la caída de presión,* $Q(x)$ *es el caudal en dicho punto y*

$$K(x) = \frac{\rho \left(\frac{\mu}{\rho D}\right)^{0,25} \left(\frac{4}{\pi D}\right)^{1,75}}{2D} x.$$

Considérese una tubería de sección circular que va del punto P1 al punto P2 y en él se divide en dos ramas, una que va al punto P3 y otra que va al punto P4. Designando por Q *al caudal que va de P1 a P2, por* Q_1 *al que va de P2 a P3, por* Q_2 *al que va de P2 a P4 y por* p_1, p_2, p_3 *y* p_4 *a las presiones en los puntos P1, P2, P3 y P4 respectivamente, las expresiones anteriores, junto a un balance de masa, nos conducen al sistema de ecuaciones no lineales:*

$$p_1 - p_2 = K_1 Q^{1,75}$$

$$p_2 - p_3 = K_2 Q_1^{1,75}$$

$$p_2 - p_4 = K_3 Q_2^{1,75}$$

$$Q = Q_1 + Q_2$$

Si para un fluido y una tubería concretos se han estimado los valores siguientes:

$$K_1 = 2{,}35e^{-3}, \qquad K_2 = 4{,}67e^{-3}, \qquad K_3 = 3{,}72e^{-2},$$

y

$$p_1 = 75\ psi, \qquad p_3 = 20\ psi, \qquad p_4 = 15\ psi$$

se desea estimar la presión p_2 existente en el punto $P2$ así como los caudales Q, Q_1 y Q_2 que circulan por cada una de las ramas de la red de tuberías antes descrita.

Solución:

El sistema dado puede escribirse, de acuerdo a los datos del ejercicio, como:

$$2{,}35e^{-3}Q^{1{,}75} - 75 + p_2 = 0$$

$$4{,}67e^{-3}Q_1^{1{,}75} + 20 - p_2 = 0$$

$$3{,}72e^{-2}Q_2^{1{,}75} + 15 - p_2 = 0$$

$$Q - Q_1 - Q_2 = 0$$

Este sistema de 4 ecuaciones con 4 incógnitas puede intentar resolverse tal cual está planteado mediante el método de Newton-Raphson (invitamos al lector a hacerlo). Pero el proceso puede tener problemas si alguno de los caudales (independientemente del sentido físico que ello pueda tener) se hace negativo ya que en ese caso no podrá calcularse $Q_i^{1{,}75} = \sqrt[4]{Q_i^7}$. Por dicho motivo es ventajoso en este caso utilizar la última ecuación inyectada en la primera reformulando el sistema como:

$$2{,}35e^{-3}(Q_1 + Q_2)^{1{,}75} - 75 + p_2 = 0$$

$$4{,}67e^{-3}Q_1^{1{,}75} + 20 - p_2 = 0$$

$$3{,}72e^{-2}Q_2^{1{,}75} + 15 - p_2 = 0$$

Este nuevo sistema ya sólo tiene 3 ecuaciones con 3 incógnitas. En él la función que define el sistema es:

$$\mathbf{f}(Q_1, Q_2, p_2) = \begin{pmatrix} 2{,}35e^{-3}(Q_1 + Q_2)^{1{,}75} - 75 + p_2 \\ 4{,}67e^{-3}Q_1^{1{,}75} + 20 - p_2 \\ 3{,}72e^{-2}Q_2^{1{,}75} + 15 - p_2 \end{pmatrix}$$

y la matriz jacobiana puede escribirse como:

$$[\mathbf{J_f}(Q_1, Q_2, p_2)] = \begin{bmatrix} (0{,}205(Q_1+Q_2)^{0,75}) & (0{,}205(Q_1+Q_2)^{0,75}) & 1 \\ 0{,}407Q_1^{0,75} & 0 & -1 \\ 0 & 0{,}881Q_2^{0,75} & -1 \end{bmatrix}$$

En cuanto a los valores de partida para inicializar el método, puesto que P2 es un punto intermedio entre P1 y los extremos P3 y P4, tomaremos como p_2 una presión intermedia, por ejemplo $p_2 = 50\ psi$. Para los caudales Q_1 y Q_2 no se dispone de ninguna pista que nos indique en qué entorno pueden estar. No obstante, si se considera $p_2 = 50$, de la segunda ecuación se tiene que $Q_1 \approx 16$ y, de la tercera ecuación, que $Q_2 \approx 7$ por lo que estos pueden ser valores coherentes con la presión tomada para inicializar el proceso.

Aplicando pues el algoritmo de Newton-Raphson antes descrito a esta situación, con los parámetros: $\varepsilon = 10^{-6}$, $\delta = 10^{-6}$, $maxiter$= 100, $\left\{Q_1^{(0)}, Q_2^{(0)}, p_2^{(0)}\right\}^T = \{16, 7, 50\}$ se tiene la siguiente sucesión de vectores:

$$\begin{pmatrix} Q_1^{(1)} \\ Q_2^{(1)} \\ p_2^{(1)} \end{pmatrix} = \begin{pmatrix} 14{,}0506076 \\ 10{,}4943950 \\ 43{,}4152926 \end{pmatrix} \rightarrow \begin{pmatrix} Q_1^{(2)} \\ Q_2^{(2)} \\ p_2^{(2)} \end{pmatrix} = \begin{pmatrix} 14{,}1344377 \\ 10{,}1343069 \\ 43{,}9558088 \end{pmatrix} \rightarrow$$

$$\rightarrow \begin{pmatrix} Q_1^{(3)} \\ Q_2^{(3)} \\ p_2^{(3)} \end{pmatrix} = \begin{pmatrix} 14{,}1355465 \\ 10{,}1303048 \\ 43{,}9596512 \end{pmatrix} \rightarrow \begin{pmatrix} Q_1^{(4)} \\ Q_2^{(4)} \\ p_2^{(4)} \end{pmatrix} = \begin{pmatrix} 14{,}1355467 \\ 10{,}1303043 \\ 43{,}9596517 \end{pmatrix}$$

no realizándose más iteraciones pues la norma-2 del vector diferencia entre los hallados en la 3^a y 4^a iteraciones es inferior a 10^{-6} y los valores de la función que define el sistema en este vector son:

$$\mathbf{f}(Q_1^{(4)}, Q_2^{(4)}, p_2^{(4)}) = \begin{pmatrix} 0{,}28727010^{-14} \\ 0{,}11588010^{-14} \\ 0{,}49323410^{-13} \end{pmatrix}$$

La solución buscada es:

$$Q_1 = 14{,}1355467$$

$$Q_2 = 10{,}1303043$$

$$Q = Q_1 + Q_2 = 24{,}265851$$

$$p_2 = 43{,}9596517 psi$$

Variantes del método de Newton-Raphson para sistemas: método de Newton modificado y métodos de cuasi-Newton.

El paso más costoso de la aplicación del método de Newton-Raphson consiste en la evaluación en cada paso de la matriz jacobiana $\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]$ (lo cual conlleva la evaluación de las n^2 funciones derivadas parciales primeras, que a su vez implicarán un número de operaciones elementales que dependerá de las expresiones de estas derivadas parciales) y de su inversión $\left(\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]^{-1}\right)$ o lo que es equivalente de la resolución del sistema lineal de ecuaciones algebraicas $\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]\boldsymbol{\delta}\mathbf{x}^{(i)} = \mathbf{f}(\mathbf{x}^{(i)})$, lo cual implica tener que hacer del orden de $O(\lambda n^3)$ operaciones, siendo n el número de incógnitas y ecuaciones del sistema y λ un parámetro menor o igual a (2/3) dependiendo de la estructura de la matriz jacobiana. Ello permite estimar el número de operaciones elementales en cada iteración como un valor proporcional a n^3 lo cual, cuando n toma valores elevados puede representar un coste computacional grande.

Para intentar solventar este problema en la aplicación práctica del método de Newton-Raphson se han desarrollado numerosos métodos que, con mayor o mejor fortuna según el sistema al que se aplique, tratan de aproximar ya sea $\left(\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]\right)$ o su inversa $\left(\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]^{-1}\right)$. Entre ellos señalamos los siguientes:

a) Aproximación de las derivadas mediante diferencias finitas.

En esta variante del método los valores de las derivadas que intervienen en la expresión de la matriz jacobiana se aproximan mediante fórmulas en diferencias finitas como por ejemplo:

$$\frac{\partial f_k}{\partial x_j}(\mathbf{x}^{(i)}) \approx$$

$$\frac{f_k(x_1^{(i)},..,x_{j-1}^{(i)},x_j^{(i)}+h_j^{(i)},x_{j+1}^{(i)},...,x_n^{(i)}) - f_k(x_1^{(i)},..,x_{j-1}^{(i)},x_j^{(i)},x_{j+1}^{(i)},...,x_n^{(i)})}{h_j^{(i)}}$$

donde los parámetros $h_j^{(i)}$ (j=1, .., n) son tomados en cada iteración "suficientemente" pequeños para asegurar una buena aproximación. Por ejemplo:

$$h_j^{(i)} = h^{(i)} = Inf\left(\frac{\left\|\mathbf{x}^{(i)}-\mathbf{x}^{(i-1)}\right\|}{10}, h^{(i-1)}\right)$$

Otra elección posible consiste en aproximar el valor de la derivada $\frac{\partial f_k}{\partial x_j}(\mathbf{x}^{(i)})$ con un cierto valor de $h_j^{(i)}$ y a continuación hacerlo con $\frac{h_j^{(i)}}{2}$ y comparar las dos aproximaciones obtenidas. Si la diferencia entre ambas es suficientemente pequeña se dará por buena la aproximación obtenida. Si es "elevada" se repetirá

el proceso comparando las aproximaciones obtenidas para $\frac{h_j^{(i)}}{2}$ y para $\frac{h_j^{(i)}}{4}$. Este proceso se continua hasta obtener dos aproximaciones de la derivada parcial suficientemente próximas, momento en el que una de ellas se toma como la aproximación buscada de la derivada parcial.

El proceso anterior nos permite estimar de forma aproximada la matriz jacobiana pero no nos elimina la necesidad de invertirla (o factorizarla) por lo que el número de operaciones por iteración sigue siendo proporcional a n^3.

b) Método de Newton modificado.

Esta variante consiste en utilizar durante la k primeras iteraciones (siendo k un número a predeterminar por el usuario del método) como aproximación de la matriz tangente la matriz $\left(\left[\mathbf{J_f}(\mathbf{x}^{(0)})\right]\right)$. Con ello en estas k iteraciones sólo se realiza una vez el cálculo de la matriz jacobiana y de su inversa (y si se optase por la resolución del sistema $\left[\mathbf{J_f}(\mathbf{x}^{(0)})\right]\boldsymbol{\delta}\mathbf{x}^{(i)} = \mathbf{f}(\mathbf{x}^{(i)})$ bastará con factorizar una vez la matriz $\left[\mathbf{J_f}(\mathbf{x}^{(0)})\right]$ en la primera iteración y conservar las matrices de la factorización). Realizadas estas k primeras iteraciones se calcula $\left[\mathbf{J_f}(\mathbf{x}^{(k)})\right]$ y se utiliza esta matriz en las k siguientes iteraciones tras las cuales se vuelve a actualizar obteniéndose $\left[\mathbf{J_f}(\mathbf{x}^{(2k)})\right]$ y continuando así el proceso.

Con ello, a costa de una pérdida de velocidad de convergencia, se logra que sólo las iteraciones 1^a, (k+1)-ésima, (2.k+1)-ésima, ... impliquen la realización de un número de operaciones proporcional a n^3 en tanto que el resto conllevarán del orden de n^2 operaciones.

c) Método de Jacobi

En este método la matriz tangente que interviene en cada iteración $\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]$ se sustituye por otra con la misma diagonal pero con todos sus demás elementos nulos. Más concretamente, denotando por $\left[\mathbf{D_f}(\mathbf{x}^{(i)})\right]$ a la matriz:

$$\left[\mathbf{D_f}(\mathbf{x}^{(i)})\right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}^{(i)}) & 0 & \dots & 0 \\ 0 & \frac{\partial f_2}{\partial x_2}(\mathbf{x}^{(i)}) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{\partial f_n}{\partial x_n}(\mathbf{x}^{(i)}) \end{bmatrix}$$

se utilizará el esquema iterativo:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{D_f}(\mathbf{x}^{(i)})\right]^{-1}\mathbf{f}(\mathbf{x}^{(i)}) \qquad (i = 0, 1, ...)$$

Esta forma de proceder efectivamente reduce de forma notable el número de operaciones (sólo conlleva evaluar n funciones derivadas en lugar de n^2

y además la inversión de una matriz diagonal sólo implica n operaciones). Pero sólo es válida si los elementos no diagonales de la matriz jacobiana son "pequeños" comparados con los términos diagonales.

d)Método de Gauss-Seidel

En esta variante del método de Newton-Raphson, la matriz tangente de cada iteración es sustituida por otra triangular inferior en la que los elementos de la diagonal y los que están por debajo de ella coinciden con los de la matriz jacobiana. Más concretamente, siendo $\left[\mathbf{G_f}(\mathbf{x}^{(i)})\right]$ la matriz:

$$\left[\mathbf{G_f}(\mathbf{x}^{(i)})\right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}^{(i)}) & 0 & ... & 0 \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}^{(i)}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}^{(i)}) & ... & 0 \\ ... & ... & ... & ... \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}^{(i)}) & \frac{\partial f_n}{\partial x_2}(\mathbf{x}^{(i)}) & ... & \frac{\partial f_n}{\partial x_n}(\mathbf{x}^{(i)}) \end{bmatrix}$$

el esquema que se emplea es:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{G_f}(\mathbf{x}^{(i)})\right]^{-1} \mathbf{f}(\mathbf{x}^{(i)}) \qquad (i = 0, 1, ...)$$

En esta variante del método de Newton-Raphson también se reduce de forma notable el número de operaciones (sólo conlleva evaluar $(n.(n+1)/2)$ funciones derivadas en lugar de n^2 y además la inversión de una matriz triangular sólo implica del orden de n^2 operaciones). Pero también su límite de validez lo marcará la pequeñez de los términos que se están despreciando en la matriz tangente.

e) Métodos de sobrerrelajación (SOR)

Con la misma notación empleada en la descripción de los métodos de Jacobi y de Gauss-Seidel, este método consiste en utilizar el esquema:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\rho\mathbf{D_f}(\mathbf{x}^{(i)}) + \mathbf{G_f}(\mathbf{x}^{(i)})\right]^{-1} \mathbf{f}(\mathbf{x}^{(i)}) \qquad (i = 0, 1, ...)$$

donde ρ es un parámetro que debe fijarse de antemano, llamándose parámetro de relajación al valor $\omega = \frac{1}{(1+\rho)}$. La matriz a invertir también es ahora triangular inferior por lo que el número de operaciones es similar al del método de Gauss-Seidel (que no es más que un caso particular de este cuando a ω se le da el valor 1). Pueden encontrarse detalles sobre las condiciones de convergencia de este método (y del de Gauss-Seidel) en J.M. Ortega & W.C. Rheinboldt [11].

f) Métodos de cuasi-Newton: Método de Broyden.

Este tipo de métodos generalizan el método de la secante en el sentido de

que la idea de la que parten consiste en aproximar la matriz jacobiana en cada iteración a partir de la matriz tangente utilizada en la iteración anterior. En este sentido la primera iteración del método se realiza como en el método de Newton pero a partir de la segunda iteración la matriz jacobiana $\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]$ es sustituida por otra matriz $\left[\mathbf{A}^{(i)}\right]$ que se obtiene a partir de $\left[\mathbf{A}^{(i-1)}\right]$, (siendo $\left[\mathbf{A}^{(0)}\right] = \left[\mathbf{J_f}(\mathbf{x}^{(0)})\right]$) y siguiéndose el esquema iterativo:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{A}^{(i)}\right]^{-1} \mathbf{f}(\mathbf{x}^{(i)}) \qquad (i = 0, 1, ...)$$

El más popular de este tipo de métodos es el conocido como **método de Broyden** que a continuación describimos. Este método se inspira en el método de la secante en el sentido de que en dicho método el valor de $f'(x_i)$ se aproximaba mediante la expresión:

$$f'(x_i) \approx f'_i = \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}$$

y reemplazando en el método de Newton la "tangente" $f'(x_i)$ por la "tangente aproximada" f'_i se obtenía el método de la secante. Obsérvese que de la expresión anterior se tiene que:

$$f'_i.(x_i - x_{i-1}) = f(x_i) - f(x_{i-1})$$

Utilizando esta idea, C.G. Broyden (en [2]) propuso reemplazar en cada iteración del método de Newton la matriz tangente $\left[\mathbf{J_f}(\mathbf{x}^{(i)})\right]$ por otra matriz $\left[\mathbf{A}^{(i)}\right]$ que verificase que:

$$\left[\mathbf{A}^{(i)}\right](\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}) = \mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) \qquad (1)$$

Junto a esta condición, C.G. Broyden impuso otra que pasamos a "justificar". Si se considera que $\mathbf{f}(\mathbf{x}) \in \left(C^2(D)\right)^n$ considerando desarrollos en serie de Taylor hasta el primer orden (es decir linealizando los desarrollos como se hace en el método de Newton) y siendo $\mathbf{x}^{(i-1)}$ y $\mathbf{x}^{(i)}$ dos vectores "suficientemente próximos" y $\mathbf{x}$ otro vector también suficientemente próximo a los anteriores, se tiene que:

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{x}^{(i)}) + \left[\mathbf{J_f}(\mathbf{x}^{(i)})\right](\mathbf{x} - \mathbf{x}^{(i)}) \qquad (2)$$

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{x}^{(i-1)}) + \left[\mathbf{J_f}(\mathbf{x}^{(i-1)})\right](\mathbf{x} - \mathbf{x}^{(i-1)}) \quad (3)$$

de donde restando (3) a (2) se tendrá que:

$$\mathbf{f}(\mathbf{x}^{(i)}) + \left[\mathbf{J_f}(\mathbf{x}^{(i)})\right](\mathbf{x} - \mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{J_f}(\mathbf{x}^{(i-1)})\right](\mathbf{x} - \mathbf{x}^{(i-1)}) \approx \mathbf{0} \quad (4)$$

Obviamente el razonamiento anterior sólo tendrá validez para puntos suficientemente cercanos. Es decir que si los vectores $\mathbf{x}^{(i)}$, $\mathbf{x}^{(i-1)}$ se suponen generados por el método de Newton, el razonamiento anterior sólo será válido en las cercanías de la solución y para vectores $\mathbf{x}$ también próximos a la solución.

En el método propuesto por Broyden las aproximaciones del vector solución se buscan en la forma:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{A}^{(i)}\right] \mathbf{f}(\mathbf{x}^{(i)})$$

sustituyéndose la matriz jacobiana por $\left[\mathbf{A}^{(i)}\right]$. Por ello la segunda condición, junto a (1), que impone Broyden es que se minimice para cualquier vector $\mathbf{x} \in I\!R^n$ (y en el sentido de cualquier norma) el vector:

$$\mathbf{f}(\mathbf{x}^{(i)}) + \left[\mathbf{A}^{(i)}\right](\mathbf{x} - \mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i-1)}\right](\mathbf{x} - \mathbf{x}^{(i-1)})$$

que puede escribirse de forma equivalente como:

$$\mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i)}\right](\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}) + \left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right](\mathbf{x} - \mathbf{x}^{(i-1)}) \qquad (5)$$

Si en (5) se introduce la primera condición impuesta (1) se obtiene finalmente que debe minimizarse el vector:

$$\left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right](\mathbf{x} - \mathbf{x}^{(i-1)})$$

El vector $(\mathbf{x} - \mathbf{x}^{(i-1)})$ puede a su vez expresarse como suma de un vector proporcional a $(\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)})$ más otro vector ortogonal a este es decir en la forma:

$$(\mathbf{x} - \mathbf{x}^{(i-1)}) = \alpha(\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}) + \mathbf{v} = \alpha\mathbf{u}^{(i)} + \mathbf{v}$$

donde $\mathbf{u}^{(i)} = (\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)})$ y $\mathbf{v}$ es un vector ortogonal al $\mathbf{u}^{(i)}$ es decir tal que: $\mathbf{v}^T\mathbf{u}^{(i)} = 0$. De esta forma:

$$\left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right](\mathbf{x} - \mathbf{x}^{(i-1)}) = \alpha\left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right].\mathbf{u}^{(i)} + \left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right]\mathbf{v}$$

El primer sumando de la expresión anterior es fijo pues por la condición (1) tomará el valor:

$$\alpha\left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right].\mathbf{u}^{(i)} = \alpha(\mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i-1)}\right]\mathbf{u}^{(i)})$$

Por tanto, debe minimizarse el valor de $\left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right]\mathbf{v}$ para cualquier vector $\mathbf{v}$ ortogonal a $\mathbf{u}^{(i)} = (\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)})$. Ello se logra obligando a que $\left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right]\mathbf{v} = \mathbf{0}$, es decir haciendo que todos los vectores fila de la matriz $\left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right]$ sean ortogonales a $\mathbf{v}$, o lo que es lo mismo que todas las

filas de $\left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right]$ sean proporcionales a $\mathbf{u}^{(i)}$. En otros términos la matriz $\left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right]$ debe ser de la forma $\mathbf{b}(\mathbf{u}^{(i)})^T$ donde $\mathbf{b}$ es un vector columna formado por los factores de proporcionalidad de cada fila. Para determinar el valor de las componentes del vector $\mathbf{b}$ puede volverse a recurrir a la condición (1) de forma que:

$$\left[\mathbf{A}^{(i)}\right]\mathbf{u}^{(i)} = \mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) \Rightarrow$$

$$\Rightarrow \left[\mathbf{A}^{(i)} - \mathbf{A}^{(i-1)}\right]\mathbf{u}^{(i)} = \mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i-1)}\right]\mathbf{u}^{(i)} \Rightarrow$$

$$\Rightarrow \mathbf{b}(\mathbf{u}^{(i)})^T\mathbf{u}^{(i)} = \mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i-1)}\right]\mathbf{u}^{(i)} \Rightarrow$$

$$\Rightarrow \mathbf{b} = \frac{1}{(\mathbf{u}^{(i)})^T\mathbf{u}^{(i)}}\left(\mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i-1)}\right]\mathbf{u}^{(i)}\right)$$

Siendo la anterior la expresión del vector de proporcionalidad $\mathbf{b}$ se tiene que:

$$\left[\mathbf{A}^{(i)}\right] = \left[\mathbf{A}^{(i-1)}\right] + \frac{\left(\mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i-1)}\right]\mathbf{u}^{(i)}\right)(\mathbf{u}^{(i)})^T}{(\mathbf{u}^{(i)})^T\mathbf{u}^{(i)}}$$

La evaluación de la expresión anterior, por "aparatosa" que parezca no es excesivamente costosa. En ella debe estimarse en primer lugar el vector $\left[\mathbf{A}^{(i-1)}\right]\mathbf{u}^{(i)}$ para lo que se realizan $(2n^2)$ operaciones; tras ello se calcula el vector dado por

$$\left(\mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i-1)}\right]\mathbf{u}^{(i)}\right)$$

lo que implica realizar otras $(2n)$ operaciones; posteriormente se evaluará el producto:

$$\left(\mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i-1)}\right]\mathbf{u}^{(i)}\right)(\mathbf{u}^{(i)})^T$$

lo que conllevará hacer otras (n^2) operaciones; asimismo debe evaluarse el producto escalar:

$$(\mathbf{u}^{(i)})^T\mathbf{u}^{(i)}$$

lo que podrá hacerse con otras $(2n)$ operaciones. Finalmente se realizará la suma de matrices

$$\left[\mathbf{A}^{(i-1)}\right] + \frac{1}{(\mathbf{u}^{(i)})^T\mathbf{u}^{(i)}}\left(\left(\mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i-1)}\right]\mathbf{u}^{(i)}\right)(\mathbf{u}^{(i)})^T\right)$$

lo que necesita nuevamente del orden n^2 operaciones. En resumen el método exige del orden de $O(4n^2)$ operaciones.

Ahora bien, poco habríamos ganado en número de operaciones si no se encuentra una forma eficaz de calcular $\left[\mathbf{A}^{(i)}\right]^{-1}$ pues recordemos que en el método de Broyden se realiza la operación:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left[\mathbf{A}^{(i)}\right]^{-1} f(\mathbf{x}^{(i)})$$

y el proceso de inversión nos devolvería a estar en el rango de operaciones de $O(n^3)$. Afortunadamente, la inversión también puede realizarse manteniéndose el orden de operaciones en $O(n^2)$ utilizando la expresión de Sherman-Morrison-Woodbury que se recoge en la proposición siguiente:

Proposición 2.4.2 *Siendo* $[\mathbf{A}]$ *una matriz real cuadrada de orden n regular y siendo* $\mathbf{u}$ *y* $\mathbf{v}$ *dos vectores reales de n componentes tales que:*

$$\mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v} \neq -1$$

se verifica que la matriz $\left([\mathbf{A}] + \mathbf{v}\mathbf{u}^T\right)$ *también es regular y:*

$$\left([\mathbf{A}] + \mathbf{v}\mathbf{u}^T\right)^{-1} = [\mathbf{A}]^{-1} - \frac{1}{1 + \mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}} [\mathbf{A}]^{-1} \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1}$$

Demostración: Si $[\mathbf{A}]$ es una matriz regular y $\mathbf{u}$ y $\mathbf{v}$ son dos vectores tales que

$$\mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v} \neq -1$$

se podrá calcular la matriz

$$[\mathbf{A}]^{-1} - \frac{1}{1 + \mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}} [\mathbf{A}]^{-1} \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1}.$$

Comprobemos que esta matriz es efectivamente la inversa de $\left([\mathbf{A}] + \mathbf{v}\mathbf{u}^T\right)$. Para ello:

$$\left([\mathbf{A}] + \mathbf{v}\mathbf{u}^T\right) \left([\mathbf{A}]^{-1} - \frac{1}{1 + \mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}} [\mathbf{A}]^{-1} \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1}\right) =$$

$$= [\mathbf{A}] [\mathbf{A}]^{-1} - \frac{1}{1 + \mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}} [\mathbf{A}] [\mathbf{A}]^{-1} \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} +$$

$$+ \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} - \frac{1}{1 + \mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}} \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} =$$

$$= \mathbf{I} - \frac{1}{1 + \mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}} \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} + \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} -$$

$$- \frac{\mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}}{1 + \mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}} \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} =$$

$$= \mathbf{I} - \frac{1 + \mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}}{1 + \mathbf{u}^T [\mathbf{A}]^{-1} \mathbf{v}} \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} + \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} =$$

$$= \mathbf{I} - \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} + \mathbf{v}\mathbf{u}^T [\mathbf{A}]^{-1} = \mathbf{I}$$

c.q.d.

Para aplicar la expresión de Sherman-Morrison-Woodbury a la expresión:

$$\left[\mathbf{A}^{(i)}\right] = \left[\mathbf{A}^{(i-1)}\right] + \frac{\left(\mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}) - \left[\mathbf{A}^{(i-1)}\right] \mathbf{u}^{(i)}\right) (\mathbf{u}^{(i)})^T}{(\mathbf{u}^{(i)})^T \mathbf{u}^{(i)}}.$$

Supondremos que $\left[\mathbf{A}^{(i-1)}\right]$ es regular y utilizaremos la siguiente notación:

$$\boldsymbol{\Delta}^{(i)}\mathbf{f} = \mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)}),$$

$$\mathbf{w}^{(i)} = \frac{1}{(\mathbf{u}^{(i)})^T \mathbf{u}^{(i)}} \left(\boldsymbol{\Delta}^{(i)}\mathbf{f} - \left[\mathbf{A}^{(i-1)}\right] \mathbf{u}^{(i)}\right) = \frac{1}{||\mathbf{u}^{(i)}||_2} \left(\boldsymbol{\Delta}^{(i)}\mathbf{f} - \left[\mathbf{A}^{(i-1)}\right] \mathbf{u}^{(i)}\right)$$

con lo que

$$\left[\mathbf{A}^{(i)}\right] = \left[\mathbf{A}^{(i-1)}\right] + \mathbf{w}^{(i)} (\mathbf{u}^{(i)})^T$$

por lo que finalmente

$$\left[\mathbf{A}^{(i)}\right]^{-1} = \left[\mathbf{A}^{(i-1)}\right]^{-1} - \frac{\left[\mathbf{A}^{(i-1)}\right]^{-1} \mathbf{w}^{(i)} (\mathbf{u}^{(i)})^T \left[\mathbf{A}^{(i-1)}\right]^{-1}}{1 + (\mathbf{u}^{(i)})^T \left[\mathbf{A}^{(i-1)}\right]^{-1} \mathbf{w}^{(i)}} =$$

$$= \left[\mathbf{A}^{(i-1)}\right]^{-1} - \frac{\left[\mathbf{A}^{(i-1)}\right]^{-1} \frac{1}{||\mathbf{u}^{(i)}||_2} \left(\boldsymbol{\Delta}^{(i)}\mathbf{f} - \left[\mathbf{A}^{(i-1)}\right] \mathbf{u}^{(i)}\right) (\mathbf{u}^{(i)})^T \left[\mathbf{A}^{(i-1)}\right]^{-1}}{1 + (\mathbf{u}^{(i)})^T \left[\mathbf{A}^{(i-1)}\right]^{-1} \frac{1}{||\mathbf{u}^{(i)}||_2} \left(\boldsymbol{\Delta}^{(i)}\mathbf{f} - \left[\mathbf{A}^{(i-1)}\right] \mathbf{u}^{(i)}\right)} =$$

$$= \left[\mathbf{A}^{(i-1)}\right]^{-1} - \frac{\left(\left[\mathbf{A}^{(i-1)}\right]^{-1} \boldsymbol{\Delta}^{(i)}\mathbf{f} - \mathbf{u}^{(i)}\right) (\mathbf{u}^{(i)})^T \left[\mathbf{A}^{(i-1)}\right]^{-1}}{||\mathbf{u}^{(i)}||_2 + (\mathbf{u}^{(i)})^T \left[\mathbf{A}^{(i-1)}\right]^{-1} \boldsymbol{\Delta}^{(i)}\mathbf{f} - (\mathbf{u}^{(i)})^T \mathbf{u}^{(i)}} =$$

$$= \left[\mathbf{A}^{(i-1)}\right]^{-1} + \frac{\left(\mathbf{u}^{(i)} - \left[\mathbf{A}^{(i-1)}\right]^{-1} \boldsymbol{\Delta}^{(i)}\mathbf{f}\right) (\mathbf{u}^{(i)})^T \left[\mathbf{A}^{(i-1)}\right]^{-1}}{(\mathbf{u}^{(i)})^T \left[\mathbf{A}^{(i-1)}\right]^{-1} \boldsymbol{\Delta}^{(i)}\mathbf{f}}$$

Examinemos el orden del número de operaciones que conlleva la aplicación de la fórmula anterior (supuestos conocidos la matriz $\left[\mathbf{A}^{(i-1)}\right]^{-1}$, y los vectores $\boldsymbol{\Delta}^{(i)}\mathbf{f} = \mathbf{f}(\mathbf{x}^{(i)}) - \mathbf{f}(\mathbf{x}^{(i-1)})$ y $\mathbf{u}^{(i)} = (\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)})$). El cálculo de

$$(\mathbf{z}^{(i)})^T = (\mathbf{u}^{(i)})^T \left[\mathbf{A}^{(i-1)}\right]^{-1}$$

implica realizar $(2n^2)$ operaciones elementales. El cálculo de

$$\mathbf{y}^{(i)} = \left[\mathbf{A}^{(i-1)}\right]^{-1} \boldsymbol{\Delta}^{(i)}\mathbf{f}$$

implica otras $(2n^2)$ operaciones. La estimación de

$$\mathbf{r}^{(i)} = \left(\mathbf{u}^{(i)} - \mathbf{y}^{(i)}\right)$$

conlleva n operaciones elementales más. La determinación de la matriz del numerador:

$$\left[\mathbf{M}^{(i)}\right] = \mathbf{r}^{(i)} \left(\mathbf{z}^{(i)}\right)^T$$

se realiza con n^2 operaciones elementales adicionales. El cálculo del escalar del denominador

$$\alpha^{(i)} = (\mathbf{z}^{(i)})^T \boldsymbol{\Delta}^{(i)} \mathbf{f}$$

exige realizar otras $2n$ operaciones elementales. El multiplicar el escalar $\left(\alpha^{(i)}\right)^{-1}$ por las componentes de la matriz $\mathbf{M}^{(i)}$, estimando la matriz $[\boldsymbol{\Delta}\mathbf{A}^{(i)}]$, conlleva otras n^2 operaciones elementales más. Finalmente el sumar $\left[\mathbf{A}^{(i-1)}\right]^{-1}$ y $[\boldsymbol{\Delta}\mathbf{A}^{(i)}]$ para obtener $\left[\mathbf{A}^{(i)}\right]^{-1}$ se puede realizar con n^2 operaciones elementales. En total se realizan por tanto: $7n^2 + 3n \sim O(7n^2)$ operaciones en cada iteración.

Comparando las operaciones que se realizan en cada iteración del método de Newton ($O(\frac{2}{3}n^3)$) con las necesarias en cada iteración de este método, se concluye que operacionalmente cada iteración del método de Broyden es ventajosa siempre que $\frac{2}{3}n > 7$, es decir para valores de n superiores a 10.

Según todo lo anterior un algoritmo del método de Broyden es el que se recoge a continuación:

Algoritmo del método de Broyden para sistemas no lineales

Dado el sistema de ecuaciones no lineales $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, los indicadores de precisión ε y δ, un valor máximo del número de iteraciones que se permiten realizar $(maxiter)$ y un vector $\mathbf{x}$ con el que inicializar el proceso,

$iteración \leftarrow 1$
$\mathbf{v} \leftarrow \mathbf{f}(\mathbf{x})$
Evaluar la matriz $[\mathbf{J_f}(\mathbf{x})]$
Evaluar la matriz $[\mathbf{A}] = [\mathbf{J_f}(\mathbf{x})]^{-1}$
Calcular $\mathbf{u} = -[\mathbf{A}]\,\mathbf{v}$
Calcular $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{u}$
$\mathbf{w} \leftarrow \mathbf{f}(\mathbf{x})$
$tolx \leftarrow ||\mathbf{u}||$
$tolf \leftarrow ||\mathbf{w}||$
Mientras ($(iteración < maxiter)$ **y** $((tolx > \varepsilon)$ **o** $(tolf > \delta)$ **), hacer:**
    $\boldsymbol{\delta}\mathbf{f} \leftarrow \mathbf{w} - \mathbf{v}$

$\mathbf{v} \leftarrow \mathbf{w}$
$\mathbf{z} \leftarrow [\mathbf{A}]^T\mathbf{u}$
$\mathbf{y} \leftarrow [\mathbf{A}]\boldsymbol{\delta}\mathbf{f}$
$\mathbf{r} \leftarrow \mathbf{u} - \mathbf{y}$
$\alpha \leftarrow \mathbf{z}^T\boldsymbol{\delta}\mathbf{f}$
$[\mathbf{A}] \leftarrow [\mathbf{A}] + \frac{1}{\alpha}\mathbf{r}\mathbf{z}^T$
$\mathbf{u} \leftarrow -[\mathbf{A}]\mathbf{v}$
$\mathbf{x} \leftarrow \mathbf{x} + \mathbf{u}$
$\mathbf{w} \leftarrow \mathbf{f}(\mathbf{x})$
$tolx \leftarrow ||\mathbf{u}||$
$tolf \leftarrow ||\mathbf{w}||$
$iteración \leftarrow iteración + 1$

Fin bucle condicional.
Si $((tolx < \varepsilon)$ **y** $(tolf < \delta)$) **entonces:**
tomar $\mathbf{x}$ como solución
si no:
Escribir un mensaje de error en el proceso de cálculo
fin condición.
Fin del algoritmo.

Apliquemos el algoritmo anterior al ejemplo considerado anteriormente para ilustrar el método de Newton, es decir, a la resolución del sistema de ecuaciones:

$$\mathbf{f}(Q_1, Q_2, p_2) = \begin{pmatrix} 2{,}35e^{-3}(Q_1+Q_2)^{1,75} - 75 + p_2 \\ 4{,}67e^{-3}Q_1^{1,75} + 20 - p_2 \\ 3{,}72e^{-2}Q_2^{1,75} + 15 - p_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

cuya matriz jacobiana está dada por:

$$[\mathbf{J}_\mathbf{f}(Q_1, Q_2, p_2)] = \begin{bmatrix} (0{,}205(Q_1+Q_2)^{0,75}) & (0{,}205(Q_1+Q_2)^{0,75}) & 1 \\ 0{,}407Q_1^{0,75} & 0 & -1 \\ 0 & 0{,}881Q_2^{0,75} & -1 \end{bmatrix}$$

Comenzaremos el proceso con los mismos valores iniciales que utilizamos en el método de Newton, es decir:

$$\mathbf{x}^{(0)} = \left\{Q_1^{(0)}, Q_2^{(0)}, p_2^{(0)}\right\}^T = \{16, 7, 50\}^T$$

La **primera iteración** del método de Broyden coincide con la de Newton por lo que en ella se tiene que:

$$\mathbf{v} = \mathbf{f}(16, 7, 50) = \begin{pmatrix} 3{,}2623408 \\ -2{,}3928201 \\ -19{,}8338430 \end{pmatrix}$$

con

$$\mathbf{A} = [\mathbf{J_f}(16, 7, 50)]^{-1} = \begin{bmatrix} 0{,}1379003 & 0{,}2161114 & -0{,}0782111 \\ 0{,}1183890 & -0{,}0782111 & 0{,}1966001 \\ 0{,}4488765 & -0{,}2965403 & -0{,}2545832 \end{bmatrix}$$

por lo que,

$$\mathbf{u} = -\mathbf{A}\mathbf{v} = \begin{pmatrix} -1{,}9493924 \\ 3{,}4943950 \\ -6{,}5847074 \end{pmatrix}$$

siendo la nueva aproximación del vector solución:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{u} = \begin{pmatrix} 14{,}0506076 \\ 10{,}4943950 \\ 43{,}4152926 \end{pmatrix}$$

y el valor de la función que define el sistema en este punto:

$$\mathbf{w} = \mathbf{f}(\mathbf{x}^{(1)}) = \begin{pmatrix} 8{,}32323310^{-2} \\ 2{,}92979610^{-1} \\ 2{,}3902906 \end{pmatrix}$$

siendo tanto $||\mathbf{u}||_2$ como $||\mathbf{w}||_2$ superiores a las tolerancias respectivas (10^{-6} para ambas) por lo que se comienzan las siguientes iteraciones. En la **segunda iteración**:

$$\boldsymbol{\delta}\mathbf{f} = \mathbf{w} - \mathbf{v} = (\mathbf{f}(\mathbf{x}^{(1)}) - \mathbf{f}(\mathbf{x}^{(0)})) = \begin{pmatrix} -3{,}1791085 \\ 0{,}5322596 \\ 22{,}2241337 \end{pmatrix}$$

pasando a actualizar el vector $\mathbf{v}$ mediante

$$\mathbf{v} = \mathbf{w} = \mathbf{f}(\mathbf{x}^{(1)}) = \begin{pmatrix} 8{,}32323310^{-2} \\ 2{,}92979610^{-1} \\ 2{,}3902906 \end{pmatrix}$$

Con ello los vectores $\mathbf{z}$, $\mathbf{y}$ y $\mathbf{r}$ del algoritmo de Broyden en esta iteración serán:

$$z = [\mathbf{A}]^T\mathbf{u} = \begin{pmatrix} -2{,}8108445 \\ 1{,}2580449 \\ 2{,}5158179 \end{pmatrix}$$

$$\mathbf{y} = [\mathbf{A}]\boldsymbol{\delta}\mathbf{f} = \begin{pmatrix} -2{,}0615460 \\ 3{,}9512660 \\ -7{,}2427538 \end{pmatrix}$$

$$\mathbf{r} = \mathbf{u} - \mathbf{y} = \begin{pmatrix} 0{,}1121536 \\ -0{,}4568710 \\ 0{,}6580464 \end{pmatrix}$$

Con estos vectores se tiene que:

$$\alpha = \mathbf{z}^T \boldsymbol{\delta}\mathbf{f} = 65{,}5174604$$

por lo que la matriz tangente actualizada será:

$$\mathbf{A} \leftarrow [\mathbf{A}] + \frac{1}{\alpha}\mathbf{r}\mathbf{z}^T =$$

$$\begin{bmatrix} 1{,}330887 \cdot 10^{-1} & 2{,}182650 \cdot 10^{-1} & -7{,}390446 \cdot 10^{-2} \\ 1{,}379898 \cdot 10^{-1} & -8{,}698375 \cdot 10^{-2} & 1{,}790566 \cdot 10^{-1} \\ 4{,}206449 \cdot 10^{-1} & -2{,}839047 \cdot 10^{-1} & -2{,}293147 \cdot 10^{-1} \end{bmatrix}$$

Con esta matriz se tendrá que el nuevo vector de incrementos será:

$$\mathbf{u} = \mathbf{x}^{(2)} - \mathbf{x}^{(1)} = -[\mathbf{A}]\mathbf{v} = \begin{pmatrix} 1{,}016291 \cdot 10^{-1} \\ -4{,}139981 \cdot 10^{-1} \\ 5{,}962953 \cdot 10^{-1} \end{pmatrix}$$

por lo que,

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \mathbf{u} = \begin{pmatrix} 14{,}1522367 \\ 10{,}0803968 \\ 44{,}0115879 \end{pmatrix}$$

siendo

$$\mathbf{w} = \mathbf{f}(\mathbf{x}^{(2)}) = \begin{pmatrix} -2{,}238497 \cdot 10^{-2} \\ -2{,}407702 \cdot 10^{-3} \\ -3{,}011495 \cdot 10^{-1} \end{pmatrix}$$

Puesto que tanto $||\mathbf{u}||_2$ como $||\mathbf{w}||_2$ son superiores a las tolerancias respectivas (10^{-6} para ambas) se pasará a realizar la **tercera iteración**. En ella:

$$\boldsymbol{\delta}\mathbf{f} = \mathbf{w} - \mathbf{v} = (\mathbf{f}(\mathbf{x}^{(2)}) - \mathbf{f}(\mathbf{x}^{(1)})) = \begin{pmatrix} -1{,}056173 \cdot 10^{-1} \\ -2{,}953853 \cdot 10^{-1} \\ -2{,}6914401 \end{pmatrix}$$

pasando a actualizar el vector $\mathbf{v}$ mediante:

$$\mathbf{v} = \mathbf{w} = \mathbf{f}(\mathbf{x}^{(2)}) = \begin{pmatrix} -2{,}238497 \cdot 10^{-2} \\ -2{,}407702 \cdot 10^{-3} \\ -3{,}011495 \cdot 10^{-1} \end{pmatrix}$$

Con ello los vectores $\mathbf{z}$, $\mathbf{y}$ y $\mathbf{r}$ del algoritmo de Broyden en esta iteración serán:

$$z = [\mathbf{A}]^T\mathbf{u} = \begin{pmatrix} 2{,}0722672 \\ -1{,}1109786 \\ -2{,}1837922 \end{pmatrix}$$

$$\mathbf{y} = [\mathbf{A}]\boldsymbol{\delta}\mathbf{f} = \begin{pmatrix} 1{,}2038069 \\ -4{,}7080045 \\ 6{,}5662072 \end{pmatrix}$$

$$\mathbf{r} = \mathbf{u} - \mathbf{y} = \begin{pmatrix} -1{,}875159 \cdot 10^{-2} \\ 5{,}680227 \cdot 10^{-2} \\ -6{,}032545 \cdot 10^{-2} \end{pmatrix}$$

Con estos vectores se tiene que:

$$\alpha = \mathbf{z}^T \boldsymbol{\delta}\mathbf{f} = 0{,}5986845$$

por lo que la matriz tangente actualizada será:

$$\mathbf{A} \leftarrow [\mathbf{A}] + \frac{1}{\alpha}\mathbf{r}\mathbf{z}^T =$$

$$\begin{bmatrix} 1{,}265981 \cdot 10^{-1} & 2{,}217447 \cdot 10^{-1} & -6{,}706453 \cdot 10^{-2} \\ 1{,}576511 \cdot 10^{-1} & -9{,}752455 \cdot 10^{-2} & 1{,}583371 \cdot 10^{-1} \\ 3{,}9976401 \cdot 10^{-1} & -2{,}727101 \cdot 10^{-1} & -2{,}073101 \cdot 10^{-1} \end{bmatrix}$$

Con esta matriz se tendrá que el nuevo vector de incrementos será:

$$\mathbf{u} = \mathbf{x}^{(3)} - \mathbf{x}^{(2)} = -[\mathbf{A}]\mathbf{v} = \begin{pmatrix} -1{,}68286610^{-2} \\ 5{,}09773410^{-2} \\ -5{,}41392310^{-2} \end{pmatrix}$$

por lo que,

$$\mathbf{x}^{(3)} = \mathbf{x}^{(2)} + \mathbf{u} = \begin{pmatrix} 14{,}1354080 \\ 10{,}1313741 \\ 43{,}9574486 \end{pmatrix}$$

siendo

$$\mathbf{w} = \mathbf{f}(\mathbf{x}^{(3)}) = \begin{pmatrix} -1{,}184048 \cdot 10^{-4} \\ 1{,}791805 \cdot 10^{-3} \\ 7{,}555445 \cdot 10^{-3} \end{pmatrix}$$

Puesto que tanto $||\mathbf{u}||_2$ como $||\mathbf{w}||_2$ son superiores a las tolerancias respectivas (10^{-6} para ambas) se pasará a realizar la **cuarta iteración**. En ella:

$$\boldsymbol{\delta}\mathbf{f} = \mathbf{w} - \mathbf{v} = (\mathbf{f}(\mathbf{x}^{(3)}) - \mathbf{f}(\mathbf{x}^{(2)})) = \begin{pmatrix} 2{,}226656 \cdot 10^{-2} \\ 4{,}199507 \cdot 10^{-3} \\ 3{,}087049 \cdot 10^{-1} \end{pmatrix}$$

pasando a actualizar el vector $\mathbf{v}$ mediante:

$$\mathbf{v} = \mathbf{w} = \mathbf{f}(\mathbf{x}^{(3)}) = \begin{pmatrix} -1{,}184048 \cdot 10^{-4} \\ 1{,}791805 \cdot 10^{-3} \\ 7{,}555445 \cdot 10^{-3} \end{pmatrix}$$

Con ello los vectores $\mathbf{z}$, $\mathbf{y}$ y $\mathbf{r}$ del algoritmo de Broyden en esta iteración serán:

$$\mathbf{z} = [\mathbf{A}]^T\mathbf{u} = \begin{pmatrix} -1{,}573676 \cdot 10^{-2} \\ 6{,}061109 \cdot 10^{-3} \\ 2{,}042382 \cdot 10^{-2} \end{pmatrix}$$

$$\mathbf{y} = [\mathbf{A}]\boldsymbol{\delta}\mathbf{f} = \begin{pmatrix} -1{,}695303 \cdot 10^{-2} \\ 5{,}198024 \cdot 10^{-2} \\ -5{,}624153 \cdot 10^{-2} \end{pmatrix}$$

$$\mathbf{r} = \mathbf{u} - \mathbf{y} = \begin{pmatrix} 1{,}243689 \cdot 10^{-4} \\ -1{,}002896 \cdot 10^{-3} \\ 2{,}102297 \cdot 10^{-3} \end{pmatrix}$$

Con estos vectores se tiene que:

$$\alpha = \mathbf{z}^T\boldsymbol{\delta}\mathbf{f} = 5{,}979984 \cdot 10^{-3}$$

por lo que la matriz tangente actualizada será:

$$\mathbf{A} \leftarrow [\mathbf{A}] + \frac{1}{\alpha}\mathbf{r}\mathbf{z}^T =$$

$$\begin{bmatrix} 1{,}26270810^{-1} & 2{,}21870810^{-1} & -6{,}66397710^{-2} \\ 1{,}60290310^{-1} & -9{,}85410510^{-2} & 1{,}54911810^{-1} \\ 3{,}94231710^{-1} & -2{,}70579310^{-1} & -2{,}03184810^{-1} \end{bmatrix}$$

Con esta matriz se tendrá que el nuevo vector de incrementos será:

$$\mathbf{u} = \mathbf{x}^{(4)} - \mathbf{x}^{(3)} = -[\mathbf{A}]\mathbf{v} = \begin{pmatrix} 1{,}208950 \cdot 10^{-4} \\ -9{,}748824 \cdot 10^{-4} \\ 2{,}043575 \cdot 10^{-3} \end{pmatrix}$$

por lo que,

$$\mathbf{x}^{(4)} = \mathbf{x}^{(3)} + \mathbf{u} = \begin{pmatrix} 14{,}1355289 \\ 10{,}1303992 \\ 43{,}9594922 \end{pmatrix}$$

siendo

$$\mathbf{w} = \mathbf{f}(\mathbf{x}^{(4)}) = \begin{pmatrix} 1{,}343715 \cdot 10^{-5} \\ 1{,}068316 \cdot 10^{-4} \\ 6{,}345653 \cdot 10^{-4} \end{pmatrix}$$

Puesto que tanto $||\mathbf{u}||_2 = 2{,}267424 \cdot 10^{-3}$ como $||\mathbf{w}||_2 = 6{,}436355 \cdot 10^{-4}$ son superiores a las tolerancias respectivas (10^{-6} para ambas) se pasará a realizar la **quinta iteración**. En ella:

$$\boldsymbol{\delta}\mathbf{f} = \mathbf{w} - \mathbf{v} = (\mathbf{f}(\mathbf{x}^{(4)}) - \mathbf{f}(\mathbf{x}^{(3)})) = \begin{pmatrix} 1{,}318420 \cdot 10^{-4} \\ -1{,}684974 \cdot 10^{-3} \\ -6{,}920880 \cdot 10^{-3} \end{pmatrix}$$

pasando a actualizar el vector $\mathbf{v}$ mediante:

$$\mathbf{v} = \mathbf{w} = \mathbf{f}(\mathbf{x}^{(4)}) = \begin{pmatrix} 1{,}343715 \cdot v10^{-5} \\ 1{,}068316 \cdot 10^{-4} \\ 6{,}345653 \cdot 10^{-4} \end{pmatrix}$$

Con ello los vectores $\mathbf{z}$, $\mathbf{y}$ y $\mathbf{r}$ del algoritmo de Broyden en esta iteración serán:

$$\mathbf{z} = [\mathbf{A}]^T\mathbf{u} = \begin{pmatrix} 6{,}646434 \cdot 10^{-4} \\ -4{,}300603 \cdot 10^{-4} \\ -5{,}680580 \cdot 10^{-4} \end{pmatrix}$$

$$\mathbf{y} = [\mathbf{A}]\boldsymbol{\delta}\mathbf{f} = \begin{pmatrix} 1{,}040072 \cdot 10^{-4} \\ -8{,}849542 \cdot 10^{-4} \\ 1{,}892971 \cdot 10^{-3} \end{pmatrix}$$

$$\mathbf{r} = \mathbf{u} - \mathbf{y} = \begin{pmatrix} 1{,}688776 \cdot 10^{-5} \\ -8{,}992822 \cdot 10^{-5} \\ 1{,}506046 \cdot 10^{-4} \end{pmatrix}$$

Con estos vectores se tiene que:

$$\alpha = \mathbf{z}^T\boldsymbol{\delta}\mathbf{f} = 4{,}743729 \cdot 10^{-6}$$

por lo que la matriz tangente actualizada será:

$$\mathbf{A} \leftarrow [\mathbf{A}] + \frac{1}{\alpha}\mathbf{r}\mathbf{z}^T =$$

$$\begin{bmatrix} 1{,}286370 \cdot 10^{-1} & 2{,}203397 \cdot 10^{-1} & -6{,}866206 \cdot 10^{-2} \\ 1{,}476905 \cdot 10^{-1} & -9{,}038827 \cdot 10^{-2} & 1{,}656807 \cdot 10^{-1} \\ 4{,}153328 \cdot 10^{-1} & -2{,}842329 \cdot 10^{-1} & -2{,}181648 \cdot 10^{-1} \end{bmatrix}$$

Con esta matriz se tendrá que el nuevo vector de incrementos será:

$$\mathbf{u} = \mathbf{x}^{(5)} - \mathbf{x}^{(4)} = -[\mathbf{A}]\mathbf{v} = \begin{pmatrix} 1{,}830280 \cdot 10^{-5} \\ -9{,}746342 \cdot 10^{-5} \\ 1{,}632240 \cdot 10^{-4} \end{pmatrix}$$

por lo que,

$$\mathbf{x}^{(5)} = \mathbf{x}^{(4)} + \mathbf{u} = \begin{pmatrix} 14{,}1355472 \\ 10{,}1303018 \\ 43{,}9596554 \end{pmatrix}$$

siendo

$$\mathbf{w} = \mathbf{f}(\mathbf{x}^{(5)}) = \begin{pmatrix} -5{,}450506 \cdot 10^{-7} \\ -2{,}101924 \cdot 10^{-6} \\ -1{,}624559 \cdot 10^{-5} \end{pmatrix}$$

Puesto que tanto $||\mathbf{u}||_2 = 1{,}909874 \cdot 10^{-4}$ como $||\mathbf{w}||_2 = 1{,}639007 \cdot 10^{-5}$ son superiores a las tolerancias respectivas (10^{-6} para ambas) se pasará a realizar la **sexta iteración**. En ella:

$$\boldsymbol{\delta}\mathbf{f} = \mathbf{w} - \mathbf{v} = (\mathbf{f}(\mathbf{x}^{(5)}) - \mathbf{f}(\mathbf{x}^{(4)})) = \begin{pmatrix} -1{,}398220 \cdot 10^{-5} \\ -1{,}089335 \cdot 10^{-4} \\ -6{,}508109 \cdot 10^{-4} \end{pmatrix}$$

pasando a actualizar el vector $\mathbf{v}$ mediante:

$$\mathbf{v} = \mathbf{w} = \mathbf{f}(\mathbf{x}^{(5)}) = \begin{pmatrix} -5{,}450506 \cdot 10^{-7} \\ -2{,}101924 \cdot 10^{-6} \\ -1{,}624559 \cdot 10^{-5} \end{pmatrix}$$

Con ello los vectores $\mathbf{z}$, $\mathbf{y}$ y $\mathbf{r}$ del algoritmo de Broyden en esta iteración serán:

$$\mathbf{z} = [\mathbf{A}]^T\mathbf{u} = \begin{pmatrix} 5{,}575227 \cdot 10^{-5} \\ -3{,}355124 \cdot 10^{-5} \\ -5{,}301423 \cdot 10^{-5} \end{pmatrix}$$

$$\mathbf{y} = [\mathbf{A}]\boldsymbol{\delta}\mathbf{f} = \begin{pmatrix} 1{,}888501 \cdot 10^{-5} \\ -1{,}000455 \cdot 10^{-4} \\ 1{,}671392 \cdot 10^{-4} \end{pmatrix}$$

$$\mathbf{r} = \mathbf{u} - \mathbf{y} = \begin{pmatrix} -5{,}822047 \cdot 10^{-7} \\ 2{,}582090 \cdot 10^{-6} \\ -3{,}915274 \cdot 10^{-6} \end{pmatrix}$$

Con estos vectores se tiene que:

$$\alpha = \mathbf{z}^T\boldsymbol{\delta}\mathbf{f} = 3{,}737755 \cdot 10^{-8}$$

por lo que la matriz tangente actualizada será:

$$\mathbf{A} \leftarrow [\mathbf{A}] + \frac{1}{\alpha}\mathbf{r}\mathbf{z}^T =$$

$$\begin{bmatrix} 1{,}277686 \cdot 10^{-1} & 2{,}208623 \cdot 10^{-1} & -6{,}783630 \cdot 10^{-2} \\ 1{,}515419 \cdot 10^{-1} & -9{,}270604 \cdot 10^{-2} & 1{,}620184 \cdot 10^{-1} \\ 4{,}094919 \cdot 10^{-1} & -2{,}807185 \cdot 10^{-1} & -2{,}126116 \cdot 10^{-1} \end{bmatrix}$$

Con esta matriz se tendrá que el nuevo vector de incrementos será:

$$\mathbf{u} = \mathbf{x}^{(6)} - \mathbf{x}^{(5)} = -[\mathbf{A}]\mathbf{v} = \begin{pmatrix} -5{,}681645 \cdot 10^{-7} \\ 2{,}519821 \cdot 10^{-6} \\ -3{,}820860 \cdot 10^{-6} \end{pmatrix}$$

por lo que,

$$\mathbf{x}^{(6)} = \mathbf{x}^{(5)} + \mathbf{u} = \begin{pmatrix} 14{,}1355467 \\ 10{,}1303043 \\ 43{,}9596516 \end{pmatrix}$$

siendo

$$\mathbf{w} = \mathbf{f}(\mathbf{x}^{(6)}) = \begin{pmatrix} 2{,}998971 \cdot 10^{-9} \\ 3{,}361990 \cdot 10^{-8} \\ 1{,}813023 \cdot 10^{-7} \end{pmatrix}$$

Puesto que $||\mathbf{u}||_2 = 4{,}612076 \cdot 10^{-6}$ es superior a la tolerancia permitida (10^{-6}) a pesar de que ahora $||\mathbf{w}||_2 = 1{,}844175 \cdot 10^{-7}$ es inferior a la tolerancia en el valor de la función vectorial que define el sistema (también 10^{-6} para ambas) se pasará a realizar la **séptima iteración**. En ella:

$$\boldsymbol{\delta}\mathbf{f} = \mathbf{w} - \mathbf{v} = (\mathbf{f}(\mathbf{x}^{(6)}) - \mathbf{f}(\mathbf{x}^{(5)})) = \begin{pmatrix} 5{,}480496 \cdot 10^{-7} \\ 2{,}135544 \cdot 10^{-6} \\ 1{,}642689 \cdot 10^{-5} \end{pmatrix}$$

pasando a actualizar el vector $\mathbf{v}$ mediante:

$$\mathbf{v} = \mathbf{w} = \mathbf{f}(\mathbf{x}^{(6)}) = \begin{pmatrix} 2{,}998971 \cdot 10^{-9} \\ 3{,}361990 \cdot 10^{-8} \\ 1{,}813023 \cdot 10^{-7} \end{pmatrix}$$

Con ello los vectores $\mathbf{z}$, $\mathbf{y}$ y $\mathbf{r}$ del algoritmo de Broyden en esta iteración serán:

$$\mathbf{z} = [\mathbf{A}]^T\mathbf{u} = \begin{pmatrix} -1{,}255348 \cdot 10^{-6} \\ 7{,}134958 \cdot 10^{-7} \\ 1{,}259157 \cdot 10^{-6} \end{pmatrix}$$

$$\mathbf{y} = [\mathbf{A}]\boldsymbol{\delta}\mathbf{f} = \begin{pmatrix} -5{,}726548 \cdot 10^{-7} \\ 2{,}546533 \cdot 10^{-6} \\ -3{,}867612 \cdot 10^{-6} \end{pmatrix}$$

$$\mathbf{r} = \mathbf{u} - \mathbf{y} = \begin{pmatrix} 4{,}490337 \cdot 10^{-9} \\ -2{,}671202 \cdot 10^{-8} \\ 4{,}675664 \cdot 10^{-8} \end{pmatrix}$$

Con estos vectores se tiene que:

$$\alpha = \mathbf{z}^T\boldsymbol{\delta}\mathbf{f} = 2{,}151975 \cdot 10^{-11}$$

por lo que la matriz tangente actualizada será:

$$\mathbf{A} \leftarrow [\mathbf{A}] + \frac{1}{\alpha}\mathbf{r}\mathbf{z}^T =$$

$$\begin{bmatrix} 1{,}275066 \cdot 10^{-1} & 2{,}210112 \cdot 10^{-1} & -6{,}757356 \cdot 10^{-2} \\ 1{,}531002 \cdot 10^{-1} & -9{,}359168 \cdot 10^{-2} & 1{,}604554 \cdot 10^{-1} \\ 4{,}067653 \cdot 10^{-1} & -2{,}791682 \cdot 10^{-1} & -2{,}098756 \cdot 10^{-1} \end{bmatrix}$$

Con esta matriz se tendrá que el nuevo vector de incrementos será:

$$\mathbf{u} = \mathbf{x}^{(7)} - \mathbf{x}^{(6)} = -[\mathbf{A}]\mathbf{v} = \begin{pmatrix} 4{,}438482 \cdot 10^{-9} \\ -2{,}640355 \cdot 10^{-8} \\ 4{,}621670 \cdot 10^{-8} \end{pmatrix}$$

por lo que,

$$\mathbf{x}^{(7)} = \mathbf{x}^{(6)} + \mathbf{u} = \begin{pmatrix} 14{,}1355467 \\ 10{,}1303043 \\ 43{,}9596517 \end{pmatrix}$$

siendo

$$\mathbf{w} = \mathbf{f}(\mathbf{x}^{(7)}) = \begin{pmatrix} 4{,}551615 \cdot 10^{-11} \\ 5{,}687885 \cdot 10^{-10} \\ 2{,}995289 \cdot 10^{-9} \end{pmatrix}$$

Puesto que $||\mathbf{u}||_2 = 5{,}34118910^{-8}$ y $||\mathbf{w}||_2 = 3{,}04915510^{-9}$ son inferiores a la tolerancia dada para ellas se detiene el proceso iterativo dando como solución:

$$Q_1 = 14{,}1355467,\ Q_2 = 10{,}1303043,\ Q = Q_1 + Q_2 = 24{,}265851$$

y:

$$p_2 = 43{,}9596517$$

es decir, los mismos valores que los obtenidos con el método de Newton-Raphson (hallados ahora en 7 iteraciones en lugar de en 4 pero no habiendo tenido que invertir ni calcular la matriz jacobiana en cada iteración). Con todo en este caso en el que el número de ecuaciones es $3(< 10)$ el método de Broyden es más costoso por iteración que el de Newton. Esperamos de la benevolencia del lector que entienda por ello que lo anterior simplemente pretende ilustrar la metodología seguida en el método de Broyden, pues para apreciar el ahorro computacional debería haberse acudido a un sistema con mayor número de ecuaciones bastante más "pesado" de transcribir en el papel.

Observación 2.4.7 *Se observa lo siguiente:*

1. *El análisis detallado de la convergencia del método de Broyden escapa a la disponibilidad temporal de este curso por lo que remitimos al lector a la bibliografía que se cita al final del capítulo (véase por ejemplo De La Fuente O'Connor [7]). En dicha referencia bibliográfica puede encontrar el lector la justificación detallada de que el proceso debido Broyden es el que introduce menores cambios en la matriz* $[\mathbf{A}^{(i-1)}]$ *y que, cuando la matriz jacobiana es lipschitciana, el "deterioro" de la matriz jacobiana al ser sustituida esta por su aproximación es lo suficientemente "lento" como para poder probar la convergencia local del método sobre dominios* D *convexos.*

2. *Para el caso particular de que la matriz jacobiana sea simétrica se puede modificar el método anterior reduciendo aun más su coste computacional. Ello se hace por ejemplo en el método BFGS (debido a Broyden, Fletcher, Goldfab y Shano).*

2.4.3. Algunos comentarios sobre los métodos de resolución de sistemas de ecuaciones no lineales

1^o) El método de Newton-Raphson y sus variantes pueden escribirse en la forma:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + \mathbf{d}^{(i)} \qquad (i = 0, 1, ...)$$

donde

$$\mathbf{d}^{(i)} = -\left[\mathbf{A}^{(i)}\right] \mathbf{f}(\mathbf{x}^{(i)})$$

siendo $\left[\mathbf{A}^{(i)}\right]$ la matriz tangente utilizada en el método (es decir, la jacobiana o una aproximación de ella en el punto $\mathbf{x}^{(i)}$). Ello da pie a considerar familias de métodos de la forma:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + \alpha^{(i)}\mathbf{d}^{(i)} \qquad (i = 0, 1, ...)$$

donde $\mathbf{d}^{(i)}$ es la denominada **dirección de descenso** y $\alpha^{(i)}$ es el **parámetro de descenso**. En general la dirección de descenso $\mathbf{d}^{(i)}$ puede ser una dirección seguida para pasar de un punto $\mathbf{x}^{(i)}$ a otro $\mathbf{x}^{(i+1)}$ (por ejemplo la seguida en el método de Newton-Raphson o en alguna de sus variantes) y el parámetro de descenso $\alpha^{(i)}$ se toma en cada iteración de forma que se minimice el valor de

$$\mathbf{f}(\mathbf{x}^{(i)} + \alpha^{(i)}\mathbf{d}^{(i)})$$

lo que implica que para su determinación deba resolverse el sistema:

$$\frac{d\mathbf{f}}{d\alpha}(\mathbf{x}^{(i)} + \alpha\mathbf{d}^{(i)}) = \mathbf{0}$$

Para ello se han ideado diferentes estrategias. La más simple, debida a Armijo, consiste en, conocidos $\mathbf{x}^{(i)}$ y $\mathbf{d}^{(i)}$, evaluar:

$$\left\|\mathbf{f}(\mathbf{x}^{(i)}+\alpha^{(i)}\mathbf{d}^{(i)})\right\|$$

para los valores de $\alpha = 1,\ \frac{1}{2},\ \frac{1}{4},$ ....hasta determinar un valor de α para el que se satisfaga la relación:

$$\left\|\mathbf{f}(\mathbf{x}^{(i)}+\alpha^{(i)}\mathbf{d}^{(i)})\right\| \leq \left(1-\frac{\alpha}{2}\right)\left\|\mathbf{f}(\mathbf{x}^{(i)})\right\|$$

momento en el que se tomará dicho valor de α como parámetro de descenso. En De La Fuente O'Connor [7], por ejemplo, pueden encontrarse los detalles que justifican esta forma de proceder.

2^o) La estructura que se ha dado a los métodos de descenso, contemplados en el comentario anterior, nos conduce a los denominados **métodos de tipo gradiente** basados en la teoría de optimización. En ellos dado el sistema no lineal $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ se considera la función residuo

$$r(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{n} f_i^2(x_1, x_2, ..., x_n) = \frac{1}{2}\left\|\mathbf{f}(\mathbf{x})\right\|_2^2$$

Esta función residuo siempre tomará valores positivos o nulos. Será precisamente en los puntos en los que se verifique el sistema de ecuaciones en los que el residuo tome el valor nulo. Por ello el problema se reduce a buscar los mínimos de la función residuo $r(\mathbf{x})$. Para ello partiendo de un punto $\mathbf{x}^{(0)}$ tomado arbitrariamente se sigue el esquema iterativo:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + \rho_i \mathbf{d}^{(i)} \qquad (i = 0, 1, ...)$$

La dirección de descenso que ahora se sigue está relacionada con el gradiente de la función residuo (y de ahí el nombre de este tipo de métodos):

$$\nabla r(\mathbf{x}^{(i)}) = \begin{pmatrix} \dfrac{\partial r}{\partial x_1}(\mathbf{x}^{(i)}) \\ \dfrac{\partial r}{\partial x_2}(\mathbf{x}^{(i)}) \\ ..... \\ \dfrac{\partial r}{\partial x_n}(\mathbf{x}^{(i)}) \end{pmatrix}$$

El parámetro de descenso ρ_i en cada iteración se determina minimizando el valor de:

$$r(\mathbf{x}^{(i)} + \rho \mathbf{d}^{(i)})$$

mediante técnicas similares a la de Armijo antes descrita o interpolando (en 3 o 4 valores dados del parámetro de descenso) la función residuo mediante una parábola (algoritmo de Powell) o mediante un polinomio de grado 3 (algoritmo de Davidon).

Uno de los problemas que plantea esta forma de proceder es que los métodos de tipo gradiente son métodos de tipo diferencial, es decir que buscan puntos en los que la diferencial del residuo se anula (o lo que es más preciso, en los que se anula el gradiente del residuo). Ello nos puede conducir a mínimos *locales* del residuo en los que este no toma valor nulo (por lo que no se satisface en ellos el sistema de ecuaciones). Nuevamente, remitimos al lector a la bibliografía sobre el tema (por ejemplo Burden & Faires [3]) para un estudio detallado de estos métodos.

3^o) Dada la relación existente entre la búsqueda de soluciones del sistema y la minimización de funciones (por ejemplo de la función residuo contemplada en el comentario anterior) en las últimas décadas se han desarrollado métodos diferenciales de optimización local (como el de Marquardt & Levenberg que se describe en De La Fuente O'Connor [7]) o métodos de búsqueda directa para optimización global (como los basados en algoritmia genética) que también pueden adaptarse fácilmente a la determinación de soluciones de sistemas de ecuaciones no lineales. El estudio de estos métodos corresponde a la materia denominada como Optimización y desborda los objetivos del presente curso.

4^o) Otra familia de métodos para la resolución de sistemas de ecuaciones no lineales son los métodos de continuación en los que el vector $\mathbf{x}$ que interviene en la definición del sistema $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ se hace depender de un parámetro λ que toma valores entre 0 y 1 y de tal forma que $\mathbf{x}(1)$ sea una solución $\mathbf{x}^*$ del sistema. Más concretamente dado un vector inicial $\mathbf{x}^{(0)}$ se puede considerar la función

$$\mathbf{f}(\mathbf{x}(\lambda)) = (1 - \lambda)\mathbf{f}(\mathbf{x}^{(0)})$$

para la que se verifica que:

$$\mathbf{f}(\mathbf{x}(0)) = \mathbf{f}(\mathbf{x}^{(0)})$$

$$\mathbf{f}(\mathbf{x}(1)) = \mathbf{0}$$

por lo que $\mathbf{x}^* = \mathbf{x}(1)$. La expresión considerada para $\mathbf{f}(\mathbf{x}(\lambda))$ define implícitamente la función $\mathbf{x}(\lambda)$. Asumiendo condiciones suficientes de regularidad puede

derivarse dicha expresión obteniendo:

$$\left\{ \begin{array}{c} [\mathbf{J_f}(\mathbf{x})] \begin{pmatrix} \frac{dx_1}{d\lambda}(\lambda) \\ \frac{dx_2}{d\lambda}(\lambda) \\ ... \\ \frac{dx_n}{d\lambda}(\lambda) \end{pmatrix} = -\mathbf{f}(\mathbf{x}^{(0)}) \\ \mathbf{x}(0) = \mathbf{x}^{(0)} \end{array} \right\} \qquad \lambda \in [0,1]$$

Este sistema diferencial ordinario de primer orden puede entonces ser resuelto (por ejemplo utilizando algún método numérico como los que se estudiarán en el tema siguiente) determinándose una aproximación del valor de $\mathbf{x}^* = \mathbf{x}(1)$.

5º) Determinado un punto solución del sistema de ecuaciones no lineales, otras soluciones se pueden buscar eliminando dichas soluciones del sistema mediante una estrategia de deflacción análoga a la descrita en el sexto comentario sobre los métodos para la resolución de una única ecuación no lineal. Dejamos al lector el desarrollo detallado de esta estrategia de deflacción en el caso n-dimensional.

6º) En Press et al. [12] (y en diferentes "sitios" de internet) pueden encontrarse bibliotecas de códigos que implementan los métodos tratados en este tema y muchos otros. Asimismo, los paquetes informáticos MAPLE, MATLAB, MATHEMATICA, y otros incluyen poderosas rutinas de resolución de sistemas no lineales.

7º) También para el caso de sistemas no lineales puede utilizarse la técnica de aceleración de Aitken (algoritmo conocido como método Δ^2 modificado que básicamente aplica la técnica vista para una ecuación a cada una de las ecuaciones del sistema). No obstante, como se recoge por ejemplo en O.T. Hanna y O.C. Sandall [9], a diferencia de lo que sucedía en el caso de una única ecuación, esta técnica no siempre funciona sobre los sistemas de varias ecuaciones, pudiendo incluso darse el caso de que el método de aproximaciones sucesivas por sí solo converja en tanto que si le combina con la técnica Δ^2 diverja. Remitimos al lector interesado a la referencia citada para obtener un mayor detalle al respecto.

2.4.4. Un programa FORTRAN para la resolución de sistemas no lineales. Aplicación a la resolución del sistema lineal planteado en la motivación de este tema

Al lector que haya seguido estas ciento y muchas páginas sobre los métodos de resolución de sistemas de ecuaciones lineales no le podemos dejar sin la solución del sistema lineal al que nos conducía el ejemplo, tomado de Hanna & Sandal [9] que recogíamos como motivación a este tema. Recordemos que dicho sistema venía dado por las ecuaciones:

$$\frac{x_1(x_1 - x_2 + 2x_4)D^2}{(3 - 3x_1 + x_2 - 5x_4)^3(1 - x_1 - x_2 + 2x_3 - 2x_4)} = 69{,}18$$

$$\frac{(x_2 - x_3)(3 - 3x_1 + x_2 - 5x_4)}{(1 - x_1 - x_2 + 2x_3 - 2x_4)(x_1 - x_2 + 2x_4)} = 4{,}68$$

$$\frac{(1 - x_1 - x_2 + 2x_3 - 2x_4)^2}{(x_2 - x_3)D} = 0{,}0056$$

$$\frac{x_4(x_1 - x_2 + 2x_4)^2D^4}{(3 - 3x_1 + x_2 - 5x_4)^5(1 - x_1 - x_2 + 2x_3 - 2x_4)^2} = 0{,}141$$

donde x_1, x_2, x_3 y x_4 eran las coordenadas de reacción de las 4 reacciones relevantes del equilibrio químico a 500^oC de una mezcla de 1 mol de CO y de 3 moles de H_2, siendo $D = (4 - 2.x_1 + x_3 - 4.x_4)$ y estando relacionadas estas coordenadas de reacción con las fracciones molares de las distintas especies presentes en el equilibrio a través de las expresiones:

$$\begin{aligned}
y_1 &= (3 - 3x_1 + x_2 - 5x_4)/D \\
y_2 &= (1 - x_1 - x_2 + 2x_3 - 2x_4)/D \\
y_3 &= x_1/D \\
y_4 &= (x_1 - x_2 + 2x_4)/D \\
y_5 &= (x_2 - x_3)/D \\
y_6 &= x_4/D
\end{aligned}$$

habiéndose designado por especie 1 al H_2, por especie 2 al CO, por especie 3 al CH_4, por especie 4 al H_2O, por especie 5 al CO_2 y por especie 6 al C_2H_6.

Para evitar "problemas" en el sistema a resolver en el caso de que algunos denominadores puedan tomar valores muy próximos a 0, rescribiremos el sistema en la forma:

$$f_1(\mathbf{x}) = x_1(x_1 - x_2 + 2x_4)D^2 -$$

$$-69{,}18(3 - 3x_1 + x_2 - 5x_4)^3(1 - x_1 - x_2 + 2x_3 - 2x_4) = 0$$

$$f_2(\mathbf{x}) = (x_2 - x_3)(3 - 3x_1 + x_2 - 5x_4) -$$

$$-4{,}68(1 - x_1 - x_2 + 2x_3 - 2x_4)(x_1 - x_2 + 2x_4) = 0$$

$$f_3(\mathbf{x}) = (1 - x_1 - x_2 + 2x_3 - 2x_4)^2 -$$

$$-0{,}0056(x_2 - x_3)D = 0$$

$$f_4(\mathbf{x}) = x_4(x_1 - x_2 + 2x_4)^2 D^4 -$$

$$-0{,}141(3 - 3x_1 + x_2 - 5x_4)^5(1 - x_1 - x_2 + 2x_3 - 2x_4)^2 = 0$$

Este sistema compuesto por las cuatro funciones anteriores puede ser resuelto mediante el método de Newton. Puesto que el estimar la expresión de la matriz jacobiana puede ser laborioso, se ha optado por aproximar la matriz tangente mediante diferencias finitas. Asimismo, puesto que realizar manualmente las operaciones es una tarea muy tediosa se ha optado por realizar un "pequeño" programa en FORTRAN que recoja esta estrategia.

El programa, que se incluye a continuación, refleja fielmente el método de Newton-Raphson (véase la subrutina NEWTON y las que desde ella se llaman) incluyendo también la estimación de las fracciones molares. Fácilmente adaptará el lector este programa al caso general para que pueda servirle en los ejercicios que se le puedan plantear. En todo caso, hemos escrito en letra itálica las sentencias que por ser específicas de este ejemplo, pueden ser eliminadas para adaptarlo al caso general.

```
C*********************************************************************
C                  PROGRAMA SNLNEWT *
C*********************************************************************
C OBJETO:
* C -------
* C     RESOLUCION DE UN SISTEMA DE ECUACIONES NO LINEALES
MEDIANTE EL  * C     METODO DE NEWTON UTILIZANDO LAS EXPRESIONES
ANALITICAS DE LAS   * C     DERIVADAS O UNA APROXIMACION DE ELLAS
EN LOS PUNTOS EN QUE SE   * C     NECESITEN
* C     REFERENCIA: C. Conde y G. Winter. (1.990). "Métodos y
* C     Algoritmos B\'asicos del Algebra Num\'erica''. Ed. Revert\'e.
*
C*********************************************************************
C DICCIONARIO DE VARIABLES:
* C -------------------------
* C  * VARIABLES DE ENTRADA:
* C      EPS    --> PARAMETRO USADO PARA DETENER EL PROCESO
ITERATIVO   * C                 CUANDO ESTE CONVERGE.
* C      EPSF   --> PARAMETRO USADO PARA DETENER EL PROCESO
ITERATIVO   * C                 CUANDO ESTE CONVERGE.
* C      EPSD   --> PARAMETRO DE TOLERANCIA USADO EN LA
```

```
APROXIMACION DE * C                  LAS DERIVADAS PARCIALES QUE
FORMAN LA MATRIZ JACO-  * C                  BIANA. SOLO ES
UTILIZADO CUANDO IOPCION TOMA UN     * C                  VALOR
DIFERENTE A LA UNIDAD.                          * C      FICDAT -->
NOMBRE DEL FICHERO CONTENIENDO LOS DATOS DE ENTRADA.* C FICOUT -->
NOMBRE DEL FICHERO DE RESULTADOS.                    * C IOPCION-->
INDICADOR QUE TOMA EL VALOR:                         * C 1: SI SE
EVALUA LA MATRIZ JACOBIANA DE FORMA    * C EXACTA
* C OTRO VALOR: SI LA MATRIZ JACOBIANA SE EVALUA    * C DE FORMA
APROXIMADA.                          * C      MAXIT  --> NUMERO
MAXIMO DE ITERACIONES QUE SE PERMITE REALIZAR* C      N --> NUMERO
DE ECUACIONES NO LINEALES QUE FORMAN EL      * C SISTEMA.
* C      SF --> FACTOR DE SEGURIDAD UTILIZADO EN LA ESTIMACION DE
LA* C MATRIZ JACOBIANA. SOLO ES USADO SI OPCION TOMA UN   * C
VALOR DIFERENTE A LA UNIDAD.                         * C      X -->
VECTOR CON EL QUE SE INICIALIZA EL METODO.           * C
* C  * VARIABLES DE SALIDA:
* C    --------------------
* C      ITER   --> NUMERO DE ITERACIONES REALIZADAS.
* C      X      --> VECTOR SOLUCION DEL SISTEMA.
* C      TOL    --> NORMA-2 DEL VECTOR DIFERENCIA ENTRE LOS
VECTORES    * C                 HALLADOS EN LAS DOS ULTIMAS
ITERACIONES.            * C
* C      NOTA: Estos resultados se proporcionaran solo en el caso
de    * C            haberse obtenido una solucion suficientemente
precisa    * C            en un numero de iteraciones menor o
igual que MAXIT.     * C            En caso contrario se emitirá
un mensaje de error.        * C
* C  * OTRAS VARIABLES UTILIZADAS:
* C    --------------------------
* C      I      --> VARIABLE DE CONTROL EN LOS BUCLES.
*
C***********************************************************************
C SUBPROGRAMAS UTILIZADOS:
* C ------------------------
* C     SUBROUTINE NEWTON (N, MAXIT, IOPCION, EPS, EPSF, H, X,
ITER, C                       TOL,TOLF,EPSD,SF)
* C     Esta subrutina es en la que se realizan todos los cálculos
del  * C     metodo de Newton.
* C
* C     FUNCTION F(I, X)
* C     Este subprograma permite evaluar la funcion que define el
```

```
siste-* C     ma en el vector X. La variable I indica la
componente de la fun-* C     cion que se evalua. *
C*********************************************************************
C FICHEROS UTILIZADOS:
* C --------------------
* C  * LOS DATOS SON LEIDOS DE UN FICHERO CUYO NOMBRE ES
SOLICITADO AL  * C    USUARIO (Y ALMACENADO EN LA VARIABLE
FICDAT). EN LAS SUCESIVAS   * C    LINEAS DE ESTE FICHERO SE
GRABARAN LOS VALORES DE LAS SIGUIENTES * C    VARIABLES:
* C        Linea 1:   N, MAXIT, IOPCION
* C        Linea 2:   EPS, EPSF, EPSD, SF
* C        Linea 3:   X(1)
* C        Linea 4:   X(2)
* C        Linea 5:   X(3)
* C        .......    ...
* C        Linea N+2: X(N)
* C
* C  * LOS RESULTADOS SE GRABAN EN UN FICHERO CUYO NOMBRE ES
SOLICITADO * C    AL USUARIO (Y ALMACENADO EN AL VARIABLE FICOUT).
EN EL SE GRABAN * C    DEBIDAMENTE IDENTIFICADOS LOS VALORES DE
ITER, TOL Y LAS COMPO-  * C    NENTES DEL VECTOR SOLUCION. *
C*********************************************************************
C  PROGRAMADORES Y FECHA DE PROGRAMACION:
* C  ------------------------------------
* C     CARLOS CONDE LAZARO Y EMANUELE SCHIAVI
* C     ESCUELA SUPERIOR DE CIENCIAS EXPERIMENTALES Y TECNOLOGIA.
* C     UNIVERSIDAD REY JUAN CARLOS.
* C     JULIO DE 1.998
*
C*********************************************************************
C C  Definicion de variables C
      IMPLICIT NONE
      INTEGER*4 I, IOPCION, ITER, MAXIT, N
      REAL*8 EPS, EPSF, TOL, TOLF, X[ALLOCATABLE](:), EPSD, F, SF
      CHARACTER*12 FICDAT, FICOUT
C C  Escritura en pantalla de la cabecera del programa C
      WRITE(*, 100)
      WRITE(*, 101)
      READ(*,*)
C C  Lectura de teclado de los nombres de los ficheros de datos y
C  resultados y apertura del fichero de datos. C
      WRITE(*,*) 'TECLEE EL NOMBRE DEL FICHERO DE DATOS:'
```

```
      READ(*, '(A)') FICDAT
      WRITE(*,*) 'TECLEE EL NOMBRE DEL FICHERO DE RESULTADOS:'
      READ(*,'(A)') FICOUT
      OPEN (1, FILE = FICDAT, STATUS = 'OLD')
C C  Lectura del fichero de los datos del programa y cierre del
fichero C  de datos. C
      READ(1, *) N, MAXIT, IOPCION
      ALLOCATE (X(N))
      IF (IOPCION.EQ.1) THEN
         READ(1, *) EPS, EPSF
      ELSE
         READ(1, *) EPS, EPSF, EPSD, SF
      END IF
      DO I = 1, N
         READ(1,*) X(I)
      END DO
      CLOSE (1)
C C  Apertura del fichero de resultados C
      OPEN (2, FILE = FICOUT, STATUS = 'UNKNOWN')
C C  Llamada a la subrutina de calculo C
      CALL NEWTON (N,MAXIT,IOPCION,EPS,EPSF,X,ITER,TOL,TOLF,EPSD,SF)
C C  Escritura de resultados C

      WRITE(2,105)
      IF ((TOL.LE.EPS).AND.(TOLF.LE.EPSF)) THEN
         WRITE(2, 102) ITER, TOL
         WRITE(*, 102) ITER, TOL
         DO I = 1, N
            WRITE(2, 103) I, X(I), F(I,X)
            WRITE(*, 103) I, X(I), F(I,X)
         END DO
      ELSE
         WRITE(2, 104)
         WRITE(*, 104)
         WRITE(2, 102) ITER, TOL
         WRITE(*, 102) ITER, TOL
         DO I = 1, N
            WRITE(2, 103) I, X(I), F(I,X)
            WRITE(*, 103) I, X(I), F(I,X)
         END DO
      END IF
C C  Cierre del fichero de resultados, formatos de escritura y C
```

```
finalizacion del programa C
      CLOSE (2)
      STOP
  100 FORMAT(20X,'PROGRAMA SNLNEWT'//3X, 'ESTE PROGRAMA RESUELVE UN SIST
     &EMA DE ECUACIONES NO LINEALES'/3X,'MEDIANTE EL METODO DE NEWTON'//
     &3X,'PARA SU EJECUCION ES NECESARIO QUE PROGRAME LA FUNCION VECTORI
     &AL'/3X,'QUE DEFINE EL SISTEMA F(X) = 0 EN EL SUBPROGRAMA FUNCTION
     &F(I,X)'//3X,'ASIMISMO DEBE GRABAR UN FICHERO DE DATOS CON LA ESTRU
     &CTURA QUE'/3X,'SE DESCRIBE EN LA CABECERA DEL LISTADO DE ESTE PROG
     &RAMA'//)
  101 FORMAT(3X,'ESTE PROGRAMA HA SIDO ESCRITO POR:'/6X,'CARLOS CONDE LA
     &ZARO'/6X,'ESCUELA SUPERIOR DE CIENCIAS EXPERIMENTALES Y TECNOLOGIA
     &'/6X,'UNIVERSIDAD REY JUAN CARLOS'/6X,'MOSTOLES (MADRID)'//6X,'ema
     &il: cconde@escet.urjc.es'//10X,'PULSE LA TECLA DE RETURN PARA CONTI
     &NUAR'/)
  102 FORMAT(20X,'METODO DE NEWTON ----SOLUCIONES'/20X,'================
     &================'///5X,'NUMERO DE ITERACIONES REALIZADAS: ',I4,/
     &5X,'PARAMETRO DE TOLERANCIA DE ERROR: 'G20.10//15X,'VECTOR SOLUCIO
     &N'/15X,'---------------'//)
  103 FORMAT(7X,'X(',I3,') = ',G15.6,4X,'(RESIDUO = ',G15.6,')')
  104 FORMAT(5X,'ATENCION --> EN LAS ITERACIONES PERMITIDAS NO SE PUDO'/
     &18X,'HALLAR LA SOLUCION CON LA PRECISION DESEADA')
  105 FORMAT(20X,'PROGRAMA SNLNEWT'//3X, 'ESTE PROGRAMA RESUELVE UN SIST
     &EMA DE ECUACIONES NO LINEALES'/3X,'MEDIANTE EL METODO DE NEWTON'//
     &/3X,'Programadores: CARLOS CONDE LAZARO y EMANUELE SCHIAVI'/16X,
     &'ESCUELA SUPERIOR DE CIENCIAS EXPERIMENTALES Y TECNOLOGIA'/16X,
     &'UNIVERSIDAD REY JUAN CARLOS'/16X,'MOSTOLES (MADRID)'//16X,'email:
     & carlos.conde@upm.es'/16X,'email:emanuele.schiavi@urjc.es'///)
      END
C
C**********************************************************************
C                         MODULO ACTUALIZA
*
C**********************************************************************
C  OBJETO:
* C  -------
* C     RESTAR AL VECTOR X EL VECTOR B.
*
C**********************************************************************
C  DICCIONARIO DE VARIABLES:
* C  -------------------------
* C    * VARIABLES DE ENTRADA:
```

```
* C         B      --> VECTOR QUE SE SUMA AL VECTOR X.
* C         N      --> NUMERO DE COMPONENTES DE LOS VECTORES.
* C         X      --> VECTOR AL QUE SE SUMA B.
* C
* C    * VARIABLES DE SALIDA:
* C         X      --> VECTOR ACTUALIZADO.
* C
* C    * OTRAS VARIABLES UTILIZADAS:
* C         I      --> VARIABLE DE CONTROL EN BUCLES
* C
*
C**********************************************************************
C  SUBPROGRAMAS UTILIZADOS:
* C  ========================
* C     NINGUNO
*
C**********************************************************************
C FICHEROS UTILIZADOS:
* C --------------------
* C     NINGUNO.
*
C**********************************************************************
C  PROGRAMADORES Y FECHA DE PROGRAMACION:
* C  -------------------------------------
* C     CARLOS CONDE LAZARO Y EMANUELE SCHIAVI
* C     ESCUELA SUPERIOR DE CIENCIAS EXPERIMENTALES Y TECNOLOGIA.
* C     UNIVERSIDAD REY JUAN CARLOS.
* C     JULIO DE 1.998
*
C**********************************************************************
C
      SUBROUTINE ACTUALIZA (N, X, B)
C C  Definicion de variables C
      IMPLICIT NONE
      INTEGER*4 I, N
      REAL*8 B(*),X(*)
C C  Actualizacion del vector X C
      DO I = 1, N
         X(I) = X(I) - B(I)
      END DO
C C  Finalizacion C
      RETURN
```

```
      END
C
C*********************************************************************
C                        MODULO EVALF
*
C*********************************************************************
C  OBJETO:
* C  -------
* C      EVALUACION DE LOS VALORES DE UNA FUNCION VECTORIAL.
*
C*********************************************************************
C  DICCIONARIO DE VARIABLES:
* C  ------------------------
* C    * VARIABLES DE ENTRADA:
* C          N   --> NUMERO DE COMPONENTES DE LA FUNCION VECTORIAL
QUE  * C                   SE EVALUA Y DEL VECTOR EN EL QUE SE
EVALUA.         * C          X   --> VECTOR EN EL QUE SE EVALUA LA
FUNCION.              * C    * VARIABLES DE SALIDA:
* C          B   --> VECTOR EN EL QUE SE ALMACENAN LOS VALORES
* C                  CALCULADOS.
* C    * OTRAS VARIABLES UTILIZADAS:
* C          I   --> VARIABLE UTILIZADA EN EL CONTROL DE BUCLES.
*
C*********************************************************************
C  SUBPROGRAMAS UTILIZADOS:
* C  ========================
* C     REAL*8 FUNCTION F(I,X) En este modulo se define la funcion
a    * C            evaluar de tal forma que para un valor
concreto de I,    * C           F(I,X) es la expresion analitica
de la I-esima componen- * C            te de la funcion.
*
C*********************************************************************
C FICHEROS UTILIZADOS:
* C -------------------
* C    NINGUNO.
*
C*********************************************************************
C  PROGRAMADORES Y FECHA DE PROGRAMACION:
* C  -------------------------------------
* C     CARLOS CONDE LAZARO Y EMANUELE SCHIAVI
* C     ESCUELA SUPERIOR DE CIENCIAS EXPERIMENTALES Y TECNOLOGIA.
* C     UNIVERSIDAD REY JUAN CARLOS.
```

```
* C     JULIO DE 1.998
*
C**********************************************************************
C
      SUBROUTINE EVALF (N, X, B)
C C  Definicion de variables C
      IMPLICIT NONE
      INTEGER*4 I, N
      REAL*8 B(*), F, X(*)
C C  Evaluacion del vector de valores de la funcion vectorial C
      DO I = 1, N
         B(I) = F(I,X)
      END DO
C C  Finalizacion C
      RETURN
      END
C
C**********************************************************************
C                     MODULO GAUSS
*
C**********************************************************************
C  OBJETO:
* C  -------
* C       RESOLUCION DE UN SISTEMA LINEAL DE ECUACIONES POR EL
METODO  * C       DE GAUSS CON PIVOTE PARCIAL Y CON UN
REFINAMIENTO DE LA      * C       SOLUCION.
*
C**********************************************************************
C  DICCIONARIO DE VARIABLES:
* C  -------------------------
* C  * VARIABLES DE ENTRADA:
* C        A     --> MATRIZ DEL SISTEMA.
* C        B     --> VECTOR DE SEGUNDOS MIEMBROS.
* C        N     --> NUMERO DE ECUACIONES DEL SISTEMA.
* C
* C  * VARIABLES DE SALIDA:
* C        B     --> VECTOR SOLUCION.
* C
* C  * OTRAS VARIABLES DE INTERES:
* C        AA    --> MATRIZ EN QUE SE GUARDA LA MATRIZ DEL SISTEMA
* C                  ANTES DE TRIANGULARIZARLA.
* C        FILA  --> VECTOR INDICANDO LAS PERMUTACIONES DE FILAS
```

```
REA-  * C                  LIZADAS EN EL PROCESO DE PIVOTAJE
PARCIAL.        * C        RES   --> VECTOR DE RESIDUOS TRAS LA
PRIMERA APROXIMACION   * C                 DE LA SOLUCION.
*
C***********************************************************************
C  FICHEROS UTILIZADOS:
* C  --------------------
* C      Ninguno.
*
C***********************************************************************
C SUBPROGRAMAS A LOS QUE SE LLAMA:
* C -------------------------------
* C      Ninguno.
*
C***********************************************************************
C  NOTAS:
* C  ------
* C    * Mediante el metodo de Gauss con pivotaje parcial se
calcula la * C      primera aproximacion {X'}de la solucion del
sistema:           * C                         [A].{X} = {B}.
* C      Para refinarla se estima el vector residuo
{RES}={B}-[A].{X'}  * C      y se resuelve, aprovechando la
triangularizacion ya hecha, el  * C      sistema [A].{EPS}={RES},
refinandose la solucion mediante:     * C                {X} =
{X'}+{EPS}                                    * C
* C     * La descripcion del metodo de Gauss puede consultarse en
* C        C.CONDE & G. WINTER 'Metodos y Algoritmos Basicos del
Algebra* C        Numerica'. Ed. Reverte (1.990)
*
C***********************************************************************
C  PROGRAMADORES Y FECHA DE PROGRAMACION:
* C  ------------------------------------
* C     CARLOS CONDE LAZARO Y EMANUELE SCHIAVI
* C      Escuela Superior de Ciencias Experimentales y Tecnologia
* C      Universidad Rey Juan Carlos
* C      MOSTOLES (MADRID)
* C      Julio, 1.998
*
C***********************************************************************
C C   Definicion de variables C C
      SUBROUTINE GAUSS (N, A, B)
      IMPLICIT NONE
```

```
      INTEGER*4 I, J, K, N, FILA[ALLOCATABLE](:)
      REAL*8 A(N,N), AUX, B(*), PIV
      REAL*8 AA[ALLOCATABLE](:,:), RES[ALLOCATABLE](:)
      ALLOCATE (AA(N,N),RES(N))
      ALLOCATE (FILA(N))
C C   Salvaguarda de la matriz y vector del sistema C
      DO I = 1, N
         RES(I) = B(I)
         DO J = 1, N
            AA(I,J) = A(I,J)
         END DO
      END DO
C C   Proceso de Gauss C
      DO K = 1, (N-1)
C C       Busqueda de la posicion del pivote parcial C
         PIV = DABS(A(K,K))
         FILA(K) = K
         DO I = K+1, N
            AUX = DABS(A(I,K))
            IF (PIV.LT.AUX) THEN
               FILA(K) = I
               PIV = AUX
            END IF
         END DO
C C        Intercambio de filas C
         IF (FILA(K).NE.K) THEN
            J = FILA(K)
            DO I = K, N
               PIV = A(K,I)
               A(K,I) = A(J,I)
               A(J,I) = PIV
            END DO
            PIV = B(K)
            B(K) = B(J)
            B(J) = PIV
         END IF
C C        Triangularizacion en la columna K C
         DO I = K+1, N
            PIV = A(I,K) / A(K,K)
            DO J = K+1, N
               A(I,J) = A(I,J) - PIV * A(K,J)
            END DO
```

```
            B(I) = B(I) - PIV * B(K)
         END DO
      END DO
C C         Proceso de remonte C
      B(N) = B(N) / A(N,N)
      DO I = N-1, 1, -1
         PIV = 0.D0
         DO J = I+1, N
            PIV = PIV + A(I,J) * B(J)
         END DO
         B(I) = (B(I) - PIV) / A(I,I)
      END DO
C C         Estimacion de residuos C
      DO I = 1, N
         PIV = 0.D0
         DO J = 1, N
            PIV = PIV + AA(I,J)*B(J)
         END DO
         RES(I) = RES(I) - PIV
      END DO
      DO K = 1, N-1
         J = FILA(K)
         IF (J.NE.K) THEN
            PIV = RES(J)
            RES(J) = RES(K)
            RES(K) = PIV
         END IF
      END DO
C C      Correccion de la solucion. C
      RES(N) = RES(N) / A(N,N)
      DO I = N-1, 1, -1
         PIV = 0.D0
         DO J = I+1, N
            PIV = PIV + A(I,J)*RES(J)
         END DO
         RES(I) = (RES(I) - PIV)/A(I,I)
         B(I) = B(I) + RES(I)
      END DO
C C    Finalizacion C
      DEALLOCATE (FILA)
      DEALLOCATE (AA, RES)
      RETURN
```

```
      END
C C
C**********************************************************************
C               MODULO JACOBIANA
*
C**********************************************************************
C  OBJETO:
* C  -------
* C        ESTIMACION DE LA MATRIZ JACOBIANA DE UNA FUNCION F(X).
SE    * C        EVALUARA DE FORMA EXACTA (SI IOPCION = 1) O
UTILIZANDO UNA   * C        APROXIMACION MEDIANTE FORMULAS
PROGRESIVAS CON CONTROL DEL   * C        PASO DE DERIVACION.
*
C**********************************************************************
C  DICCIONARIO DE VARIABLES:
* C  -------------------------
* C  * VARIABLES DE ENTRADA:
* C
* C     EPSD     --> TOLERANCIA EN LA APROXIMACION DE LAS
DERIVADAS QUE * C                  FORMAN LA MATRIZ JACOBIANA.
* C     IOPCION  --> INDICADOR DE LA FORMA DE CALCULAR LA MATRIZ
JACO-  * C                   BIANA DE F(Y). PUEDE TOMAR LOS VALORES
* C                       1 ... LA MATRIZ JACOBIANA SE EVALUA DE
FORMA  * C                                EXACTA A PARTIR DE LAS
EXPRESIONES ANA- * C                                 LITICAS
PROGRAMADAS EN EL MODULO        * C
REAL*8 FUNCTION DF(I,J,X,Y)             * C
OTRO VALOR ... LA MATRIZ JACOBIANA SE EVALUA  * C
APROXIMANDO LAS DERIVADAS MEDIANTE UNA  * C
FORMULA PROGRESIVA CON CONTROL DE PASO  * C     N        -->
NUMERO DE COMPONENTES DE LA FUNCION Y DE VARIABLES * C
INDEPENDIENTES.                                    * C     SF
--> FACTOR DE SEGURIDAD USADO EN EL CONTROL DE LA      * C
LONGITUD DEL PASO DE DERIVACION.                   * C     VAL
--> VECTOR DE N COMPONENTES CONTENIENDO LOS VALORES DE * C
LAS DISTINTAS COMPONENTES DE LA FUNCION CON LA QUE * C
SE TRABAJA EVALUADAS EN EL PUNTO X.                * C     X
--> PUNTO EN EL QUE SE EVALUA LA MATRIZ JACOBIANA.     * C
* C  * VARIABLES DE SALIDA:
* C     JAC      --> MATRIZ JACOBIANA.
* C
* C
```

```
*
C***********************************************************************
C  FICHEROS UTILIZADOS:
* C  --------------------
* C      Ninguno.
*
C***********************************************************************
C SUBPROGRAMAS A LOS QUE SE LLAMA:
* C -------------------------------
* C      REAL*8 FUNCTION F(I,X). Este modulo recoge las
expresiones     * C            analiticas de las funciones que
componen la funcion vec- * C            torial F(X), de manera que
F(I,X) devuelva el valor de   * C            la I-esima componente
de la funcion F(X).                * C      REAL*8 FUNCTION
DF(I,J,X). Este modulo solo es llamado si el   * C
valor de IOPCION es 1. En el se recogeran las expresio-  * C
nes de las derivadas parciales primeras de la funcion    * C
F(X), de tal forma que F(I,J,X) devuelva el valor en X   * C
de la derivada respecto a la componente J-esima de la    * C
variable independiente de la I-esima componente de la    * C
funcion F(X)
C***********************************************************************
C  NOTAS:
* C  ------
* C     * El metodo de calculo aproximado para la derivada de la
compo- * C       nente I-esima respecto a la J-esima variable
consiste en divi-* C       dir [el valor de F en el vector X con
su J-esima componente   * C       incrementada en el valor H menos
el valor  de F en X] entre   * C       el valor que tenga H. Esta
manera de proceder se realiza con  * C       un valor H = PASO y
con otro valor H = PASO2, siendo PASO2    * C       menor que
PASO. Si entre las dos estimaciones asi obtenidas   * C       hay
un ERROR relativo menor o igual que la tolerancia EPSD    * C
se da por buena la estimacion realizada con PASO2. En caso    * C
contrario se reduce el valor de PASO y se repite el proceso.  * C
* C    * Se permite repetir el proceso de calculo anterior hasta
15 ve- * C      ces. En caso de no obtenerse una precision como la
deseada en  * C      estas iteraciones se detiene el programa y se
avisa de ello al * C      usuario. Segun la longitud final del
PASO usado el usuario pue-* C      de modificar el programa
aumentando el numero de iteraciones   * C      permitidas.
*
```

```
C*********************************************************************
C  PROGRAMADORES Y FECHA DE PROGRAMACION:
* C  ----------------------------------
* C     CARLOS CONDE LAZARO Y EMANUELE SCHIAVI
* C      Escuela Superior de Ciencias Experimentales y Tecnologia
* C      Universidad Rey Juan Carlos
* C      MOSTOLES (MADRID)
* C
* C      Julio, 1.998
*
C*********************************************************************
C C   Definicion de variables C
      SUBROUTINE JACOBIANA(N, JAC, X, VAL, EPSD, SF, IOPCION)
      IMPLICIT NONE
      INTEGER*4 I,  IOPCION, ITER, J,  N
      REAL*8 AUX,  DF, EPSD, ERROR, ESTIMP, ESTIMP2, F, JAC(N,N)
      REAL*8 VAL(*)
      REAL*8 PASO, PASO2, REDUC, SF,  X(*)
C C   Estimacion de la matriz Jacobiana de forma exacta. C
      IF (IOPCION.EQ.1) THEN
         DO I = 1, N
            DO J = 1, N
               JAC(I,J) = DF(I,J,X)
            END DO
         END DO
      ELSE
C C   Estimacion de la matriz Jacobiana de forma aproximada. C
         DO J = 1, N
C C        Para la J-esima variable... C
            DO I = 1, N
C C           ... y para la I-esima componente de la funcion, se C
inicializa el valor del PASO y del PASO2 a usar .... C
               PASO = 1.D-3 * X(J)
               ERROR = 2.D0*EPSD
               IF (DABS(X(J)).LE.1.D0) THEN
                  PASO = 1.D-3
               END IF
               REDUC=5.D-1
               PASO2 = REDUC*PASO
C C           ... se estima la derivada con el valor de PASO ....
C
               X(J) = X(J) + PASO
```

```
               ESTIMP = (F(I,X) - VAL(I))/PASO
               X(J) = X(J) - PASO
               ITER = 0
               DO WHILE ((ITER.LT.15).AND.(ERROR.GT.EPSD))
                  ITER = ITER + 1
C C              ... y con el valor PASO2 C
                  X(J) = X(J) + PASO2
                  ESTIMP2 = (F(I,X) - VAL(I)) / PASO2
                  X(J) = X(J) - PASO2
C C              ... se halla el error relativo entre ellas .... C
                  ERROR = (ESTIMP2 - ESTIMP) * REDUC / (1.D0 - REDUC)
C C              ... se corrige la derivada ..... C
                  AUX = ESTIMP2 + ERROR
                  ERROR = DABS(ERROR )/(1.D-3+DABS(AUX))
C C              ... y si el error es mayor que la tolerancia
permitida C                  se disminuye el valor del PASO y del
PASO2 y se C                  repite el proceso .... C
                  IF (ERROR.GT.EPSD) THEN
                     ESTIMP = ESTIMP2
                     PASO = PASO2
                     PASO2 = PASO2*EPSD/(SF*DABS(ERROR))
                     REDUC = PASO2 / PASO
                  ELSE
C C              ... mientras que en caso contrario se da por
bueno C                  el valor esimado C
                     JAC(I,J) = AUX
                  END IF
               END DO
C C     Si en 15 iteraciones no se obtuvo una buena aproximacion C
de la derivada se detiene el program C
               IF ((ITER.EQ.15).AND.(ERROR.GT.EPSD)) THEN
                  WRITE(*,*) 'NO SE PUDO CALCULAR LA MATRIZ JACOBIANA'
                  STOP
               END IF
            END DO
         END DO
      END IF
C C   Finalizacion. C
      RETURN
      END

C
```

```
C*********************************************************************
C                   MODULO NEWTON
*
C*********************************************************************
C OBJETO:
* C -------
* C     RESOLUCION DE UN SISTEMA DE ECUACIONES NO LINEALES
MEDIANTE EL  * C     METODO DE NEWTON UTILIZANDO LAS EXPRESIONES
ANALITICAS DE LAS   * C     DERIVADAS O UNA APROXIMACION DE ELLAS
EN LOS PUNTOS EN QUE SE   * C     NECESITEN Y USANDO LA NORMA-2 EN
EL CRITERIO DE DETENCION DEL   * C     PROCESO ITERATIVO.
* C     REFERENCIA: C. Conde y G. Winter. (1.990). "Métodos y
* C     Algoritmos Básicos del Algebra Numerica". Ed. Reverté.
*
C*********************************************************************
C DICCIONARIO DE VARIABLES:
* C ------------------------
* C  * VARIABLES DE ENTRADA:
* C      EPS    --> PARAMETRO USADO PARA DETENER EL PROCESO
ITERATIVO   * C                 CUANDO ESTE CONVERGE.
* C      EPSD   --> PARAMETRO DE TOLERANCIA USADO EN LA
APROXIMACION DE * C                 LAS DERIVADAS PARCIALES QUE
FORMAN LA MATRIZ JACO-  * C                 BIANA. SOLO ES
UTILIZADO CUANDO IOPCION TOMA UN     * C                 VALOR
DIFERENTE A LA UNIDAD.                          * C      EPSF   -->
PARAMETRO USADO PARA DETENER EL PROCESO ITERATIVO   * C
CUANDO ESTE CONVERGE.                               * C
IOPCION--> INDICADOR QUE TOMA EL VALOR:                          * C
1: SI SE EVALUA LA MATRIZ JACOBIANA DE FORMA    * C
EXACTA                                          * C
OTRO VALOR: SI LA MATRIZ JACOBIANA SE EVALUA    * C
DE FORMA APROXIMADA.                            * C      MAXIT  -->
NUMERO MAXIMO DE ITERACIONES QUE SE PERMITE REALIZAR* C      N
--> NUMERO DE ECUACIONES NO LINEALES QUE FORMAN EL      * C
SISTEMA.                                          * C      SF
--> FACTOR DE SEGURIDAD UTILIZADO EN LA ESTIMACION DE LA* C
MATRIZ JACOBIANA. SOLO ES USADO SI OPCION TOMA UN   * C
VALOR DIFERENTE A LA UNIDAD.                      * C      X
--> VECTOR CON EL QUE SE INICIALIZA EL METODO.          * C
* C  * VARIABLES DE SALIDA:
* C    --------------------
* C      ITER   --> NUMERO DE ITERACIONES REALIZADAS.
```

```
* C      X      --> VECTOR SOLUCION DEL SISTEMA.
* C      TOL    --> NORMA-2 DEL VECTOR DIFERENCIA ENTRE LOS
VECTORES    * C                 HALLADOS EN LAS DOS ULTIMAS
ITERACIONES.            * C      TOL    --> NORMA-2 DEL VECTOR
RESIDUO                          * C
* C
* C  * OTRAS VARIABLES UTILIZADAS:
* C    ---------------------------
* C      B      --> VECTOR AUXILIAR DE N ELEMENTOS.
* C      JAC    --> MATRIZ AUXILIAR DE (N,N) ELEMENTOS QUE
CONTENDRA    * C                 LA MATRIZ JACOBIANA.
*
C***********************************************************************
C SUBPROGRAMAS UTILIZADOS:
* C ------------------------
* C     SUBROUTINE EVALF (N, X, B) Esta subrutina permite evaluar
el    * C          vector de N componentes B = F(X), siendo F la
funcion      * C          vectorial de N componente que define el
sistema no lineal  * C          F(X)=0.
* C     SUBROUTINE JACOBIANA (N, JAC, X, B, EPSD, SF, IOPCION)
* C          Esta subrutina permite evaluar la matriz Jacobiana de
una  * C          funcion vectorial F(X) en un punto X. Segun el
valor de    * C          IOPCION el calculo se realiza de forma
exacta (IOPCION = 1)* C          o aproximando las derivadas
parciales que intervienen      * c          (IOPCION = Cualquier
otro valor).                                    * C
SUBROUTINE GAUSS (N, JAC, B) Esta subrutina permite resolver un *
C          sistema de N ecuaciones lineales cuya matriz del
sistema   * C          sea la matriz JAC y el vector de segundos
miembros se B.   * C          La solucion del sistema se almacena
en el propio vector B. * C     SUBROUTINE NORMA (N, B, TOL) Esta
subrutina evalua la norma-2   * C          de un vector B de N
componentes. El valor de la norma-2 se * C          almacena en la
variable TOL.                          * C     SUBROUTINE
ACTUALIZA (N, X, B) Esta subrutina actualiza el valor* c
de las N componentes del vector X, restandole el vector B. *
C***********************************************************************
C FICHEROS UTILIZADOS:
* C --------------------
* C    NINGUNO.
*
C***********************************************************************
```

```
C  PROGRAMADORES Y FECHA DE PROGRAMACION:
* C  -----------------------------------
* C     CARLOS CONDE LAZARO Y EMANUELE SCHIAVI
* C     ESCUELA SUPERIOR DE CIENCIAS EXPERIMENTALES Y TECNOLOGIA.
* C     UNIVERSIDAD REY JUAN CARLOS.
* C     JULIO DE 1.998
*
C**********************************************************************
C
      SUBROUTINE NEWTON(N,MAXIT,IOPCION,EPS,EPSF,X,ITER,TOL,TOLF,EPSD,
     &SF)
C C  Definicion de variables C
      IMPLICIT NONE
      INTEGER*4 I, IOPCION, ITER, MAXIT, N
      REAL*8 A, B[ALLOCATABLE](:), EPS, EPSD, EPSF
      REAL*8 JAC[ALLOCATABLE](:,:)
      REAL*8 SF, TOL, TOLF, X(*)
C C (La instruccion siguiente es especifica para el ejemplo
resuelto y C debe eliminarse para la realizacion de un programa
general) C

          real*8 Y(6), D, sum

C
      ALLOCATE (JAC(N,N), B(N))
C C  Inicializacion del contador de iteraciones y del valor de TOL
C
      ITER = 0
      TOL = 2.D0*EPS
      TOLF = 2.D0*EPSF
      A = 5.D-1
C C  Mientras no se supere el limite maximo de iteraciones
permitidas C  y no se obtengan soluciones suficientemente precisas
.... C
      DO WHILE ((ITER.LT.MAXIT).AND.((TOL.GT.EPS).AND.(TOLF.GT.EPSF))
C C     ..... se evalua el vector B = F(X) ..... C
         CALL EVALF (N, X, B)
         CALL NORMA (N, B, TOLF)
C C     ..... se calcula la matriz Jacobiana en X .... C
         CALL JACOBIANA (N, JAC, X, B, EPSD, SF, IOPCION)
C C     ..... se calcula el incremento de cada componente del C
vector solucion X resolviendo el sistema lineal C           de
matriz del sistema JAC y de vector de segundo C           termino
```

```
B. El vector incremento se almacena en el C           propio
vector B ..... C
         CALL GAUSS (N, JAC, B)
C C      ..... se evalua la norma-2 del vector de incrementos B
..... C
         CALL NORMA (N, B, TOL)
C C      ..... se actualiza el vector solucion X restandole el
vector C           de incrementos B C           ..... C
         CALL ACTUALIZA (N, X, B)
C C      ..... y se actualiza el valor del contador de iteraciones
ITER .. C
         ITER = ITER + 1
         write(*,*) 'Iteracion n.: ',ITER , '   Tolerancia: ', TOL
         write(2,*) 'Iteracion n.: ',ITER , '   Tolerancia: ', TOL
C C (Las 7 instrucciones siguientes son especificas para el
ejemplo C resuelto y deben eliminarse para la realizacion de un
programa C general) C
```

*D = 4.D0 - 2.D0*X(1) + X(3) - 4.D0*X(4)*
*Y(1) = (3.D0 - 3.D0*X(1) + X(2) - 5.D0*X(4))/D*
*Y(2) = (1.D0 - X(1) - X(2) + 2.D0*X(3) - 2.D0*X(4))/D*
Y(3) = X(1)/D
*Y(4) = (X(1) - X(2) + 2.D0*X(4))/D*
Y(5) = (X(2) - X(3))/D
Y(6) = X(4)/D

```
C
         DO I = 1, N
            WRITE(*,*) 'X(',I,') = ',X(I)
            WRITE(2,*) 'X(',I,') = ',X(I)
         END DO
C C (Las 6 instrucciones siguientes son especificas para el
ejemplo C resuelto y deben eliminarse para la realizacion de un
programa C general) C
```

WRITE(2,) 'Fracciones molares: '*
sum = 0.d0
do I = 1, 6
write(2,) 'Y(',I,') = ',Y(I)*
sum = sum + Y(I)
end do

```
C
```

```
         write(2,*)
C C (La instruccion siguiente es especifica para el ejemplo
resuelto y C debe eliminarse para la realizacion de un programa
general) C
           write(2,*) 'D = ', D,' SUMA = ',sum
C
         write(2,*)
         write(2,*)
C C     ..... para, si procede, poder realizar la siguiente
iteracion. C
      END DO
C C  Finalizacion C
      RETURN
      END
C
C*********************************************************************
C                        MODULO NORMA
*
C*********************************************************************
C  OBJETO:
* C  -------
* C     EVALUACION DE LA NORMA-2 DE UN VECTOR
*
C*********************************************************************
C  DICCIONARIO DE VARIABLES:
* C  ------------------------
* C    * VARIABLES DE ENTRADA:
* C         N      --> NUMERO DE COMPONENTES DEL VECTOR.
* C         V      --> VECTOR DEL QUE SE EVALUA LA NORMA.
* C
* C    * VARIABLES DE SALIDA:
* C         NORMA2 --> VALOR DE LA NORMA-2 DE V.
* C
* C    * OTRAS VARIABLES UTILIZADAS:
* C         I      --> VARIABLE DE CONTROL EN BUCLES
* C
*
C*********************************************************************
C  SUBPROGRAMAS UTILIZADOS:
* C  ========================
* C     NINGUNO
```

```
*
C*********************************************************************
C FICHEROS UTILIZADOS:
* C --------------------
* C     NINGUNO.
*
C*********************************************************************
C  PROGRAMADORES Y FECHA DE PROGRAMACION:
* C  ------------------------------------
* C     CARLOS CONDE LAZARO Y EMANUELE SCHIAVI
* C     ESCUELA SUPERIOR DE CIENCIAS EXPERIMENTALES Y TECNOLOGIA.
* C     UNIVERSIDAD REY JUAN CARLOS.
* C     JULIO DE 1.998
*
C*********************************************************************
C
      SUBROUTINE NORMA (N, V, NORMA2)
C C  Definicion de variables C
      IMPLICIT NONE
      INTEGER*4 I, N
      REAL*8 NORMA2, V(*)
C C  Puesta a cero del valor de NORMA2 C
      NORMA2 = 0.D0
C C  Evaluacion de la suma de los cuadrados de las componentes del
vector C
      DO I = 1, N
         NORMA2 = NORMA2 + V(I)*V(I)
      END DO
C C  Estimacion de la norma C
      NORMA2 = DSQRT(NORMA2)
C C  Finalizacion C
      RETURN
      END
C
C*********************************************************************
C                        MODULO FUNCTION F(I,X)
*
C*********************************************************************
C  OBJETO:
* C  -------
* C      CALCULO DEL VALOR DE LA I-ESIMA COMPONENTE DE UNA FUNCION
* C      VECTORIAL EN EL VECTOR X.
```

```
*
C***********************************************************************
C  DICCIONARIO DE VARIABLES:
* C  -------------------------
* C    * VARIABLES DE ENTRADA:
* C         I      --> INDICADOR DE LA COMPONENTE QUE SE DEFINE.
* C         X      --> VECTOR DE VARIABLES INDEPENDIENTES.
* C
* C    * VARIABLES DE SALIDA:
* C         F      --> VALOR CALCULADO.
* C
*
C***********************************************************************
C  SUBPROGRAMAS UTILIZADOS:
* C  ========================
* C     NINGUNO
*
C***********************************************************************
C FICHEROS UTILIZADOS:
* C --------------------
* C     NINGUNO.
*
C***********************************************************************
C  PROGRAMADORES Y FECHA DE PROGRAMACION:
* C  -------------------------------------
* C     CARLOS CONDE LAZARO Y EMANUELE SCHIAVI
* C     ESCUELA SUPERIOR DE CIENCIAS EXPERIMENTALES Y TECNOLOGIA.
* C     UNIVERSIDAD REY JUAN CARLOS.
* C     JULIO DE 1.998
*
C***********************************************************************
C C NOTA: C ==== C  En este modulo deben programarse las
expresiones de las funciones C  que definen el sisyema no lineal
de ecuaciones. Para cada sistema C  a resolver, el usuario debe
programar aqui sus expresiones tal C  cual se hace a continuacion
con las resultantes del ejemplo tomado C  de O. T. Hanna & O. C.
Shandall ''Computational Methods in Chemical C  Engineering'' Ed.
Prentice Hall (1.995) pags.170-172. C C  Definicion de variables
utilizadas C
      REAL*8 FUNCTION F(I,  X)
      IMPLICIT NONE
      INTEGER*4 I
```

```
      REAL*8 D, X(*), Y(6)
C C  Programacion de las expresiones de las distintas componentes
C  de la funcion vectorial C
      D = 4.D0 - 2.D0*X(1) + X(3) - 4.D0*X(4)
      IF (I.EQ.1) THEN
         Y(1) = (3.D0 - 3.D0*X(1) + X(2) - 5.D0*X(4))
         Y(2) = (1.D0 - X(1) - X(2) + 2.D0*X(3) - 2.D0*X(4))
         Y(3) = X(1)
         Y(4) = (X(1) - X(2) + 2.D0*X(4))
        F = Y(3)*Y(4)*D*D  - 6918.D-2 *(Y(1)**3)*Y(2)
      ELSE IF (I.EQ.2) THEN
         Y(1) = (3.D0 - 3.D0*X(1) + X(2) - 5.D0*X(4))
         Y(2) = (1.D0 - X(1) - X(2) + 2.D0*X(3) - 2.D0*X(4))
         Y(4) = (X(1) - X(2) + 2.D0*X(4))
         Y(5) = (X(2) - X(3))
         F = Y(5)*Y(1)   - 468.D-2*Y(2)*Y(4)
      ELSE IF (I.EQ.3) THEN
         Y(2) = (1.D0 - X(1) - X(2) + 2.D0*X(3) - 2.D0*X(4))
         Y(5) = (X(2) - X(3))
         F = Y(2)*Y(2)  - 0.0056d0 * Y(5)*D
      ELSE IF (I.EQ.4) THEN
         Y(1) = (3.D0 - 3.D0*X(1) + X(2) - 5.D0*X(4))
         Y(2) = (1.D0 - X(1) - X(2) + 2.D0*X(3) - 2.D0*X(4))
         Y(4) = (X(1) - X(2) + 2.D0*X(4))
         Y(6) = X(4)
         F = Y(6)*Y(4)*Y(4)*(D**4) - 0.141d0*(Y(1)**5)*Y(2)*Y(2)
      ELSE
         WRITE(*,*) 'ERROR EN LA DEFINICION DE LA FUNCION'
      END IF
C C   Finalizacion C
      RETURN
      END
C
C*********************************************************************
C                        MODULO FUNCTION DF(I,J,X)
*
C*********************************************************************
C  OBJETO:
* C  -------
* C      CALCULO DEL VALOR DE LA DERIVADA DE LA COMPONENTE I-ESIMA
DE   * C      UNA FUNCION VECTORIAL RESPECTO A LA J-ESIMA VARIABLE
EN EL     * C      VECTOR X.
```

```
*
C**********************************************************************
C  DICCIONARIO DE VARIABLES:
* C  -------------------------
* C    * VARIABLES DE ENTRADA:
* C         I      --> INDICADOR DE LA COMPONENTE QUE SE DERIVA.
* C         J      --> INDICADOR DE LA VARIABLE RESPECTO A LA QUE
SE    * C                    DERIVA
* C         X      --> VECTOR DE VARIABLES INDEPENDIENTES.
* C
* C    * VARIABLES DE SALIDA:
* C         DF     --> VALOR CALCULADO.
* C
*
C**********************************************************************
C  SUBPROGRAMAS UTILIZADOS:
* C  ========================
* C     NINGUNO
*
C**********************************************************************
C FICHEROS UTILIZADOS:
* C --------------------
* C     NINGUNO.
*
C**********************************************************************
C  PROGRAMADORES Y FECHA DE PROGRAMACION:
* C  -----------------------------------
* C     CARLOS CONDE LAZARO Y EMANUELE SCHIAVI
* C     ESCUELA SUPERIOR DE CIENCIAS EXPERIMENTALES Y TECNOLOGIA.
* C     UNIVERSIDAD REY JUAN CARLOS.
* C     JULIO DE 1.998
*
C**********************************************************************
C
      REAL*8 FUNCTION DF(I,J,X)
C C  Definicion de variables C
      IMPLICIT NONE
      INTEGER*4 I,J
      REAL*8 X(*)
C C  Programacion de las expresiones de las distintas derivadas C
de las componentes de la funcion vectorial C C  NOTA: C  ===== C
En el caso del ejemplo al que se va a aplicar, la matriz C
```

```
jacobiana se aproximara por diferencias finitas, por lo que C  no
se utilizara este modulo (aunque es necesaria su exis- C  tencia a
efectos de compilacion no es llamado). Por ello las C  expresiones
de las derivadas que aparecen no se ajustan a las C  las
expresiones de las derivadas de las funciones del sistema C
      IF (I.EQ.1) THEN
         IF (J.EQ.1) THEN
             DF = 1.75D0* 2.35d0*dexp(-3.d0)*((X(1)+X(2))**0.75D0)
         ELSE IF (J.EQ.2) THEN
             DF = 1.75D0* 2.35d0*dexp(-3.d0)*((X(1)+X(2))**0.75D0)
         ELSE
            DF = 1.D0
         END IF
      ELSE IF (I.EQ.2) THEN
         IF (J.EQ.1) THEN
            DF = 1.75D0*4.67d0*dexp(-3.d0)*(X(1)**0.75D0)
         ELSE IF (J.EQ.2) THEN
            DF = 0.D0
         ELSE
            DF = -1.D0
         END IF
      ELSE
         IF (J.EQ.1) THEN
           DF = 0.D0
         ELSE IF (J.EQ.2) THEN
           DF = 3.72D0*DEXP(-2)*1.75D0*(X(2)**0.75)
         ELSE
            DF = -1.D0
         END IF
      END IF
C C   Finalizacion C
      RETURN
      END
C
```

El programa anterior se ha ejecutado a partir de los valores iniciales suministrados por Hanna & Sandall ($x_1 = 0{,}1$, $x_2 = 0{,}2$, $x_3 = 0{,}3$ y $x_4 = 0{,}4$) con los siguientes valores para la tolerancia en los tests de control: EPS = 10^{-6}, EPSF = 10^{-6}, EPSD = 10^{-3}. Junto a ellos se dieron además los valores siguientes MAXIT = 1000 (se permitían hasta 1000 iteraciones en el proceso), IOPCION = 2 (es decir que se aproxime la matriz jacobiana mediante diferencias finitas) y SF = 2 (parámetro usado en el test de convergencia de los valores aproximados de las derivadas parciales). Con estos datos el programa, en **9 iteraciones**, nos conduce a las soluciones:

$$x_1^* = 0{,}681604, \quad x_2^* = 0{,}0158962, \quad x_3^* = -0{,}128703, \quad x_4^* = 0{,}140960 \cdot 10^{-4}$$

que se corresponden con las fracciones molares:

$$y_1 = 3{,}871616 \cdot 10^{-1}$$

$$y_2 = 1{,}796845 \cdot 10^{-2}$$

$$y_3 = 2{,}717684 \cdot 10^{-1}$$

$$y_4 = 2{,}654415 \cdot 10^{-1}$$

$$y_5 = 5{,}765447 \cdot 10^{-2}$$

$$y_6 = 5{,}620353 \cdot 10^{-6}$$

Esta es la solución que proporcionan Hanna y Sandall en [9]. Para los valores de las coordenadas de reacción calculados los residuos en las funciones que definen el sistema fueron:

$$f_1(x_1^*, x_2^*, x_3^*, x_4^*) = -0{,}215626 \cdot 10^{-14}$$

$$f_2(x_1^*, x_2^*, x_3^*, x_4^*) = -0{,}979848 \cdot 10^{-16}$$

$$f_3(x_1^*, x_2^*, x_3^*, x_4^*) = 0{,}245258 \cdot 10^{-17}$$

$$f_4(x_1^*, x_2^*, x_3^*, x_4^*) = -0{,}276799 \cdot 10^{-15}$$

por lo que el test de parada basado en los valores que toma la función es superado ampliamente. Asimismo el valor de la norma-2 de la diferencia entre las dos últimas soluciones halladas fue de $4{,}778881 \cdot 10^{-12}$ por lo que también se satisfizo ampliamente el otro test de parada.

Observación 2.4.8 *1. El sistema dado admite otras soluciones que matemáticamente, son tan válidas como la anterior. Serán otros criterios (de tipo químico en este caso) los que puedan ayudar a elegir entre la mejor de ellas. Así si el programa se inicializa con los mismos parámetros pero*

con los valores iniciales $x_1 = x_2 = x_3 = x_4 = 1$ *se obtienen (al cabo de 158 iteraciones) las soluciones:*

$$x_1^* = 0{,}103513, \quad x_2^* = 2{,}000043, \quad x_3^* = 2{,}000043, \quad x_4^* = 1{,}448265$$

que se corresponden con las fracciones molares (sin sentido químico):

$$y_1 = 59220{,}669601$$

$$y_2 = 2{,}447693 \cdot 10^{-11}$$

$$y_3 = -2402{,}257205$$

$$y_4 = -23207{,}216399$$

$$y_5 = -5{,}153037 \cdot 10^{-11}$$

$$y_6 = -33610{,}195997$$

pero que también satisfacen el sistema de ecuaciones dado pues los residuos resultan ser:

$$f_1(x_1^*, x_2^*, x_3^*, x_4^*) = 0{,}190986 \cdot 10^{-9}$$

$$f_2(x_1^*, x_2^*, x_3^*, x_4^*) = -0{,}730130 \cdot 10^{-15}$$

$$f_3(x_1^*, x_2^*, x_3^*, x_4^*) = 0{,}535803 \cdot 10^{-21}$$

$$f_4(x_1^*, x_2^*, x_3^*, x_4^*) = 0{,}499293 \cdot 10^{-17}$$

2. *El método de Broyden, con los mismos datos que el método de Newton, y aproximando en la primera iteración las derivadas que intervienen en la matriz jacobiana mediante diferencias finitas, también nos conduce a la misma solución que el método de Newton-Raphson pero en este caso son necesarias ...* 263 *iteraciones. Como además el número de ecuaciones no es muy elevado (sólo* 4*) no es aconsejable en este caso este método frente al de Newton.*

2.5. Bibliografía

- Axelsson, O., (1.996). Iterative solution methods. Ed. Cambridge University Press.
- Broyden, C.G., (1.965). *A class of methods for solving nonlinear simultaneous equations*. Mathematics of Computation, 19, págs. 577-593.
- Burden, R.L. y Faires, J.D., (1.998). Análisis numérico. (6^a ed.). Ed. International Thomson editores.
- Ciarlet, P.G. y Lions, J.L. (eds)., (1.990). Handbook of Numerical Analysis. Volume 1: Finite difference methods (part 1); Solution of equations in IR^n (part 1). Ed. North Holland Publishing Company.
- Conde, C. y Winter, G., (1.990) Métodos y algoritmos básicos del álgebra numérica. Ed. Reverté.
- Durand, E., (1.972). Solutions numériques des équations algébriques. Tomes I y II. Ed. Masson et Cie.
- de la Fuente O'Connor, J.L., (1.998). Técnicas de cálculo para sistemas de ecuaciones, programación lineal y programación entera. (2^a ed.). Ed. Reverté.
- Fletcher, R., (1.987). Practical methods of Optimization. Ed. John Wiley and Sons.
- Hanna, O.T. y Sandall, O.C., (1.995). Computational methods in chemical engineering. Ed. Prentice Hall International editions.
- Kincaid, D. y Cheney, W., (1.994). Análisis numérico. Las matemáticas del cálculo científico. Ed. Addison-Wesley Iberoamericana.
- Ortega, J.M. y Rheinboldt, W.C., (1.970). Iterative solution of nonlinear equations in several variables. Ed. Academic Press, Inc.
- Press, W.H., Flannery, B.P., Teukolsky, S.A. y Vetterling, W.T., (1.986). Numerical recipes in FORTRAN. The art of scientific computing. Ed. Cambridge University Press.
- Schiavi, E., Muñoz Montalvo, A.I., Conde, C. (2012). Métodos Matemáticos para los Grados en Ingeniería. Primera parte: teoría. Ed. Dykinson, Textos Docentes 31, Universidad Rey Juan Carlos, ISBN: 978-84-15454-58-8.

- Shampine, L.F., Allen, R.C. y Pruess, S., (1.997). Fundamentals of numerical computing. Ed. John Wiley and Sons.
- Stoer, J. y Bulirsch, R., (1.980). Introduction to numerical analysis. Ed. Springer Verlag.

Capítulo 3

Métodos Numéricos para la resolución de Problemas de Valor Inicial.

3.1. Planteamiento y generalidades.

Sea I un intervalo abierto de la recta real, no reducido a un único punto, y sea t_0 un punto fijado de él. Supongamos además que f es una función definida y continua en $I \times \mathbb{R}$ y con valores en $\mathbb{R}$, y sea y_0 un valor dado de $\mathbb{R}$. En estas condiciones nos planteamos el siguiente problema:

Hallar una función continua y diferenciable $y(t)$ definida en I y con valores en $\mathbb{R}$ verificando:

$$(P.C.) \quad \begin{cases} y'(t) &= f(t, y(t)), \qquad \forall t \in I, \\ y(t_0) &= y_0 \in \mathbb{R}. \end{cases} \tag{3.1}$$

Este tipo de problemas se conoce con el nombre de **problema de Cauchy** para la ecuación diferencial (3.1). La condición asociada se denomina **condición de Cauchy**. Toda función $y(t)$ verificando la EDO del problema de Cauchy se denomina **integral de la EDO** o solución particular de la EDO. En general las soluciones de la EDO estarán dadas por la expresión formal:

$$y(t) = C + \int f(t, y(t))dt$$

donde C es una constante arbitrariamente elegida. A esta familia de soluciones se la denomina **solución general de la EDO**. En algunos casos puede haber además otras soluciones que no pertenezcan a la familia de la solución general. Dichas soluciones se denominarán **soluciones singulares de la EDO**.

A aquella solución particular que, además de satisfacer la EDO del problema de Cauchy, verifique la condición de Cauchy se la denomina **solución del problema de Cauchy.**

Observación 3.1.1 *1^a) Por no ser objeto de este tema el estudio teórico de los problemas de Cauchy no se han considerado diferentes tipos de soluciones vinculadas al estudio de las ecuaciones diferenciales: solución local, maximal, global,...*

2^a) Por el mismo motivo no entraremos a estudiar los diferentes teoremas que aseguran, bajo ciertas hipótesis, la existencia, unicidad o regularidad de las soluciones de los problemas de Cauchy. No obstante recordamos a continuación uno de los teoremas de existencia y unicidad más clásicos, el Teorema de Cauchy-Lipschitz.

Teorema 3.1.1 (de Cauchy-Lipschitz) *Suponiendo que la función $f(t,y)$ es una función continua sobre $I \times \mathbb{R}$ y que es lipschitciana respecto a la segunda de sus variables, es decir:*

$$\exists k > 0 \quad / \quad | f(t,y) - f(t,z) | \leq k||y-z|| \quad \forall t \in I, \ \ \forall y, z \in \mathbb{R}$$

entonces el problema de Cauchy (P.C.) definido en (3.1) admite una única solución.

Observación 3.1.2 *En la bibliografía sobre ecuaciones diferenciales ordinarias (Crouzeix y Mignot[1], Guzmán[2], Marcellán et al.[3], Martínez y Sanz[4],....) puede encontrarse el estudio general de este tipo de problemas con detalle.*

Observación 3.1.3 *La condición de que $f(t,y(t))$ sea lipschitciana toma su nombre del matemático alemán Rudolph Lipschitz que fue el primero en utilizar dicha condición en una publicación del año 1876. Es una condición relativamente "suave" sobre la función. A fin de cuentas, si la derivada parcial de $f(x,y)$ respecto a su segunda variable fuese una función continua, la condición de Lipschitz se traduciría en que dicha derivada permaneciera acotada.*

En numerosos problemas físicos la variable t representa el tiempo. Además el intervalo I se suele considerar de la forma $[t_0, t_0+T]$ y por ello al instante t_0

[1] Crouzeix, M., Mignot A.L. (1984) "Analyse numérique des équations différentielles", Ed. Masson.

[2] Guzmán, M. de (1987). "Ecuaciones diferenciales ordinarias. Teoría de estabilidad y control" (3^a reimpresión). Ed. Alhambra Universidad.

[3] Marcellán, F., Casasús, L. y Zarzo, A. (1991). "Ecuaciones diferenciales. Problemas lineales y aplicaciones". Ed. McGraw Hill.

[4] Martínez, C. y Sanz, M.A. (1991). "Introducción a las ecuaciones diferenciales ordinarias". Ed. Reverté.

se le llama entonces **instante inicial** y a la condición impuesta en este instante se la denomina **condición inicial**. En esta situación a los problemas de Cauchy correspondientes se les denomina habitualmente **problemas de valor inicial** (P.V.I.), extendiéndose esta terminología también a los problemas en los que t no se identifique con la variable física "tiempo".

Los problemas de Cauchy que no sean problemas de valor inicial por no ser t_0 el extremo izquierdo del intervalo I pueden reducirse a problemas de valor inicial mediante sencillos cambios de variable. En efecto si $I = [a, b]$ y t_0 es un punto intermedio de I el problema de Cauchy correspondiente puede descomponerse en dos problemas de la forma:

$$(P.C,1) \quad \begin{cases} y'(t) &= f(t, y(t)), \quad \forall t \in [a, t_0], \\ y(t_0) &= y_0 \in I\!R, \end{cases}$$

$$(P.C,2) \quad \begin{cases} y'(t) &= f(t, y(t)), \quad \forall t \in [t_0, b], \\ y(t_0) &= y_0 \in I\!R. \end{cases}$$

El problema $(P.C,2)$ es un problema de valor inicial. Y el problema $(P.C,1)$ puede transformarse en un problema de valor inicial si se realiza el cambio de variable independiente: $t' = -t$.

Por este motivo los métodos numéricos que se estudiarán en este tema se plantearán sobre problemas de valor inicial. Para ellos, además, teniendo en cuenta la expresión de la solución general de la EDO, se tendrá que la expresión formal de la solución del problema de Cauchy está dada por:

$$y(t) = y_0 + \int_{t_0}^{t} f(x, y(x))dx.$$

Los P.V.I. pueden formularse de una forma más general para **sistemas** de ecuaciones diferenciales de primer orden. Más concretamente, siendo I un intervalo de $I\!R$ de la forma $[t_0, t_0 + T]$ y siendo $\mathbf{f}$ una función continua definida en $I \times I\!R^m$ y con valores en $I\!R^m$ puede considerarse el problema siguiente:

Hallar una función continua y diferenciable $\mathbf{y}(t)$ *definida en* I *y con valores en* $I\!R^m$ *verificando:*

$$(P.V.I.) \quad \begin{cases} \mathbf{y}'(t) &= \mathbf{f}(t, \mathbf{y}(t)), \quad \forall t \in I, \\ \mathbf{y}(t_0) &= \mathbf{y}^0 \in I\!R^m. \end{cases}$$

La solución de este tipo de problemas, formalmente, estará dada por el vector solución:

$$\mathbf{y}(t) = \mathbf{y}^0 + \int_{t_0}^{t} \mathbf{f}(x, \mathbf{y}(x))dx$$

Por simplicidad nosotros desarrollaremos los métodos numéricos que configuren este tema en el caso de problemas de valor inicial en los que interviene una única EDO. La extensión de los métodos numéricos a problemas en los que intervenga un sistema de ecuaciones diferenciales ordinarias es automática y, en cuanto a su análisis, serán válidos los teoremas y propiedades que se presenten con la precaución de emplear la métrica (normas) correspondiente en $\mathbb{R}^m$.

Cabe señalar además que numerosos problemas de valor inicial se formulan mediante ecuaciones diferenciales ordinarias de orden superior a 1 de la forma:

$$\begin{cases} y^{(p}(t) & = & f(t, y(t), y'(t), y''(t), ..., y^{(p-1}(t)), \qquad t \in [t_0, t_0 + T] \\ y(t_0) & = & y_0 \\ y'(t_0) & = & y_0^{(1)} \\ y''(t_0) & = & y_0^{(2)} \\ .... & ... & .... \\ y^{(p-1}(t_0) & = & y_0^{(p-1)} \end{cases}$$

Tales tipos de P.V.I de orden superior a 1 pueden ser reducidos a sistemas de p ecuaciones diferenciales de primer orden denominando:

$$z_1(t) = y(t), \;\; z_2(t) = y'(t), \;\; z_3(t) = y''(t), \;\; ...., \;\; z_p(t) = y^{(p-1}(t)$$

con lo que el problema anterior se reescribe como:

$$(P.V.I.) \quad \begin{cases} \mathbf{z}'(t) & = & \mathbf{f}(t, \mathbf{z}(t)), \quad \forall t \in I, \\ \mathbf{z}(t_0) & = & \mathbf{z}^0 \in \mathbb{R}^m \end{cases}$$

siendo,

$$\mathbf{z}(t) = (z_1(t), \; z_2(t), \; ...., \; z_p(t))^T$$

$$\mathbf{z}^{(0)} = (z_1(t_0), \; z_2(t_0), \; ...., \; z_p(t_0))^T = \Big(y(t_0), \; y'(t_0), \; ...., \; y^{(p-1}(t_0)\Big)^T$$

y

$$\mathbf{f}(t, \mathbf{z}(t)) = \begin{pmatrix} z_2(t) \\ z_3(t) \\ .... \\ z_p(t) \\ f(t, z_1(t), z_2(t), ..., z_p(t)) \end{pmatrix}$$

Por este motivo los métodos numéricos se plantean habitualmente sobre problemas de valor inicial regidos por ecuaciones diferenciales ordinarias de primer orden.

Ejemplo 3.1.1 *En el estudio de algunas capas límite de flujos laminares aparece la denominada ecuación de Blasius que es una EDO de tercer orden de la forma:*

$$y'''(x) + y(x)y''(x) = 0, \qquad x \geq 0.$$

Nótese que aquí hemos modificado la notación de la variable independiente llamándola x pues en esta ocasión la variable independiente no es el tiempo sino que es una variable adimensional relacionada con la coordenada espacial. La ecuación de Blasius se acompaña de las tres condiciones iniciales $y(0) = y'(0) = 0$, $y''(0) = \alpha$ donde α es un valor conocido ($\alpha = 0{,}47$). Con ello se puede plantear el problema de valor inicial de Blasius de la forma siguiente:

Hallar una función $y(x)$ verificando:

$$\begin{cases} y'''(x) + y(x)y''(x) = 0, & x \geq 0 \\ y(0) = 0, \;\; y'(0) = 0, \;\; y''(0) = \alpha \end{cases}$$

Este problema es equivalente al P.V.I. de primer orden:

$$\begin{cases} z_1'(x) = z_2(x), & x \geq 0 \\ z_2'(x) = z_3(x), & x \geq 0 \\ z_3'(x) = -z_1(x).z_3(x), & x \geq 0 \\ z_1(0) = 0, \quad z_2(0) = 0, \quad z_3(0) = \alpha \end{cases}$$

donde se ha denotado por $z_1(x) = y(x)$, $z_2(x) = y'(x)$, $z_3(x) = y''(x)$.

Abreviadamente el problema anterior puede escribirse entonces como:

$$\begin{cases} \mathbf{z}(x) = \mathbf{f}(x, \mathbf{z}(x)), & x \geq 0 \\ \mathbf{z}(0) = (0, 0, \alpha)^T \end{cases}$$

donde

$$\mathbf{z}(x) = (z_1(x), z_2(x), z_3(x))^T$$

y

$$\mathbf{f}(x, \mathbf{z}(x)) = (z_2(x), z_3(x), -z_1(x)z_3(x))^T$$

Observación 3.1.4 *Existe otra gran familia de problemas regidos por ecuaciones diferenciales ordinarias que es la formada por los problemas de contorno unidimensionales. Este tipo de problemas, en el caso de estar regidos por una EDO de segundo orden puede formularse de la forma siguiente:*

Hallar una función $y(t)$ verificando:

$$\begin{cases} y''(t) = f(t, y(t), y'(t)), & t \in [a, b], \\ c_{1,1}y(a) + c_{1,2}y'(a) = \alpha, \\ c_{2,1}y(b) + c_{2,2}y'(b) = \beta \end{cases} \tag{3.2}$$

donde a, b, $c_{1,1}, c_{1,2}, c_{2,1}, c_{2,2}$ y α, β son constantes conocidas y $f(t, y(t), y'(t))$ es una función dada.

Existen métodos específicos para abordar este tipo de problemas (como por ejemplo los "métodos de tiro"), aunque aquí no los abordaremos. Su tratamiento numérico, no obstante, puede realizarse mediante métodos en diferencias finitas (véase el siguiente capítulo). Un estudio detallado de métodos numéricos para problemas de contorno regidos por EDO puede encontrarse en algunas de las referencias que se citan en la bibliografía de este tema. Nótese finalmente que desde el punto de vista físico los problemas de contorno del II orden del tipo (3.2) representan, en fluido-dinámica y en fenómenos de transporte procesos de difusión, convección y reacción.

3.1.1. Tipos de métodos de resolución de problemas de valor inicial

Existen numerosos tipos de métodos para la resolución de problemas de valor inicial. Sin pretender ser exhaustivos en la descripción de ellos pueden citarse los siguientes:

A) Métodos gráficos.

Se basan, fundamentalmente, en buscar el significado geométrico de la EDO a resolver construyendo lo que habitualmente se conoce como el campo de direcciones, esto es, los valores de la pendiente de la solución (los valores de $f(t, y)$) en un dominio más o menos amplio del espacio $\mathbb{R}^2$. Con ello, partiendo de (t_0, y_0) se reconstruye gráficamente la solución buscada. El más popular de estos métodos es el método conocido como **método de las isoclinas** (del cual puede encontrarse una descripción, por ejemplo, en el libro del Pr. Guzmán[5]). Este tipo de métodos, que fueron relativamente populares hasta los comienzos

[5]Guzmán, M. de (1987). "Ecuaciones diferenciales ordinarias. Teoría de estabilidad y control" (3^a reimpresión). Ed. Alhambra Universidad.

del siglo XX, está ya prácticamente en desuso pues, aparte de ser bastante poco precisos, no son aplicables a problemas en los que aparecen sistemas de ecuaciones diferenciales con más de 2 ecuaciones. No obstante, sus fundamentos geométricos están en los orígenes de muchos de los métodos numéricos empleados hoy en día.

B) Métodos analíticos.

Son los métodos clásicos de resolución de ecuaciones diferenciales ordinarias (factor integrante, separación de variables, cambios de variables, reducción a homogéneas, método de la mayorante, etc...). En numerosas referencias (Guzmán[6], Marcellán et al.[7], Martínez y Sanz[8], etc...) pueden encontrarse magníficas descripciones y análisis de este tipo de métodos ... y de sus límites de aplicabilidad. Su principal inconveniente es que "cada ecuación diferencial ordinaria tiene su método analítico de resolución" y su aplicación puede ser complicada (determinación de un cambio de variable "afortunado", conocimiento previo de una solución particular de la EDO, búsqueda de un factor integrante, etc...). Ello por no citar que existen ecuaciones diferenciales ordinarias de las que no se conocen métodos analíticos que puedan resolverlas. No obstante, desde el punto de vista numérico este tipo de métodos analíticos son de gran utilidad para, sobre problemas "test", disponer de soluciones exactas con las que contrastar el comportamiento de los métodos numéricos.

C) Métodos basados en desarrollos de Taylor

En síntesis, este tipo de métodos consisten en expresar el valor de la función solución en un punto t a través del desarrollo en serie de Taylor de la función solución $y(t)$ en torno a t_0. Más concretamente:

$$y(t) = y(t_0) + (t-t_0)y'(t_0) + \frac{(t-t_0)^2}{2}y''(t_0) + ... + \frac{(t-t_0)^m}{m!}y^{(m}(t_0) + ...$$

estimándose los valores de la función $y(t)$ y de sus derivadas en t_0 a partir de la condición inicial y de la propia función $f(t, y(t))$. Así se tiene que:

$$y(t_0) = y_0$$

luego considerando la EDO y evaluando en $t = t_0$ se deduce

$$y'(t) = f(t, y(t)) \quad \Rightarrow \quad y'(t_0) = f(t_0, y_0).$$

[6]Guzmán, M. de (1987). "Ecuaciones diferenciales ordinarias. Teoría de estabilidad y control" (3^a reimpresión). Ed. Alhambra Universidad.

[7]Marcellán, F., Casasús, L. y Zarzo, A. (1991). "Ecuaciones diferenciales. Problemas lineales y aplicaciones". Ed. McGraw Hill.

[8]Martínez, C. y Sanz, M.A. (1991). "Introducción a las ecuaciones diferenciales ordinarias". Ed. Reverté.

Si derivamos la EDO, aplicamos la regla de la cadena y evaluamos en $t = t_0$ se tiene además

$$y''(t) = \frac{df}{dt}(t, y(t)) = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t))\frac{dy}{dt}(t) \quad \Rightarrow$$

$$\Rightarrow \quad y''(t_0) = \frac{\partial f}{\partial t}(t_0, y_0) + f(t_0, y_0)\frac{\partial f}{\partial y}(t_0, y_0)$$

y así sucesivamente.

Los problemas principales de estos métodos se pueden resumir en, por una parte, la exigencia de regularidad elevada a la función solución $y(t)$ para que pueda ser desarrollada en serie de Taylor hasta el grado que se desee y, por otra, en el esfuerzo de cálculo necesario que va incrementándose paulatinamente para aproximar las derivadas de orden superior que intervienen en el desarrollo. No obstante, en los fundamentos de este tipo de métodos se basan muchos estudios sobre el análisis del "error" de los métodos numéricos que abordaremos más adelante.

Ejemplo 3.1.2 *Al problema de valor inicial de Blasius que citábamos en un ejemplo anterior:*

$$\begin{cases} y'''(x) + y(x)y''(x) = 0, & x \geq 0 \\ y(0) = 0, \quad y'(0) = 0, \quad y''(0) = 0{,}47 \end{cases}$$

bajo hipótesis de regularidad de la solución, puede aplicársele el método de los desarrollos en serie de Taylor.

En efecto, al aplicar el método de los desarrollos en serie de Taylor resulta que

$$\begin{aligned} y(t) \quad = \quad & y(0) + ty'(0) + \frac{t^2}{2}y''(0) + \frac{t^3}{3!}y'''(0) + \frac{t^4}{4!}y^{(iv}(0) + \\ & + \frac{t^5}{5!}y^{(v}(0) + \frac{t^6}{6!}y^{(vi}(0) + \frac{t^7}{7!}y^{(vii}(0) + \frac{t^8}{8!}y^{(viii}(0) + ... \end{aligned}$$

donde

$$y(0) = 0, \quad y'(0) = 0, \quad y''(0) = 0{,}47$$

y

$$y'''(t) = -y(t)\, y''(t) \quad \Rightarrow \quad y'''(0) = -y(0)\, y''(0) = (0)\, 0{,}47 = 0$$

$$\begin{aligned} y^{(iv}(t) \quad = \quad & -(y'(t)\, y''(t) + y(t)\, y'''(t)) \quad \Rightarrow \\ \Rightarrow \quad & y^{(iv}(0) = -(y'(0)\, y''(0) + y(0)\, y'''(0)) = 0 \end{aligned}$$

$$y^{(v}(t) \quad = \quad -\left(\left(y''(t)\right)^2 + 2\,y'(t)\,y'''(t) + y(t)\,y^{(iv}(t)\right) \quad \Rightarrow$$

$$\Rightarrow \quad y^{(v}(0) = -\left(\left(y''(0)\right)^2 + 2\,y'(0)\,y'''(0) + y(0)\,y^{(iv}(0)\right) = -(0{,}47)^2$$

$$y^{(vi}(0) = 0, \qquad y^{(vii}(0) = 0$$

$$y^{(viii}(0) = 11\ \left(y''(0)\right)^3 = 11\,(0{,}47)^3$$

La introducción de estos valores en el desarrollo de Taylor de la solución nos permite escribir esta como:

$$y(t) = 0{,}235\,t^2 - 1{,}84083\,10^{-3}\,t^5 + 2{,}8325\,10^{-5}\,t^8 + ...$$

D) Método de Picard (o de aproximaciones sucesivas)

El método de Picard toma su nombre del matemático francés Emile Picard (1856 a 1941) quien utilizó la técnica de aproximaciones sucesivas para estudiar la existencia y unicidad de soluciones de problemas de valor inicial. En síntesis el método consiste en generar, a partir del valor y_0 con el que se define la condición inicial y habida cuenta de la expresión formal de la solución del problema, la sucesión de funciones:

$$\begin{array}{ll} y_0(t) & = y_0 \\ y_1(t) & = y_0 + \displaystyle\int_{t_0}^{t} f(x, y_0(x))dx \\ .... & ..... \\ y_{n+1}(t) & = y_0 + \displaystyle\int_{t_0}^{t} f(x, y_n(x))dx \\ .... & ..... \end{array}$$

Bajo ciertas hipótesis de regularidad sobre la función $f(t,y)$, como las recogidas en el teorema siguiente, puede afirmarse que este método converge hacia la solución del problema de Cauchy:

Teorema 3.1.2 *Siendo (t^*, y^*) un punto del plano $\mathbb{R}^2$ y suponiendo que en un dominio $D = \{(t,y)/\ |t-t^*| \leq \varepsilon,\ \ |y-y^*| < \delta\}$ la función $f(t,y)$ es continua, está acotada y verifica la condición de Lipschitz respecto a su segunda variable, y denotando por M a un valor tal que*

$$|f(x,y)| \leq M, \qquad \forall\,(x,y) \in D$$

por L al valor:

$$L = \text{ínf}\left(\varepsilon, \frac{\delta}{M}\right)$$

y por I al intervalo $I = (t^* - L, t^* + L)$, entonces existe una única función $y(t)$ definida en I que sea solución del problema de Cauchy:

$$\begin{cases} y'(t) &= f(t, y(t)), \qquad \forall t \in I \\ y(t^*) &= y^* \end{cases}$$

Además dicha función puede obtenerse como límite de la sucesión de funciones generada mediante:

$$\begin{cases} y_0(t) &= y^* \\ .... & .... \\ y_{n+1}(t) &= y^* + \displaystyle\int_{t^*}^{t} f(x, y_n(x))dx \\ .... & .... \end{cases}$$

Demostración: Consúltese, por ejemplo, T.M. Apostol[9].

Los principales inconvenientes para la aplicación práctica de este método son que el cálculo de las integrales $\displaystyle\int_{t_0}^{t} f(x, y_n(x))dx$ suele volverse cada vez más complicado, debiendo estimarse en ocasiones dichas integrales, para valores concretos de t, mediante aproximaciones numéricas ... con lo que se acaba obteniendo aproximaciones de los valores de la solución en ciertos puntos t_i en lugar de la expresión de la solución. Y ello, como veremos, puede realizarse de forma más eficiente con los métodos numéricos basados en diferencias finitas.

A favor del método de Picard puede señalarse que es una herramienta de enorme utilidad en el estudio teórico de la existencia y unicidad de la solución de los problemas de Cauchy. Además algunos métodos que estudiaremos pueden interpretarse como variantes del método de Picard.

Para poner de manifiesto tanto la forma operativa del método como sus inconvenientes prácticos, ilustremos el método de Picard con algunos ejemplos extraídos de T.M. Apostol[9] tal cual aparecen en dicha referencia o con ligeras modificaciones de los mismos:

[9] Apostol, T. M. (1997) "Linear Algebra. A first course with applications to differential equations". Ed. John Wiley & Sons, Inc.

Ejemplo 3.1.3 *Determínese mediante el método de Picard la solución del P.V.I.:*

$$\begin{cases} y'(t) &= t^2 + y^2, \qquad t \geq 0 \\ y(0) &= 0 \end{cases}$$

Generemos la sucesión del método:

$$\begin{aligned}
y_0(t) &= y(0) && = 0 \\
y_1(t) &= 0 + \int_0^t \left(x^2 + 0^2\right) dx && = \frac{t^3}{3} \\
y_2(t) &= 0 + \int_0^t \left(x^2 + \left(x^3/3\right)^2\right) dx && = \frac{t^3}{3} + \frac{t^7}{63} \\
y_3(t) &= 0 + \int_0^t \left(x^2 + (x^3/3 + x^7/63)^2\right) dx && = \frac{t^3}{3} + \frac{t^7}{63} + \frac{2t^{11}}{2079} + \frac{t^{15}}{59535}
\end{aligned}$$

Como puede apreciarse a medida que aumenta el valor de n también aumenta la complejidad en la evaluación de la correspondiente integral. Se deja como ejercicio al lector comprobar que $y_4(t)$ es un polinomio de grado 31 y que $y_5(t)$ resultará un polinomio de grado 63. Y si bien es cierto que, al crecer el valor de los denominadores de los coeficientes de dicho polinomio, con "pocos" términos de él se podría obtener una aproximación razonablemente buena de la solución para valores de t inferiores a 1, no es menos cierto que para valores mayores de t la aproximación de la función exige calcular las expresiones de $y_n(t)$ para valores elevados de n pues la variable t aparece elevada a exponentes cada vez mayores. Por ejemplo, para $t = 3$, el término en t^{11} toma el valor $\frac{354294}{2079} = 170{,}4155844$ que no es depreciable (y para la potencia 15 el valor es aún mayor: 241,0163265).

Ejemplo 3.1.4 *Determínese mediante el método de Picard la solución del P.V.I.:*

$$\begin{cases} y'(t) &= 2t + e^y \qquad t \geq 0 \\ y(0) &= 0 \end{cases}$$

Generemos la sucesión del método:

$$\begin{aligned}
y_0(t) &= y(0) && = 0 \\
y_1(t) &= 0 + \int_0^t (2x + e^0) dx && = t^2 + t \\
y_2(t) &= 0 + \int_0^t (2x + e^{x^2+x}) dx && = t^2 + \int_0^t e^{x^2+x} dx
\end{aligned}$$

La última integral que aparece no puede evaluarse en términos de funciones elementales. En efecto:

$$\int_0^t e^{x^2+x}dx = -\frac{i}{2}\,\mathrm{erf}(i\,t+\frac{i}{2})\,\sqrt{\pi}\,e^{-1/4} + \frac{i}{2}\,\mathrm{erf}(\frac{i}{2})\,\sqrt{\pi}\,e^{-1/4}$$

No obstante, para un valor prefijado $t = t^*$ la integral puede aproximarse mediante alguna fórmula de integración numérica (ver C. Conde y E. Schiavi[10]).

Ejemplo 3.1.5 *Determínese mediante el método de Picard la solución del P.V.I.:*

$$\begin{cases} y'(t) &= 2\,e^t - y \qquad t \geq 0 \\ y(0) &= 1 \end{cases}$$

Generemos la sucesión del método:

$$\begin{aligned}
y_0(t) &= y(0) && = 1 \\
y_1(t) &= 1 + \int_0^t (2e^x - 1)\,dx && = 2\,e^t - t - 1 \\
y_2(t) &= 1 + \int_0^t (2\,e^x - 2\,e^x + x + 1)\,dx && = \tfrac{t^2}{2} + t + 1 \\
y_3(t) &= 1 + \int_0^t (2\,e^x - \frac{x^2}{2} - x - 1)\,dx && = 2\,e^t - \tfrac{t^3}{3!} - \tfrac{t^2}{2} - t - 1 \\
y_4(t) &= 1 + \int_0^t \left(2\,e^x - 2\,e^x + \frac{x^2}{2} + x + 1\right)\,dx && = \frac{t^4}{4!} + \frac{t^3}{3!} + \frac{t^2}{2} + t + 1
\end{aligned}$$

En general se verifica:

$$y_{2\,n}(t) = \sum_{i=0}^{2\,n} \frac{t^i}{i!} \qquad \Rightarrow \qquad \lim_{n\to\infty} y_{2\,n}(t) = \lim_{n\to\infty} \sum_{i=0}^{2\,n} \frac{t^i}{i!} = e^t$$

$$y_{2\,n+1}(t) = 2\,e^t - \sum_{i=0}^{2\,n+1} \frac{t^i}{i!} \quad \Rightarrow \quad \lim_{n\to\infty} y_{2\,n+1}(t) = 2e^t - \lim_{n\to\infty} \sum_{i=0}^{2\,n+1} \frac{t^i}{i!} = e^t$$

por lo que

$$y(t) = \lim_{n\to\infty} y_n(t) = e^t.$$

[10]Conde, C. y Schiavi, E. (2000) "Elementos de Matemáticas: Guiones de los temas de la asignatura". Apuntes. Universidad Rey Juan Carlos.

Ejemplo 3.1.6 *Determínese mediante el método de Picard la solución del P.V.I.:*

$$\begin{cases} y'(t) &= \sup\{t, y(t)\}, \qquad -1 \leq t \leq 1 \\ y(0) &= 1 \end{cases}$$

El problema de Cauchy anterior puede descomponerse en los dos problemas siguientes:

$$(P-1)\begin{cases} y'(t) &= \sup\{t, y(t)\} \qquad -1 \leq t \leq 0 \\ y(0) &= 1 \end{cases}$$

$$(P-2)\begin{cases} y'(t) &= \sup\{t, y(t)\} \qquad 0 \leq t \leq 1 \\ y(0) &= 1 \end{cases}$$

Realizando el cambio de variable $t = -\widetilde{t}$, con lo que $y(t) = y(-\widetilde{t})$ por lo que se verifica que $y'(t) = -y'(\widetilde{t})$, el problema $(P-1)$ puede formularse como:

$$(\widetilde{P}-1) \quad \begin{cases} y'(\widetilde{t}) &= -\sup\left\{-\widetilde{t}, y(-\widetilde{t})\right\} \qquad 0 \leq \widetilde{t} \leq 1 \\ y(0) &= 1 \end{cases}$$

Generemos la sucesión del método de Picard para $(\widetilde{P}-1)$:

$$\begin{aligned} y_0(\widetilde{t}) &= y(0) && = 1 \\ y_1(\widetilde{t}) &= 1 - \int_0^{\widetilde{t}} \sup(-x, 1)\, dx && = 1 - \widetilde{t} \\ y_2(\widetilde{t}) &= 1 - \int_0^{\widetilde{t}} \sup(-x, 1-x)\, dx && = 1 - \widetilde{t} + \frac{\widetilde{t}^2}{2} \\ y_3(\widetilde{t}) &= 1 - \int_0^{\widetilde{t}} \sup\left(-x, 1 - x + \frac{x^2}{2}\right) dx && = 1 - \widetilde{t} + \frac{\widetilde{t}^2}{2} - \frac{\widetilde{t}^3}{3!} \end{aligned}$$

de esta forma se tiene, en general, que

$$\widetilde{y}_n(\widetilde{t}) = \sum_{i=0}^{n} (-1)^i \frac{\widetilde{t}^i}{i!} \qquad \Rightarrow \qquad y_n(t) = y_n(-\widetilde{t}) = \sum_{i=0}^{n} \frac{t^i}{i!}$$

por lo que

$$y(t) = \lim_{n\to\infty} y_n(t) = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{t^i}{i!} = e^t, \qquad -1 \leq t \leq 0.$$

Análogamente para el problema $(P-2)$ se tendrá que:

$$\begin{aligned}
y_0(t) &= y(0) && = 1\\
y_1(t) &= 1 + \int_0^t \sup(x,1)\,dx && = 1+t\\
y_2(t) &= 1 + \int_0^t \sup(x,1+x)\,dx && = 1+t+\frac{t^2}{2}\\
y_3(t) &= 1 + \int_0^t \sup\left(x,1+x+\frac{x^2}{2}\right)\,dx && = 1+t+\frac{t^2}{2}+\frac{t^3}{3!}
\end{aligned}$$

luego

$$y(t) = \lim_{n\to\infty} y_n(t) = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{t^i}{i!} = e^t, \qquad 0 \le t \le 1.$$

En resumen, la solución del problema de Cauchy planteado, en el intervalo $[-1,\,1]$ es $y(t) = e^t$.

E) Métodos numéricos (o en diferencias finitas).

Este tipo de métodos, a diferencia de los anteriores, no persiguen la determinación de la expresión de la función solución $y(t)$ de un problema de valor inicial. Tan sólo persiguen calcular, de forma aproximada, el valor de la solución en ciertos puntos t_i para $i = 0, 1, 2, ..., N$ seleccionados (bien de antemano por el usuario de dichos métodos, bien de forma "automática" a través de algoritmos diseñados para su cálculo con el objeto de asegurar ciertas tolerancias de error entre el valor exacto (desconocido) y el valor aproximado por el método). Para su descripción, siendo $I = [t_0, t_0 + T]$ consideremos dados los puntos en los que se va a determinar la solución y que denotaremos por:

$$t_0 < t_1 < t_2 < .... < t_n < t_{n+1} < .... < t_N = t_0 + T$$

Los métodos en diferencias finitas, en general, se basan en, conocidos los valores aproximados de la solución en los puntos $\{t_{n-k},\ t_{n-k+1},\ t_{n-k+2},\ .....,\ t_{n-1}\}$, que denotaremos como:

$$y_{n-k} \approx y(t_{n-k}),\ \ y_{n-k+1} \approx y(t_{n-k+1}),\ .....,\ y_{n-1} \approx y(t_{n-1})$$

interpolar, habitualmente mediante un polinomio, la función $y(t)$ sobre el soporte

$$\{t_{n-k}, ..., t_n\}.$$

Por ejemplo, si se utilizara interpolación polinómica de Lagrange el polinomio interpolador estaría dado por:

$$y(t) \approx p_{k,n}(t) = \sum_{j=0}^{k} y_{n-j}.L_{j,n}(t)$$

donde

$$L_{j,n}(t) = \prod_{i=0,\, i\neq j}^{k} \frac{(t - t_{n-i})}{(t_{n-j} - t_{n-i})}, \qquad (j = 0, ..., k)$$

son los correspondiente polinomios de base de Lagrange. Con ello:

$$y'(t) \approx p'_{k,n}(t) = \sum_{j=0}^{k} y_{n-j}.L'_{j,n}(t)$$

por lo que la EDO del P.V.I. podría aproximarse por:

$$\sum_{j=0}^{k} y_{n-j}.L'_{j,n}(t) \approx f(t, \sum_{j=0}^{k} y_{n-j}.L_{j,n}(t))$$

Esta expresión puede particularizarse en algún valor de la variable t (por ejemplo para $t = t_n$) y ello nos permite escribir, despejando y_n del lado izquierdo de la igualdad anterior que:

$$y_n = \Phi(y_{n-k}, y_{n-k+1}, ..., y_n)$$

En el proceso anterior se ha utilizado interpolación polinómica de Lagrange y se ha particularizado la expresión resultante en $t = t_n$. Obviamente muchas otras opciones serían posibles lo que nos conduciría a expresiones diferentes de la función Φ. Pero en todo caso serían expresiones en las que el valor aproximado y_n se estimaría en función de los valores aproximados, previamente calculados, $y_{n-k}, y_{n-k+1}, ..., y_{n-1}$ y eventualmente el propio valor y_n que se desea aproximar. Por ello, en general, se entenderá por **método numérico para resolver un problema de valor inicial** todo esquema de cálculo (algoritmo) que nos permita evaluar y_n mediante una expresión de la forma:

$$y_n = \Phi(y_{n-k}, y_{n-k+1}, ..., y_{n-1})$$

o

$$y_n = \Phi(y_{n-k}, y_{n-k+1}, ..., y_n).$$

Además tales tipos de métodos diremos que son métodos de k pasos en el sentido de que para evaluar y_n utilizan los k últimos valores previamente calculados. En el caso de que $k = 1$ los métodos correspondientes se dicen

métodos de pasos libres o **métodos unipaso**. Por el contrario si $k > 1$ los métodos correspondientes se dicen **métodos de pasos ligados** o **métodos multipaso**. Los métodos multipaso de k pasos sólo podrán utilizarse a partir del conocimiento de los k primeros valores: $y_0, y_1, ..., y_{k-1}$. Para la determinación de estos deberán emplearse métodos de un menor número de pasos. Y la combinación de unos y otros no es trivial pues interesa no "estropear" la precisión que pueda conseguirse con el método de k pasos por utilizar métodos menos precisos en las etapas iniciales. Además, conviene advertir ya, desde el comienzo, que la consideración de un número mayor de pasos no implica necesariamente la mejoría de las aproximaciones obtenidas. Es relativamente frecuente que un método de un paso bien elegido conduzca a soluciones más baratas de obtener e igual o más precisas que un método multipaso.

Observación 3.1.5 *Recordando la expresión formal de la solución de un P.V.I.:*

$$y(t) = y_0 + \int_{t_0}^{t} f(x, y(x))\, dx$$

se podría expresar el valor de la solución en un instante $t = t_n$ como:

$$y(t_n) = y_0 + \int_{t_0}^{t_n} f(x, y(x))\, dx = y_0 + \int_{t_0}^{t_{n-1}} f(x, y(x))\, dx + \int_{t_n-1}^{t_n} f(x, y(x))\, dx \quad \Rightarrow$$

$$\Rightarrow \quad y(t_n) = y(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(x, y(x))\, dx$$

En este sentido, una interpretación de la expresión anterior (recogida por ejemplo en Shampine et al.[11]*) consiste en decir que el valor exacto de la solución en el instante t_n depende sólamente del valor en t_{n-1} y no de valores anteriores. En otros téminos, lo anterior puede resumirse diciendo que la solución analítica de un P.V.I. no tiene "memoria" de lo que sucede en instantes anteriores al instante t_{n-1}. Ello es utilizado por los defensores de los métodos de pasos libres para justificar la bondad de estos métodos frente a los de pasos ligados ya que, en los primeros, las soluciones aproximadas tienen un comportamiento similar (sin "memoria") al de la solución exacta.*

Por otra parte, si el método se formula en la forma:

$$y_n = \Phi(y_{n-k}, y_{n-k+1}, ..., y_{n-1})$$

[11] Shampine, L.F., Allen, R.C. Jr. y Pruess, S. (1997). "Fundamentals of numerical computing". Ed. John Wiley & Sons, Inc.

diremos que es un método **explícito** pues el cálculo de y_n tan sólo implica evaluar la función $\Phi(y_{n-k}, y_{n-k+1}, ..., y_{n-1})$. En el caso de que el método se formule como:

$$y_n = \Phi(y_{n-k}, y_{n-k+1}, ..., y_n)$$

el método se dirá **implícito** pues el valor de y_n depende del propio y_n. Por ello, la determinación de y_n exigirá resolver la ecuación algebraica:

$$y_n - \Phi(y_{n-k}, y_{n-k+1}, ..., y_n) = 0$$

por ejemplo, utilizando alguna de las técnicas estudiadas en el capítulo anterior de estos apuntes. En general son métodos más costosos en cada uno de los pasos de integración (aunque puedan presentar también ventajas como el permitir considerar pasos de integración de mayor tamaño, como más adelante detallaremos).

A analizar el comportamiento de este tipo de métodos numéricos dedicaremos el resto de este tema. Y para ilustrarlos con más detalle comenzaremos abordando, en el apartado siguiente, el más simple de todos y alguna de sus variantes.

3.2. El método de Euler y variantes de él

Consideramos de nuevo el problema de valor inicial

$$\begin{cases} y'(t) &= f(t, y(t)), \quad t \in [t_0, t_0 + T] \\ y(t_0) &= y_0 \end{cases}$$

y una subdivisión del intervalo $[t_0, t_0 + T]$ mediante los puntos:

$$t_0 < t_1 < t_2 < .... < t_n < t_{n+1} < .... < t_N = t_0 + T$$

Designaremos por h_n a los valores $(t_{n+1} - t_n)$ y por y_n a las aproximaciones que se vayan obteniendo de $y(t_n)$, $(n = 0, 1, ..., N)$.

3.2.1. El esquema de cálculo del método de Euler y algunas variantes

En primer lugar, examinemos cómo se puede justificar el esquema de cálculo que se conoce con el nombre de método de Euler mediante diferentes interpretaciones: usando fórmulas de integración numérica; mediante desarrollos en serie de Taylor; de forma gráfica; mediante interpolación polinómica; etc.

Obtención del método de Euler mediante fórmulas de integración numérica

Es conocido que la solución del problema de valor inicial verificará:

$$\int_{t_n}^{t_{n+1}} y'(t)dt = \int_{t_n}^{t_{n+1}} f(t, y(t))dt \Rightarrow y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))dt$$

El problema que plantea en general la expresión anterior es la evaluación de la integral que en ella aparece. Una idea para resolver este problema puede ser evaluarla mediante una fórmula de integración numérica (véase por ejemplo C. Conde y E. Schiavi[12]). Al no ser exacta dicha fórmula, nos conducirá a un valor más o menos cercano al valor exacto de la solución. Una primera elección de la fórmula de integración numérica a emplear podría ser la fórmula del rectángulo con soporte en el extremo izquierdo del intervalo de integración, es decir:

$$y(t_{n+1}) \approx y(t_n) + h_n\, f(t_n, y(t_n)) \Rightarrow y_{n+1} = y_n + h_n\, f(t_n, y_n)$$

Con ello puede plantearse el siguiente algoritmo de cálculo:

$Dado\ y_0 = y(t_0)$:

$$y_{n+1} = y_n + h_n\, f(t_n, y_n) \qquad (n = 0, 1, 2, ..., N-1)$$

Este esquema de cálculo de las soluciones aproximadas del P.V.I. se conoce con el nombre **método de Euler explícito** en honor al matemático Leonard Euler (nacido en Basilea (Suiza) en 1707 y muerto en San Petersburgo (Rusia) en 1783). En el proceso de obtención del método se ha utilizado una de las múltiples formas de aproximar la integral que aparece en la expresión formal de la solución. Pero podrían haberse considerado muchas otras. Por ejemplo, si se hubiera utilizado la fórmula del rectángulo soportada en el extremo derecho del intervalo se obtendría:

$$y(t_{n+1}) \approx y(t_n) + h_n f(t_{n+1}, y(t_{n+1})) \Rightarrow y_{n+1} - h_n f(t_{n+1}, y_{n+1}) = y_n$$

con lo que podría plantearse el denominado **método de Euler implícito (o retrógrado)** consistente en:

$Dado\ y_0 = y(t_0)$:

$$Resolver:\ y_{n+1} - h_n\, f(t_{n+1}, y_{n+1}) = y_n \qquad (n = 0, 1, 2, ..., N-1)$$

[12] Conde, C. y Schiavi, E. (2000) "Elementos de Matemáticas: Guiones de los temas de la asignatura". Apuntes. Universidad Rey Juan Carlos.

Este esquema de cálculo exige, en cada paso, resolver una ecuación algebraica. Para ello puede utilizarse alguno de los métodos planteados en el tema 4 de estos apuntes.

Análogamente, si se hubiera evaluado la integral anterior mediante el método del trapecio, se obtendría el esquema:

$$y(t_{n+1}) \approx y(t_n) + \frac{h_n}{2}\left(f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))\right) \quad \Rightarrow$$

$$\Rightarrow \quad y_{n+1} - \frac{h_n}{2} f(t_{n+1}, y_{n+1}) = y_n + \frac{h_n}{2} f(t_n, y_n)$$

del que se infiere el algoritmo de cálculo:

Dado $y_0 = y(t_0)$:

Resolver: $y_{n+1} - \dfrac{h_n}{2} f(t_{n+1}, y_{n+1}) = y_n + \dfrac{h_n}{2} f(t_n, y_n), \quad (n = 0, 1, 2, ..., N-1)$

El esquema, también implícito, dado por la expresión anterior se designa habitualmente con el nombre de esquema de **Crank-Nicholson**.

Los esquemas de Euler explícito, de Euler implícito y de Crank-Nicholson son un caso particular de la familia de esquemas que se obtienen ponderando mediante un parámetro $\theta \in [0, 1]$ el peso dado a $f(t_n, y(t_n))$ frente al dado a $f(t_{n+1}, y(t_{n+1}))$ en la fórmula de integración numérica con lo que se obtendrían las expresiones:

$$y(t_{n+1}) \approx y(t_n) + h_n\left((1-\theta) f(t_n, y(t_n)) + \theta f(t_{n+1}, y(t_{n+1}))\right) \quad \Rightarrow$$

$$\Rightarrow \quad y_{n+1} - \theta h_n \left(f(t_{n+1}, y_{n+1}) = y_n + (1-\theta) h_n f(t_n, y_n)\right.$$

que corresponden a los métodos conocidos con el nombre de **θ-métodos**. Para el caso en que $\theta = 0$ se recupera el método de Euler explícito, para el caso $\theta = 1$ el de Euler implícito y para $\theta = 0{,}5$ el de Crank-Nicholson. En general si $\theta \neq 0$ el método es implícito y exige, en cada paso, la resolución de una ecuación de tipo algebraico para determinar y_{n+1}.

Esta forma de introducir el método de Euler y sus variantes muestra muchas otras formas de proceder (aproximando el valor de la integral de f mediante muy diferentes fórmulas de integración) que dan origen a muchos otros métodos numéricos de un paso. En los apartados siguientes presentaremos algunos otros métodos de un paso muy utilizados en la práctica. Pero antes de ello examinemos otras maneras diferentes (aunque equivalentes) de obtener los esquemas anteriores y analicemos cuál es el comportamiento del método de Euler y las variantes presentadas.

Obtención del método de Euler mediante desarrollos de Taylor

Si se supone que la función $y(t)$ es al menos de clase $C^2([t_0, t_N])$, puede expresarse el valor de $y(t)$ en el punto t_{n+1} en función del valor de $y(t)$ en t_n mediante el siguiente desarrollo en serie de Taylor:

$$y(t_{n+1}) = y(t_n + h_n) = y(t_n) + h_n\, y'(t_n) + \frac{h_n^2}{2}\, y''(t_n) + \ \ s \ \Rightarrow$$

$$\Rightarrow y(t_{n+1}) = y(t_n) + h_n\, f(t_n, y(t_n)) + \frac{h_n^2}{2}\, y''(t_n) + \ s$$

Si h_n es suficientemente pequeño, el desarrollo anterior podrá aproximarse despreciando los términos en los que aparezcan potencias de h_n superiores o iguales a 2, obteniéndose así nuevamente el esquema de cálculo del **método de Euler:**

$$y_{n+1} = y_n + h_n\, f(t_n, y_n)$$

Si en lugar de expresar el valor de $y(t_{n+1})$ en función del valor de la función $y(t)$ y de sus derivadas en t_n, se hiciera al revés y se expresara el valor en t_n en función del valor de $y(t)$ y de sus derivadas en t_{n+1} se obtendría la expresión siguiente:

$$y(t_n) = y(t_{n+1} - h_n) = y(t_{n+1}) - h_n\, y'(t_{n+1}) + \frac{h_n^2}{2}\, y''(t_{n+1}) + \ \ s \ \Rightarrow$$

$$\Rightarrow y(t_{n+1}) - h_n\, f(t_{n+1}, y(t_{n+1})) = y(t_n) - \frac{h_n^2}{2}\, y''(t_{n+1}) + \ \ s$$

de la que, despreciando los términos cuadráticos y posteriores en h_n, se obtiene la fórmula que constituye el esquema de cálculo del **método de Euler implícito (o retrógrado):**

$$y_{n+1} - h_n\, f(t_{n+1}, y_{n+1}) = y_n$$

Si se combinan los desarrollos en serie anteriores, ponderando el segundo por el parámetro $\theta \in [0, 1]$ y el primero por $(1 - \theta)$ se obtendría la familia de $\theta-$ **métodos**.

Esta forma de interpretar el método de Euler permitirá obtener las expresiones del " error de consistencia" de una forma sencilla (como el término de mayor orden que hemos despreciado en el desarrollo en serie). Además, la combinación de desarrollos en serie en torno a distintos puntos puede dar origen a otros métodos de muy diversa índole.

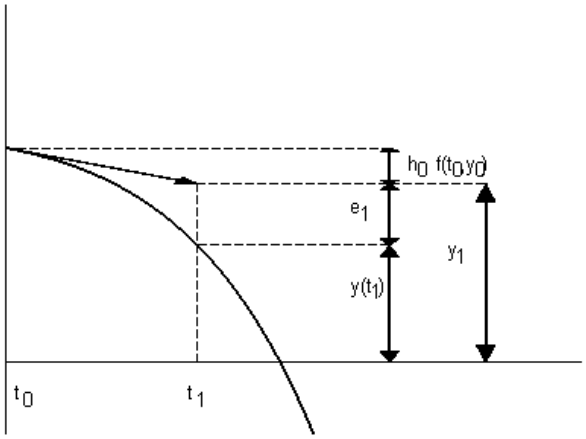


Figura 3.1:

Obtención del método de Euler de forma gráfica

Gráficamente, el primer paso del **método de Euler** consiste en seguir el camino señalado en la figura siguiente:

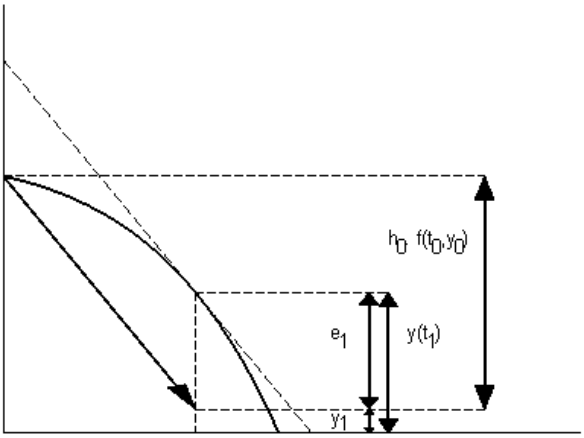
En dicha figura se representa como y_1 es evaluado siguiendo la tangente a la curva en t_0 (cuya pendiente está dada por $f(t_0, y_0)$). El punto (t_1, y_1) hallado es distinto en general de $(t_1, y(t_1))$. Por ello se ha elegido el punto inicial para ilustrar gráficamente el proceso ya que así no se consideran los errores acumulados en iteraciones anteriores. La idea es entonces repetir el proceso, partiendo de (t_1, y_1) y siguiendo la recta de dirección dada por $f(t_1, y_1)$.

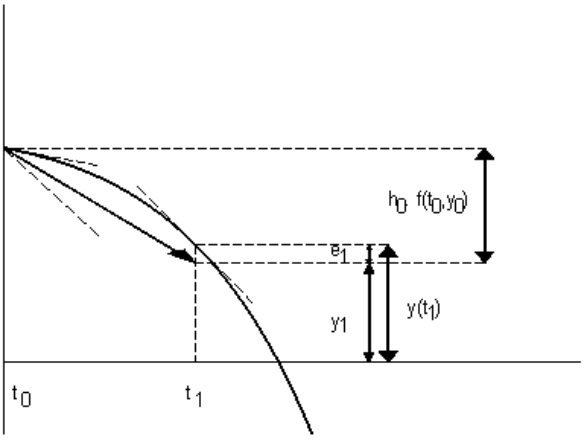
El **método de Euler implícito** puede ilustrarse de una forma muy similar pero siguiendo ahora una dirección paralela a la marcada por la tangente a la curva $y(t)$ en el punto t_1:

Observación 3.2.1 *Obsérvese que realmente la recta seguida en el método de Euler implícito es la que pasando por el punto* (t_0, y_0) *tiene por pendiente* $f(t_1, y_1)$ *(que en general será distinta de* $f(t_1, y(t_1))$*). En este sentido la figura anterior es engañosa pues la recta que se sigue no tiene la pendiente correcta.*

Y los θ-**métodos** se pueden justificar gráficamente por el hecho de seguir una dirección comprendida entre la marcada por la tangente a $y(t)$ en t_0 y la recta que pasando por $(t_0, y(t_0))$ es paralela a la tangente de $y(t)$ en t_1:

Esta forma de interpretar el método de Euler y sus variantes da un sentido geométrico al esquema.


$h_0\ f(t_0,y_0)$
e_1
$y(t_1)$
y_1


$h_0\ f(t_0,y_0)$
e_1
y_1
$y(t_1)$
t_0
t_1

Obtención del método de Euler mediante interpolación polinómica

El polinomio interpolador de Lagrange de $y(t)$ sobre el soporte formado por los dos puntos t_n y t_{n+1} puede expresarse mediante la fórmula:

$$y(t) \approx p_{1,n}(t) = y(t_n)\,\frac{t - t_{n+1}}{t_n - t_{n+1}} + y(t_{n+1})\,\frac{t - t_n}{t_{n+1} - t_n} \quad \forall t \in [t_n, t_{n+1}]$$

Por ello la primera derivada de $y(t)$ puede aproximarse por:

$$y'(t) \approx p'_{1,n}(t) = \frac{y(t_{n+1}) - y(t_n)}{h_n} \quad \forall t \in [t_n, t_{n+1}]$$

lo que nos permite aproximar en el intervalo $[t_n, t_{n+1}]$ la ecuación diferencial del P.V.I por:

$$p'_{1,n}(x) = \frac{y(t_{n+1}) - y(t_n)}{h_n} \approx y'(t) = f(t, y(t)), \quad \forall t \in [t_n, t_{n+1}].$$

Y particularizando esta expresión para $t = t_n$ vuelve a obtenerse la expresión utilizada en el **método de Euler explícito**:

$$y(t_{n+1}) \approx y(t_n) + h_n\, f(t_n, y(t_n)) \to y_{n+1} = y_n + h_n\, f(t_n, y_n)$$

Si en lugar de particularizar la expresión anterior en el instante $t = t_n$ se particularizase en el instante $t = t_{n+1}$ se obtendría la fórmula utilizada en el **método de Euler implícito**:

$$y(t_{n+1}) \approx y(t_n) + h_n\, f(t_{n+1}, y(t_{n+1})) \to y_{n+1} - h_n\, f(t_{n+1}, y_{n+1}) = y_n$$

Sumando las dos expresiones, multiplicando la primera por $(1 - \theta)$ y la segunda por θ, donde $\theta \in [0, 1]$, se obtiene la familia de los θ**−métodos.**

Esta forma de introducir el método de Euler y sus variantes principales nos vincula la teoría de interpolación con el método de Euler. El considerar un polinomio de interpolación u otro (variando la posición de los puntos del soporte, el número de ellos o el tipo de interpolación realizada) puede conducir a muy diferentes métodos numéricos de resolución de problemas de valor inicial. Por otra parte, en esta forma de obtener el método, al aproximar la derivada primera de $y(t)$ en t_n (respectivamente en t_{n+1}) por la expresión:

$$y'(t_n) \;\approx\; \frac{y(t_{n+1}) - y(t_n)}{h_n} \quad \left(\text{resp. } y'(t_{n+1}) \;\approx\; \frac{y(t_{n+1}) - y(t_n)}{h_n}\right)$$

se está utilizando la denominada **diferencia finita progresiva de primer orden** de $y(t)$ en t_n (respectivamente **diferencia finita regresiva (o retrógrada) de primer orden** de $y(t)$ en t_{n+1}) lo que da origen al nombre de este tipo de métodos numéricos. Puede consultarse, por ejemplo, C. Conde y E. Schiavi[13] para un estudio más detallado de las diferencias finitas y su uso en la obtención del polinomio interpolador.

Obtención del método de Euler mediante derivación numérica

Una forma equivalente de obtener el método de Euler consiste en considerar que, en un punto t^* la derivada de la función $y(t)$ está dada por:

$$y'(t^*) = \lim_{h\to 0}\frac{y(t^*+h)-y(t^*)}{h}$$

Para valores de h suficientemente pequeños, puede aproximarse la expresión anterior, prescindiendo del "límite" , es decir por:

$$y'(t^*) \approx \frac{y(t^*+h)-y(t^*)}{h}$$

Aplicando esta fórmula de derivación numérica para $t^* = t_n$ y con $h = h_n$ se tendrá que:

$$y'(t_n) \approx \frac{y(t_n+h_n)-y(t_n)}{h_n} = \frac{y(t_{n+1})-y(t_n)}{h_n}$$

con lo que la EDO del problema de valor inicial puede aproximarse por:

$$\frac{y(t_n+h_n)-y(t_n)}{h_n} \approx f(t_n, y(t_n))$$

de donde se obtiene nuevamente la expresión del **método de Euler explícito:**

$$y_{n+1} = y_n + h_n\, f(t_n, y_n)$$

Si en lugar de aplicar la fórmula de derivación numérica para $t^* = t_n$ se aplicase para $t^* = t_{n+1}$ y tomando ahora $h = -h_n$ se tendría que:

$$\frac{y(t_{n+1}-h_n)-y(t_{n+1})}{-h_n} \approx f(t_{n+1}, y(t_{n+1}))$$

de donde se podría obtener nuevamente la expresión del **método de Euler implícito:**

$$y_{n+1} - h_n\, f(t_{n+1}, y_{n+1}) = y_n.$$

[13] Conde, C. y Schiavi, E. (2000) "Elementos de Matemáticas: Guiones de los temas de la asignatura". Apuntes. Universidad Rey Juan Carlos.

Como en otras ocasiones, sumando a la expresión del método implícito multiplicada por θ la expresión del método explícito multiplicada por $(1-\theta)$ se recupera la expresión de los **θ-métodos.**

Esta forma de proceder nos permite atisbar que empleando diferentes formas de aproximar la derivada de una función (consúltese por ejemplo C. Conde y E. Schiavi[14]) se podrán obtener muy distintos esquemas de resolución del problema de valor inicial.

3.2.2. El algoritmo recogiendo el método de Euler y sus variantes

Resumiendo los esquemas numéricos hasta ahora planteados, todos ellos pueden englobarse dentro de la familia de los **θ-métodos** asignándose a θ un valor comprendido entre 0 y 1 y correspondiendo los valores extremos de este parámetro al **método de Euler explícito** ($\theta = 0$) o **implícito** ($\theta = 1$) y denominándose **método de Crank-Nicholson** al esquema que se obtiene al dar a θ el valor 0,5. La extensión de estos esquemas de cálculo al caso de problemas de valor inicial regidos por un sistema de ecuaciones diferenciales de primer orden es inmediata bastando para ello sustituir las funciones escalares por funciones vectoriales. Por ello el algoritmo que presentaremos a continuación lo plantearemos sobre el caso más general de un problema de valor inicial formulado como:

Conocidas la función $\mathbf{f} : I \times \mathbb{R}^m \to \mathbb{R}^m$ *y dado un vector de* $\mathbb{R}^m$ *denotado por* $\mathbf{y}^{(0)}$*, determinar la función* $\mathbf{y} : I \to \mathbb{R}^m$ *solución del problema de valor inicial:*

$$\begin{cases} \mathbf{y}'(t) & = \mathbf{f}(t, \mathbf{y}(t)), \qquad t \in I = [t_0, t_0 + T] \\ \mathbf{y}(t_0) & = \mathbf{y}^{(0)} \end{cases}$$

Con ello, un primer algoritmo de la familia de los θ-métodos puede escribirse como sigue:

INICIO DEL ALGORITMO

Dados:

θ, t_0, t_1, ..., $t_N = t_0 + T$, la expresión de la función $\mathbf{f}(t, \mathbf{y}(t))$ y el vector $\mathbf{y}^{(0)}$

Para $n = 0$, **hasta** $n = N - 1$, **con paso** 1, **hacer:**

$h_n \leftarrow t_{n+1} - t_n$

Resolver la ecuación:

[14] Conde, C. y Schiavi, E. (2000) "Elementos de Matemáticas: Guiones de los temas de la asignatura". Apuntes. Universidad Rey Juan Carlos.

$$\mathbf{y}^{(n+1)} - h_n\,\theta\,\mathbf{f}(t_{n+1},\mathbf{y}^{(n+1)}) = \mathbf{y}^{(n)} + (1-\theta)\,\mathbf{f}(t_n,\mathbf{y}^{(n)})$$

Escribir el vector $\mathbf{y}^{(n+1)}$ como solución aproximada en t_{n+1}.

Fin bucle en n.

FIN DEL ALGORITMO.

Observación 3.2.2 *El algoritmo anterior puede (y en la práctica debe) modificarse con alguna estrategia que permita controlar el valor de h_n en cada etapa con el objeto de minimizar el error entre la solución aproximada ofrecida por el método y la solución exacta del P.V.I. A estas estrategias de control del paso de integración nos referiremos más adelante.*

La pregunta que podemos formularnos ahora es: ¿Entre los esquemas planteados, cuál es "mejor"?. En otros términos ¿qué valor conviene asignar al parámetro θ ?. La respuesta a esta cuestión dependerá de qué entendamos por "mejor". Si por tal entendemos, por ejemplo, "lo que menos operaciones exige en cada paso" es evidente que la respuesta sería el método de Euler explícito pues los implícitos conllevan, en general, la necesidad de resolver una ecuación (o sistema) no lineal. No obstante, los esquemas implícitos, como se verá posteriormente, pueden permitir considerar pasos de integración mayores, lo que se traduce en realizar menos etapas de cálculo. En resumen los esquemas implícitos son más costosos por paso pero necesitan realizar menos pasos. Por ello, si por "mejor" se entendiese "el que menos operaciones realice en total" la respuesta no es evidente. Por otra parte, si por "mejor" entendemos el que se cometa menos error entre el valor aproximado y el valor exacto en cada punto la respuesta podrá ser diferente. Al análisis de los métodos planteados nos referiremos un poco más adelante y, entonces, trataremos de responder a esta cuestión. Pero antes de ello, para acabar de ilustrar el funcionamiento de los métodos vistos, y poner de manifiesto algunas de las limitaciones que tienen, apliquémoslos a algunos ejemplos.

3.2.3. Algunos ejemplos ilustrativos de aplicación del método de Euler

Primer ejemplo: aplicación a un problema regido por una única ecuación de variables separadas.

Con el objeto de poder comparar la solución aproximada con la solución exacta comencemos considerando un problema de valor inicial sencillo: regido por una EDO de variables separadas. Ello nos permitirá ilustrar la práctica del método y realizar algunas primeras consideraciones.

Considérese el P.V.I. formulado por:

$$\begin{cases} y'(t) &= 2\,t + e^t, \qquad t \in [0,\ 1] \\ y(0) &= 0 \end{cases}$$

La solución analítica de este P.V.I. está dada por:

$$y(t) = -1 + t^2 + e^t$$

y con ella podremos comparar el comportamiento de los métodos planteados. Resolvamos el problema por el método de Euler explícito determinando los valores de la solución aproximada en los instantes $t_0 = 0{,}0$, $t_1 = 0{,}1$, $t_2 = 0{,}2$, $t_3 = 0{,}3$, $t_4 = 0{,}4$, ..., $t_9 = 0{,}9$, y $t_{10} = 1$, siendo la longitud del paso de integración en todas las etapas de cálculo $h = 0{,}1$. Con ello el esquema de cálculo, según vimos anteriormente puede resumirse en:

$$\begin{array}{ll} y_0 & = 0 \\ y_{n+1} & = y_n + 0{,}1\,(2\,t_n + e^{t_n}) \qquad (n = 0, 1, 2, ..., 9) \end{array}$$

Por tanto, los primeros valores calculados serán:

$$y_0 = 0$$

$$y_1 = y_0 + 0{,}1\,(2\,t_0 + e^{t_0}) = 0 + 0{,}1\,(2\,0 + e^0) = 0{,}1$$

$$y_2 = y_1 + 0{,}1\,(2\,t_1 + e^{t_1}) = 0{,}1 + 0{,}1\,(2\,0{,}1 + e^{0{,}1}) = 0{,}23052$$

$$y_3 = y_2 + 0{,}1\,(2\,t_2 + e^{t_2}) = 0{,}23052 + 0{,}1\,(2\,0{,}2 + e^{0{,}2}) = 0{,}39265$$

$$y_4 = y_3 + 0{,}1\,(2\,t_3 + e^{t_3}) = 0{,}39265 + 0{,}1\,(2\,0{,}3 + e^{0{,}3}) = 0{,}58764$$

Con el mismo paso y para los mismos instantes de cálculo puede obtenerse la solución mediante el método de Euler implícito:

$$\begin{array}{ll} y_0 & = 0 \\ y_{n+1} & = y_n + 0{,}1\,(2\,t_{n+1} + e^{t_{n+1}}) \qquad (n = 0, 1, 2, ..., 9) \end{array}$$

obteniéndose

$$y_0 = 0$$

$$y_1 = y_0 + 0{,}1\,(2\,t_1 + e^{t_1}) = 0 + 0{,}1\,(2\,0{,}1 + e^{0{,}1}) = 0{,}13052$$

$$y_2 = y_1 + 0{,}1\,(2\,t_2 + e^{t_2}) = 0{,}1352 + 0{,}1\,(2\,0{,}2 + e^{0{,}2}) = 0{,}29266$$

$$y_3 = y_2 + 0{,}1\,(2\,t_3 + e^{t_3}) = 0{,}29266 + 0{,}1\,(2\,0{,}3 + e^{0{,}3}) = 0{,}48764$$

$$y_4 = y_3 + 0{,}1\,(2\,t_4 + e^{t_4}) = 0{,}48764 + 0{,}1\,(2\,0{,}4 + e^{0{,}4}) = 0{,}65182$$

valores que son sensiblemente mayores que los obtenidos por el método explícito.

Resolvamos ahora por el método de Crank-Nicholson. Este esquema, sobre este P.V.I. se formula:

$$\begin{array}{ll} y_0 & = 0 \\ y_{n+1} & = y_n + 0{,}05\left[(2t_n + e^{t_n}) + (2t_{n+1} + e^{t_{n+1}})\right], \qquad (n = 0, 1, 2, ..., 9) \end{array}$$

y los primeros valores a los que nos conduce son:

$$y_0 = 0$$

$$\begin{aligned} y_1 &= y_0 + 0{,}05(2\,(t_0 + t_1) + e^{t_0} + e^{t_1}) = \\ &= 0 + 0{,}05\,(0{,}2 + 1 + e^{0{,}1}) = 0{,}11526 \end{aligned}$$

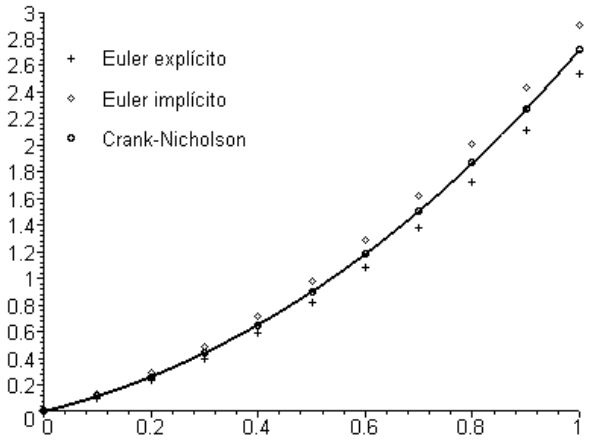
$$\begin{aligned} y_2 &= y_1 + 0{,}05(2\,(t_1 + t_2) + e^{t_1} + e^{t_2}) = \\ &= 0{,}11526 + 0{,}05\,(0{,}6 + e^{0{,}1} + e^{0{,}2}) = 0{,}26158 \end{aligned}$$

$$\begin{aligned} y_3 &= y_2 + 0{,}05(2\,(t_2 + t_3) + e^{t_2} + e^{t_3}) = \\ &= 0{,}26158 + 0{,}05\,(1{,}0 + e^{0{,}2} + e^{0{,}3}) = 0{,}44015 \end{aligned}$$

$$\begin{aligned} y_4 &= y_3 + 0{,}05(2\,(t_3 + t_4) + e^{t_3} + e^{t_4}) = \\ &= 0{,}26158 + 0{,}05\,(1{,}4 + e^{0{,}3} + e^{0{,}4}) = 0{,}65182 \end{aligned}$$

Estos valores son intermedios entre los obtenidos por los métodos anteriores. En este caso es sencillo verificar cuáles son los valores más precisos pues la solución analítica es conocida. Podemos representar los valores obtenidos junto a la solución exacta obteniendo:

Puede observarse en la figura anterior que el método de Crank-Nicholson es el que mejor se ajusta a la solución entre los tres planteados mientras que el de Euler explícito aproxima la solución "por defecto" en tanto que el implícito lo hace "por exceso". Pero también puede apreciarse que los errores van creciendo a medida que avanza el instante de cálculo. Ello ya nos indica que no sólo deberemos intentar conocer el error que se comete al predecir con un método la solución aproximada y_{n+1} partiendo del valor exacto en el instante anterior $y(t_n)$ sino que deberá también prestarse atención a cómo los errores cometidos en una etapa influyen en el error de las siguientes. Al fin y al cabo, cuando obtenemos y_{n+1} no partimos de $y(t_n)$ sino de y_n y, en general $y_n \neq y(t_n)$.



Segundo ejemplo: un problema modelo típico que nos introduce en los límites de estabilidad de los esquemas.

Un problema de valor inicial frecuentemente utilizado para testear los esquemas numéricos (por razones que más adelante se explicarán) es el siguiente:

$$\begin{cases} y'(t) &= -ky(t), & t>0 \\ y(0) &= 1 \end{cases}$$

donde k es una constante positiva. Este sencillo problema (que es una EDO de variables separadas), admite como solución analítica:

$$y(t) = e^{-kt}$$

Obsérvese que la solución analítica es decreciente y siempre positiva, características que sería deseable que poseyesen también las soluciones aproximadas que nos generen los esquemas numéricos al aplicarlos a este caso concreto.

La aplicación de los esquemas de Euler explícito, Euler implícito y θ-métodos en general al P.V.I. considerado, suponiendo de momento que la longitud entre los instantes de cálculo es constante, h, se reduce a:

a) Euler explícito:

$$y_{n+1} = y_n + h\,(-k\,y_n) \Rightarrow y_{n+1} = (1-k\,h)\,y_n$$

y por recursión

$$y_{n+1} = (1-k\,h)^{n+1}\,y_0$$

b) Euler implícito:

$$y_{n+1} = y_n + h\,(-k\,y_{n+1}) \Rightarrow y_{n+1} = \left(\frac{1}{1+k\,h}\right)\,y_n$$

y por recursión

$$y_{n+1} = \left(\frac{1}{1+k\,h}\right)^{n+1}\,y_0$$

c) θ-métodos:

$$\begin{aligned} y_{n+1} &= y_n - h\,((1-\theta)\,k\,y_n + \theta\,k\,y_{n+1}) \quad \Rightarrow \\ \Rightarrow \quad y_{n+1} &= \left(\frac{1-(1-\theta)\,k\,h}{1+\theta\,k\,h}\right)\,y_n \end{aligned}$$

y por recursión

$$y_{n+1} = \left(\frac{1-(1-\theta)\,k\,h}{1+\theta\,k\,h}\right)^{n+1}\,y_0$$

Examinemos cuando estas soluciones son decrecientes y positivas.

En el primer caso, método de Euler explícito, la solución aproximada en cada instante de tiempo t_{n+1} es la condición inicial ($y(0) = y_0 = 1$) multiplicada por α^{n+1} donde hemos denotado por α al valor $\alpha = (1 - k\,h)$, es decir escribimos el esquema de Euler explícito en la forma: $y_n = \alpha^n\,y_0 = \alpha^n$. Para que la solución aproximada efectivamente decrezca (al menos en valor absoluto) deberá verificarse que $|\alpha| < 1$. Puesto que hemos supuesto $k > 0$ y la longitud del paso también es positiva, un pequeño cálculo nos conduce a que el decaimiento en los valores absolutos de la solución se preservará si:

$$0 < k\,h < 2.$$

Lo anterior nos limita el tamaño del paso que podemos utilizar si deseamos que nuestra solución aproximada decrezca con el transcurrir del tiempo, debiendo verificarse para ello que:

$$h < \frac{2}{k}$$

En el caso de que $h = 2/k$ la solución numérica no "explotará" con el transcurrir del tiempo pero como en ese caso $\alpha = -1$ la solución numérica oscilará en cada instante de cálculo pasando del valor -1 al valor 1.

No obstante se puede ser un poco más sutil en el examen de la longitud del paso de cálculo a utilizar. En efecto, si se desea que además la solución aproximada sea siempre positiva la condición a imponer es que:

$$0 < \alpha < 1$$

lo que nos conduce a que

$$h \in \left]0,\ \frac{1}{k}\right[$$

Nótese que el intervalo al que deben pertenecer los valores de h es un intervalo abierto. En efecto, en sus extremos la solución numérica no oscilará entre valores negativos y positivos, pero también adolecerá de graves defectos. Así el valor $h = 0$ no puede tomarse (pues todos los puntos de cálculo se reducirían al instante inicial t_0). Y si se tomase $h = 1/k$ el valor de α sería nulo y la solución aproximada en instantes positivos será siempre 0. El problema entonces es saber cuál es el mejor valor que podríamos tomar para el paso de integración. El responder a esta pregunta, en este caso concreto, es muy simple. Puesto que la solución analítica es conocida, se tendría un error nulo si:

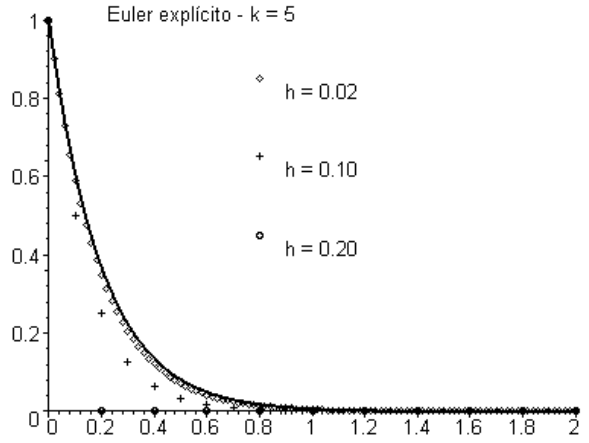
$$(1 - h\,k)^n = e^{t_n} = e^{nh} \Rightarrow 1 - hk = e^h \Rightarrow h = 0$$

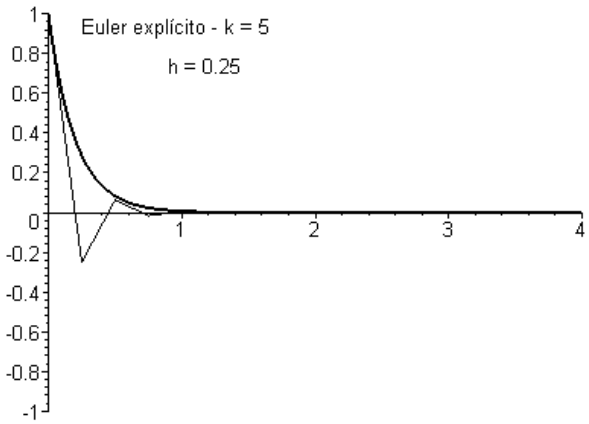
Pero es absurdo pensar en tomar $h = 0$ (pues ello haría que todos los instantes de cálculo fuesen el mismo: t_0). No obstante, lo anterior nos indica que cuanto menor sea el paso de integración menor será el error cometido. Debe matizarse no obstante que aquí nos estamos refiriendo sólo a un tipo de error: el que se comete con la aplicación del método sin "redondear" los valores obtenidos. La consideración de que al trabajar con un programa de cálculo se introducen además errores de redondeo al no poderse almacenar infinitos decimales introducirá (como se verá más adelante) un límite inferior a los valores de h que puedan considerarse. Pero olvidándonos por el momento de los errores de redondeo, efectivamente, cuanto menor sea el tamaño del paso de integración más precisa será la aproximación obtenida.

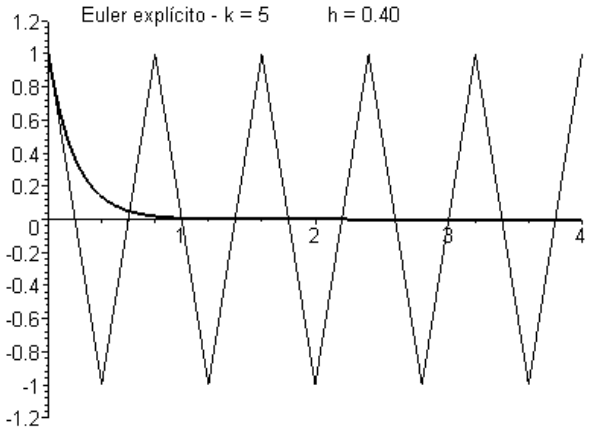
Ilustremos lo anterior presentando la solución analítica del problema y las soluciones aproximadas mediante el método de Euler para distintos valores del tamaño del paso de integración y tomando como valor de $k = 5$ (lo que hace que para pasos menores que $1/5 = 0{,}2$ el método produzca soluciones siempre positivas que decrecen en valor absoluto con el paso del tiempo):

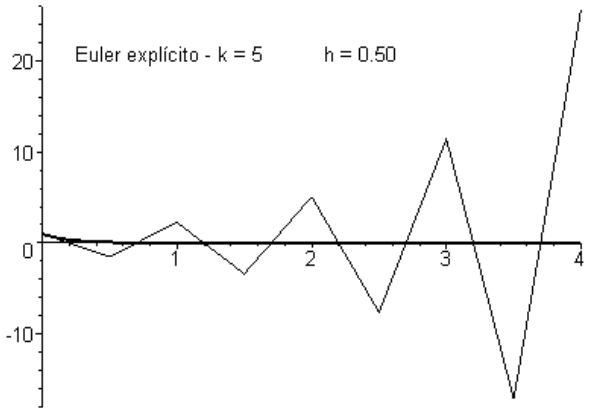
Obsérvese como para $h = 0{,}2$ la solución obtenida es, efectivamente, siempre nula y como la solución obtenida con $h = 0{,}02$ es más precisa que la obtenida para $h = 0{,}1$ y esta, a su vez, es más precisa que la obtenida para $h = 0{,}2$. Si tomásemos mayores valores del paso de integración comenzarían a aparecer soluciones oscilantes pero si no sobrepasamos el valor crítico $h = 2/k = 0{,}4$ la solución decrece en valor absoluto por lo que para tiempos grandes acaba convergiendo hacia la solución exacta como muestra la figura siguiente:

Si a h le asignamos el valor crítico $h = 0{,}4$ la solución permanece acotada, no "explota" para tiempos grandes, pero tampoco decrece en valor absoluto y oscila entre -1 y 1 como se ve en la gráfica siguiente:


Euler explícito - k = 5
h = 0.02
h = 0.10
h = 0.20
1
0.8
0.6
0.4
0.2
0
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2


Euler explícito - k = 5
h = 0.25
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
1 2 3 4


Euler explícito - k = 5
h = 0.40
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
1 2 3 4



Finalmente si a h le asignamos valores superiores a 0,4 la solución aproximada, en valor absoluto, crecerá por lo que tenderá a "explotar" para tiempos elevados:

Diremos en este caso que el esquema considerado se hace **inestable**.

Examinemos ahora los métodos implícitos (es decir con $\theta > 0$). En este caso la solución aproximada también puede expresarse como:

$$y_n = \alpha^n y_0 \qquad (n = 1, 2, ..., N)$$

siendo ahora

$$\alpha = \left(\frac{1 - (1-\theta)\, k\, h}{1 + \theta\, k\, h} \right) = \frac{1 + \theta\, k\, h - k\, h}{1 + \theta\, k\, h} = 1 - \frac{k\, h}{1 + \theta\, k\, h}$$

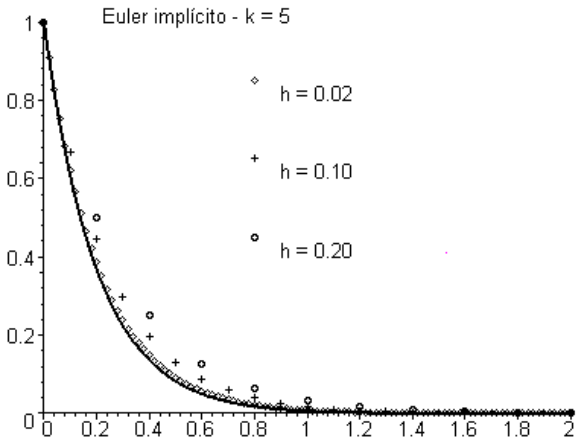
Puesto que $0 < \theta \leq 1$, y $k > 0$ se verificará que α siempre será inferior a 1. Para garantizar que también es superior a -1 se debe verificar que:

$$\frac{k\, h}{1 + \theta\, k\, h} < 2 \Rightarrow (1 - 2\,\theta)\, k\, h < 2$$

En el caso de que $\theta \geq 0{,}5$ la expresión anterior se verificará para cualquier valor que se tome para h : el esquema en ese caso es incondicionalmente estable. Y si $\theta < 0{,}5$ la condición a imponer sobre el paso de integración para garantizar la estabilidad, será:

$$h < \frac{2}{k\,(1 - 2\,\theta)}$$

Estas elecciones del valor del paso de integración nos aseguran que el valor de la solución aproximada que se obtenga también va a decaer hacia 0 para tiempos grandes. Si además se desease asegurar la positividad de las soluciones



aproximadas se debería "hilar un poco más fino". En efecto para que esto suceda se debe obligar a que $0 < \alpha < 1$. Ello nos conduce a que:

$$\frac{k\,h}{1+\theta\,k\,h} < 1 \quad \Rightarrow \quad (1-\theta)\,k\,h < 1 \quad \Rightarrow \quad h < \frac{1}{k\,(1-\theta)}$$

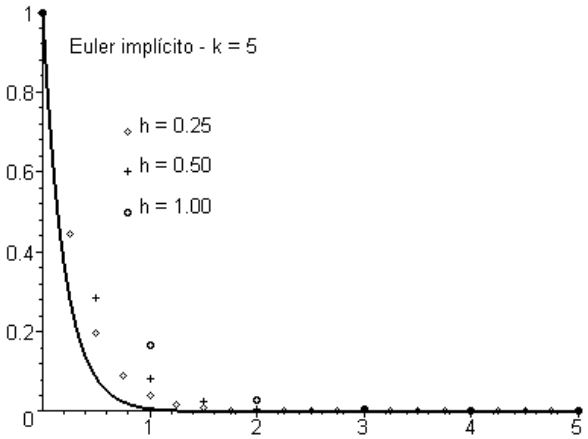
Obsérvese que cuanto más se aproxime θ al valor 1 mayor será el tamaño de paso máximo que puede ser tomado para garantizar tanto la "estabilidad" de la solución aproximada (es decir que no explote para tiempos grandes) como su positividad. Particularmente para $\theta = 1$ (esquema de Euler implícito) cualquier tamaño del paso de integración nos serviría para asegurar ambas cosas. Para el esquema de Crank-Nicholson la "estabilidad" nos la garantizaría cualquier tamaño del paso pero la positividad nos obligaría a considerar tamaños inferiores a $2/k$. Ello nos amplía el tamaño del paso que puede ser tomado respecto al esquema explícito (en el que la positividad y la estabilidad se aseguraban con tamaños inferiores a $1/k$). Ilustremos lo anterior con la aplicación de los esquemas de Euler implícito y de Crank-Nicholson al problema:

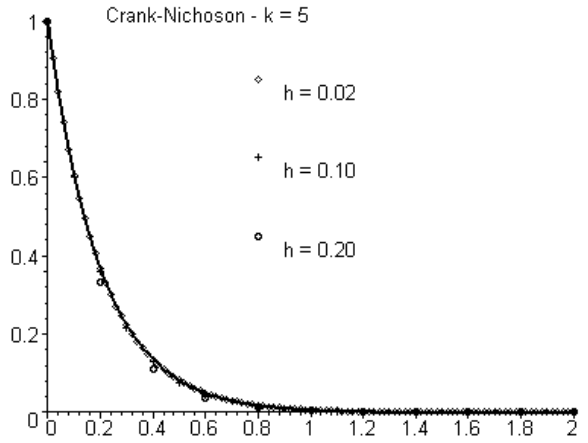
$$\begin{cases} y'(t) &= -5.y(t), \qquad t > 0 \\ y(0) &= 1 \end{cases}$$

para distintos valores del paso de integración. Comencemos con el método de Euler implícito que, en este caso se formulará como:

$$\begin{aligned} y_0 &= 1 \\ y_n &= \left(\frac{1}{1+5\,h}\right)^n, \qquad (n = 1, 2, 3...) \end{aligned}$$

y que será estable y positivo para todos los valores de h que se consideren.



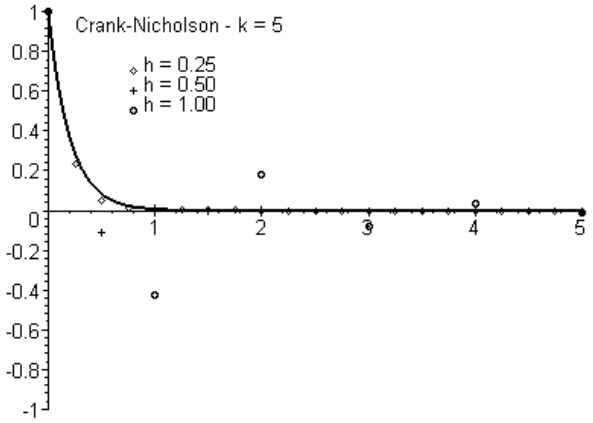


Como puede apreciarse las soluciones, efectivamente siempre permanecen positivas y decrecen con el tiempo. Pero también puede advertirse como su precisión es más pobre a medida que se aumenta el paso de integración temporal. Además la solución es aproximada en "exceso", es decir, con valores mayores que la exacta, a diferencia de lo que sucedía en el método explícito.

Presentemos ahora algunos resultados obtenidos con el método de Crank-Nicholson:

$$\begin{aligned} y_0 &= 1 \\ y_n &= \left(\frac{1-0{,}5\,5\,h}{1+0{,}5\,5\,h}\right)^n \qquad (n=1,2,3...) \end{aligned}$$

Obsérvese como, de momento, todas las soluciones permanecen positivas y, además, el método no parece tan sensible al paso de integración en cuanto a su precisión. No obstante si sobrepasamos el límite $h = 2/k = 2/5 = 0{,}4$ para el



tamaño del paso comenzaremos a perder la positividad y a notar la influencia del valor de h.

Observación 3.2.3 *Esta cualidad de los esquemas implícitos, poder considerar pasos de integración más amplios que en los explícitos sin perder la estabilidad, permite pensar en abordar problemas que se planteen en dominios* $[t_0, t_0 + T]$ *en los que* T *tome valores elevados. Piénsese, por ejemplo, que la EDO* $y' = -k\,y$ *es la que rige procesos de desintegración de isótopos radiactivos que, por su peligrosidad hacia el entorno, pueden requerir estudiarlos durante millones de años lo que haría inviable la consideración de pasos de integración excesivamente pequeños. Pero, lamentablemente, no todo es la estabilidad y la positividad de los esquemas de cálculo sino que, además, habrá que prestar atención a su "precisión" como se hará un poco más adelante. Y a otros problemas como el que se plantea en el ejemplo siguiente.*

Tercer ejemplo: En los problemas implícitos puede no haber unicidad de la solución aproximada.

El ejemplo anterior puede llevarnos a la conclusión de que los esquemas implícitos son "mejores" que los explícitos y a plantearnos entonces el por qué de estos últimos. No obstante recordemos que los esquemas implícitos exigen, en general, resolver una ecuación algebraica de tipo no lineal. Ello, aparte del esfuerzo computacional que pueda requerir, plantea algunos problemas adicionales tales como: ¿con qué valor se inicializa el método de resolución de ecuaciones no lineales? y dado que la ecuación algebraica puede tener más de una solución ¿cómo podemos obtener entre ellas la "más conveniente"?. Esta situación es la que pretendemos ilustrar en este ejemplo resolviéndola a través

de la consideración de los denominados esquemas "predictor-corrector".

Considérese el problema de valor inicial formulado mediante:

$$\begin{cases} y'(t) &= -y^2(t), \quad t > 0 \\ y(0) &= 1 \end{cases}$$

La solución analítica de este problema, cuya EDO es una ecuación de tipo Bernoulli, puede obtenerse realizando el cambio de variable:

$$u(t) = \frac{1}{y(t)}$$

lo que nos conduce a que

$$u'(t) = \frac{-y(t)}{y^2(t)}$$

lo cual nos permite integrar la EDO dada como sigue:

$$y'(t) = -y^2(t) \quad \Rightarrow \quad \frac{y'(t)}{y^2(t)} = -1 \quad \Rightarrow \quad u'(t) = 1 \quad \Rightarrow$$

$$u(t) = C + t \quad \Rightarrow \quad y(t) = \frac{1}{C+t}.$$

Obligando ahora a que se verifique la condición inicial se obtiene finalmente como solución analítica del P.V.I. planteado:

$$y(t) = \frac{1}{1+t}$$

Ejercicio 3.2.1 *La EDO anterior también podría considerarse como una EDO de variables separadas. Te dejamos como ejercicio propuesto el resolverla de esta manera obteniendo la misma solución del problema de valor inicial.*

Si aplicamos al problema de este ejemplo el método de Euler implícito, con paso de integración constante (h), obtendremos:

$$y_{n+1} + h\, y_{n+1}^2 = y_n \quad (n = 0, 1, 2, ..N-1)$$

Tomemos, por ejemplo, $h = 0{,}1$ y calculemos el valor aproximado de la solución en $t_1 = 0{,}1$.

$$y_1 + 0{,}1\, y_1^2 = 1 \Rightarrow 0{,}1\, y_1^2 + y_1 - 1 = 0$$

ecuación de segundo grado que admite por soluciones:

$$y_1 = \frac{-1 \pm \sqrt{1 + 0{,}4}}{0{,}2}$$

es decir, los valores:

$$-10{,}91607978....,\qquad 0{,}91607978....$$

En este caso, en el que conocemos la solución exacta ($y(0{,}1) = 1/1{,}1 = 0{,}909090...$), es evidente que debería optarse por la segunda de las aproximaciones obtenidas. Pero, en un caso real en el que no se conociese la solución exacta ¿cómo se podría saber cuál es la solución a considerar?. En asusencia de criterios físico-químicos o técnicos no se pude descartar una solución frente a otra. Por ello, en estos casos, lo aconsejable es acudir a los **métodos predictor-corrector** en los que, en síntesis, se utiliza un método explícito para (predecir) obtener el valor de partida con el que aplicar el método de resolución de ecuaciones no lineales a la ecuación obtenida del esquema implícito que nos permite refinar (corregir) el valor aproximado de la solución finalmente obtenida. La aplicación de esta forma de proceder a nuestro problema nos conduciría en el primer paso de integración a:

Fase de predicción mediante Euler explícito:

$$y_1^{(0)} = y_0 - h\, y_0^2 = 1 - 0{,}1\,(1)^2 = 0{,}9$$

Fase de corrección mediante Euler implícito (resolviendo por el método de aproximaciones sucesivas):

$$y_1^{(iter+1)} = y_0 - h\,\left(y_1^{(iter)}\right)^2 \qquad (iter = 0, 1, 2, ...)$$

que nos conduce a que

$$y_1^{(1)} = 1 - 0{,}1\,(0{,}9)^2 = 0{,}919$$

$$y_1^{(2)} = 1 - 0{,}1\,(0{,}919)^2 = 0{,}9155439$$

$$y_1^{(3)} = 1 - 0{,}1\,(0{,}9155439)^2 = 0{,}916177$$

$$y_1^{(4)} = 1 - 0{,}1\,(0{,}916177)^2 = 0{,}9160617$$

$$y_1^{(5)} = 1 - 0{,}1\,(0{,}9160617)^2 = 0{,}9160838$$

$$y_1^{(6)} = 1 - 0{,}1\,(0{,}9160838)^2 = 0{,}9160792$$

$$y_1^{(7)} = 1 - 0{,}1\,(0{,}9160792)^2 = 0{,}9160798$$

por lo que tomaremos como $y_1 = 0{,}9160798$.

En el segundo paso de integración se tendría:

Fase de predicción mediante Euler explícito:

$$y_2^{(0)} = y_1 - h\, y_1^2 = 0{,}9160798 - 0{,}1\,(0{,}9160798)^2 = 0{,}8321595$$

Fase de corrección mediante Euler implícito (resolviendo por el método de aproximaciones sucesivas):

$$y_2^{(iter+1)} = y_1 - h\,\left(y_2^{(iter)}\right)^2, \qquad (iter = 0, 1, 2, ...)$$

que nos conduce a que

$$y_2^{(1)} = 0{,}9160798 - 0{,}1\,(0{,}8321595)^2 = 0{,}8468308$$

$$y_2^{(2)} = 0{,}9160798 - 0{,}1\,(0{,}8468308)^2 = 0{,}8443675$$
$$y_2^{(3)} = 0{,}9160798 - 0{,}1\,(0{,}8443675)^2 = 0{,}8447841$$
$$y_2^{(4)} = 0{,}9160798 - 0{,}1\,(0{,}8447841)^2 = 0{,}8447137$$
$$y_2^{(5)} = 0{,}9160798 - 0{,}1\,(0{,}8447137)^2 = 0{,}8447256$$
$$y_2^{(6)} = 0{,}9160798 - 0{,}1\,(0{,}8447256)^2 = 0{,}8447236$$
$$y_2^{(7)} = 0{,}9160798 - 0{,}1\,(0{,}8447236)^2 = 0{,}8447239$$

por lo que tomaremos como $y_2 = 0{,}8447239$.

El tercer paso de integración que nos permitirá calcular $y_3 \approx y(0{,}3)$ consistirá en:

Fase de predicción mediante Euler explícito:

$$y_3^{(0)} = y_2 - h\, y_2^2 = 0{,}8447239 - 0{,}1\,(0{,}8447239)^2 = 0{,}7733681$$

Fase de corrección mediante Euler implícito (resolviendo por el método de aproximaciones sucesivas):

$$y_3^{(iter+1)} = y_2 - h\,\left(y_2^{(iter)}\right)^2, \qquad (iter = 0, 1, 2, ...)$$

que nos conduce a que

$$y_3^{(1)} = 0{,}8447239 - 0{,}1\,(0{,}7733681)^2 = 0{,}7849141$$

$$y_3^{(2)} = 0{,}8447239 - 0{,}1\,(0{,}7849141)^2 = 0{,}7831149$$
$$y_3^{(3)} = 0{,}8447239 - 0{,}1\,(0{,}7831149)^2 = 0{,}783397$$
$$y_3^{(4)} = 0{,}8447239 - 0{,}1\,(0{,}783397)^2 = 0{,}7833529$$
$$y_3^{(5)} = 0{,}8447239 - 0{,}1\,(0{,}7833529)^2 = 0{,}7833598$$
$$y_3^{(6)} = 0{,}8447239 - 0{,}1\,(0{,}7833598)^2 = 0{,}7833587$$
$$y_3^{(7)} = 0{,}8447239 - 0{,}1\,(0{,}7833587)^2 = 0{,}7833589$$

por lo que tomaremos como $y_3 = 0{,}7833589$.

En sucesivos pasos de integración temporal la estrategia predictora-correctora que acabamos de realizar nos conduce a que $y_4 = 0{,}7300601$, $y_5 = 0{,}6833618$, ...

Esta estrategia puede extenderse a esquemas numéricos muy diferentes, combinando un esquema explícito de predicción y un esquema implícito de corrección. Por ejemplo, si se deseara utilizar un θ -método con $\theta > 0$ para la fase correctora y el método de Euler explícito para la fase predictora, el esquema de cálculo sería:

Fase predictora (Euler explícito):

$$y_{n+1}^{(0)} = y_n + h_n\, f(t_n, y_n)$$

Fase correctora (θ -método combinado con aproximaciones sucesivas):

$$y_{n+1}^{(iter+1)} = y_n + (1-\theta)\, h_n\, f(t_n, y_n) + \theta\, h_n\, f(t_{n+1}, y_{n+1}^{(iter)}) \qquad (iter = 0, 1, ...)$$

Así por ejemplo, se deja como ejercicio propuesto verificar que la estrategia de cálculo anterior con $\theta = 0{,}5$ aplicada al problema de valor inicial:

$$\left\{ \begin{array}{ll} y'(t) & = -y^2(t), \quad t > 0 \\ y(0) & = 1 \end{array} \right.$$

conduce a los valores

$$y_1 = 0{,}9087121, \qquad y_2 = 0{,}8327505, \qquad y_3 = 0{,}7685438,$$
$$y_4 = 0{,}7135529, \qquad y_5 = 0{,}6659224, ...$$

Observación 3.2.4 *De hecho algunos métodos de cálculo muy utilizados en la práctica, y que posteriormente nos reencontraremos, podrían interpretarse como un esquema predictor corrector en los que se realizan "pocas" iteraciones del método de aproximaciones sucesivas en la fase de corrección. Es el caso por ejemplo de combinar el método de Euler explícito en la fase de predicción y realizar una sóla iteración en la fase correctora utilizando el método de Crank-Nicholson, lo que nos conduce a:*

$$y_{n+1}^{(0)} = y_n + h_n\, f(t_n, y_n)$$

$$y_{n+1} = y_n + \frac{h_n}{2}\left(f(t_n, y_n) + f(t_{n+1}, y_{n+1}^{(0)})\right)$$

por lo que, resumiendo las dos etapas en una única expresión resulta:

$$y_{n+1} = y_n + \frac{h_n}{2}\left(f(t_n, y_n) + f(t_{n+1}, y_n + h_n\, f(t_n, y_n))\right)$$

Este esquema de cálculo recibe el nombre de **método de Heun** *y lo volveremos a deducir al examinar los métodos de Runge-Kutta.*

Cuarto ejemplo: Un problema regido por un sistema de ecuaciones diferenciales ordinarias

Consideraremos ahora un ejemplo que se puede encontrar en el libro de Hanna y Sandall[15])

En la ingeniería en general, y en la ingeniería química en particular, es muy frecuente toparse con problemas de valor inicial regidos por sistemas de ecuaciones diferenciales ordinarias. Como botón de muestra baste el siguiente ejemplo en el que se considera una reación química no lineal semejante a las que tienen lugar en un reactor químico durante la fase transitoria de una reacción a volumen constante de la forma:

$$A + B \leftrightarrow C \rightarrow D + E$$

donde A, B, C, D y E representan diferentes compuestos que interaccionan entre sí. En una reacción química como la planteada, la concentración de la especie A, que denotaremos por y_1, y la de la especie C que donotaremos por y_2 se relacionan entre sí mediante:

$$\begin{cases} \dfrac{dy_1}{dt} &= -k_1\, y_1\,(y_1 - K) + k_2\, y_2 \\ \dfrac{dy_2}{dt} &= -(k_2 + k_3)\, y_2 + k_1\, y_1\,(y_1 - K) \end{cases}$$

[15]Hanna, O.T. y Sandall, O.C. (1995). "Computational Methods in Chemical Engineering". Ed. Prentice Hall International Editions.

donde k_1, k_2 y k_3 son las constantes de reacción de las reacciones químicas $A + B \to C$, $C \to A + B$ y $C \to D + E$ respectivamente. Por otra parte K es una constante que depende de la composición inicial de la mezcla que se deja reaccionar. Al sistema anterior debe acompañársele de las condiciones iniciales $y_1^{(0)}$ e $y_2^{(0)}$. Si consideramos el caso en que $k_1 = k_2 = k_3 = 1$, $K = 0$, $y_1^{(0)} = 1$ e $y_2^{(0)} = 0$ el sistema resultante será:

$$\left\{ \begin{array}{ll} \mathbf{y}'(t) & = \mathbf{f}(t, \mathbf{y}(t)) \quad t > 0 \\ \mathbf{y}(0) & = \mathbf{y}^{(0)} \end{array} \right.$$

donde

$$\mathbf{y}(t) = \left\{ \begin{array}{l} y_1(t) \\ y_2(t) \end{array} \right\}, \quad \mathbf{y}'(t) = \left\{ \begin{array}{l} y_1'(t) \\ y_2'(t) \end{array} \right\}, \quad \mathbf{y}^{(0)} = \left\{ \begin{array}{l} y_1^{(0)} \\ y_2^{(0)} \end{array} \right\} = \left\{ \begin{array}{l} 1 \\ 0 \end{array} \right\}$$

y

$$\mathbf{f}(t, \mathbf{y}(t)) = \left\{ \begin{array}{l} -y_1^2(t) + y_2(t) \\ -2\, y_2(t) + y_1^2(t) \end{array} \right\}$$

Si tomamos $t_0 = 0$, $t_1 = 0{,}1$, $t_2 = 0{,}2$, $t_3 = 0{,}3$, ...La aplicación del método de Euler explícito a este sistema nos conduce a que:

$$\mathbf{y}^{(1)} = \mathbf{y}^{(0)} + 0{,}1\, \mathbf{f}(t_0, \mathbf{y}^{(0)}) = \left\{ \begin{array}{l} 1 \\ 0 \end{array} \right\} + 0{,}1 \left\{ \begin{array}{l} -1^2 + 0 \\ -2\,0 + 1^2 \end{array} \right\} = \left\{ \begin{array}{l} 0{,}9 \\ 0{,}1 \end{array} \right\}$$

$$\mathbf{y}^{(2)} = \mathbf{y}^{(1)} + 0{,}1\, \mathbf{f}(t_1, \mathbf{y}^{(1)}) = \left\{ \begin{array}{l} 0{,}9 \\ 0{,}1 \end{array} \right\} + 0{,}1 \left\{ \begin{array}{l} -(0{,}9)^2 + 0{,}1 \\ -2\,0{,}1 + (0{,}9)^2 \end{array} \right\} = \left\{ \begin{array}{l} 0{,}829 \\ 0{,}161 \end{array} \right\}$$

$$\begin{array}{rcl} \mathbf{y}^{(3)} & = & \mathbf{y}^{(2)} + 0{,}1\, \mathbf{f}(t_2, \mathbf{y}^{(2)}) = \\ & = & \left\{ \begin{array}{l} 0{,}829 \\ 0{,}161 \end{array} \right\} + 0{,}1 \left\{ \begin{array}{l} -(0{,}829)^2 + 0{,}161 \\ -2\,0{,}161 + (0{,}829)^2 \end{array} \right\} = \left\{ \begin{array}{l} 0{,}7763759 \\ 0{,}1975241 \end{array} \right\} \end{array}$$

$$\mathbf{y}^{(4)} = \mathbf{y}^{(3)} + 0{,}1\, \mathbf{f}(t_3, \mathbf{y}^{(3)}) =$$

$$\left\{ \begin{array}{l} 0{,}776375 \\ 0{,}197524 \end{array} \right\} + 0{,}1 \left\{ \begin{array}{l} -(0{,}776375)^2 + 0{,}197524 \\ -2\,0{,}197524 + +(0{,}776375)^2 \end{array} \right\} = \left\{ \begin{array}{l} 0{,}73585235 \\ 0{,}21829523 \end{array} \right\}$$

Para posteriores instantes de cálculo se van obteniendo los vectores de concentración:

$$\mathbf{y}^{(5)} = \left\{ \begin{array}{l} 0{,}7035340106 \\ 0{,}2287840561 \end{array} \right\}$$

$$\mathbf{y}^{(6)} = \left\{ \begin{array}{l} 0{,}6769164058 \\ 0{,}2325232553 \end{array} \right\}$$

$$\mathbf{y}^{(7)} = \left\{ \begin{array}{l} 0{,}6543471493 \\ 0{,}2318401863 \end{array} \right\}$$

$$\mathbf{y}^{(8)} = \left\{ \begin{array}{l} 0{,}6347141488 \\ 0{,}2282891682 \end{array} \right\}$$

$$\mathbf{y}^{(9)} = \left\{ \begin{array}{l} 0{,}6172568606 \\ 0{,}2229175396 \end{array} \right\}$$

$$\mathbf{y}^{(10)} = \left\{ \begin{array}{l} 0{,}6014480114 \\ 0{,}2164346349 \end{array} \right\}$$

Si se desease aplicar un esquema implícito, en cada paso deberá resolverse un sistema de 2 ecuaciones no lineales. Una alternativa a ello puede ser el uso del método de Heun que presentábamos en la observación realizada al ejemplo anterior, que en este caso quedaría de la forma:

$$\mathbf{y}^{(n+1)} = \mathbf{y}^{(n)} + \frac{h_n}{2} \left(\mathbf{f}(t_n, \mathbf{y}^{(n)}) + \mathbf{f}(t_{n+1}, \mathbf{y}^{(n)} + h_n \, \mathbf{f}(t_n, \mathbf{y}^{(n)})) \right)$$

y que, operacionalmente, estructuraremos de la forma siguiente:

$$\begin{array}{rcl} \mathbf{W}^{(n,1)} & = & \mathbf{f}(t_n, \mathbf{y}^{(n)}) \\ \mathbf{W}^{(n,2)} & = & \mathbf{f}(t_{n+1}, \mathbf{y}^{(n)} + h_n \, \mathbf{W}^{(n,1)}) \\ \mathbf{y}^{(n+1)} & = & \mathbf{y}^{(n)} + \dfrac{h_n}{2} \left(\mathbf{W}^{(n,1)} + \mathbf{W}^{(n,2)} \right) \end{array}$$

Todo ello (con la misma longitud de paso de integración) nos conduce en el caso del sistema considerado a que:

Estimación de la solución en $t = 0{,}1$:

$$\mathbf{W}^{(0,1)} = \mathbf{f}(t_0, \mathbf{y}^{(0)}) = \left\{ \begin{array}{l} -(1)^2 + 0 \\ -2\,0 + (1)^2 \end{array} \right\} = \left\{ \begin{array}{l} -1 \\ 1 \end{array} \right\}$$

$$\mathbf{W}^{(0,2)} = \mathbf{f}(t_1, \mathbf{y}^{(0)} + h_n \, \mathbf{W}^{(0,1)}) = \left\{ \begin{array}{l} -(1 + 0{,}1\,(-1))^2 + (0 + 0{,}1\,(1)) \\ -2\,(0 + 0{,}1\,(1)) + (1 + 0{,}1\,(-1))^2 \end{array} \right\} =$$

$$= \left\{ \begin{array}{l} -(0{,}9)^2 + 0{,}1 \\ -0{,}2 + (0{,}9)^2 \end{array} \right\} = \left\{ \begin{array}{l} -0{,}71 \\ 0{,}61 \end{array} \right\}$$

$$\mathbf{y}^{(1)} = \mathbf{y}^{(0)} + 0{,}05\,(\mathbf{W}^{(0,1)} + \mathbf{W}^{(0,2)}) = \left\{ \begin{array}{l} 0{,}914500000 \\ 0{,}080500000 \end{array} \right\}$$

Estimación de la solución en $t = 0{,}2$:

$$\mathbf{W}^{(1,1)} = \mathbf{f}(t_1, \mathbf{y}^{(1)}) = \left\{ \begin{array}{l} -(0{,}9145)^2 + 0{,}0805 \\ -2\,0{,}0805 + (0{,}9145)^2 \end{array} \right\} = \left\{ \begin{array}{l} -0{,}75581025 \\ 0{,}67531025 \end{array} \right\}$$

$$\mathbf{W}^{(1,2)} = \mathbf{f}(t_2, \mathbf{y}^{(1)} + h_n\,\mathbf{W}^{(1,1)}) =$$

$$= \left\{ \begin{array}{l} -(0{,}9145 + 0{,}1\,(-{,}7558102500))^2 + (0{,}0805 + 0{,}1\,(0{,}67531025)) \\ -2\,(0{,}0805 + 0{,}1\,(0{,}67531025)) + (0{,}9145 + 0{,}1\,(-{,}7558102500)^2) \end{array} \right\} =$$

$$= \left\{ \begin{array}{l} -({,}838918975)^2 + 0{,}148031025 \\ -2\,0{,}148031025 + ({,}838918975)^2 \end{array} \right\} = \left\{ \begin{array}{l} -0{,}5557540216 \\ 0{,}4077229966 \end{array} \right\}$$

$$y^{(2)} = \mathbf{y}^{(1)} + 0{,}05\,(\mathbf{W}^{(1,1)} + \mathbf{W}^{(1,2)}) = \left\{ \begin{array}{l} 0{,}8489217864 \\ 0{,}1346516624 \end{array} \right\}$$

Estimación de la solución en $t = 0{,}3$:

$$\mathbf{W}^{(2,1)} = \mathbf{f}(t_2, \mathbf{y}^{(2)}) = \left\{ \begin{array}{l} -(0{,}8489217864)^2 + 0{,}1346516624 \\ -2\,0{,}1346516624 + (0{,}8489217864)^2 \end{array} \right\} =$$

$$= \left\{ \begin{array}{l} -0{,}5860165370 \\ 0{,}4513648746 \end{array} \right\}$$

$$\mathbf{W}^{(2,2)} = \mathbf{f}(t_3, \mathbf{y}^{(2)} + h_n\,\mathbf{W}^{(2,1)}) =$$

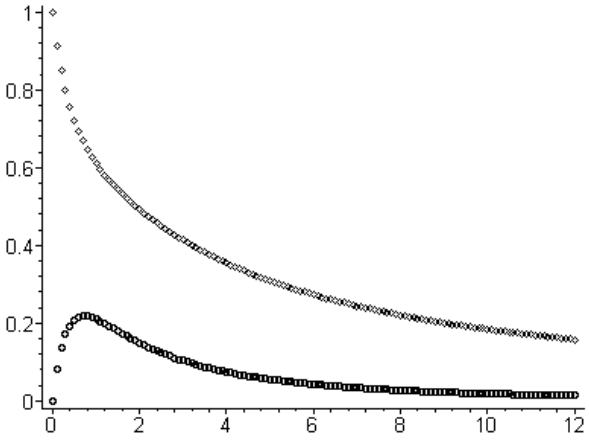
$$= \left\{ \begin{array}{l} -(0{,}84892 + 0{,}1(-0{,}58601))^2 + (0{,}18331 + 0{,}1(0{,}451364)) \\ -2(0{,}18331 + 0{,}1(0{,}45136)) + (0{,}84892 + 0{,}1(-0{,}58601)^2) \end{array} \right\} =$$

$$= \left\{ \begin{array}{l} -(0{,}7903201327)^2 + 0{,}1797881499 \\ -2\,0{,}1797881499 + (0{,}7903201327)^2 \end{array} \right\} = \left\{ \begin{array}{l} -0{,}4448177623 \\ 0{,}2650296124 \end{array} \right\}$$

$$y^{(3)} = \mathbf{y}^{(2)} + 0{,}05\,(\mathbf{W}^{(2,1)} + \mathbf{W}^{(2,2)}) = \left\{ \begin{array}{l} 0{,}7973800715 \\ 0{,}1704713868 \end{array} \right\}$$

Para posteriores instantes de cálculo se van obteniendo las soluciones:

$$\mathbf{y}^{(4)} = \left\{ \begin{array}{l} 0{,}7559223581 \\ 0{,}1934076005 \end{array} \right\}$$



$$\mathbf{y}^{(5)} = \left\{ \begin{array}{l} 0{,}7218302976 \\ 0{,}2072358839 \end{array} \right\}$$

$$\mathbf{y}^{(6)} = \left\{ \begin{array}{l} 0{,}6931987138 \\ 0{,}2146110432 \end{array} \right\}$$

$$\mathbf{y}^{(7)} = \left\{ \begin{array}{l} 0{,}6686718377 \\ 0{,}2174203031 \end{array} \right\}$$

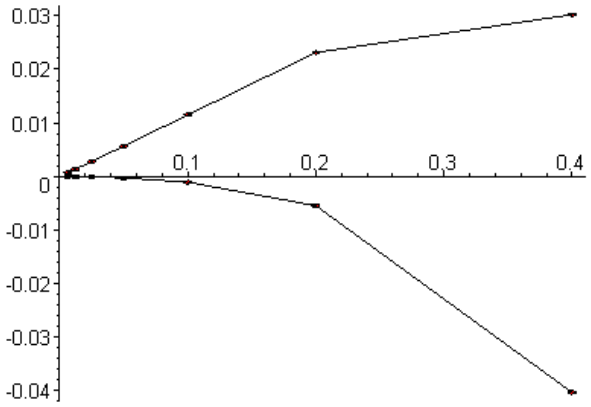
$$\mathbf{y}^{(8)} = \left\{ \begin{array}{l} 0{,}6472726418 \\ 0{,}2170160616 \end{array} \right\}$$

$$\mathbf{y}^{(9)} = \left\{ \begin{array}{l} 0{,}6282894584 \\ 0{,}2143729901 \end{array} \right\}$$

$$\mathbf{y}^{(10)} = \left\{ \begin{array}{l} 0{,}6111990088 \\ 0{,}2101961324 \end{array} \right\}$$

En la figura siguiente se representa la evolución de las concentraciones calculadas mediante este esquema de Heun hasta el tiempo $t = 12$ con paso $h = 0{,}1$.

Estos mismos esquemas pueden utilizarse para evaluar la solución con distintos pasos de tiempo. Si se comparan las soluciones obtenidas con los esquemas anteriores para diferentes pasos de tiempo con el valor de la solución exacta de $y_1(t)$ en el instante $t = 0{,}8$ (tomando $y(0{,}8) = 0{,}646234$ de la refe-



rencia Hanna y Sandall[16]) se obtendría una tabla como la siguiente:

h	Euler explícito	Heun
0,4	0,030234	−0,040291
0,2	0,023161	−0,005432
0,1	0,011520	−0,001039
0,05	0,005704	−0,000231
0,025	0,002835	−0,000055
0,0125	0,001413	−0,000013
0,00625	0,000705	−0,000003

La evolución de los errores de cada método con el tamaño de paso presentada en la tabla anterior se recoge en la gráfica siguiente (correspondiendo los valores positivos al método de Euler explícito y los negativos al de Heun):

En la evolución de los errores anterior pueden realizarse algunas consideraciones interesantes. En ambos métodos el error disminuye a medida que lo hace el tamaño de paso de integración utilizado. Pero en el caso del método de Heun, aun comenzando con un error mayor en valor absoluto para $h = 0{,}4$, el decrecimiento del error es más rápido. En efecto la reducción del error en el método de Euler es "grosso modo" lineal en el sentido de que reducir el tamaño de paso de un valor h a la mitad $h/2$ se traduce en una reducción del error a (un orden de) la mitad. En el método de Heun el error no se reduce linealmente sino que una reducción del paso a $\frac{1}{2}$ del valor que tenía se traduce en una reducción del orden de $\frac{1}{4} = \left(\frac{1}{2}\right)^2$ (o mayor) del error que se cometía. Es una reducción del error "cuadrática". Esta forma de depender el error del

[16] Hanna, O.T. y Sandall, O.C. (1995). "Computational Methods in Chemical Engineering". Ed. Prentice Hall International Editions.

tamaño del paso de integración es diferente en los distintos métodos numéricos que puedan considerarse y debe ser analizada con mayor detalle y de forma más rigurosa como se hace en el subapartado siguiente.

3.2.4. Análisis del método de Euler.

Consideremos el esquema de Euler explícito:

$$\begin{array}{rcll} & & y_0 \; dado & \\ y_{n+1} & = & y_n + h_n \, f(t_n, y_n) & (n = 0, 1, 2, ..., N-1) \end{array}$$

aplicado al problema de valor inicial:

$$(P.V.I.) \left\{ \begin{array}{ll} y'(t) & = f(t, y(t)) \quad t \in [t_0, t_N] \\ y(0) & = y_0 \end{array} \right\}$$

siendo $t_0 < t_1 < .... < t_N$ instantes preseleccionados en los que se estimará el valor y_n que aproxima al valor de la solución exacta $y(t_n)$. Denotemos, además por h al valor:

$$h = Sup_{0 \leq n \leq N-1} \{h_n\} = Sup_{0 \leq n \leq N-1} \{|t_{n+1} - t_n|\}$$

Con esta notación se tiene la siguiente definición:

Definición 3.2.1 *Se denomina* ***error de consistencia*** *del método de Euler explícito en el punto* t_{n+1} *al valor:*

$$E_{n+1} = y(t_{n+1}) - y(t_n) - h_n \, f(t_n, y(t_n)) \quad (n = 0, 1, ..., N-1)$$

Obsérvese que el error de consistencia en el instante t_{n+1} nos proporciona la diferencia entre el valor exacto de la solución en dicho punto $(y(t_{n+1}))$ y el valor que se obtendría al aproximar este valor mediante el método de Euler si se partiese de la solución exacta en el instante t_n (es decir $y(t_n) + h_n \, f(t_n, y(t_n))$). No obstante, como en la aplicación del método no se parte de este valor sino de y_n el error de consistencia, por sí solo, no nos acabará de ilustrar correctamente el funcionamiento del método. Es por ello que, como se detallará más adelante, será necesario considerar otros tipos de error (de los que el error de consistencia será sólo una de sus partes). Pero antes de ello introduzcamos ya la noción de consistencia de un método.

Definición 3.2.2 *Se dice que el método de Euler explícito es* ***consistente*** *con el problema* $(P.V.I.)$ *si se verifica que:*

$$\lim_{h \to 0} \left(\sum_{n=1}^{N} |E_n| \right) = 0$$

Además, si para cualquier valor de h positivo e inferior a un cierto paso máximo h^*, la suma de los errores de consistencia verifica que:

$$\sum_{n=1}^{N} |E_n| \leq C.h^k$$

donde C es una constante, se dice que el método de Euler explícito es ***consistente de orden*** *k.*

Proposición 3.2.1 *Si el problema de valor inicial (P.V.I.) admite una solución de clase $C^2([t_0, t_N])$, entonces el método de Euler es consistente al menos de primer orden.*

Demostración:

Desarrollando en serie de Taylor se tiene que:

$$y(t_{n+1}) = y(t_n) + h_n\, y'(t_n) + \frac{h_n^2}{2}\, y"(\xi_n) \qquad \xi_n \in [t_n, t_{n+1}]$$

de donde

$$|E_{n+1}| = |y(t_{n+1}) - y(t_n) - h_n\, f(t_n, y(t_n))| \leq K\, h\, h_n$$

donde se ha designado por K al valor:

$$K = \frac{1}{2}\text{máx}_{t\in[t_0,t_N]}\,(|y"(t)|)$$

Por tanto,

$$\sum_{n=1}^{N} |E_n| \leq K\, h \sum_{n=0}^{N-1} h_n = C\, h$$

donde $C = K(t_N - t_0)$.

c.q.d.

Nótese que la consistencia de un método con un problema de valor inicial dado no es más que una forma de "medir" en qué grado la solución exacta del problema de valor inicial satisface el esquema numérico considerado (en este caso el esquema de Euler explícito). Pero como se ha señalado anteriormente el error entre el valor calculado por el método y el valor de la solución exacta no es únicamente el error de consistencia. Por ello debemos introducir ya otro tipo de error.

Definición 3.2.3 *Se denomina* ***error*** *(de truncatura)* ***del método de Euler explícito*** *en el punto t_n al valor dado por:*

$$e_n = y(t_n) - y_n, \qquad (n = 0, 1, 2, ..., N)$$

Definición 3.2.4 *Se dice que el método de Euler explícito es* ***convergente*** *hacia la solución del problema (P.V.I.) cuando se verifica:*

$$\lim_{h\to 0}\left(\max_{0\leq n\leq N}|e_n|\right)=0$$

Como ya se apuntó anteriormente, entre el error de consistencia y el error del método existe una relación que se recoge en la siguiente proposición.

Proposición 3.2.2 *Entre el error de consistencia del método de Euler explícito en el punto t_{n+1} , E_{n+1}, y el error de dicho método en ese mismo punto, e_{n+1}, y en el punto de cálculo anterior, e_n, se verifica la siguiente relación:*

$$e_{n+1}=E_{n+1}+e_n+h_n(y'(t_n)-y'_n)$$

donde se ha denotado por y'_n al valor $f(t_n,y_n)$.

Demostración:

Según la definición de error de consistencia se verificará:

$$y(t_{n+1})=y(t_n)+h_n\,f(t_n,y(t_n))+E_{n+1}$$

Por otra parte el método de Euler explícito nos conduce a que:

$$y_{n+1}=y_n+h_n\,f(t_n,y_n)$$

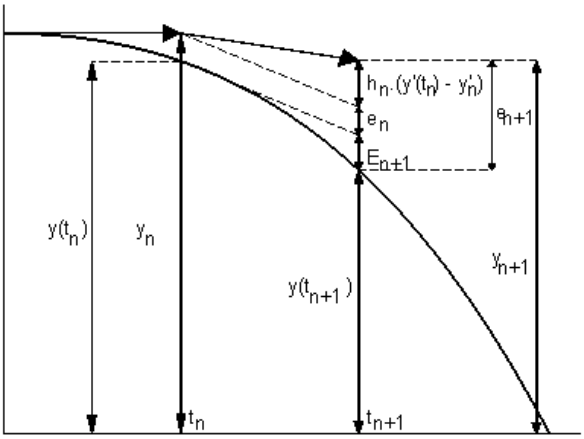
Restando ambas expresiones se tiene que:

$$y(t_{n+1})-y_{n+1}=y(t_n)-y_n+h_n\,\left(f(t_n,y(t_n))-f(t_n,y_n)\right)+E_{n+1}\Rightarrow$$
$$\Rightarrow e_{n+1}=e_n+h_n(y'(t_n)-y'_n)+E_{n+1}$$

c.q.d.

La propiedad anterior pone de manifiesto el hecho de que el error cometido en la evaluación aproximada de la solución en el punto t_{n+1} no es sólo la suma del error de consistencia en dicho punto más el que se hubiera cometido hasta la evaluación de la solución aproximada en t_n. En efecto, debe tenerse en cuenta además que, según la interpretación gráfica antes dada, la dirección de la tangente a la curva que se sigue al partir de (t_n,y_n) es diferente de la que se seguiría si se hubiese partido de $(t_n,y(t_n))$. Este hecho es el que se ilustra en la figura.

El análisis de la convergencia del método de Euler puede realizarse entonces analizando los errores de consistencia y los valores de $h_n\,(f(t_n,y(t_n))-f(t_n,y_n)$ en cada paso. Esto último nos exigirá tener que imponer alguna condición a la función f como se hace en el siguiente teorema. Pero para realizar la demostración del citado teorema será necesario tener en cuenta la propiedad que se recoge en el siguiente lema:



Lema 3.2.1 *Siendo A y B dos números reales no negativos y siendo $\{a_n\}$ una sucesión de números reales no negativos tales que satisfacen la desigualdad:*

$$a_{n+1} \leq (1+A)\, a_n + B$$

se verifica alguna de las dos desigualdades siguientes:

$$a_n \leq a_0 + n\, B \qquad (Si\ A = 0)$$

$$a_n \leq a_0\, e^{n\,A} + \frac{e^{n\,A} - 1}{A}\, B \quad (Si\ A > 0)$$

Demostración:

Si $A = 0$ la hipótesis supuesta se reduce a que:

$$a_n \leq a_{n-1} + B \leq a_{n-2} + 2\, B \leq ..... \leq a_0 + n\, B$$

por lo que en este caso el lema es evidente.

Si suponemos que $A > 0$, considerando el desarrollo en serie de la función exponencial:

$$e^A = 1 + A + \frac{A^2}{2} + \frac{A^3}{3!} + ...$$

se deduce que

$$A - e^A = -1 - \frac{A^2}{2} - \frac{A^3}{3!} - .... \leq -1 \Rightarrow A \leq e^A - 1$$

Por tanto, y habida cuenta de la hipótesis realizada sobre los elementos de la sucesión $\{a_n\}$ se tiene que:

$$a_1 \leq (1+A)\, a_0 + B \leq e^A\, a_0 + B \leq e^A\, a_0 + \frac{e^A - 1}{A}\, B$$

Luego el lema es cierto, al menos, para el caso en que $n = 1$. Procedamos entonces por inducción esto es supongamos que el lema es cierto para un valor dado de n y demostremos que entonces también es cierto para el valor $n + 1$. De esta forma si admitimos que se verifica:

$$a_n \leq a_0\, e^{n\,A} + \frac{e^{n\,A} - 1}{A}\, B$$

se tendrá que

$$a_{n+1} \leq (1 + A)\, a_n + B \leq (1 + A)\, \left(a_0\, e^{n\,A} + \frac{e^{n\,A} - 1}{A}\, B \right) + B \leq$$

$$\leq e^A\, \left(a_0\, e^{n\,A} + \frac{e^{n\,A} - 1}{A}\, B \right) + B \leq$$

$$\leq e^{(n+1)\,A}\, a_0 + \frac{e^{(n+1)\,A}}{A}\, B + \left(1 - \frac{e^A}{A} \right)\, B =$$

$$= e^{(n+1)\,A}\, a_0 + \frac{e^{(n+1)\,A}}{A}\, B + \left(\frac{A - e^A}{A} \right)\, B \leq$$

$$\leq e^{(n+1)\,A}\, a_0 + \frac{e^{(n+1)\,A}}{A}\, B - \frac{1}{A}\, B =$$

$$= e^{(n+1)\,A}\, a_0 + \frac{e^{(n+1)\,A} - 1}{A}\, B$$

c.q.d.

Procedamos ya a presentar el teorema que acota el error del método de Euler explícito.

Teorema 3.2.1 *Si la función $f(t, y)$ que interviene en la definición del problema de valor inicial (P.V.I.) es una función de clase $C^1([t_0, t_N], R)$ y es lipschitciana respecto a su segunda variable, es decir:*

$$\exists L > 0 \;\; / \;\; |f(t, y) - f(t, z)| \leq L.\, |y - z| \quad \forall t \in [t_0, t_N], \;\; \forall y, z \in R$$

entonces el método de Euler explícito inicializado con el valor $y_0 = y(t_0)$ es convergente y además se verifica que:

$$e_n \leq C(h)h \qquad (n = 0, 1, ..., N)$$

donde $C(h)$ es una constante que sólo depende del valor de h y que tiende a 0 cuando h tiende hacia 0.

Demostración:

Por ser f una función de clase C^1 la función $y(t)$ será de clase C^2 en el intervalo en el que está definida. Por ello es admisible el desarrollo en serie de Taylor:

$$y(t_{n+1}) = y(t_n) + h_n\, y'(t_n) + \frac{h_n}{2}\, y''(\xi_n) =$$

$$= y(t_n) + h_n\, f(t_n, y(t_n)) + \frac{h_n}{2}\, y''(\xi_n) \quad ,\ \xi_n \in [t_n, t_{n+1}]$$

Por otra parte, al aplicar el método de Euler explícito se tiene que:

$$y_{n+1} = y_n + h_n\, f(t_n, y_n)$$

Combinando ambas expresiones, y teniendo en cuenta que f es lipschitciana, se tiene:

$$y(t_{n+1}) - y_{n+1} = y(t_n) - y_n + h_n\ (f(t_n, y(t_n)) - f(t_n, y_n)) + \frac{h_n^2}{2}\, y''(\xi_n) \Rightarrow$$

$$\Rightarrow e_{n+1} = e_n + h_n\, L\, (y(t_n) - y_n) + \frac{h_n^2}{2}\, y''(\xi_n) \Rightarrow$$

$$\Rightarrow e_{n+1} = (1 + h_n\, L)\, e_n + \frac{h_n^2}{2}\, y''(\xi_n) \Rightarrow$$

$$\Rightarrow |e_{n+1}| \leq (1 + h\, L)\ |e_n| + h^2\, B$$

donde $h = Sup_{0 \leq n \leq N-1}\, |t_{n+1} - t_n|$ y $B = \frac{1}{2}\, Sup_{\xi \in [t_0, t_N]}\, |y''(\xi)|$. Por verificarse esta relación entre los distintos elementos de la sucesión de valores $\{|e_n|\}_{n=0}^{N}$ se tendrá en función del lema anteriormente demostrado que:

$$|e_n| \leq e^{n\, L\, h}\ |e_0| + \frac{e^{n\, L\, h} - 1}{L\, h}\, h^2\, B = e^{n\, L\, h}\ |e_0| + \frac{e^{n\, L\, h} - 1}{L}\, h\, B$$

Por tanto, si se tiene en cuenta que $e_0 = y(t_0) - y_0 = 0$ resultará que:

$$|e_n| \leq C(h)\, h$$

donde se ha denotado por $C(h)$ a la expresión:

$$C(h) = \frac{e^{n\, L\, h} - 1}{L}\, B$$

que, efectivamente, verifica:

$$\lim_{h \to 0} C(h) = B \lim_{h \to 0} - \frac{e^{n\, L\, h} - 1}{L} = 0$$

c.q.d.

El teorema anterior demuestra el comportamiento "lineal" que tenía el error en el método de Euler explícito y que se puso de manifiesto en el cuarto de los ejemplos considerado en el subapartado anterior.

Una consecuencia de lo anterior es que:

$$\text{máx}_{0\leq n\leq N}\,|e_n| \leq \frac{e^{n\,L\,h}-1}{L}\,B\,h \leq \frac{e^{L\,(t_N-t_0)}-1}{L}\,B\,h = M\,h$$

lo que pone de manifiesto el carácter lineal de la cota de error (por lo que en ocasiones se utiliza esta expresión como cota de error del método de Euler).

Observación 3.2.5 *Un tratamiento similar al realizado para el método de Euler explícito puede realizarse para el método de Euler implícito y para los θ-métodos. Dejamos al lector interesado la tarea de demostrar que, en general, estos métodos tienen una convergencia lineal salvo en el caso en que $\theta = 0{,}5$ en el que, suponiendo que la solución exacta $y(t)$ es de clase $C^3(I)$ se puede demostrar que la convergencia es cuadrática, es decir que el error del método puede acotarse mediante $|e_n| \leq C(h)h^2$ donde $C(h)$ tiende hacia 0 cuando lo hace h.*

Observación 3.2.6 *En el estudio del error que se ha realizado con los teoremas anteriores nos ha faltado un "invitado" importante: el error de redondeo. Este error será el culpable en la práctica de que los valores de y_n obtenidos por el método se conviertan realmente en $\widetilde{y}_n$ al no poderse almacenar, trabajando con un número finito de decimales, los números de forma exacta. De hecho, en el último teorema demostrado se suponía como hipótesis que el método de Euler se inicializaba con el valor $y_0 = y(t_0)$. Y ni esto será posible en numerosas aplicaciones (tómese por ejemplo $y(t_0) = \sqrt{2}$). Ello hará que el método se inicialice con $\widetilde{y}_0 = y_0 + \delta_0$ dependiendo el valor de δ_0 de la precisión que tenga la máquina con que se trabaje (del número de decimales con los que permita trabajar). Con ello, la primera etapa del método de Euler consistirá en:*

$$\widetilde{y}_1 = y_0 + h_n\,f(t_0, y_0) + h_n\,\varepsilon_0 + \delta_0$$

donde $\varepsilon_0 = f(t_0, y_0 + \delta_0) - f(t_0, y_0)$. De esta forma en una etapa genérica del método se estará calculando, en lugar de y_{n+1} el valor:

$$\widetilde{y}_{n+1} = y_n + h_n\,f(t_n, y_n) + h_n\,\varepsilon_n + \delta_n$$

Admitiendo que $|\delta_n| \leq \delta$ y que $|\varepsilon_n| \leq \varepsilon$ para todo $h_n \leq h^*$, la acotación de error del teorema anterior debe modificarse resultando:

$$\max_{0\leq n\leq N}\,|y(t_n) - \widetilde{y}_n| \leq \frac{e^{L\,(t_N-t_0)}-1}{L}\,\left(B.h + \varepsilon + \frac{\delta}{h}\right)$$

La cota anterior pone de manifiesto el hecho de que disminuciones excesivas del tamaño del paso de integración pueden ser "contraproducentes" en el sentido de que el error realmente cometido con el método aumenta para disminuciones del paso de integración.

3.2.5. El control del tamaño del paso de integración

El análisis del método de Euler realizado en el apartado anterior parece indicarnos que cuanto más se reduzca el paso de integración (al menos hasta un límite inferior h_* determinado por la precisión de la máquina con la que se trabaje) más preciso será el método. Una conclusión similar podría obtenerse para todos los métodos hasta ahora planteados. Pero, obviamente, la reducción del paso no puede ser arbitrariamente pequeña pues impediría en la práctica resolver problemas planteados en intervalos "grandes". Además, debe advertirse que en la acotación del error realizada se ha sido bastante "grosero" tomando, por ejemplo, como constante B en la acotación la mitad del mayor valor de $y''(t)$ en todo el intervalo de trabajo.

Es por ello que, en la práctica, los programas que recogen el método de Euler no trabajan con el simple algoritmo que hasta ahora hemos utilizado sino que además ajustan el tamaño del paso de integración permitiendo que en dominios en los que la solución es relativamente "suave" se tomen pasos de integración "grandes" en tanto que en zonas en las que la solución presente cambios de pendiente bruscos ajustan el tamaño del paso de integración a pequeños valores.

Una de las técnicas que permite realizar este ajuste de paso se basa en técnicas de extrapolación y concretamente en la denominada **extrapolación de Richardson**. Veamos en qué consiste su aplicación al método de Euler.

En primer lugar los tamaños de paso de integración a utilizar en cada etapa se toman de la forma:

$$h_n = \mu(t_n)\, h^*$$

donde h^* es un tamaño de paso de integración máximo y $\mu(t)$ es una función continua a trozos que toma valores positivos y tales que

$$0 < \mu_* < \mu(t) \leq 1 \qquad \forall t \in [t_0, t_N]$$

Con ello la técnica de control del tamaño de paso se basa en el siguiente teorema:

Teorema 3.2.2 *Suponiendo que:*

a) el problema de valor inicial:

$$\begin{cases} y'(t) & = f(t, y(t)), \quad t \in [t_0, t_N] \\ y(t_0) & = y_0 \end{cases}$$

admite una solución $y(t)$,

b) f es una función de clase $C^2(]t_0, t_N[\times \mathbb{R})$ lipschitciana respecto a su segunda variable, y

c) el método de Euler se inicializa para cada tamaño de paso h con un valor $y_{0,h}$ satisfaciéndose:

$$\exists \alpha \in \mathbb{R} \,/ \quad y(t_0) - y_{0,h} = \alpha\, h + O(h^2)$$

entonces se verifica que:

$$y(t_n) - y_{n,h} = h\, g(t_n) + O(h^2)$$

donde $y_{n,h}$ es la solución aproximada obtenida mediante el método de Euler con un tamaño de paso h, y $g(t)$ es solución del problema de valor inicial:

$$\begin{cases} g'(t) & = g(t)\, \dfrac{\partial f}{\partial y}(t, y(y)) + \dfrac{1}{2}\, \mu(t)\, y''(t), \quad t \in [t_0, t_N] \\ g(0) & = \alpha \end{cases}$$

Demostración:

Consúltese Crouzeix y Mignot[17].

Supongamos entonces que con un tamaño de paso h se ha obtenido una aproximación $y_{n,h}$ del valor de la solución en el instante de cálculo t_n. Según el teorema precedente se tendrá que:

$$y(t_n) = y_{n,h} + h\, g(t_n) + O(h^2)$$

Si en el mismo punto t_n se calculase la aproximación mediante el método de Euler utilizando un tamaño de paso de longitud $q\,h$ se tendría que:

$$y(t_n) = y_{m,q\,h} + q\, h\, g(t_n) + O(h^2)$$

Combinando las dos expresiones anteriores se obtiene que:

$$\frac{q\, y_{n,h} - y_{m,q\,h}}{q-1} = y(t_n) + O(h^2)$$

y que

$$y_{n,h} - y(t_n) = \frac{y_{m,q\,h} - y_{n,h}}{q-1} + O(h^2)$$

La última de las expresiones obtenidas nos permite considerar que:

$$\frac{y_{m,q\,h} - y_{n,h}}{q-1}$$

[17] Crouzeix, M., Mignot A.L. (1984) "Analyse numérique des équations différentielles", Ed. Masson.

es una estimación del error cometido al actuar con paso h.

Con estas consideraciones una técnica para controlar automáticamente el paso de cálculo consiste en calcular el valor $y_{n+1,h}$ a partir de y_n (es decir con un tamaño de paso h_n) y calcular $y_{n+1,h'}$ a partir de y_{n-1} (es decir con tamaño de paso $h' = h_{n-1} + h_n$) denotando por q al valor:

$$q = \frac{h_n + h_{n+1}}{h_n}$$

Si, siendo ε una tolerancia de error admisible y dada por el usuario, se verificase que:

$$\left| \frac{y_{n+1,h'} - y_{n+1,h}}{q-1} \right| \leq \varepsilon$$

el error cometido entre la aproximación obtenida con paso h y la solución exacta es un error aceptable y podría intentar operarse con un tamaño de paso $q\,h_n$ para etapas posteriores del método. Además el valor aproximado puede mejorarse mediante:

$$\widehat{y}_{n+1} = y_{n+1,h} - \frac{y_{n+1,h'} - y_{n+1,h}}{q-1} = \frac{q\,y_{n+1,h} - y_{n+1,h'}}{q-1}$$

Si por el contrario resultara que:

$$\left| \frac{y_{n+1,h'} - y_{n+1,h}}{q-1} \right| > \varepsilon$$

entonces el tamaño de paso h_n no nos garantiza un error aceptable en la estimación de la solución aproximada obtenida por el método de Euler. Ello aconsejará reducir el tamaño del paso utilizado (por ejemplo a la mitad) pasando a denominarse t_{n+1} al punto $t_n + \frac{h_n}{2}$ considerando entonces que h_n toma un valor igual a la mitad que el que tenía. Con este nuevo tamaño de paso puede volver a repetirse el mismo proceso.

3.3. Estudio de un método general de pasos libres

Consideremos nuevamente un problema de valor inicial que formularemos abreviadamente por:

$$(P.V.I.) \left\{ \begin{array}{lll} y'(t) & = f(t, y(t)) & t \in [t_0, t_N] \\ y(t_0) & = y_0 & \end{array} \right\}$$

y consideremos los puntos de cálculo:

$$t_0 < t_1 < t_2 < .... < t_n < ..... < t_N$$

denotando por h_n a los valores $h_n = (t_{n+1} - t_n)$ $(n = 0, 1, ..., N-1)$ y siendo:

$$h = Sup - 0 \leq n < N \{h_n\}$$

Con esta notación, un método de pasos libres genérico consiste en un esquema de cálculo que, partiendo del valor $y_0 = y(t_0)$ permite estimar los valores aproximados de la solución del problema $(P.V.I.)$ mediante:

$$y_{n+1} = y_n + h_n\, g(t_n, y_n, h_n) \tag{3.3}$$

donde g es una función continua en $[t_0, t_N] \times I\!R \times]0, H]$ y que está definida a partir de la función f que interviene en $(P.V.I.)$, e y_n representa los valores calculados por el método que aproximan el valor exacto de $y(t_n)$.

Ejemplo 3.3.1 *En el método de Euler explícito:*

$$y_{n+1} = y_n + h_n\, f(t_n, y_n)$$

se tiene que $g(t, y, h) = f(t, y)$.

Ejemplo 3.3.2 *Si se considerase el método de Euler implícito:*

$$y_{n+1} - h_n\, f(t_n + h_n, y_{n+1}) = y_n$$

el valor de y_{n+1} *podrá expresarse también en la forma anterior, si bien ahora la función* g *debería buscarse, en general, mediante la relación implícita que define la ecuación no lineal anterior.*

Sobre un método genérico de un paso pueden introducirse los mismos conceptos que se introdujeron al analizar el método de Euler. Así se tiene que:

Definición 3.3.1 *Se denomina* ***error de consistencia*** *del método (3.3) en el punto* t_{n+1} *al valor:*

$$E_{n+1} = y(t_{n+1}) - y(t_n) - h_n\, g(t_n, y(t_n), h_n)$$

Esta definición del error de consistencia nos permite observar que por tal concepto se entiende el error que se cometería con el método si en el instante t_n se partiera de la solución exacta en lugar de la solución aproximada y, por tanto, sin tener en consideración los errores cometidos en etapas anteriores del método. Puesto que la solución aproximada verifica:

$$y_{n+1} - y_n - h_n\, g(t_n, y_n, h_n) = 0$$

la definición dada de error de consistencia puede interpretarse también como un indicador de en qué medida la solución exacta del problema de valor inicial satisface el esquema de cálculo, es decir de hasta que punto el esquema utilizado puede reemplazar a la EDO del problema $(P.V.I.)$ o, lo que es lo mismo, si el esquema es consistente o no con la EDO dada. Es por ello que es natural considerar también la siguiente definición:

Definición 3.3.2 *Se dice que el método (3.3) es* ***consistente*** *con la ecuación diferencial del problema (P.V.I.) cuando para toda solución de dicha ecuación diferencial se verifica que:*

$$\lim_{h\to 0}\left(\sum_{n=1}^{N}|E_n|\right)=0$$

Asimismo se dice que el método (3.3) es ***consistente de orden*** k *con la EDO del problema (P.V.I.) cuando para todo valor de h perteneciente a un intervalo $]0,H[$ se verifica:*

$$\exists C\in I\!R/\ \sum_{n=1}^{N}|E_n|\leq C\,h^k$$

donde la constante C depende sólamente de $y(t)$ y de $f(t,y(t))$.

Cuanto mayor sea el orden de consistencia de los métodos numéricos, en general, mayores tamaños de paso de discretización h podrán tomarse para asegurar una determinada precisión. No obstante también es habitual que el incremento en el orden de consistencia vaya acompañado de un mayor coste computacional. Pero, al igual que se señaló para el método de Euler, no todo el error es el error de consistencia. Debe tenerse en cuenta que al evaluar y_{n+1} no se parte del valor exacto $y(t_n)$ sino de uno aproximado y_n. En este sentido se debe definir un nuevo error conocido con el nombre de error del método. Ello es lo que hacemos en la definición siguiente:

Definición 3.3.3 *Se denomina* ***error del método*** *(3.3) en el punto t_n al valor:*

$$e_n=y(t_n)-y_n\qquad(n=0,1,....N)$$

El error de un método nos representa el error total existente en un punto de cálculo entre la solución exacta y la aproximada (a falta de la consideración de los errores de redondeo debidos a una representación numérica con un número finito de decimales en las máquinas de calcular). En este sentido también es natural dar la siguiente definición:

Definición 3.3.4 *Se dice que el método (3.3) es* ***convergente*** *cuando se verifica que:*

$$\lim_{h\to 0}\left(Sup_{0\leq n\leq N}\{|e_n|\}\right)=0$$

Entre el error de consistencia y el error del método existe una relación pues el primero es una parte del segundo como se demuestra en la propiedad siguiente:

Proposición 3.3.1 *Si la función $g(t,y,h)$ que interviene en la definición del método (3.3) es una función lipschitciana respecto a su segunda variable, es decir:*

$$\exists L \in I\!R / \left| g(t,y,h) - g(t,z,h) \right| \leq L \left| y - z \right| \ \forall t \in [t_0, t_N], \forall h \in]0, H[, \forall y, z \in I\!R$$

entonces se verifica la siguiente relación entre el error de consistencia y el error del método en cada punto de cálculo:

$$e_{n+1} \leq (1 + h_n L)\, e_n + E_{n+1}$$

Demostración:

De la definición de error de consistencia se tiene que:

$$y(t_{n+1}) = y(t_n) + h_n\, g(t_n, y(t_n), h_n) + E_{n+1}$$

y de (3.3) se tiene que:

$$y_{n+1} = y_n + h_n\, g(t_n, y_n, h_n)$$

de donde, restando la segunda igualdad de la primera resultará:

$$e_{n+1} = e_n + h_n \left(g(t_n, y(t_n), h_n) - g(t_n, y_n, h_n) \right) + E_{n+1}$$

de donde, tomando valores absolutos y considerando la hipótesis de Lipschitz hecha en el enunciado, resulta:

$$e_{n+1} \leq (1 + h_n L)\, e_n + E_{n+1}$$

c.q.d.

La expresión anterior nos permitiría realizar, de forma similar a como se hizo para el método de Euler explícito, una acotación del error del método. Pero como ya se señaló anteriormente esta acotación del error dejaría fuera la incidencia de los errores de redondeo en el funcionamiento del esquema de cálculo. Es por ello que en el análisis de los métodos conviene introducir un nuevo concepto del que hasta ahora sólo habíamos hablado ilustrándolo sobre ejemplos. Este es el concepto de estabilidad que pasamos a definir de una forma más rigurosa.

Definición 3.3.5 *Se dice que el método (3.3) es* ***estable*** *cuando para cualquier terna de sucesiones $\{y_n\}_{n=0}^N$, $\{z_n\}_{n=0}^N$ y $\{E_n\}_{n=1}^N$ verificando las relaciones:*

$$y_{n+1} = y_n + h_n\, g(t_n, y_n, h_n)$$

$$z_{n+1} = z_n + h_n\, g(t_n, z_n, h_n) + E_{n+1}$$

se puede encontrar una constante M *tal que:*

$$Sup_{0\leq n\leq N}\{|z_n - y_n|\} \leq M\left[|z_0 - y_0| + \sum_{n=1}^{N}|E_n|\right] \qquad \forall h \in]0, H]$$

Antes de utilizar esta definición de estabilidad interpretemos qué nos está diciendo. Por una parte, si un método es estable y consistente en el sentido de la definición que acaba de darse se puede afirmar que si la solución exacta de $(P.V.I.)$ no "explota" la solución aproximada tampoco "explotará" pues su diferencia con la exacta estará acotada por una cantidad finita. Esta interpretación es la que utilizábamos en alguno de los ejemplos con los que ilustrábamos el comportamiento del método de Euler. Pero, por otra parte, si en la definición dada se toma como $z_n = y(t_n)$ (es decir el valor exacto de la solución), como y_n el valor aproximado y por E_n se designa al error de consistencia del método, el que un método sea estable querrá decir que el máximo error cometido en uno de los puntos de cálculo t_n puede acotarse por M veces la suma de los errores de consistencia más el error (de redondeo) que se cometa con la estimación del valor inicial. Por ello si se logra demostrar que un método es estable, bastará con demostrar que además es consistente para asegurar la convergencia del método. Esta lectura que hemos hecho de la definición de estabilidad es la que se recoge en el siguiente teorema:

Teorema 3.3.1 *Si el método definido en (3.3) es un método estable y consistente entonces si el esquema de cálculo se inicializa con el valor* $y_0 = y(t_0)$ *el método también es convergente.*

Demostración:

De la definición del esquema (3.3) y de la definición de error de consistencia se tiene que:

$$y_{n+1} = y_n + h_n\, g(t_n, y_n, h_n)$$

$$y(t_{n+1}) = y(t_n) + h_n\, g(t_n, y(t_n), h_n) + E_{n+1}$$

por lo que si el método fuese estable se deduciría que:

$$\exists M/\ Sup_{0\leq n\leq N}\{|y(t_n) - y_n|\} \leq M\left[|y(t_0) - y_0| + \sum_{n=1}^{N}|E_n|\right] \qquad \forall h \in]0, H]$$

por lo que si el método se inicializa con el valor $y_0 = y(t_0)$ la consistencia implica:

$$\lim_{h\to 0}\left(Sup_{0\leq n\leq N}\left\{|y(t_n)-y_n|\right\}\right)\leq \lim_{h\to 0}\left(\sum_{n=1}^{N}|E_n|\right)=0$$

lo que demuestra el teorema.

c.q.d.

Observación 3.3.1 *Obsérvese que, inversamente, la convergencia del método garantiza también la consistencia pues de la relación que obtuvimos en la propiedad que relacionaba el error de consistencia con el error del método:*

$$e_{n+1}=e_n+h_n\left(g(t_n,y(t_n),h_n)-g(t_n,y_n,h_n)\right)+E_{n+1}$$

puede inferirse que

$$\sum_{n=1}^{N}|E_n|=\left|e_N-e_0+\sum_{n=0}^{N}h_n\left(g(t_n,y(t_n),h_n)-g(t_n,y_n,h_n)\right)\right|$$

de donde, considerando que la convergencia del método implica que el mayor de los errores del método en cada punto tiende hacia 0 *con* h *resulta que:*

$$\sum_{n=1}^{N}|E_n|\leq|e_N|+|e_0|+\sum_{n=0}^{N}h\ |g(t_n,y(t_n),h_n)-g(t_n,y_n,h_n)|\Rightarrow$$

$$\Rightarrow \lim_{h\to 0}\left(\sum_{n=1}^{N}|E_n|\right)\leq\lim_{h\to 0}\left(|e_N|+|e_0|\right)=0$$

Asimismo, la convergencia también garantiza la estabilidad del esquema pues si la mayor de las diferencias entre la solución exacta y la solución aproximada tiende a 0 *con* h*, la solución aproximada, cuando* h *tienda a* 0 *permanecerá acotada siempre lo haga la solución exacta.*

El teorema anterior nos marca el camino usual para estudiar la convergencia de los métodos numéricos: analizar la consistencia y la estabilidad de los mismos. En este sentido presentamos a continuación algunos teoremas que pueden facilitarnos este tipo de estudios.

Teorema 3.3.2 *Una condición necesaria y suficiente para que el método dado por (3.3) sea consistente con la EDO del problema (P.V.I.) es que se verifique:*

$$g(t,y,0)=f(t,y)$$

Demostración:

A) Demostremos en primer lugar que la condición dada es suficiente. Para ello supongamos que se verifica dicha condición demostremos que entonces el esquema es consistente. En efecto, de la EDO de $(P.V.I.)$ se tiene que:

$$y'(t) = f(t, y(t)) \Rightarrow \lim_{h_n \to 0} \frac{y(t_n + h_n) - y(t_n)}{h_n} = f(t_n, y(t_n)) \Rightarrow$$

$$\lim_{h \to 0} [y(t_n + h_n) - y(t_n) - h_n\, f(t_n, y(t_n))] = 0$$

por lo que si se verifica la condición expresada en el enunciado se puede escribir que:

$$\lim_{h \to 0} \left(\sum_{n=1}^{N} |E_n| \right) = \lim_{h \to 0} \left(\sum_{n=0}^{N-1} |y(t_n + h_n) - y(t_n) - h_n\, g(t_n, y(t_n), h_n)| \right) \leq$$

$$\leq \sum_{n=0}^{N-1} \lim_{h \to 0} |y(t_n + h_n) - y(t_n) - h_n\, g(t_n, y(t_n), h_n)| =$$

$$= \sum_{n=0}^{N-1} \lim_{h \to 0} |y(t_n + h_n) - y(t_n) - h_n\, f(t_n, y(t_n))| = 0$$

lo que asegura la consistencia del método.

B) Demostremos ahora que, además, la condición es necesaria. Para ello supongamos que el método es consistente y demostremos que entonces se ha de verificar la condición del enunciado. En efecto, puesto que:

$$E_{n+1} = y(t_{n+1}) - y(t_n) - h_n\, g(t_n, y(t_n), h_n)$$

la aplicación del teorema de incrementos finitos nos asegura que:

$$\exists \xi_n \in]t_n, t_{n+1}[\ / \ E_{n+1} = h_n \ (f(t_n, y(\xi_n)) - g(t_n, y(t_n), h_n))$$

por lo que,

$$E_{n+1} = h_n \ (f(t_n, y(\xi_n)) - g(t_n, y(t_n), 0) + g(t_n, y(t_n), 0) - g(t_n, y(t_n), h_n)) =$$

$$= \alpha_n + h_n\, \beta_n$$

donde

$$\alpha_n = h_n \ (f(t_n, y(\xi_n)) - g(t_n, y(t_n), 0))$$

y

$$\beta_n = (g(t_n, y(t_n), 0) - g(t_n, y(t_n), h_n))$$

Recordando el concepto de integral de Riemann resulta que:

$$\lim_{h\to 0}\left(\sum_{n=0}^{N-1}|\alpha_n|\right)=\int_{t_0}^{t_N}|f(\xi,y(\xi))-g(\xi,y(\xi),0)|\;d\xi$$

y, por otra parte

$$|\beta_n|\leq\beta(h)=Sup_{0\leq\mu\leq h,\,|t-\tau|\leq h}\left\{|g(t,y(t),0)-g(\tau,y(\tau),\mu)|\right\}$$

de donde, por continuidad de la función g se tiene que:

$$\lim_{h\to 0}\beta(h)=0$$

por lo que

$$\sum_{n=0}^{N-1}|h_n\,\beta_n|\leq(t_N-t_0)\,\beta(h)\Rightarrow\lim_{h\to 0}\left(\sum_{n=0}^{N-1}|h_n\,\beta_n|\right)=0$$

En resumen:

$$\lim_{h\to 0}\left(\sum_{n=0}^{N-1}|E_{n+1}|\right)=\int_{t_0}^{t_N}|f(\xi,y(\xi))-g(\xi,y(\xi),0)|\;d\xi$$

por lo que si se supone que el método es consistente se debe verificar, por la continuidad de g que:

$$0=\int_{t_0}^{t_N}|f(\xi,y(\xi))-g(\xi,y(\xi),0)|\;d\xi\Rightarrow f(t,y(t))=g(t,y(t),0)$$

c.q.d.

Antes de demostrar las condiciones que aseguran la estabilidad de un método genérico de pasos libres, demostremos un lema previo que utilizaremos más adelante.

Lema 3.3.1 *Siendo A una constante positiva y dadas tres sucesiones de números reales no negativos $\{a_n\}_{n=0}^{N}$, $\{b_n\}_{n=1}^{N}$ y $\{h_n\}_{n=0}^{N-1}$ satisfaciendo la relación:*

$$a_{n+1}\leq(1+A\,h_n)\,a_n+b_{n+1}$$

se verifica para todo valor de n entre 0 y $N-1$ que:

$$a_{n+1}\leq e^{A\,(t_{n+1}-t_0)}\,a_0+\sum_{i=0}^{n}e^{A\,(t_{n+1}-t_{t+1})}\,b_{i+1}$$

donde se ha denotado por $t_n=t_0+\sum_{j=0}^{n-1}h_j$, siendo t_0 un valor arbitrario.

Demostración:

Para $n = 0$ la hipótesis realizada sobre las sucesiones nos conduce a que:

$$a_1 \leq (1 + A\,h_0)\,a_0 + b_1 \leq e^{A\,h_0}\,a_0 + b_1 = e^{A\,(t_1 - t_0)}\,a_0 + \sum_{i=0}^{0} e^{A\,(t_1 - t_{i+1})}\,b_1$$

por lo que para $n = 0$ se verifica el lema.

Procedamos ahora por inducción, suponiendo que el lema se verifica para un cierto valor de $(n-1)$ y demostremos que entonces también se verifica para el valor n. En efecto si el lema es cierto para $(n-1)$ se tendrá que siendo $n > 1$:

$$a_n \leq e^{A\,(t_n - t_0)}\,a_0 + \sum_{i=0}^{n-1} e^{A\,(t_n - t_{t+1})}\,b_{i+1}$$

por lo que, utilizando la hipótesis realizada sobre las sucesiones, se tiene que:

$$a_{n+1} \leq (1 + A\,h_n)\,a_n + b_{n+1} \leq$$

$$\leq (1 + A\,h_n)\left(e^{A\,(t_n - t_0)}\,a_0 + \sum_{i=0}^{n-1} e^{A\,(t_n - t_{t+1})}\,b_{i+1}\right) + b_{n+1} =$$

$$= (1 + A\,h_n)\,e^{A\,(t_n - t_0)}\,a_0 + (1 + A\,h_n)\sum_{i=0}^{n-1}\left(e^{A\,(t_n - t_{t+1})}\,b_{i+1}\right) + b_{n+1} \leq$$

$$\leq e^{A\,(t_{n+1} - t_0)}\,a_0 + e^{A\,(t_{n+1} - t_n)}\sum_{i=0}^{n-1}\left(e^{A\,(t_n - t_{t+1})}\,b_{i+1}\right) + b_{n+1} =$$

$$= e^{A\,(t_{n+1} - t_0)}\,a_0 + \sum_{i=0}^{n-1}\left(e^{A\,(t_{n+1} - t_{t+1})}\,b_{i+1}\right) + e^{A\,(t_{n+1} - t_{n+1})}\,b_{n+1} =$$

$$= e^{A\,(t_{n+1} - t_0)}\,a_0 + \sum_{i=0}^{n}\left(e^{A\,(t_{n+1} - t_{t+1})}\,b_{i+1}\right)$$

c.q.d.

Con ayuda del lema previo podemos abordar ya el estudio de las condiciones que garantizan la estabilidad del método general de un paso. Ello lo hacemos en el siguiente teorema:

Teorema 3.3.3 *Una condición suficiente para que el método dado por la expresión (3.3) sea estable es que g sea lipschitciana respecto a su segunda variable, es decir que:*

$$\exists L \in I\!R / \, |g(t,y,h) - g(t,z,h)| \leq L \, |y - z| \;\; \forall t \in [t_0, t_N] \, , \forall h \in]0, H[\, , \forall y, z \in I\!R$$

Además en dicho caso, la constante M que interviene en la definición de estabilidad dada en la definición (3.3.5) está dada por:

$$M = e^{L\,(t_N - t_0)}$$

Demostración:

Supongamos que $\{y_n\}_{0 \leq n \leq N}$, $\{z_n\}_{0 \leq n \leq N}$, y $\{E_n\}_{1 \leq n \leq N}$ son tres sucesiones verificando:

$$y_{n+1} = y_n + h_n \, g(t_n, y_n, h_n)$$

$$z_{n+1} = z_n + h_n \, g(t_n, z_n, h_n) + E_{n+1}$$

Se tiene entonces que:

$$|y_{n+1} - z_{n+1}| \leq |y_n - z_n| + h_n \, |g(t_n, y_n, h_n) - g(t_n, z_n, h_n)| + |E_{n+1}|$$

y por ser g lipschitciana respecto a su segunda variable:

$$|y_{n+1} - z_{n+1}| \leq (1 + L\, h_n) \, |y_n - z_n| + |E_{n+1}|$$

por lo que aplicando el lema precedente con $A = L$, $a_n = |y_n - z_n|$ y $b_n = |E_n|$ se deduce que:

$$|y_{n+1} - z_{n+1}| \leq e^{L\,(t_{n+1} - t_0)} \, |y_0 - z_0| + \sum_{i=0}^{n} e^{L\,(t_{n+1} - t_{i+1})} \, |E_{n+1}| \quad (0 \leq n \leq N-1)$$

De la expresión anterior se obtiene:

$$|y_{n+1} - z_{n+1}| \leq e^{L\,(t_{n+1} - t_0)} \, |y_0 - z_0| + e^{L\,(t_{N+1} - t_0)} \sum_{i=0}^{n} |E_{n+1}| \quad (0 \leq n \leq N-1) \Rightarrow$$

$$|y_{n+1} - z_{n+1}| \leq e^{L\,(t_{n+1} - t_0)} \left(|y_0 - z_0| + \sum_{i=0}^{n} |E_{n+1}| \right) \quad (0 \leq n \leq N-1) \Rightarrow$$

$$Sup_{0 \leq n \leq N} \, |y_n - z_n| \leq e^{L\,(t_N - t_0)} \left(|y_0 - z_0| + \sum_{i=0}^{n} |E_{n+1}| \right)$$

c.q.d.

Los dos teoremas anteriores se pueden resumir en el siguiente:

Teorema 3.3.4 *Dado el problema de valor inicial:*

$$(P.V.I.)\begin{cases} y'(t) &= f(t,y(t)), \qquad t \in [t_0, t_0+T] \\ y(t_0) &= y_0 \end{cases}$$

y considerando el método numérico de pasos libres:

$$\begin{aligned} & y_0 \quad dado \\ y_{n+1} &= y_n + h_n\, g(t_n, y_n, h_n) \qquad (n = 0, 1, ...N-1) \end{aligned}$$

entonces, bajo las hipótesis:

$$\begin{array}{l} 1^a) g(t,y,0) = f(t,y) \\ 2^a) \exists L \in I\!R / \, |g(t,y,h) - g(t,z,h)| \leq L\, |y-z| \\ \forall t \in [t_0, t_N], \forall h \in]0, H[, \forall y, z \in I\!R \end{array}$$

el método de pasos libres es convergente.

Demostración:

Es una consecuencia inmediata de los dos teoremas anteriores.

c.q.d.

Teorema 3.3.5 *Bajo las hipótesis:*

$$\begin{array}{l} 1^a) g(t,y,0) = f(t,y) \\ 2^a) \exists L \in I\!R / \, |g(t,y,h) - g(t,z,h)| \leq L\, |y-z| \\ \forall t \in [t_0, t_N], \forall h \in]0, H[, \forall y, z \in I\!R \end{array}$$

el error del método de pasos libres (3.3) puede acotarse mediante la expresión:

$$|y(t_n) - y_n| \leq e^{L\,(t_n - t_0)}\, |y(t_0) - y_0| + C(h)\, \frac{1}{L} \left(e^{L\,(t_n - t_0)} - 1 \right)$$

donde $C(h)$ sólo depende de $h = Sup_{0 \leq n \leq N-1} \{h_n\}$.

Demostración:

Nótese que si en el teorema 3.3.3 se considera $z_n = y(t_n)$, entonces se tiene que para $0 \leq n \leq N-1$,

$$|y(t_{n+1}) - y_{n+1}| \leq e^{L\,(t_{n+1} - t_0)}\, |y(t_0) - y_0| + \sum_{i=0}^{n} e^{L\,(t_{n+1} - t_{i+1})}\, |E_{n+1}|$$

donde E_{n+1} es el error de consistencia que está dado por:

$$E_{n+1} = y(t_{n+1}) - y(t_n) - h_n\, g(t_n, y(t_n), h_n)$$

Por otra parte, habida cuenta de la EDO que define el problema de valor inicial y aplicando el teorema de incrementos finitos se tiene que:

$$\exists \xi_n \in]t_n, t_{n+1}[\ / \ y(t_{n+1}) - y(t_n) = h_n \, f(\xi_n, y(\xi_n))$$

por lo que:

$$\exists \xi_n \in]t_n, t_{n+1}[\ / \ E_{n+1} = h_n \, [f(\xi_n, y(\xi_n)) - g(t_n, y(t_n), h_n)] \Rightarrow$$

$$\Rightarrow \qquad \exists \xi_n \in]t_n, t_{n+1}[\ / \ E_{n+1} = h_n \, (f(\xi_n, y(\xi_n)) - g(\xi_n, y(\xi_n), 0)) +$$

$$+ h_n \, (g(\xi_n, y(\xi_n), 0) - g(t_n, y(t_n), h_n))$$

y por verificarse la primera de las hipótesis supuestas en el enunciado:

$$\exists \xi_n \in]t_n, t_{n+1}[\ / \ E_{n+1} = h_n \, (g(\xi_n, y(\xi_n), 0) - g(t_n, y(t_n), h_n))$$

de donde, denotando por

$$C(h) = Sup_{\{0 \leq h' \leq h, \, |t-\xi| \leq h\}} \left[\left| g(t, y(t), 0) - g(\xi, y(\xi), h') \right| \right]$$

resulta que

$$|E_{n+1}| \leq C(h) h_n$$

Volviendo ahora a nuestra desigualdad inicial, para todo valor de n comprendido entre 0 y $N-1$ se tiene que:

$$|y(t_{n+1}) - y_{n+1}| \leq e^{L\,(t_{n+1}-t_0)} \, |y(t_0) - y_0| + C(h) \sum_{i=0}^{n} e^{L\,(t_{n+1}-t_{i+1})} \, h_i \leq$$

$$\leq e^{L\,(t_{n+1}-t_0)} \, |y(t_0) - y_0| + C(h) \sum_{i=0}^{n} \int_{t_i}^{t_{i+1}} e^{L\,(t_{n+1}-t)} dt \ \Rightarrow$$

$$\Rightarrow |y(t_{n+1}) - y_{n+1}| \leq e^{L\,(t_{n+1}-t_0)} \, |y(t_0) - y_0| + C(h) \, \frac{e^{L\,(t_{n+1}-t_0)} - 1}{L}$$

c.q.d.

Una vez analizada la convergencia de un método general de pasos libres podemos presentar un teorema que nos permite determinar de manera sencilla el orden de convergencia de método (3.3).

Teorema 3.3.6 *Suponiendo que la función f que interviene en el problema (P.V.I) es una función k veces continuamente diferenciable en $]t_0, t_N[\times\mathbb{R}$ y siendo la función g que interviene en la definición del método dado por (3.3) una función k veces continuamente diferenciable en $]t_0, t_N[\times\mathbb{R}\times]0, H[$, una condición necesaria y suficiente para que el método sea convergente de orden k es que se verifiquen para todo par de valores (t, y) de $]t_0, t_N[\times\mathbb{R}$ las k igualdades siguientes:*

$$
\begin{array}{lll}
g(t,y,0) & = f(t,y) & \\
\dfrac{\partial g}{\partial h} g(t,y,0) & = \dfrac{1}{2}\left[\dfrac{\partial f}{\partial x}(t,y) + \dfrac{\partial f}{\partial x}(t,y)\, f(t,y)\right] & = \dfrac{1}{2}\,\dfrac{df}{dx}(t,y) \\
\dfrac{\partial^2 g}{\partial h^2}(t,y,0) & = \dfrac{1}{3}\,[\dfrac{\partial^2 f}{\partial x^2} + 2\,\dfrac{\partial^2 f}{\partial x \partial y}\, f + & \\
 & + \dfrac{\partial f}{\partial x}\,\dfrac{\partial f}{\partial x} + \left(\dfrac{\partial^2 f}{\partial y^2}\right)^2 f + \dfrac{\partial^2 f}{\partial y^2}\, f^2] & = \dfrac{1}{3}\dfrac{d^2 f}{dx^2}(t,y) \\
\ldots. & \ldots\ldots & \ldots\ldots \\
\dfrac{\partial^{(k-1)} g}{\partial h^{(k-1)}}(t,y,0) & = \dfrac{1}{k}\,\dfrac{d^{(k-1)} f}{dx^{(k-1)}}(t,y) &
\end{array}
$$

Demostración:
Consúltese Crouzeix y Mignot[18].

Observación 3.3.2 *Las mayoraciones realizadas en los teoremas precedentes son, a menudo, muy "pesimistas", es decir utilizando magnitudes sobremayoradas. En muchos casos (imponiendo más condiciones a las funciones $f(t, y)$ y $g(t, y, h)$) pueden obtenerse acotaciones más precisas. Por ejemplo, en Crouzeix y Mignot[19] pueden encontrarse algunas de ellas. En esa misma referencia puede encontrarse la demostración del siguiente teorema:*

Teorema 3.3.7 *Bajo las hipótesis:*

a) Si el método (3.3) es estable y consistente de orden k, con $k \geq 1$,

b) la función f se supone $(k+1)$ veces continuamente diferenciable en $]t_0, t_N[\times\mathbb{R}$

c) siendo la función g que interviene en la definición del método una función $(k+1)$ veces continuamente diferenciable en $]t_0, t_N[\times\mathbb{R}\times]0, H[$

d) la longitud de los intervalos de integración h_n puede expresarse como una función $h_n = h(\theta(t_n) + 0(h))$, siendo $\theta(t)$ una función lipschitciane en $[t_0, t_N]$ y tal que $0 < \theta(t) \leq 1$ para todo valor de t

[18] Crouzeix, M., Mignot A.L. (1984) "Analyse numérique des équations différentielles", Ed. Masson.

e) si para los diferentes valores de h que puedan considerarse la aproximación $y_{0,h}$ que se tome del valor inicial y_0 satisface que $|y_0 - y_{0,h}| = O(h^{k+1})$.

entonces el error del método en t_n satisface la expresión:

$$e_n = y(t_n) - y_n = h^k\, z_1(t_n) + (y_0 - y_{0,h})\, z_0(t_n) + O(h^{k+1})$$

donde $z_0(t)$ y $z_1(t)$ son las soluciones de los problemas de valor inicial siguientes:

$$\begin{cases} z_0'(t) & = \dfrac{\partial f}{\partial y}(t, y(t)) z_0(t) \\ z_0(t_0) & = 1 \end{cases}$$

y

$$\begin{cases} z_1'(t) & = \dfrac{\partial f}{\partial y}(t, y(t))\, z_1(t) + \left[\dfrac{1}{(k+1)!} \dfrac{d^k f}{dx^k}(t, y(t)) - \dfrac{1}{k!} \dfrac{d^k g}{dh^k}(t, y(t), 0) \right] (\theta(t))^k \\ z_1(t_0) & = 0 \end{cases}$$

Observación 3.3.3 *El teorema anterior puede ser utilizado, entre otras cosas, para diseñar estrategias de control del tamaño de paso basadas en la extrapolación de Richardson. En efecto, en dicho teorema se establece que dado un método estable de orden k (y bajo hipótesis de regularidad suficientes) en un punto t^* puede obtenerse una aproximación con un tamaño de paso h que denotaremos por $y_{n,h}$ y otra aproximación $y_{m,q\,h}$ utilizando un tamaño de paso $q\,h$ verificándose entonces que:*

$$y(t_*) = y_{n,h} + h^k\, z_1(t^*) + O(h^{k+1})$$

$$y(t_*) = y_{m,q\,h} + q^k\, h^k\, z_1(t^*) + O(h^{k+1})$$

Combinando las dos expresiones anteriores se obtiene que:

$$\frac{q^k\, y_{n,h} - y_{m,q\,h}}{q^k - 1} = y(t^*) + O(h^{k+1})$$

y que

$$y_{n,h} - y(t_*) = \frac{y_{m,q\,h} - y_{n,h}}{q^k - 1} + O(h^{k+1})$$

La última de las expresiones obtenidas nos permite considerar que:

$$\frac{y_{m,q\,h} - y_{n,h}}{q^k - 1}$$

es una estimación del error cometido al actuar con paso h.

Con estas consideraciones una técnica para controlar automáticamente el paso de cálculo consiste en calcular el valor $y_{n+1,h}$ a partir de y_n (es decir con un tamaño de paso h_n) y calcular $y_{n+1,h'}$ a partir de y_{n-1} (es decir con tamaño de paso $h' = h_{n-1} + h_n$) denotando por q al valor:

$$q = \frac{h_n + h_{n+1}}{h_n}$$

Si, siendo ε una tolerancia de error asumible y dada por el usuario, se verificase que:

$$\left| \frac{y_{n+1,h'} - y_{n+1,h}}{q^k - 1} \right| \leq \varepsilon$$

el error cometido entre la aproximación obtenida con paso h y la solución exacta es un error aceptable y podría intentar operarse con un tamaño de paso $q\,h_n$ para etapas posteriores del método. Además el valor aproximado puede mejorarse mediante:

$$\widehat{y}_{n+1} = y_{n+1,h} - \frac{y_{n+1,h'} - y_{n+1,h}}{q^k - 1} = \frac{q^k\, y_{n+1,h} - y_{n+1,h'}}{q^k - 1}$$

Si por el contrario resultara que:

$$\left| \frac{y_{n+1,h'} - y_{n+1,h}}{q - 1} \right| > \varepsilon$$

resultará que el tamaño de paso h_n no nos garantiza un error aceptable en la estimación de la solución aproximada obtenida por el método de Euler. Ello aconsejará reducir el tamaño del paso utilizado (por ejemplo a la mitad) pasando a denominarse t_{n+1} al punto $t_n + \dfrac{h_n}{2}$ considerando entonces que h_n toma un valor igual a la mitad que el que tenía. Con este nuevo tamaño de paso puede volver a repetirse el mismo proceso.

3.4. Los métodos de Runge-Kutta

Los métodos de Runge-Kutta representan el ejemplo más clásico de los métodos de pasos libres.

3.4.1. Descripción.

Consideramos el problema de valor inicial (*P.V.I.*) y la subdivisión del intervalo $[t_0, t_N]$ generando los puntos $t_0 < t_1 < .... < t_n < .... < t_N$, y con subintervalos $[t_n, t_{n+1}]$ de longitud h_n ($n = 0, 1, ..., N-1$).

Siendo $y(t_n)$ el valor de la solución en t_n, el valor exacto de la solución en t_{n+1} puede estimarse mediante la expresión:

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \tag{3.4}$$

Una forma de obtener aproximaciones de dicho valor consistirá en aproximar la integral que aparece en 3.4 mediante una fórmula de integración numérica:

$$y_{n+1} = y_n + h_n \sum_{j=1}^{p} a_j\, f(t_{n,j}, y_{n,j}) \tag{3.5}$$

donde p es el número de puntos de integración usados en la fórmula escogida, $h_n a_j$ ($j = 1, ..., p$) son los pesos de dicha fórmula, $t_{n,j}$ ($j = 1, ..., p$) son los p puntos pertenecientes al intervalo $[t_n, t_{n+1}]$ que actúan como soporte para la fórmula de integración y, finalmente, $y_{n,j}$ son los valores (o una aproximación de ellos) de $y(t)$ en los puntos $t_{n,j}$.

El problema que plantea el uso de (3.5) es cómo evaluar $y_{n,j}$. Para ello, de forma similar a (3.4) se considerará que:

$$y(t_{n,j}) = y(t_n) + \int_{t_n}^{t_{n,j}} f(t, y(t))\, dt, \qquad j = 1,\, 2\ s,\, p \tag{3.6}$$

por lo que una aproximación de dichos valores puede obtenerse, nuevamente, aproximando la integral anterior. Y en los métodos de Runge-Kutta esta aproximación se realiza utilizando los mismos puntos que en la expresión (3.5), es decir:

$$y_{n,j} = y_n + h_n \sum_{i=1}^{p} b_{j,i}\, f(t_{n,i}, y_{n,i}) \quad (j = 1,\ s,\, p) \tag{3.7}$$

donde $h_n b_{j,k}$ ($j, k = 1, ..., p$) es el peso otorgado al punto $t_{n,k}$ en la fórmula de integración numérica utilizada para aproximar el valor de $y(t)$ en $t_{n,j}$.

Las expresiones dadas por (3.7) más la expresión 3.5 forman un sistema de $(p+1)$ ecuaciones en las que las incógnitas son los p valores $y_{n,j}$ ($j = 1, ..., p$) y el valor y_{n+1}. El uso de una u otra fórmula de integración numérica en las expresiones (3.5) y (3.7) conduce a muy distintos esquemas de tipo Runge-Kutta. Por ello para definir una fórmula de Runge-Kutta se debe concretar cuáles son los puntos $t_{n,j}$ ($j = 1, ..., p$) utilizados en las fórmulas de integración numérica, cuáles son los coeficientes $b_{j,k}$ ($j, k = 1, ..., p$) usados en las fórmulas (3.7) y cuáles son los coeficientes a_j ($j = 1, ..., p$) que intervienen en la expresión (3.5). Habitualmente, los puntos $t_{n,j}$ se suelen definir mediante una expresión del tipo:

$$t_{n,j} = t_n + c_j\, h_n \;\; (j \;=\; 1,\, ...,\, p)$$

y el esquema se define mediante la tabla siguiente:

$$\begin{array}{c|cccc} c_1 & b_{1,1} & b_{1,2} & s & b_{1,p} \\ c_2 & b_{2,1} & b_{2,2} & s & b_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_p & b_{p,1} & b_{p,2} & s & b_{p,p} \\ \hline & a_1 & a_2 & s & a_p \end{array}$$

En dicha tabla designaremos por vector $\mathbf{a}$, matriz $\mathbf{B}$ y matriz $\mathbf{C}$ al vector y matrices siguientes:

$$\mathbf{a} \;=\; \left\{ \begin{array}{c} a_1 \\ a_2 \\ \vdots \\ a_p \end{array} \right\},\;\; \mathbf{B} \;=\; \left[\begin{array}{cccc} b_{1,1} & b_{1,2} & s & b_{1,p} \\ b_{2,1} & b_{2,2} & s & b_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ b_{p,1} & b_{p,2} & s & b_{p,p} \end{array} \right],\; \mathbf{C} \;=\; \left[\begin{array}{cccc} c_1 & 0 & s & 0 \\ 0 & c_2 & s & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & s & c_p \end{array} \right] \tag{3.8}$$

3.4.2. Ejemplos de métodos.

Un primer ejemplo.

Puede diseñarse un método de Runge-Kutta en el que la fórmula que nos proporcione y_{n+1} utilice la fórmula de Simpson:

$$\int_a^b \phi(x)dx \approx \frac{b-a}{6}\,\left(\phi(a) + 4\,\phi\left(\frac{a+b}{2}\right) + \phi(b)\right)$$

Para ello, se tomará $p = 3$ y en el intervalo $[t_n, t_{n+1}]$ los puntos de integración serán:

$$\begin{array}{rcll} t_{n,1} & = & t_n & (\rightarrow c_1 = 0) \\ t_{n,2} & = & t_n + 0{,}5\, h_n & (\rightarrow c_2 = 0{,}5) \\ t_{n,3} & = & t_n + h_n & (\rightarrow c_3 = 1) \end{array}$$

y los pesos de integración en la fórmula considerada serán:

$$a_1 = \frac{1}{6},\;\; a_2 = \frac{4}{6},\;\; a_3 = \frac{1}{6}$$

resultando:

$$y_{n+1} = y_n + h_n\,\left(\frac{1}{6}\,f(t_{n,1}, y_{n,1}) + \frac{4}{6}\,f(t_{n,2}, y_{n,2}) + \frac{1}{6}\,f(t_{n,3}, y_{n,3})\right)$$

Para emplear la fórmula anterior es necesario evaluar: $y_{n,1}$ $(= y_n)$, $y_{n,2}$ e $y_{n,3}$. Para ello se sabe que:

$$\begin{aligned} y_{n,1} &= y_n \\ y_{n,2} &= y_n + \int_{t_n}^{t_{n,2}} f(t, y(t))\, dt \\ y_{n,3} &= y_n + \int_{t_n}^{t_{n,3}} f(t, y(t))\, dt \end{aligned}$$

y evaluando la primera de las integrales mediante la fórmula del trapecio:

$$\begin{aligned} \int_{t_n}^{t_{n,2}} f(t, y(t))\, dt &\approx \frac{t_{n,2} - t_n}{2}\, (f(t_n, y_n) + f(t_{n,2}, y_{n,2})) \\ &= \frac{h_n}{4}\, (f(t_n, y_n) + f(t_{n,2}, y_{n,2})) \end{aligned}$$

y la segunda mediante el método del punto medio:

$$\begin{aligned} \int_{t_n}^{t_{n,3}} f(t, y(t))\, dt &\approx (t_{n,3} - t_n)\, f(t_{n,2}, y_{n,2}) \\ &= h_n\, f(t_{n,2}, y_{n,2}) \end{aligned}$$

resultará

$$\begin{aligned} y_{n,1} &= y_n && (\to b_{1,1} = 0,\ b_{1,2} = 0,\ b_{1,3} = 0) \\ y_{n,2} &= y_n + \frac{h_n}{4}\, (f(t_n, y_n) + f(t_{n,2}, y_{n,2})) && \left(\to b_{2,1} = \frac{1}{4},\ b_{2,2} = \frac{1}{4},\ b_{2,3} = 0\right) \\ y_{n,3} &= y_n + h_n\, f(t_{n,2}, y_{n,2}) && (\to b_{3,1} = 0,\ b_{3,2} = 1,\ b_{3,3} = 0) \end{aligned}$$

La segunda de las expresiones anteriores es una ecuación (en general no lineal) que deberá resolverse mediante algún método numérico como los presentados en el tema anterior.

El esquema en forma de tabla será:

0	0	0	0
1/2	1/4	1/4	0
1	0	1	0
	1/6	4/6	1/6

El método de Euler explícito

El **método de Euler**, estudiado en el apartado 3, puede considerarse como un caso particular de los métodos de Runge-Kutta en el que $p = 1$ y:

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

El método de Euler modificado y el método de Heun.

Siendo α un número real, el método dado por:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1\text{-}1/(2\,\alpha) & 1/(2\,\alpha) \end{array}$$

se traduce en las expresiones:

$$\begin{aligned} y_{n,1} &= y_n \\ y_{n,2} &= y_n + \alpha h_n f(t_n, y_n) \\ y_{n+1} &= y_n + h_n \left(\left(1 - \frac{1}{2\alpha} \right) f(t_n, y_n) + \frac{1}{2\alpha} f(t_n + \alpha h_n, y_{n,2}) \right) \end{aligned}$$

Para el caso $\alpha = 0{,}5$, el método se denomina de **Método de Euler modificado**:

$$y_{n+1} = y_n + h_n f\left(t_n + \frac{h_n}{2}, y_n + \frac{h_n}{2} f(t_n, y_n) \right) \tag{3.9}$$

Para el caso $\alpha = 1{,}0$, el método se denomina de **Método de Heun**:

$$y_{n+1} = y_n + \frac{h_n}{2} \left(f(t_n, y_n) + f(t_n + h_n, y_n + h_n f(t_n, y_n)) \right) \tag{3.10}$$

El método de Runge-Kutta clásico (de orden 4)

El **método de Runge-Kutta clásico** responde a la tabla:

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 2/6 & 2/6 & 1/6 \end{array}$$

Por tanto, en este método, los puntos intermedios escogidos en el paso n serán:

$$\begin{aligned} t_{n,1} &= t_n \\ t_{n,2} &= t_n + 0{,}5\, h_n \\ t_{n,3} &= t_n + 0{,}5\, h_n \ \ (= t_{n,2}) \\ t_{n,4} &= t_n + h_n \end{aligned}$$

y como vemos se repite el punto medio. Con los puntos así tomados, se considera $y_{n,1} = y_n$. El valor de $y_{n,2}$ se evalúa aproximando la integral correspondiente mediante la fórmula del rectángulo con soporte en el extremo izquierdo del intervalo de integración, es decir:

$$y(t_{n,2}) = y(t_n) + \int_{t_n}^{t_{n,2}} f(t, y(t))dt \Rightarrow y_{n,2} = y_n + \frac{h_n}{2} f(t_n, y_n)$$

El valor de $y_{n,3}$ se calcula aproximando la integral correspondiente mediante la fórmula del rectángulo pero esta vez con el punto de soporte en el extremo derecho del intervalo y tomando precisamente $y_{n,2}$ como valor aproximado de la función $y(t)$ en dicho extremo derecho:

$$y(t_{n,3}) = y(t_n) + \int_{t_n}^{t_{n,3}} f(t, y(t))dt \Rightarrow y_{n,3} = y_n + \frac{h_n}{2} f(t_n + \frac{h_n}{2}, y_{n,2})$$

En cuanto al valor de $y_{n,4}$ lo evaluaremos aproximando la integral correspondiente mediante una fórmula de punto medio, considerando como valor de $y(t)$ en el punto medio del intervalo en el que se plantea la fórmula el valor $y_{n,3}$ que se acaba de calcular, es decir:

$$y(t_{n,4}) = y(t_n) + \int_{t_n}^{t_n+h_n} f(t, y(t))dt \Rightarrow y_{n,4} = y_n + h_n f(t_n + \frac{h_n}{2}, y_{n,3})$$

Por último, el valor de y_{n+1} se evalúa, con ayuda de los valores anteriores, aproximando la integral correspondiente mediante una fórmula de Simpson en la cual el peso asignado al punto medio del intervalo se reparte por igual entre las dos aproximaciones del valor de $y(t)$ que hemos obtenido en dicho punto medio ($y_{n,2}$ e $y_{n,3}$):

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_n+h_n} f(t, y(t))dt \ \Rightarrow$$

$$y_{n+1} = y_n + \frac{h_n}{6}\left[f(t_n, y_n) + 2f(t_{n,2}, y_{n,2}) + 2f(t_{n,3}, y_{n,3}) + f(t_{n,4}, y_{n,4})\right] =$$

$$y_n + \frac{h_n}{6}\left[f(t_n, y_n) + 2\left(f(t_n + \frac{h_n}{2}, y_{n,2}) + f(t_n + \frac{h_n}{2}, y_{n,3})\right) + f(t_n + h_n, y_{n,4})\right]$$

NOTA: *Este método es atribuido a Carle David Tolmé Runge (nacido en Bremen (Alemania) en 1856 y fallecido en Göttingen (Alemania) en 1927) y a Martin Wilhelm Kutta (nacido en Pitschen (Polonia) en 1867 y fallecido en Fürstenfeldbruck (Alemania) en 1944). De él toman el nombre la familia de métodos numéricos que estamos presentando.*

Otra variante del método de Euler.

Si se considera el método de Runge-Kutta definido mediante la tabla:

$$\begin{array}{c|c} 1 & \boxed{1} \\ \hline & 1 \end{array}$$

se tiene el método:

$$\left\{\begin{array}{rcl} y_{n,1} & = & y_n + h_n f(t_{n+1}, y_{n,1}) \\ y_{n+1} & = & y_n + h_n f(t_{n+1}, y_n + h_n f(t_{n+1}, y_{n,1})) \end{array}\right. \tag{3.11}$$

que es el denominado **método de Euler implícito mejorado**.

Un método de orden 3.

Si se considera el valor:

$$\alpha = \frac{1}{2} + \frac{1}{2\sqrt{3}}$$

se puede considerar el método definido mediante la tabla:

$$\begin{array}{c|cc} \alpha & \alpha & 0 \\ 1-\alpha & 1-2\alpha & \alpha \\ \hline & 1/2 & 1/2 \end{array}$$

que conduce al esquema:

$$y_{n,1} = y_n + \alpha\, h_n\, f(t_n + \alpha h_n, y_{n,1}) \tag{3.12}$$

$$y_{n,2} = y_n + h_n\, ((1-2\alpha)\, f(t_n + \alpha h_n, y_{n,1}) + \alpha\, f(t_n + (1-\alpha)h_n, y_{n,2})) \tag{3.13}$$

$$y_{n+1} = y_n + \frac{h_n}{2}\left(f(t_n + \alpha\, h_n, y_{n,1}) + f(t_n + (1-\alpha)\, h_n, y_{n,2})\right) \qquad (3.14)$$

que es un método que exige, en cada paso, resolver las ecuaciones (en general no lineales) (3.12) y (3.13) para poder estimar y_{n+1} mediante (3.14). Este método es de orden de consistencia 3.

Observación 3.4.1 *Como puedes ver el método de Euler y algunas de las variantes que anteriormente te habíamos presentado pueden obtenerse como un caso particular de los métodos de Runge-Kutta. En el apartado dedicado al método de Euler obtuvimos dicho método de formas diferentes (mediante desarrollos de Taylor, mediante interpolación polinómica, etc...). De hecho los métodos de Runge-Kutta en general podrían obtenerse de manera diferente a como aquí se ha planteado. En este sentido es muy frecuente en la literatura sobre este tema el obtener los métodos de Runge-Kutta partiendo de desarrollos en serie de Taylor. Así, por ejemplo, se hace en Burden y Faires*[19] *o Shampine y Allen y Pruess*[20].

3.4.3. Un algoritmo del método de Runge-Kutta clásico.

Considérese un método de Runge-Kutta genérico definido, en cada paso, por las expresiones:

$$t_{n,j} = t_n + c_j\, h_n \;\; (j = 1, ..., p)$$

$$y_{n,j} = y_n + h_n \sum_{i=1}^{p} b_{j,i}\, f(t_{n,i}, y_{n,i}), \qquad (j = 1,\ s, p)$$

$$y_{n+1} = y_n + h_n \sum_{j=1}^{p} a_j\, f(t_{n,j}, y_{n,j})$$

A la hora de realizar un algoritmo de un método como el anterior, computacionalmente es más eficaz denotar por:

$$W_{n,j} = f(t_{n,j}, y_n + h_n . \sum_{i=1}^{p} b_{j,i}\, W_{n,i}) \qquad (j = 1, ...p), \qquad (3.15)$$

con lo que:

$$y_{n+1} = y_n + h_n \sum_{j=1}^{p} a_j\, W_{n,j}. \qquad (3.16)$$

[19] Burden R.L., Douglas Faires J. (1998) "Análisis Numérico", International Thomson Editores, 6. edición.

[20] Shampine, L.F., Allen, R.C. Jr. y Pruess, S. (1997). "Fundamentals of numerical computing". Ed. John Wiley & Sons, Inc.

La expresión (3.15) representa un sistema de p ecuaciones (en general no lineales) que nos proporcionará los valores de $W_{n,j}$ en cada paso. Estos valores introducidos en (3.16) nos determinarán la solución aproximada y_{n+1}. Fácilmente puede comprobarse que ambas expresiones son equivalentes a las antes utilizadas para describir el método.

Utilizando esta forma de proceder, se recoge a continuación un algoritmo del método de Runge-Kutta clásico (que es el más frecuentemente utilizado) para la resolución de un problema de valor inicial regido por un sistema de ecuaciones diferenciales ordinarias como el siguiente:

$$\begin{cases} \mathbf{y}'(t) & = \mathbf{f}(t, \mathbf{y}(t)), \quad t_0 \leq t \leq t_N \\ \mathbf{y}(t_0) & = \mathbf{y}^{(0)} \end{cases}$$

INICIO DEL ALGORITMO

Dados:

t_0, t_1, ..., $t_N = t_0 + T$, la expresión de la función $\mathbf{f}(t, \mathbf{y}(t))$ y el vector $\mathbf{y}^{(0)}$

Para $n = 0$, **hasta** $n = N - 1$, **con paso** 1, **hacer:**

$h_n \leftarrow t_{n+1} - t_n$

$\mathbf{W1} \leftarrow \mathbf{f}(t_n, \mathbf{y}^{(n)})$

$\mathbf{W2} \leftarrow \mathbf{f}(t_n + \dfrac{h_n}{2}, \mathbf{y}^{(n)} + \dfrac{h_n}{2}\,\mathbf{W1})$

$\mathbf{W3} \leftarrow \mathbf{f}(t_n + \dfrac{h_n}{2}, \mathbf{y}^{(n)} + \dfrac{h_n}{2}\,\mathbf{W2})$

$\mathbf{W4} \leftarrow \mathbf{f}(t_n + h_n, \mathbf{y}^{(n)} + h_n\,\mathbf{W3})$

$\mathbf{y}^{(n+1)} \leftarrow \mathbf{y}^{(n)} + \dfrac{h_n}{6}\,(\mathbf{W1} + 2\,(\mathbf{W2} + \mathbf{W3}) + \mathbf{W4})$

Escribir el vector $\mathbf{y}^{(n+1)}$ como solución aproximada en t_{n+1}.

Fin bucle en n.

FIN DEL ALGORITMO.

Observación 3.4.2 *El algoritmo anterior puede (y en la práctica debe) modificarse con alguna estrategia que permita controlar el valor de h_n en cada etapa con el objeto de minimizar el error entre la solución aproximada ofrecida por el método y la solución exacta del P.V.I. A estas estrategias de control automático del tamaño del paso de integración, que suelen combinar dos tipos de métodos de Runge-Kutta, nos referiremos más adelante.*

3.4.4. Clasificación y análisis.

Consideremos un método de Runge-Kutta definido por la tabla que se obtiene a partir del vector **a** y de las matrices **B** y **C** como las definidas en (3.8).

Si la matriz **B** fuese estrictamente triangular inferior, cada valor $y_{n,j}$ puede evaluarse únicamente a partir de y_n y de los valores $y_{n,k}$ $(k = 1, ..., j-1)$ previamente calculados. Es decir que en este caso:

$$\begin{aligned} y_{n,1} &= y_n \\ y_{n,j} &= y_n + h_n \sum_{k=1}^{j-1} b_{j,k}\, f(t_{n,k}, y_{n.k}), \qquad (j = 2, ..., p) \\ y_{n+1} &= y_n + h_n \sum_{j=1}^{p} a_j\, f(t_{n,j}, y_{n,j}) \end{aligned}$$

por lo que el valor de y_{n+1} puede determinarse mediante las operaciones aritméticas que se acaban de describir y no es necesario resolver ninguna ecuación intermedia. Este tipo de métodos se denomina **métodos de Runge-Kutta explícitos.**

Si la matriz **B** fuese triangular inferior con alguno de sus elementos diagonales no nulo, ello querría decir que algunos valores intermedios $y_{n,j}$ exigen para su cálculo la resolución de una ecuación ya que la expresión (3.7) que los proporciona los deja dependiendo de sí mismos. Si por ejemplo $b_{j,j} \neq 0$ se tiene que:

$$y_{n,j} = y_n + h_n \sum_{k=1}^{j-1} b_{j,k}\, f(x_{n,k}, y_{n,k}) + h_n\, b_{j,j}\, f(t_{n,j}, y_{n,j}) \Rightarrow$$

$$\Rightarrow y_{n,j} - h_n\, b_{j,j}\, f(t_{n,j}, y_{n,j}) = y_n + h_n \sum_{k=1}^{j-1} b_{j,k}\, f(t_{n,k}, y_{n,k}).$$

Luego el uso de este tipo de métodos exige resolver tantas ecuaciones (en general no lineales) como coeficientes diagonales $b_{j,j}$ no nulos existan. Los valores de $y_{n,k}$ para los que $b_{k,k}$ sea nulo se pueden determinar sin necesidad de resolver ecuación alguna. Además, en todo caso, las ecuaciones que deben resolverse son independientes entre sí y pueden irse resolviendo de una en una a partir de la que haga intervenir puntos con menor subíndice. Este tipo de métodos se denomina **métodos de Runge-Kutta semi-implícitos.**

Por último, si la matriz **B** no fuese triangular inferior en el cálculo de alguno de los valores $y_{n,j}$ intervendría el valor de $y_{n,k}$ con $k > j$. En dicho

caso la determinación de los valores intermedios exigiría resolver un sistema de ecuaciones (en general no lineales). Por dicho motivo a este tipo de métodos se les denomina **métodos de Runge-Kutta implícitos.**

Observación 3.4.3 *Suponer que la matriz* $\mathbf{B}$ *puede ser triangular superior equivale a suponerla triangular inferior sin más que ordenar los segundos subíndices de los puntos intermedios* $t_{n,j}$ *en sentido inverso al que se considere. Por tanto sólo se distinguen las tres situaciones que se acaban de señalar: que* $\mathbf{B}$*:* **1)** *sea estrictamente triangular inferior;* **2)** *sea triangular inferior;* **3)** *no sea triangular inferior, asumiéndose en este caso que tampoco es triangular superior. Es por eso que en el tercer caso se ha afirmado que la aplicación del método exige resolver un sistema de ecuaciones (cuando si se hubiese considerado la posibilidad de que* $\mathbf{B}$ *fuese triangular superior podría darse el caso de sólo tener que resolver ecuaciones independientes entre sí (si* $\mathbf{B}$ *es triangular superior con algún elemento diagonal no nulo) o sin necesidad de tener que resolver ninguna ecuación (si* $\mathbf{B}$ *fuese estríctamente triangular superior)).*

La consideración de métodos semi-implícitos e implícitos plantea una cuestión adicional en el análisis de los métodos. Dicha cuestión consiste en determinar si la ecuación o ecuaciones a resolver en cada paso tienen solución y, en caso de tenerla, si dicha solución es única o existen varias soluciones verificando las ecuaciones intermedias. Esta última cuestión cae dentro del análisis de ecuaciones no lineales y de los métodos numéricos de resolución de las mismas.

En general escribiremos el método de Runge-Kutta en la forma:

$$\{\tilde{\mathbf{y}}_n\} - h_n \left[\mathbf{B}\right] \{\tilde{\mathbf{y}}_n\} = \{\mathbf{y}_n\} \tag{3.17}$$

$$y_{n+1} = y_n + h_n g(t_n, \{\tilde{\mathbf{y}}_n\}, h_n) = y_n + h_n \left(\sum_{j=1}^{p} a_j f(t_j + c_j h_n, \tilde{y}_{n,j}) \right) \tag{3.18}$$

donde se ha designado por:

$$\{\tilde{\mathbf{y}}_n\} = \left\{ \begin{array}{c} \tilde{y}_{n,1} \\ \tilde{y}_{n,2} \\ \vdots \\ \tilde{y}_{n,p} \end{array} \right\} = \left\{ \begin{array}{c} y_{n,1} \\ y_{n,2} \\ \vdots \\ y_{n,p} \end{array} \right\}, \quad \{\mathbf{y}_n\} = \left\{ \begin{array}{c} y_n \\ y_n \\ \vdots \\ y_n \end{array} \right\}$$

Supondremos además en este apartado que la función $f(t, y)$ que interviene en el problema de valor inicial es lipschitciana, de razón L, respecto a su segunda variable, es decir que verifica:

$$\exists L > 0 / \forall t \in [t_0, t_N], \quad \forall y, z \in I\!R : \quad | f(t,y) - f(t,z) | \leq L \, | y - z | \quad (3.19)$$

En estas condiciones puede demostrarse el siguiente teorema:

Teorema 3.4.1 *Si se designa por $\rho(\mathbf{B})$ al radio espectral de la matriz B con la que se define el método de Runge-Kutta y se verifica que $h_n \, L \, \rho(\mathbf{B}) < 1$, entonces el sistema dado por (3.7) admite una solución única.*

Demostración:

Es una consecuencia del teorema de punto fijo demostrado en el tema anterior (consultar el Crouzeix y Mignot[21]).

c.q.d.

En cuanto al análisis de la estabilidad de los métodos de Runge-Kutta comencemos analizando el caso de los métodos explícitos. Con el objeto de introducir al lector de forma sencilla en la técnica del análisis de estabilidad estudiamos en primer lugar la estabilidad de un método de Runge-Kutta explícito con dos puntos "intermedios" $t_{n,1} = t_n + c_1 \, h_n$ y $t_{n,2} = h_n + c_2 \, h_n$. La tabla de este método de Runge-Kutta puede escribirse entonces como:

c_1	0	0
c_1	$b_{2,1}$	0
	a_1	a_2

con lo que una etapa genérica del método se puede escribir (véase los comentarios antes realizados al describir la forma de construir un algoritmo de estos métodos) como:

$$W_{n,1} = f(t_{n,1}, y_n)$$

$$W_{n,2} = f(t_{n,2}, y_n + h_n \, b_{2,1} \, W1)$$

$$y_{n+1} = y_n + h_n \, (a_1 \, W_{n,1} + a_2 \, W_{n,2})$$

La idea para estudiar las condiciones en que este método es estable consiste en verificar las hipótesis del teorema (3.3.3). Por ello supondremos que el método se aplica a un problema de valor inicial:

$$(P.V.I.) \begin{cases} y'(t) & = f(t, y(t)) \quad t \in [t_0, t_N] \\ y(t_0) & = y_0 \end{cases}$$

[21] Crouzeix, M., Mignot A.L. (1984) "Analyse numérique des équations différentielles", Ed. Masson.

en el que la función $f(t,y)$ es una función lipschitciana (de razón L) respecto a su segunda variable y llamaremos función $g(t,y,h)$ a la función:

$$g(t,y,h) = a_1\, f(t + c_1\, h, y) + a_2\, f(t + c_2\, h, y + h\, b_{2,1}\, f(t + c_1\, h, y)) =$$

$$= a_1\, W1(t,y,h) + a_2\, W2(t,y,h)$$

con

$$W1(t,y,h) = f(t + c_1\, h, y), \quad W2(t,y,h) = f(t + c_2\, h, y + h\, b_{2,1}\, W1(t,y,h))$$

Utilizando esta notación es evidente que para todo par de valores de y y z reales:

$$|W1(t,y,h) - W1(t,z,h)| = |f(t + c_1\, h, y) - f(t + c_1\, h, z)| \leq$$

$$\leq L\ |y - z|$$

y

$$|W2(t,y,h) - W2(t,z,h)| =$$

$$|f(t + c_2\, h, y + h\, b_{2,1}\, W1(t,y,h)) - f(t + c_2\, h, z + h\, b_{2,1}\, W1(t,z,h))| \leq$$

$$\leq L\ |(y + h\, b_{2,1}\, W1(t,y,h)) - (z + h\, b_{2,1}\, W1(t,z,h))| =$$

$$= L\ |(y - z) + h\, b_{2,1}\ (W1(t,y,h) - W1(t,z,h))| \leq$$
$$\leq L.\, |y - z| + h\, L\ |b_{2,1}|\ |W1(t,y,h) - W1(t,z,h)| \leq$$

$$\leq L.\, |y - z| + h\, L\ |b_{2,1}|\ L\ |y - z| =$$

$$= \left(L + h\, L^2\ |b_{2,1}|\right)\ |y - z|$$

Por tanto, para cualquier par de valores y y z:

$$|g(t,y,h) - g(t,z,h)| =$$

$$|(a_1\, W1(t,y,h) + a_2\, W2(t,y,h)) - (a_1\, W1(t,z,h) + a_2\, W2(t,z,h))| =$$

$$= |a_1\ (W1(t,y,h) - W1(t,z,h)) + a_2\ (W2(t,y,h) - W2(t,z,h))| \leq$$

$$\leq |a_1|\ |W1(t,y,h) - W1(t,z,h)| + |a_2|\ |W2(t,y,h) - W2(t,z,h)| \leq$$

$$\leq |a_1|\ L\ |y-z| + |a_2|\ \left(L + h\,L^2\ |b_{2,1}|\right)\ |y-z| =$$
$$= \left(|a_1|\ L + |a_2|\ \left(L + h\,L^2\ |b_{2,1}|\right)\right) |y-z| =$$
$$= L'\ |y-z|$$

donde hemos designado por $\widetilde{L}$ al valor:

$$L' = L\ (|a_1| + |a_2|\ (1 + h\,L\ |b_{2,1}|))$$

El único problema que nos plantea la constante anterior es que depende del propio valor de h. En este sentido basta con considerar que h se escoge en un intervalo $0 < h \leq H$ para poder afirmar que:

$$|g(t,y,h) - g(t,z,h)| \leq \widetilde{L}\ |y-z|$$

donde se ha denotado por:

$$\widetilde{L} = L\ (|a_1| + |a_2|\ (1 + H\,L\ |b_{2,1}|))$$

Ello, en virtud del teorema (3.3.3) nos demuestra que el método es estable y que para cualquier terna de sucesiones $\{y_n\}_{n=0}^N$, $\{z_n\}_{n=0}^N$ y $\{E_n\}_{n=1}^N$ verificando las relaciones:

$$y_{n+1} = y_n + h_n\,g(t_n, y_n, h_n)$$

$$z_{n+1} = z_n + h_n\,g(t_n, z_n, h_n) + E_{n+1}$$

se verifica

$$Sup_{0\leq n\leq N}\,\{|z_n - y_n|\} \leq e^{\widetilde{L}\,(t_N - t_0)}\ \left[|z_0 - y_0| + \sum_{n=1}^{N} |E_n|\right] \qquad \forall h / 0 < h < H$$

El resultado anterior puede generalizarse al caso más general como se hace en el siguiente teorema:

Teorema 3.4.2 *Considérese el método de Runge-Kutta explícito definido mediante:*

$$W_{n,1} = f(t_n + c_1\,h_n, y_n)$$

$$W_{n,j} = f(t_n + c_j\, h_n, y_n + \sum_{k=1}^{j-1} b_{j,k}\, W_{n,k}) \qquad (j = 2, ..., p)$$

$$y_{n+1} = y_n + h_n \sum_{j=1}^{p} a_j\, W_{n,j}$$

y aplicado a la resolución del problema (P.V.I) en el que la función $f(t, y(t))$ es lipschitciana de razón L respecto a su segunda variable. En estas condiciones el método es estable y se verifica que:

$$Sup_{0\leq n\leq N}\{|z_n - y_n|\} \leq e^{\widetilde{L}\,(t_N - t_0)} \left[|z_0 - y_0| + \sum_{n=1}^{N} |E_n|\right] \qquad \forall h/0 < h \leq H$$

donde se ha denotado por $\widetilde{L}$ al valor:

$$\widetilde{L} = L \left(\boldsymbol{\alpha}^T\, (\mathbf{I} + \sum_{j=1}^{p-1} h^j\, L^j\, (\boldsymbol{\beta})^j)\, \mathbf{e}\right)$$

habiéndose designado por:

$$\boldsymbol{\alpha}^T = \left\{ \begin{array}{cccc} |a_1| & |a_2| & ... & |a_p| \end{array} \right\}$$

$$\boldsymbol{\beta} = \left[\begin{array}{cccc} |b_{1,1}| & |b_{1,2}| & s & |b_{1,p}| \\ |b_{2,1}| & |b_{2,2}| & s & |b_{2,p}| \\ \vdots & \vdots & \vdots & \vdots \\ |b_{p,1}| & |b_{p,2}| & s & |b_{p,p}| \end{array} \right]$$

$$\mathbf{e} = \left\{ \begin{array}{c} 1 \\ 1 \\ .... \\ 1 \end{array} \right\}$$

Demostración:

Consiste en generalizar el desarollo realizado anteriormente para el caso de un método con dos puntos intermedios.

c.q.d.

El teorema anterior nos asegura que si $(t_N - t_0)$ tiene un valor finito y el método es consistente (cosa de la que nos ocuparemos posteriormente) y la función f es lipschitciana respecto a su segunda variable, la máxima diferencia entre los valores generados por el método de Runge-Kutta y la solución

analítica permanecerá acotada siempre que trabajemos con un paso inferior a H. Esto está bien (pues, obviamente, no deseamos que nuestras soluciones aproximadas "exploten" si no lo hacen las soluciones analíticas) pero no nos dice mucho acerca del funcionamiento de los métodos.

Baste observar para justificar la afirmación anterior que para valores elevados de H o de $(t_N - t_0)$ o de L, aunque el método sea consistente, pequeños errores (por ejemplo de redondeo) al estimar $y_0 \approx y(t_0) = z_0$ se traducen en una cota de error elevada. Y si se deseara asegurar que la cota de error no se hace muy grande, puesto que L y $(t_N - t_0)$ no está en nuestra mano elegirlos, nos veremos obligados a tener que tomar valores del tamaño máximo del paso H muy reducidos.

Observación 3.4.4 *Puesto que en la acotación anterior interveniene una exponencial, los métodos numéricos se suelen someter al test dado por el problema de valor inicial:*

$$\begin{cases} y'(t) &= -Ky(t), \quad t > 0 \\ y(0) &= 1 \end{cases} \qquad K > 0$$

que admite como solución una función exponencial $y(t) = e^{-Kt}$. Ello, en primer lugar, puede darnos una idea bastante buena de la evolución de los errores del método. Asimismo, puesto que la solución analítica es siempre positiva y decreciente puede permitirnos extraer cotas sobre los tamaños máximos de los pasos de integración a utilizar si se desean garantizar la positividad y el decrecimiento en valor absoluto de las soluciones aproximadas.

Ejemplo 3.4.1 *Considérese el método de Heun:*

$$y_{n+1} = y_n + \frac{h_n}{2}\left(f(t_n, y_n) + f(t_n + h_n, y_n + h_n\, f(t_n, y_n))\right)$$

y supóngase que se aplica con longitud de paso de integración constante h a la resolución del problema:

$$\begin{cases} y'(t) &= -K.y(t) \quad 0 < t < 10 \\ y(0) &= 1 \end{cases}, \qquad K > 0$$

Al estar trabajando en un intervalo acotado, los errores del método de Heun (que como veremos posterormente es consistente) van a permanecer acotados sea cual sea el valor que asignemos a h. Pero la solución aproximada puede no tener el más remoto parecido con la analítica si no limitamos este tamaño de paso de integración. ¿A qué valor debemos limitar este tamaño máximo?.

La forma de razonar consiste en analizar una etapa genérica del método aplicado al problema anterior, es decir:

$$y_{n+1} = y_n + \frac{h}{2}\left(f(t_n, y_n) + f(t_n + h, y_n + h\, f(t_n, y_n)\right) =$$

$$= y_n + \frac{h}{2}\left(-K\, y_n - K\,\left(y_n + h\,\left(-K\, y_n\right)\right)\right) =$$

$$= \left(1 - \frac{K\, h}{2} - \frac{K\, h}{2} + \frac{K^2\, h^2}{2}\right) y_n =$$

$$= \left(1 - K\, h + \frac{K^2\, h^2}{2}\right) y_n = \alpha\, y_n$$

donde

$$\alpha = \left(1 - K\, h + \frac{K^2\, h^2}{2}\right)$$

La expresión anterior, aplicada de forma recursiva, nos permite escribir que:

$$y_n = \alpha^n\, y_0 = \alpha^n$$

Obsévese que en este caso α siempre será positivo (al ser K y h positivos). En efecto, si se denota por $\xi = K\, h$ se tiene que $\alpha = 1 - \xi + \xi^2/2$ que es la ecuación de una parábola que nunca corta al eje de abscisas (no tiene raíces reales) y que, al tomar valor positivo para $\xi = 0$, siempre toma valores positivos. Por tanto, el método de Heun garantiza la positividad de la solución aproximada para cualquier elección del paso de integración. Pero ¿y el decrecimiento?. Para ello se debería exigir que:

$$\alpha < 1 \Rightarrow 1 - K\, h + \frac{K^2\, h^2}{2} < 1 \Rightarrow K\, h\,(1 - \frac{K\, h}{2}) > 0 \Rightarrow$$

$$\Rightarrow h < \frac{2}{K}$$

Es decir, que tenemos restricciones del tamaño de paso similares a las que encontrábamos en el método de Euler.

Observación 3.4.5 *Obsérvese que realmente estamos combinando dos conceptos de estabilidad. Uno, el que hemos definido anteriormente, que se refiere a que la solución aproximada permanezca acotada cuando se trabaja en intervalos acotados, aunque se utilicen valores del tamaño de paso muy pequeños y con ello se realicen un número elevado de pasos de integración. Y, otro, el que las soluciones aproximadas no exploten cuando se trabaja en intervalos de longitud tendente a infinito si no lo hacen las soluciones analíticas. Ello lleva a distinguir el concepto de estabilidad que hasta ahora hemos manejado del concepto de estabilidad absoluta cuyo tratamiento riguroso desborda los objetivos de este curso.*

Observación 3.4.6 *Utilizando palabras del profesor Euvrard, notemos que no sería "moral" que un método explícito (poco costoso por paso) fuese incondicionalmente estable para todo tipo de tamaños de pasos de integración y en tiempos tendientes al infinito, pues dejaría en muy mal lugar a los métodos implícitos.*

El análisis de la estabilidad realizado para los métodos de Runge-Kutta explícitos puede extenderse fácilmente a los métodos semi-implícitos e implícitos. Nosotros nos limitaremos a enunciar algunos de los teoremas más clásicos al respecto. En ellos denotaremos por matriz $\mathbf{B}^*$ a la matriz:

$$\mathbf{B}^* = \begin{bmatrix} |b_{1,1}| & |b_{1,2}| & s & |b_{1,p}| \\ |b_{2,1}| & |b_{2,2}| & s & |b_{2,p}| \\ \vdots & \vdots & \vdots & \vdots \\ |b_{p,1}| & |b_{p,2}| & s & |b_{p,p}| \end{bmatrix}$$

Teorema 3.4.3 *Suponiendo que* $\mathbf{B}$ *es una matriz para la que se verifica que* $H\,L\,\rho(\mathbf{B}^*) < 1$ *entonces el método de Runge-Kutta correspondiente es estable para todo valor del paso de integración* h *que verifique:* $0 < h < H$.

Demostración:

Es una consecuencia de los teoremas vistos para un método genérico de un paso. La demostración detallada puede consultarse, por ejemplo, en Crouzeix & Mignot[22].

c.q.d.

Observación 3.4.7 *El teorema anterior establece condiciones suficientes para la estabilidad. Al no ser necesarias, el límite del valor de paso de integración que asegura la estabilidad puede ser aún mayor que el recogido en el teorema.*

Observación 3.4.8 *Obsérvese que el teorema anterior nos vueleve a conducir a que los métodos de Runge-Kutta explícitos son siempre estables ya que en este tipo de métodos la matriz* $\mathbf{B}$ *es estrictamente triangular inferior por lo que* $\rho(\mathbf{B}^*) = 0$ *y, para cualquier valor de* H *se verifica que:* $\rho(\mathbf{B}^*)\,L\,H < 1$.. *Pero recuérdese lo que, en este contexto quería decir que el método fuese estable.*

Una vez presentadas las condiciones que aseguran la estabilidad de los esquemas de Runge-Kutta, pasemos a examinar las que garantizan la consistencia de tales métodos. Con ello se tendrá analizada la convergencia pues, como se detalló en el estudio de un método genérico, "consistencia más estabilidad implica convergencia".

[22] Crouzeix, M., Mignot A.L. (1984) "Analyse numérique des équations différentielles", Ed. Masson.

Teorema 3.4.4 *A) Una condición necesaria y suficiente para que un método de de Runge-Kutta definido por (3.5) y (3.7) sea consistente de primer orden es que:*

$$\sum_{j=1}^{p} a_j = 1. \tag{3.20}$$

B) Una condición necesaria y suficiente para que un método de Runge-Kutta definido por (3.5) y (3.7) sea consistente de segundo orden es que se verifique (3.20) y:

$$\sum_{j=1}^{p} a_j \, c_j = \sum_{j=1}^{p} a_j \, \left(\sum_{k=1}^{p} b_{j,k} \right) = \frac{1}{2}. \tag{3.21}$$

C) Suponiendo que:

$$\sum_{k=1}^{p} b_{j,k} = c_j, \tag{3.22}$$

entonces una condición necesaria y suficiente para que el método de Runge-Kutta correspondiente sea de tercer orden es que se verifiquen (3.20), (3.21) y:

$$\sum_{j=1}^{p} a_j \, \left(\sum_{k=1}^{p} b_{j,k} \, c_k \right) = \frac{1}{6}, \tag{3.23}$$

y

$$\sum_{j=1}^{p} a_j \, c_j^2 = \frac{1}{3} \tag{3.24}$$

D) Suponiendo que se verifique (3.22) una condición necesaria y suficiente para que el método de Runge-Kutta correspondiente sea de cuarto orden es que se verifiquen (3.20), (3.21), (3.23), (3.24) y:

$$\sum_{j=1}^{p} a_j \, \left(\sum_{k=1}^{p} b_{j,k} \, c_k^2 \right) = \frac{1}{2} \tag{3.25}$$

$$\mathbf{a}^T \, \mathbf{B} \, \mathbf{B} \, \mathbf{C} \, \mathbf{e} = \frac{1}{8} \tag{3.26}$$

y

$$\mathbf{a}^T \, \mathbf{C} \, \mathbf{B} \, \mathbf{C} \, \mathbf{e} = \frac{1}{8} \tag{3.27}$$

donde $\mathbf{e}$ *es el vector de* $\mathbb{R}^p$ *que tiene todas sus componentes iguales a la unidad.*

Demostración:

Es una consecuencia de los teoremas vistos para un método genérico. Una demostración detallada puede encontrarse, por ejemplo, en Crouzeix y Mignot[23].

c.q.d.

Teorema 3.4.5 *Si las fórmulas de integración utilizadas en la definición de los valores intermedios* $y_{n,j}$ $(j = 1, ..., p)$ *en la expresión (3.7) corresponden a fórmulas de integración de orden* $(k-2)$ *y la fórmula de integración utilizada en la determinación del valor* y_{n+1} *en la expresión (3.5) es una fórmula de orden* $(k-1)$ *entonces el método de Runge-Kutta correspondiente es de orden* k.

Demostración:

Es una consecuencia del teorema anterior. Una demostración detallada puede encontrarse, por ejemplo, en Crouzeix y Mignot[24].

c.q.d.

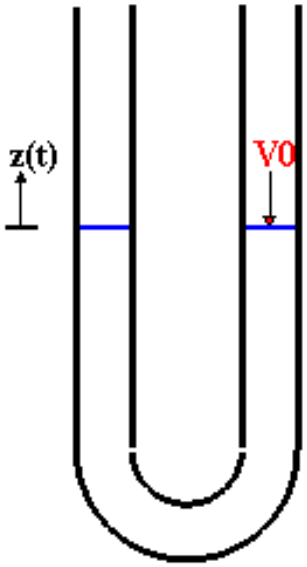
Los teoremas anteriores nos llevan a la conclusión de que, por ejemplo, el método de Runge-Kutta clásico descrito en el apartado 3.4.2 es de orden 4 y el ejemplo tratado en el apartado (3.4.2) es un método de tercer orden.

3.4.5. Aplicación del método de Runge-Kutta clásico a la resolución de un ejemplo.

Un líquido de densidad ρ y viscosidad dinámica μ se encuentra en reposo contenido en un tubo de sección recta circular de diámetro constante d y con forma de U. La longitud que ocupa el fluido, medida en el eje del tubo, es denotada por L.

En el instante $t_0 = 0$ (s) un pistón golpea una de las superficies libres del fluido imprimiéndole una velocidad V_0. Como consecuencia de ello la otra superficie libre comienza a ascender hasta que, por efecto de la fuerza gravitatoria, se detiene su ascenso. En ese momento el líquido de esa rama del tubo comienza a descender rebasando la posición en que inicialmente se encontraba en equilibrio hasta que el efecto de la gravedad sobre la otra rama del fluido detiene el descenso y el líquido vuelve a ascender. Se genera así un movimiento oscilatorio del fluido en torno a su posición de equilibrio inicial (que, con el transcurrir del tiempo, volverá a recuperarse). Si se denota por g a la aceleración gravitatoria y por $z(t)$ al desplazamiento de la superficie libre del fluido, despreciando el efecto del rozamiento del líquido con las paredes

[23] Crouzeix, M., Mignot A.L. (1984) "Analyse numérique des équations différentielles", Ed. Masson.



del tubo, el proceso anterior puede modelizarse mediante el problema de valor inicial siguiente:

$$\begin{cases} z''(t) + A.z'(t) + B.z(t) & = 0, \quad t > 0 \\ z(0) = 0, \qquad z'(0) = V_0 \end{cases}$$

donde A y B son dos constantes dadas por:

$$A = \frac{32\,\mu}{\rho\, d^2}, \quad B = \frac{2\,g}{L}$$

La figura siguiente recoge un esquema del proceso descrito:

El problema de valor inicial antes planteado puede transformarse en otro regido por un sistema de dos ecuaciones diferenciales de primer orden en la forma:

$$\begin{cases} y_1'(t) & = y_2(t) & t > 0 \\ y_2'(t) & = -B\, y_1(t) - A\, y_2(t) & t > 0 \\ y_1(0) & = 0, \qquad y_2(0) = V_0 \end{cases}$$

donde se ha denotado por $y_1(t) = z(t)$ (la función de desplazamiento) y por $y_2(t) = z'(t)$ (la función de velocidad).

Consideremos un fluido para el que $\mu = 0{,}001\,\mathrm{Kg/(m\,s)}$, $\rho = 1\,\mathrm{Kgm^{-3}}$ y tomemos como valor de $g = 9{,}81\,\mathrm{ms^{-2}}$, $d = 0{,}2\mathrm{m}$, $L = 3\mathrm{m}$ y $V_0 = 1\mathrm{m/s}$

y apliquemos el algoritmo del método de Runge-Kutta clásico antes descrito, con paso de integración constante $h = 0{,}05$, a la resolución de este problema.

Para ello llamaremos:

$$A = \frac{32\,0{,}001}{1\,0{,}04} = 0{,}8, \qquad B = \frac{2\,9{,}81}{3} = 6{,}54$$

$$\mathbf{y}^{(0)} = \left\{ \begin{array}{c} 0 \\ 1 \end{array} \right\}, \qquad \mathbf{f}(t, \mathbf{y}(t)) = \left\{ \begin{array}{l} y_2(t) \\ -6{,}54\, y_1(t) - 0{,}8\, y_2(t) \end{array} \right\}$$

El primer paso del algoritmo nos conduce a:

$$\mathbf{W1} = \mathbf{f}(0, \{0,1\}^T) = \left\{ \begin{array}{l} 1 \\ -6{,}54\,0 - 0{,}8\,1 \end{array} \right\} = \left\{ \begin{array}{c} 1 \\ -0{,}8 \end{array} \right\}$$

$$\mathbf{W2} = \mathbf{f}\left(0 + \frac{0{,}05}{2}, \left\{ \begin{array}{c} 0 \\ 1 \end{array} \right\} + \frac{0{,}05}{2} \left\{ \begin{array}{c} 1 \\ -0{,}8 \end{array} \right\}\right) = \mathbf{f}\left(0{,}025, \left\{ \begin{array}{c} 0{,}025 \\ 0{,}98 \end{array} \right\}\right)$$

$$= \left\{ \begin{array}{l} 0{,}98 \\ -6{,}54\,0{,}025 - 0{,}8\,0{,}98 \end{array} \right\} = \left\{ \begin{array}{c} 0{,}98 \\ -0{,}9475 \end{array} \right\}$$

$$\mathbf{W3} = \mathbf{f}\left(0 + \frac{0{,}05}{2}, \left\{ \begin{array}{c} 0 \\ 1 \end{array} \right\} + \frac{0{,}05}{2} \left\{ \begin{array}{c} 0{,}98 \\ -0{,}9475 \end{array} \right\}\right) = \mathbf{f}\left(0{,}025, \left\{ \begin{array}{c} 0{,}0245 \\ 0{,}9763125 \end{array} \right\}\right)$$

$$= \left\{ \begin{array}{l} 0{,}9763125 \\ -6{,}54\,0{,}0245 - 0{,}8\,0{,}9763125 \end{array} \right\} = \left\{ \begin{array}{c} 0{,}9763125 \\ -0{,}94128 \end{array} \right\}$$

$$\mathbf{W4} = \mathbf{f}\left(0 + 0{,}05, \left\{ \begin{array}{c} 0 \\ 1 \end{array} \right\} + 0{,}05 \left\{ \begin{array}{c} 0{,}9763125 \\ -0{,}94128 \end{array} \right\}\right) = \mathbf{f}\left(0{,}05, \left\{ \begin{array}{c} 0{,}048815625 \\ 0{,}952936 \end{array} \right\}\right)$$

$$= \left\{ \begin{array}{l} 0.{,}952936 \\ -6{,}54\,0{,}952936 - 0{,}8\,0{,}952936 \end{array} \right\} = \left\{ \begin{array}{c} 0{,}952936 \\ -1{,}0816029875 \end{array} \right\}$$

con lo que finalmente

$$\mathbf{y}^{(1)} = \mathbf{y}^{(0)} + \frac{h_0}{6}\,(\mathbf{W1} + 2\,(\mathbf{W2} + \mathbf{W3}) + \mathbf{W4}) = \left\{ \begin{array}{c} 0 \\ 1 \end{array} \right\} +$$

$$\frac{0{,}05}{6}\left(\left\{ \begin{array}{c} 1 \\ -0{,}8 \end{array} \right\} + 2\left(\left\{ \begin{array}{c} 0{,}98 \\ -0{,}9475 \end{array} \right\} + \left\{ \begin{array}{c} 0{,}97631 \\ -0{,}94128 \end{array} \right\}\right) + \left\{ \begin{array}{c} 0{,}95293 \\ -1{,}08160 \end{array} \right\}\right) =$$

$$= \left\{ \begin{array}{c} 0{,}04887 \\ 0{,}95284 \end{array} \right\}$$

Es decir que al cabo de 0,05 segundos el líquido habrá ascendido 0,04887 metros y se estará moviendo con una velocidad de 0,95284 m/s.

El segundo paso consistirá en:

$$\mathbf{W1} = \mathbf{f}\left(0{,}05, \left\{ \begin{array}{l} 0{,}04887 \\ 0{,}95284 \end{array} \right\}\right) =$$

$$\left\{ \begin{array}{l} 1 \\ -6{,}54 \cdot 0{,}04887 - 0{,}8 \cdot 0{,}95284 \end{array} \right\} = \left\{ \begin{array}{l} 0{,}95284 \\ -1{,}08194 \end{array} \right\}$$

$$\mathbf{W2} = \mathbf{f}\left(0{,}05 + \frac{0{,}05}{2}, \left\{ \begin{array}{l} 0{,}04887 \\ 0{,}95284 \end{array} \right\} + \frac{0{,}05}{2} \left\{ \begin{array}{l} 0{,}95284 \\ -1{,}08194 \end{array} \right\}\right) =$$

$$= \mathbf{f}\left(0{,}075, \left\{ \begin{array}{l} 0{,}07270 \\ 0{,}92579 \end{array} \right\}\right) =$$

$$\left\{ \begin{array}{l} 0{,}92579 \\ -6{,}54 \cdot 0{,}07270 - 0{,}8 \cdot 0{,}92579 \end{array} \right\} = \left\{ \begin{array}{l} 0{,}92579 \\ -1{,}21609 \end{array} \right\}$$

$$\mathbf{W3} = \mathbf{f}\left(0{,}05 + \frac{0{,}05}{2}, \left\{ \begin{array}{l} 0{,}048879 \\ 0{,}95284 \end{array} \right\} + \frac{0{,}05}{2} \left\{ \begin{array}{l} 0{,}92579 \\ -1{,}21609 \end{array} \right\}\right) =$$

$$= \mathbf{f}\left(0{,}075, \left\{ \begin{array}{l} 0{,}07202 \\ 0{,}92243 \end{array} \right\}\right) =$$

$$\left\{ \begin{array}{l} 0{,}92243 \\ -6{,}54 \cdot 0{,}07202 - 0{,}8 \cdot 0{,}92243 \end{array} \right\} = \left\{ \begin{array}{l} 0{,}92243 \\ -1{,}20899 \end{array} \right\}$$

$$\mathbf{W4} = \mathbf{f}\left(0{,}05 + 0{,}05, \left\{ \begin{array}{l} 0{,}048879 \\ 0{,}95284 \end{array} \right\} + 0{,}05 \left\{ \begin{array}{l} 0{,}92243 \\ -1{,}20899 \end{array} \right\}\right) =$$

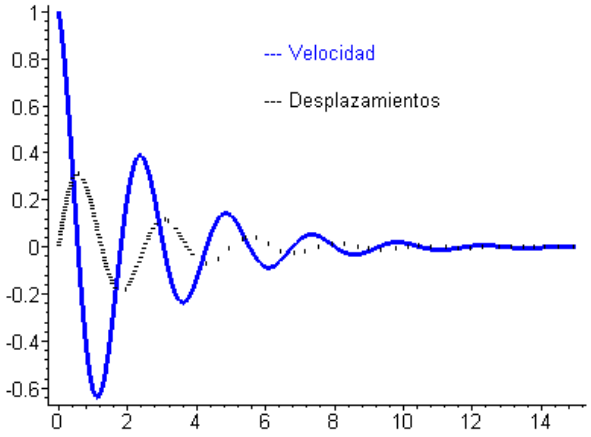
$$= \mathbf{f}\left(0{,}1, \left\{ \begin{array}{l} 0{,}09500 \\ 0{,}89239 \end{array} \right\}\right) =$$

$$\left\{ \begin{array}{l} 0{,}89239 \\ -6{,}54 \cdot 0{,}09500 - 0{,}8 \cdot 0{,}89239 \end{array} \right\} = \left\{ \begin{array}{l} 0{,}89239 \\ -1{,}33522 \end{array} \right\}$$

con lo que finalmente

$$\mathbf{y}^{(2)} = \mathbf{y}^{(2)} + \frac{h_2}{6} \left(\mathbf{W1} + 2\left(\mathbf{W2} + \mathbf{W3}\right) + \mathbf{W4}\right) =$$

$$= \left\{ \begin{array}{l} 0{,}04887 \\ 0{,}95284 \end{array} \right\} + \frac{0{,}05}{6} \left\{ \begin{array}{l} 0{,}95284 \\ -1{,}08194 \end{array} \right\} +$$



$$+\frac{0{,}05}{6}\left(2\left(\begin{Bmatrix} 0{,}92579 \\ -1{,}21609 \end{Bmatrix}+\begin{Bmatrix} 0{,}92243 \\ -1{,}20899 \end{Bmatrix}\right)+\begin{Bmatrix} 0{,}89239 \\ -1{,}33522 \end{Bmatrix}\right)=$$

$$=\begin{Bmatrix} 0{,}09506 \\ 0{,}89227 \end{Bmatrix}$$

Es decir, que al cabo de 0,1 segundos el líquido habrá ascendido 0,09506 metros y se estará moviendo con una velocidad de 0,89227 m/s.

Con ayuda de un programa recogiendo este método se han realizado 300 pasos de tiempo obteniéndose como valores aproximados de los desplazamientos y velocidades los que se recogen en la gráfica siguiente (en punteado los desplazamientos y en trazo continuo las velocidades del fluido):

Puedes observar como efectivamente el sistema tiende a volver a la situación de reposo (desplazamiento nulo y velocidad nula).

3.5. Introducción a los métodos multipaso

3.5.1. Introducción.

Los métodos vistos hasta ahora para resolver numéricamente el problema de encontrar valores aproximados en ciertos instantes t^n de la solución al problema:

$$(P.V.I.)\begin{cases} y'(t) & = f(t, y(t) \quad t \in [t_0, t_N] \\ y(t_0) & = y_0 \end{cases}$$

utilizaban únicamente el valor de $y_n \approx y(t_n)$ para estimar $y_{n+1} \approx y(t_{n+1})$. De forma natural surge entonces la cuestión de si es posible aprovechar la información procedente de las soluciones (aproximadas) obtenidas en los instantes

t_{n-1}, t_{n-2}, ..., t_{n-p} para "mejorar" la solución obtenida en t_{n+1}. Surgen de esta manera los denominados **métodos multipaso** (o métodos de pasos ligados o, también, métodos de pasos múltiples). De ellos los más clásicos son los denominados métodos de Adams y sobre ellos centraremos nuestra atención.

Definition 2 *Se denomina método de r pasos a aquel que para obtener la solución aproximada y_{n+1} del problema (P.V.I.) en el instante t_{n+1} utiliza los valores aproximados de la solución obtenidos en los r instantes anteriores, es decir, y_n, y_{n-1}, ..., y_{n-r}.*

3.5.2. Los métodos de Adams.

Consideremos nuevamente el problema $(P.V.I.)$ donde supondremos que la función f es lipschitciana respecto a su segunda variable. Sea ademas dada una colección de instantes intermedios:

$$t_0 < t_1 < ... < t_n < ... < t_N$$

y denotemos como hemos venido haciendo anteriormente por $h_n = t_{n+1} - t_n$.

Según se vio en apartados anteriores la solución $y(t)$ en el instante t_{n+1} puede estimarse mediante:

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt$$

expresión que podrá aproximarse por:

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} p_n(t)\, dt$$

donde $p_n(t)$ es una cierta función que aproxime a la función $f(t, y(t))$. En los métodos de Adams esta función $p_n(t)$ es un polinomio soportado en los instantes de cálculo precedentes. Según el polinomio escogido, los métodos de Adams se dividen en dos grandes grupos como a continuación se detalla.

Métodos de Adams-Bashforth de (r+1) pasos.

Suponiendo conocidos los valores $f_{n-r} = f(t_{n-r}, y_{n-r})$, $f_{n-r+1} = f(t_{n-r+1}, y_{n-r+1})$, ..., $f_n = f(t_n, y_n)$ los métodos de Adams-Bashforth de $(r+1)$ pasos construyen el polinomo $p_n(t)$ como el polinomio interpolador de Lagrange de la función f en el soporte $t_{n-r}, ..., t_n$, es decir como aquel polinomio de grado r que verifica:

$$p_n(t_{n-i}) = f_{n-i} \quad (i = 0, ..., r)$$

NOTA:

Consúltese Conde y Schiavi[24] *para la determinación del polinomio interpolador de Lagrange.*

Definition 3 *Se denominan métodos de Adams-Basforth a aquellos que se obtienen al desarrollar la fórmula:*

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} p_n(t)\, dt$$

donde $p_n(t)$ es el polinomio de grado r que verifica las igualdades:

$$p_n(t_{n-i}) = f_{n-i} \quad (i = 0, ..., r)$$

El polinomio $p_n(t)$ anterior puede expresarse como:

$$p_n(t) = \sum_{i=0}^{r} f_{n-i}\, L_{n,i,r}(t)$$

donde

$$L_{n,i,r}(t) = \prod_{j=0//j\neq i}^{r} \frac{t - t_{n-j}}{t_{n-i} - t_{n-j}}$$

por lo que el método de Adams-Basforth de $(r+1)$ pasos puede escribirse como:

$$y_{n+1} = y_n + h_n \sum_{i=0}^{r} b_{n,i,r}\, f_{n-i}$$

donde

$$b_{n,i,r} = \frac{1}{h_n} \int_{t_n}^{t_{n+1}} L_{n,i,r}(t)\, dt$$

NOTA:

Obsérvese que los métodos de Adams-Bashforth son métodos explícitos y que necesitan para ser aplicados conocer, al menos, los valores aproximados

[24] Conde, C. y Schiavi, E. (2000) "Elementos de Matemáticas: Guiones de los temas de la asignatura". Apuntes. Universidad Rey Juan Carlos.

y_0, y_1, ..., y_r. *Estos deberán ser obtenidos por métodos que utilicen un menor número de pasos.*

En el caso particular de que todos los pasos temporales tengan el mismo tamaño h los coeficientes $b_{n,i,r}$ de las expresiones anteriores se hacen independientes del índice n (pues en todos los pasos de tiempo se obtendrán los mismos coeficientes). En dicho caso los valores de los coeficientes aparecen tabulados en numerosas referencias (como por ejemplo Gear[25] o Crouzeix y Mignot[26]). Algunos de ellos se recogen en la tabla siguiente:

	$b_{0,r}$	$b_{1,r}$	$b_{2,r}$	$b_{3,r}$	$b_{4,r}$	$b_{5,r}$	$b_{6,r}$	$\sum_{i=0}^{r} \lvert b_{i,r}\rvert$
$r=0$	1							1
$r=1$	$\frac{3}{2}$	$\frac{-1}{2}$						2
$r=2$	$\frac{23}{12}$	$\frac{-4}{3}$	$\frac{5}{12}$					3,66..
$r=3$	$\frac{55}{24}$	$\frac{-59}{24}$	$\frac{37}{24}$	$\frac{-3}{8}$				6,66...
$r=4$	$\frac{1901}{720}$	$\frac{-1387}{360}$	$\frac{109}{30}$	$\frac{-637}{360}$	$\frac{251}{729}$			12,24..
$r=5$	$\frac{4277}{1440}$	$\frac{-7923}{1440}$	$\frac{4991}{720}$	$\frac{-3649}{720}$	$\frac{959}{480}$	$\frac{-95}{288}$		22,14..
$r=6$	$\frac{198721}{60480}$	$\frac{-18637}{2520}$	$\frac{2235183}{20160}$	$\frac{-10754}{945}$	$\frac{135713}{20160}$	$\frac{-5603}{2520}$	$\frac{1987}{60480}$	43,75..

Métodos de Adams-Moulton de (r+1) pasos

En las mismas condiciones que en el apartado anterior, los métodos de Adams-Moulton consideran el polinomio interpolador $p_n(t)$ de grado $(r+1)$ y verificando ahora que:

$$p_n(t_{n-i}) = f_{n-i} \qquad (i = 0, 1, ..., r)$$

$$p_n(t_{n+1}) = f_{n+1} = f(t_{n+1}, y_{n+1})$$

Obsérvese que ahora el polinomio se hace depender del propio valor que se quiere estimar. Por ello estos esquemas constituyen una familia de métodos implícitos.

[25] Gear, C. W. (1971) "Numerical Initial Value Problems in Ordinary Differential Equations". Ed. Prentice Hall.

[26] Crouzeix, M., Mignot A.L. (1984) "Analyse numérique des équations différentielles", Ed. Masson.

También ahora los métodos podrán escribirse en la forma:

$$y_{n+1} - h_n\, c_{n,-1,r}\, f(t_{n+1}, y_{n+1}) = y_n + h_n \sum_{i=0}^{r} c_{n,i,r}\, f_{n-i}$$

donde

$$c_{n,i,r} = \frac{1}{h_n} \int_{t_n}^{t_{n+1}} L_{n,i,r}(t)\, dt \quad (i = -1, 0, 1, ..., r)$$

siendo

$$L_{n,i,r}(t) = \prod_{j=-1 j \neq i}^{r} \frac{t - t_{n-j}}{t_{n-i} - t_{n-j}} \qquad (i = -1, 0, 1, ..., r)$$

NOTAS:

El análisis de los métodos multipaso puede realizarse también a través del estudio de la consistencia y estabilidad de dichos esquemas. Su estudio detallado desborda la disponibilidad de tiempo para este curso por lo que remitimos al lector interesado a la bibliografía sobre el tema para tal menester.

En Crouzeix y Mignot[27] *o en Gear*[28] *pueden encontrarse para los métodos de Adams-Moulton tablas similares a la antes escrita para los métodos de Adams-Bashforth recogiendo los valores de los coeficientes de los esquemas para distintos valores de r en el caso de paso de discretización constante.*

[27] Crouzeix, M., Mignot A.L. (1984) "Analyse numérique des équations différentielles", Ed. Masson.

[28] Gear, C. W. (1971) "Numerical Initial Value Problems in Ordinary Differential Equations". Ed. Prentice Hall.

3.6. Bibliografía.

- Apostol, T. M. (1997) "Linear Algebra. A first course with applications to differential equations". Ed. John Wiley & Sons, Inc.
- Burden R.L., Douglas Faires J. (1998) "Análisis Numérico", International Thomson Editores, 6. edición.
- Conde, C. y Schiavi, E. (2000) "Elementos de Matemáticas: Guiones de los temas de la asignatura". Apuntes. Universidad Rey Juan Carlos.
- Crouzeix, M., Mignot A.L. (1984) "Analyse numérique des équations différentielles", Ed. Masson.
- Gear, C. W. (1971) "Numerical Initial Value Problems in Ordinary Differential Equations". Ed. Prentice Hall.
- Guzmán, M. de (1987). "Ecuaciones diferenciales ordinarias. Teoría de estabilidad y control" (3^a reimpresión). Ed. Alhambra Universidad.
- Hanna, O.T. y Sandall, O.C. (1995). "Computational Methods in Chemical Engineering". Ed. Prentice Hall International Editions.
- Kincaid, D. y Cheney, W. (1994). "Análisis numérico. Las matemáticas del cálculo científico". Ed. Addison-Wesley Iberoamericana.
- Marcellán, F., Casasús, L. y Zarzo, A. (1991). "Ecuaciones diferenciales. Problemas lineales y aplicaciones". Ed. McGraw Hill.
- Martínez, C. y Sanz, M.A. (1991). "Introducción a las ecuaciones diferenciales ordinarias". Ed. Reverté.
- Schiavi, E., Muñoz Montalvo, A.I., Conde, C. (2012). Métodos Matemáticos para los Grados en Ingeniería. Primera parte: teoría. Ed. Dykinson, Textos Docentes 31, Universidad Rey Juan Carlos, ISBN: 978-84-15454-58-8.
- Shampine, L.F. (1.994) "Numerical solution of ordinary differential equations". Ed. Chapman&Hall Mathematics.
- Shampine, L.F., Allen, R.C. Jr. y Pruess, S. (1997). "Fundamentals of numerical computing". Ed. John Wiley & Sons, Inc.
- Sibony, M. y Mardon, J. Cl. (1982). "Analyse numérique II: Approximation et equations differentielles". Ed. Hemann.

- Stoer, J. y Bulirsch, R. (1993)"Introduction to Numerical Analysis" (2^a edición). Ed. Springer Verlag.
- Zill, D. G. (1997). Ecuaciones diferenciales con aplicaciones de modelado. (VI edición) Ed. International Thomson editores.

Capítulo 4

Métodos en diferencias finitas para la resolución de problemas de contorno

4.1. Presentación y generalidades

4.1.1. El problema modelo sobre el que se plantearán los esquemas numéricos.

En este tema estudiaremos métodos en diferencias finitas para la resolución de problemas de transporte formulados como sigue:

Hallar una función $u(x,t)$ *que satisfaga:*

$$\begin{cases} \dfrac{\partial\left(a(\mathbf{x},t,u)u(\mathbf{x},t)\right)}{\partial t} - \nabla \bullet \left(\left[\mathbf{D}(\mathbf{x},t,u)\right]\nabla u(\mathbf{x},t)\right) + \\ +\nabla \bullet \left(\overrightarrow{\mathbf{V}}(\mathbf{x},t,u)u(\mathbf{x},t)\right) + q(x,t,u)u(\mathbf{x},t) = f(\mathbf{x},t,u) & \mathbf{x}\in\Omega, t\in(0,T) \\ u(\mathbf{x},t) = u_D(\mathbf{x},t) & \mathbf{x}\in\Gamma_D\times(0,T) \\ \left(-\left[\mathbf{D}(\mathbf{x},t,u)\right]\nabla u(\mathbf{x},t) + \overrightarrow{\mathbf{V}}(\mathbf{x},t,u)u(\mathbf{x},t)\right)\bullet\overrightarrow{\mathbf{n}} = g(\mathbf{x},t,u) & \mathbf{x}\in\Gamma_N\times(0,T) \\ u(\mathbf{x},0) = u^0(\mathbf{x}) & \mathbf{x}\in\overline{\Omega} \end{cases}$$

donde $\mathbf{x}\in\Omega\subset\mathbb{R}^d$ $(d=1,2\ o\ 3)$, *siendo* Ω *un abierto de frontera* $\Gamma=\Gamma_D\cup\Gamma_N$ *donde* Γ_D *y* Γ_N *tienen interiores vacíos,* $a(\mathbf{x},t,u)$ *es la función denominada "factor de retardo",* $[\mathbf{D}(\mathbf{x},t,u)]$ *es la representación mediante una matriz*

de dimensiones $(d,\ d)$ *del tensor (de segundo orden) de "difusión-dispersión",* $\overrightarrow{\mathbf{V}}(\mathbf{x},t,u)$ *es el campo de velocidades de convección,* $q(\mathbf{x},t,u)$ *es la función de "reacción química",* $f(\mathbf{x},t,u)$ *es la función de "términos fuentes",* $u_D(\mathbf{x},t)$ *es la función que nos permite imponer la condiciones de tipo Dirichlet en el tramo de frontera* Γ_D, $g(\mathbf{x},t,u)$ *es la función que nos permite imponer la condición de flujo en el tramo de frontera* Γ_N *y* $u^0(\mathbf{x})$ *es la función que asigna valores iniciales en el dominio. Todas estas funciones se supondrán conocidas. Por último,* $\overrightarrow{\mathbf{n}}$ *es el vector normal unitario exterior al dominio* Ω *en cada punto de la frontera y* T *es el instante de cálculo hasta el cual se extiende nuestro estudio.*

NOTA:

Obviamente existen muchas otras ecuaciones en derivadas parciales de interés en la Ingeniería (ecuaciones de Navier-Stokes para el estudio de los movimientos de fluidos, ecuaciones de Maxwell para el estudio de fenómenos electromagnéticos, ecuación de ondas para el movimiento ondulatorio, ecuaciones de la elastodinámica para el estudio del comportamiento mecánico de los materiales, ecuaciones de Euler para el estudio de fluidos compresibles, etc...). El estudio detallado de todas ellas desborda ampliamente los objetivos de este curso. No obstante, a estas ecuaciones serán aplicables muchas de las técnicas (y de los fundamentos en que se basan) que presentaremos sobre la ecuación del transporte.

Será fecuente que el comportamiento de las distintas formas de aproximar la solución del problema anterior se estudie sobre casos particulares de la ecuación. Entre ellos distinguiremos:

a) Problemas estacionarios: es decir, independientes del tiempo, que serán de la forma:

$$-\nabla \bullet \left([\mathbf{D}(\mathbf{x},u)]\,\nabla u(\mathbf{x})\right) + \nabla \bullet \left(\overrightarrow{\mathbf{V}}(\mathbf{x},u)u(\mathbf{x})\right) + q(\mathbf{x})u(\mathbf{x}) = f(\mathbf{x},u),\ \mathbf{x}\in\Omega$$

acompañados de las oportunas condiciones de contorno. En ocasiones será interesante simplificar aún más el problema estacionario reduciéndolo a uno de los siguientes:

a-1) Problema difusivo:

$$-\nabla \bullet \left([\mathbf{D}(\mathbf{x},u)]\,\nabla u(\mathbf{x})\right) = f(\mathbf{x}) \quad \mathbf{x}\in\Omega$$

acompañado de las condiciones de contorno pertinentes. En el caso de que el tensor de difusividad-dispersividad pudiera reducirse a una constante (D) multiplicando a la matiz identidad, se tendría la ecuación de Poisson:

$$-\Delta u(\mathbf{x}) = \frac{1}{D} f(x) \quad \mathbf{x}\in\Omega$$

que, si $f(x) = 0$ es la conocida ecuación de Laplace. Este tipo de ecuaciones modeliza el estado estacionario que, en su caso, alcanzarán los procesos de transporte conductivo de calor o el transporte difusivo de materia y es el "prototipo" que suele escogerse para el estudio de las denominadas ecuaciones en derivadas parciales de tipo elíptico. La aproximación de este tipo de ecuaciones por distintos esquemas nos permitirá analizar de forma sencilla si tales esquemas satisfacen teoremas y leyes que verifican las respectivas soluciones analíticas (tales como el principio del máximo que puedes encontrar en el anexo).

a-2) Problema convectivo:

$$\nabla \bullet \left(\overrightarrow{\mathbf{V}}(\mathbf{x})u(\mathbf{x})\right) = f(\mathbf{x}) \quad \mathbf{x} \in \Omega$$

acompañado, obviamente, de las pertinentes condiciones de contorno. Esta ecuación es una EDP de primer orden y modeliza el estado estacionario que, en su caso, alcanzarán los procesos de transporte convectivo de calor o materia. La aproximación de este tipo de ecuaciones por distintos esquemas nos permitirá analizar de forma sencilla las diferentes estrategias (centradas y descentradas) que pueden seguirse para modelizar fenómenos en los que se produce un "arrastre" por un fluido.

a-3) Problema difusivo-convectivo:

$$-\nabla \bullet \left([\mathbf{D}(\mathbf{x},u)]\,\nabla u(\mathbf{x})\right) + \nabla \bullet \left(\overrightarrow{\mathbf{V}}(\mathbf{x})u(\mathbf{x})\right) = f(\mathbf{x}) \quad \mathbf{x} \in \Omega$$

acompañada de las pertinentes condiciones de contorno. La aproximación de este tipo de ecuaciones por distintos esquemas nos permitirá poner de manifiesto el comportamiento de los distintos esquemas ante "capas límite" y el papel que juega la relación entre el "peso" que tenga en la ecuación el término convectivo frente al término convectivo.

b) Problemas evolutivos. En ellos el tiempo t será una variable independiente más. Será frecuente considerar también en este caso los fenómenos de difusión y los de convección de forma separada antes de "juntarlos" en una única ecuación. En este sentido se considerará las EDP:

b-1) Ecuación de difusión:

$$\frac{\partial u}{\partial t}(\mathbf{x},t) - \nabla \bullet \left([\mathbf{D}(\mathbf{x},t,u)]\,\nabla u(\mathbf{x},t)\right) = f(\mathbf{x},t,u) \quad x \in \Omega,\ \ 0 < t < T$$

El caso en el que el tensor de difusividad pueda reducirse a una constante D, que se trabaje en dominios de una dimensión espacial y que la función de términos fuentes $f(\mathbf{x},t,u)$ sea nula, nos conduce a la ecuación "prototipo" de las

denominadas ecuaciones en derivadas parciales parabólicas y que es conocida con el nombre de "ecuación del calor":

$$\frac{\partial u}{\partial t}(x,t) = D\frac{\partial^2 u}{\partial x^2}(x,t),\ 0 < x < L,\ 0 < t < T$$

Sobre este tipo de ecuaciones podrán ponerse de manifiesto las muy diferentes formas que hay de combinar discretizaciones espaciales y discretizaciones temporales, analizándose las relaciones que debe haber entre el paso de discretización temporal y el paso de discretización espacial, para que los esquemas verifiquen ciertas propiedades que cumplen las soluciones analíticas (tales como el principio del máximo) o simplemente para que las soluciones aproximadas no "exploten" con el transcurrir del tiempo (es decir, permanezcan estables).

b-2) Ecuación de convección:

$$\frac{\partial u}{\partial t}(\mathbf{x},t) + \nabla \bullet \left(\overrightarrow{\mathbf{V}}(\mathbf{x},t)u(\mathbf{x},t)\right) = f(\mathbf{x},t,u),\ x \in \Omega,\ 0 < t < T.$$

Este tipo de ecuaciones son ecuaciones en derivadas parciales hiperbólicas de primer orden. Su estudio nos permitirá introducir de forma simple el método de las características y plantear un amplio número de esquemas (centrados, descentrados, explícitos o implícitos) para la resolución de este tipo de ecuaciones. En el caso de considerar dominios espaciales unidimensionales, términos fuentes nulos y velocidades constantes en espacio se obtendrá la ecuación de advección unidimensional:

$$\frac{\partial u}{\partial t}(x,t) + V(x)\frac{\partial u}{\partial x}(x,t) = 0,\ 0 < x < L,\ 0 < t < T,$$

que es el "prototipo" de las ecuaciones hiperbólicas de primer orden.

b-3) Ecuación de difusión-convección:

$$\frac{\partial u}{\partial t}(x,t) - \nabla \bullet \left([\mathbf{D}(\mathbf{x},t,u)]\,\nabla u(\mathbf{x},t)\right) + \nabla \bullet \left(\overrightarrow{\mathbf{V}}(\mathbf{x},t)u(\mathbf{x},t)\right) = 0,$$
$$0 < x < L,\ 0 < t < T.$$

En este tipo de ecuaciones, y especialmente en el caso de dominios espaciales unidimensionales, será cómodo ilustrar la influencia entre el tamaño de paso espacial, el temporal y los pesos asignados al transporte convectivo frente al difusivo (a través de los denominados números de Courant y de Peclet que posteriormente se introducirán).

Para aligerar la escritura, a la hora de escribir las ecuaciones omitiremos con frecuencia las variables de las que depende cada parámetro o función. En este

sentido, a veces escribiremos q en lugar de $q(\mathbf{x},t,u)$, $\overrightarrow{\mathbf{V}}$ en lugar de $\overrightarrow{\mathbf{V}}(\mathbf{x},t,u)$ y f en lugar de $f(\mathbf{x},t,u)$.

Una última observación a realizar consiste en que, cuando el factor de retardo a sea independiente de u (retardo lineal) la ecuación de transporte que se recogía en la formulación anteriormente hecha, habitualmente la modificaremos según el proceso siguiente:

$$\frac{\partial(au)}{\partial t} - \nabla \bullet ([\mathbf{D}]\nabla u) + \nabla \bullet \left(\overrightarrow{\mathbf{V}} u\right) + qu = f \Rightarrow$$

$$a\frac{\partial u}{\partial t} + \frac{\partial a}{\partial t}u - \sum_{i=1}^{d}\left(\sum_{j=1}^{d}\left(D_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} + \frac{\partial D_{i,j}}{\partial x_i}\frac{\partial u}{\partial x_j}\right)\right) +$$

$$+\sum_{i=1}^{d}\left(V_i\frac{\partial u}{\partial x_i} + \frac{\partial V_i}{\partial x_i}u\right) + qu = f \Rightarrow$$

$$a\frac{\partial u}{\partial t} - \sum_{i=1}^{d}\left(\sum_{j=1}^{d} D_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j}\right) + \sum_{i=1}^{d}\left(\left(V_i - \sum_{j=1}^{d}\frac{\partial D_{i,j}}{\partial x_i}\right)\frac{\partial u}{\partial x_i}\right) +$$

$$+\left(q + \frac{\partial a}{\partial t} + \sum_{i=1}^{d}\frac{\partial V_i}{\partial x_i}\right)u = f \Rightarrow$$

$$\Rightarrow a\frac{\partial u}{\partial t} - \sum_{i=1}^{d}\left(\sum_{j=1}^{d} D_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j}\right) + \overrightarrow{\widetilde{\mathbf{V}}}\nabla u + \widetilde{q}u = f$$

donde

$$\widetilde{q} = \left(q + \frac{\partial a}{\partial t} + \sum_{i=1}^{d}\frac{\partial V_i}{\partial x_i}\right)$$

y

$$\overrightarrow{\widetilde{\mathbf{V}}} = \left\{\begin{array}{c} V_1 - \sum\limits_{j=1}^{d}\dfrac{\partial D_{1,j}}{\partial x_i} \\ V_2 - \sum\limits_{j=1}^{d}\dfrac{\partial D_{2,j}}{\partial x_i} \\ V_3 - \sum\limits_{j=1}^{d}\dfrac{\partial D_{3,j}}{\partial x_i} \end{array}\right\}$$

La peculiaridad que tiene la formulación anterior es que las derivadas que en ella aparecen explícitamente actúan únicamente sobre la función u. Incluso, cuando sea posible, será cómodo dividir toda la ecuación por a obteniendo:

$$\frac{\partial u}{\partial t} - \sum_{i=1}^{d}\left(\sum_{j=1}^{d}\frac{D_{ij}}{a}\frac{\partial^2 u}{\partial x_i \partial x_j}\right) + \frac{1}{a}\overrightarrow{\widetilde{\mathbf{V}}}\nabla u + \frac{\widetilde{q}}{a}u = \frac{f}{a}$$

que es la forma más usual en la que se escribe la ecuación de transporte:

$$\frac{\partial u}{\partial t} - \sum_{i=1}^{d}\left(\sum_{j=1}^{d}\widehat{D}_{i,j}\frac{\partial^2 u}{\partial x_i \partial x_j}\right) + \vec{\widehat{\mathbf{V}}}\nabla u + \widehat{q}u = \widehat{f}$$

y que por aligerar la notación, en numerosas ocasiones, cuando ello no conduzca a confusión escribiremos prescindiendo del símbolo " $\widehat{}$ " en los coeficientes.

4.1.2. Obtención de fórmulas en diferencias finitas para la aproximación de derivadas de funciones.

A) Derivadas de funciones de una variable.

La forma más usual para obtener expresiones en diferencias finitas de los valores de las derivadas de funciones consiste en combinar de forma adecuada desarrollos en serie de Taylor de dichas funciones. Ello exige, obviamente, admitir cierta regularidad para las funciones con las que se opere en el sentido de exigir que la función pueda desarrollarse en serie de Taylor hasta el término del desarrollo que nos interese. En este sentido, comenzando por el caso de funciones de una sola variable, recordamos que si $u(x)$ es una función de clase $C^{m+1}(]0, L[)$ y h es un valor real, el valor de la función en el punto $(x^* + h)$ puede expresarse mediante un desarrollo en serie de Taylor en la forma:

$$u(x^* + h) = u(x^*) + h\frac{du}{dx}(x^*) + \frac{h^2}{2}\frac{d^2u}{dx^2}(x^*) + ...$$
$$+\frac{h^m}{m!}\frac{d^m u}{dx^m}(x^*) + \frac{h^{m+1}}{(m+1)!}\frac{d^{m+1}u}{dx^{m+1}}(\xi)$$

donde ξ es un punto comprendido entre x^* y x^*+h. En el desarrollo anterior, el último sumando se suele designar como "el resto" del desarrollo y, para valores de h suficientemente pequeños, puede obtenerse una buena aproximación del valor $u(x^* + h)$ mediante:

$$u(x^* + h) \approx u(x^*) + h\frac{du}{dx}(x^*) + \frac{h^2}{2}\frac{d^2u}{dx^2}(x^*) + ... + \frac{h^m}{m!}\frac{d^m u}{dx^m}(x^*)$$

cometiéndose un error dado por el resto del desarrollo y que, abreviadamente, escribiremos como un error $O(h^{m+1})$ (es decir, un error del orden de h^{m+1} o un error de orden $(m+1)$). Este error es debido al truncamiento del desarrollo en serie de Taylor en su sumando $(m+1)$-ésimo y por ello es habitual designarlo con el nombre de **error de truncamiento (o de truncatura)**.

La aproximación anterior nos permite ya obtener fórmulas que, a su vez, aproximen el valor de la primera derivada de una función. En efecto, si consideramos $m = 1$ y un valor $h > 0$ el desarrollo anterior se reescribe como:

$$u(x^* + h) \approx u(x^*) + h\frac{du}{dx}(x^*)$$

de donde

$$\frac{du}{dx}(x^*) \approx \frac{u(x^*+h)-u(x^*)}{h}$$

fórmula en la que se está cometiendo un error de truncamiento dado por:

$$E_{tr} = -\frac{h}{2}\frac{d^2u}{dx^2}(\xi) \rightarrow E_{tr} = O(h)$$

La expresión que acabamos de obtener se conoce como una **fórmula en diferencias finitas progresiva** de primer orden para aproximar la primera derivada de una función (o, más brevemente, fórmula descentrada en adelanto) pues, suponiendo $h > 0$, para aproximar el valor de la primera derivada en x^* utiliza el valor de la función en x^* y en un punto (x^*+h) "adelantado" respecto a x^*.

Obviamente, siendo $h > 0$, podría considerarse el desarrollo:

$$u(x^*-h) = u(x^*) - h\frac{du}{dx}(x^*) + \frac{h^2}{2}\frac{d^2u}{dx^2}(\xi)$$

del que, siguiendo el mismo proceso anterior, se obtendría la **fórmula en diferencias finitas regresiva** de primer orden para aproximar la primera derivada de una función (o, más brevemente, fórmula descentrada en retroceso):

$$\frac{du}{dx}(x^*) \approx \frac{u(x^*)-u(x^*-h)}{h}$$

en la que se comete un error de truncamiento:

$$E_{tr} = \frac{h}{2}\frac{d^2u}{dx^2}(\xi) \rightarrow E_{tr} = O(h).$$

Pero también podríamos haber combinado desarrollos en serie. Así, si consideramos los dos desarrollos:

$$u(x^*+h) = u(x^*) + h\frac{du}{dx}(x^*) + \frac{h^2}{2}\frac{d^2u}{dx^2}(x^*) + \frac{h^3}{6}\frac{d^3u}{dx^3}(\xi')$$

$$u(x^*-h) = u(x^*) - h\frac{du}{dx}(x^*) + \frac{h^2}{2}\frac{d^2u}{dx^2}(x^*) - \frac{h^3}{6}\frac{d^3u}{dx^3}(\xi'')$$

y al primero le restamos el segundo, se tendrá que:

$$u(x^*+h) - u(x^*-h) = 2h\frac{du}{dx}(x^*) + \frac{h^3}{6}\left(\frac{d^3u}{dx^3}(\xi') + \frac{d^3u}{dx^3}(\xi'')\right) \Rightarrow$$

$$\frac{du}{dx}(x^*) = \frac{u(x^*+h)-u(x^*-h)}{2h} - \frac{h^2}{6}\frac{d^3u}{dx^3}(\xi)$$

donde ξ es un punto intermedio entre $(x^* - h)$ y $(x^* + h)$. De esta expresión puede obtenerse entonces la **fórmula en diferencias finitas centrada**:

$$\frac{du}{dx}(x^*) \approx \frac{u(x^* + h) - u(x^* - h)}{2h}$$

en la que el error de truncatura estará dado por:

$$E_{tr} = -\frac{h^2}{6}\frac{d^3u}{dx^3}(\xi) \rightarrow E_{tr} = O(h^2)$$

NOTA:

En el proceso anterior se ha utilizado la igualdad:

$$\left(\frac{d^3u}{dx^3}(\xi') + \frac{d^3u}{dx^3}(\xi'')\right) = 2\frac{d^3u}{dx^3}(\xi)$$

Ello es posible gracias a la suposición de que u es una función de clase C^3. En efecto, en estos casos, que nos surgirán frecuentemente, será de aplicación el siguiente teorema cuya demostración (consecuencia inmediata del teorema del valor medio) puedes encontrar, por ejemplo, en Michavila y Conde[1]*:*

Theorem 3 *Siendo f una función continua en $[a, b]$, siendo $\{\xi_i\}_{i=1}^n$ un conjunto de puntos de $[a, b]$y siendo $\{\alpha\}_{i=1}^n$ un conjunto de números reales del mismo signo y no nulos existe un punto $\xi \in [a, b]$ tal que:*

$$\sum_{i=1}^{n} \alpha_i f(\xi_i) = \left(\sum_{i=1}^{n} \alpha_i\right) f(\xi)$$

NOTA:

Obsérvese que la fórmula centrada que aproxima la derivada primera presenta un error de truncatura de segundo orden en tanto que las descentradas presentaban un error de primer orden. Ello nos indica que reducciones en el valor de h conllevarán una disminución de error cometido con dichas fórmulas que será mayor al usar la fórmula centrada que las descentradas. Siendo esto así, conviene señalar ya que no todo es el error de truncatura y que las fórmulas centradas pueden tener un "peor comportamiento" (en un sentido que más adelante se matizará) que las fórmulas descentradas.

Al igual que se han obtenido las fórmulas en diferencias que aproximan el valor de la primera derivada en un punto, podríamos obtener expresiones que

[1] F. Michavila y C. Conde. (1987). Métodos de aproximación. Ed. Universidad Politécnica de Madrid.

aproximasen el valor de derivadas de órdenes superiores. Así por ejemplo si sumamos los desarrollos en serie:

$$u(x^* + h) = u(x^*) + h\frac{du}{dx}(x^*) + \frac{h^2}{2}\frac{d^2u}{dx^2}(x^*) + \frac{h^3}{6}\frac{d^3u}{dx^3}(x^*) + \frac{h^4}{24}\frac{d^4u}{dx^4}(\xi')$$

$$u(x^* - h) = u(x^*) - h\frac{du}{dx}(x^*) + \frac{h^2}{2}\frac{d^2u}{dx^2}(x^*) - \frac{h^3}{6}\frac{d^3u}{dx^3}(x^*) + \frac{h^4}{24}\frac{d^4u}{dx^4}(\xi")$$

tendremos que:

$$u(x^* + h) + u(x^* - h) = 2u(x^*) + h^2\frac{d^2u}{dx^2}(x^*) + \frac{h^4}{24}\left(\frac{d^4u}{dx^4}(\xi') + \frac{d^4u}{dx^4}(\xi")\right)$$

de donde se obtendría una fórmula en **diferencias finitas centrada** para aproximar la derivada segunda de la función u en el punto x^* :

$$\frac{d^2u}{dx^2}(x^*) \approx \frac{u(x^* - h) - 2u(x^*) + u(x^* + h)}{h^2}$$

en la que se comete un error de truncatura dado por:

$$E_{tr} = -\frac{h^2}{12}\frac{d^4u}{dx^4}(\xi) \rightarrow E_{tr} = O(h^2)$$

Hasta aquí sólo se han considerado desarrollos en serie de Taylor de los valores $u(x^* + h)$ y $u(x^* - h)$. Se podrían poner en juego desarrollos de los valores en otros puntos ($(x \pm 2h)$, $(x \pm 3h)$, ...) para obtener otras fórmulas (centradas o descentradas) aproximando los valores de las derivadas (primeras, segundas, terceras, etc...) de la función con diferentes soportes.

NOTA:

De hecho las expresiones anteriores pueden obtenerse de otras formas distintas. Una de ellas es enmarcarlas dentro de la teoría de la interpolación y obtenerlas como fórmulas de derivación numérica, esto es, sustituyendo la derivada de orden m de la función u en x^* por la derivada de orden m en dicho punto del polinomio interpolador $p(x)$ de la función u. Consúltese para mayores detalles, por ejemplo, Conde & Schiavi[2].

Ejercicios propuestos:

1^o) Obténgase la fórmula de diferencias finitas:

$$\frac{d^2u}{dx^2}(x^*) \approx$$

[2]C. Conde y E. Schiavi. (2000). Guiones de la asignatura de Elementos de Matemáticas. Universidad Rey Juan Carlos.

$$\frac{-2u(x^*-2h)+32u(x^*-h)-60u(x^*)+32u(x+h)-2u(x+2h)}{24h^2}$$

y demuéstrese que el error de truncatura de la misma es $O(h^4)$.

2^o) Obténgase una fórmula en diferencias finitas que permita aproximar $\frac{du}{dx}(x^*)$ haciendo intervenir los valores $u(x^*-2h)$, $u(x^*-h)$, $u(x^*)$, $u(x^*+h)$ y $u(x^*+2h)$ y que sea de orden $O(h^4)$.

3^o) Deducir la fórmula:

$$\begin{aligned}\frac{du}{dx}(x^*) &= \frac{1}{12h}[-25u(x^*)+48u(x^*+h)-36u(x^*+2h)+\\ &\quad +16u(x^*+3h)-3u(x^*+4h)]\end{aligned}$$

y encontrar la expresión de su error de truncatura.

Cabe señalar también que en los desarrollos utilizados en las expresiones anteriores se han considerado puntos que distaban de x^* múltiplos enteros de h. Ello en ocasiones no podrá ser así y deberán considerarse "soportes no equidistantes". La forma de actuar con este tipo de soportes no diferirá de la antes presentada aunque, obviamente, estaremos conducidos a fórmulas y expresiones del error diferentes. Por centrar las ideas, siendo γ_1 y γ_2 dos números estríctamente positivos menores o iguales que 1, podemos considerar los desarrollos:

$$u(x^*+\gamma_1 h)=u(x^*)+\gamma_1 h\frac{du}{dx}(x^*)+\frac{\gamma_1^2h^2}{2}\frac{d^2u}{dx^2}(x^*)+\frac{\gamma_1^3h^3}{6}\frac{d^3u}{dx^3}(\xi')$$

$$u(x^*-\gamma_2 h)=u(x^*)-\gamma_2 h\frac{du}{dx}(x^*)+\frac{\gamma_2^2h^2}{2}\frac{d^2u}{dx^2}(x^*)-\frac{\gamma_2^3h^3}{6}\frac{d^3u}{dx^3}(\xi")$$

Restando el segundo desarrollo del primero obtendremos:

$$u(x^*+\gamma_1 h)-u(x^*-\gamma_2 h)=(\gamma_1+\gamma_2)h\frac{du}{dx}(x^*)+\frac{(\gamma_1^2-\gamma_2^2)h^2}{2}\frac{d^2u}{dx^2}(x^*)+$$

$$+\frac{(\gamma_1^3+\gamma_2^3)h^3}{6}\frac{d^3u}{dx^3}(\xi")$$

de donde

$$\frac{du}{dx}(x^*)\approx\frac{u(x^*+\gamma_1 h)-u(x^*-\gamma_2 h)}{(\gamma_1+\gamma_2)h}$$

cometiéndose un error de truncatura dado por:

$$E_{tr}=\frac{(\gamma_2^2-\gamma_1^2)h}{2(\gamma_1+\gamma_2)}\frac{d^2u}{dx^2}(x^*)-\frac{(\gamma_1^3+\gamma_2^3)h^2}{6(\gamma_1+\gamma_2)}\frac{d^3u}{dx^3}(\xi").$$

Se tiene así la correspondiente fórmula centrada para aproximar la primera derivada pero que ahora será de orden $O(h)$ salvo que $\gamma_1=\gamma_2$ en cuyo caso

será de orden $O(h^2)$. No obstante los mismos desarrollos podrían haberse combinado de forma que se anularan los términos en derivadas segundas. Para ello basta con restar a γ_2^2 veces el primer desarrollo γ_1^2 veces el segundo desarrollo, obteniendo en este caso:

$$\gamma_2^2 u(x^* + \gamma_1 h) - \gamma_1^2 u(x^* - \gamma_2 h) = (\gamma_2^2\gamma_1 + \gamma_1^2\gamma_2)h\frac{du}{dx}(x^*) +$$

$$+\frac{(\gamma_2^2\gamma_1^3 + \gamma_1^2\gamma_2^3)h^3}{6}\frac{d^3u}{dx^3}(\xi")$$

de donde se obtendría la fórmula en diferencias finitas:

$$\frac{du}{dx}(x^*) \approx \frac{\gamma_2^2 u(x^* + \gamma_1 h) - \gamma_1^2 u(x^* - \gamma_2 h)}{(\gamma_2^2\gamma_1 + \gamma_1^2\gamma_2)h}$$

con un error de truncatura:

$$E_{tr} = -\frac{(\gamma_2^2\gamma_1^3 + \gamma_1^2\gamma_2^3)h^2}{6(\gamma_2^2\gamma_1 + \gamma_1^2\gamma_2)}\frac{d^3u}{dx^3}(\xi")$$

que ya es de orden $O(h^2)$.

De forma similar podría procederse para la aproximación de derivadas de mayor orden. Así, sumando γ_2 veces el primer desarrollo a γ_1 veces el segundo desarrollo se obtendría:

$$\gamma_2 u(x^* + \gamma_1 h) + \gamma_1 u(x^* - \gamma_2 h) = (\gamma_1 + \gamma_2)u(x^*) + \frac{(\gamma_2\gamma_1^2 + \gamma_1\gamma_2^2)h^2}{2}\frac{d^2u}{dx^2}(x^*) +$$

$$+\frac{(\gamma_2\gamma_1^3 - \gamma_1\gamma_2^3)h^3}{6}\frac{d^3u}{dx^3}(\xi)$$

de donde se obtiene la fórmula:

$$\frac{d^2u}{dx^2}(x^*) \approx 2\frac{\gamma_2 u(x^* + \gamma_1 h) - (\gamma_1 + \gamma_2)u(x^*) + \gamma_1 u(x^* - \gamma_2 h)}{(\gamma_2\gamma_1^2 + \gamma_1\gamma_2^2)h^2}$$

con la que se comete un error de truncatura:

$$E_{tr} = \frac{(\gamma_2\gamma_1^3 - \gamma_1\gamma_2^3)h}{3(\gamma_2\gamma_1^2 + \gamma_1\gamma_2^2)}\frac{d^3u}{dx^3}(\xi) \rightarrow E_{tr} = O(h)$$

B) Derivadas de funciones de varias variables.

La forma de proceder anterior puede extenderse al caso de funciones de varias variables. Por simplicidad nosotros la expondremos sobre funciones de 2 variables y dejamos al lector la sencilla tarea de aplicar estos procedimientos a

funciones de 3 o más variables. Para ello, de forma similar a cómo procedíamos anteriormente, suponiendo la suficiente regularidad a nuestras funciones, y siendo $h > 0$ y $k > 0$ dos números positivos consideraremos los desarrollos en serie:

$$u(x^* + h, y^*) = u(x^*, y^*) + h\frac{\partial u}{\partial x}(x^*, y^*) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x^*, y^*) + \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x^*, y^*) +$$

$$ch^4 24\frac{\partial^4 u}{\partial x^4}(x^*, y^*) + ... \tag{4.1}$$

$$u(x^* - h, y^*) = u(x^*, y^*) - h\frac{\partial u}{\partial x}(x^*, y^*) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x^*, y^*) - \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x^*, y^*) +$$

$$+\frac{h^4}{24}\frac{\partial^4 u}{\partial x^4}(x^*, y^*) + ... \tag{4.2}$$

$$u(x^*, y^* + k) = u(x^*, y^*) + k\frac{\partial u}{\partial y}(x^*, y^*) + \frac{k^2}{2}\frac{\partial^2 u}{\partial y^2}(x^*, y^*) + \frac{k^3}{6}\frac{\partial^3 u}{\partial y^3}(x^*, y^*) +$$

$$+\frac{k^4}{24}\frac{\partial^4 u}{\partial y^4}(x^*, y^*) + ... \tag{4.3}$$

$$u(x^*, y^* - k) = u(x^*, y^*) - k\frac{\partial u}{\partial y}(x^*, y^*) + \frac{k^2}{2}\frac{\partial^2 u}{\partial y^2}(x^*, y^*) - \frac{k^3}{6}\frac{\partial^3 u}{\partial y^3}(x^*, y^*) +$$

$$+\frac{k^4}{24}\frac{\partial^4 u}{\partial y^4}(x^*, y^*) + ... \tag{4.4}$$

$$u(x^* + h, y^* + k) = u(x^*, y^*) + h\frac{\partial u}{\partial x}(x^*, y^*) + k\frac{\partial u}{\partial y}(x^*, y^*) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x^*, y^*) +$$

$$+hk\frac{\partial^2 u}{\partial x \partial y}(x^*, y^*) + \frac{k^2}{2}\frac{\partial^2 u}{\partial y^2}(x^*, y^*) + \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x^*, y^*) + \frac{h^2 k}{2}\frac{\partial^3 u}{\partial x^2 \partial y}(x^*, y^*) +$$

$$+\frac{hk^2}{2}\frac{\partial^3 u}{\partial x \partial y^2}(x^*, y^*) + \frac{k^3}{6}\frac{\partial^3 u}{\partial y^3}(x^*, y^*) + ... \tag{4.5}$$

$$u(x^* + h, y^* - k) = u(x^*, y^*) + h\frac{\partial u}{\partial x}(x^*, y^*) - k\frac{\partial u}{\partial y}(x^*, y^*) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x^*, y^*) +$$

$$-hk\frac{\partial^2 u}{\partial x \partial y}(x^*,y^*)+\frac{k^2}{2}\frac{\partial^2 u}{\partial y^2}(x^*,y^*)+\frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x^*,y^*)-\frac{h^2 k}{2}\frac{\partial^3 u}{\partial x^2 \partial y}(x^*,y^*)+$$

$$+\frac{hk^2}{2}\frac{\partial^3 u}{\partial x \partial y^2}(x^*,y^*)-\frac{k^3}{6}\frac{\partial^3 u}{\partial y^3}(x^*,y^*)+... \tag{4.6}$$

$$u(x^*-h,y^*+k)=u(x^*,y^*)-h\frac{\partial u}{\partial x}(x^*,y^*)+k\frac{\partial u}{\partial y}(x^*,y^*)+\frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x^*,y^*)+$$

$$-hk\frac{\partial^2 u}{\partial x \partial y}(x^*,y^*)+\frac{k^2}{2}\frac{\partial^2 u}{\partial y^2}(x^*,y^*)-\frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x^*,y^*)+\frac{h^2 k}{2}\frac{\partial^3 u}{\partial x^2 \partial y}(x^*,y^*)-$$

$$-\frac{hk^2}{2}\frac{\partial^3 u}{\partial x \partial y^2}(x^*,y^*)+\frac{k^3}{6}\frac{\partial^3 u}{\partial y^3}(x^*,y^*)+... \tag{4.7}$$

y

$$u(x^*-h,y^*-k)=u(x^*,y^*)-h\frac{\partial u}{\partial x}(x^*,y^*)-k\frac{\partial u}{\partial y}(x^*,y^*)+\frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x^*,y^*)+$$

$$+hk\frac{\partial^2 u}{\partial x \partial y}(x^*,y^*)+\frac{k^2}{2}\frac{\partial^2 u}{\partial y^2}(x^*,y^*)-\frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x^*,y^*)-\frac{h^2 k}{2}\frac{\partial^3 u}{\partial x^2 \partial y}(x^*,y^*)-$$

$$-\frac{hk^2}{2}\frac{\partial^3 u}{\partial x \partial y^2}(x^*,y^*)-\frac{k^3}{6}\frac{\partial^3 u}{\partial y^3}(x^*,y^*)+... \tag{4.8}$$

La combinación de los 8 desarrollos anteriores nos puede conducir a muy diferentes formas de aproximar las derivadas parciales (de primer y segundo orden) de la función $u(x,y)$ en el punto (x^*,y^*). En efecto, de la expresión (4.1) podría despejarse la primera derivada parcial de u respecto a x obteniéndose:

$$\frac{\partial u}{\partial x}(x^*,y^*)\approx\frac{u(x^*+h,y^*)-u(x^*,y^*)}{h} \tag{4.9}$$

fórmula en diferencias finitas progresiva que nos permite aproximar la derivada parcial respecto a x de la función u en (x^*,y^*) con un error de truncatura:

$$E_{tr}=-\frac{h}{2}\frac{\partial^2 u}{\partial x^2}(\xi,y^*)\rightarrow E_{tr}=O(h).$$

De (4.2) también podría despejarse la derivada parcial respecto a x obteniendo en este caso:

$$\frac{\partial u}{\partial x}(x^*, y^*) \approx \frac{u(x^*, y^*) - u(x^* - h, y^*)}{h} \tag{4.10}$$

fórmula en diferencias finitas regresiva que nos permite aproximar la derivada parcial respecto a x de la función u en (x^*, y^*) con un error de truncatura:

$$E_{tr} = \frac{h}{2}\frac{\partial^2 u}{\partial x^2}(\xi, y^*) \to E_{tr} = O(h).$$

Restando (2) de (1) obtendríamos la **fórmula centrada**:

$$\frac{\partial u}{\partial x}(x^*, y^*) \approx \frac{u(x^* + h, y^*) - u(x^* - h, y^*)}{2h} \tag{4.11}$$

que tiene un error de truncatura dado por:

$$E_{tr} = -\frac{h^2}{3}\frac{\partial^3 u}{\partial x^3}(\xi, y^*) \to E_{tr} = O(h^2).$$

Haciendo las mismas operaciones con los desarrollos (4.3) y (4.4) se obtienen para aproximar la primera derivada parcial respecto a y las fórmulas y errores:

progresiva:

$$\frac{\partial u}{\partial y}(x^*, y^*) \approx \frac{u(x^*, y^* + k) - u(x^*, y^*)}{k} \tag{4.12}$$

$$E_{tr} = -\frac{k}{2}\frac{\partial^2 u}{\partial y^2}(x^*, \zeta) \to E_{tr} = O(k)$$

regresiva:

$$\frac{\partial u}{\partial y}(x^*, y^*) \approx \frac{u(x^*, y^*) - u(x^*, y^* - k)}{k} \tag{4.13}$$

$$E_{tr} = \frac{k}{2} \cdot \frac{\partial^2 u}{\partial y^2}(x^*, \zeta) \to E_{tr} = O(k)$$

centrada:

$$\frac{\partial u}{\partial y}(x^*, y^*) \approx \frac{u(x^*, y^* + k) - u(x^*, y^* - k)}{2k} \tag{4.14}$$

$$E_{tr} = -\frac{k^2}{3}\frac{\partial^3 u}{\partial y^3}(x^*, \zeta) \to E_{tr} = O(k^2).$$

Para aproximar la segunda derivada parcial de u respecto a x en el punto (x^*, y^*) podría, por ejemplo, sumarse (4.1) y (4.2) lo que nos conduciría a la fórmula **centrada**:

$$\frac{\partial^2 u}{\partial x^2}(x^*, y^*) \approx \frac{u(x^* - h, y^*) - 2u(x^*, y^*) + u(x^* + h, y^*)}{h^2} \tag{4.15}$$

con la que se comete un error de truncatura:

$$E_{tr} = -\frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi, y^*) \rightarrow E_{tr} = O(h^2).$$

Y sumando (4.3) y (4.4) se obtendría:

$$\frac{\partial^2 u}{\partial y^2}(x^*, y^*) \approx \frac{u(x^*, y^* - h) - 2u(x^*, y^*) + u(x^*, y^* + k)}{k^2} \tag{4.16}$$

con la que se comete un error de truncatura:

$$E_{tr} = -\frac{k^2}{12}\frac{\partial^4 u}{\partial y^4}(x^*, \zeta) \rightarrow E_{tr} = O(k^2).$$

Y si se deseara aproximar la segunda derivada cruzada podrían, por ejemplo, sumarse (4.5) y (4.8) y a su resultado restarles (4.6) y (4.7). Ello nos conduciría a:

$$\frac{\partial^2 u}{\partial x \partial y}(x^*, y^*) \approx \frac{u(x^* + h, y^* + k) - u(x^* - h, y^* + k)}{4hk}$$

$$+\frac{-u(x^* + h, y^* - k) + u(x^* - h, y^* - k)}{4hk}$$

con un error (se deja como ejercicio propuesto al lector el determinarlo) del orden:

$$E_{tr} = O(\frac{h^3}{k}, h^2, hk, k^2, \frac{k^3}{h})$$

Obsérvese que en esta última expresión del orden del error intervienen distintos términos y, entre ellos, aparecen las fracciones h^3/k y k^3/h. Ello nos indica que reduciones del valor h que no vayan acompañadas de la correspondiente reducción del valor k (o viceversa) pueden empeorar la aproximación ralizada (pues k^3/h crecerá).

Obviamente muchas otras fórmulas en diferencias finitas podrían construirse considerando puntos no equidistantes, otros puntos o combinando los desarrollos en serie de Taylor de formas diferentes a como se ha hecho más arriba. Algunas de estas fórmulas las iremos presentando a lo largo del tema. Otras podrán encontrarse en la bibliografía que sobre el tema se cita. En todo caso, la aproximación de los operadores diferenciales que intervienen en las EDP podrá hacerse ya combinando estas u otras fórmulas en diferencias finitas. Pero no todas las fórmas posibles de combinar distintas fórmulas nos conducirán a esquemas de cálculo que tengan un buen comportamiento. Por ello, en los apartados siguientes, mostraremos la forma de proceder para analizar cómo se comportan distintos algoritmos numéricos que se obtienen de la combinación de fórmulas en diferencias finitas.

Ejercicios propuestos:

1^o) Obtén fórmulas centradas y descentradas para aproximar las siguientes derivadas:

$$\frac{\partial u}{\partial x}(x^*, y^*, z^*), \quad \frac{\partial u}{\partial y}(x^*, y^*, z^*), \quad \frac{\partial u}{\partial z}(x^*, y^*, z^*)$$

basadas en los desarrollos de $u(x^* \pm \Delta x, y^*, z^*)$, $u(x^*, y^* \pm \Delta y, z^*)$ y $u(x^*, y^*, z^* \pm \Delta z)$ en torno al punto (x^*, y^*, z^*) Determina el error de trucamiento en cada una de ellas.

2^o) Con los desarrollos anteriores y los correspondientes a:

$$u(x^* \pm \Delta x, y^* \pm \Delta y, z^* \pm \Delta z)$$

deduce fórmulas para aproximar las derivadas segundas de u en (x^*, y^*, z^*), determinando el error de truncatura de cada una de ellas.

4.2. Aproximación mediante esquemas en diferencias de problemas de transporte estacionarios.

4.2.1. El problema de transporte estacionario en dominios unidimensionales.

A) Problemas con coeficientes constantes y esquemas con mallados equidistantes.

Por centrar ideas comencemos considerando el siguiente problema:

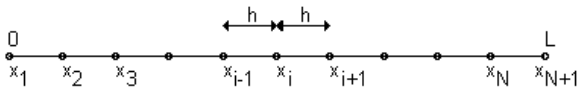
Siendo D, V y q tres constantes conocidas y tales que $D > 0$, y siendo conocida la función $f(x)$ y estando dados los valores u_{IZQ} y u_{DER}, encontrar una función $u(x)$ que verifique:

$$\begin{cases} -Du''(x) + Vu'(x) + qu(x) = f(x), & 0 < x < L \\ u(0) = u_{IZQ} \\ \\ u(L) = u_{DER} \end{cases}$$

Para la resolución mediante un esquema en diferencias finitas de este tipo de problemas comenzaremos introduciendo una **discretización espacial** del dominio $[0, L]$ en el que se plantea el problema. Ello consiste simplemente en seleccionar en el intervalo $[0, L]$ un conjunto de $(N + 1)$ puntos:

$$0 = x_1 < x_2 < ... < x_{i-1} < x_i < x_{i+1} < ... < x_N < x_{N+1} = L$$

en los que se calculará el valor (aproximado) de la solución del problema planteado. A estos puntos seleccionados los designaremos como **nodos** siendo x_i el i-ésimo nodo. Denotaremos por u_i al valor aproximado de $u(x_i)$ (que



se calculará mediante el esquema en diferencias finitas que se aplique). A estos valores u_i les denominaremos **valores nodales**. Al conjunto de nodos introducidos se le denominará en este caso **mallado**. Además, por simplicidad, consideraremos por el momento que las distancias entre dos nodos consecutivos x_i y x_{i+1} tiene siempre el valor h (cosa que expresaremos diciendo que el soporte está formado por nodos equidistantes o, más brevemente, diciendo que el mallado es equidistante o uniforme).

Los métodos numéricos de resolución del problema planteado no persiguen determinar la expresión analítica de la solución. Simplemente persiguen determinar los valores nodales de la solución, es decir, una aproximación de los valores que la solución toma en los nodos del mallado realizado. Obsérvese que por las condiciones de contorno consideradas ya se conocen los valores nodales u_1 y u_{N+1} siendo desconocidos $u_2, u_3, ..., u_N$. Para determinar estos valores nodales puede, en primer lugar, plantearse la EDP del problema en cada uno de los nodos del mallado en los que no se conoce el valor que toma la solución, lo que escribiremos como:

$$-Du"(x_i) + Vu'(x_i) + qu(x_i) = f(x_i) \qquad (i = 2, ..., N)$$

y tras ello aproximar las derivadas en cada punto mediante alguna fórmula en diferencias finitas. Así, si por ejemplo utilizamos fórmulas centradas para aproximar la primera y segunda derivada, obtendremos:

$$-D\frac{u(x_i - h) - 2u(x_i) + u(x_i + h)}{h^2} + O(h^2) + V\frac{u(x_i + h) - u(x_i - h)}{2h} + O(h^2) +$$

$$+qu(x_i) = f(x_i) \qquad (i = 2, ..., N)$$

de donde en cada nodo x_i puede plantearse la ecuación:

$$\left(-\frac{D}{h^2} - \frac{V}{2h}\right) u(x_{i-1}) + \left(\frac{2D}{h^2} + q\right) u(x_i) + \left(-\frac{D}{h^2} + \frac{V}{2h}\right) u(x_{i+1}) + O(h^2) = f(x_i)$$

y, prescindiendo del término debido a los errores de truncamiento:

$$\left(-\frac{D}{h^2}-\frac{V}{2h}\right)u_{i-1}+\left(\frac{2D}{h^2}+q\right)u_i+\left(-\frac{D}{h^2}+\frac{V}{2h}\right)u_{i+1}=f_i \quad (i=2,...,N) \tag{4.17}$$

que expresaremos brevemente como:

$$\alpha u_{i-1}+\beta u_i+\gamma u_{i+1}=f_i \qquad (i=2,...,N)$$

donde

$$\alpha=-\left(\frac{D}{h^2}+\frac{V}{2h}\right), \quad \beta=\left(\frac{2D}{h^2}+q\right), \quad \gamma=\left(\frac{V}{2h}-\frac{D}{h^2}\right)$$

Las ecuaciones (1) forman un sistema con $(N-1)$ ecuaciones y $(N+1)$ incógnitas. A estas ecuaciones se le pueden añadir las ecuaciones deducidas de las condiciones de contorno:

$$u_1=u_{IZQ}, \qquad u_{N+1}=u_{DER}$$

obteniéndose así un sistema de $(N+1)$ ecuaciones con el mismo número de incógnitas. Tal sistema lo representaremos de forma resumida como:

$$[\mathbf{A}]\{\mathbf{u}\}=\{\mathbf{b}\}$$

y una vez resuelto nos permitirá obtener una estimación de los valores nodales de la solución.

Con ello un algoritmo para calcular los valores aproximados de la solución del problema antes planteado consistirá en:

INICIO DEL ALGORITMO

Definir los valores de D, V, q, u_{IZQ}, u_{DER}, h *y* N, *y calcular los puntos:*

$$x_i \leftarrow (i-1)+h \qquad (i=1,2,...,N+1)$$

Definir la función $f(x)$ *y calcular los valores:*

$$f_i \leftarrow f(x_i) \qquad (i=2,...,N)$$

Calcular:

$$\alpha \leftarrow -\left(\frac{D}{h^2}+\frac{V}{2h}\right), \quad \beta \leftarrow \left(\frac{2D}{h^2}+q\right), \quad \gamma \leftarrow \left(\frac{V}{2h}-\frac{D}{h^2}\right)$$

$[\mathbf{A}] \leftarrow [\mathbf{0}]$

Para $i=2$ *hasta* $i=N$, *con paso* 1, *hacer:*

$$A(i,i-1) \leftarrow \alpha, \qquad A(i,i) \leftarrow \beta, \qquad A(i,i+1) \leftarrow \gamma, \quad b(i) = f_i$$

Fin bucle en i.

$A(1,1) \leftarrow 1, \quad A(N+1,N+1) \leftarrow 1, \quad b(1) \leftarrow u_{IZQ}, \quad b(N+1) \leftarrow u_{DER}$

Resolver el sistema $[\mathbf{A}] \cdot \{\mathbf{u}\} = \{\mathbf{b}\}$

Escribir los valores de las componentes u_i $(i = 1, ..., N+1)$ *del vector* $\{\mathbf{u}\}$.

FIN ALGORITMO.

Ejemplo:

Resolver mediante el esquema en diferencias finitas anterior el problema:

$$\begin{cases} -u''(x) + Vu'(x) & = 0, \quad 0 < x < 1 \\ u(0) = 0 \\ u(1) = 1 \end{cases}$$

para distintos valores de la velocidad V. Tómese como valor del parámetro de discretización $h = 0{,}1$.

En este caso se tendrán los nodos:

$$x_1 = 0, \quad x_2 = 0{,}1, \quad x_3 = 0{,}2, \quad x_4 = 0{,}3, \quad x_5 = 0{,}4,$$

$$x_6 = 0{,}5, \quad x_7 = 0{,}6, \quad x_8 = 0{,}7, \quad x_9 = 0{,}8, \quad x_{10} = 0{,}9, \quad x_{11} = 1{,}0$$

siendo $f_i = 0 \quad (i = 2, ..., 10)$. Puesto que $D = 1$, $q = 0$ y V queda libre para asignarle distintos valores, se tiene que:

$$\alpha = -100 - 5V, \quad \beta = 200, \quad \gamma = 5V - 100$$

con lo que el sistema de ecuaciones resultante se podrá escribir como:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \alpha & \beta & \gamma & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha & \beta & \gamma & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha & \beta & \gamma & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha & \beta & \gamma & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \alpha & \beta & \gamma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \alpha & \beta & \gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha & \beta & \gamma & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & \beta & \gamma & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & \beta & \gamma \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10} \\ u_{11} \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{Bmatrix}$$

sistema que para diferentes valores de V nos conduce a las siguientes soluciones:

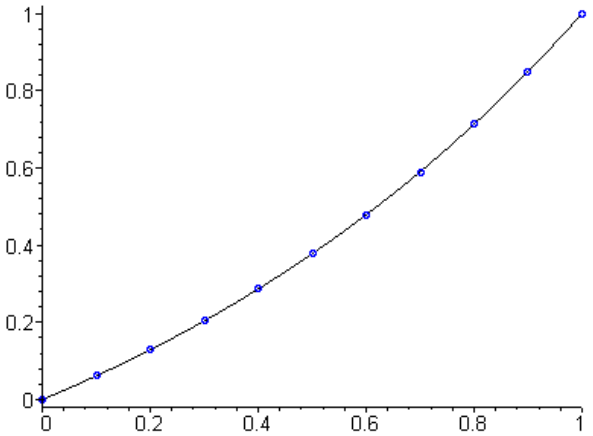


Figura 4.1: Resultados obtenidos para $V = 1$.

$\underline{\mathbf{V = 1}}$:

$u_1 = 0$, $u_2 = 0{,}06117989669$, $u_3 = 0{,}1287997825$, $u_4 = 0{,}2035375511$, $u_5 = 0{,}2861424532$, $u_6 = 0{,}3774426083$, $u_7 = 0{,}4783533060$, $u_8 = 0{,}5898861825$, $u_9 = 0{,}7131593618$, $u_{10} = 0{,}8494086650$, $u_{11} = 1$.

Esta solución puede compararse con la solución exacta dada por:

$$u(x) = \frac{e^x - 1}{e - 1}$$

observándose una buena coincidencia entre los valores exactos y los aproximados como se recoge en la gráfica siguiente en la que a trazo continuo se representa la solución y por puntos los valores nodales aproximados:

$\underline{\mathbf{V = 25}}$

$u_1 = 0$, $u_2 = -0{,}2867971998 \cdot 10^{-8}$, $u_3 = 0{,}2294377598 \cdot 10^{-7}$, $u_4 = -0{,}2093619558 \cdot 10^{-6}$, $u_5 = 0{,}1881389630 \cdot 10^{-5}$, $u_6 = -0{,}1693537464 \cdot 10^{-4}$, $u_7 = 0{,}1524155037 \cdot 10^{-3}$, $u_8 = -0{,}1371742401 \cdot 10^{-2}$, $u_9 = 0{,}1234567873 \cdot 10^{-1}$, $u_{10} = 0{,}1111111114$, $u_{11} = 1$.

Esta solución puede compararse con la solución exacta dada por:

$$u(x) = \frac{e^{25x} - 1}{e^{25} - 1}$$

observándose que los valores aproximados (representados en trazo grueso en la gráfica siguiente) y los exactos tienen un comportamiento muy diferente en el entorno de la capa límite:

Esta disparidad entre la aproximación obtenida y el valor exacto se acentúa conforme se va elevando el valor de la velocidad como se muestra en la gráfica obtenida para una velocidad $V = 100$.

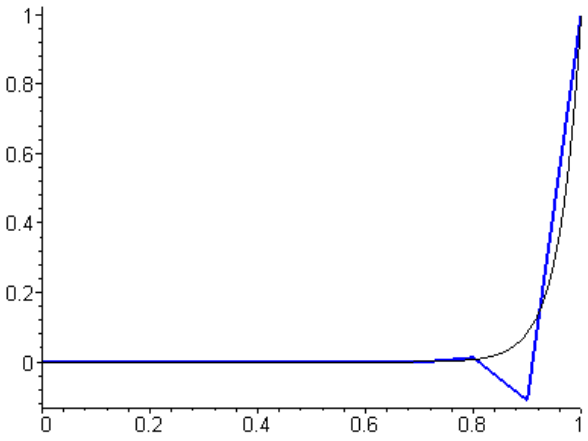


Figura 4.2: Resultados obtenidos para $V = 25$.

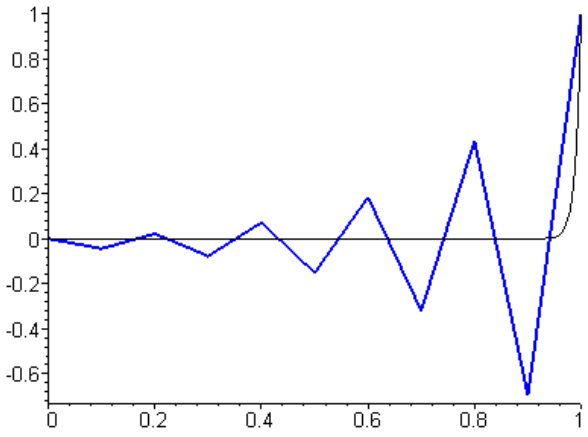


Figura 4.3: Resultados obtenidos para $V = 100$.

La justificación "física" de lo anterior puede buscarse en que el proceso de convección es un proceso "descentrado" mientras que la fórmula empleada para aproximar el término convectivo (la primera derivada $u'(x)$) era una fórmula centrada. Por ello cuanto mayor importancia cobra este término "peor" se vuelve la aproximación obtenida. Con independencia de que un poco más adelante justifiquemos este hecho de una forma matemática más rigurosa, modifiquemos el esquema obtenido usando una aproximación descentrada del término convectivo. Para ello llamemos:

$$\rho = \frac{1}{2} + \frac{|V|}{2V}$$

Obsérvese que si la velocidad es positiva $\rho = 1$ mientras que si la velocidad es negativa $\rho = 0$. Con ello aproximaremos la primera derivada $u'(x_i)$ mediante:

$$u'(x_i) \approx \rho \frac{u_i - u_{i-1}}{h} + (1-\rho)\frac{u_{i+1} - u_i}{h} =$$

$$\frac{-\rho u_{i-1} + (2\rho - 1)u_i + (1-\rho)u_{i+1}}{h}$$

es decir, utilizando el valor nodal u_i y el valor nodal "aguas arriba". Con ello el esquema en cada nodo puede escribirse ahora en la forma siguiente:

$$-D\frac{u_{i+1} - 2u_i + u_{i+1}}{h^2} +$$

$$V\frac{-\rho u_{i-1} + (2\rho - 1)u_i + (1-\rho)u_{i+1}}{h} + qu_i = f_i \qquad (4.18)$$

que también podrá expresarse en la forma:

$$\alpha u_{i-1} + \beta u_i + \gamma u_{i+1} = f_i \qquad (i = 2, ..., N)$$

donde

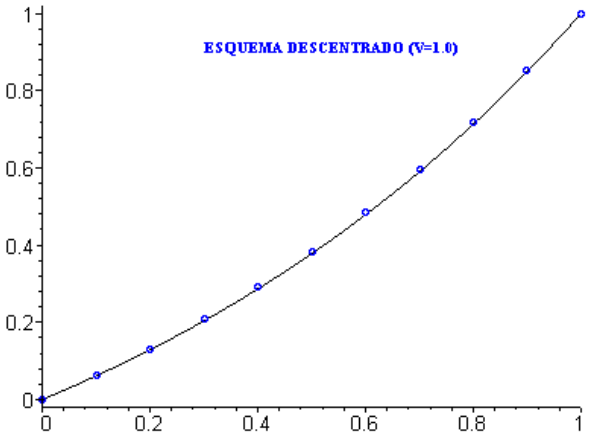
$$\alpha = -\left(\frac{D}{h^2} + \frac{\rho V}{h}\right), \quad \beta = \left(\frac{2D}{h^2} + \frac{2\rho - 1}{h}V + q\right), \quad \gamma = \left(\frac{(1-\rho)V}{h} - \frac{D}{h^2}\right)$$

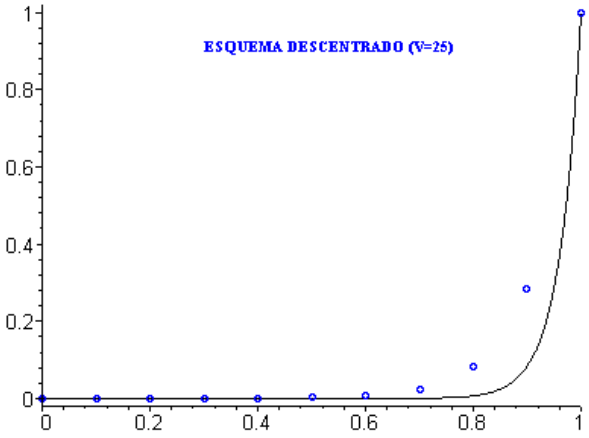
El mismo algoritmo anterior, pero sustituyendo en él las expresiones de α, β y γ por las que se acaban de obtener nos servirá para el uso de este nuevo esquema.

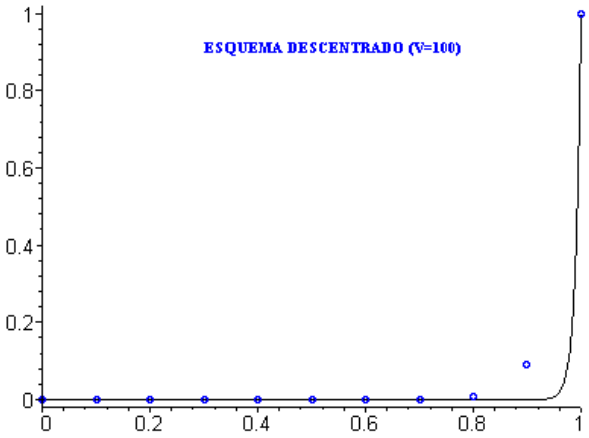
La aplicación de este esquema a los casos anteriormente resueltos nos conduce a las gráficas siguientes:

Como se ve la aproximación obtenida mejora ostensiblemente aunque persisten diferencias en las cercanías de la capa límite.

NOTA:


ESQUEMA DESCENTRADO (V=1.0)
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1


ESQUEMA DESCENTRADO (V=25)
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1


ESQUEMA DESCENTRADO (V=100)
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

Existen otras formas, además de la descrita, para "estabilizar" la solución aproximada de forma que no "oscile" en torno a la solución exacta aun para valores muy elevados de la velocidad.

¿De qué depende el que la solución aproximada se comporte en un caso de una manera y en otro caso de forma diferente?. ¿Hasta qué valores de V podría utilizarse el esquema centrado?. ¿Si el descentrado funciona mejor, qué interés puede tener el esquema centrado?. A éstas y otras cuestiones intentaremos responder a continuación.

B) Orden de los esquemas.

Una primera cuestión a plantearnos es la relativa al orden de los esquemas que acabamos de introducir. En ambos casos los esquemas los formulábamos, en cada nodo, como:

$$\alpha u_{i-1} + \beta u_i + \gamma u_{i+1} = f_i \qquad (i = 2, ..., N)$$

Es evidente que los valores nodales (aproximados) satisfacerán por tanto la relación:

$$\alpha u_{i-1} + \beta u_i + \gamma u_{i+1} - f_i = 0 \qquad (i = 2, ..., N)$$

Pero si en esta expresión sustituimos los valores nodales aproximados por los valores que en el nodo correspondiente tome la solución exacta, la igualdad anterior ya no tendrá por qué verificarse. En general se tendrá que:

$$E_i = \alpha u(x_{i-1}) + \beta u(x_i) + \gamma u(x_{i+1}) - f_i \neq 0 \qquad (i = 2, ..., N)$$

Definition 4 *Se denomina* ***error de consistencia local*** *en el nodo x_i al valor:*

$$E_i = \alpha u(x_{i-1}) + \beta u(x_i) + \gamma u(x_{i+1}) - f_i.$$

Según lo anterior, los errores de consistencia de un método nos indican en qué medida la solución analítica del problema a resolver satisface el esquema numérico con el que se la está aproximando.

Examinemos cómo son estos errores en el caso de los esquemas antes considerados. Para ello, con el primero de los presentados (el dado por la expresión (1)), suponiendo que $u(x)$ es suficientemente regular y considerando los desarrollos en serie de Taylor pertinentes, se tendrá en cada nodo que:

$$0 = -Du"(x_i) + Vu'(x_i) + qu(x_i) - f(x_i) =$$

$$= \alpha u(x_{i-1}) + \beta u(x_i) + \gamma u(x_{i+1}) - f_i + D\frac{h^2}{12}u^{(iv}(\xi_{1,i}) - V\frac{h^2}{6}u'''(\xi_{2,i})$$

de donde

$$E_i = \alpha u(x_{i-1}) + \beta u(x_i) + \gamma u(x_{i+1}) - f_i = -D\frac{h^2}{12}u^{(iv}(\xi_{1,i}) + V\frac{h^2}{6}u^{'''}(\xi_{2.i})$$

En otros términos el error de consistencia en cada nodo, utilizando el esquema (4.17), es de orden $O(h^2)$ (lo que expresaremos diciendo que es un **esquema de orden de consistencia local** 2).

Los mismos cálculos para el caso del esquema en el que la convección se aproximaba de forma descentrada (esquema (4.18)) nos conducen a que dicho esquema presenta un error de consistencia en cada nodo de orden $O(h)$. En efecto, en ese caso, y siempre bajo hipótesis de regularidad suficiente para la función $u(x)$ se tiene que:

$$0 = -Du"(x_i) + Vu'(x_i) + qu(x_i) - f(x_i) =$$

$$= \alpha u(x_{i-1}) + \beta u(x_i) + \gamma u(x_{i+1}) - f_i + D\frac{h^2}{12}u^{(iv}(\xi_{1,i}) +$$

$$+\rho V\frac{h}{6}u"(\xi_{2,i}) - (1-\rho)V\frac{h}{6}u"(\xi_{3,i})$$

de donde

$$E_i = \alpha u(x_{i-1}) + \beta u(x_i) + \gamma u(x_{i+1}) - f_i = -D\frac{h^2}{12}u^{(iv}(\xi_{1,i}) -$$

$$-\rho V\frac{h}{6}u"(\xi_{2.i}) + (1-\rho)V\frac{h}{6}u"(\xi_{3,i})$$

lo que confirma que este segundo esquema es un esquema de **orden de consistencia local** 1.

Lo anterior nos indica que una reducción del tamaño de h a la mitad se traduce en una reducción del error cometido a (del orden de) la cuarta parte en el caso del esquema (4.17) y a (del orden de) la mitad en el caso del esquema (4.18). En otros términos, si se trabaja con tamaños de paso suficientemente pequeños será preferible el primero de los esquemas.

Pero el problema es ¿qué valores son suficientemente pequeños para el paso de discretización espacial de h?. A ello vamos a contestar a continuación.

C) Verificación del principio del máximo discreto.

Si D y V no son nulos la solución analítica del problema

$$\left\{\begin{array}{ll} -Du"(x) + Vu'(x) & = 0, \quad 0 < x < L \\ u(0) = u_{IZQ} & \\ u(L) = u_{DER} & \end{array}\right\}$$

está dada por:

$$u(x) = u_{IZQ} + \frac{u_{DER} - u_{IZQ}}{e^{\frac{V}{D}L} - 1} \left(e^{\frac{V}{D}x} - 1 \right)$$

que fácilmente puedes comprobar que verifica el principio del máximo, esto es, la solución analítica, en cualquier punto x tal que $0 < x < L$ toma valores intermedios entre u_{IZQ} y u_{DER}.

Examinemos si los valores nodales que toma la solución aproximada mediante los esquemas anteriores también son intermedios entre u_{IZQ} y u_{DER}.

Para este problema (término fuente nulo) los esquemas se escribirán:

$$\alpha u_{i-1} + \beta u_i + \gamma u_{i+1} = 0 \qquad (i = 2, ..., N)$$

de donde

$$u_i = a u_{i-1} + b u_{i+1} \qquad (i = 2, ..., N)$$

siendo

$$a = \frac{-\alpha}{\beta}, \qquad b = \frac{-\gamma}{\beta}$$

Para el esquema (4.17) (en el que no se descentra la aproximación del término convectivo), al ser $q = 0$, se tendrá que:

$$\alpha = -\left(\frac{D}{h^2} + \frac{V}{2h} \right), \quad \beta = \left(\frac{2D}{h^2} \right), \quad \gamma = \left(\frac{V}{2h} - \frac{D}{h^2} \right)$$

por lo que en ese caso,

$$a = \frac{1}{2} + \frac{Vh}{2D}, \qquad b = \frac{1}{2} - \frac{Vh}{2D}$$

Obsérvese que $a + b = 1$. Pero nada asegura que los valores de a y b estén comprendidos entre 0 y 1. En resumidas cuentas no puede afirmarse que el valor u_i sea "una combinación convexa" de los valores de u_{i-1} y de u_{i+1} (lo cual garantizaría que es un valor intermedio). Para poder asegurar lo anterior debería obligarse a que:

$$0 \leq \frac{1}{2} + \frac{Vh}{2D} \leq 1$$

y

$$0 \leq \frac{1}{2} - \frac{Vh}{2D} \leq 1$$

es decir,

$$\frac{|V|\, h}{2D} \leq \frac{1}{2} \Rightarrow h \leq \frac{D}{|V|}$$

Para estos valores de h sí se puede afirmar que u_i es un valor intermedio entre u_{i-1} y u_{i+1}. Como esta conclusión es válida para todo nodo interior del mallado, según cómo sean los valores de u_{IZQ} y de u_{DER}, se verificará, bajo la hipótesis anterior sobre h, alguna de las dos situaciones siguientes:

$$u_{IZQ} = u_1 \leq u_2 \leq u_3 \leq ..... \leq u_{i-1} \leq u_i \leq u_{i+1} \leq ..... \leq u_{N+1} = u_{DER}$$

o

$$u_{IZQ} = u_1 \geq u_2 \geq u_3 \geq ..... \geq u_{i-1} \geq u_i \geq u_{i+1} \geq ..... \geq u_{N+1} = u_{DER}$$

por lo que, siempre bajo la hipótesis $h \leq D/|V|$, se satisfacerá el principio del máximo.

Para valores superiores de h el esquema viola dicho principio. Ello demuestra el siguiente:

Theorem 4 *El esquema (4.17) satisface el principio del máximo sólamente si se verifica que:*

$$h \leq \frac{D}{|V|}$$

NOTA:

El resultado del teorema anterior te permitirá saber por qué el esquema (4.17) *tenía un comportamiento "oscilante" sobre el ejemplo al que lo aplicamos anteriormente. En efecto, puesto que el paso utilizado fue* $h = 0{,}1$, *y* $D = 1$, *cuando la velocidad sobrepasaba el valor* $V = 10$, *el esquema no verificaba el principio del máximo. Aparecían valores negativos (inferiores a* u_{IZQ}*) y ello conllevaba la aparición de las oscilaciones.*

Examinemos ahora el segundo esquema dado por (4.18). Por simplicidad consideremos $V > 0$ con lo que $\rho = 1$ (si la velocidad fuese negativa puedes realizar los desarrollos análogos para comprobar la veracidad de la conclusión a la que lleguemos). En este caso:

$$\alpha = -\left(\frac{D}{h^2} + \frac{V}{h}\right), \quad \beta = \left(\frac{2D}{h^2} + \frac{V}{h}\right), \quad \gamma = \left(-\frac{D}{h^2}\right)$$

por lo que

$$a = \frac{D + hV}{2D + hV}, \qquad b = \frac{D}{2D + hV}$$

Obsérvese que en este caso se verifica que:

$$a + b = 1, \quad 0 < a < 1, \quad 0 < b < 1$$

es decir, que el valor

$$u_i = a u_{i-1} + b u_{i+1}$$

es una combinación convexa de u_{i-1} y u_{i+1} por lo que u_i será un valor intermedio entre u_{i-1} y u_{i+1}. En resumen se tiene así demostrado el siguiente:

Theorem 5 *El esquema (4.18) verifica el principio del máximo para todo valor del paso de discretización h.*

NOTA:

Observa que en el caso de velocidad positiva, cuanto mayor sea $|V|$ en comparación con D más próximo a 1 será el valor del coeficiente a y más próximo a 0 será el valor del coeficiente b. En otros términos, puesto que

$$u_i = au_{i-1} + bu_{i+1}$$

más se parecerá u_i a u_{i-1}. Ello explica el por qué en los ejemplos tratados, cuanto mayor era la velocidad más tendía hacia 0 la solución aproximada, salvo en un entorno del extremo derecho en el que se imponía como valor de la solución el valor 1. ¿Podrías realizar un razonamiento parecido en el caso de que $V < 0$?.

En resumidas cuentas, será preferible el uso del esquema 4.17 para valores de $h \leq D/|V|$ pues en ese caso se verifica el principio del máximo y el orden de consistencia local es 2. Pero si se necesita utilizar un paso de discretización mayor (lo que puede ser necesario en problemas fuertemente convectivos en los que $|V| >> D$ para no tener un número "excesivo" de nodos) será más conveniente el uso del esquema (4.18) para asegurar el cumplimiento del principio del máximo.

D) Tamaños de paso variables.

En ocasiones será conveniente no utilizar mallados equidistantes pues en el dominio de estudio pueden existir zonas en las que la solución sea "suave" y zonas en las que presente cambios bruscos. Por ejemplo, en el problema antes considerado, cuando la velocidad tomaba valores elevados, la solución era prácticamente nula en todo el dominio salvo en un entorno de la capa límite. Podría pensarse en ese caso en utilizar distancias "grandes" entre los primeros nodos y distancias "pequeñas" entre los últimos.

La forma de proceder en ese caso es análoga a la antes considerada, si bien, como vimos en el apartado 1.2 de este capítulo, las fórmulas a emplear variarán. Más concretamente consideremos nuevamente el problema:

Siendo D, V y q tres constantes conocidas y tales que $D > 0$, y siendo conocida la función $f(x)$ y estando dados los valores u_{IZQ} y u_{DER}, encontrar una función $u(x)$ que verifique:

$$\left\{ \begin{array}{ll} -Du"(x) + Vu'(x) + qu(x) & = f(x) \quad 0 < x < L \\ u(0) = u_{IZQ} & \\ u(L) = u_{DER} & \end{array} \right\}$$

y denotemos por:

$$0 = x_1 < x_2 < x_3 < ... < x_i < x_{i+1} < ... < x_N < x_{N+1} = L$$

a los nodos del mallado, designando ahora por:

$$h_i = x_{i+1} - x_i \quad (i = 1, 2, ..., N)$$

por

$$h = 1 \leq i \leq N \underset{1 \leq i \leq N}{\text{máx}} \{h_i\}$$

y por

$$\mu_i = \frac{h_i}{h} \quad (i = 1, 2, ..., N)$$

Aproximando en cada nodo todas las derivadas de forma centrada, se obtiene el siguiente desarrollo:

$$-Du"(x_i) + Vu'(x_i) + qu(x_i) = f(x_i) \Rightarrow$$

$$\Rightarrow -D2\frac{\mu_i u(x_i - \mu_{i-1}h) - (\mu_{i-1} + \mu_i)u(x_i) + \mu_{i-1}u(x_i + \mu_i h)}{(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2\mu_i)h^2} + O(h) +$$

$$+V\frac{u(x_i + \mu_i h) - u(x_i - \mu_{i-1}h)}{(\mu_{i-1} + \mu_i)h} + O(h) + qu(x_i) = f(x_i), \quad (i = 2, ..., N)$$

de donde en cada nodo x_i puede plantearse la ecuación:

$$\left(-\frac{2\mu_i D}{(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2\mu_i)h^2} - \frac{V}{(\mu_{i-1} + \mu_i)h}\right)u(x_{i-1}) +$$

$$\left(\frac{2(\mu_{i-1} + \mu_i)D}{(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2\mu_i)h^2} + q\right)u(x_i) +$$

$$+\left(-\frac{2\mu_{i-1}D}{(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2\mu_i)h^2} + \frac{V}{(\mu_{i-1} + \mu_i)h}\right)u(x_{i+1}) + O(h) = f(x_i)$$

y, prescindiendo del término debido a los errores de truncamiento:

$$\left(-\frac{2\mu_i D}{(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2\mu_i)h^2} - \frac{V}{(\mu_{i-1} + \mu_i)h}\right)u_{i-1} +$$

$$\left(\frac{2(\mu_{i-1} + \mu_i)D}{(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2\mu_i)h^2} + q\right)u_i +$$

$$+\left(-\frac{2\mu_{i-1}D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}+\frac{V}{(\mu_{i-1}+\mu_i)h}\right)u_{i+1}=f_i,\ i=2,...,N) \quad (4.19)$$

que expresaremos brevemente como:

$$\alpha_{i,i-1}u_{i-1}+\alpha_{i,i}u_i+\alpha_{i,i+1}u_{i+1}=f_i \qquad (i=2,...,N)$$

donde

$$\alpha_{i,i-1}=\left(-\frac{2\mu_iD}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}-\frac{V}{(\mu_{i-1}+\mu_i)h}\right)$$

$$\alpha_{i,i}=\left(\frac{2(\mu_{i-1}+\mu_i)D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}+q\right)$$

$$\alpha_{i,i+1}=\left(-\frac{2\mu_{i-1}D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}+\frac{V}{(\mu_{i-1}+\mu_i)h}\right)$$

En este caso el orden del error de consistencia local de esquema será $O(h)$. No obstante otras aproximaciones de las derivadas serían posibles.

Si optamos por utilizar una fórmula descentrada para la aproximación del término convectivo, siendo, al igual que antes,

$$\rho=\frac{1}{2}+\frac{|V|}{2V}$$

obtendremos,

$$-Du"(x_i)+Vu'(x_i)+qu(x_i)=f(x_i)\Rightarrow$$

$$\Rightarrow -D2\frac{\mu_iu(x_i-\mu_{i-1}h)-(\mu_{i-1}+\mu_i)u(x_i)+\mu_{i-1}u(x_i+\mu_ih)}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}+O(h)+$$

$$+\rho V\frac{u(x_i)-u(x_i-\mu_{i-1}h)}{\mu_{i-1}h}+O(h)+(1-\rho)V\frac{u(x_i+\mu_ih)-u(x_i)}{\mu_ih}+O(h)+$$

$$+qu(x_i)=f(x_i) \qquad (i=2,...,N)$$

de donde en cada nodo x_i puede plantearse la ecuación:

$$\left(-\frac{2\mu_iD}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}-\frac{\rho V}{\mu_{i-1}h}\right)u(x_{i-1})+$$

$$+\left(\frac{2(\mu_{i-1}+\mu_i)D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}+\frac{\rho V}{\mu_{i-1}h}+\frac{(\rho-1)V}{\mu_i h}+q\right)u(x_i)+$$

$$+\left(-\frac{2\mu_{i-1}D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}+\frac{(1-\rho)V}{\mu_i h}\right)u(x_{i+1})+O(h)=f(x_i)$$

y, prescindiendo del término debido a los errores de truncamiento,

$$\left(-\frac{2\mu_i D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}-\frac{\rho V}{\mu_{i-1}h}\right)u_{i-1}+$$

$$+\left(\frac{2(\mu_{i-1}+\mu_i)D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}+\frac{\rho V}{\mu_{i-1}h}+\frac{(\rho-1)V}{\mu_i h}+q\right)u_i+$$

$$+\left(-\frac{2\mu_{i-1}D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}+\frac{(1-\rho)V}{\mu_i h}\right)u_{i+1}=f_i,\ (i=2,...,N) \qquad (4.20)$$

que también expresaremos brevemente como:

$$\alpha_{i,i-1}u_{i-1}+\alpha_{i,i}u_i+\alpha_{i,i+1}u_{i+1}=f_i,\ (i=2,...,N)$$

donde

$$\alpha_{i,i-1}=\left(-\frac{2\mu_i D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}-\frac{\rho V}{\mu_{i-1}h}\right)$$

$$\alpha_{i,i}=\left(\frac{2(\mu_{i-1}+\mu_i)D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}+\frac{\rho V}{\mu_{i-1}h}+\frac{(\rho-1)V}{\mu_i h}+q\right)$$

$$\alpha_{i,i+1}=\left(-\frac{2\mu_{i-1}D}{(\mu_{i-1}\mu_i^2+\mu_{i-1}^2\mu_i)h^2}+\frac{(1-\rho)V}{\mu_i h}\right)$$

Este esquema en el que se descentra el término convectivo también presenta errores de consistencia local de orden 1.

Ejercicio propuesto:

Diseña un algoritmo sobre los esquemas de cálculo (4.19) *y* (4.20).

Ya hemos comentado anteriormente el orden de los errores de consistencia local de los esquemas (4.19) y (4.20) Examinemos si verifican el principio del máximo. Para ello consideraremos el problema:

$$\left\{\begin{array}{ll} -Du"(x)+Vu'(x) & =0 \quad 0<x<L \\ u(0)=u_{IZQ} & \\ u(L)=u_{DER} & \end{array}\right\}$$

y, procediendo de forma análoga a la que seguimos en el caso de paso constante, podemos escribir cada ecuación en cada nodo como:

$$u_i = a_i u_{i-1} + b_i u_{i+1}$$

donde

$$a_i = \frac{-\alpha_{i-1,i}}{\alpha_{i,i}}, \qquad b_i = \frac{-\alpha_{i,i+1}}{\alpha_{i,i}}$$

y podremos asegurar el cumplimiento del principio del máximo cuando $0 \leq a_i \leq 1$, $0 \leq b_i \leq 1$ y además $a_i + b_i = 1$ (pues en ese caso u_i será una combinación convexa de los valores u_{i-1} y u_{i+1}).

Para el esquema (4.19) se tiene que:

$$a_i = \frac{\mu_i}{(\mu_{i-1} + \mu_i)} + \frac{\mu_{i-1}\mu_i}{(\mu_{i-1} + \mu_i)} \frac{V}{D} h$$

$$b_i = \frac{\mu_{i-1}}{(\mu_{i-1} + \mu_i)} - \frac{\mu_{i-1}\mu_i}{(\mu_{i-1} + \mu_i)} \frac{V}{D} h$$

Nuevamente se puede concluir que la condición $a_i + b_i = 1$ se verifica; pero estos coeficientes no tienen por qué tomar valores comprendidos entre 0 y 1. Para que ello fuese así se debería verificar que:

$$h \leq 2 \leq i \leq N \inf_{2 \leq i \leq N} \left(\frac{1}{\mu_{i-1}} \left| \frac{D}{V} \right|, \frac{1}{\mu_i} \left| \frac{D}{V} \right| \right)$$

o lo que es lo mismo,

$$h_i \leq \left| \frac{D}{V} \right| \qquad (i = 1, 2, ..., N)$$

Para el esquema (4.20) si se supone $V > 0$ (es decir, $\rho = 1$) se tendrá que:

$$a_i = \frac{\dfrac{2\mu_i D}{(\mu_i + \mu_{i-1})} + V\mu_i h}{2D + V\mu_i h}$$

$$b_i = \frac{\dfrac{2\mu_{i-1} D}{(\mu_i + \mu_{i-1})}}{2D + V\mu_i h}$$

Puede observarse que ahora $0 < a_i < 1$, $0 < b_i < 1$, y que $a_i + b_i = 1$ para todos los valores de $i = 1, 2, ..., N$. Por tanto en este caso sí que se verificará el principio del máximo independientemente de cual sea el valor de h tomado.

NOTA:

También en este caso, si $V > 0$, cuanto mayor sea la relación V/D más próximo a 1 será el valor de a_i y más próximo a 0 el de b_i. En otros términos, u_i será más próximo a u_{i-1} cuanto mayor sea la relación V/D. En el caso de que $V < 0$ puedes comprobar que es b_i el coeficiente que más se acercará al valor 1 mientras que a_i tenderá hacia 0.

Ejercicio propuesto:

Deducir bajo qué condiciones se verifica el principio del máximo para el esquema (4.20) en el caso de que $V < 0$.

E) Consideración de coeficientes variables.

Consideremos ahora el caso de problemas de transporte estacionarios planteados en dominios unidimensionales pero con coeficientes no constantes. Más concretamente consideremos el problema:

Siendo $D(x)$, $\widetilde{V}(x)$ y $\widetilde{q}(x)$ tres funciones conocidas y tales que $D(x) > 0 \ \forall x$, y siendo conocida la función $f(x)$ y estando dados los valores u_{IZQ} y u_{DER}, encontrar una función $u(x)$ que verifique:

$$\left\{ \begin{array}{ll} -\dfrac{d}{dx}\left(D(x)u'(x)\right) + \dfrac{d}{dx}\left(\widetilde{V}(x)u(x)\right) + \widetilde{q}(x)u(x) & = f(x) \quad 0 < x < L \\ u(0) = u_{IZQ} & \\ u(L) = u_{DER} & \end{array} \right\}$$

El tratamiento de este tipo de problemas se realizará de forma muy semejante a los casos antes considerados pero "operando" previamente en la EDP que interviene en el problema. En efecto, dicha ecuación es quivalente a:

$$-D(x)u"(x) - D'(x)u'(x) + \widetilde{V}(x)u'(x) + \widetilde{V}'(x).u(x) + \widetilde{q}(x)u(x) = f(x) \Rightarrow$$

$$\Rightarrow -D(x).u"(x) + (\widetilde{V}(x) - D'(x))u'(x) + (\widetilde{q}(x) + \widetilde{V}'(x))u(x) = f(x) \Rightarrow$$

$$\Rightarrow -D(x)u"(x) + V(x)u'(x) + q(x)u(x) = f(x)$$

donde hemos denotado por $V(x) = \widetilde{V}(x) - D'(x)$ y por $q(x) = \widetilde{q}(x) + \widetilde{V}'(x)$.

Con ello la aplicación de los esquemas antes descritos puede realizarse de forma similar a los casos con coeficientes constantes pero particularizando el valor de los coeficientes $D(x)$, $V(x)$ y $q(x)$ en el nodo x_i en el que se esté discretizando la ecuación. Si, por simplicidad, consideramos un mallado equidistante, el tratamiento centrado del término convectivo nos conduciría en cada nodo interior del dominio a la ecuación:

$$\alpha_{i,i-1}u_{i-1} + \alpha_{i,i}u_i + \alpha_{i,i+1}u_{i+1} = f_i \qquad (i = 2, ..., N)$$

donde ahora

$$\alpha_{i,i-1} = -\left(\frac{D_i}{h^2} + \frac{V_i}{2h}\right), \quad \alpha_{i,i} = \left(\frac{2D_i}{h^2} + q_i\right), \quad \alpha_{i,i+1} = \left(\frac{V_i}{2h} - \frac{D_i}{h^2}\right)$$

habiéndose designado por $D_i = D(x_i)$, $V_i = V(x_i)$ y por $q_i = q(x_i)$.

Si el esquema tratase el término convectivo de forma centrada las ecuaciones nodales tendrían la misma estructura pero los coeficientes de ellas serían:

$$\alpha_{i,i-1} = -\left(\frac{D_i}{h^2} + \frac{\rho_i V_i}{h}\right), \quad \alpha_{i,i} = \left(\frac{2D_i}{h^2} + \frac{(2\rho_i - 1)V_i}{h} + q_i\right),$$

$$\alpha_{i,i+1} = \left(\frac{(1-\rho_i)V_i}{h} - \frac{D_i}{h^2}\right)$$

donde D_i, V_i y q_i tienen el mismo significado que anteriormente siendo

$$\rho_i = \frac{1}{2} + \frac{|V_i|}{2V_i}.$$

Si los pasos fuesen de tamaño variable las expresiones de los coeficientes $\alpha_{i,i-1}$, $\alpha_{i,i}$ y $\alpha_{i,i+1}$ estarían dadas, para el esquema en el que no se descentra la aproximación del término convectivo por:

$$\alpha_{i,i-1} = \left(-\frac{2D_i}{\mu_{i-1}(\mu_{i-1}\mu_i + \mu_{i-1}\mu_i)h^2} - \frac{V_i}{(\mu_{i-1} + \mu_i)h}\right)$$

$$\alpha_{i,i} = \left(\frac{2D_i}{\mu_{i-1}\mu_i h^2} + q_i\right)$$

$$\alpha_{i,i+1} = \left(-\frac{2D_i}{\mu_i(\mu_{i-1}\mu_i + \mu_{i-1}\mu_i)h^2} + \frac{V_i}{(\mu_{i-1} + \mu_i)h}\right)$$

y para el esquema en el que el término convectivo se trataba mediante una aproximación descentrada por:

$$\alpha_{i,i-1} = \left(-\frac{2D_i}{\mu_{i-1}(\mu_{i-1}\mu_i + \mu_{i-1}\mu_i)h^2} - \frac{\rho_i V_i}{\mu_{i-1}h}\right)$$

$$\alpha_{i,i} = \left(\frac{2D_i}{\mu_{i-1}\mu_i h^2} + \frac{\rho_i V_i}{\mu_{i-1}h} + \frac{(\rho_i - 1)V_i}{\mu_i h} + q_i\right)$$

$$\alpha_{i,i+1} = \left(-\frac{2D_i}{\mu_i(\mu_{i-1}\mu_i + \mu_{i-1}\mu_i)h^2} + \frac{(1-\rho_i)V_i}{\mu_i h}\right)$$

Ejercicio propuesto:

Analiza el error de consistencia local de estos esquemas y las condiciones en que satisfacen el principio del máximo.

F) Condiciones de contorno sobre el flujo.

Hasta ahora en los extremos del intervalo $(0, L)$ hemos considerado condiciones de contorno de tipo Dirichlet mediante las cuales se imponía el valor de la función en los nodos x_1 y x_{N+1}. Pero en numerosos problemas de la ingeniería encontrarás condiciones de contorno diferentes, del tipo:

$$\left\{ \begin{array}{ll} c_{1,1}u'(0) + c_{1,2}u(0) & = g_{IZQ} \\ c_{2,1}u'(L) + c_{2,2}u(L) & = g_{DER} \end{array} \right\}$$

donde $c_{1,1}, c_{1,2}, c_{2,1}$, $c_{2,2}$ g_{IZQ} y g_{DER} son constantes conocidas.

Si $c_{11} \neq 0$ (en otro caso la condición en el extremo izquierdo se vuelve condición de Dirichlet) y $c_{21} \neq 0$ (por el mismo motivo para el extremo derecho) la imposición de este tipo de condiciones de contorno lleva a plantearse el esquema de cálculo en los nodos x_1 y x_{N+1}. Pero para ello deberían considerarse dos nodos ficticios (el x_0 y el x_{N+2}) que no existen en nuestro mallado. No obstante, de las condiciones de contorno anteriores podrá inferirse un valor "ficticio" para estos dos nodos ficticios. Detallemos este proceso. Supongamos que siendo h_1 la distancia entre el nodo x_1 y el nodo x_2 se considera un nodo ficticio x_0 en la posición:

$$x_0 = x_1 - h_1$$

La primera de las condiciones de contorno antes escritas puede reescribirse como:

$$u'(x_1) = \frac{(g_{IZQ} - c_{1,2}u(x_1))}{c_{11}} = r_1 + s_1 u(x_1)$$

que aproximaremos mediante:

$$u_1^{'} = \frac{(g_{IZQ} - c_{1,2}u_1)}{c_{11}} = r_1 + s_1 u_1 \tag{4.21}$$

siendo $r_1 = g_{IZQ}/c_{1,1}$ y $s_1 = -c_{1,2}/c_{1,1}$. Puede entonces considerarse un polinomio de segundo grado $p(x)$ que en el nodo ficticio x_0 tomase el valor (ficticio) u_0, en el nodo x_1 el valor u_1 y en el nodo x_2 el valor u_2. Tal polinomio estaría dado por

$$p(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}u_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}u_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}u_2$$

y su primera derivada en x_1 estará dada por

$$p'(x_1) = \frac{1}{2h_1}(u_2 - u_0)$$

La idea, entonces, consiste en aproximar la derivada de la función u en x_1 por la derivada de este polinomio que la interpola. Ello nos conduce a

$$\frac{1}{2h_1}(u_2 - u_0) = r_1 + s_1 u_1 \Rightarrow$$

$$\Rightarrow u_0 = -2h_1 s_1 u_1 + u_2 - 2h_1 r_1$$

Una vez calculado el valor nodal ficticio en x_0 podemos plantear en x_1 el esquema de cálculo como sigue:

$$-D_1 u"(x_1) + V_1 u'(x_1) + q_1 u(x_1) = f_1 \Rightarrow$$

$$-D_1 \frac{u_0 - 2u_1 + u_2}{h_1^2} + V_1 \left(r_1 + s_1 u_1\right) + q_1 u_1 = f_1 \Rightarrow$$

$$\Rightarrow \left(\frac{2D_1}{h_1^2} + s_1 \left(V_1 + \frac{D_1}{h_1}\right) + q_1\right) u_1 + \left(\frac{-2D_1}{h_1^2}\right) u_2 = f_1 - r_1 \left(V_1 + \frac{2D_1}{h_1}\right)$$

o deshaciendo los cambios de notación,

$$\left(\frac{2D_1}{h_1^2} - \frac{c_{12}}{c_{11}} \left(V_1 + \frac{D_1}{h_1}\right) + q_1\right) u_1 + \left(\frac{-2D_1}{h_1^2}\right) u_2 = f_1 - \frac{g_{IZQ}}{c_{11}} \left(V_1 + \frac{2D_1}{h_1}\right)$$

ecuación que escribiremos en la forma,

$$\alpha_{1,1} u_1 + \alpha_{1,2} u_2 = b_1$$

con

$$\alpha_{1,1} = \frac{2D_1}{h_1^2} - \frac{c_{12}}{c_{11}} \left(V_1 + \frac{D_1}{h_1}\right) + q_1$$

$$\alpha_{1,2} = \frac{-2D_1}{h_1^2}$$

y

$$b_1 = f_1 - \frac{g_{IZQ}}{c_{11}} \left(V_1 + \frac{2D_1}{h_1}\right).$$

Esta ecuación (que se añadirá a las antes encontradas para los nodos interiores) será la primera ecuación del sistema lineal a resolver.

En el nodo x_{N+1} se seguirá un procedimiento análogo al que se acaba de describir y se deja como ejercicio propuesto al lector la tarea de detallarlo para obtener:

$$\left(\frac{-2D_{N+1}}{h_N^2}\right) u_N + \left(\frac{2D_{N+1}}{h_N^2} - \frac{c_{22}}{c_{21}} \left(V_{N+1} - \frac{D_{N+1}}{h_N}\right) + q_1\right) u_{N+1} =$$

$$= f_{N+1} - \frac{g_{DER}}{c_{21}} \left(V_{N+1} - \frac{2D_{N+1}}{h_N}\right)$$

4.2.2. El problema de transporte estacionario en dominios bidimensionales.

A) Caso de coeficientes constantes en dominios rectangulares.

Siendo $\Omega = (0, L_x) \times (0, L_y)$ un rectángulo abierto de $I\!R^2$, de frontera $\partial\Omega$ consideremos ahora el problema formulado por la EDP:

$$-D_1 \frac{\partial^2 u}{\partial x^2}(x,y) - D_2 \frac{\partial^2 u}{\partial y^2}(x,y) + V_1 \frac{\partial u}{\partial x}(x,y) + V_2 \frac{\partial u}{\partial y}(x,y) +$$

$$+qu(x,y) = f(x,y) \qquad (x,y) \in \Omega$$

con la condición de contorno:

$$u(x,y) = u_D(x,y) \qquad (x,y) \in \partial\Omega$$

La forma aplicar un esquema en diferencias finitas a este tipo de problemas consiste en introducir un mallado considerando para ello las abscisas:

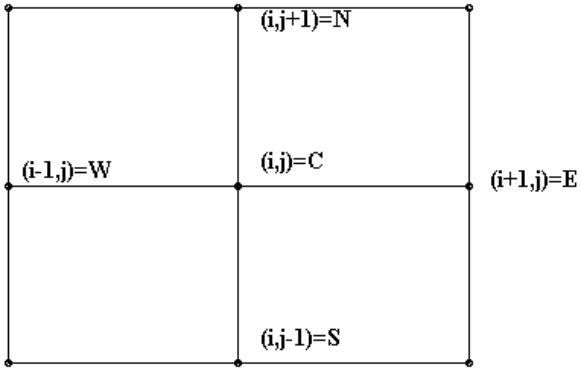
$$0 = x_1 < x_2 < .... < x_{i-1} < x_i < ..... < x_{NX} < x_{NX+1} = L_x$$

las ordenadas:

$$0 = y_1 < y_2 < .... < y_{i-1} < y_i < ..... < y_{NY} < y_{NY+1} = L_y$$

y con ello los nodos $n_{i,j} = (x_i, y_j) \quad (i = 1, 2, ..., NX + 1;\ j = 1, 2, .., NY + 1)$. Denotaremos por $NN = (NX + 1)(NY + 1)$ al número total de nodos y por $NF = 2(NX + NY + 1)$ al número de ellos que están ubicados en la frontera. Nótese que en estos últimos se conocerá el valor de la solución por las condiciones de contorno de tipo Dirichlet que hemos considerado. Por simplicidad, en un primer momento supongamos que la distancia entre dos abscisas cualesquiera consecutivas de las consideradas es siempre la misma h, y que la distancia entre dos ordenadas consecutivas cualesquiera de las consideradas también es siempre la misma k. Con ello en cada uno de los nodos interiores a Ω puede plantearse la EDP que interviene en la formulación del problema y aproximar las derivadas que en ella intervienen mediante alguna de las fórmulas (4.9)-(4.16), presentadas en la sección 1.2 de este capítulo. Por aligerar la notación denotemos por C (de Centro) al número del nodo $n_{i,j}$ en el que se plantee la ecuación, por N (de Norte) al nodo $n_{i,j+1}$ que está encima de $n_{i,j}$, por S (de Sur) al nodo $n_{i,j-1}$ del nodo que está debajo de $n_{i,j}$, por W (de Oeste) al nodo $n_{i-1,j}$ y por E (de Este) al nodo $n_{i+1,j}$.

Con ello, la obtención de un esquema centrado puede realizarse planteando la EDP en el nodo C:



$$-D_1\left(\frac{\partial^2 u}{\partial x^2}\right)_C - D_2\left(\frac{\partial^2 u}{\partial y^2}\right)_C + V_1\left(\frac{\partial u}{\partial x}\right)_C + V_2\left(\frac{\partial u}{\partial y}\right)_C + qu_C = f_C$$

y aproximándola mediante algunas de las fórmulas en diferencias. Por ejemplo utilizando fórmulas centradas:

$$\rightsquigarrow -D_1\frac{u_W - 2u_C + u_E}{h^2} - D_2\frac{u_S - 2u_C + u_N}{k^2} + V_1\frac{u_E - u_W}{2h} +$$

$$+V_2\frac{u_N - u_S}{2k} + qu_C = f_C \Rightarrow$$

$$\Rightarrow \left(\frac{-D_2}{k^2} - \frac{V_2}{2k}\right)u_S + \left(\frac{-D_1}{h^2} - \frac{V_1}{2h}\right)u_W + \left(\frac{2D_1}{h^2} + \frac{2D_2}{k^2} + q\right)u_C +$$

$$+\left(\frac{-D_1}{h^2} + \frac{V_1}{2h}\right)u_E + \left(\frac{-D_2}{k^2} + \frac{V_2}{2k}\right)u_N = f_C$$

ecuación que escribiremos en forma abreviada como:

$$a_S u_S + a_W u_W + a_C u_C + a_E u_E + a_N u_N = f_C$$

donde fácilmente pueden identificarse los coeficientes con los de la expresión anterior.

La ecuación anterior se podrá plantear para todos y cada uno de los nodos interiores al dominio Ω obteniéndose así $(NN - NF)$ ecuaciones. Sumando a estas las NF ecuaciones que dan valor a los NF nodos ubicados en la frontera se forma nuevamente un sistema de NN ecuaciones con NN incógnitas (los

valores nodales) cuya resolución nos permitirá obtener una aproximación de los valores de la solución en los nodos.

NOTA:

Al igual que antes, la aproximación realizada para el término convectivo presentará problemas cuando $|\overrightarrow{\mathbf{V}}|$ sea elevado frente al valor de $|D|$. Es por ello que también en este caso podrían contemplarse aproximaciones descentradas de los términos convectivos que ahora deberán realizarse según el sentido de cada componente de la velocidad. Te dejamos como ejercicio propuesto el desarrollo de un esquema descentrado (que podrás encontrar aplicado en un ejemplo posterior) y que conduce al esquema:

$$a_S u_S + a_W u_W + a_C u_C + a_E u_E + a_N u_N = f_C$$

donde ahora

$$a_S = \left(-\frac{D_2}{k^2} - \frac{\rho_2 V_2}{k}\right), \quad a_W = \left(-\frac{D_1}{h^2} - \frac{\rho_1 V_1}{h}\right)$$

$$a_C = \left(\frac{2D_1}{h} + \frac{2D_2}{k} + \frac{(2\rho_1 - 1)V_1}{h} + \frac{(2\rho_2 - 1)V_2}{k} + q\right)$$

$$a_E = \left(\frac{(1-\rho_1)V_1}{h} - \frac{2D_1}{h}\right), \quad a_N = \left(\frac{(1-\rho_2)V_2}{k} - \frac{2D_2}{k}\right)$$

siendo $\rho_1 = (1/2) + |V_1|/(2V_1)$ *y* $\rho_2 = (1/2) + |V_2|/(2V_2)$.

B) Caso particular: aproximación del operador laplaciano: Esquemas de 5 puntos y de 9 puntos.

Como caso particular del problema antes planteado, consideremos el problema:

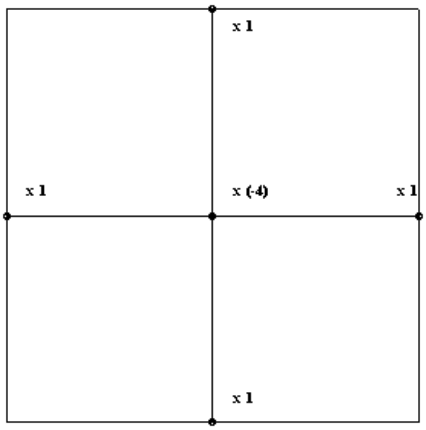
$$\left\{ \begin{array}{lll} \Delta u(x,y) & = f(x,y) & en\ \Omega \\ u(x,y) & = u_D(x,y) & en\ \partial\Omega \end{array} \right\}$$

La aproximación antes realizada, recuperando la notación $C = (i,j)$, $N = (i,j+1)$, $S = (i,j-1)$, $W = (i-1,j)$ y $E = (i+1,j)$, nos conduciría en este caso particular al esquema de cálculo:

$$\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{k^2} = f_{i,j}$$

que reescribiremos como

$$\frac{1}{k^2}u_{i,j-1} + \frac{1}{h^2}u_{i-1,j} - \left(\frac{2}{h^2} + \frac{2}{k^2}\right)u_{i,j} + \frac{1}{h^2}u_{i+1,j} + \frac{1}{k^2}u_{i,j+1} = f_{i,j}$$



El esquema anterior se conoce con el nombre de **esquema de 5 puntos** (en cruz) y es uno de los más frecuentemente utilizados para la aproximación de este tipo de operadores. En el caso de que además se verifique que $h = k$ el esquema se formula como:

$$\frac{u_{i,j-1} + u_{i-1,j} - 4u_{i,j} + u_{i+1,j} + u_{i,j+1}}{h^2} = f_{i,j}$$

y es frecuente representarlo gráficamente mediante

Otras formas de aproximar el operador laplaciano surgen de combinar los desarrollos en serie (4.1) a (4.8) de la primera sección de este capítulo. Así por ejemplo, continuando en el caso de que $h = k$, de sumar los desarrollos de $u(x+h, y+h)$, $u(x+h, y-h)$, $u(x-h, y+h)$ y de $u(x-h, y-h)$ se puede obtener el esquema de 5 puntos en cruz siguiente:

$$f_{i,j} = \Delta u(x_i, y_j) \approx \frac{u_{i-1,j-1} + u_{i+1,j-1} - 4u_{i,j} + u_{i-1,j+1} + u_{i+1,j+1}}{2h^2}$$

que tiene un error de consistencia local del orden de $O(h^2)$ y que se representa gráficamente por:

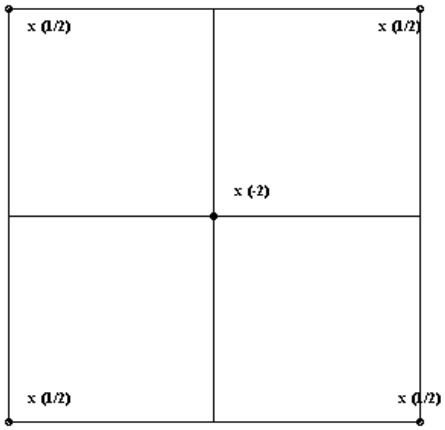
Ejercicio propuesto:

Obtener el esquema de nueve puntos:

$$\Delta u(x_i, y_j) \approx \frac{u_{i-1,j-1} + u_{i+1,j-1} + u_{i-1,j+1} + u_{i+1,j+1}}{6h^2} +$$

$$+ \frac{4\left(u_{i,j-1} + u_{i-1,j} + u_{i+1,j} + u_{i,j+1}\right) - 20u_{i,j}}{6h^2}$$

y demostrar que presenta un orden de consistencia local del orden $O(h^2)$.



C) Verificación del principio del máximo para el esquema de 5 puntos en cruz.

Según se detalla en el anexo, la ecuación:

$$\left\{ \begin{array}{lll} \Delta u(x,y) & = 0 & en\ \ \Omega \\ u(x,y) & = u_D(x,y) & en\ \ \partial\Omega \end{array} \right\}$$

planteada sobre un dominio abierto Ω que sea conexo y acotado, satisface el principio del máximo, esto es, los valores de la solución $u(x,y)$ en los puntos del abierto Ω están comprendidos siempre entre el valor mínimo y el valor máximo de $u(x,y)$ en la frontera $\partial\Omega$. Sería deseable que los esquemas numéricos que se utilizaran para resolver ecuaciones de este tipo también gozasen de la misma propiedad. Ello hará que, al aplicar el método numérico a problemas en los que la solución analítica $u(x,y)$ tome valores comprendidos entre $u_{\text{ínf}}$ y $u_{máx}$, las soluciones aproximadas que se obtengan también estén entre estas cotas.

Pero no todos los esquemas numéricos que pueden construirse utilizando fórmulas en diferencias finitas para aproximar el operador laplaciano verificarán el principio del máximo discreto. Por ello, a continuación, detallamos cómo se puede analizar si se cumple o no este principio con uno de los esquemas antes presentados: el esquema de cinco puntos en cruz. Puesto que hasta ahora sólo hemos tratado el caso de dominios rectangulares supondremos que Ω es un dominio rectangular en el que se introduce un mallado de nodos (x_i, y_j) $(i = 1, ..., NX+1;\ j = 1, ..., NY+1)$ y denotaremos a la distancia entre dos abscisas consecutivas (que supondremos siempre la misma) por h y a la distancia entre dos ordenadas conscutivas (que también supondremos constante) por k.

Con ello el esquema de cinco puntos en cruz consiste en plantear sobre cada nodo interior a Ω (en los de la frontera la solución se conoce), la ecuación en

diferencias:

$$\frac{u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{h^2}+\frac{u_{i,j-1}-2u_{i,j}+u_{i,j+1}}{k^2}=0$$

que reescribiremos como,

$$\frac{1}{k^2}u_{i,j-1}+\frac{1}{h^2}u_{i-1,j}-\left(\frac{2}{h^2}+\frac{2}{k^2}\right)u_{i,j}+\frac{1}{h^2}u_{i+1,j}+\frac{1}{k^2}u_{i,j+1}=0\Rightarrow$$

$$u_{i,j}=\alpha u_{i,j-1}+\beta u_{i-1,j}+\beta u_{i+1,j}+\alpha u_{i,j+1}$$

donde

$$\alpha=\frac{1}{\left(\frac{2}{h^2}+\frac{2}{k^2}\right)k^2}=\frac{h^2}{2(h^2+k^2)}$$

$$\beta=\frac{1}{\left(\frac{2}{h^2}+\frac{2}{k^2}\right)h^2}=\frac{k^2}{2(h^2+k^2)}$$

Puesto que $\alpha+\beta+\beta+\alpha=1$ y $0<\alpha<1$ y $0<\beta<1$ la expresión que nos proporciona el valor de $u_{i,j}$ nos conduce a que este es una combinación convexa de los valores de la solución en los nodos vecinos "Norte", "Sur", "Este" y "Oeste" del propio nodo (x_i,y_j). Por tanto $u_{i,j}$ será un valor intermedio entre los valores en esos cuatro nodos. Extendiendo esta forma de razonar a todos los nodos interiores del mallado se concluye que los valores en los nodos interiores serán valores intermedios entre el mayor y el menor de los valores de la solución en los nodos del mallado ubicados sobre la frontera. Y estos, a su vez, estarán comprendidos (salvo errores de redondeo) entre el mayor y el menor valor que tome la solución analítica $u(x,y)$ en los puntos de $\partial\Omega$. Por tanto, este esquema verifica el principio del máximo discreto.

Ejercicio propuesto:

Te dejamos como ejercicio propuesto analizar si los esquemas de 5 puntos en diagonal y de 9 puntos verifican el principio del máximo.

D) Mallados no uniformes.

Si la distancia entre abscisas o entre ordenadas no fuese constante la forma de proceder es análoga a la que antes considerábamos. Concretamente, si se considera un mallado en el que $h_i=x_{i+1}-x_i$ $(i=1,...,NX)$ y $k_j=y_{j+1}-y_j$ $(j=1,...,NY)$ denominaremos $h=\sup_{1\leq i\leq NX}\{h_i\}$ y $k=\sup_{1\leq j\leq NX}\{h_j\}$.

Con la ayuda de estos valores definimos además los números (comprendidos entre 0 y 1) :

$$\mu_i = \frac{h_i}{h} \quad (i = 1,\ ..., NX), \qquad \eta_j = \frac{k_j}{k} \quad (j = 1, ..., NY)$$

Con ello se tendrá que:

$$x_{i+1} = x_i + \mu_i h, \qquad x_{i-1} = x_i - \mu_{i-1} h$$

$$y_{j+1} = y_j + \eta_j k, \qquad y_{j-1} = y_{,j} - \eta_{j-1} k$$

con lo que podemos plantearnos nuevamente (bajo hipótesis de regularidad adecuadas) los correspondientes desarrollos en serie de Taylor de la función $u(x, y)$ en torno al punto (x_i, y_j). Con ello se tendrían, entre otros, los siguientes desarrollos:

$$u(x_{i+1}, y_j) = u(x_i + \mu_i h, y_j) = u(x_i, y_j) + \mu_i h \frac{\partial u}{\partial x}(x_i, y_j) + \frac{\mu_i^2 h^2}{2} \frac{\partial^2 u}{\partial x^2}(x_i, y_j) +$$

$$+ \frac{\mu_i^3 h^3}{6} \frac{\partial^3 u}{\partial x^3}(x_i, y_j) + ... \tag{4.22}$$

$$u(x_{i-1}, y_j) = u(x_i - \mu_{i-1} h, y_j) = u(x_i, y_j) - \mu_{i-1} h \frac{\partial u}{\partial x}(x_i, y_j) + \frac{\mu_{i-1}^2 h^2}{2} \frac{\partial^2 u}{\partial x^2}(x_i, y_j) +$$

$$- \frac{\mu_{i-1}^3 h^3}{6} \frac{\partial^3 u}{\partial x^3}(x_i, y_j) + ... \tag{4.23}$$

$$u(x_i, y_{j+1}) = u(x_i, y_j + \eta_j k) = u(x_i, y_j) + \eta_j k \frac{\partial u}{\partial y}(x_i, y_j) + \frac{\eta_j^2 h^2}{2} \frac{\partial^2 u}{\partial y^2}(x_i, y_j) +$$

$$+ \frac{\eta_j^3 h^3}{6} \frac{\partial^3 u}{\partial y^3}(x_i, y_j) + ... \tag{4.24}$$

$$u(x_i, y_{j-1}) = u(x_i, y_j - \eta_{j-1} k) = u(x_i, y_j) - \eta_{j-1} k \frac{\partial u}{\partial y}(x_i, y_j) + \frac{\eta_{j-1}^2 h^2}{2} \frac{\partial^2 u}{\partial y^2}(x_i, y_j) +$$

$$- \frac{\eta_{j-1}^3 h^3}{6} \frac{\partial^3 u}{\partial y^3}(x_i, y_j) + ... \tag{4.25}$$

Restando (7) de (6) se tiene que:

$$u(x_{i+1},y_j)-u(x_{i-1},y_j)=(\mu_i h+\mu_{i-1}h)\frac{\partial u}{\partial x}(x_i,y_j)+\frac{(\mu_i^2-\mu_{i-1}^2)h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_i,y_j)+$$

$$+\frac{(\mu_i^3+\mu_{i-1}^3)h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_i,y_j)+.....\Rightarrow$$

$$\Rightarrow \frac{\partial u}{\partial x}(x_i,y_j)=\frac{u(x_{i+1},y_j)-u(x_{i-1},y_j)}{h_{i-1}+h_i}-\frac{(\mu_i^2-\mu_{i-1}^2)h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_i,y_j)-$$

$$-\frac{(\mu_i^3+\mu_{i-1}^3)h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_i,y_j)-...$$

de donde se infiere la **fórmula en diferencias centrada** para aproximar $\frac{\partial u}{\partial x}(x_i,y_j)$ siguiente:

$$\frac{\partial u}{\partial x}(x_i,y_j)\approx\frac{u_{i+1,j}-u_{i-1,j}}{h_{i-1}+h_i} \tag{4.26}$$

en la que se comete un error de truncamiento dado por:

$$E_{i,j}=-\frac{(\mu_i^2-\mu_{i-1}^2)h}{2(\mu_i+\mu_{i-1})}\frac{\partial^2 u}{\partial x^2}(x_i,y_j)-...\rightarrow E_{i,j}\sim O(h). \tag{4.27}$$

Obsérvese que en el caso $h_{i-1}=h_i$ (o lo que es lo mismo $\mu_i=\mu_{i-1}$) el error de truncamiento pasa a ser de orden $O(h^2)$.

Nótese además que se podría obtener también una fórmula centrada de orden 2 si los desarrollos (4.22) y (4.23) se combinan de otra forma. En efecto si la ecuación (4.22) se multiplicase por μ_{i-1}^2 y a la expresión resultante se le restara (4.23) multiplicada por μ_i^2 se tendría que:

$$\mu_{i-1}^2u(x_{i+1},y_j)-\mu_i^2u(x_{i-1},y_j)=(\mu_{i-1}^2-\mu_i^2)u(x_i,y_j)+(\mu_{i-1}^2\mu_i+\mu_{i-1}\mu_i^2)h\frac{\partial u}{\partial x}(x_i,y_j)+$$

$$+\frac{(\mu_{i-1}^2\mu_i^3+\mu_{i-1}^3\mu_i^2)h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_i,y_j)+...$$

de donde se infiere la **fórmula en diferencias centrada** para aproximar $\frac{\partial u}{\partial x}(x_i,y_j)$ siguiente:

$$\frac{\partial u}{\partial x}(x_i,y_j)\approx\frac{\mu_{i-1}^2u_{i+1,j}-(\mu_{i-1}^2-\mu_i^2)u(x_i,y_j)-\mu_i^2u_{i-1,j}}{(\mu_{i-1}^2\mu_i+\mu_{i-1}\mu_i^2)h} \tag{4.28}$$

en la que se comete un error de truncamiento dado por:

$$E_{i,j} = -\frac{(\mu_{i-1}^2\mu_i^3 + \mu_{i-1}^3\mu_i^2)h^2}{6(\mu_{i-1}^2\mu_i + \mu_{i-1}\mu_i^2)}\frac{\partial^3 u}{\partial x^3}(x_i, y_j) - ... \rightarrow E_{i,j} \sim O(h^2) \tag{4.29}$$

Análogamente se podría haber despejado de (4.22) para obtener la **fórmula en diferencias finitas progresiva** que aproxime a $\frac{\partial u}{\partial x}(x_i, y_j)$ de la forma siguiente:

$$\frac{\partial u}{\partial x}(x_i, y_j) \approx \frac{u_{i+1,j} - u_{i,j}}{h_i} \tag{4.30}$$

con un error de truncamiento:

$$E_{i,j} = -\frac{\mu_i h}{2}\frac{\partial^2 u}{\partial x^2}(x_i, y_j) \rightarrow E_{i,j} \sim O(h) \tag{4.31}$$

También podría haberse despejado de (7) para obtener la **fórmula en diferencias finitas regresiva** que aproxime a $\frac{\partial u}{\partial x}(x_i, y_j)$ de la forma siguiente:

$$\frac{\partial u}{\partial x}(x_i, y_j) \approx \frac{u_{i,j} - u_{i-1,j}}{h_{i-1}} \tag{4.32}$$

con un error de truncamiento:

$$E_{i,j} = \frac{\mu_{i-1} h}{2}\frac{\partial^2 u}{\partial x^2}(x_i, y_j) \rightarrow E_{i,j} \sim O(h) \tag{4.33}$$

Procediendo de la misma manera con las expresiones (4.24) y (4.25) se aproxima $\frac{\partial u}{\partial y}(x_i, y_j)$ mediante una **fórmula en diferencias centrada:**

$$\frac{\partial u}{\partial y}(x_i, y_j) \approx \frac{u_{i,j+1} - u_{i,j-1}}{k_{j-1} + k_j} \tag{4.34}$$

con error de truncamiento:

$$E_{i,j} = -\frac{(\eta_j^2 - \eta_{j-1}^2)k}{2(\eta_j + \eta_{j-1})}\frac{\partial^2 u}{\partial x^2}(x_i, y_j) - ... \rightarrow E_{i,j} \sim O(k) \tag{4.35}$$

o la **fórmula en diferencias finitas progresiva:**

$$\frac{\partial u}{\partial y}(x_i, y_j) \approx \frac{u_{i,j+1} - u_{i,j}}{k_j} \tag{4.36}$$

con un error de truncamiento:

$$E_{i,j} = -\frac{\eta_j k}{2}\frac{\partial^2 u}{\partial x^2}(x_i, y_j) \rightarrow E_{i,j} \sim O(k) \tag{4.37}$$

o la **fórmula en diferencias finitas regresiva:**

$$\frac{\partial u}{\partial y}(x_i, y_j) \approx \frac{u_{i,j} - u_{i,j-1}}{k_{j-1}} \tag{4.38}$$

con un error de truncamiento:

$$E_{i,j} = -\frac{\eta_{j-1} k}{2} \frac{\partial^2 u}{\partial x^2}(x_i, y_j) \to E_{i,j} \sim O(k) \tag{4.39}$$

Ejercicio propuesto:

Se deja como ejercicio obtener, de forma análoga a como se hizo para la derivada parcial primera respecto a x, una fórmula centrada que aproxime la primera derivada con respecto a y con un error de truncamiento de orden 2.

De los desarrollos en serie considerados también podríamos obtener aproximaciones de derivadas segundas. Así, sumando μ_{i-1} veces (4.22) a μ_i veces (4.23) obtendríamos:

$$\mu_{i-1} u(x_{i+1}, y_j) + \mu_i u(x_{i-1}, y_j) = (\mu_{i-1} + \mu_i) u(x_i, y_j) +$$

$$\frac{(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2 \mu_i) h^2}{2} \frac{\partial^2 u}{\partial x^2}(x_i, y_j) + \frac{(\mu_{i-1}\mu_i^3 - \mu_{i-1}^3 \mu_i) h^3}{6} \frac{\partial^3 u}{\partial x^3}(x_i, y_j) + ... \Rightarrow$$

$$\Rightarrow \frac{\partial^2 u}{\partial x^2}(x_i, y_j) = 2. \frac{\mu_{i-1} u(x_{i+1}, y_j) - (\mu_{i-1} + \mu_i) u(x_i, y_j) + \mu_i u(x_{i-1}, y_j)}{(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2 \mu_i) h^2} +$$

$$+ \frac{(\mu_{i-1}^3 \mu_i - \mu_{i-1}\mu_i^3) h}{3(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2 \mu_i)} \frac{\partial^3 u}{\partial x^3}(x_i, y_j) + ...$$

de donde se infiere la **fórmula en diferencias finita centrada** para aproximar $\dfrac{\partial^2 u}{\partial x^2}(x_i, y_j)$ siguiente:

$$\frac{\partial^2 u}{\partial x^2}(x_i, y_j) \approx 2 \frac{\mu_{i-1} u_{i+1,j} - (\mu_{i-1} + \mu_i) u_{i,j} + \mu_i u_{i-1,j}}{(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2 \mu_i) h^2} \tag{4.40}$$

con la que se comete un error de truncamiento dado por:

$$E_{i,j} = \frac{(\mu_{i-1}^3 \mu_i - \mu_{i-1}\mu_i^3) h}{3(\mu_{i-1}\mu_i^2 + \mu_{i-1}^2 \mu_i)} \frac{\partial^3 u}{\partial x^3}(x_i, y_j) + ... \to E_{i,j} = O(h) \tag{4.41}$$

Obsérvese que en el caso de que $\mu_{i-1} = \mu_i$ el error de truncamiento pasa a ser de orden mayor $O(h^2)$.

Análogamente, sumando η_{j-1} veces (4.24) a η_j veces (4.25) obtendríamos la **fórmula en diferencias finita centrada** para aproximar $\dfrac{\partial^2 u}{\partial y^2}(x_i, y_j)$ siguiente:

$$\frac{\partial^2 u}{\partial y^2}(x_i, y_j) \approx 2\frac{\eta_{j-1}u_{i,j+1} - (\eta_{j-1} + \eta_j)u_{i,j} + \eta_j u_{i,j-1}}{(\eta_{j-1}\eta_j^2 + \eta_{j-1}^2\eta_j)k^2} \tag{4.42}$$

con la que se comete un error de truncamiento:

$$E_{i,j} = \frac{(\eta_{j-1}^3\eta_j - \eta_{j-1}\eta_{j-1}^3)k}{3(\eta_{j-1}\eta_{j-1}^2 + \eta_{j-1}^2\eta_j)}\frac{\partial^3 u}{\partial x^3}(x_i, y_j) + ... \to E_{i,j} = O(k) \tag{4.43}$$

que también pasaría a ser de segundo orden si $\eta_{j-1} = \eta_j$.

Otras aproximaciones de estas derivadas podrían obtenerse con adecuadas combinaciones de los desarrollos en serie:

$$u(x_{i+1}, y_{j+1}) = u(x_i + \mu_i h, y_j + \eta_j k) = u_{i,j} + \mu_i h\left(\frac{\partial u}{\partial x}\right)_{i,j} + \eta_j k\left(\frac{\partial u}{\partial y}\right)_{i,j} +$$

$$+\frac{\mu_i^2 h^2}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} + \mu_i\eta_j hk\left(\frac{\partial^2 u}{\partial x \partial y}\right)_{i,j} + \frac{\eta_j^2 k^2}{2}\left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j} +$$

$$+\frac{\mu_i^3 h^3}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_{i,j} + \frac{\mu_i^2\eta_j h^2 k}{2}\left(\frac{\partial^3 u}{\partial x^2 \partial y}\right)_{i,j} + \frac{\mu_i\eta_j^2 hk^2}{2}\left(\frac{\partial^3 u}{\partial x \partial y^2}\right)_{i,j} +$$

$$+\frac{\eta_j^3 k^3}{6}\left(\frac{\partial^3 u}{\partial y^3}\right)_{i,j} + ... \tag{4.44}$$

$$u(x_{i-1}, y_{j+1}) = u(x_i - \mu_{i-1}h, y_j + \eta_j k) = u_{i,j} - \mu_{i-1}h\left(\frac{\partial u}{\partial x}\right)_{i,j} + \eta_j k\left(\frac{\partial u}{\partial y}\right)_{i,j} +$$

$$+\frac{\mu_{i-1}^2 h^2}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} - \mu_{i-1}\eta_j hk\left(\frac{\partial^2 u}{\partial x \partial y}\right)_{i,j} + \frac{\eta_j^2 k^2}{2}\left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j} +$$

$$-\frac{\mu_{i-1}^3 h^3}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_{i,j} + \frac{\mu_{i-1}^2\eta_j h^2 k}{2}\left(\frac{\partial^3 u}{\partial x^2 \partial y}\right)_{i,j} - \frac{\mu_{i-1}\eta_j^2 hk^2}{2}\left(\frac{\partial^3 u}{\partial x \partial y^2}\right)_{i,j} +$$

$$+\frac{\eta_j^3 k^3}{6}\left(\frac{\partial^3 u}{\partial y^3}\right)_{i,j}+... \tag{4.45}$$

$$u(x_{i+1},y_{j-1})=u(x_i+\mu_i h,y_j-\eta_{j-1}k)=u_{i,j}+\mu_i h\left(\frac{\partial u}{\partial x}\right)_{i,j}-\eta_{j-1}k\left(\frac{\partial u}{\partial y}\right)_{i,j}+$$

$$+\frac{\mu_i^2 h^2}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j}-\mu_i\eta_{j-1}hk\left(\frac{\partial^2 u}{\partial x\partial y}\right)_{i,j}+\frac{\eta_{j-1}^2 k^2}{2}\left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j}+$$

$$+\frac{\mu_i^3 h^3}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_{i,j}-\frac{\mu_i^2\eta_{j-1}h^2k}{2}\left(\frac{\partial^3 u}{\partial x^2\partial y}\right)_{i,j}+\frac{\mu_i\eta_{j-1}^2hk^2}{2}\left(\frac{\partial^3 u}{\partial x\partial y^2}\right)_{i,j}+$$

$$-\frac{\eta_{j-1}^3 k^3}{6}\left(\frac{\partial^3 u}{\partial y^3}\right)_{i,j}+... \tag{4.46}$$

y

$$u(x_{i-1},y_{j-1})=u(x_i-\mu_{i-1}h,y_j-\eta_{j-1}k)=u_{i,j}-\mu_{i-1}h\left(\frac{\partial u}{\partial x}\right)_{i,j}-\eta_{j-1}k\left(\frac{\partial u}{\partial y}\right)_{i,j}+$$

$$+\frac{\mu_{i-1}^2 h^2}{2}\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j}+\mu_{i-1}\eta_{j-1}hk\left(\frac{\partial^2 u}{\partial x\partial y}\right)_{i,j}+\frac{\eta_{j-1}^2 k^2}{2}\left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j}+$$

$$-\frac{\mu_{i-1}^3 h^3}{6}\left(\frac{\partial^3 u}{\partial x^3}\right)_{i,j}-\frac{\mu_{i-1}^2\eta_{j-1}h^2k}{2}\left(\frac{\partial^3 u}{\partial x^2\partial y}\right)_{i,j}-\frac{\mu_{i-1}\eta_{j-1}^2hk^2}{2}\left(\frac{\partial^3 u}{\partial x\partial y^2}\right)_{i,j}+$$

$$-\frac{\eta_{j-1}^3 k^3}{6}\left(\frac{\partial^3 u}{\partial y^3}\right)_{i,j}+... \tag{4.47}$$

Ejercicio propuesto:

Dejamos como ejercicio propuesto encontrar en el caso de mallados no uniformes una aproximación de las derivadas primeras y segundas de la función $u(x,y)$ *partiendo de los desarrollos* (4.44)-(4.47) *que se acaban de escribir.*

Combinando estas fórmulas en diferencias pueden entonces aproximarse operadores diferenciales aplicados a nodos del mallado. Así por ejemplo, si se considera el problema modelo formulado por la EDP:

$$-D_1\frac{\partial^2 u}{\partial x^2}(x,y) - D_2\frac{\partial^2 u}{\partial y^2}(x,y) + V_1\frac{\partial u}{\partial x}(x,y) + V_2\frac{\partial u}{\partial y}(x,y) +$$

$$+qu(x,y) = f(x,y) \qquad (x,y) \in \Omega$$

con la condición de contorno:

$$u(x,y) = u_D(x,y) \qquad (x,y) \in \partial\Omega$$

sobre cada nodo (x_i, y_j) (que donotaremos por nodo C) del mallado, que sea interior a Ω, pueden utilizarse las expresiones (4.27), (4.40), (4.42) y la equivalente a esta última pero para la derivada respecto a y (que se dejo como ejercicio propuesto obtenerla) para construir una ecuación aprocimada en dicho nodo de la forma:

$$a_{C,S}u_S + a_{C,W}u_W + a_{C,C}u_C + a_{C,E}u_E + a_{C,N}u_N = f_C$$

donde

$$a_{C,S} = \left(\frac{-2\eta_j D_2}{\eta_{j-1}\eta_j(\eta_{j-1}+\eta_j)k^2} - \frac{\eta_j^2 V_2}{\eta_{j-1}\eta_j(\eta_{j-1}+\eta_j)k}\right)$$

$$a_{C,W} = \left(\frac{-2\mu_i D_1}{\mu_{i-1}\mu_i(\mu_{i-1}+\mu_i)h^2} - \frac{\mu_i^2 V_1}{\mu_{i-1}\mu_i(\mu_{i-1}+\mu_i)h}\right)$$

$$a_{C,C} = \left(\frac{2(\mu_{i-1}+\mu_i)D_1}{\mu_{i-1}\mu_i(\mu_{i-1}+\mu_i)h^2} + \frac{2(\eta_{j-1}+\eta_j)D_2}{\eta_{j-1}\eta_j(\eta_{j-1}+\eta_j)k^2}\right) +$$

$$+\left(\frac{(\mu_i^2-\mu_{i-1}^2)V_1}{\mu_{i-1}\mu_i(\mu_{i-1}+\mu_i)h} + \frac{(\eta_j^2-\eta_{j-1}^2)V_2}{\eta_{j-1}\eta_j(\eta_{j-1}+\eta_j)k} + q\right)$$

$$a_{C,E} = \left(\frac{-2\mu_{i-1}D_1}{\mu_{i-1}\mu_i(\mu_{i-1}+\mu_i)h^2} + \frac{\mu_{i-1}^2 V_1}{\mu_{i-1}\mu_i(\mu_{i-1}+\mu_i)h}\right)$$

$$a_{C,S} = \left(\frac{-2\eta_{j-1}D_2}{\eta_{j-1}\eta_j(\eta_{j-1}+\eta_j)k^2} - \frac{\eta_{j-1}^2 V_2}{\eta_{j-1}\eta_j(\eta_{j-1}+\eta_j)k}\right)$$

que es la expresión correspondiente al **esquema de 5 puntos en cruz con convección centrada y mallado no uniforme.**

Y si se quisiera descentrar las aproximaciones realizadas para los términos convectivos (usando para aproximar las derivadas primeras las expresiones (4.30), (4.32), (4.36) y (4.38) en lugar de las anteriores) la ecuación resultante tendría la misma forma pero ahora:

$$a_{C,S} = \left(\frac{-2\eta_j D_2}{\eta_{j-1}\eta_j(\eta_{j-1}+\eta_j)k^2} - \frac{\rho_2 V_2}{\eta_{j-1}k} \right)$$

$$a_{C,W} = \left(\frac{-2\mu_i D_1}{\mu_{i-1}\mu_i(\mu_{i-1}+\mu_i)h^2} - \frac{\rho_1 V_1}{\mu_{i-1}h} \right)$$

$$a_{C,C} = \left(\frac{2(\mu_{i-1}+\mu_i)D_1}{\mu_{i-1}\mu_i(\mu_{i-1}+\mu_i)h^2} + \frac{2(\eta_{j-1}+\eta_j)D_2}{\eta_{j-1}\eta_j(\eta_{j-1}+\eta_j)k^2} \right) +$$

$$+ \left(\frac{V_1}{h}\left(\frac{\rho_1}{\mu_{i-1}} + \frac{(\rho_1-1)}{\mu_i} \right) + \frac{V_2}{k}\left(\frac{\rho_2}{\eta_{j-1}} + \frac{(\rho_2-1)}{\eta_j} \right) + q \right)$$

$$a_{C,E} = \left(\frac{-2\mu_i D_1}{\mu_{i-1}\mu_i(\mu_{i-1}+\mu_i)h^2} + \frac{(1-\rho_1)V_1}{\mu_i h} \right)$$

$$a_{C,S} = \left(\frac{-2\eta_j D_2}{\eta_{j-1}\eta_j(\eta_{j-1}+\eta_j)k^2} + \frac{(1-\rho_2)V_2}{\eta_j k} \right)$$

que es la expresión correspondiente al **esquema de 5 puntos en cruz con convección descentrada y mallado no uniforme.**

Ejercicio propuesto:

Determínense las expresiones de los esquemas de 5 *puntos en diagonal y de* 9 *puntos, tanto con convección centrada como con convección descentrada para mallados no uniformes.*

E) Problemas con coeficientes no constantes.

Consideremos ahora el problema:

$$-\frac{\partial}{\partial x}\left(D_1(x,y)\frac{\partial u}{\partial x}(x,y) \right) - \frac{\partial}{\partial y}\left(D_2(x,y)\frac{\partial u}{\partial y}(x,y) \right) + \frac{\partial}{\partial x}\left(\widetilde{V}_1(x,y)u(x,y) \right) +$$

$$+\frac{\partial}{\partial y}\left(\widetilde{V}_2(x,y)u(x,y) \right) + \widetilde{q}(x,y)u(x,y) = f(x,y) \qquad (x,y) \in \Omega$$

con la condición de contorno:

$$u(x,y) = u_D(x,y) \qquad (x,y) \in \partial\Omega$$

De forma similar al proceso seguido en el caso unidimensional, será cómodo antes de discretizar la EDP anterior en cada nodo proceder a manipularla ligeramente como sigue:

$$-D_1(x,y)\frac{\partial^2 u}{\partial x^2}(x,y) - D_2(x,y)\frac{\partial^2 u}{\partial y^2}(x,y)+$$

$$+\left(\widetilde{V}_1(x,y) - \frac{\partial D_1}{\partial x}(x,y)\right)\frac{\partial u}{\partial x}(x,y) + \left(\widetilde{V}_2(x,y) - \frac{\partial D_2}{\partial y}(x,y)\right)\frac{\partial u}{\partial y}(x,y)+$$

$$+\left(\widetilde{q}(x,y) + \frac{\partial \widetilde{V}_1}{\partial x}(x,y) + \frac{\partial \widetilde{V}_2}{\partial y}(x,y)\right)u(x,y) = f(x,y) \qquad (x,y)\in\Omega$$

que escribiremos brevemente como:

$$-D_1(x,y)\frac{\partial^2 u}{\partial x^2}(x,y) - D_2(x,y)\frac{\partial^2 u}{\partial y^2}(x,y) + V_1(x,y)\frac{\partial u}{\partial x}(x,y)+$$

$$+V_2(x,y)\frac{\partial u}{\partial y}(x,y) + q(x,y)u(x,y) = f(x,y) \qquad (x,y)\in\Omega$$

donde

$$V_1(x,y) = \left(\widetilde{V}_1(x,y) - \frac{\partial D_1}{\partial x}(x,y)\right), \qquad V_2(x,y) = \left(\widetilde{V}_2(x,y) - \frac{\partial D_2}{\partial y}(x,y)\right)$$

y

$$q(x,y) = \left(\widetilde{q}(x,y) + \frac{\partial \widetilde{V}_1}{\partial x}(x,y) + \frac{\partial \widetilde{V}_2}{\partial y}(x,y)\right).$$

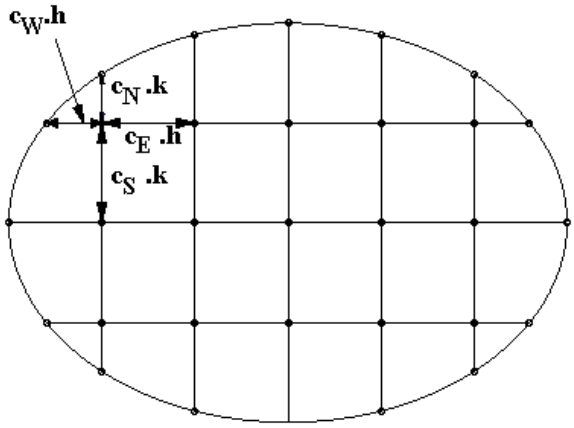
Con esta forma de proceder la aplicación de cualquiera de los esquemas en diferencias finitas antes tratados a la ecuación con coeficientes variables se realiza sin más que sustituir en las fórmulas obtenidas D_1 por $D_{i,j}^{(1)} = D_1(x_i, y_j)$, D_2 por $D_{i,j}^{(2)} = D_2(x_i, y_j)$, V_1 por $V_{i,j}^{(1)} = V_1(x_i, y_j)$, D_1 por $V_{i,j}^{(2)} = V_2(x_i, y_j)$ y q por $q_{i,j} = q(x_i, y_j)$.

Ejercicio propuesto:

Te dejamos como ejercicio propuesto la sencilla tarea de escribir las expresiones resultantes en el caso del esquemas de 5 *puntos en cruz, tanto con convección descentrada como con convección centrada.*

F) Tratamiento de dominios no rectangulares.

La consideración de mallados no uniformes nos permite tratar fácilmente dominios no rectangulares. En efecto, utilizando la notación seguida en el



apartado E) anterior y siendo Ω un abierto conexo de IR^2, consideremos el problema estacionario:

$$-D_1(x,y)\frac{\partial^2 u}{\partial x^2}(x,y) - D_2(x,y)\frac{\partial^2 u}{\partial y^2}(x,y) + V_1(x,y)\frac{\partial u}{\partial x}(x,y)+$$

$$+V_2(x,y)\frac{\partial u}{\partial y}(x,y) + q(x,y)u(x,y) = f(x,y) \qquad (x,y) \in \Omega$$

con la condición de contorno:

$$u(x,y) = u_D(x,y) \qquad (x,y) \in \partial\Omega$$

En este caso puede realizarse un mallado no uniforme que se adapte al contorno $\partial\Omega$ del dominio. Esta situación es la que se ilustra en la figura siguiente.

Puesto que en los nodos de la frontera se conoce el valor de la función solución al estar dada por la condición de tipo Dirichlet considerada (a continuación trataremos el caso de condiciones de contorno más generales) no habrá dificultad conceptual alguna (aunque puede haberla en cuanto a los cálculos a realizar) para aplicar los esquemas en diferencias antes planteados con mallados no uniformes a este tipo de dominios.

G) Imposición de condiciones de contorno más generales.

Hasta ahora hemos estado trabajando sobre problemas de contorno estacionarios en los que se imponían condiciones de frontera de tipo Dirichlet (es decir, se asignaba el valor de la solución en los puntos de la frontera). En numerosos problemas de la ingeniería estas condiciones no son realistas pues no se puede predeterminar el valor de la solución (la temperatura o la concentración de un soluto) en toda la frontera del dominio estudiado. En esos casos lo que

sí será factible hacer es medir el flujo de la función u en la parte de la frontera en que no se conozca el valor de dicha función. En este sentido las condiciones sobre la frontera pueden generalizarse considerándose entonces condiciones del tipo siguiente:

$$\left([\boldsymbol{\alpha}(x,y)]\nabla u(x,y) + \overrightarrow{\boldsymbol{\beta}}(x,y)u(x,y)\right) \bullet \overrightarrow{\mathbf{n}} = g(x,y) \qquad \forall (x,y) \in \partial\Omega$$

donde $[\boldsymbol{\alpha}(x,y)]$ es una función matricial "dependiente" de la función matricial $[\mathbf{D}(x,y)]$ que representaba al tensor de difusividad (y habitualmente coincidente con ésta), $\overrightarrow{\boldsymbol{\beta}}(x,y)$ es un campo vectorial "dependiente" al campo de velocidades $\overrightarrow{\mathbf{V}}(x,y)$ que interviene en la formulación de nuestra EDP (y habitualmente coincidente con dicho campo), $\overrightarrow{\mathbf{n}}$ representa el vector normal unitario saliente de Ω en el punto de la frontera $\partial\Omega$ en el que se aplique la condición de contorno y $g(x,y)$ es una función conocida (y suficientemente regular) definida en los puntos de la frontera $\partial\Omega$.

NOTA:

Puedes consultar en C. Conde y E. Schiavi[3] *cómo se determina en cada punto de* $\partial\Omega$ *el vector normal* $\overrightarrow{\mathbf{n}}$.

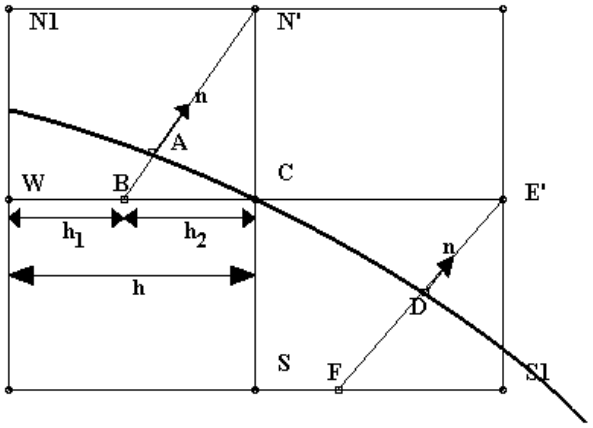
Cuando se consideran condiciones de contorno de este tipo, el valor de la solución en los nodos de la frontera debe calcularse pues es desconocido. Ello nos obliga a tener que plantear el esquema en diferencias finitas también en los nodos de la frontera con la dificultad de que dichos nodos no tienen todos los nodos "vecinos" (N, S, E y W en el caso, por ejemplo de un esquema de 5 puntos) en los que soportar el esquema en diferencias. Estos nodos, que sin formar parte del mallado real deberán ser considerados para poder plantear el esquema en los nodos de la frontera, serán denominados **nodos ficticios** (o pseudonodos) y el valor en ellos se deberá calcular a partir de las condiciones de contorno como una combinación de valores en los nodos reales de nuestro mallado.

Ilustraremos la forma de proceder en este caso sobre el esquema de 5 puntos en cruz. La figura siguiente recoge el caso de un mallado en el que de los 5 puntos "asociados" a un nodo C de la frontera y los nodos S y W pertenecen al dominio Ω pero no así los nodos N y E (que por eso se han denotado como N' y E').

Si en el nodo C se aplica el esquema en diferencias finitas se obtendrá la ecuación:

$$a_{C,S}u_S + a_{C,W}u_W + a_{C,C}u_C + a_{C,E'}u_{E'} + a_{C,N'}u_{N'} = f_C$$

[3] C. Conde y E. Schiavi. (2000). Guiones de la asignatura de Elementos de Matemáticas. Universidad Rey Juan Carlos.



Los valores nodales $u_{N'}$ y $u_{E'}$ no forman parte de las incógnitas ni los nodos E' y N' pertenecen a Ω. Serán por tanto dos nodos ficticios (o pseudonodos) y a los valores a ellos asignados les denominaremos valores nodales ficticios. Como se señaló más arriba, los valores nodales en ellos serán calculados en función de los valores en otros nodos "reales" pertenecientes a Ω a través de las condiciones de contorno. Por ejemplo, para asignar un valor a $u_{N'}$ deberemos considerar la recta que pasando por N' es normal a la frontera $\partial\Omega$. Esta recta cortará a la frontera en el punto A y a la recta que une los nodos W y C en el punto B (véase la figura anterior).

NOTA:

Deliberadamente, para no complicar el proceso que se está exponiendo, se ha sido poco riguroso con el planteamiento anterior. En efecto, según como sea la frontera podría haber más de una recta normal a $\partial\Omega$ que pasara por el punto N'. Para eliminar esta posibilidad deberían exigirse ciertas hipótesis sobre la curva que define la frontera Ω. Dichas hipótesis, que resumiremos diciendo que la frontera $\partial\Omega$ sea suficientemente regular, supondremos en lo que sigue que son satisfechas por la función que defina $\partial\Omega$.

Si se plantea la condición de contorno en el punto A se podrá escribir:

$$([\boldsymbol{\alpha}(x_A, y_A)]\,(\nabla u(x,y))_A) \bullet \overrightarrow{\mathbf{n}} = g(x_A, y_A) - (\overrightarrow{\boldsymbol{\beta}}(x_A, y_A) \bullet \overrightarrow{\mathbf{n}}) u(x_A, y_A) \Rightarrow$$

$$\Rightarrow \left(\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\right)_A = \frac{g(x_A, y_A) - (\overrightarrow{\boldsymbol{\beta}}(x_A, y_A) \bullet \overrightarrow{\mathbf{n}}) u(x_A, y_A)}{\left([\boldsymbol{\alpha}(x_A, y_A)]\overrightarrow{\mathbf{n}}\right) \bullet \overrightarrow{\mathbf{n}}} = \alpha + \gamma u(x_A, y_A).$$

En esta expresión, a su vez, se puede aproximar la derivada normal mediante:

$$\left(\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\right) \approx \frac{u_{N'} - u_B}{d_{N'B}}$$

donde $d_{N'B}$ es la distancia entre los puntos N' y B. Con ello se tendría que:

$$\frac{u_{N'} - u_B}{d_{N'B}} = \alpha + \gamma u(x_A, y_A).$$

Pero ni los valores en B ni en A son tampoco valores nodales (pues ni A ni B son nodos de nuestro mallado). Por ello estos valores deben, nuevamente, ser aproximados. En el caso de u_B puede procederse por interpolación lineal de forma que, según la notación de la figura resultaría:

$$u_B \approx \frac{h_2}{h} u_W + \frac{h_1}{h} u_C$$

Y en cuanto al valor en A podrá interpolarse sobre la propia curva que define $\partial\Omega$ o sobre la propia recta que nos une B con N'. Usando esta última opción se tendrá que:

$$u(x_A, y_A) \approx \frac{d_{N'A}}{d_{N'B}} u_B + \frac{d_{AB}}{d_{N'B}} u_{N'} = \frac{d_{N'A}}{d_{N'B}} \frac{h_2}{h} u_W + \frac{d_{N'A}}{d_{N'B}} \frac{h_1}{h} u_C + \frac{d_{AB}}{d_{N'B}} u_{N'}$$

De estas tres últimas expresiones se tiene que:

$$u_{N'} = \left(\frac{h_2(1 + d_{N'A})}{h(1 - d_{AB})} \right) u_W + \left(\frac{h_1(1 + d_{N'A})}{h(1 - d_{AB})} \right) u_C + \frac{d_{N'B}\gamma}{(1 - d_{AB})} =$$

$$\beta_W u_W + \beta_C u_C + \beta_{N'}$$

Esta expresión del valor nodal ficticio $u_{N'}$ puede entonces sustituirse en la expresión del esquema obteniendo:

$$a_{C,S} u_S + a_{C,W} u_W + a_{C,C} u_C + a_{C,E'} u_{E'} + a_{C,N'} u_{N'} = f_C \Rightarrow$$

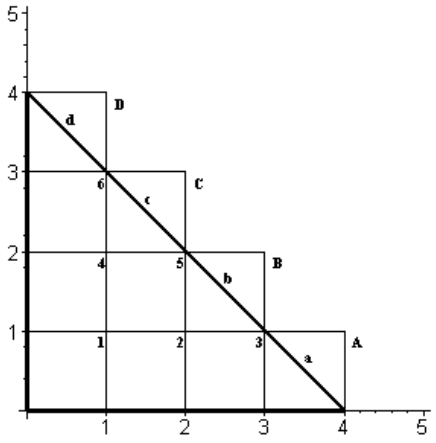
$$a_{C,S} u_S + a_{C,W} u_W + a_{C,C} u_C + a_{C,E'} u_{E'} + a_{CN'}(\beta_W u_W + \beta_C u_C) = f_C - a_{CN'}\beta_{N'} \Rightarrow$$

$$a_{C,S} u_S + (a_{C,W} + a_{C,N'}\beta_W) u_W + (a_{C,C} + a_{C,N'}\beta_C) u_C + a_{C,E'} u_{E'} = f_C - a_{CN'}\beta_{N'}$$

Ejercicio propuesto:

Realiza un proceso similar al anterior para "eliminar" en la ecuación anterior $u_{E'}$ expresándolo en función de valores nodales pertenecientes a Ω.

Ilustremos todo lo anterior mediante un ejemplo.



H) Un ejemplo.

Sea Ω el triángulo rectángulo de vértices $(0,0)$, $(4,0)$ y $(0,4)$. En él se considerará el problema:

$$\begin{cases} -\nabla \bullet ([\mathbf{D}]\,\nabla u(x,y)) + \nabla \bullet \left(\overrightarrow{V}(x,y)u(x,y)\right) & = f(x,y) \quad en\ \ \Omega \\ u(x,0) = 0, & u(0,y) = 0, \\ \left(-[\mathbf{D}]\nabla u(x,y) + \overrightarrow{V}(x,y)u(x,y)\right) \bullet \overrightarrow{n} = g(x,y,u) & en\ \ L \end{cases}$$

donde

$$[\mathbf{D}] = \begin{bmatrix} 5 & 0 \\ 0 & 3 \end{bmatrix}, \quad \overrightarrow{V}(x,y) = \begin{Bmatrix} -y \\ x \end{Bmatrix}, \quad f(x,y) = x^2(x^2 - 3y^2) - 30xy$$

$$g(x,y,u) = \frac{\sqrt{2}}{2}\left(u\,(x-y) - 3x^3 - 15x^2y\right)$$

y L es el lado que une los vértices $(4,0)$ y $(0,4)$ del triángulo.

Se pide obtener una aproximación del valor que toma la solución en los nodos 1, 2, 3, 4, 5 y 6 del mallado que se representa en la figura siguiente mediante un esquema de diferencias finitas de 5 puntos en cruz para el término difusivo y descentrado en retroceso según el valor de la correspondiente componente de la velocidad para el término convectivo.

Solución:

Reescribamos la ecuación diferencial en la forma:

$$-\nabla \bullet ([\mathbf{D}]\,\nabla u(x,y)) + \nabla \bullet \left(\overrightarrow{V}(x,y)u(x,y)\right) = f(x,y) \Rightarrow$$

$$\Rightarrow -\nabla\bullet([\mathbf{D}]\,\nabla u(x,y))+\left(\nabla\bullet\overrightarrow{V}(x,y)\right)u(x,y)+\overrightarrow{V}(x,y)\bullet\nabla u(x,y)=f(x,y)\Rightarrow$$

$$\Rightarrow -5\frac{\partial^2 u}{\partial x^2}-3\frac{\partial^2 u}{\partial y^2}+(-y)\frac{\partial u}{\partial x}+x\frac{\partial u}{\partial y}=f(x,y)$$

Esta ecuación es la que deberá discretizarse en cada uno de los nodos del mallado. Hagámoslo nodo a nodo:

En el nodo 1:

$$-5\frac{\partial^2 u}{\partial x^2}\approx -5(0-2u_1+u_2)=10u_1-5u_2$$

$$-3\frac{\partial^2 u}{\partial y^2}\approx -3(0-2u_1+u_4)=6u_1-3u_4$$

Puesto que la componente 1^a de la velocidad en el nodo 1 toma el valor: $\overrightarrow{V}_1(1,1)=-1$ utilizaremos una fórmula progresiva para aproximar la derivada parcial respecto a x :

$$(-y)\frac{\partial u}{\partial x}\approx(-y_1)\frac{u_2-u_1}{1}=u_1-u_2$$

Puesto que la componente 2^a de la velocidad en el nodo 1 toma el valor: $\overrightarrow{V}_2(1,1)=1$ utilizaremos una fórmula regresiva para aproximar la derivada parcial respecto a y :

$$(x)\frac{\partial u}{\partial y}\approx(x_1)\frac{u_1-0}{1}=u_1$$

En este nodo la función $f(x,y)$ toma el valor:

$$f(1,1)=-32.$$

Por tanto la ecuación a plantear en el nodo 1 será:

$$10u_1-5u_2+6u_1-3u_4+u_1-u_2+u_1=-32\Rightarrow$$

$$\Rightarrow 18u_1-6u_2-3u_4=-32\qquad(N1)$$

En el nodo 2:

$$-5\frac{\partial^2 u}{\partial x^2}\approx -5(u_1-2u_2+u_3)=-5u_1+10u_2-5u_3$$

$$-3\frac{\partial^2 u}{\partial y^2}\approx -3(0-2u_2+u_5)=6u_2-3u_5$$

Puesto que la componente 1^a de la velocidad en el nodo 2 toma el valor: $\overrightarrow{V}_1(2,1) = -1$ utilizaremos una fórmula progresiva para aproximar la derivada parcial respecto a x :

$$(-y)\frac{\partial u}{\partial x} \approx (-y_2)\frac{u_3 - u_2}{1} = u_2 - u_3$$

Puesto que la componente 2^a de la velocidad en el nodo 2 toma el valor: $\overrightarrow{V}_2(2,1) = 2$ utilizaremos una fórmula regresiva para aproximar la derivada parcial respecto a y :

$$(x)\frac{\partial u}{\partial y} \approx (x_2)\frac{u_2 - 0}{1} = 2u_2$$

En este nodo la función $f(x,y)$ toma el valor:

$$f(2,1) = -56.$$

Por tanto la ecuación a plantear en el nodo 2 será:

$$-5u_1 + 10u_2 - 5u_3 + 6u_2 - 3u_5 + u_2 - u_3 + 2u_2 = -56 \Rightarrow$$

$$\Rightarrow -5u_1 + 19u_2 - 6u_3 - 3u_5 = -56 \qquad (N2)$$

En el nodo 3:

$$-5\frac{\partial^2 u}{\partial x^2} \approx -5(u_2 - 2u_3 + u_A) = -5u_2 + 10u_3 - 5u_A$$

$$-3\frac{\partial^2 u}{\partial y^2} \approx -3(0 - 2u_3 + u_B) = 6u_3 - 3u_B$$

Puesto que la componente 1^a de la velocidad en el nodo 3 toma el valor: $\overrightarrow{V}_1(3,1) = -1$ utilizaremos una fórmula progresiva para aproximar la derivada parcial respecto a x :

$$(-y)\frac{\partial u}{\partial x} \approx (-y_3)\frac{u_A - u_3}{1} = u_3 - u_A$$

Puesto que la componente 2^a de la velocidad en el nodo 3 toma el valor: $\overrightarrow{V}_2(3,1) = 3$ utilizaremos una fórmula regresiva para aproximar la derivada parcial respecto a y :

$$(x)\frac{\partial u}{\partial y} \approx (x_3)\frac{u_3 - 0}{1} = 3u_3$$

En este nodo la función $f(x,y)$ toma el valor:

$$f(3,1) = -36.$$

Por tanto la ecuación a plantear en el nodo 3 será:

$$-5u_2 + 10u_3 - 5u_A + 6u_3 - 3u_B + u_3 - u_A + 3u_3 = -36 \Rightarrow$$

$$\Rightarrow -5u_2 + 20u_3 - 6u_A - 3u_B = -36 \qquad (N3)$$

En el nodo 4:

$$-5\frac{\partial^2 u}{\partial x^2} \approx -5(0 - 2u_4 + u_5) = 10u_4 - 5u_5$$

$$-3\frac{\partial^2 u}{\partial y^2} \approx -3(u_1 - 2u_4 + u_6) = -3u_1 + 6u_4 - 3u_6$$

Puesto que la componente 1^a de la velocidad en el nodo 4 toma el valor: $\overrightarrow{V}_1(1,2) = -2$ utilizaremos una fórmula progresiva para aproximar la derivada parcial respecto a x :

$$(-y)\frac{\partial u}{\partial x} \approx (-y_4)\frac{u_5 - u_4}{1} = 2u_4 - 2u_5$$

Puesto que la componente 2^a de la velocidad en el nodo 4 toma el valor: $\overrightarrow{V}_2(1,2) = 1$ utilizaremos una fórmula regresiva para aproximar la derivada parcial respecto a y :

$$(x)\frac{\partial u}{\partial y} \approx (x_4)\frac{u_4 - u_1}{1} = u_4 - u_1$$

En este nodo la función $f(x,y)$ toma el valor:

$$f(1,2) = -71.$$

Por tanto, la ecuación a plantear en el nodo 4 será:

$$10u_4 - 5u_5 - 3u_1 + 6u_4 - 3u_6 + 2u_4 - 2u_5 + u_4 - u_1 = -71 \Rightarrow$$

$$\Rightarrow -4u_1 + 19u_4 - 7u_5 - 3u_6 = -71 \qquad (N4)$$

En el nodo 5:

$$-5\frac{\partial^2 u}{\partial x^2} \approx -5(u_4 - 2u_5 + u_B) = -5u_4 + 10u_5 - 5u_B$$

$$-3\frac{\partial^2 u}{\partial y^2} \approx -3(u_2 - 2u_5 + u_C) = -3u_2 + 6u_5 - 3u_C$$

Puesto que la componente 1^a de la velocidad en el nodo 5 toma el valor: $\overrightarrow{V}_1(2,2) = -2$ utilizaremos una fórmula progresiva para aproximar la derivada parcial respecto a x :

$$(-y)\frac{\partial u}{\partial x} \approx (-y_5)\frac{u_B - u_5}{1} = 2u_5 - 2u_B$$

Puesto que la componente 2^a de la velocidad en el nodo 5 toma el valor: $\overrightarrow{V}_2(2,2) = 2$ utilizaremos una fórmula regresiva para aproximar la derivada parcial respecto a y :

$$(x)\frac{\partial u}{\partial y} \approx (x_5)\frac{u_5 - u_2}{1} = 2u_5 - 2u_2$$

En este nodo la función $f(x,y)$ toma el valor:

$$f(2,2) = -152.$$

Por tanto, la ecuación a plantear en el nodo 5 será:

$$-5u_4 + 10u_5 - 5u_B - 3u_2 + 6u_5 - 3u_C + 2u_5 - 2u_B +$$

$$+2u_5 - 2u_2 = -152 \Rightarrow$$

$$\Rightarrow -5u_2 - 5.u_4 + 20u_5 - 7u_B - 3u_C = -152 \qquad (N5)$$

<u>En el nodo 6</u>:

$$-5\frac{\partial^2 u}{\partial x^2} \approx -5(0 - 2u_6 + u_C) = 10u_6 - 5u_C$$

$$-3\frac{\partial^2 u}{\partial y^2} \approx -3(u_4 - 2u_6 + u_D) = -3u_4 + 6u_6 - 3u_D$$

Puesto que la componente 1^a de la velocidad en el nodo 6 toma el valor: $\overrightarrow{V}_1(1,3) = -3$ utilizaremos una fórmula progresiva para aproximar la derivada parcial respecto a x :

$$(-y)\frac{\partial u}{\partial x} \approx (-y_6)\frac{u_C - u_6}{1} = 3u_6 - 3u_C$$

Puesto que la componente 2^a de la velocidad en el nodo 6 toma el valor: $\overrightarrow{V}_2(1,3) = 1$ utilizaremos una fórmula regresiva para aproximar la derivada parcial respecto a y :

$$(x)\frac{\partial u}{\partial y} \approx (x_6)\frac{u_6 - u_4}{1} = u_6 - u_4$$

En este nodo la función $f(x,y)$ toma el valor:

$$f(1,3) = -116.$$

Por tanto, la ecuación a plantear en el nodo 6 será:

$$10u_6 - 5u_C - 3u_4 + 6u_6 - 3u_D + 3u_6 - 3u_C + u_6 - u_4 = -116 \Rightarrow$$

$$\Rightarrow -4u_4 + 20u_6 - 8u_C - 3u_D = -116 \qquad (N6)$$

El conjunto de ecuaciones $(N1)$, $(N2)$, $(N3)$, $(N4)$, $(N5)$ y $(N6)$ forma un sistema de 6 ecuaciones algebraicas con 10 incógnitas (los valores en los nodos del dominio u_1,...., u_6 y los cuatro valores en los nodos ficticios u_A, u_B, u_C y u_D). Para eliminar cuatro incógnitas debemos considerar la condición de Neumann en la frontera L. Para ello sabemos que:

$$\left(-[\mathbf{D}]\nabla u(x,y) + \overrightarrow{V}(x,y)u(x,y)\right) \bullet \overrightarrow{n} = g(x,y,u) \Rightarrow$$

$$\Rightarrow -\left([\mathbf{D}]\nabla u(x,y)\right) \bullet \overrightarrow{n} = g(x,y,u) - \left(\overrightarrow{V}(x,y) \bullet \overrightarrow{n}\right) u(x,y)$$

Y como

$$\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}} = \nabla u \bullet \overrightarrow{\mathbf{n}}$$

se verificará que

$$\nabla u = \frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\,\overrightarrow{\mathbf{n}}$$

por lo que

$$-\left([\mathbf{D}]\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\,\overrightarrow{\mathbf{n}}\right) \bullet \overrightarrow{\mathbf{n}} = g(x,y,u) - \left(\overrightarrow{V}(x,y) \bullet \overrightarrow{\mathbf{n}}\right) u(x,y)$$

de donde la condición de contorno dada puede transformarse en

$$\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}} = \frac{\left(\overrightarrow{V}(x,y) \bullet \overrightarrow{\mathbf{n}}\right) u(x,y) - g(x,y,u)}{\left([\mathbf{D}]\overrightarrow{\mathbf{n}}\right) \bullet \overrightarrow{\mathbf{n}}}$$

En nuestro ejemplo, teniendo en cuenta que:

$$\overrightarrow{\mathbf{n}} = \left\{ \begin{array}{c} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{array} \right\}$$

y habida cuenta de las expresiones de [**D**], $\overrightarrow{V}(x,y)$ y $g(x,y,u)$ resultará que en cualquier punto p de la frontera L:

$$\left(\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\right)_p =$$

$$\frac{\left(\left\{\begin{array}{c} -y_p \\ x_p \end{array}\right\} \bullet \left\{\begin{array}{c} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{array}\right\}\right) u_p - \frac{1}{\sqrt{2}}(x_p - y_p)u_p + \frac{1}{\sqrt{2}}x_p^2(3x_p + 15y_p)}{\left(\left[\begin{array}{cc} 5 & 0 \\ 0 & 3 \end{array}\right] \left\{\begin{array}{c} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{array}\right\}\right) \bullet \left\{\begin{array}{c} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{array}\right\}}$$

$$\left(\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\right)_p = \frac{\frac{1}{\sqrt{2}}(x_p - y_p)u_p - \frac{1}{\sqrt{2}}(x_p - y_p)u_p + \frac{1}{\sqrt{2}}x_p^2(3x_p + 15y_p)}{4}$$

$$\left(\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\right)_p = \frac{x_p^2(3x_p + 15y_p)}{4\sqrt{2}} = \widetilde{g}(x_p, y_p)$$

Esta ecuación sobre los puntos de la frontera L puede particularizarse para los "pseudonodos" a, b, c y d representados en la figura del mallado, obteniéndose que:

*) En el pseudonodo n_a:

$$\left(\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\right)_a = \widetilde{g}(x_a, y_a) = \widetilde{g}(3{,}5, 0{,}5) = \frac{441}{8\sqrt{2}}$$

de donde aproximando la derivada normal en a resulta:

$$\frac{u_A - 0}{\sqrt{2}h} = \frac{441}{8\sqrt{2}} \Rightarrow u_A = \frac{441}{8}$$

*) En el pseudonodo n_b:

$$\left(\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\right)_b = \widetilde{g}(x_b, y_b) = \widetilde{g}(2{,}5, 1{,}5) = \frac{375}{8\sqrt{2}}$$

de donde aproximando la derivada normal en a resulta:

$$\frac{u_B - u_2}{\sqrt{2}h} = \frac{375}{8\sqrt{2}} \Rightarrow u_B = u_2\frac{375}{8}$$

*) En el pseudonodo n_c:

$$\left(\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\right)_c = \widetilde{g}(x_c, y_c) = \widetilde{g}(1{,}5, 2{,}5) = \frac{189}{8\sqrt{2}}$$

de donde aproximando la derivada normal en a resulta:

$$\frac{u_C - u_4}{\sqrt{2}h} = \frac{189}{8\sqrt{2}} \Rightarrow u_C = u_4 + \frac{189}{8}$$

*) En el pseudonodo n_d:

$$\left(\frac{\partial u}{\partial \overrightarrow{\mathbf{n}}}\right)_d = \widetilde{g}(x_d, y_d) = \widetilde{g}(0{,}5, 3{,}5) = \frac{27}{8\sqrt{2}}$$

de donde aproximando la derivada normal en a resulta:

$$\frac{u_D - 0}{\sqrt{2}h} = \frac{27}{8\sqrt{2}} \Rightarrow u_D = \frac{27}{8}$$

Introduciendo estas expresiones en las ecuaciones $(N1) - (N6)$ se obtiene finalmente el sistema:

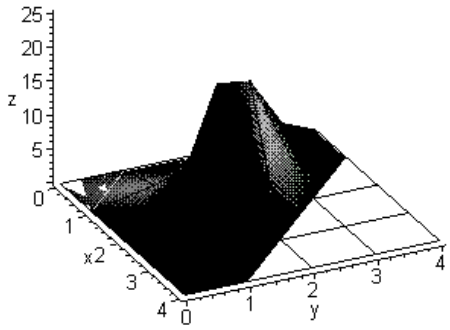
$$\left\{\begin{array}{lll}
18u_1 - 6u_2 - 3u_4 & = -32 & (Ec{,}1) \\
-5u_1 + 19u_2 - 6u_3 - 3u_5 & = -56 & (Ec{,}2) \\
-8u_2 + 20u_3 & = -36 + 6\dfrac{441}{8} + 3\dfrac{375}{8} & (Ec{,}3) \\
-4u_1 + 19u_4 - 7u_5 - 3u_6 & = -71 & (Ec{,}4) \\
-12u_2 - 8.u_4 + 20u_5 & = -152 - 7\dfrac{375}{8} + 3\dfrac{189}{16} & (Ec{,}5) \\
-8u_4 + 20u_6 & = -116 + 189 + 3\dfrac{27}{16} & (Ec{,}6)
\end{array}\right\}$$

Este sistema ya tiene 6 incógnitas y 6 ecuaciones y puede procederse a resolverlo mediante cualquiera de los métodos de resolución de ecuaciones lineales. Si así se hace se obtiene la solución que se escribe en la tabla siguiente (en la que también se recogen los valores de la solución exacta $u(x, y) = yx^3$ para poder compararlos):

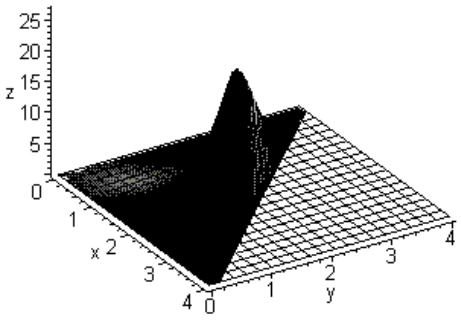
$Nodo(i)$	x_i	y_i	u_i	$u(x_i, y_i)$	$u(x_i, y_i) - u_i$
1	1	1	1,9173978	1.	−0,9173978
2	2	1	8,5962066	8.	−0,5962066
3	3	1	25,207233	27.	1,792767
4	1	2	4,9786397	2.	−2,9786397
5	2	2	19,499180	16.	−3,499180
6	1	3	7,1434338	3.	−4,1434338

Observa que la precisión es bastante mala. Ello es debido al uso de un mallado "grosero" (con una distancia entre puntos nodales excesiva). La representación de las soluciones exacta y aproximada es la siguiente:

SOLUCION APROXIMADA (h = 1.0)



SOLUCION EXACTA



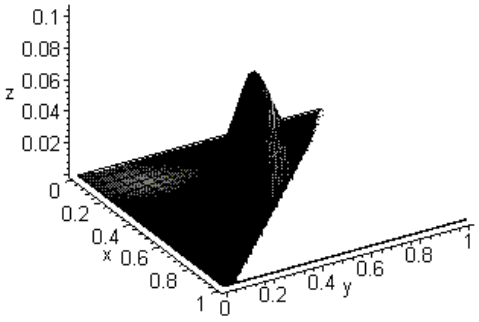
A continuación te presentamos un programa escrito en MAPLE con el que puedes introducir más nodos en el dominio triangular. En él se recogen las gráficas obtenidas con un valor de la distancia entre abscisas y ordenadas dado por $h = 0{,}05$ para el mismo problema planteado sobre el triángulo de vértices $(0,0)$, $(1,0)$ y (0, 1).

```
>restart;
>with(linalg): with(plots):
>Digits:=8;
>NX:=20; LONG:=1.;
>ub:=(x,y)->0.;
>ud:=(x,y)->0.;
>g:=(x,y)->x*x*(3*x+15*y)/(4*sqrt(2));
>f:=(x,y)->x*x*(x*x-3*y*y)-30*x*y;
>difusion1:=(x,y)->5; difusion2:=(x,y)->3;
>v1:=(x,y)->-y; v2:=(x,y)->x;
>reac:=(x,y)->0;
>h:=LONG/NX;
>NN:=(NX+2)*(NX+1)/2;
>A:=array(sparse,1..NN,1..NN); b:=vector(NN);
>for i from 1 to NX+1 by 1 do:
    abscisa[i]:=(i-1)*h;
    ordenada[i]:=(i-1)*h;
od;
>k:=0:
for j from 1 to NX+1 by 1 do
     for i from 1 to NX+2 - j by 1 do
        k:=k+1:
         x[k]:=abscisa[i]:
        y[k]:=ordenada[j]:
    od:
od:
>print(k);
>for k from 1 to NN by 1 do
    print(k,x[k],y[k]);
 od;
>k:=0:
for j from 1 to NX+1 by 1 do
       if j = 1 then
          for i from 1 to NX+1 by 1 do
             k:=k+1:
             A[k,k]:=1.:
             b[k]:=ub(x[k],y[k]):
```

```
    od:
elif (j<NX+1) then
   for i from 1 to (NX+2 - j)by 1 do
       k:= k+1:
        if i = 1 then
           A[k,k]:=1:
            b[k]:=ud(x[k],y[k]):
        elif i<(NX+2 - j) then
           velx:=v1(x[k],y[k]):
           vely:=v2(x[k],y[k]):
           dx:=difusion1(x[k],y[k]):
           dy:=difusion2(x[k],y[k]):
           q:=reac(x[k],y[k]):
           alfa:=(1/2)+ abs(velx)/(2*velx):
           beta:=(1/2)+abs(vely)/(2*vely):
           alfa1:=1-alfa: beta1:=1-beta:
           w:=k-1: e:=k+1: s:=k-(NX+3-j):
            n:= k+(NX+2-j):
            A[k,k]:=((2*(alfa*velx+beta*vely+
                     ((dx+dy)/h))-velx-vely)/h)+q:
           A[k,w]:= -((dx/h)+alfa*velx)/h:
           A[k,e]:= (alfa1*velx-(dx/h))/h:
           A[k,s]:= -((dy/h)+beta*vely)/h:
           A[k,n]:= (beta1*vely-(dy/h))/h:
           b[k]:=f(x[k],y[k]):
       else
           velx:=v1(x[k],y[k]):
           vely:=v2(x[k],y[k]):
           dx:=difusion1(x[k],y[k]):
           dy:=difusion2(x[k],y[k]):
           dis:=sqrt(2)*h:
           q:=reac(x[k],y[k]):
           alfa:=(1/2)+ abs(velx)/(2*velx):
           beta:=(1/2)+abs(vely)/(2*vely):
           alfa1:=1-alfa: beta1:=1-beta:
           w:=k-1: s:=k-(NX+3-j):
           xmedn:=x[k]-(h/2):
           ymedn:=y[k]+(h/2):
           xmeds:=x[k]+(h/2):
           ymeds:=y[k]-(h/2):
           A[k,k]:=((2*(alfa*velx+beta*vely+
                    ((dx+dy)/h))-velx-vely)/h)+q:
```

```
                A[k,w]:= (beta1*vely-alfa*velx-
                        ((dx+dy)/h))/h:
                A[k,s]:= (alfa1*velx-beta*vely-
                        ((dx+dy)/h))/h:
                b[k]:=f(x[k],y[k])-
                    ((alfa1*velx-(dx/h))/h)*dis*g(xmeds,ymeds)-
                    ((beta1*vely-(dy/h))/h)*dis*g(xmedn,ymedn):
            fi:
        od:
    else
        k:=k+1:
        A[k,k]:=1:
        b[k]:=ud(x[k],y[k]):
    fi:
od:
>print(A);
>print(b);
>u:=(x,y)->y * x^3;
>for k from 1 to NN by 1 do
    uexac[k]:=evalf(u(x[k],y[k])):
od:
>uap:=linsolve(A,b):
>mayor:=0.;kpos:=0;
>for k from 1 to NN by 1 do
    difer[k]:=uexac[k]-uap[k]:
     if (abs(difer[k])>mayor) then
        mayor:=abs(difer[k]):
        kpos:=k:
    fi:
od:
>for k from 1 to NN by 1 do
    print(uap[k],uexac[k],difer[k]);
od;
>print(kpos,x[kpos],y[kpos],uap[kpos],uexac[kpos],mayor);
>for k from 1 to NN by 1 do
     print(uap[k],uexac[k],difer[k]);
od;
>data:=[seq([x[k],y[k],uap[k]],k=1..NN)]:
>k:=0:
 for j from 1 to NX+1 by 1 do
    for i from 1 to NX+2 - j by 1 do
        k:=k+1:
```

SOLUCION EXACTA

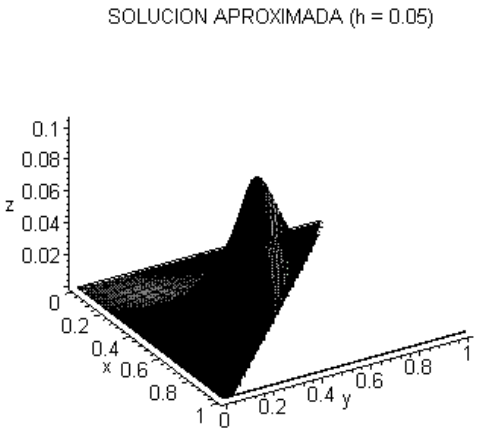



```
        datos[i,j]:=uap[k]:
      od:
    if j >1 then
      for i from NX+3-j to NX+1 by 1 do
        datos[i,j]:=0.:
      od:
    fi:
  od:
  >data:=[seq([seq([abscisa[i],ordenada[j],datos[i,j]],i=1..NX+1)],j=1..NX+1)]:
  > diba:=surfdata(data,axes=framed,labels=[x,y,z],orientation=[-31,52],
title='SOLUCIÓN APROXIMADA (h = 0.05)'):
  >uuu:= proc(x,y) if y+x<=1 then y*x^3 else 0 fi: end;
  >dibe:=plot3d(uuu,0..1,0..1,grid=[20,20],axes=framed,
labels=[x,y,z],orientation=[-31,52],title='SOLUCIÓN EXACTA'):
  >display(dibe);
  >display(diba);
  >fin;
```

En este caso la mayor diferencia entre los valores nodales de la solución aproximada y de la solución exacta se produce en el nodo 186 de coordenadas $x_{186} = 0{,}45$, $y_{186} = 0{,}55$ resultando que $u_{186} = 0{,}061088534$ y $u(0{,}45, 0{,}55) = 0{,}05011875$ por lo que $|u_{186} - u(0{,}45, 0{,}55)| = 0{,}010969784$.

I) Algunos comentarios sobre los esquemas en diferencias para problemas estacionarios.

1^o) Los esquemas que hemos presentado anteriormente son los más usuales y básicos. No obstante existen muchos otros esquemas en los que se hace intervenir en la discretización de la EDP en cada nodo a un cierto número de nodos



vecinos. En la bibliografía a este tema encontrarás referencias (como Lapidus y Pinder[4], Smith[5], Morton y Mayers[6], etc...) en las que consultar muy diferentes esquemas en diferencias.

2º) Los esquemas en diferencias presentados se han obtenido a partir de la combinación de desarrollos en serie de Taylor de los valores de una función suficientemente regular en nodos vecinos a aquel en el que se plantea la EDP. Existen otras formas diferentes para obtener estos esquemas partiendo de "formulaciones integrales" del problema. En Morton y Mayers[7], por ejemplo puedes encontrar detalles sobre esta forma de proceder.

3º) Los esquemas en diferencias finitas que permiten aproximar problemas estacionarios acaban formulando una ecuación algebraica en cada nodo del mallado formando así un sistema con tantas ecuaciones como nodos y el mismo número de incógnitas. A partir de ese momento, el problema consiste en resolver **de forma eficiente** el sistema algebraico obtenido. Además, en cada una de las ecuaciones, todos los coeficientes de la ecuación serán nulos salvo los que multipliquen a valores nodales correspondientes a nodos "vecinos" del nodo al que se asocie dicha ecuación algebraica (entendiendo por "vecinos" aquellos que se utilizan para discretizar las derivadas en dicho nodo). En otros términos la matriz del sistema algebraico tendrá la mayor parte de sus elementos nulos. Ello se expresará diciendo que es una **matriz hueca**. Esto plantea problemas tanto de almacenamiento (¿para qué almacenar los coeficientes que "a priori"

[4]L. Lapidus y G.F. Pinder.(1982) Numerical solution of partial differential equations in science and engineering. Ed.: John Wiley.

[5]G.D. Smith (1.985) Numerical Solution of partial Differential Equations. Finite Difference Methods. (3^a edición, 4^a reimpresión (1996)). Ed. Clarendon Press.

[6]K.W. Morton y D.F. Mayers. (1994) Numerical solution of Partial Differential Equations. Ed. Cambridge University Press.

[7]K.W. Morton y D.F. Mayers. (1994) Numerical solutioon of Partial Differential Equations. Ed. Cambridge University Press.

se sabe que va a ser nulos) como de número de operaciones (¿para qué vamos a sumar, restar o multiplicar valores que son nulos si sabemos de antemano que el resultado de esas operaciones va a ser 0?). Todo lo anterior obliga a modificar los métodos clásicos de resolución de sistemas adaptándolos a esta situación. Incluso, en la práctica, es preferible el uso de técnicas iterativas de resolución de sistemas lineales (tales como los métodos de tipo gradiente o métodos de relajación) antes que los métodos directos. Para el estudio de estos métodos iterativos, que aquí no abordaremos, nuevamente te remitimos a la bibliografía indicada al final de este capítulo (por ejemplo Smith[8] o Morton y Mayers[9], o Conde y Winter[10]) para su estudio detallado.

4^o) No hemos tratado problemas formulados en dominios tridimensionales. En todo caso la forma de proceder para deducirlos es análoga y las propiedades vistas para los esquemas en dominios bidimensionales pueden extenderse sin dificultad a dominios tridimensionales. Lo único que sucede es que las fórmulas obtenidas son un poco más pesadas de manipular. Es decir se incrementa el volumen de cálculos a hacer pero la dificultad conceptual en 3 (o en general en n) dimensiones es la misma que en 2 dimensiones.

5^o) Todos los ejemplos considerados anteriormente, así como el planteamiento de los esquemas presentados, han considerado sólo el caso en que la e.d.p. a resolver era lineal. En el último ejemplo (desarrollado en el apartado H) anterior) se consideró no obstante una condición de contorno no lineal (en donde la función definida en el contorno g dependía de la propia incógnita u). En general podrían contemplarse problemas en los que tanto la difusividad como el campo de velocidades como el coeficiente del término reactivo así como las funciones que intervienen en la definición de las condiciones de contorno dependieran de la propia función incógnita. En el caso bidimensional estaríamos tratando problemas con una EDP de la forma:

$$-\frac{\partial}{\partial x}\left(D_1(x,y,u)\frac{\partial u}{\partial x}(x,y)\right)-\frac{\partial}{\partial y}\left(D_2(x,y,u)\frac{\partial u}{\partial y}(x,y)\right)+$$

$$\frac{\partial}{\partial x}\left(V_1(x,y,u)u(x,y)\right)+\frac{\partial}{\partial y}\left(V_2(x,y,u)u(x,y)\right)+q(x,y,u)u(x,y)=$$

$$f(x,y,u),\ (x,y)\in\Omega$$

Ahora no es tan simple proceder a "manipular" previamente nuestra ecuación y, en un primer planteamiento, puede procederse a discretizarla tal cual está

[8] G.D. Smith (1.985) Numerical Solution of partial Differential Equations. Finite Difference Methods. (3^a edición, 4^a reimpresión (1996)). Ed. Clarendon Press.

[9] K.W. Morton y D.F. Mayers. (1994) Numerical solutioon of Partial Differential Equations. Ed. Cambridge University Press.

[10] C. Conde y G. Winter (1.991) Métodos y algoritmos básicos del álgebra numérica. Ed. Reverté.

escrita sobre cada nodo C. Así, utilizando la notación antes introducida y con la aplicación de fórmulas centradas para los términos difusivos y descentradas para los convectivos, para mallados uniformes, puede procederse como sigue:

$$-\frac{\left(D_1(x,y,u)\frac{\partial u}{\partial x}(x,y)\right)_W - \frac{\partial}{\partial x}\left(D_1(x,y,u)\frac{\partial u}{\partial x}(x,y)\right)_E}{2h} -$$

$$-\frac{\left(D_2(x,y,u)\frac{\partial u}{\partial x}(x,y)\right)_N - \frac{\partial}{\partial x}\left(D_2(x,y,u)\frac{\partial u}{\partial x}(x,y)\right)_S}{2k} +$$

$$+\rho_1\frac{(V_1(x,y,u)u(x,y))_C - (V_1(x,y,u)u(x,y))_W}{h} +$$

$$+(1-\rho_1)\frac{(V_1(x,y,u)u(x,y))_C - (V_1(x,y,u)u(x,y))_W}{h}$$

$$+\rho_2\frac{(V_2(x,y,u)u(x,y))_C - (V_2(x,y,u)u(x,y))_S}{k} +$$

$$+(1-\rho_2)\frac{(V_2(x,y,u)u(x,y))_N - (V_2(x,y,u)u(x,y))_C}{k} +$$

$$+q(x_C,y_C,u_C) = f(x_C,y_C,u_C)$$

en donde volviendo a aproximar las derivadas primeras que quedan en el término convectivo mediante fórmulas descentradas, se obtiene:

$$-\frac{D_W^{(1)}(u_W)\frac{u_C-u_W}{h} - D_E^{(1)}(u_E)\frac{u_E-u_C}{h}}{2h}$$

$$-\frac{D_N^{(2)}(u_N)\frac{u_N-u_C}{k} - D_S^{(2)}(u_S)\frac{u_C-u_S}{k}}{2k} +$$

$$+\rho_1\frac{V_C^{(1)}(u_C)u_C - V_W^{(1)}(u_W)u_W}{h} + (1-\rho_1)\frac{V_E^{(1)}(u_E)u_E - V_C^{(1)}(u_C)u_C}{h} +$$

$$+\rho_2\frac{V_C^{(2)}(u_C)u_C - V_S^{(2)}(u_S)u_S}{k} + (1-\rho_2)\frac{V_N^{(2)}(u_N)u_N - V_C^{(2)}(u_C)u_C}{k} +$$

$$+q_C(u_C)u_C = f_C(u_C)$$

En la expresión anterior el superíndice de los coeficientes indica la componente que se considera, el subíndice el punto (x, y) en el que se evalúa y se ha dejado explícita la dependencia respecto al propio valor nodal (u_C, u_S, u_W, u_E o u_N) de los citados coeficientes. La expresión obtenida en cada nodo puede escribirse entonces de forma resumida como:

$$a_{C,S}(u_S)u_S + a_{C,W}(u_W)u_W + a_{C,C}(u_S, u_W, u_C, u_E, u_N)u_C +$$

$$a_{C,E}(u_E)u_E + a_{C,N}(u_N)u_N = f_C(u_C)$$

donde los coeficientes de la ecuación dependen de los propios valores nodales que se tratan de resolver. En resumen, es una ecuación no lineal.

La consideración de todas las ecuaciones algebraicas planteadas en los distintos nodos del mallado nos conduce de esta forma a un sistema de tantas ecuaciones e incógnitas como nodos haya en el mallado que podrá escribirse de la forma:

$$[\mathbf{A}(\{\mathbf{u}\}]\{\mathbf{u}\} = \{\mathbf{b}(\{\mathbf{u}\})\}$$

Este es un sistema no lineal que deberá tratarse según las técnicas analizadas en el primer capítulo del libro.

6^o) En algunos casos el tratamiento de dominios no rectangulares puede realizarse, de forma más cómoda que la relatada para el caso general, realizando un cambio de coordenadas. Es el caso, por ejemplo de trabajar sobre dominios circulares o elípticos en los que un cambio a coordenadas polares o elípticas puede simplificar el tratamiento numérico de los problemas y mejorar la eficiacia de los esquemas utilizados. Puede consultarse, por ejemplo, Morton y Mayers[11] o Smith[12] para un estudio detallado de esta forma de proceder.

4.3. Generalidades sobre el tratamiento de problemas evolutivos.

Siendo $u(x, y, t)$ una función definida en $\overline{\Omega} \times [0, T] = \{\Omega \cup \partial\Omega\} \times [0, T]$ consideremos el problema evolutivo siguiente:

$$\left\{ \begin{array}{ll} \dfrac{\partial(u)}{\partial t} - \nabla \bullet ([\mathbf{D}]\nabla u) + \nabla \bullet \left(\vec{\mathbf{V}} u\right) + qu = f & (x, y) \in \Omega, t \in (0, T) \\ u(x, y, t) = u_D(x, y, t) & (x, y) \in \partial\Omega, t \in (0, T) \\ u(x, y, 0) = u^{(0)}(x, y) & (x, y) \in \overline{\Omega} \end{array} \right\}$$

[11] K.W. Morton y D.F. Mayers. (1994) Numerical solutioon of Partial Differential Equations. Ed. Cambridge University Press.

[12] G.D. Smith (1.985) Numerical Solution of partial Differential Equations. Finite Difference Methods. (3^a edición, 4^a reimpresión (1996)). Ed. Clarendon Press.

cuya EDP escribiremos en forma abreviada como:

$$\frac{\partial u}{\partial t} = L(u) \tag{4.48}$$

donde hemos denotado por:

$$L(u) = f + \nabla \bullet ([\mathbf{D}]\,\nabla u) - \nabla \bullet \left(\overrightarrow{\mathbf{V}} u\right) - qu. \tag{4.49}$$

El tratamiento numérico mediante diferencias finitas de este tipo de problemas se fundamenta en dos pasos: en primer lugar se procede a realizar una **discretización temporal** del problema (4.48) como si de un problema de valor inicial se tratase, es decir por alguna de las técnicas descritas en el capítulo referente a la resolución de problemas de valor inicial, y tras ello se realiza una **discretización espacial** aproximando el operador diferencial (en derivadas espaciales) $L(u)$ dado por (4.49) mediante un esquema en diferencias finitas como los esquemas planteados en el apartado anterior (u otros similares).

Por ilustrar lo que se acaba de decir podrían considerarse diferentes instantes de cálculo:

$$0 = t^0 < t^1 < .... < t^n < .... < t^N = T$$

que por simplicidad supondremos inicialmente que son equidistantes y designaremos por $\Delta t = t^n - t^{n-1}$ $(n = 1, 2, ..., N)$, y proceder a discretizar (4.48) mediante un θ -esquema como sigue:

$$\frac{\partial u}{\partial t} = L(u) \rightarrow \frac{u^{n+1}(x,y) - u^n(x,y)}{\Delta t} = (1-\theta)L(u^n(x,y)) + \theta L(u^{n+1}(x,y))$$

donde hemos denotado por $u^n(x,y)$ a una aproximación de la función (dependiente sólo de las coordenadas espaciales) $u(x,y,t^n)$. La expresión anterior puede reescribirse como:

$$u^{n+1}(x,y) - \theta\Delta t L(u^{n+1}(x,y)) = u^n + (1-\theta)\Delta t L(u^n(x,y))$$

o, teniendo en cuenta cómo era la expresión del operador $L(u)$:

$$u^{n+1}(x,y) - \theta\Delta t \left(\nabla \bullet \left([\mathbf{D}]\,\nabla u^{n+1}\right) - \nabla \bullet \left(\overrightarrow{\mathbf{V}} u^{n+1}\right) - qu^n\right) =$$

$$= u^n + (1-\theta)\Delta t \left(f(x,y,t^n) + \nabla \bullet ([\mathbf{D}]\,\nabla u^n) - \nabla \bullet \left(\overrightarrow{\mathbf{V}} u^n\right) - qu^n\right)$$

$$+\theta\Delta t f(x,y,t^{n+1})$$

En ella ya sólo intervienen derivadas espaciales que podrán aproximarse como se acaba de detallar para los problemas estacionarios, y que nos conducirán a partir del conocimiento de los valores nodales $u_{i,j}^n$ (aproximación de $u(x_i, y_j, t^n)$) a los valores de $u_{i,j}^{n+1}$ (aproximación de $u(x_i, y_j, t^{n+1})$). La idea, por tanto, puede resumirse en estimar los valores $u_{i,j}^0$ que aproximen en los nodos a $u^0(x_i, y_j)$ y a partir de ellos, mediante la estrategia anterior determinar $u_{i,j}^1$, y de estos $u_{i,j}^2$, etc... hasta obtener $u_{i,j}^N$.

Siendo esta la sencilla idea en la que se sustentan los esquemas numéricos en diferencias finitas para la resolución aproximada de problemas evolutivos como el antes planteado, aparecen numerosas cuestiones que será necesario analizar con detalle tales como ¿qué esquemas conviene emplear tanto para la discretización temporal como para la espacial?, ¿qué relaciones debe haber entre el tamaño de paso temporal y los tamaños de discretización espacial para garantizar una buena convergencia?, ¿qué esfuerzo computacional conllevan las diferentes estrategias que puedan emplearse?, etc... A estas cuestiones y a presentar algunos de los esquemas en diferencias finitas más clásicos, dedicaremos los siguientes apartados. Por simplicidad y por cuestiones de tiempo disponible, realizaremos este estudio sobre casos en los que el dominio espacial sea unidimensional. En la bibliografía podrás encontrar referencias en las que dichos esquemas se extienden a 2 y 3 dimensiones espaciales.

Además, por su distinta naturaleza, trataremos separadamente los términos difusivos (propios de los problemas denominados parabólicos) de los términos convectivos (propios de los problemas denominados hiperbólicos de primer orden). El tratamiento de la ecuación de transporte completa incluyendo ambos términos podrá realizarse a partir de cuanto digamos para cada uno de ellos.

4.4. Esquemas centrados para la ecuación de difusión evolutiva en una dimensión espacial.

4.4.1. Algunos esquemas explícitos e implícitos.

Consideremos el problema:

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) = D\dfrac{\partial^2 u}{\partial x^2}(x,t) & 0 < x < L \quad t > 0 \\ u(0,t) = u_I(t) \;\; t > 0, & u(L,t) = u_D(t) \quad t > 0 \\ u(x,0) = u^0(x) & 0 \le x \le L \end{cases}$$

donde supondremos que D es una constante, que $u(x,t)$ es suficientemente regular y que, para evitar problemas de discontinuidades, $u_I(0) = u^0(0)$ y $u_D(0) = u^0(L)$.

Consideremos además un conjunto de valores de t de la forma

$$t^n = n\Delta t \qquad (n = 0, 1, 2, ...)$$

y dividamos el intervalo $[0, L]$ en N subintervalos de la forma $[x_i, x_{i+1}]$ $(i = 1, 2, ..., N)$ donde:

$$\Delta x = \frac{L}{N} \quad \text{y} \quad x_i = (i-1)\Delta x \quad (i = 1, 2, ..., N)$$

Denotaremos por $U_i^n = u(x_i, t^n)$ y por u_i^n a una aproximación del valor $u(x_i, t^n)$ obtenida mediante alguno de los esquemas en diferencias finitas que a continuación plantearemos. Obsérvese que con esta notación los valores $u_i^0 \approx u(x_i, 0) = u^0(x_i)$ serán conocidos a través de la condición inicial del problema formulado anteriormente. Asimismo serán conocidos, a través en este caso de las condiciones de contorno, los valores $u_1^n \approx u_I(t^n)$ y $u_{N+1}^n \approx u_D(t^n)$.

Planteemos un primer esquema explícito en diferencias finitas para obtener las aproximaciones u_i^n de la solución de nuestro problema. Para ello, de forma similar a como hacíamos en el capítulo anterior con el método de Euler, puede aproximarse en el punto x_i y en un instante $t^* \in \left[t^n, t^{n+1}\right]$ la derivada temporal mediante:

$$\frac{\partial u}{\partial t}(x_i, t^*) \approx \frac{u_i^{n+1} - u_i^n}{\Delta t}$$

Con ello nuestra EDP puede aproximarse (semidiscretizada en tiempo) en el punto x_i y en un instante de tiempo $t^* \in \left[t^n, t^{n+1}\right]$ por:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} \approx D\frac{\partial^2 u}{\partial x^2}(x_i, t^*)$$

Según se escoja el instante t^* obtendremos un esquema explícito (si $t^* = t^n$, pues en ese caso la solución u_i^{n+1} se hará depender de la solución, previamente calculada, en el instante t^n) o implícito (con cualquier otra elección de t^*). Comencemos considerando por simplicidad el caso explícito y ya nos ocuparemos de los esquemas implícitos un poco más adelante. En esta situación la ecuación anterior se reescribe como:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} \approx D\frac{\partial^2 u}{\partial x^2}(x_i, t^n)$$

y puede aproximarse la derivada espacial segunda mediante un esquema centrado resultando en este caso que:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = D\frac{u_{i-1}^n - 2u_i^n + u_{i+1}^n}{(\Delta x)^2} \Rightarrow$$

$$\Rightarrow u_i^{n+1} = D\alpha u_{i-1}^n + (1 - 2D\alpha)u_i^n + D\alpha u_{i+1}^n \tag{4.50}$$

donde se ha llamado $\alpha = \dfrac{\Delta t}{(\Delta x)^2}$. Habida cuenta de que las aproximaciones iniciales se conocen y que las soluciones en el contorno también son conocidas, puede plantearse el siguiente esquema de cálculo explícito:

COMIENZO DEL ALGORITMO

Definir el mallado concretando el valor de Δt *y de* Δx *y de* $\{x_i\}_{i=1}^{N+1}$. *Definir el valor de* D.

$\alpha \leftarrow \Delta t/(\Delta x)^2$

Evaluar $u_i^0 \leftarrow u^0(x_i) \quad (i = 1, ..., N+1)$

Para $n = 0, 1, 2, ...$ *Conocidos los valores* $\{u_i^n\}_{i=1}^{N+1}$

Para $i = 2$ *hasta* $i = N$ *con paso* 1 *hacer:*

$u_j^{n+1} \leftarrow u_i^n + D\alpha(u_{i-1}^n - 2u_i^n + u_{i+1}^n)$

Fin bucle en i.

$u_1^{n+1} \leftarrow u_I(t^{n+1})$, $u_{N+1}^{n+1} \leftarrow u_D(t^{n+1})$

Fin bucle en n.

Escritura de resultados

FIN ALGORITMO.

Obsérvese que el cálculo del vector $\{\mathbf{u}^{n+1}\}$ conteniendo los valores nodales aproximados en el instante t^{n+1} puede expresarse en la forma:

$$\{\mathbf{u}^{n+1}\} = [\mathbf{A}].\{\mathbf{u}^n\} + \{\mathbf{b}^n\}$$

donde $[\mathbf{A}]$ es una matriz de la forma:

$$[\mathbf{A}] = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ D\alpha & (-2D\alpha) & D\alpha & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & D\alpha & (-2D\alpha) & D\alpha & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & D\alpha & (-2D\alpha) & \dots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & D\alpha & (-2D\alpha) & D\alpha \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \end{bmatrix}$$

y

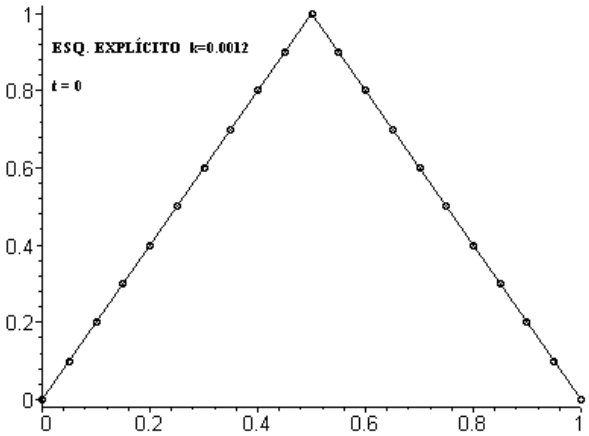
$$\{\mathbf{b}^n\} = \left\{u_I(t^{n+1}),\ \ 0\ \ ,0,\ \ ....,\ 0,\ u_D(t^{n+1})\right\}^T$$

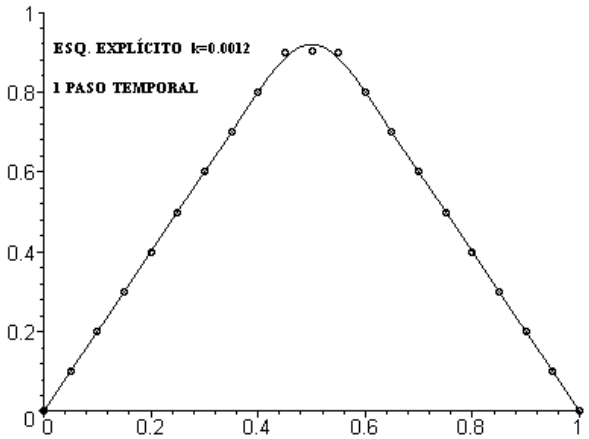
$$\{\mathbf{u}^n\} = \left\{u_1^n, u_2^n, u_3^n, ..., u_N^n,\ u_{N+1}^n\right\}^T$$

Apliquemos este esquema a un ejemplo concreto tomado de Morton y Mayers[13]. Sea el problema:

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) & = \dfrac{\partial^2 u}{\partial x^2}(x,t) \quad 0 < x < 1 \quad t > 0 \\ u(0,t) & = 0 \ \ t > 0, \quad u(L,t) \quad = 0 \quad t > 0 \\ u(x,0) & = u^0(x) \qquad 0 \le x \le 1 \end{cases}$$

[13] K.W. Morton y D.F. Mayers. (1994) Numerical solution of Partial Differential Equations. Ed. Cambridge University Press.





donde

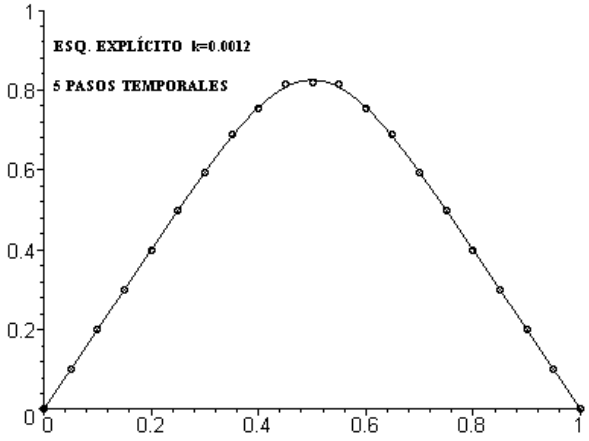
$$u^0(x) = \left\{ \begin{array}{ll} 2x & si \;\; x \leq 0{,}5 \\ 2(1-x) & si \;\; x > 0{,}5 \end{array} \right\}$$

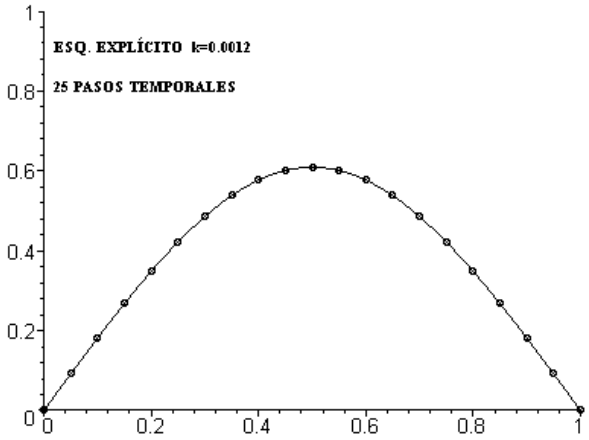
Consideremos una discretización espacial de tamaño de paso 0,05 y apliquemos el esquema anterior con paso temporal $\Delta t = 0{,}0012$. Los resultados que se van obteniendo tras diferentes pasos de tiempo se recogen en las figuras siguientes (donde en trazo continuo se representa la solución exacta y con puntos la solución aproximada obtenida):

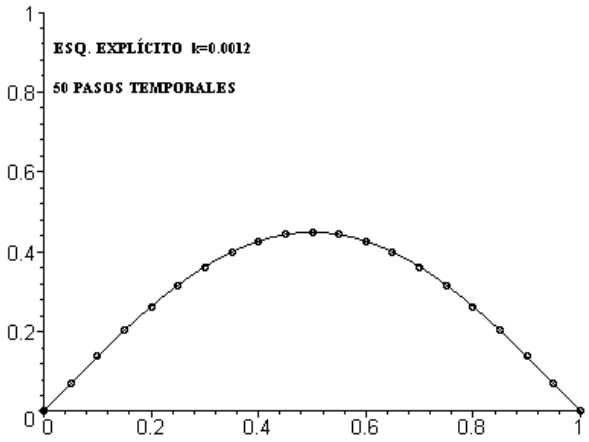
Como vemos se aprecia una buena concordancia entre la solución aproximada y solución analítica. Nótese que en este caso $\alpha = \dfrac{\Delta t}{(\Delta x)^2} = 0{,}48$.

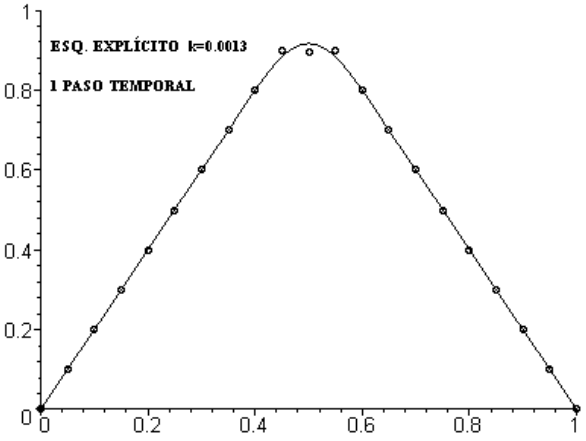
Incrementemos el paso temporal a $\Delta t = 0{,}0013$ (con lo que $\alpha = \dfrac{\Delta t}{(\Delta x)^2} = 0{,}52$) y representemos las soluciones (véanse las figuras de la página siguiente).

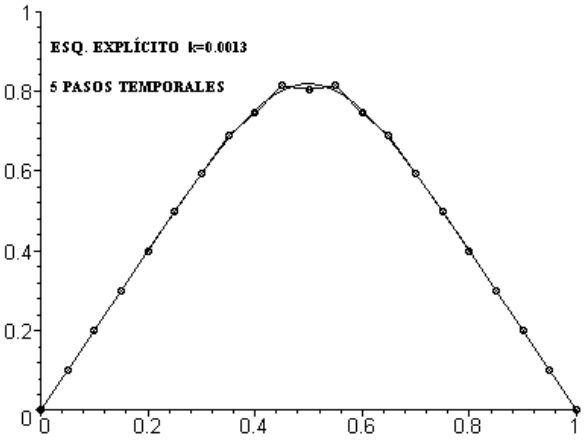
Bastó un pequeño incremento en el paso de tiempo para que el esquema

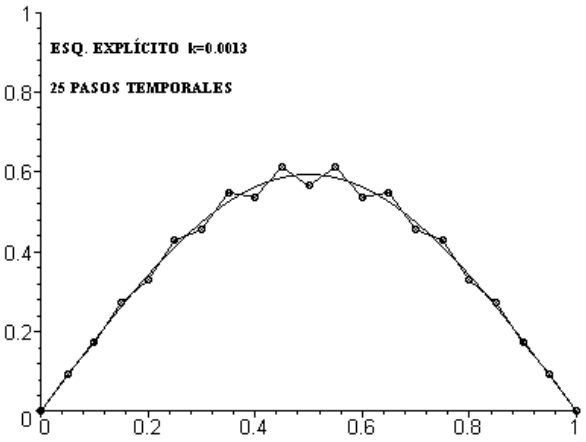

ESQ. EXPLÍCITO k=0.0012
5 PASOS TEMPORALES
0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1

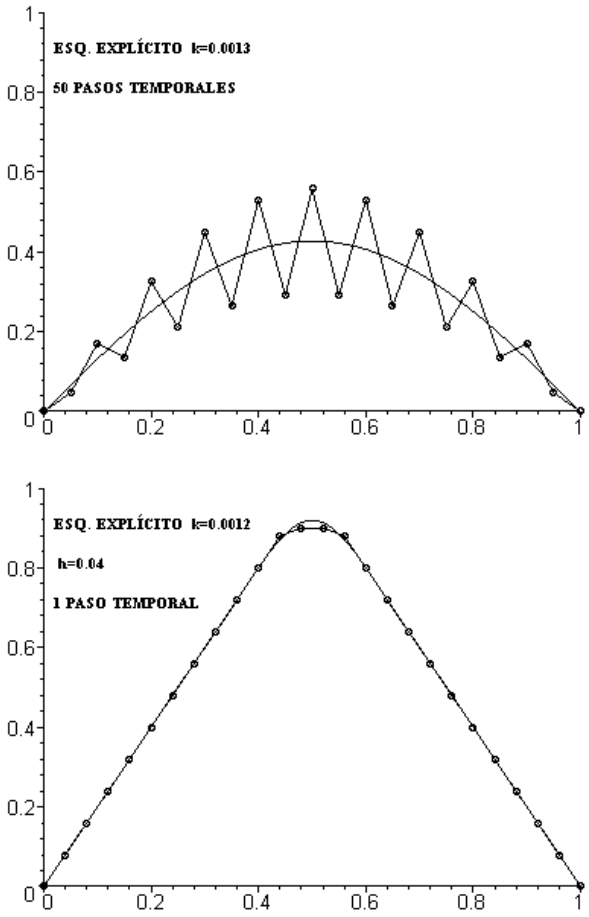

ESQ. EXPLÍCITO k=0.0012
25 PASOS TEMPORALES
0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1


ESQ. EXPLÍCITO k=0.0012
50 PASOS TEMPORALES
0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1


ESQ. EXPLÍCITO k=0.0013
1 PASO TEMPORAL
0
0.2
0.4
0.6
0.8
1


ESQ. EXPLÍCITO k=0.0013
5 PASOS TEMPORALES
0
0.2
0.4
0.6
0.8
1

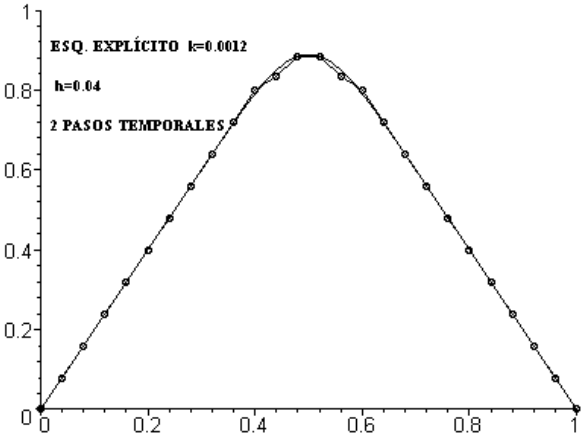

ESQ. EXPLÍCITO k=0.0013
25 PASOS TEMPORALES
0
0.2
0.4
0.6
0.8
1

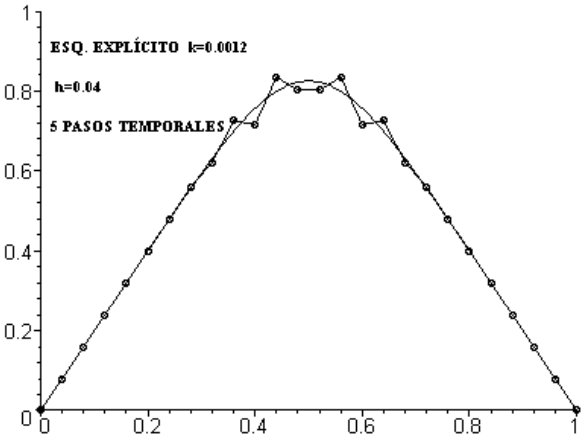


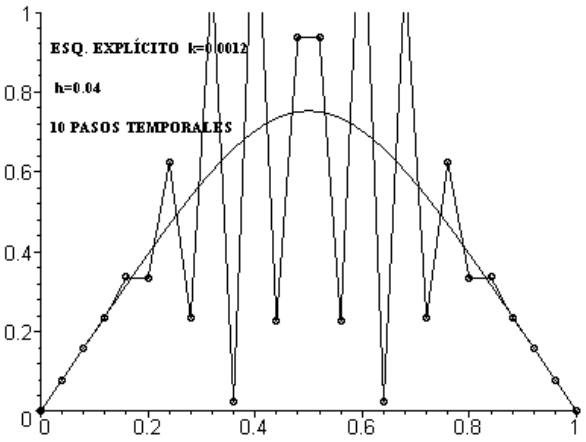
se volviese inestable. Puede pensarse (erróneamente) que el motivo de ello se debe simplemente a estar utilizando una discretización más grosera que en el caso anterior.

En efecto, repitamos el proceso manteniendo el paso de discretización temporal del primer caso, es decir trabajando con $\Delta t = 0{,}0012$ y con una discretización más fina espacialmente, con $\Delta x = 0{,}04$. A priori podría pensarse que los resultados obtenidos deberían ser tan "buenos" al menos como los obtenidos en el primer caso. Y sin embargo las gráficas que se obtienen son las siguientes:

Todavía peor que en el segundo caso (observa que las soluciones mostradas han sido calculadas en un número de pasos temporales menor que en el segundo caso). Y esto parece un contrasentido: ¡discretizamos más fino y el esquema funciona peor!. Pero tiene su explicación como más adelante demostraremos. De hecho, como luego justificaremos, la estabilidad de este esquema no está gobernada por el tamaño de paso de discretización sino por la relación


ESQ. EXPLÍCITO k=0.0012
h=0.04
2 PASOS TEMPORALES
0
0.2
0.4
0.6
0.8
1


ESQ. EXPLÍCITO k=0.0012
h=0.04
5 PASOS TEMPORALES
0
0.2
0.4
0.6
0.8
1


ESQ. EXPLÍCITO k=0.0012
h=0.04
10 PASOS TEMPORALES
0
0.2
0.4
0.6
0.8
1

$\alpha = \Delta t/(\Delta x)^2$ de tal forma que para $\alpha > 1/2$ el esquema se hace inestable. Y en este tercer caso $\alpha = \dfrac{\Delta t}{(\Delta x)^2} = 0{,}75$. En otros términos, el **refinar la discretización disminuyendo Δt e Δx mejora la solución siempre y cuando se verifique que** $\alpha = \Delta t/(\Delta x)^2 \leq 0{,}5$.

Una alternativa a lo anterior consiste en buscar un método de discretización temporal un poco más sofisiticado que el método de Euler explícito. Por ejemplo, podría pensarse en un θ-esquema. Ello nos conduciría a la siguiente ecuación discretizada temporalmente:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} \approx \theta D \frac{\partial^2 u}{\partial x^2}(x_i, t^{n+1}) + (1-\theta)\frac{\partial^2 u}{\partial x^2}(x_i, t^n)$$

y aproximando la derivada espacial segunda mediante un esquema centrado resulta en este caso que:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = (1-\theta)D\frac{u_{i-1}^n - 2u_i^n + u_{i+1}^n}{(\Delta x)^2}$$

$$+\theta D\frac{u_{i-1}^{n+1} - 2u_i^{n+1} + u_{i+1}^{n+1}}{(\Delta x)^2} \Rightarrow$$

$$\Rightarrow -\theta D\alpha u_{i-1}^{n+1} + (1 + 2\theta D\alpha)u_i^{n+1} - \theta D\alpha u_{i+1}^{n+1} =$$

$$(1-\theta)D\alpha u_{i-1}^n + (1 - 2(1-\theta)D\alpha)u_i^n + (1-\theta)D\alpha u_{i+1}^n,$$

donde nuevamente se ha designado por $\alpha = \Delta t/(\Delta x^2)$. Obsérvese que en este caso la expresión anterior por sí sola no nos permite obtener una aproximación del valor de u_i^{n+1} a partir de los valores de u_j^n $(j = 1, ..., N+1)$. Será el conjunto de todas las ecuaciones que se puedan plantear sobre los distintos nodos del mallado las que formen un sistema de ecuaciones que, una vez resuelto, nos proporcione los valores nodales de la solución en el instante t^{n+1}. Más concretamente, reescribamos la ecuación en la forma:

$$a_{i,i-1}u_{i-1}^{n+1} + a_{i,i}u_i^{n+1} + a_{i,i+1}u_{i+1}^{n+1}$$

$$= b_{i,i-1}u_{i-1}^n + b_{i,i}u_i^n + b_{i,i+1}u_{i+1}^n$$

donde

$$a_{i,i-1} = a_{i,i+1} = -\theta D\alpha, \quad a_{i,i} = (1 + 2\theta D\alpha)$$

$$b_{i,i-1} = b_{i,i+1} = (1-\theta)D\alpha, \quad b_{i,i} = (1 - 2(1-\theta)D\alpha)$$

Con ello el conjunto de ecuaciones que pueden plantearse (incluyendo como primera ecuación la que da valor a u_1^{n+1} mediante la condición de contorno en

$x = 0$ y como última la que da valor a u_{N+1}^{n+1} a través de la condición de contorno en $x = L$) forma el sistema:

$$[\mathbf{A}]\{\mathbf{u}^{n+1}\} = [\mathbf{B}]\{\mathbf{u}^n\} + \{\mathbf{c}^n\}$$

donde $[\mathbf{A}]$ es una matriz de la forma:

$$[\mathbf{A}] = \begin{bmatrix} 1 & 0 & 0 & 0 & ... & 0 & 0 & 0 & 0 \\ a_{2,1} & a_{2,2} & a_{2,3} & 0 & ... & 0 & 0 & 0 & 0 \\ 0 & a_{3,2} & a_{3,3} & a_{3,4} & ... & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{4,3} & a_{4,4} & ... & 0 & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 0 & a_{N,N-1} & a_{N,N} & a_{N,N+1} \\ 0 & 0 & 0 & 0 & ... & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$[\mathbf{B}] = \begin{bmatrix} 0 & 0 & 0 & 0 & ... & 0 & 0 & 0 & 0 \\ b_{2,1} & b_{2,2} & b_{2,3} & 0 & ... & 0 & 0 & 0 & 0 \\ 0 & b_{3,2} & b_{3,3} & b_{3,4} & ... & 0 & 0 & 0 & 0 \\ 0 & 0 & b_{4,3} & b_{4,4} & ... & 0 & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 0 & b_{N,N-1} & b_{N,N} & b_{N,N+1} \\ 0 & 0 & 0 & 0 & ... & 0 & 0 & 0 & 0 \end{bmatrix}$$

y

$$\{\mathbf{c}^n\} = \{u_I(t^{n+1}),\ \ 0\ \ , 0,\ \ ....., \ 0,\ u_D(t^{n+1})\}^T$$

$$\{\mathbf{u}^n\} = \left\{u_1^n, u_2^n, u_3^n, ..., u_N^n,\ u_{N+1}^n\right\}^T$$

La resolución del sistema anterior nos proporcionará los valores nodales de la solución en el instante t^{n+1}.

Obsévese que una ligera manipulación del sistema lineal anteriormente escrito nos conduce a que:

$$[\mathbf{A}]\{\mathbf{u}^{n+1}\} = [\mathbf{B}]\{\mathbf{u}^n\} + \{\mathbf{c}^n\} \Rightarrow \{\mathbf{u}^{n+1}\} = [\mathbf{A}]^{-1}[\mathbf{B}]\{\mathbf{u}^n\} + [\mathbf{A}]^{-1}\{\mathbf{c}^n\} \Rightarrow$$

$$\Rightarrow \{\mathbf{u}^{n+1}\} = [\mathbf{M}]\{\mathbf{u}^n\} + \{\mathbf{r}^n\}$$

Según lo anterior un algoritmo que nos permite resolver problemas difusivos como el antes planteado mediante este método consistiría en:

COMIENZO DEL ALGORITMO
Definir el mallado concretando el valor de Δt *y de* Δx *y de* $\{x_i\}_{i=1}^{N+1}$
Definir el valor de D *y del parámetro* θ.
$\alpha \leftarrow \Delta t/(\Delta x)^2$
Evaluar $u_i^0 \leftarrow u^0(x_i) \quad (i = 1, ..., N+1)$
Asignar:

$$[\mathbf{A}] \leftarrow [\mathbf{0}], \quad [\mathbf{B}] \leftarrow [\mathbf{0}]$$

Para $i = 2$ *hasta* $i = N$, *con paso* 1 *hacer:*

$$A(i, i-1) \leftarrow -\theta D\alpha,\ A(i,i) \leftarrow (1 + 2\theta D\alpha),\ A(i, i+1) \leftarrow -\theta D\alpha$$

$$B(i, i-1) \leftarrow (1-\theta)D\alpha,\ B(i,i) \leftarrow (1 - 2(1-\theta)D\alpha),\ B(i, i+1) \leftarrow (1-\theta)D\alpha$$

Fin bucle en i.

$A(1,1) \leftarrow 1, \quad A(N+1, N+1) \leftarrow 1$
Para $n = 0, 1, 2, ...$
Conocidos los valores $\{u_i^n\}_{i=1}^{N+1}$
Para $i = 2$ *hasta* $i = N$ *con paso* 1 *hacer:*
$r_i \leftarrow B(i, i-1)u_{i-1}^n + B(i,i)u_i^n + B(i, i+1)u_{i+1}^n$
Fin bucle en i.
$r_1 \leftarrow u_I(t^{n+1})$
$r_{N+1} \leftarrow u_D(t^{n+1})$
Resolver el sistema lineal $[\mathbf{A}]\{\mathbf{u}^{n+1}\} = \{\mathbf{r}\}$
Fin bucle en n.
Escritura de resultados
FIN ALGORITMO.

El comportamiento de estos esquemas dependerá del valor que asignemos al parámetro θ. Te dejamos como ejercicio propuesto el construir un procedimiento en MAPLE (o en cualquier lenguaje de programación que prefieras) y analizar experimentalmente el comportamiento de estos esquemas. Nosotros lo aplicaremos un poco más adelante a la resolución de algún ejemplo. Pero previamente realicemos un análisis de estos métodos para conocer con qué valores de los parámetros nos interesará "jugar".

NOTA:

Obsérvese que el esquema explícito inicialmente planteado es un caso particular de los θ *-esquemas que acabamos de considerar en el que a* θ *se le asigna el valor* $\theta = 0$.

4.4.2. Consistencia de los esquemas.

Obviamente los valores nodales de la solución aproximada que se encuentren en un instante de tiempo satisfacerán la expresión mediante la cual han sido evaluados, es decir, utilizando la notación introducida en el apartado anterior:

$$a_{i,i-1}u_{i-1}^{n+1} + a_{i,i}u_i^{n+1} + a_{i,i+1}u_{i+1}^{n+1} = b_{i,i-1}u_{i-1}^n + b_{i,i}u_i^n + b_{i,i+1}u_{i+1}^n$$

o lo que es lo mismo:

$$a_{i,i-1}u_{i-1}^{n+1} + a_{i,i}u_i^{n+1} + a_{i,i+1}u_{i+1}^{n+1} - \left(b_{i,i-1}u_{i-1}^n + b_{i,i}u_i^n + b_{i,i+1}u_{i+1}^n\right) = 0$$

Pero la solución analítica (exacta) no tiene por qué satisfacer en cada nodo x_i e instante de cálculo t^n la igualdad anterior. En general para la solución analítica se verificará que:

$$E_i^{n+1} = a_{i,i-1}U_{i-1}^{n+1} + a_{i,i}U_i^{n+1} + a_{i,i+1}U_{i+1}^{n+1} - \left(b_{i,i-1}U_{i-1}^n + b_{i,i}U_i^n + b_{i,i+1}U_{i+1}^n\right)$$

donde como ya se señaló anteriormente se ha denotado por $U_i^n = u(x_i, t^n)$.

Ello, de forma análoga a lo que se dijo para los problemas estacionarios, nos permite introducir un primer concepto sobre los errores del método en diferencias:

Definition 5 *Se denomina* ***error de consistencia*** *del esquema en el instante* t^{n+1} *y en el nodo* x_i *al valor* E_i^{n+1} *dado por:*

$$E_i^{n+1} = a_{i,i-1}U_{i-1}^{n+1} + a_{i,i}U_i^{n+1} + a_{i,i+1}U_{i+1}^{n+1} - \left(b_{i,i-1}U_{i-1}^n + b_{i,i}U_i^n + b_{i,i+1}U_{i+1}^n\right).$$

En este sentido, además, se dirá que un método es **consistente** cuando, al hacer tender Δt e Δx hacia cero los errores de consistencia también tiendan hacia 0.

El análisis de la consistencia de un método numérico en diferencias finitas suele realizarse mediante el uso de desarrollos en serie de Taylor, de tal forma que el orden al que aparecen elevados Δt e Δx en el error de truncamiento de los desarrollos nos marca lo que se denomina el **orden de consistencia** (local) del método. Examinemos sobre los θ-métodos antes planteados el orden de consistencia. Para ello, puesto que el esquema se plantea en el intervalo temporal $[t^n, t^{n+1}]$ debe elegirse un isntante en torno al cual desarrollar las funciones en serie de Taylor. El resultado final será el mismo sea cual sea el instante elegido pero la laboriosidad de los cálculos a realizar puede verse determinada por la elección que hagamos. En el caso de los θ-métodos es cómodo tomar como tal instante el instante medio que denotaremos como $t^{n+1/2} = (t^n + t^{n+1})/2$. Con esta elección pueden considerarse los siguientes desarrollos en serie de Taylor:

$$U_i^{n+1} = U_i^{n+1/2} + \frac{1}{2}\Delta t\frac{\partial u}{\partial t}(x_i, t^{n+1/2}) + \frac{1}{8}(\Delta t)^2\frac{\partial^2 u}{\partial t^2}(x_i, t^{n+1/2}) +$$

$$+\frac{1}{48}(\Delta t)^3\frac{\partial^3 u}{\partial t^3}(x_i,t^{n+1/2})+...$$

$$U_i^n=U_i^{n+1/2}-\frac{1}{2}\Delta t\frac{\partial u}{\partial t}(x_i,t^{n+1/2})+\frac{1}{8}(\Delta t)^2\frac{\partial^2 u}{\partial t^2}(x_i,t^{n+1/2})-$$

$$-\frac{1}{48}(\Delta t)^3\frac{\partial^3 u}{\partial t^3}(x_i,t^{n+1/2})+...$$

de donde

$$\frac{U_i^{n+1}-U_i^n}{\Delta t}=\frac{\partial u}{\partial t}(x_i,t^{n+1/2})+\frac{1}{24}(\Delta t)^2\frac{\partial^3 u}{\partial t^3}(x_i,t^{n+1/2})+...$$

Análogamente:

$$\frac{U_i^{n+1}-2U_i^{n+1}+U_{i+1}^{n+1}}{\Delta x^2}=\frac{\partial^2 u}{\partial x^2}(x_i,t^{n+1})+\frac{1}{12}(\Delta x)^2\frac{\partial^4 u}{\partial x^4}(x_i,t^{n+1})+$$

$$+\frac{1}{360}(\Delta x)^4\frac{\partial^6 u}{\partial t^6}(x_i,t^{n+1})+...=$$

$$=\left(\frac{\partial^2 u}{\partial x^2}(x_i,t^{n+1/2})+\frac{\Delta t}{2}\frac{\partial^3 u}{\partial t\partial x^2}(x_i,t^{n+1/2})+\frac{(\Delta t)^2}{8}\frac{\partial^4 u}{\partial t^2\partial x^2}(x_i,t^{n+1/2})+...\right)+$$

$$\frac{(\Delta x)^2}{12}\left(\frac{\partial^4 u}{\partial x^4}(x_i,t^{n+1/2})+\frac{\Delta t}{2}\frac{\partial^5 u}{\partial t\partial x^4}(x_i,t^{n+1/2})+\frac{(\Delta t)^2}{8}\frac{\partial^6 u}{\partial t^2\partial x^4}(x_i,t^{n+1/2})+...\right)+$$

$$\frac{(\Delta x)^4}{360}\left(\frac{\partial^6 u}{\partial x^6}(x_i,t^{n+1/2})+\frac{\Delta t}{2}\frac{\partial^7 u}{\partial t\partial x^6}(x_i,t^{n+1/2})+\frac{(\Delta t)^2}{8}\frac{\partial^8 u}{\partial t^2\partial x^6}(x_i,t^{n+1/2})+...\right)+...$$

$$=\left[\frac{\partial^2 u}{\partial x^2}(x_i,t^{n+1/2})+\frac{(\Delta x)^2}{12}\frac{\partial^4 u}{\partial x^4}(x_i,t^{n+1/2})+\frac{(\Delta x)^4}{360}\frac{\partial^6 u}{\partial x^6}(x_i,t^{n+1/2})+...\right]+$$

$$\frac{\Delta t}{2}\left[\frac{\partial^3 u}{\partial t\partial x^2}(x_i,t^{n+1/2})+\frac{(\Delta x)^2}{12}\frac{\partial^5 u}{\partial t\partial x^4}(x_i,t^{n+1/2})+\frac{(\Delta x)^4}{360}\frac{\partial^7 u}{\partial t\partial x^6}(x_i,t^{n+1/2})+...\right]+$$

$$+\frac{(\Delta t)^2}{8}\left[\frac{\partial^4 u}{\partial t^2\partial x^2}(x_i,t^{n+1/2})+\frac{(\Delta x)^2}{12}\frac{\partial^6 u}{\partial t^2\partial x^4}(x_i,t^{n+1/2})\right]+$$

$$\frac{(\Delta t)^2}{8}\left[\frac{(\Delta x)^4}{360}\frac{\partial^8 u}{\partial t^2 \partial x^6}(x_i, t^{n+1/2}) + ...\right] + ...$$

Y del mismo modo:

$$\frac{U_i^n - 2U_i^n + U_{i+1}^n}{\Delta x^2} =$$

$$\left[\frac{\partial^2 u}{\partial x^2}(x_i, t^{n+1/2}) + \frac{(\Delta x)^2}{12}\frac{\partial^4 u}{\partial x^4}(x_i, t^{n+1/2}) + \frac{(\Delta x)^4}{360}\frac{\partial^6 u}{\partial x^6}(x_i, t^{n+1/2}) + ...\right]$$

$$-\frac{\Delta t}{2}\left[\frac{\partial^3 u}{\partial t \partial x^2}(x_i, t^{n+1/2}) + \frac{(\Delta x)^2}{12}\frac{\partial^5 u}{\partial t \partial x^4}(x_i, t^{n+1/2}) + \frac{(\Delta x)^4}{360}\frac{\partial^7 u}{\partial t \partial x^6}(x_i, t^{n+1/2}) + ...\right] +$$

$$\frac{(\Delta t)^2}{8}\left[\frac{\partial^4 u}{\partial t^2 \partial x^2}(x_i, t^{n+\frac{1}{2}}) + \frac{(\Delta x)^2}{12}\frac{\partial^6 u}{\partial t^2 \partial x^4}(x_i, t^{n+\frac{1}{2}}) + \frac{(\Delta x)^4}{360}\frac{\partial^8 u}{\partial t^2 \partial x^6}(x_i, t^{n+\frac{1}{2}}) + ...\right.$$

Por tanto,

$$(1-\theta)\frac{U_i^n - 2U_i^n + U_{i+1}^n}{\Delta x^2} + \theta\frac{U_i^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}}{\Delta x^2} =$$

$$= \left[\frac{\partial^2 u}{\partial x^2}(x_i, t^{n+1/2}) + \frac{(\Delta x)^2}{12}\frac{\partial^4 u}{\partial x^4}(x_i, t^{n+1/2}) + \frac{(\Delta x)^4}{360}\frac{\partial^6 u}{\partial x^6}(x_i, t^{n+1/2}) + ...\right] +$$

$$+\left(\theta - \frac{1}{2}\right)\Delta t\left[\frac{\partial^3 u}{\partial t \partial x^2}(x_i, t^{n+1/2}) + \frac{(\Delta x)^2}{12}\frac{\partial^5 u}{\partial t \partial x^4}(x_i, t^{n+1/2}) + ...\right] +$$

$$+\frac{(\Delta t)^2}{2}\left[\frac{\partial^4 u}{\partial t^2 \partial x^2}(x_i, t^{n+1/2}) + ...\right]$$

Finalmente, aplicando el θ-esquema a la solución analítica se puede concluir que:

$$E_i^{n+1} = \frac{U_i^{n+1} - U_i^n}{\Delta t} - D\left[(1-\theta)\frac{U_i^n - 2U_i^n + U_{i+1}^n}{\Delta x^2} + \theta\frac{U_i^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}}{\Delta x^2}\right] =$$

$$= \left(\frac{\partial u}{\partial t}(x_i, t^{n+1/2}) - D\frac{\partial^2 u}{\partial x^2}(x_i, t^{n+1/2})\right) +$$

$$\left(\left(\frac{1}{2} - \theta\right)\Delta t\frac{\partial^3 u}{\partial t \partial x^2}(x_i, t^{n+1/2}) - D\frac{(\Delta x)^2}{12}\frac{\partial^4 u}{\partial x^4}(x_i, t^{n+1/2})\right) +$$

$$+\left(\frac{(\Delta t)^2}{12}\frac{\partial^3 u}{\partial t^3}(x_i,t^{n+1/2})-\frac{D}{8}(\Delta t)^2\frac{\partial^4 u}{\partial t^2\partial x^2}(x_i,t^{n+1/2})\right)+$$

$$+\left(\frac{\Delta t(\Delta x)^2}{12}D\left(\frac{1}{2}-\theta\right)\frac{\partial^5 u}{\partial t\partial x^4}(x_i,t^{n+\frac{1}{2}})-\frac{D}{360}(\Delta x)^4\frac{\partial^6 u}{\partial x^6}(x_i,t^{n+\frac{1}{2}})\right)+...$$

De los términos de este error de consistencia local, el primero de ellos es la propia ecuación diferencial planteada en el punto $(x_i,t^{n+1/2})$. Y como $u(x,t)$ es la solución analítica de ella este primer sumando se anulará. Es decir, que el error de consistencia local del método será:

Si $\theta\neq 1/2:\quad E_i^{n+1}=O(\Delta t,\Delta x^2)$.

Si $\theta=1/2:\quad E_i^{n+1}=O(\Delta t^2,\Delta x^2)$.

Este segundo caso en el que $\theta=1/2$, que presenta un mayor orden del error de consistencia local en cuanto al paso temporal se conoce, por analogía con el correspondiente método para la resolución de problemas de valor inicial, con el nombre de **método de Crank-Nicolson**.

Obsérvese también que, en el caso de que $\theta\neq 1/2$, el término principal del error de consistencia está dado por:

$$T_i^{n+1}=\left(\frac{1}{2}-\theta\right)\Delta t\frac{\partial^3 u}{\partial t\partial x^2}(x_i,t^{n+1/2})-D\frac{(\Delta x)^2}{12}\frac{\partial^4 u}{\partial x^4}(x_i,t^{n+1/2})$$

de donde podremos escribir que:

$$\frac{1}{\left(\frac{1}{2}-\theta\right)\Delta t}T_i^{n+1}=\frac{\partial^3 u}{\partial t\partial x^2}(x_i,t^{n+1/2})-\frac{1}{\left(\frac{1}{2}-\theta\right)12}\frac{(\Delta x)^2}{\Delta t}D\frac{\partial^4 u}{\partial x^4}(x_i,t^{n+1/2})$$

Pero, suponiendo que $u(x,t)$ es suficientemente regular, de la propia EDP puede inferirse, derivándola una vez respecto al tiempo, que:

$$\frac{\partial u}{\partial t}(x,t)-D\frac{\partial^2 u}{\partial x^2}(x,t)=0\Rightarrow\frac{\partial^2 u}{\partial t^2}(x,t)-D\frac{\partial^3 u}{\partial t\partial x^2}(x,t)=0$$

por lo que si

$$\frac{1}{\left(\frac{1}{2}-\theta\right)12}\frac{(\Delta x)^2}{\Delta t}=1$$

se tendrá que $T_i^{n+1}=0$. La condición anterior puede reescribirse como

$$\alpha=\frac{\Delta t}{(\Delta x)^2}=\frac{1}{6\,(1-2\theta)}$$

y los razonamientos anteriores muestran que trabajar con esta relación entre los pasos de discretización espacial y temporal aumenta el orden de consistencia del esquema.

NOTA:

Obsérvese que si $\theta \geq 1/2$ la relación anterior no puede satisfacerse pues tanto Δt como Δx deben ser estrictamente positivos.

Los (un tanto tediosos) desarrollos anteriores ponen de manifiesto que los esquemas planteados son consistentes. Pero no todo es la consistencia ya que, en cada paso temporal, no partiremos de la solución analítica en el instante t^n sino de otra previamente aproximada. En este sentido debe prestarse atención, al igual que con los métodos para problemas de valor inicial, a otro concepto importante: la estabilidad. De él pasamos a ocuparnos a continuación.

4.4.3. Análisis de la estabilidad de los esquemas.

El concepto de estabilidad de un esquema numérico nos permite analizar cómo, pequeños errores iniciales, se transmiten en las sucesivas etapas del cálculo de la solución aproximada. En este sentido, en numerosas ocasiones los valores u_i^0 $(i = 1, ..., N+1)$ con los que se inicializa el proceso no serán exactamente $u^0(x_i)$ pues los errores de redondeo, inherentes al hecho de no poder manipular infinitos decimales cuando se trabaja en aritmética decimal finita, serán responsables de pequeños errores iniciales. Incluso cuando esto no sea así, en la primera etapa de cálculo ya se introducirá una diferencia entre los valores exactos de la solución y los valores numéricos obtenidos (debido tanto a la aproximación de los operadores diferenciales por fórmulas en diferencias finitas como, nuevamente, a los errores de redondeo provocados por la representación con un número de decimales finitos de los resultados de nuestras operaciones aritméticas).

En resumen, el concepto de estabilidad estima cómo serán las diferencias entre los valores obtenidos partiendo de los valores iniciales $\{u_1^0, u_2^0, ..., u_{N+1}^0\}^T = \{\mathbf{u}^0\}$ y los que se obtendrían si se hubiera partido del vector $\{U_1^0, U_2^0, ..., U_{N+1}^0\}^T = \{\mathbf{U}^0\}$ suponiendo que entre ellos hay diferencias dadas por el vector $\{\boldsymbol{\varepsilon}^0\} = \{\mathbf{u}^0\} - \{\mathbf{U}^0\}$. Si este vector fuese nulo, todo cuanto se diga podría trasladarse a la etapa dada por t^1.

Más concretamente: Siendo $\{u_1^n, u_2^n, ..., u_{N+1}^n\}^T = \{\mathbf{u}^n\}$ el vector de valores nodales obtenidos para el tiempo t^n con un esquema numérico inicializado con el vector $\{\mathbf{u}^0\}$ y siendo $\{U_1^n, U_2^n, ..., U_{N+1}^n\}^T = \{\mathbf{U}^n\}$ el vector de valores nodales obtenidos para el tiempo t^n con un esquema numérico inicializado con otro vector cualquiera $\{\mathbf{U}^0\}$ (que estará destinado a jugar el papel de los valores nodales de la solución exacta, aunque cuanto se diga debe ser válido

para cualquier elección de este vector) denotaremos por

$$\{\boldsymbol{\varepsilon}^n\} = \{\mathbf{u}^n\} - \{\mathbf{U}^n\} = \{u_1^n - U_1^n, u_2^n - U_2^n, ..., u_{N+1}^n - U_{N+1}^n\}^T$$

y diremos que el método numérico es estable cuando:

$$\exists M \; / \quad \|\{\boldsymbol{\varepsilon}^n\}\| \leq M \quad \forall n$$

En otros términos, cuando las diferencias entre los valores nodales de una y otra sucesión permanecen acotadas para cualquier etapa de cálculo por la constante M.

Existen diferentes formas (equivalentes entre sí) de analizar si un esquema numérico es estable o no. Entre ellas presentaremos aquí la que se fundamenta en el análisis espectral de las matrices que aparecen en la formulación de los esquemas numéricos.

NOTA:

Cuando tratemos los problemas convectivos, introduciremos otra técnica de análisis de la estabilidad de los esquemas conocida con el nombre de método de von Neumann. También esta segunda técnica podría aplicarse al análisis de los esquemas hasta ahora considerados. En Morton y Mayers[14] *puedes consultar cómo se aplica a los θ-métodos.*

La idea básica del estudio de la estabilidad mediante análisis espectral se basa en formular los esquemas en la forma:

$$\left\{\mathbf{u}^{n+1}\right\} = [\mathbf{M}]\{\mathbf{u}^n\} + \{\mathbf{r}^n\}$$

En los apartados anteriores puedes consultar cómo son la matriz $[\mathbf{M}]$ y el vector $\{\mathbf{r}^n\}$ para los θ -esquemas planteados. La consideración de forma recursiva del esquema así formulado nos conduce a que:

$$\{\mathbf{u}^n\} = [\mathbf{M}]\{\mathbf{u}^{n-1}\} + \{\mathbf{r}^{n-1}\} = [\mathbf{M}]\left([\mathbf{M}]\{\mathbf{u}^{n-2}\} + \{\mathbf{r}^{n-2}\}\right) + \{\mathbf{r}^n\} =$$

$$= [\mathbf{M}]^2\{\mathbf{u}^{n-2}\} + [\mathbf{M}]\{\mathbf{r}^{n-2}\} + \{\mathbf{r}^n\} = [\mathbf{M}]^2\left([\mathbf{M}]\{\mathbf{u}^{n-3}\} + \{\mathbf{r}^{n-3}\}\right) +$$

$$[\mathbf{M}]\{\mathbf{r}^{n-2}\} + \{\mathbf{r}^n\} = [\mathbf{M}]^3\{\mathbf{u}^{n-3}\} + [\mathbf{M}]^2\{\mathbf{r}^{n-3}\} + [\mathbf{M}]\{\mathbf{r}^{n-2}\} + \{\mathbf{r}^n\} =$$

$$[\mathbf{M}]^n\{\mathbf{u}^0\} + [\mathbf{M}]^{n-1}\{\mathbf{r}^1\} + ... + [\mathbf{M}]\{\mathbf{r}^{n-2}\} + \{\mathbf{r}^n\}$$

[14] K.W. Morton y D.F. Mayers. (1994) Numerical solutioon of Partial Differential Equations. Ed. Cambridge University Press.

Puesto que ni la matriz $[\mathbf{M}]$ ni los vectores $\{\mathbf{r}^n\}$ dependen de la solución de la que se parta (pues la matriz sólo depende del coeficiente de difusión D, del valor escogido para θ, y de los tamaños de los pasos de discretización Δx y Δt, y los vectores $\{\mathbf{r}^n\}$ dependen de los mismos parámetros más de las condiciones de contorno), el mismo esquema de cálculo partiendo de $\{\mathbf{U}^0\}$ nos conduce a que:

$$\{\mathbf{U}^n\} = [\mathbf{M}]^n\{\mathbf{U}^0\} + [\mathbf{M}]^{n-1}\{\mathbf{r}^1\} + ... + [\mathbf{M}]\{\mathbf{r}^{n-2}\} + \{\mathbf{r}^n\}$$

Por tanto,

$$\{\boldsymbol{\varepsilon}^n\} = \{\mathbf{u}^n\} - \{\mathbf{U}^n\} = [\mathbf{M}]^n\left(\{\mathbf{u}^0\} - \{\mathbf{U}^0\}\right) = [\mathbf{M}]^n\{\boldsymbol{\varepsilon}^0\} \quad (n > 0)$$

La expresión anterior nos determinará el vector $\{\boldsymbol{\varepsilon}^n\}$ en función del valor de n y del vector de diferencias iniciales $\{\boldsymbol{\varepsilon}^0\}$.

Supongamos ahora que la matriz $[\mathbf{M}]$ es diagonalizable y tiene por valores propios a λ_1, λ_2,..., λ_{N+1} siendo una base de $\mathbf{C}^{N+1}$ formada por los vectores propios: $\left\{\mathbf{v}^{(1)}\right\}$, $\left\{\mathbf{v}^{(2)}\right\}$, ..., $\left\{\mathbf{v}^{(N+1)}\right\}$.

NOTA:

En Conde y Schiavi[15] *puedes consultar cómo se determinaban los valores y vectores propios de una matriz, qué propiedades verifican, cuándo esta es diagonalizable y cómo construir una base de* $\mathbf{C}^{N+1}$ *formada por vectores propios de la matriz si esta es diagonalizable.*

En estas condiciones el vector $\{\boldsymbol{\varepsilon}^0\}$ puede expresarse en la base formada por vectores propios como:

$$\{\boldsymbol{\varepsilon}^0\} = c_1\left\{\mathbf{v}^{(1)}\right\} + c_2\left\{\mathbf{v}^{(2)}\right\} + ... + c_{N+1}\left\{\mathbf{v}^{(N+1)}\right\}$$

expresión de la que se obtiene

$$[\mathbf{M}]\{\boldsymbol{\varepsilon}^0\} = c_1[\mathbf{M}]\left\{\mathbf{v}^{(1)}\right\} + c_2[\mathbf{M}]\left\{\mathbf{v}^{(2)}\right\} + ... + c_{N+1}[\mathbf{M}]\left\{\mathbf{v}^{(N+1)}\right\} =$$

$$= c_1\lambda_1\left\{\mathbf{v}^{(1)}\right\} + c_2\lambda_2\left\{\mathbf{v}^{(2)}\right\} + ... + c_{N+1}\lambda_{N+1}\left\{\mathbf{v}^{(N+1)}\right\}$$

$$[\mathbf{M}]^2\{\boldsymbol{\varepsilon}^0\} = c_1\lambda_1[\mathbf{M}]\left\{\mathbf{v}^{(1)}\right\} + c_2\lambda_2[\mathbf{M}]\left\{\mathbf{v}^{(2)}\right\} + ... + c_{N+1}\lambda_{N+1}[\mathbf{M}]\left\{\mathbf{v}^{(N+1)}\right\} =$$

$$= c_1\lambda_1^2\left\{\mathbf{v}^{(1)}\right\} + c_2\lambda_2^2\left\{\mathbf{v}^{(2)}\right\} + ... + c_{N+1}\lambda_{N+1}^2\left\{\mathbf{v}^{(N+1)}\right\}$$

[15] C. Conde y E. Schiavi. (2000). Guiones de la asignatura de Elementos de Matemáticas. Universidad Rey Juan Carlos.

y en general,

$$[\mathbf{M}]^n\{\boldsymbol{\varepsilon}^0\} = c_1\lambda_1^n\left\{\mathbf{v}^{(1)}\right\} + c_2\lambda_2^n\left\{\mathbf{v}^{(2)}\right\} + ... + c_{N+1}\lambda_{N+1}^n\left\{\mathbf{v}^{(N+1)}\right\}$$

Por tanto,

$$\{\boldsymbol{\varepsilon}^n\} = [\mathbf{M}]^n\{\boldsymbol{\varepsilon}^0\} = c_1\lambda_1^n\left\{\mathbf{v}^{(1)}\right\} + c_2\lambda_2^n\left\{\mathbf{v}^{(2)}\right\} + ... + c_{N+1}\lambda_{N+1}^n\left\{\mathbf{v}^{(N+1)}\right\}$$

y, tomando normas,

$$\|\{\boldsymbol{\varepsilon}^n\}\| \leq |c_1|\,|\lambda_1|^n\left\|\left\{\mathbf{v}^{(1)}\right\}\right\| + |c_2|\,|\lambda_2|^n\left\|\left\{\mathbf{v}^{(2)}\right\}\right\| + ... + |c_{N+1}|\,|\lambda_{N+1}|^n\left\|\left\{\mathbf{v}^{(N+1)}\right\}\right\|$$

De la expresión anterior puede concluirse que una condición suficiente para que $\|\{\boldsymbol{\varepsilon}^n\}\|$ permanezca acotado es que:

$$|\lambda_j| \leq 1 \qquad (j = 1, 2, ..., N+1)$$

o, lo que es lo mismo, que el radio espectral de la matriz $[\mathbf{M}]$, que denotaremos por $\rho(\mathbf{M})$, verifique:

$$\rho(\mathbf{M}) \leq 1.$$

NOTA:

Recuérdese que el radio espectral de una matriz es el mayor de los módulos de los valores propios de dicha matriz.

En resumen, el análisis de la estabilidad de los esquemas antes planteados puede realizarse analizando cómo es el radio espectral de la matriz del método $[\mathbf{M}]$.

En el caso que nos ocupa dicha matriz está dada por:

$$[\mathbf{M}] = [\mathbf{A}]^{-1}[\mathbf{B}]$$

donde $[\mathbf{A}]$ y $[\mathbf{B}]$ eran matrices tridiagonales simétricas.

NOTA:

Para el estudio de los valores propios en este caso son de interés las propiedades siguientes (cuya demostración omitiremos para no desviarnos en exceso de nuetro estudio y pueden encontrarse por ejemplo en el libro de Smith[16]):

Proposition 6 *Si una matriz* $[\mathbf{A}]$ *es regular y admite como valor propio al valor* λ *con vector propio asociado* $\{\mathbf{v}\}$*, entonces la matriz* $[\mathbf{A}]^{-1}$ *admite a* $\dfrac{1}{\lambda}$ *como valor propio con el mismo vector* $\{\mathbf{v}\}$ *como vector propio a él asociado.*

[16] G.D. Smith (1.985) Numerical Solution of partial Differential Equations. Finite Difference Methods. (3^a edición, 4^a reimpresión (1996)). Ed. Clarendon Press.

Proposition 7 *Una matriz tridiagonal de dimensiones* (m, m) *de la forma:*

$$\begin{bmatrix} \alpha & \beta & 0 & 0 & ... & 0 & 0 \\ \gamma & \alpha & \beta & 0 & ... & 0 & 0 \\ 0 & \gamma & \alpha & \beta & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & \gamma & \alpha & \beta \\ 0 & 0 & 0 & ... & 0 & \gamma & \alpha \end{bmatrix}$$

admite por valores propios los números:

$$\lambda_j = \alpha + 2\sqrt{\beta\gamma}\cos\left(\frac{j\pi}{m+1}\right) \qquad (j = 1, 2, ..., m).$$

Una vez presentadas las bases del método de análisis espectral para el estudio de la estabilidad de los esquemas numéricos, pasemos a analizar los θ-métodos antes introducidos. Para ello, habida cuenta de que los valores en los extremos del dominio espacial son conocidos, escribiremos el sistema de ecuaciones que proporciona la solución en los nodos interiores a $(0, L)$ en la forma:

$$[\mathbf{A}]\{\mathbf{u}^{n+1}\} = [\mathbf{B}]\{\mathbf{u}^n\} + \{c^n\}$$

donde ahora denotamos por $[\mathbf{A}]$ a una matriz de dimensiones $(N-1, N-1)$ dada por:

$$[\mathbf{A}] = \begin{bmatrix} a_d & a_s & 0 & 0 & ... & 0 & 0 & 0 & 0 \\ a_s & a_d & a_s & 0 & ... & 0 & 0 & 0 & 0 \\ 0 & a_s & a_s & a_s & ... & 0 & 0 & 0 & 0 \\ 0 & 0 & a_s & a_d & ... & 0 & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 0 & a_s & a_d & a_s \\ 0 & 0 & 0 & 0 & ... & 0 & 0 & a_s & a_d \end{bmatrix}$$

con

$$a_s = -\theta D\alpha, \quad a_d = (1 + 2\theta D\alpha)$$

por

$$[\mathbf{B}] = \begin{bmatrix} b_d & b_s & 0 & 0 & ... & 0 & 0 & 0 & 0 \\ b_s & b_d & b_s & 0 & ... & 0 & 0 & 0 & 0 \\ 0 & b_s & b_d & b_s & ... & 0 & 0 & 0 & 0 \\ 0 & 0 & b_s & b_d & ... & 0 & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 0 & b_s & b_d & b_s \\ 0 & 0 & 0 & 0 & ... & 0 & 0 & b_s & b_d \end{bmatrix}$$

con

$$b_s = (1-\theta)D\alpha, \quad b_d = (1-2(1-\theta)D\alpha)$$

y por

$$\{\mathbf{c}^n\} = \left\{ \left(b_s u_{IZQ}(t^n) - a_s u_{IZQ}(t^{n+1})\right), 0, 0, ..., 0, \left(b_s u_{DER}(t^n) - a_s u_{DER}(t^{n+1})\right) \right\}$$

Esta expresión del esquema nos indica que se podrá asegurar que el esquema será estable si el radio espectral de la matriz:

$$[\mathbf{M}] = [\mathbf{A}]^{-1}[\mathbf{B}]$$

es menor o igual que 1. Examinemos cómo es el radio espectral de esta matriz. Para ello observemos, en primer lugar, que tanto la matriz [**A**] como la matriz [**B**] pueden expresarse como sigue:

$$[\mathbf{A}] = [\mathbf{I}] + \tau[\mathbf{T}], \qquad [\mathbf{B}] = [\mathbf{I}] - \sigma[\mathbf{T}]$$

donde $\tau = \theta D\alpha = \theta D\dfrac{\Delta t}{(\Delta x)^2}$, $\sigma = (1-\theta)D\alpha = (1-\theta)D\dfrac{\Delta t}{(\Delta x)^2}$, [**I**] es la matriz identidad de dimensiones $((N-1),(N-1))$ y:

$$[\mathbf{T}] = \begin{bmatrix} 2 & -1 & 0 & 0 & ... & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & ... & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & ... & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & ... & 0 & 0 & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & 0 & ... & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & ... & 0 & 0 & -1 & 2 \end{bmatrix}$$

Por otra parte se tiene que:

Proposition 8 *Si λ es un valor propio de la matriz $[\mathbf{T}]$ y $\{\mathbf{v}\}$ es un vector propio asociado a dicho valor propio, entonces $(1+\mu\lambda)$ es un vector propio de la matriz $[[\mathbf{I}]+\mu[\mathbf{T}]]$ que tiene a $\{\mathbf{v}\}$ como vector propio asociado.*

Demostración:
Basta con tener en cuenta que:

$$[[\mathbf{I}]+\mu[\mathbf{T}]]\{\mathbf{v}\} = [\mathbf{I}]\{\mathbf{v}\}+\mu[\mathbf{T}]\{\mathbf{v}\} = [\mathbf{I}]\{\mathbf{v}\}+\mu\lambda\{\mathbf{v}\} = (1+\mu\lambda)\{\mathbf{v}\}$$

c.q.d.

Puesto que los valores propios de $[\mathbf{T}]$ están dados por:

$$\lambda_j(\mathbf{T}) = 2+2\cos\left(\frac{j\pi}{N}\right) = 2\left(2\cos^2\left(\frac{j\pi}{2N}\right)\right) = 4\cos^2\left(\frac{j\pi}{2N}\right) \qquad (j=1,2,...,N-1)$$

resultará que los de $[\mathbf{A}]$ y los de $[\mathbf{B}]$ responderán a las expresiones:

$$\lambda_j(\mathbf{A}) = 1+(\theta D\alpha)\left(4\cos^2\left(\frac{j\pi}{2N}\right)\right) \qquad (j=1,2,...,N)$$

$$\lambda_j(\mathbf{B}) = 1-((1-\theta)D\alpha)\left(4\cos^2\left(\frac{j\pi}{2N}\right)\right) \qquad (j=1,2,...,N)$$

siendo los autoespacios de $[\mathbf{A}]$ y $[\mathbf{B}]$ coincidentes.

Pero los valores propios que buscamos no son los de las matrices $[\mathbf{A}]$ y $[\mathbf{B}]$ sino los de la matriz $[\mathbf{M}]$. Para determinarlos será necesario tener en cuenta la siguiente:

Proposition 9 *Siendo $[\mathbf{A}]$ y $[\mathbf{B}]$ dos matrices cuadradas de la misma dimensión, que admiten como vector propio al vector $\{\mathbf{v}\}$ estando asociado a los valores propios $\lambda(\mathbf{A})$ y $\lambda(\mathbf{B})$ respectivamente, y siendo $[\mathbf{A}]$ una matriz no singular, se verifica que $(\lambda(\mathbf{B})\ /\ \lambda(\mathbf{A}))$ es un valor propio de la matriz $[\mathbf{M}]=[\mathbf{A}]^{-1}[\mathbf{B}]$ que tiene a $\{\mathbf{v}\}$ como vector propio asociado.*

Demostración:
Basta con observar que en las condiciones de la proposición:

$$[\mathbf{A}]\{\mathbf{v}\} = \lambda(\mathbf{A})\{\mathbf{v}\} \Rightarrow \frac{1}{\lambda(\mathbf{A})}\{\mathbf{v}\} = [\mathbf{A}]^{-1}\{\mathbf{v}\}$$

de donde

$$[\mathbf{A}]^{-1}[\mathbf{B}]\{\mathbf{v}\} = [\mathbf{A}]^{-1}\lambda(\mathbf{B})\{\mathbf{v}\} = \frac{\lambda(\mathbf{B})}{\lambda(\mathbf{A})}\{\mathbf{v}\}.$$

c.q.d.

Según esta proposición se tiene que:

$$\lambda_j(\mathbf{M}) = \frac{1 - 4((1-\theta)D\alpha)\cos^2\left(\frac{j\pi}{2N}\right)}{1 + 4(\theta D\alpha)\cos^2\left(\frac{j\pi}{2N}\right)} \qquad (j = 1, 2, ..., N-1).$$

Para que $|\lambda_j(\mathbf{M})| \leq 1$ se debe verificar que:

$$-1 \leq \frac{1 - 4((1-\theta)D\alpha)\cos^2\left(\frac{j\pi}{2N}\right)}{1 + 4(\theta D\alpha)\cos^2\left(\frac{j\pi}{2N}\right)} \leq 1.$$

La segunda de las desigualdades anteriores nos conduce a que para todo j tal que $1 \leq j \leq (N-1)$:

$$1 - 4((1-\theta)D\alpha)\cos^2\left(\frac{j\pi}{2N}\right) \leq 1 + 4(\theta D\alpha)\cos^2\left(\frac{j\pi}{2N}\right) \Rightarrow$$

$$\Rightarrow -D\alpha\cos^2\left(\frac{j\pi}{2N}\right) \leq 0$$

que, habiendo supuesto $D > 0$ y pasos de discretización positivos siempre se verificará.

De la otra desigualdad se tendrá que para todo j tal que $1 \leq j \leq (N-1)$:

$$-1 - 4(\theta D\alpha)\cos^2\left(\frac{j\pi}{2N}\right) \leq 1 - 4((1-\theta)D\alpha)\cos^2\left(\frac{j\pi}{2N}\right) \Rightarrow$$

$$\Rightarrow -2 \leq 4D\alpha\cos^2\left(\frac{j\pi}{2N}\right)(2\theta - 1) \Rightarrow$$

$$\Rightarrow (1 - 2\theta) \leq \frac{1}{2D\alpha\cos^2\left(\frac{j\pi}{2N}\right)}$$

Esta última desigualdad será verificada siempre para valores de $\theta \geq 1/2$. Para valores de este parámetro inferiores a $1/2$ la desigualdad anterior puede reescribirse en la forma:

$$\alpha \leq \frac{1}{2D(1-2\theta)\cos^2\left(\frac{j\pi}{2N}\right)} \quad (j = 1, 2, ..., N-1).$$

Puesto que para los valores que puede tomar j se verificará que:

$$2D(1-2\theta)\cos^2\left(\frac{j\pi}{2N}\right) < 2D(1-2\theta)$$

la estabilidad del esquema quedará garantizada siempre que:

$$\alpha = \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2D(1-2\theta)}$$

En resumidas cuentas, puede afirmarse que los θ -esquemas considerados son incondicionalmente estables siempre que $\theta \geq 1/2$ o que siendo $\theta < 1/2$ la relación entre los tamaños de paso de la discretización temporal y la discretización espacial verifique que:

$$\frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2D(1-2\theta)}.$$

NOTA:

En la bibliografía que se recoge al final de este tema podrás encontrar referencias en las que se incluyen diferentes propiedades y teoremas que ayudan a acotar el radio espectral de matrices como por ejemplo los tres teoremas de Gerschgorin. Con todo el análisis de métodos genéricos en diferencias finitas mediante esta técnica puede ser relativamente laborioso debido a la necesidad de estudiar el radio espectral de las matrices que en dichos métodos se ven involucradas. Como se dijo anteriormente, al abordar los problemas hiperbólicos de primer orden (el tratamiento del término convectivo) introduciremos otra técnica equivalente para el estudio de la estabilidad de los esquemas que en numerosas ocasiones se muestra más sencilla de aplicar.

4.4.4. La equivalencia entre convergencia y estabilidad más consistencia: el teorema de Lax.

Como se ilustró anteriormente, el que un esquema sea consistente con un problema de contorno nos garantizará que, para pasos de discretización suficientemente pequeños, la solución exacta del problema verificará el esquema numérico con un error tan pequeño como se desee. Pero nos falta por acabar de justificar aún el que, para pasos de discretización suficientemente pequeños los valores:

$$e_i^n = u_i^n - u(x_i, t^n) = u_i^n - U_i^n$$

también puedan hacerse suficientemente pequeños en cualquier paso de tiempo.

En este sentido se introducen las siguientes definiciones:

Definition 6 *Se denomina* ***error de convergencia*** *del equema numérico considerado en el punto x_i y en el instante t^n al valor $e_i = u_i^n - U_i^n$. Y se dice que el esquema numérico es* ***convergente*** *hacia la solución exacta del problema de contorno que se quiera resolver cuando:*

$$\Delta t \to 0, \Delta x \to 0 \text{lím}_{\Delta t \to 0, \Delta x \to 0}(\sup_{1 \leq i \leq N+1, n \geq 0} |e_i^n|) = 0.$$

El estudio de la consistencia de los sistemas planteados junto al análisis de su estabilidad nos deja a las puertas de poder concluir la convergencia de los valores nodales de las soluciones aproximadas obtenidas hacia los valores que en los nodos tome la solución exacta. En efecto, habiendo demostrado que los esquemas son consistentes se tiene asegurado que para valores suficientemente pequeños de los pasos de discretización la solución exacta verificará (con errores tan pequeños como se desee) el esquema de cálculo. Y si además el esquema es estable se tiene asegurado que los errores debidos a no partir de la solución exacta en cada paso de tiempo no se amplifican. Es más, si $\alpha = \dfrac{\Delta t}{(\Delta x)^2} < 1$ los desarrollos antes realizados demuestran que dichos errores tienden a anularse al hacer tender el número de pasos temporales hacia infinito. De una forma intuitiva parece que en esta situación "no se nos pueden escapar" los valores de la solución exacta.

Esta idea intuitiva es correcta en el caso de abordar problemas lineales (es decir cuando el coeficiente D no depende de la propia solución $u(x,t)$) y fue demostrada rigurosamente por P.D. Lax en el año 1953 a través del siguiente teorema:

Theorem 10 *(Teorema de equivalencia de Lax). La condición necesaria y suficiente para que un esquema numérico consistente con un problema evolutivo lineal sea convergente es que sea estable en el sentido anteriormente descrito.*

Una demostración rigurosa de este teorema excede los objetivos del presente curso por lo que remitimos al lector interesado a la bibliografía de este tema (por ejemplo a la obra de Ritchmyer y Morton[17]).

Consecuencia del teorema de Lax es que los θ -esquemas antes planteados, en las condiciones que garantizan su estabilidad, son convergentes.

4.4.5. El principio del máximo.

Otro aspecto que interesará analizar sobre los esquemas de resolución de problemas difusivos es si verifican o no el principio del máximo. En este sentido

[17] R.D. Ritchmyer y K.W. Morton (1967) Difference Methods for Initial Value Problems. 2^a edición. Ed.:Wiley-Interscience. reimpreso en 1994 por Ed. Kreiger.

es sabido que el problema:

$$\left\{\begin{array}{lll} \dfrac{\partial u}{\partial t}(x,t) & = \dfrac{\partial^2 u}{\partial x^2}(x,t) & 0 < x < L, \;\; t > 0 \\ u(0,t) & = u_I(t) & t > 0 \\ u(L,t) & = u_D(t) & t > 0 \\ u(x,0) & = u^0(x) & 0 \leq x \leq L \end{array}\right\}$$

tiene una solución analítica $u(x,t)$ que toma valores comprendidos entre u_m y u_M dados por:

$$u_m = \text{mín}\left\{\inf_{t>0}(u_I(t)), \inf_{t>0}(u_D(t)), \inf_{0\leq x\leq L}(u^0(x))\right\}$$

$$u_m = máx\left\{\sup_{t>0}(u_I(t)), \sup_{t>0}(u_D(t)), \sup_{0\leq x\leq L}(u^0(x))\right\}$$

En este sentido, al igual que se dijo para el caso de problemas estacionarios, sería conveniente que los esquemas numéricos satisfacieran una propiedad similar.

Examinemos bajo qué condiciones se verifica tal propiedad. Para ello recordemos que, en cada nodo x_i y en cada instante t^{n+1} el esquema podía escribirse como:

$$a_{i,i-1}u_{i-1}^{n+1} + a_{i,i}u_i^{n+1} + a_{i,i+1}u_{i+1}^{n+1} = b_{i,i-1}u_{i-1}^n + b_{i,i}u_i^n + b_{i,i+1}u_{i+1}^n$$

donde

$$a_{i,i-1} = a_{i,i+1} = -\theta D\alpha, \quad a_{i,i} = (1 + 2\theta D\alpha)$$

$$b_{i,i-1} = b_{i,i+1} = (1-\theta)D\alpha, \quad b_{i,i} = (1 - 2(1-\theta)D\alpha)$$

De estas expresiones se tiene que:

$$u_i^{n+1} = \frac{\theta D\alpha}{(1+2\theta D\alpha)}(u_{i-1}^{n+1} + u_{i+1}^{n+1}) + \frac{(1-\theta)D\alpha}{(1+2\theta D\alpha)}(u_{i-1}^n + u_{i+1}^n) +$$

$$+\frac{(1-2(1-\theta)D\alpha)}{(1+2\theta D\alpha)}u_i^n = \xi_1 u_{i-1}^{n+1} + \xi_2 u_{i+1}^{n+1} + \xi_3 u_{i-1}^n + \xi_4 u_{i+1}^n + \xi_5 u_i^n$$

donde

$$\xi_1 = \xi_2 = \frac{\theta D\alpha}{(1+2\theta D\alpha)}, \; \xi_3 = \xi_4 = \frac{(1-\theta)D\alpha}{(1+2\theta D\alpha)}, \; \xi_5 = \frac{(1-2(1-\theta)D\alpha)}{(1+2\theta D\alpha)}$$

Fácilmente se verifica que

$$\xi_1 + \xi_2 + \xi_3 + \xi_4 + \xi_5 = 1$$

Por tanto, para que u_i^{n+1} sea una combinación convexa de los valores u_{i+1}^{n+1}, u_{i-1}^{n+1}, u_{i-1}^{n}, u_i^n y u_{i+1}^n bastará con asegurar que todos los coeficiente que los multiplican están comprendidos entre 0 y 1. Del examen de la expresión de los coeficientes se deduce, recordando que $\alpha > 0$ y que $D > 0$, y que $0 \leq \theta \leq 1$, que éstos siempre serán inferiores a 1. Además, por lo mismos motivos ξ_1, ξ_2, ξ_3 y ξ_4 siempre serán no negativos. Para que también lo sea ξ_5 debe verificarse que:

$$0 \leq 1 - 2(1-\theta)D\alpha \Rightarrow \alpha \leq \frac{1}{2D(1-\theta)}$$

En resumen se tiene así demostrado el siguiente teorema:

Theorem 11 *Los θ -esquemas, planteados en el apartado* 4,1. *de este capítulo, aplicados al problema de transporte puramente difusivo con coeficiente de difusión constante verifican el principio del máximo discreto siempre que la relación $\alpha = \Delta t/(\Delta x)^2$ satisfaga la desigualdad:*

$$\alpha \leq \frac{1}{2D(1-\theta)}$$

Obsérvese que, como consecuencia del teorema anterior, cuanto más próximo a 1 sea el valor de θ mayor será el valor que puede tomar α. Asimismo, si $\theta < 1/2$ se tendrá que:

$$\alpha = \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2D(1-\theta)} \leq \frac{1}{2D(1-2\theta)}$$

por lo que la verificación del principio del máximo garantiza también la estabilidad del θ -esquema.

4.4.6. Comentarios finales sobre los esquemas en diferencias finitas para la resolución de problemas difusivos.

1^o) La forma de proceder anterior puede extenderse fácilmente al caso de coeficientes variables, es decir a problemas en los que la EDP sea de la forma:

$$\frac{\partial u}{\partial t}(x,t) = D(x,t)\frac{\partial^2 u}{\partial x^2}(x,t)$$

Bastará para ello con designar por $D_i^n = D(x_i, t^n)$ y sustituir en las ecuaciones el valor de D por el de D_i^{n+1} o el de D_i^n según el instante de cálculo en el que se esté discretizando el operador de segunda derivación espacial. Nótese

no obstante que la EDP anterior no es ya, en general, representativa de un problema de transporte difusivo pues este sería modelizado mediante:

$$\frac{\partial u}{\partial t}(x,t) = \frac{\partial}{\partial x}\left(D(x,t)\frac{\partial u}{\partial x}(x,t)\right)$$

es decir, por

$$\frac{\partial u}{\partial t}(x,t) = D(x,t)\frac{\partial^2 u}{\partial x^2}(x,t) + \left(\frac{\partial D}{\partial x}(x,y)\right)\frac{\partial u}{\partial x}(x,t)$$

apareciendo un término convectivo (en derivada espacial primera) del que nos ocuparemos más adelante.

2^o) También ahora podrían considerarse condiciones de contorno más generales. La forma de tratarlas es totalmente análoga a la descrita para los problemas estacionarios, si bien deberá diferenciarse ahora el instante de cálculo en el que se tratan. Un estudio detallado del tratamiento de condiciones de contorno más generales puede encontrarse, entre otros, en Smith[18] o en Morton y Mayers[19].

3^o) Existen muchos otros esquemas para tratar problemas como el aquí planteado, discretizando el operador de derivación temporal de muy diferentes maneras e implicando en ello tan sólo el instante t^n u otros instantes anteriores. Así por ejemplo en Lapidus y Pinder[20] o en Morton y Mayers[21] podrás encontra esquemas basados en tres niveles de tiempo.

4^o) Los problemas planteados en dominios espaciales bi o tridimensionales admiten un tratamiento que, conceptualmente, es similar al del caso unidimensional. No obstante en esos casos pueden aparecer nuevas dificultades en el tratamiento de las matrices involucradas en la formulación de los distintos esquemas. Una estrategia que se ha mostrado muy efectiva en diferentes problemas formulados en dominios bidimensionales es la recogida en los métodos de direcciones alternadas (ADI) en los que se pasa de t^n a t^{n+1} considerando un (o varios) instante intermedio y discretizando hasta él las derivadas en una dirección espacial de forma implícita (es decir, con valores nodales en dicho instante intermedio) y en la otra dirección espacial de forma explícita (es decir en t^n) para posteriormente pasar del instante intermedio a t^{n+1} cambiando los papeles a las discretizaciones empleadas según cada dirección espacial. Ello

[18] G.D. Smith (1.985) Numerical Solution of partial Differential Equations. Finite Difference Methods. (3^a edición, 4^a reimpresión (1996)). Ed. Clarendon Press.

[19] K.W. Morton y D.F. Mayers. (1994) Numerical solutioon of Partial Differential Equations. Ed. Cambridge University Press.

[20] L. Lapidus y G.F. Pinder.(1982) Numerical solution of partial differential equations in science and engineering. Ed.: John Wiley.

[21] K.W. Morton y D.F. Mayers. (1994) Numerical solutioon of Partial Differential Equations. Ed. Cambridge University Press.

transforma el problema bidimensional en sucesivos problemas unidimensionales, con el consiguiente ahorro en lo que a esfuerzo computacional se refiere, y sin pérdidas significativas de precisión ni reducciones en los límites de estabilidad. En las referencias antes aludidas encontrarás estudios detallados sobre los métodos ADI.

4.5. Esquemas en diferencias finitas para el tratamiento de problemas convectivos

4.5.1. Generalidades.

Considérese la ecuación de advección lineal en una dimensión espacial:

$$\frac{\partial u}{\partial t}(x,t) + V(x,t)\frac{\partial u}{\partial x}(x,t) = 0, \qquad t > 0$$

acompañada de la condición inicial:

$$u(x,0) = u^0(x)$$

Como se detalló en el primer tema de estos apuntes y se ilustra con abundantes ejemplos en el anexo la solución analítica de este problema tiene la propiedad de ser constante sobre las curvas características. La curva característica $x(t)$ que pase por el punto $(\xi, 0)$ a su vez estará dada como solución del problema de valor inicial:

$$\begin{cases} \dfrac{dx}{dt}(t) & = V(x(t),t) \quad t > 0 \\ x(0) & = \xi \end{cases}$$

Las consideraciones anteriores nos permiten ya plantear un primer método numérico para la resolución de los problemas de tipo convectivo: resolver numéricamente el problema de valor inicial que nos proporciona las curvas características para distintos valores ξ_i y transportar sobre ellas la función que define la condición inicial. Este tipo de métodos son denominados **métodos de características**.

NOTA:

Obviamente, cuando sea factible evaluar de forma exacta las curvas características, la solución exacta podrá estimarse analíticamente y la aplicación de los métodos numéricos a tales tipos de problemas (algunos de los cuales puedes encontrarlos en los anexos de este tema) servirá simplemente para verificar el buen comportamiento (o no) de los esquemas que se estén estudiando.

No obstante la forma usual de aplicar los métodos de características consiste en definir un mallado de la forma:

$$x_0 < x_1 < x_2 < .... < x_i < ..... < x_N < x_{N+1}$$

y estimar las soluciones en el punto (x_i, t^{n+1}) siguiendo en retroceso la característica que pasa por dicho punto hasta determinar el punto (x_i^*, t^n) en que dicha característica corta a la ordenada t^n.

Esta forma de proceder puede extenderse fácilmente a la consideración de problemas un poco más complicados en los que la EDP es sustituida por:

$$a(x,t,u)\frac{\partial u}{\partial t} + V(x,t,u)\frac{\partial u}{\partial x} = f(x,t,u).$$

pues como se vio en el primer tema de esta asignatura, en este caso las curvas características y la solución sobre ellas podrán deducirse a partir de las igualdades:

$$\frac{dt}{a(x,t,u)} = \frac{dx}{V(x,t,u)} = \frac{du}{f(x,t,u)}$$

Ilustremos la forma de proceder anterior sobre un ejemplo tomado de Smith[22]

Ejemplo:

Considérese que la función $u(x,t)$ satisface la EDP

$$u\frac{\partial u}{\partial t} + \sqrt{x}\frac{\partial u}{\partial x} = -u^2 \quad x \in I\!R; \quad t > 0$$

y que $u(x,0) = 1 \ \forall x \in I\!R$. Siendo $x^* \geq 0$ se desea determinar la curva característica que pasa por $(x^*, 0)$. Asimismo, tomando $x^* = 1$ se desea aproximar el valor de t para el que la curva característica anterior pasa por el punto $P(1,1,t_P)$ y obtener mediante el método de las características un valor aproximado u_P de la solución en él.

Para ello se tiene en este caso que:

$$\frac{dt}{u} = \frac{dx}{\sqrt{x}} = \frac{du}{-u^2}$$

De la igualdad:

$$\frac{dt}{u} = \frac{du}{-u^2}$$

se deduce que

$$dt = \frac{-1}{u}du \Rightarrow t = -\ln(Ku)$$

y como para $t = 0$ se debe verificar que $u(x,0) = 1$ se tendrá que

$$0 = -\ln(K1) \Rightarrow K = 1$$

[22] G.D. Smith (1.985) Numerical Solution of partial Differential Equations. Finite Difference Methods. (3^a edición, 4^a reimpresión (1996)). Ed. Clarendon Press.

es decir,

$$t = -\ln(u) = \ln(\frac{1}{u})$$

lo que nos indica que sobre la curva característica que estamos buscando se satisface la igualdad:

$$u = e^{-t}$$

De forma similar se tiene que

$$\frac{dx}{\sqrt{x}} = \frac{du}{-u^2} \Rightarrow 2\sqrt{x} = \frac{1}{u} + C$$

y como en $(x^*, 0)$ se verifica que $u(x^*, 0) = 0$ resultará que

$$2\sqrt{x^*} = 1 + C \Rightarrow C = 2\sqrt{x^*} - 1$$

por lo que

$$\frac{1}{u} = 2\left(\sqrt{x} - \sqrt{x^*}\right) + 1$$

En resumen la curva característica buscada responderá a la expresión:

$$t = \ln\left(\frac{1}{2\left(\sqrt{x} - \sqrt{x^*}\right) + 1}\right)$$

y sobre ella la solución $u(x,t)$ satisfacerá la expresión

$$u(x,t) = e^{-t} = \frac{1}{2\left(\sqrt{x} - \sqrt{x^*}\right) + 1}$$

De las expresiones analíticas de la solución y de la curva característica ya sería inmediato obtener los valores exactos pedidos en el enunciado de este ejercicio para el instante t_P en el que la característica que pase por $(x^*, 0) = (1, 0)$ pasa por $(1{,}1, t_P)$ y del valor de u en dicho punto. No obstante veamos cómo se podrían aproximar dichos valores mediante el método de las características. Para ello consideramos nuevamente que:

$$\sqrt{x}dt = udx$$

y

$$\sqrt{x}du = -u^2dx$$

que discretizadas nos conducirán a que:

$$\sqrt{x^*}(t_P^{(1)} - 0) \approx u^*(x_P - x^*) \Rightarrow t_P^{(1)} \approx \frac{u^*(x_P - x^*)}{\sqrt{x^*}} = \frac{1(1{,}1 - 1)}{\sqrt{1}} = 0{,}1$$

y

$$\sqrt{x^*}(u_P^{(1)} - u^*) \approx -(u^*)^2(x_P - x^*) \Rightarrow u_P^{(1)} \approx$$

$$u^* - \frac{(u^*)^2(x_P - x^*)}{\sqrt{x^*}} = 1 - \frac{1^2(1{,}1 - 1)}{\sqrt{1}} = 0{,}9$$

Estos valores aproximados podrían refinarse realizando diferentes ponderaciones entre ellos. Así, por ejemplo, se podría plantear:

$$\sqrt{x}dt = udx \rightarrow \frac{\sqrt{x^*} + \sqrt{x_P}}{2}(t_P^{(2)} - 0) \approx \frac{u^* + u_P^{(1)}}{2}(x_P - x^*) \Rightarrow$$

$$\Rightarrow t_P^{(2)} = \frac{u^* + u_P^{(1)}}{\sqrt{x^*} + \sqrt{x_P}}(x_P - x^*) = \frac{1 + 0{,}9}{1 + \sqrt{1{,}1}}0{,}1 = 0{,}0927$$

y

$$\frac{\sqrt{x^*} + \sqrt{x_P}}{2}(u_P^{(2)} - u^*) \approx -\frac{(u^*)^2 + (u_P^{(1)})^2}{2}(x_P - x^*) \Rightarrow$$

$$\Rightarrow u_p^{(2)} = u^* - \frac{(u^*)^2 + (u_P^{(1)})^2}{\sqrt{x^*} + \sqrt{x_P}}(x_P - x^*) = 1 - \frac{1 + (0{,}9)^2}{1 + \sqrt{1{,}1}}0{,}1 = 0{,}9117$$

Un tercer (y posteriores) refinamiento podría realizarse de estos valores. No obstante pueden compararse los obtenidos en esta segunda aproximación con los exactos ($t_P = 0{,}0934$, $u_P = 0{,}9111$) observándose una buena concordancia con las aproximaciones realizadas.

En la lista bibliográfica que se cita al final de este tema podrás encontrar diferentes variantes del método de las caracterísiticas (así como su extensión a problemas de segundo orden). Nosotros, siempre por la falta de tiempo, no entraremos en sus detalles y nos limitaremos a presentarte otros métodos numéricos numéricos de amplio uso basados en la discretizaciñon de las derivadas que aparecen en el problema convectivo mediante distintas estrategias en diferencias finitas. Y puesto que ya se ilustró en el caso estacionario la conveniencia de introducir descentrajes en la aproximación de los términos convectivos nos centraremos en un esquema descentrado (habitualmente conocido como esquema "upwind").

4.5.2. El esquema "upwind" explícito.

Consideremos inicialmente el problema advectivo:

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) & +V\dfrac{\partial u}{\partial x}(x,t)=0 \quad 0<x<L \quad t>0 \\ u(0,t) & = u_I(t) \;\; t>0, \\ u(x,0) & = u^0(x) \qquad\qquad 0\le x\le L \end{cases}$$

donde supondremos que V es una constante positiva.

NOTAS:

1ª) Obsérvese que en este caso la solución analítica en el punto (x^*,t^*) puede obtenerse fácilmente siguiendo en retroceso la curva característica que por él pasa y que en este caso será una recta que tendrá por ecuación:

$$x = x^* + V(t-t^*)$$

Si esta recta corta antes al eje de abscisas (eje X) que al de ordenadas (eje de tiempos) la solución estará dada por:

$$u(x^*,t^*) = u^0(x^* - Vt^*)$$

Y si cortase antes al eje de ordenadas que al de abscisas la solución sería:

$$u(x^*,t^*) = u_I(t^* - \frac{x^*}{V})$$

2ª) Te dejamos como ejercicio propuesto adaptar todo cuanto se diga para el caso de velocidad de advección positiva al caso en que $V<0$. A lo largo de la descripción de este método te iremos dando pistas que faciliten esta tarea.

Consideremos además un conjunto de valores de t de la forma

$$t^n = n\Delta t \qquad (n=0,1,2,.....)$$

y dividamos el intervalo $[0,L]$ en N subintervalos de la forma $[x_i,x_{i+1}]$ $(i=1,2,...,N)$ donde:

$$\Delta x = \frac{L}{N} \qquad \text{y} \qquad x_i = (i-1)\Delta x \qquad (i=1,2,...,N)$$

El esquema "upwind" consiste en plantear la EDP de advección sobre cada uno de los nodos interiores del mallado considerado y aproximar la derivada temporal en (x_i,t^n) mediante una fórmula progresiva y la derivada espacial mediante una fórmula regresiva. Más concretamente:

$$\frac{\partial u}{\partial t}(x_i,t^n) + V\frac{\partial^2 u}{\partial x^2}(x_i,t^n) = 0 \rightsquigarrow \frac{u_i^{n+1}-u_i^n}{\Delta t} + V\frac{u_i^n - u_{i-1}^n}{\Delta x} = 0$$

lo que nos permitirá expresar

$$u_i^{n+1} = u_i^n + V\frac{\Delta t}{\Delta x}(u_{i-1}^n - u_i^n) = cu_{i-1}^n + (1-c)u_i^n$$

donde hemos denotado por c al valor

$$c = V\frac{\Delta t}{\Delta x}$$

que habitualmente se conoce con el nombre de **número de Courant** del esquema.

NOTA:

Si $V < 0$ el esquema resultante sería:

$$u_i^{n+1} = (1+c)u_i^n - cu_{i+1}^n$$

Con ello se puede plantear el esquema de cálculo recogido en el algoritmo siguiente:

COMIENZO DEL ALGORITMO
Definir el mallado concretando el valor de Δt y de Δx y de $\{x_i\}_{i=1}^{N+1}$
Definir el valor de V
$c \leftarrow V\Delta t/\Delta x$
Evaluar $u_i^0 \leftarrow u^0(x_i) \quad (i = 1, ..., N+1)$
Para $n = 0, 1, 2, ...$
Conocidos los valores $\{u_i^n\}_{i=1}^{N+1}$
Para $i = 2$ *hasta* $i = N+1$ *con paso* 1 *hacer:*
$u_j^{n+1} \leftarrow cu_{i-1}^n + (1-c)u_i^n$
Fin bucle en i.
$u_1^{n+1} \leftarrow u_I(t^{n+1})$
Fin bucle en n.
Escritura de resultados
FIN ALGORITMO.

Apliquemos el esquema a la resolución de un ejemplo. Para ello consideremos el problema:

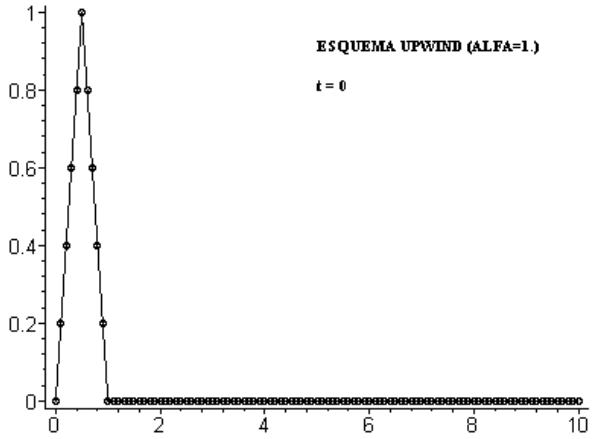
$$\left\{\begin{array}{lll} \dfrac{\partial u}{\partial t}(x,t) & +\dfrac{\partial u}{\partial x}(x,t) = 0 & 0 < x < 10 \quad t > 0 \\ u(0,t) & = u_I(t) \;\; t > 0, & \\ u(x,0) & = u^0(x) & 0 \le x \le L \end{array}\right\}$$

siendo $u_I(t) = 0$ y

$$u^0(x) = \left\{\begin{array}{ll} 2x & si\ x \in [0, 0{,}5] \\ 2(1-x) & si\ x \in [0{,}5, 1] \\ 0 & si\ x \notin [0, 1] \end{array}\right\}$$

Consideremos inicialmente que $\Delta x = \Delta t = 0{,}1$. Con ello se obtendrán como puntos del mallado:

$$x_1 = 0, \;\; x_2 = 0{,}1, \;\; x_3 = 0{,}2, ..., x_i = (i-1)0{,}1, ..., x_{101} = 10.$$



y como valores iniciales con los que arrancar el proceso de cálculo:

$$u_1^0 = 0,\ \ u_2^0 = 0{,}2,\ \ u_3^0 = 0{,}4,\ \ u_4^0 = 0{,}6,\ u_5^0 = 0{,}8,\ \ u_6^0 = 1{,}0,\ \ u_7^0 = 0{,}8,$$

$$u_8^0 = 0{,}6,\ \ u_9^0 = 0{,}4,\ \ u_{10}^0 = 0{,}2,\ \ u_{11}^0 = 0,\ u_{12}^0 = ... = u_{101}^0 = 0$$

Además en este caso se tiene que $c = 1$ por lo que el esquema de cálculo se resume en:

$$u_i^{n+1} = u_{i-1}^n$$

En el instante de cálculo $t^1 = 0{,}1$ se irán obteniendo las siguientes aproximaciones de la solución:

$$u_1^1 = u_I(0{,}1) = 0,\ \ u_2^1 = u_1^0 = 0.,\ \ u_3^1 = u_2^0 = 0{,}2,\ u_4^1 = u_3^0 = 0{,}4,\ \ u_5^1 = u_4^0 = 0{,}6, ...$$

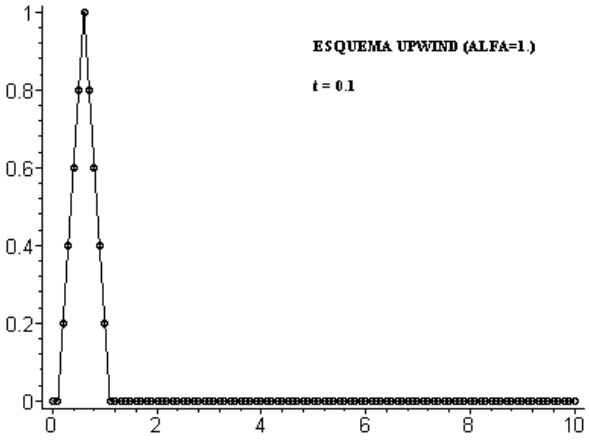
Una vez conocidas las soluciones en el intante t^1 podrán calcularse la del instante $t^2 = 0{,}2$ mediante el mismo esquema: $u_i^2 = u_{i-1}^1 = u_{i-1}^0$. En las figuras siguientes se recogen las gráficas de la solución aproximada y de la solución analítica que como ves son coincidentes.

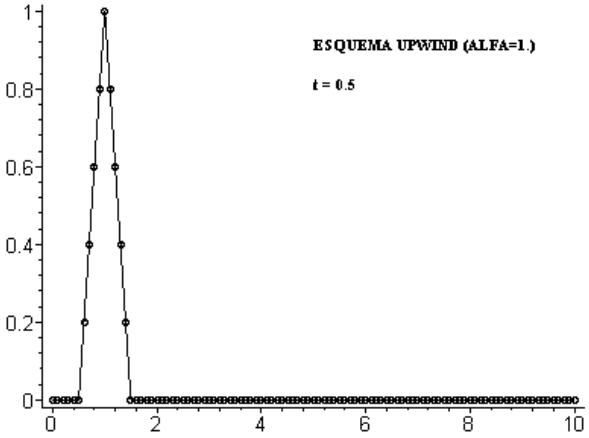
Volvamos a repetir el proceso anterior manteniendo $\Delta x = 0{,}1$ pero disminuyendo el valor de Δt al valor $\Delta t = 0{,}05$. Ello nos conduce a que el número de Courant en este caso es:

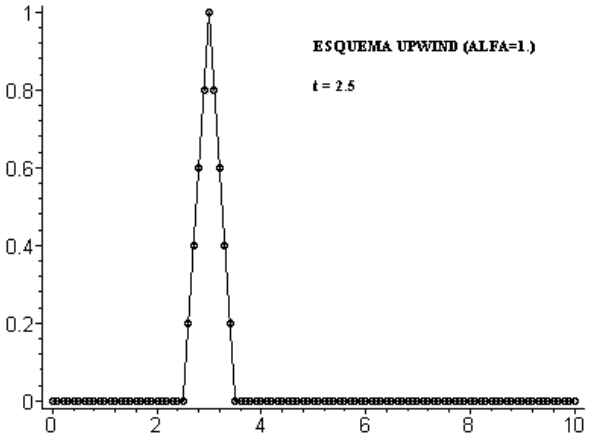
$$c = V\frac{\Delta t}{\Delta x} = 0{,}5$$

por lo que el esquema de cálculo se reduce a

$$u_i^{n+1} = \frac{u_{i-1}^n + u_i^n}{2}$$


ESQUEMA UPWIND (ALFA=1.)
t = 0.1
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10


ESQUEMA UPWIND (ALFA=1.)
t = 0.5
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10


ESQUEMA UPWIND (ALFA=1.)
t = 2.5
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10

lo que, partiendo de los mismos valores iniciales nos conduce para el instante $t^1 = 0{,}05$ a los valores:

$$u_1^1 = u_I(0{,}05) = 0$$

$$u_2^1 = \frac{u_1^0 + u_2^0}{2} = \frac{0 + 0{,}2}{2} = 0{,}1$$

$$u_3^1 = \frac{u_2^0 + u_3^0}{2} = \frac{0{,}2 + 0{,}4}{2} = 0{,}3$$

$$u_4^1 = \frac{u_3^0 + u_4^0}{2} = \frac{0{,}4 + 0{,}6}{2} = 0{,}5$$

$$u_5^1 = \frac{u_4^0 + u_5^0}{2} = \frac{0{,}6 + 0{,}8}{2} = 0{,}7$$

.....

Una vez obtenidos los valores en $t^1 = 0{,}05$ podremos pasar a estimar los valores nodales en el instante $t^2 = 0{,}1$ mediante:

$$u_1^2 = u_I(0{,}1) = 0$$

$$u_2^2 = \frac{u_1^1 + u_2^1}{2} = \frac{0 + 0{,}1}{2} = 0{,}05$$

$$u_3^2 = \frac{u_2^1 + u_3^1}{2} = \frac{0{,}1 + 0{,}3}{2} = 0{,}2$$

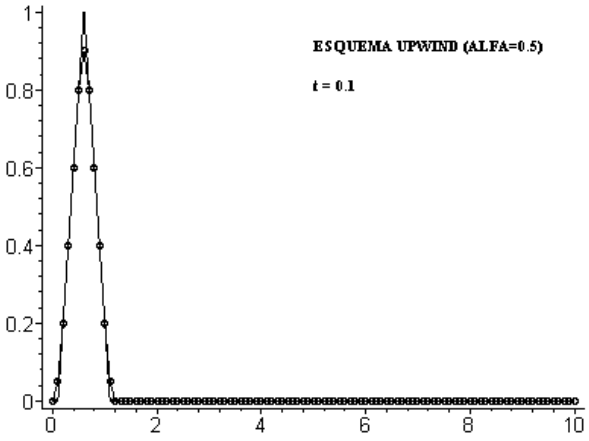
$$u_4^2 = \frac{u_3^1 + u_4^1}{2} = \frac{0{,}3 + 0{,}5}{2} = 0{,}4$$

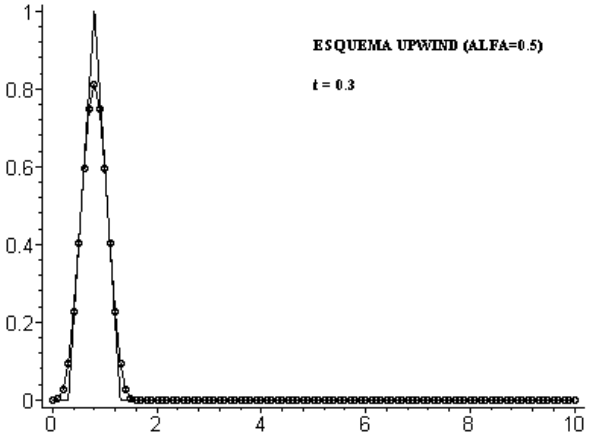
$$u_5^2 = \frac{u_4^1 + u_5^1}{2} = \frac{0{,}5 + 0{,}7}{2} = 0{,}6$$

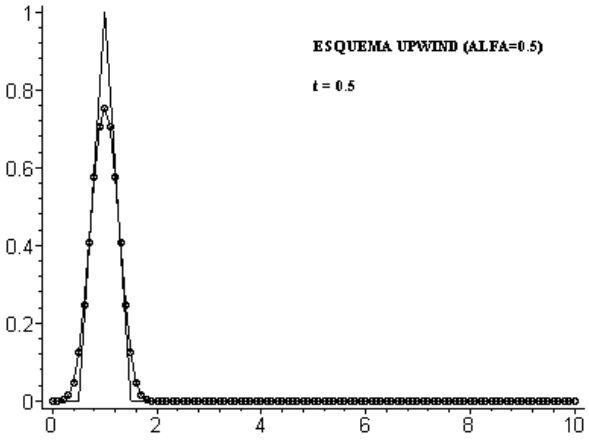
.....

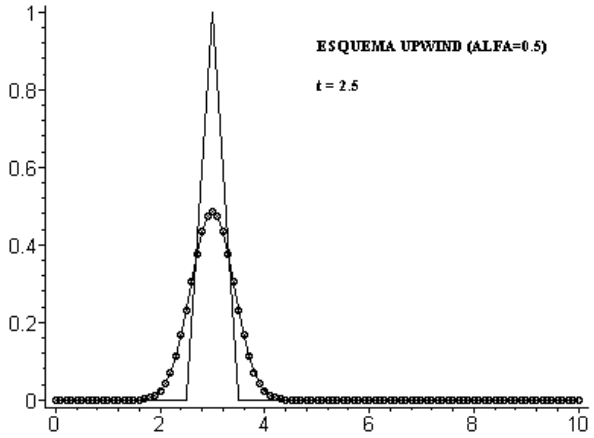
Si representamos los valores que se van obteniendo en sucesivos instantes de tiempo obtendremos las siguientes gráficas:

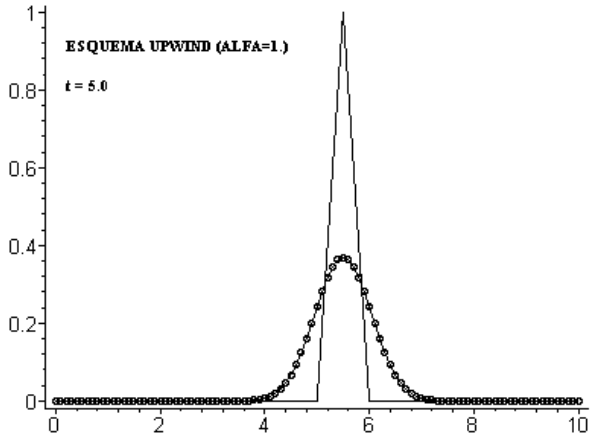
Puede observarse que, al igual que sucedía en los problemas difusivos, una discretización más fina en la variable temporal ha sido contraproducente para los resultados obtenidos.


ESQUEMA UPWIND (ALFA=0.5)
t = 0.1
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10


ESQUEMA UPWIND (ALFA=0.5)
t = 0.3
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10


ESQUEMA UPWIND (ALFA=0.5)
t = 0.5
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10


ESQUEMA UPWIND (ALFA=0.5)
t = 2.5
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10


ESQUEMA UPWIND (ALFA=1.)
t = 5.0
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10

Busquemos la justificación de este hecho. Un primer razonamiento que puede hacerse es que en el caso en que el número de Courant $c = 1$ el esquema resultante sigue exactamente la recta caracterísitica ya que, si $x_i - t^n > 0$:

$$u_i^n = u_{i-1}^{n-1} = ... = u_{i-n}^0 = u^0(x_i - n\Delta x) = u^0(x_i - n\Delta t) = u^0(x_i - t^n) = u(x_i, t^n)$$

Obviamente para otros valores de α el esquema no seguirá la curva característica de forma exacta.

Pero profundicemos un poco más en cómo variará la solución en función del número de Courant c. Es decir, analicemos el esquema.

4.5.3. Orden de consistencia del esquema "upwind".

Al igual que se hizo para la aproximación de problemas difusivos comencemos estudiando el orden de consistencia del método. Para ello se tiene que, desarrollando la solución analítica en torno al punto (x_i, t^n) :

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} = \frac{\partial u}{\partial t}(x_i, t^n) + \frac{(\Delta t)}{2}\frac{\partial^2 u}{\partial t^2}(x_i, t^n) + \frac{(\Delta t)^2}{6}\frac{\partial^3 u}{\partial t^3}(x_i, t^n) + ...$$

y

$$\frac{U_i^n - U_{i-1}^n}{\Delta t} = \frac{\partial u}{\partial x}(x_i, t^n) - \frac{(\Delta x)}{2}\frac{\partial^2 u}{\partial x^2}(x_i, t^n) + \frac{(\Delta x)^2}{6}\frac{\partial^3 u}{\partial x^3}(x_i, t^n) + ...$$

de donde,

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} + V\frac{U_i^n - U_{i-1}^n}{\Delta t} = \left(\frac{\partial u}{\partial t}(x_i, t^n) + V\frac{\partial u}{\partial x}(x_i, t^n)\right) +$$

$$+ \left(\frac{(\Delta t)}{2}\frac{\partial^2 u}{\partial t^2}(x_i, t^n) - V\frac{(\Delta x)}{2}\frac{\partial^2 u}{\partial x^2}(x_i, t^n)\right) +$$

$$+ \left(\frac{(\Delta t)^2}{6}\frac{\partial^3 u}{\partial t^3}(x_i, t^n) + V\frac{(\Delta x)^2}{6}\frac{\partial^3 u}{\partial x^3}(x_i, t^n)\right) + ...$$

Al ser $u(x, t)$ solución de la EDP de convección se verificará que:

$$\frac{\partial u}{\partial t}(x_i, t^n) + V\frac{\partial u}{\partial x}(x_i, t^n) = 0$$

por lo que el error de consistencia del esquema estará dado por

$$E_i^{n+1} = \left(\frac{(\Delta t)}{2}\frac{\partial^2 u}{\partial t^2}(x_i, t^n) - V\frac{(\Delta x)}{2}\frac{\partial^2 u}{\partial x^2}(x_i, t^n)\right) +$$

$$+\left(\frac{(\Delta t)^2}{6}\frac{\partial^3 u}{\partial t^3}(x_i,t^n)+V\frac{(\Delta x)^2}{6}\frac{\partial^3 u}{\partial x^3}(x_i,t^n)\right)+...$$

es decir, que en general el esquema será de orden $O(\Delta t,\Delta x)$.

NOTA:

Si denotamos por T_i^{n+1} a la parte principal del error de consistencia del esquema se tendrá que:

$$T_i^{n+1}=\frac{(\Delta t)}{2}\frac{\partial^2 u}{\partial t^2}(x_i,t^n)-V\frac{(\Delta x)}{2}\frac{\partial^2 u}{\partial x^2}(x_i,t^n)\Rightarrow$$

$$\Rightarrow\frac{2}{\Delta t}T_i^{n+1}=\frac{\partial^2 u}{\partial t^2}(x_i,t^n)-c\frac{\partial^2 u}{\partial x^2}(x_i,t^n)$$

Puesto que:

$$\frac{\partial u}{\partial t}(x,t)+V\frac{\partial u}{\partial x}(x,t)=0$$

si $u(x,t)$ es suficientemente regular, se tendrá que:

$$\frac{\partial^2 u}{\partial t^2}(x,t)+V\frac{\partial^2 u}{\partial t\partial x}(x,t)=0$$

y

$$\frac{\partial^2 u}{\partial x\partial t}(x,t)+V\frac{\partial^2 u}{\partial x^2}(x,t)=0$$

expresiones de las que se deduce que:

$$\frac{\partial^2 u}{\partial t^2}(x,t)-V^2\frac{\partial^2 u}{\partial x^2}(x,t)=0$$

Entrando con esta igualdad particularizada en el punto (x_i,t^n) en la expresión de T_i^{n+1} se tiene que:

$$\frac{2}{\Delta t}T_i^{n+1}=(V^2-V\frac{\Delta t}{\Delta x})\frac{\partial^2 u}{\partial x^2}(x_i,t^n)$$

por lo que si

$$\frac{\Delta t}{\Delta x}=V$$

el término principal del error de consistencia desaparecerá y el orden del esquema será superior.

4.5.4. El método de von Neumann aplicado al estudio de la estabilidad del esquema "upwind".

La estabilidad de un esquema como el que se acaba de presentar es susceptible de ser analizado mediante técnicas de análisis espectral como las desarrolladas para el caso de los problemas difusivos. No obstante, por introducir otra técnica para el estudio de la estabilidad, realizaremos este estudio mediante el método de von Neumann.

Dicha técnica consiste en comparar el comportamiento de las soluciones aproximadas sobre problemas que admitan como solución analítica algún armónico de Fourier, es decir soluciones de la forma

$$u(x,t) = Ae^{i(kx+wt)}$$

donde ahora i representa la unidad imaginaria $i = \sqrt{-1}$.

NOTA:

Veánse los anexos a este tema para un estudio detallado sobre los armónicos de Fourier y su representación física como una onda en donde se introducen conceptos tales como número de onda (k), longitud de onda ($\lambda = 2\pi/k$), pulsación (w) o periódo (T).

Para que tal función sea solución analítica de la EDP:

$$\frac{\partial u}{\partial t} + V\frac{\partial u}{\partial x} = 0$$

se debe verificar que

$$w = -Vk$$

por lo que tal armónico puede expresarse como

$$u(x,t) = Ae^{ik(x-Vt)} = e^{-iVkt}Ae^{ikx}$$

y como

$$u(x,0) = u^0(x) = Ae^{ikx}$$

la solución en cualquier punto x y en cualquier instante t responde a la expresión:

$$u(x,t) = e^{-iVkt}u^0(x)$$

Esta forma de escribir la expresión anterior nos indica que la solución analítica en un instante t puede obtenerse como el producto de la solución inicial multiplicada por e^{-iVkt}. Un razonamiento análogo nos conduciría a que:

$$u(x,t+\Delta t) = e^{-iVk\Delta t}u(x,t)$$

Es decir, es la función (compleja) $G(k\Delta t) = e^{-iVk\Delta t}$ la responsable de "modificar" nuestra solución de un instante t a otro $t+\Delta t$. En cualquier punto x y en el instante $t+\Delta t$ puede obtenerse el módulo de:

$$|u(x,t+\Delta t)| = |G(k\Delta t)u(x,t)| = |G(k\Delta t)|\,|u(x,t)|$$

Es decir, que la onda armónica que representa la solución analítica se verá amplificada o amotiguada al ser multiplicada por el valor $|G(k\Delta t)|$. Por este motivo a dicho valor se le denominará **factor de amplificación exacto**. Es obvio que en este caso:

$$|G(k\Delta t)| = 1$$

Pero cuando se trabaja con números complejos, además del módulo, debe también determinarse su argumento. En este sentido se tiene que:

$$\arg(u(x,t+\Delta t)) = \arg(G(k\Delta t)u(x,t)) = \arg(G(k\Delta t)) + \arg(u(x,t))$$

La expresión anterior nos indica que el argumento de $u(x,t+\Delta t)$ será el de $u(x,t)$ más el argumento de $G(k\Delta t)$. Por ello al valor $\arg(G(k\Delta t))$ se le denominará **factor de desfase exacto**. También es evidente que en este caso:

$$\arg(G(k\Delta t)) = Vk\Delta t$$

Cuando consideremos una discretización temporal de paso Δt y una espacial de paso Δx será cómodo expresar los parámetros anteriores en función del número de Courant

$$c = V\frac{\Delta t}{\Delta x}$$

escribiendo $G(c, k\Delta x)$ en lugar de $G(k\Delta t)$ con lo que

$$\arg(G(c,k\Delta x)) = ck\Delta x$$

NOTA:

Obsérvese que los resultados anteriores nos expresan que la solución analítica verifica:

$$u(x,t+\Delta t) = u(x-V\Delta t,t)$$

lo que, procediendo recursivamente, es otra forma de confirmar que la solución analítica del problema está dada por:

$$u(x,t+\Delta t) = u(x-V(t+\Delta t),0) = u^0(x-V(t+\Delta t))$$

Hasta aquí todo lo dicho se ha referido a la solución analítica. ¿Cómo son las soluciones aproximadas que nos produce nuestro esquema?. Estas estarán dadas por:

$$u_j^{n+1} = (1+c)u_j^n - cu_{j+1}^n$$

por lo que, para el primer paso de tiempo se tendrá que:

$$u_j^1 = cu_{j-1}^n + (1-c)u_j^n = cAe^{ik(x_j-\Delta x)} + (1-c)Ae^{ikx_j} =$$

$$= Ae^{ikx_j}(1-c(1-e^{-ik\Delta x})) = (1-c(1-e^{-ik\Delta x}))u_j^0 = g(c,k\Delta x)u_j^0$$

Este mismo razonamiento, realizado de forma recurrente, nos conduce a que:

$$u_j^{n+1} = (1-c(1-e^{-ik\Delta x}))u_j^n = g(c,k\Delta x)u_j^n$$

Definition 7 *Se denomina* ***factor de amplificación del esquema numérico*** *al valor:*

$$|g(c,k\Delta x)|\,.$$

Definition 8 *Se denomina* ***factor de desfase del esquema numérico*** *al valor:*

$$\arg(g(c,k\Delta x)).$$

Obviamente interesará que el **error de amplificación y el error de fase** definidos como:

$$E_a = |g(c,k\Delta x)| - |G(c,k\Delta x)| = |g(c,k\Delta x)| - 1$$

$$E_\varphi = \arg(G(c,k\Delta x)) - \arg(g(c,k\Delta x))$$

sean lo más próximos a 0 posible. Examinemos cómo serán en este caso. Para ello se tienen las siguientes proposiciones:

Proposition 12 *Para el esquema "upwind" se verifica que:*

$$Si\ c>1: \quad \max_k |g(c,k\Delta x)| = |g(c,\pi)| = 2c-1 > 1$$

$$Si\ c\le 1: \quad \max_k |g(c,k\Delta x)| = 1.$$

Demostración:

$$g(c, k\Delta x) = 1 - c(1 - e^{-ik\Delta x}) = 1 - c(1 - \cos(k\Delta x)) - ic\sin(k\Delta x)$$

Por tanto,

$$|g(c, k\Delta x)|^2 = (1-c)^2 + c^2\cos^2(k\Delta x) + 2c(1-c)\cos(k\Delta x) +$$

$$c^2\sin^2(k\Delta x) = (1-c)^2 + c^2 + 2c(1-c)\cos(k\Delta x) =$$

$$= (1-c)^2 + c^2 + 2c(1-c) + 2c(1-c)(\cos(k\Delta x) - 1) =$$

$$= 1 - 2c(1-c)(1 - \cos(k\Delta x))$$

de donde

$$Sup_k\,|g(c, k\Delta x)| = Máx\,(1, |1-2c|)$$

de donde es evidente el resultado de la proposición.

c.q.d.

Proposition 13 Se verifica que:

$$g(c, k\Delta x) = g(c, k'\Delta x) \qquad \forall k, k' \;/\; k\Delta x = k\Delta x + 2\pi$$

$$g(c, k\Delta x) = \overline{g}(c, -k\Delta x) \qquad \forall k \in I\!R.$$

Demostración:

Evidente si se tiene en cuenta que:

$$g(c, k\Delta x) = 1 - c(1 - e^{-ik\Delta x}) = 1 - c(1 - \cos(k\Delta x)) - ic\sin(k\Delta x)$$

c.q.d

Proposition 14 *Para pequeños valores de $K = k\Delta x$, es decir, para grandes longitudes de onda respecto a Δx, se verifica que:*

$$E_a = -\frac{c(1-c)}{2}K^2 + O(K^4)$$

$$E_\varphi = \frac{c(2c-1)(1-c)}{6}K^3 + O(K^5).$$

Demostración:

Desarrollando en serie en torno a $K = 0$ la expresión obtenida en la demostración de la proposición VI.10 para $|g(c,K)|^2$ se tiene que:

$$|g(c,K)|^2 = 1 - c(1-c)K^2 + O(K^4)$$

luego

$$|g(c,K)| = 1 - \frac{c(1-c)}{2}K^2 + O(K^4)$$

de donde se tiene la expresión del error de amplitud dada en el enunciado.

Para analizar el error de fase designemos por $\varphi = \arg(g(c,K))$. Se verificará entonces que:

$$\sin(\varphi) = \frac{c\sin(K)}{|g(c,K)|}$$

y puesto que para pequeños valores de K :

$$\sin(K) = K - \frac{K^3}{6} + O(K^5)$$

y

$$\frac{1}{|g(c,K)|} = 1 + \frac{c(1-c)}{2}K^2 + O(K^4)$$

se tendrá que:

$$\frac{c\sin(K)}{|g(c,K)|} = cK + c\left(\frac{c(1-c)}{2} - \frac{1}{6}\right)K^3 + O(K^5)$$

Si por otra parte se considera que:

$$\varphi(K) = \varphi_1 K + \varphi_3 K^3 + O(K^5)$$

se tendrá que

$$\sin(\varphi(K)) = \varphi_1 K + (\varphi_3 - \frac{\varphi_1}{6})K^3 + O(K^5)$$

e identificando ambos desarrollos se tiene que:

$$\varphi_1 = c, \quad \varphi_3 = \frac{c}{6}(1-2c)(c-1)$$

por lo que, para pequeños valores de K :

$$E_\varphi = \arg(G(c,K)) - \arg(g(c,K)) = cK - cK - \frac{c}{6}(1-2c)(c-1)K^3 + O(K^5) \Rightarrow$$

$$\Rightarrow E_\varphi = \frac{c}{6}(2c-1)(1-c)K^3 + O(K^5). \qquad \text{c.q.d}$$

Obsérvese que según lo anterior, el error de amplitud es negativo y de orden 1 en Δx. Además el error de fase es de orden 2 y es negativo si $0 < c < 1/2$ siendo positivo en otro caso. Por otra parte, el factor de amplificación numérico será superior a 1 cuando $c > 1$. En otros téminos, puesto que:

$$u_j^n = |g(c,K)|\, u_j^{n-1} = |g(c,K)|^2\, u_j^{n-2} = ... = |g(c,K)|^n\, u_j^0$$

la solución tenderá a explotar cuando n tienda a infinito si $c > 1$. Ello justifica (que no demuestra) el siguiente teorema cuya demostración rigurosa podrás encontrar en Morton y Mayers[23]:

Theorem 15 *Una condición necesaria y suficiente para que el esquema "upwind" sea estable es que se verifique la condición:*

$$c = V\frac{\Delta t}{\Delta x} \leq 1.$$

NOTA:

La condición recogida en el teorema anterior (u otra análoga en la que varíe la cota del número de Courant) aparece en el estudio de muy diferentes esquemas y se conoce con el nombre de condición de Courant-Fiedrichs-Lewy (o brevemente como condición CFL).

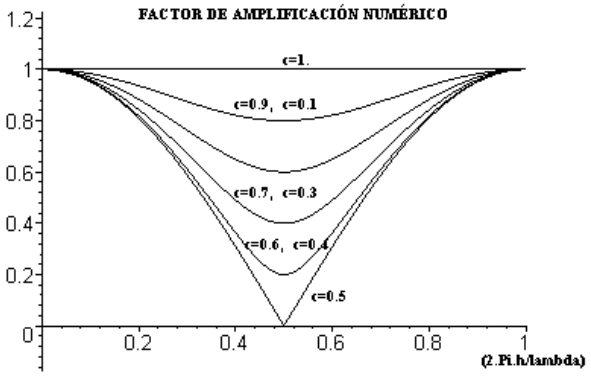
Conocida la expresión de factor de amplificación numérico puede representarse para distintos valores de $K = k\Delta x$ y distintos valores del número de Courant c. Aún más ilustrativo es tomar como eje de abscisas el valor de

$$\frac{\Delta x}{\lambda}$$

donde λ es la longitud de onda del armónico (y que como puede consultarse en el anexo a este tema se relaciona con el número de onda k por la relación: $k = 2\pi/\lambda$). Ello nos conduce a gráficas como la siguiente

Puedes observar en ella que para distintos valores de c menores que 1 el factor amortiguamiento sigue curvas muy diferentes lo cual, al realizar sucesivas etapas de cálculo nos puede conducir a soluciones que no exploten (que sean estables) pero tiendan a desaparecer. De todas formas, como parece evidente, cuando los valores de $\Delta x/\lambda$ son pequeños (mallados espaciales muy finos) el valor del factor de amplificación más se aproxima a la unidad.

[23] K.W. Morton y D.F. Mayers. (1994) Numerical solution of Partial Differential Equations. Ed. Cambridge University Press.



NOTA FINAL:

La disponibilidad de tiempo nos ha hecho tratar tan sólo uno de los esquemas en diferencias para aproximar problemas convectivos. No obstante existen muchos otros (explícitos, como el que aquí hemos considerado, e implícitos). Entre ellos merece la pena al menos citar los esquemas de Lax, Lax-Wendroff, Leap-frog, QUICK, etc... No nos queda más remedio que remitirte a la bibliografía sobre este tema para encontrar detalles sobre ellos.

4.6. Bibliografía

- A. Bamberger. (1982) Analyse numerique des équations aux derivées partielles. Support du cours de DEA. Université Pierre et marie Curie.
- R.L. Burden y J. D. Faires. (1998) Análisis numérico. (6^a edición). Ed. Thomson International.
- C. Conde y E. Schiavi (2000). Guiones de la asignatura de Elementos de Matemáticas. Universidad Rey Juan Carlos.
- C. Conde y G. Winter (1.991) Métodos y algoritmos básicos del álgebra numérica. Ed. Reverté.
- D. Euvrard. (1994) Résolution numérique des équations aux derivées partielles de la physique, de la mécanique et des sciences de l'ingénieur. Différences finies, éléments finis, problèmes en domaine non borné. (3^a edición) Ed. Masson.
- S. Godunov y V. Rabenki (1977) Schémes aux differences. Ed. Mir.
- L. Lapidus y G.F. Pinder.(1982) Numerical solution of partial differential equations in science and engineering. Ed.: John Wiley.
- G. Marchouk (1980) Méthodes de calcul numérique. Ed. Mir.
- F. Michavila y C. Conde. (1987). Métodos de aproximación. Ed. Universidad Politécnica de Madrid.
- K.W. Morton y D.F. Mayers. (1994) Numerical solution of Partial Differential Equations. Ed. Cambridge University Press.
- R.D. Ritchmyer y K.W. Morton (1967) Difference Methods for Initial Value Problems. 2^a edición. Ed.:Wiley-Interscience. Reimpreso en 1994 por Ed. Kreiger.
- Schiavi, E., Muñoz Montalvo, A.I., Conde, C. (2012). Métodos Matemáticos para los Grados en Ingeniería. Primera parte: teoría. Ed. Dykinson, Textos Docentes 31, Universidad Rey Juan Carlos, ISBN: 978-84-15454-58-8.
- M. Sibony y J. Cl. Mardon (1984) Analyse numériqe II: Approximation et équations differentielles. Ed Hermann.
- G.D. Smith (1.985) Numerical Solution of partial Differential Equations. Finite Difference Methods. (3^a edición, 4^a reimpresión (1996)). Ed. Clarendon Press.

- Zill, D. G. (1997). Ecuaciones diferenciales con aplicaciones de modelado. (VI edición) Ed. International Thomson editores.