

Miquel Grau Sánchez
Miquel Noguera Batlle

Cálculo numérico

POLITEXT 98

Cálculo numérico

POLITEXT

Miquel Grau Sánchez
Miquel Noguera Batlle

Cálculo numérico

Teoría y práctica

EDICIONS UPC

La presente obra fue galardonada en el primer concurso
"Ajuts a l'elaboració de material docent" convocado por la UPC.

Primera edición: marzo de 2001

Diseño de la cubierta: Manuel Andreu

- © Los autores, 2001
- © Edicions UPC, 2001
Edicions de la Universitat Politècnica de Catalunya, SL
Jordi Girona Salgado 31, 08034 Barcelona
Tel.: 934 016 883 Fax: 934 015 885
Edicions Virtuals: www.edicionsupc.es
E-mail: edicions-upc@upc.es

Producción: Grup Artyplan-Artympres S. A.
Agricultura 21, Nave 5, 08980 Sant Feliu de Ll. (Barcelona)

Depósito legal: B-13.283-2001
ISBN: 84-8301-455-6

Quedan rigurosamente prohibidas, sin la autorización escrita de los titulares del copyright, bajo las sanciones establecidas en las leyes, la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares de ella mediante alquiler o préstamo públicos.

Prefacio

Numerical Analysis is the effective representation of anything by anything.

Preston C. Hammer

Este libro intenta poner en las manos de los estudiantes de un primer ciclo universitario de estudios científicos o técnicos los métodos numéricos más adecuados a los problemas que aparecen actualmente, así como las técnicas de las que se sirve la Matemática Aplicada.

La incorporación de una lista amplia de problemas, así como la presentación, discusión y resolución de una práctica por capítulo, con una lista de prácticas, buscan que este libro se convierta en un instrumento útil para la introducción en el Cálculo Numérico de una manera no sólo teórica, sino también con todas las herramientas de que se pueda disponer de forma efectiva y práctica. Además de presentar ejemplos de los métodos introducidos, las prácticas propuestas no sólo son un medio para irse introduciendo en las diferentes librerías numéricas, sino también para aproximarse a problemas relacionados con las ingenierías, de las cuales se presentan modelos simplificados para facilitar el trabajo del alumno.

El principal objetivo al escribir este libro es proponer una introducción al Análisis Numérico haciendo un repaso de las técnicas que se utilizan más frecuentemente; por lo tanto, uno de los puntos que se tratan en todos los capítulos es el análisis del error, ya de los diferentes problemas, ya de los métodos presentados. Por otro lado, en cada capítulo se presenta un número considerable de ejercicios y problemas numéricos para resolver, además de los ejemplos comentados con todo detalle.

Para una total comprensión del contenido de este libro, el lector ha de tener conocimientos de álgebra lineal, y de un primer curso completo de análisis y, para algunos capítulos, de temas más especializados, como son ecuaciones diferenciales ordinarias y variable compleja (ver el esquema que se presenta al final de este prefacio).

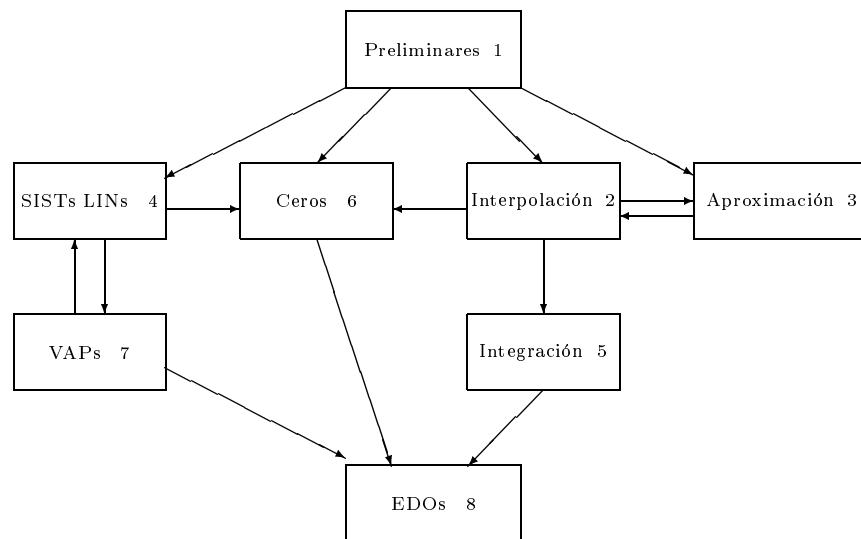
Unas palabras sobre el contenido del libro. Fundamentalmente, está dedicado a los siguientes temas: errores, interpolación y aproximación de funciones, resolución numérica de sistemas de

ecuaciones lineales, diferenciación e integración de funciones, métodos de cálculo de ceros de funciones no lineales, cálculo de valores y vectores propios, y resolución numérica de ecuaciones diferenciales ordinarias.

Una dificultad con que se encuentran los alumnos que se inician en el Cálculo Numérico es el hecho de que vayan aceptando que todos los cálculos que se hacen son aproximados, que su precisión es finita y a veces engañosa. Con la introducción de los manipuladores algebraicos, cuya precisión es tan grande como se quiera, sería conveniente, cuando se tiene que resolver un problema, hacer un estudio previo, por si fuera necesaria una capacidad de memoria inaccesible o un tiempo de ejecución imposible que haga más adecuado elegir técnicas numéricas. De todas maneras, siempre habrá problemas donde cualquier manipulador puede presentar menos dificultades y una respuesta más rápida; por lo tanto, será conveniente siempre un análisis previo.

Para terminar, desearíamos que este libro fuese el fruto de la experiencia de 25 años impartiendo asignaturas de métodos numéricos a estudiantes de ingeniería y matemáticas de la Universitat Politècnica de Catalunya y a recién licenciados que quieren aprender las técnicas básicas del Análisis Numérico para aplicarlas a su investigación.

Los autores.



Esquema de posibles relaciones entre los capítulos.

Índice General

Prefacio	3
Índice General	5
1 Preliminares	13
1.1 Introducción	13
1.1.1 Cálculo científico y campos de aplicación	13
1.1.2 Modelado matemático y solución numérica	14
1.1.3 Fuentes de error	17
1.1.4 Estabilidad de algoritmos	21
1.2 Representación aritmética en coma flotante	23
1.2.1 Conjunto de números en coma flotante	23
1.2.2 La función fl , redondeo y truncamiento	24
1.2.3 Épsilon de la máquina	24
1.3 Análisis del error	26
1.3.1 Definiciones	26
1.3.2 Errores de representación	27
1.3.3 Errores en las operaciones en coma flotante	28
1.3.4 Propagación del error	31

1.3.5	Acumulación del error	34
1.4	Cálculo de series	37
1.4.1	Métodos de comparación	38
1.4.2	Aceleración de la convergencia	39
1.5	Problemas	41
1.6	Prácticas	44
1.6.1	Práctica ejemplo	44
1.6.2	Enunciados	49
2	Interpolación polinómica	55
2.1	Introducción	55
2.2	Interpolación polinómica	55
2.2.1	Fórmula de Lagrange	56
2.2.2	Métodos de Aitken y Neville	56
2.2.3	Fórmulas de Newton. Diferencias divididas	58
2.2.4	Error en la interpolación polinómica	60
2.2.5	Elección de los nodos. Polinomios de Chebishev	61
2.2.6	Diferencias finitas. Fórmulas de Newton	63
2.2.7	Fenómeno Runge y teorema de Faber	65
2.2.8	Interpolación de Hermite	66
2.3	Interpolación por splines cúbicas	66
2.3.1	Cálculo efectivo	67
2.3.2	Curvatura mínima de las splines cúbicas	69
2.4	Problemas	71
2.5	Prácticas	74
2.5.1	Práctica ejemplo	74
2.5.2	Enunciados	77

3 Aproximación de funciones	81
3.1 Introducción	81
3.1.1 Fundamentos teóricos	82
3.1.2 Ortogonalización de Gram-Schmidt. Familias ortogonales	85
3.2 Aproximación mínimo-cuadrática polinómica	88
3.2.1 Aproximación continua por polinomios ortogonales	88
3.2.2 Aproximación discreta	90
3.3 Aproximación min-max de funciones	91
3.3.1 Aproximación polinómica continua	92
3.3.2 Aproximación polinómica discreta	93
3.4 Problemas	95
3.5 Prácticas	96
3.5.1 Práctica ejemplo	96
3.5.2 Enunciados	103
4 Sistemas Lineales	105
4.1 Introducción	105
4.2 Métodos directos	106
4.2.1 Eliminación gaussiana	106
4.2.2 Factorización LU	108
4.2.3 Métodos compactos	110
4.2.4 Cálculo de inversas	112
4.2.5 Cotas de error	113
4.3 Métodos iterativos	118
4.3.1 Método de Jacobi	120
4.3.2 Método de Gauss-Seidel	121
4.3.3 Razón de convergencia y estimación del error	122

4.3.4	Refinamiento iterativo de la solución	122
4.3.5	Métodos de sobrerelajación	124
4.4	Sistemas lineales sobre determinados	127
4.5	Problemas	129
4.6	Prácticas	132
4.6.1	Práctica ejemplo	132
4.6.2	Enunciados	138
5	Derivación e integración numérica	141
5.1	Introducción	141
5.2	Derivación interpolatoria	141
5.3	Extrapolación de Richardson	144
5.4	Integración numérica	145
5.4.1	Fórmulas de Newton–Côtes	146
5.4.2	Método de Romberg	149
5.4.3	Elección del paso de integración	151
5.4.4	Integrales impropias	152
5.5	Integración gaussiana	153
5.5.1	Gauss-Legendre	155
5.5.2	Integración gaussiana con peso	158
5.6	Problemas	162
5.7	Prácticas	165
5.7.1	Práctica ejemplo	165
5.7.2	Enunciados	172

6 Ceros de funciones no lineales	175
6.1 Introducción	175
6.2 Métodos de intervalos encajados	176
6.2.1 Método de la bisección	176
6.2.2 Método de la Regula-Falsi	177
6.3 Métodos iterativos	178
6.3.1 Método de Newton	178
6.3.2 Método de la secante	179
6.3.3 Métodos iterativos o del punto fijo	179
6.4 Orden de convergencia	182
6.5 Aceleración de la convergencia	185
6.6 Métodos de interpolación y de Taylor	186
6.7 Eficiencia de un método iterativo	188
6.8 Ceros múltiples	190
6.9 Sistemas no lineales	191
6.9.1 Método de iteración simple	191
6.9.2 Método de Newton	194
6.9.3 Métodos de continuación	196
6.10 Cálculo de las raíces de polinomios	198
6.10.1 Relación entre raíces y coeficientes	199
6.10.2 Acotación de las raíces	199
6.10.3 Separación de las raíces	200
6.10.4 Método de Newton modificado y deflación	202
6.10.5 Método de Laguerre	204
6.10.6 Método de Bairstow	206
6.11 Problemas	209

6.12 Prácticas	212
6.12.1 Práctica ejemplo	212
6.12.2 Enunciados	217
7 Valores y vectores propios	223
7.1 Introducción	223
7.2 Cotas de los valores propios	224
7.3 Transformación de matrices a forma reducida	226
7.3.1 Método de Givens	227
7.3.2 Método de Householder	229
7.3.3 Comparación de los dos métodos	232
7.4 Métodos basados en el polinomio característico	232
7.4.1 Valores y vectores propios para matrices tridiagonales simétricas	232
7.5 Métodos iterativos	235
7.5.1 Métodos de la potencia	235
7.5.2 Métodos de deflación	239
7.6 Método de Jacobi	242
7.7 Métodos de factorización	245
7.7.1 Método LR	246
7.7.2 Método QR	249
7.7.3 Traslación respecto del origen	253
7.8 Problemas	262
7.9 Prácticas	264
7.9.1 Práctica ejemplo	264
7.9.2 Enunciados	269

8 Ecuaciones diferenciales ordinarias	275
8.1 Introducción	275
8.2 Ecuaciones en diferencias	275
8.2.1 Definiciones y conceptos básicos	275
8.2.2 Ecuaciones en diferencias lineales con coeficientes constantes	276
8.3 Problema de valores iniciales	280
8.3.1 Familias de métodos	280
8.3.2 Errores, convergencia, consistencia, orden y estabilidad	283
8.3.3 Métodos lineales multipaso. Teorema de Dahlquist	287
8.3.4 Estabilidad absoluta	289
8.3.5 Ejemplos numéricos	292
8.3.6 Métodos predictor-corrector	296
8.3.7 Métodos Runge-Kutta	300
8.3.8 Comparación entre los métodos predictor-corrector y Runge-Kutta	304
8.4 Problema de valores frontera	304
8.4.1 Método del tiro simple	305
8.4.2 Método del tiro paralelo	307
8.5 Problemas	310
8.6 Prácticas	313
8.6.1 Práctica Ejemplo	313
8.6.2 Enunciados	324
A Álgebra matricial	335
A.1 Tipos de matrices	335
A.2 La forma normal de Jordan	336
A.3 Factorización de matrices	337
A.3.1 Descomposición en valores singulares	338
A.4 Normas matriciales	338

Solucionario	341
Bibliografía	347
Listado de rutinas	353
Glosario de símbolos	355
Índice	357

1 Preliminares

1.1 Introducción

El principal objetivo del Análisis Numérico consiste en suministrar métodos efectivos a fin de resolver problemas. El hecho de que los métodos numéricos puedan ser implementados en un ordenador digital hacen de éste un instrumento esencial en los estudios numéricos actuales. Así, la utilización del ordenador para resolver problemas científicos y de ingeniería es cada vez más grande, y se consiguen soluciones de modelos matemáticos que representan situaciones reales concretas.

1.1.1 Cálculo científico y campos de aplicación

El cálculo científico consiste en el conjunto de herramientas, técnicas y teorías necesarias que resuelven, actualmente con el ordenador, modelos matemáticos de problemas científicos y de ingeniería. Muchas de estas herramientas han sido desarrolladas mucho antes de la era de los ordenadores electrónicos; por ejemplo, los logaritmos y antilogaritmos de Napier, Briggs y Bürgi, la construcción de tablas trigonométricas (fundamentales para la navegación de la época) y la interpolación de valores: Harriot, Kepler y Newton entre otros ([Gol73]).

Los campos de aplicación, entre otros, son:

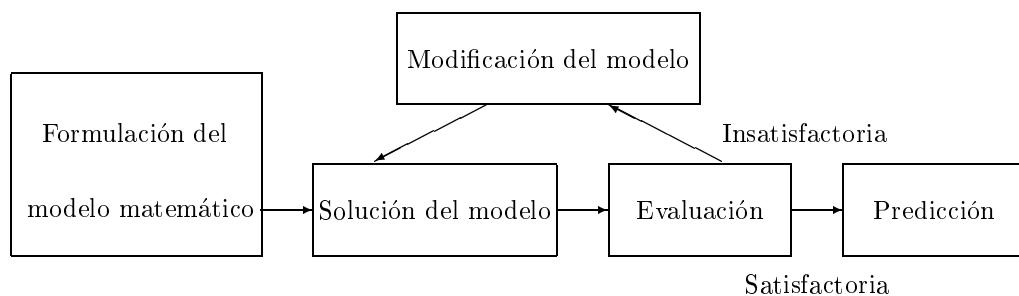
- ★ La astrodinámica: – Cálculo de trayectorias de satélites.
- ★ La mecánica celeste: – Estudio del movimiento de los astros considerando las perturbaciones creadas por sus vecinos.
- ★ La astrofísica: – Modelado de la evolución de las estrellas.
- ★ La ingeniería civil: – Estudio de las características estructurales de grandes construcciones (edificios, puentes, presas, etc.).
- ★ La meteorología: – Predicción del tiempo. Cambios climáticos.

- ★ La biología:
 - Dinámica de poblaciones.
 - Flujo de la sangre en el cuerpo humano.

- ★ Mecánica de fluidos:
 - Simulación del flujo de aire alrededor de una nave y las correspondientes presiones sobre la estructura.
 - Dispersión de contaminantes en diferentes medios.

1.1.2 Modelado matemático y solución numérica

Frente a un hecho real, el modelado matemático consiste en construir una colección de fórmulas y ecuaciones matemáticas que lo representen de la forma más fiel posible, de manera que permita realizar predicciones correctas. El método que se sigue es ([Ort70]):



Debido, muchas veces, a la complejidad del modelo, es necesario emplear técnicas numéricas para encontrar las aproximaciones de la solución del modelo. En este proceso hay que tener en cuenta:

- El hardware y el software que se utiliza.
- La representación numérica no será todo \mathbf{R} . El conjunto de números representables en un ordenador es finito (en realidad, es un subconjunto de \mathbf{Q}). Muchos números no serán nunca representados con exactitud; este hecho da lugar al **error de redondeo**.
- El número de operaciones a poder efectuar es finito, y este hecho obliga a substituir problemas continuos por discretos; esta situación tiene como consecuencia el **error de discretización**, también llamado de **truncamiento**.
- Los métodos que usaremos son iterativos y tienden a la solución; por tanto, convendrá pararlos tomando el último valor como solución aproximada, y esta situación produce el **error de convergencia** y/o el **error de truncamiento**.
- La eficiencia del método utilizado; es decir, el número de operaciones a emplear y cómo se realizan. Mientras que un sistema lineal compatible determinado tiene solución única y se puede resolver por el método de Cramer, que es impecable formalmente, si el sistema tiene

30 ecuaciones y 30 incógnitas, es más conveniente ir pensando en un método numérico, ya que el cálculo de los determinantes es muy costoso: ($25 \cdot 10^{33}$ operaciones aritméticas aproximadamente) y produce grandes errores.

Ejemplo de resolución de un problema: Deformación de una viga

La deformación, debida al propio peso, de una viga uniforme apoyada por sus dos extremos, viene definida por la ecuación diferencial siguiente (*modelo matemático*):

$$\frac{EIy''}{(1+y'^2)^{3/2}} = M(x)$$

donde $y(x)$ representa la deformación vertical en cada punto x de su longitud, E es el módulo de Young, I el momento de inercia de una sección de la viga respecto a su eje central, ambos son valores característicos del material, y $M(x)$ es el momento flector, que es la suma algebraica de los momentos de todas las fuerzas que actúan en un lado de x .

Si suponemos que la flexión es pequeña, $y'(x)$ también lo es y, por lo tanto, podemos aproximar la ecuación anterior por la siguiente (*modelo matemático aproximado*): $EIy'' = M(x)$

Si la viga tiene un peso de 60 Kg y una longitud de 5 m , el momento flector es

$$M(x) = \frac{1}{2} \frac{60}{5} x^2 - \frac{1}{2} 60x = 6x(x - 5).$$

Su módulo de Young vale $59.1716\text{E}08 \text{ Kg/m}^2$, y el momento de inercia 12.34 cm^4 . Por tanto se obtiene finalmente la ecuación siguiente: $730y'' = 6x(x - 5)$

Su solución con las condiciones frontera $y(0) = y(5) = 0$ (no hay deformación en los extremos) es (*solución exacta del modelo matemático aproximado*):

$$y(x) = \frac{1}{1460}(x^4 - 10x^3 + 125x).$$

En la figura 1.1 se ha representado esta solución evaluada en once puntos igualmente espaciados.

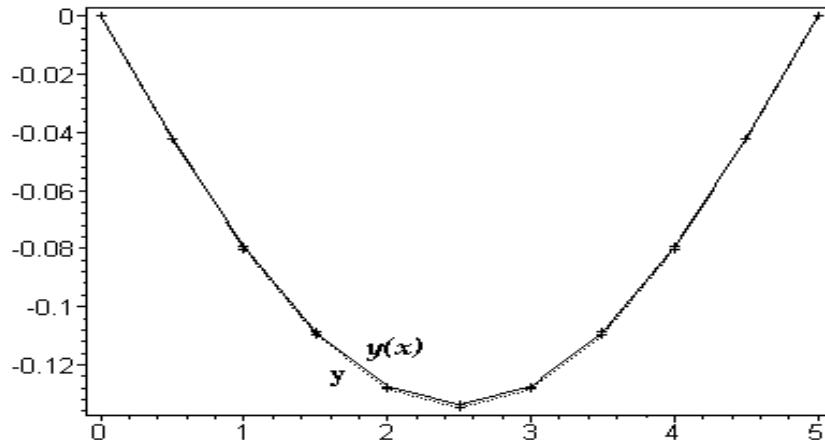


Fig. 1.1 Solución exacta y solución aproximada.

Cuando no se puede encontrar la solución analítica, es necesario integrar numéricamente la ecuación diferencial anterior. Se discretiza el intervalo $[0, 5]$ en n partes iguales (por ejemplo $n = 10$), $0 = x_0 < x_1 < \dots < x_{10} < x_{11} = 5$, $x_i = ih$, $h = 0.5$. Se aproxima la segunda derivada de y en un punto x_i por (ver el capítulo 5):

$$y''(x_i) \approx \frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2}$$

Se define $y_i = y(x_i)$ y se aplica la aproximación anterior para cada punto de la partición, y se obtiene el sistema de ecuaciones lineales siguiente:

$$\begin{aligned} 0 - 2y_1 + y_2 &= \frac{h^2}{730} 6x_1 (x_1 - 5) \\ y_{i-1} - 2y_i + y_{i+1} &= \frac{h^2}{730} 6x_i (x_i - 5) \quad \text{para } i = 2 \div 8 \\ y_8 - 2y_9 + 0 &= \frac{h^2}{730} 6x_9 (x_9 - 5) \end{aligned}$$

que, escrito en forma matricial, es (*ecuación de la solución aproximada*):

$$A \mathbf{y} = \frac{3}{1460} \mathbf{b}$$

con

$$A = \begin{pmatrix} -2 & 1 & & & & & & & \\ 1 & -2 & 1 & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & 1 & -2 & 1 & & & & \\ & & & 1 & -2 & 1 & & & \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_9 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} x_1(x_1 - 5) \\ \vdots \\ x_i(x_i - 5) \\ \vdots \\ x_9(x_9 - 5) \end{pmatrix}$$

En la figura 1.1 hay dibujada la solución de este sistema de ecuaciones lineales; comparadla con la solución exacta.

Si para la resolución de este sistema de ecuaciones lineales se aplica un método iterativo (ver el capítulo 3), que consiste en calcular iterativamente un vector \mathbf{y}^{k+1} en función del anterior \mathbf{y}^k y tal que esta sucesión de vectores converja al vector solución, finalmente se obtienen las fórmulas para calcular las componentes del vector iterado (*solución aproximada calculada*):

$$\begin{aligned} y_1^{k+1} &= \frac{1}{2}(y_2^k - \frac{3}{1460}b_1) \\ y_i^{k+1} &= \frac{1}{2}(y_{i-1}^k + y_{i+1}^k - \frac{3}{1460}b_i) \quad \text{para } i = 2 \div 8 \\ y_9^{k+1} &= \frac{1}{2}(y_8^k - \frac{3}{1460}b_9) \end{aligned}$$

En la figura 1.2 se puede observar la convergencia de este método hacia la solución exacta del sistema de ecuaciones lineales.

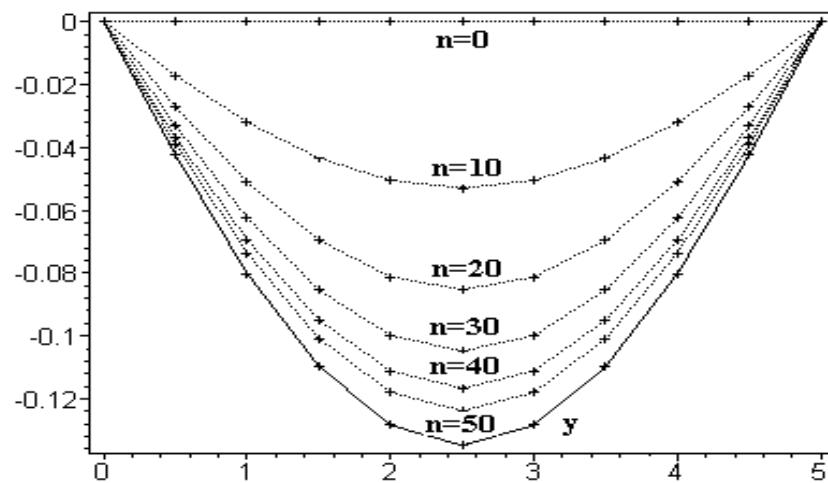


Fig. 1.2 Convergencia del método iterativo.

En la figura 1.3 se observan las tres soluciones: la exacta del modelo, la exacta de la solución aproximada y la solución calculada.

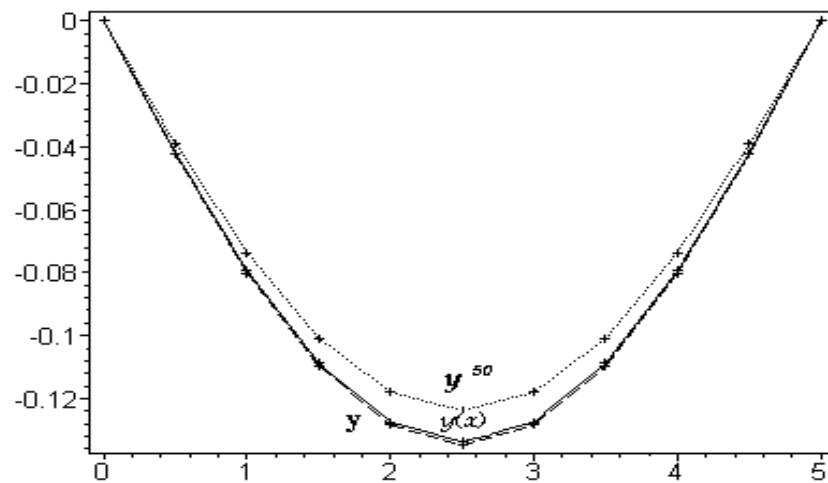


Fig. 1.3 Soluciones exacta, aproximada y calculada.

1.1.3 Fuentes de error

1. Debido a simplificaciones del fenómeno real y al empleo de un modelo poco preciso o erróneo se producirán: **errores del problema**.

2. A causa de la introducción de parámetros que son determinados aproximadamente y de constantes físicas, se tendrá un **error inicial o de medida**.
3. Los errores asociados con el sistema de representación numérica se llaman **errores de redondeo**.
4. La discretización de un proceso y tomar como solución el resultado obtenido da lugar a **errores de truncamiento**.
5. Entonces, cuando se tenga bien definido el problema, el método numérico que se utilice para resolverlo producirá errores, casi todos ellos inevitables y que pueden depender del método de resolución elegido: **errores humanos**.

Ejemplo de algoritmo numéricamente inestable

Se quiere calcular las integrales ([For77], [Var62]):

$$E_n = \int_0^1 x^n e^{x-1} dx \quad n = 1, 2, \dots$$

y se observa que $E_n = 1 - n E_{n-1}$, $n = 2, 3, \dots$ donde $E_1 = 1/e$. Si se toman 6 cifras en la representación numérica, en particular $1/e = 0.367879$, y se aplica la recurrencia, se obtiene la sucesión siguiente:

$E_1 \approx 0.367879$	$E_6 \approx 0.127120$
$E_2 \approx 0.264242$	$E_7 \approx 0.110160$
$E_3 \approx 0.207274$	$E_8 \approx 0.118720$
$E_4 \approx 0.170904$	$E_9 \approx -0.0684800$
$E_5 \approx 0.145480$	

¡Imposible! La integral de funciones positivas no puede dar nunca un valor negativo. El error de redondeo ha ido creciendo en cada paso (se ha ido multiplicando por n). El algoritmo es numéricamente inestable.

Mientras que, si se toma el algoritmo

$$E_{n-1} = \frac{1 - E_n}{n} \quad n = \dots, 3, 2$$

se obtiene

$E_{20} \approx 0.0$	$E_{14} \approx 0.0627322$
$E_{19} \approx 0.0500000$	$E_{13} \approx 0.0669477$
$E_{18} \approx 0.0500000$	$E_{12} \approx 0.0717733$
$E_{17} \approx 0.0527778$	$E_{11} \approx 0.0773523$
$E_{16} \approx 0.0557190$	$E_{10} \approx 0.0838771$
$E_{15} \approx 0.0590176$	$E_9 \approx 0.0916123$

Exacto en todas las cifras representadas, ya que, en cada paso, el error en la representación inicial de E_{20} se ha ido dividiendo por n .

Ejercicio. Se quiere calcular la integral

$$I_{20} = \int_0^1 x^{20} \sin \pi x \, dx$$

de forma recurrente $I_k = (1/\pi) - [k(k-1)/\pi^2]I_{k-2}$, $k = 2, 4, 6, \dots, 20$.

Discutir la estabilidad de la recurrencia.

Ejemplos de problemas sensibles a las condiciones iniciales

Muchos problemas son especialmente sensibles a los valores iniciales, independientemente de los errores de redondeo y del algoritmo empleado. Se presentan dos ejemplos:

a. Las raíces del polinomio ([Wil71]):

$$p(x) = \prod_{n=1}^{20} (x - n) = x^{20} - 210x^{19} + \dots + 20!$$

son $1, 2, 3, \dots, 20$. Si se modifica ligeramente **uno** de sus coeficientes

$$q(x) = p(x) + 2^{-23}x^{19}$$

las raíces de $q(x)$ se convierten en

1.000000000	$10.095266145 \pm 0.643500904i$
2.000000000	$11.793633881 \pm 1.652329728i$
3.000000000	$13.992358137 \pm 2.518830070i$
4.000000000	$16.730737466 \pm 2.812624894i$
4.999999928	$19.502439400 \pm 1.940330347i$
6.000006944	
6.999697234	
8.007267603	
8.917250249	
20.846908101	

Un ligera modificación, una unidad en el 31-ésimo dígito binario o aproximadamente una unidad en el noveno dígito decimal, da lugar a un cambio espectacular de las raíces del polinomio. Las partes imaginarias de las nuevas raíces son de tamaño considerable.

Si se considera el polinomio $f(z) = z^n + a_1 z^{n-1} + a_2 z^{n-2} + \dots + a_n$, con r raíz de $f(z)$, y se deriva respecto a_k se tiene

$$\left(\frac{\partial f}{\partial a_k} \right)_{z=r} = f'(r) \frac{\partial r}{\partial a_k} + r^{n-k} = 0$$

y, entonces $\frac{\partial r}{\partial a_k} = -r^{n-k} / f'(r)$. Este último resultado puede escribirse de la siguiente manera:

$$\left| \frac{\Delta r}{r} \right| = A_k \left| \frac{\Delta a_k}{a_k} \right| \quad \text{donde} \quad A_k = \left| \frac{a_k \cdot r^{n-k-1}}{f'(r)} \right|$$

y grandes valores de A_k hacen que pequeños cambios en los coeficientes a_k causen cambios apreciables en la raíz r . El hecho que el valor de A_k sea grande ocurrirá cuando r sea grande y $f'(r)$ pequeño. En este caso, se ha elegido $k = 1$ y $\Delta a_1 = 2^{-23}$. Si se toma $r = 16$,

$$A_1 = \frac{210 \cdot 16^{18}}{15! \cdot 4!} = 3.2 \cdot 10^{10}$$

Por tanto, es necesario tomar 10 dígitos de más respecto a la precisión deseada ([Fr 85]).

b. Si se considera el sistema ([Var62]):

$$\begin{cases} x + 2y = 3 \\ 0.499x + 1.001y = 1.5 \end{cases}$$

con solución $x = y = 1.0$. Si se substituye coeficiente 0.499 por 0.5, el nuevo sistema

$$\begin{cases} x + 2y = 3 \\ 0.5x + 1.001y = 1.5 \end{cases}$$

presenta la solución $x = 3$ y $y = 0$.

Se han presentado ejemplos de problemas en que, por más que se resuelvan correctamente, una pequeña variación en los datos de entrada hace que los resultados sean muy diferentes. Llamaremos a estos problemas, en principio, **problemas inestables**.

Ejercicios.

- Sea $J_n = \int_0^1 x^n (x^2 + 10x + 16)^{-1} dx$. Entonces, $J_0 = (1/6) \ln(4/3) = 0.047947$ y $J_1 = 0.5 \cdot \ln(27/16) - 5J_0 = 0.021889$. Comprobar que $J_{n+1} + 10 J_n + 16 J_{n-1} = 1/n$ y calcular J_n , $n = 2 \div 7$. Son fiables los resultados obtenidos?

2. Se quiere calcular la matriz inversa de la matriz de Hilbert $H = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix}$. A tal fin se considera la matriz $\mathcal{H}_\delta = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 + \delta \end{pmatrix}$ con $\delta \neq -1/12$.
- Calcular exactamente \mathcal{H}_δ^{-1} , y tomad $\delta = 0$.
 - Se considera $H_2 = \begin{pmatrix} 0.10 \cdot 10^1 & 0.50 \cdot 10^0 \\ 0.50 \cdot 10^0 & 0.33 \cdot 10^0 \end{pmatrix}$ que es \mathcal{H}_δ con $\delta = 0$ y una aritmética con dos dígitos de mantisa. Comprobar que $H_2^{-1} = \begin{pmatrix} 4.2 & -6.5 \\ -6.3 & 13 \end{pmatrix}$.

1.1.4 Estabilidad de algoritmos

Existen problemas inestables y algoritmos inestables. Un error frecuente es probar un algoritmo nuevo aplicándolo a un problema inestable. Si el método introduce pequeños errores de redondeo o truncamiento, la inestabilidad del problema producirá grandes errores. Estos errores no pueden ser inculpados al método; de hecho, el algoritmo pudo funcionar muy bien cuando se aplica a problemas estables.

Para una primera definición de algoritmo estable, se comparará la ‘solución exacta’ \mathbf{y} con la ‘solución calculada’ $\tilde{\mathbf{y}}$:

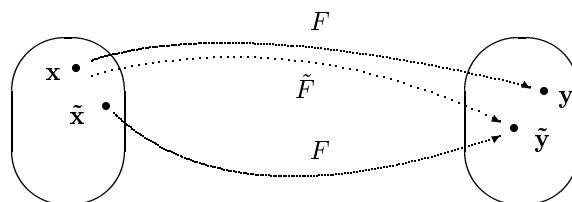
$$\begin{aligned} \mathbf{y} &= F(\mathbf{x}), \text{ solución exacta para los datos } \mathbf{x} \\ \tilde{\mathbf{y}} &= \tilde{F}(\mathbf{x}), \text{ solución calculada para los datos } \mathbf{x} \end{aligned}$$

El valor $|\mathbf{y} - \tilde{\mathbf{y}}|$ se llama **error hacia delante**. Un análisis del error hacia delante de un algoritmo es precisamente el valor $|\mathbf{y} - \tilde{\mathbf{y}}|$, pero este valor puede engañar si se tiene un problema inestable.

Otro tipo de análisis de la estabilidad de un algoritmo es el llamado análisis del **error hacia atrás**, que es independiente de la estabilidad del problema y, por lo tanto, nos indica si el método es estable.

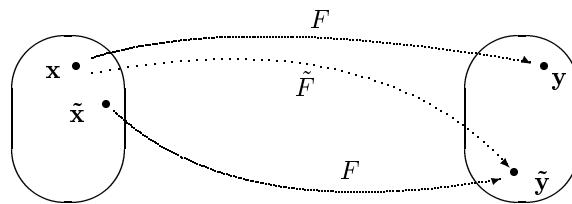
Se define $\tilde{\mathbf{x}}$ como los datos tales que $\tilde{\mathbf{y}} = F(\tilde{\mathbf{x}})$; es decir, la ‘solución calculada’, $\tilde{\mathbf{y}}$, es considerada como la ‘solución exacta’ de los datos iniciales bajo una pequeña perturbación. Se llamará al valor $|\mathbf{x} - \tilde{\mathbf{x}}|$ error hacia atrás, y el algoritmo es **estable** si este error hacia atrás es suficientemente pequeño. Algunas situaciones posibles son ([Var62]):

1. Algoritmo estable aplicado a problema estable:



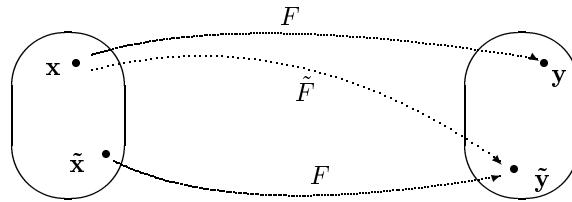
La solución calculada es próxima a la solución exacta.

2. Algoritmo estable aplicado a problema inestable:



La solución calculada no es próxima a la solución exacta.

3. Algoritmo inestable aplicado a un problema:



La inestabilidad del método se ve en el error hacia atrás (gran perturbación del problema).

Ejercicios.

1. Se considera el sistema lineal, $A\mathbf{x} = \mathbf{b}$, $\begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3.0001 \end{pmatrix}$, con solución única $\mathbf{x} = (1, 1)^T$.

Si se perturba la matriz del sistema: $\tilde{A} = \begin{pmatrix} 1 & 2 \\ 0.9999 & 2 \end{pmatrix}$, y se calcula la solución con una aritmética de 5 cifras, se obtiene $\tilde{\mathbf{x}} = (-1.0000, 2.0000)^T$. Si se perturba el vector del término independiente: $\begin{pmatrix} 3.00001 \\ 3.00003 \end{pmatrix}$, y se encuentra la solución con una aritmética de 7 cifras, se obtiene $\bar{\mathbf{x}} = (0.2000000, 1.400005)^T$. Estudiar el hecho, clasificando el algoritmo y el problema a partir de los datos que se han obtenido.

2. Resolver el sistema de ecuaciones lineales siguiente:

$$\begin{cases} 1.000x + 2.000y = 3.000 \\ 2.999x + 6.001y = 9.000 \end{cases}$$

- a) exactamente;
- b) por Cramer, utilizando cuatro cifras en todos los cálculos.
- c) Estudiar el error hacia delante y atrás.

1.2 Representación aritmética en coma flotante

1.2.1 Conjunto de números en coma flotante

Un conjunto de números, $F = F(\beta, t, L, U)$, en coma flotante se caracteriza por

- la base β , $\beta \in \mathbf{N}$, $\beta \geq 2$
- la precisión t (número de dígitos representados)
- la extensión del exponente: $e \in [L, U]$

Entonces, si $x \in F$,

$$x = \pm \left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_t}{\beta^t} \right) \cdot \beta^e$$

con $d_i \in \mathbf{N}$, $0 \leq d_i < \beta$, $i = 1 \div t$; llamaremos **fracción** f a la cantidad $\pm \left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_t}{\beta^t} \right)$.

Si cuando $x \neq 0$, $d_1 \neq 0$, se dice que F está **normalizado**; entonces $m = \beta^t f \in \mathbf{Z}$ y se llama **mantisa**. Los conjuntos en coma flotante se considerarán siempre normalizados. Así se obtiene $x = f \beta^e$ con $\beta^{-1} \leq |f| < 1$. El número de elementos representables con esta aritmética es

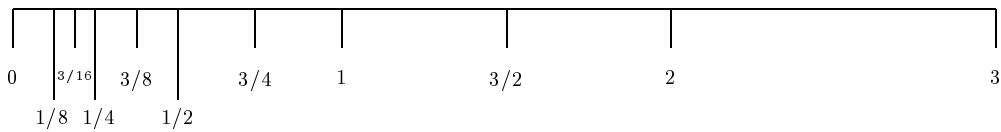
$$\text{card } F = 2(\beta - 1) \beta^{t-1} (U - L + 1) + 1$$

Si algún resultado en los cálculos en una aritmética de coma flotante resulta que $e > U$, se ha cometido un error de overflow; en el caso de que $e < L$, se comete un error de underflow.

La distribución de F sobre la recta real no es uniforme. Si se toma, por ejemplo, $F(\beta, t, L, U) = F(2, 2, -2, 2)$, se tiene:

m	e	-2	-1	0	1	2
1 0		1/8	1/4	1/2	1	2
1 1		3/16	3/8	3/4	3/2	3

y su representación gráfica es:



1.2.2 La función fl , redondeo y truncamiento

Si $x \in \mathbf{R}$, se define $fl(x)$ como el elemento de F más próximo a x . Entonces, si $a, b \in F$ se define

$$\begin{aligned} a \oplus b &= fl(a + b) \\ a \odot b &= fl(a \cdot b) \end{aligned}$$

Estas operaciones no son, en general, asociativas y distributivas; en el conjunto $F(2, 2, -2, 2)$ se tiene:

$$\left(\frac{3}{4} \oplus \frac{3}{2}\right) \oplus \frac{3}{8} = 2$$

y, en cambio,

$$\frac{3}{4} \oplus \left(\frac{3}{2} \oplus \frac{3}{8}\right) = 3$$

Hay dos maneras de reducir el número de dígitos de una cantidad a t dígitos: **redondear** y **truncar**.

Si se supone que $\beta = 10$ y se redondea: el t -ésimo decimal se aumenta en una unidad si el número que hay a la derecha del t -ésimo decimal es mayor que $0.5 \cdot 10^{-t}$. Si es más pequeño que $0.5 \cdot 10^{-t}$, entonces el t -ésimo decimal no cambia. En el caso límite de que la derecha del t -ésimo decimal es igual a $0.5 \cdot 10^{-t}$, exactamente, el t -ésimo decimales se aumenta en una unidad si t es impar y no cambia si es par.

Truncar quiere decir que todos los decimales a la derecha del t -ésimo se ignoran (no son representados).

Ejemplo. Redondear el número $\pi = 3.1415926535\dots$ a cinco, cuatro y tres dígitos quiere decir quedarse con las cantidades: $0.31416 \cdot 10$, $0.3142 \cdot 10$ y $0.314 \cdot 10$. En cambio, si se truncan se obtiene: $0.31415 \cdot 10$, $0.3141 \cdot 10$ y $0.314 \cdot 10$.

1.2.3 Épsilon de la máquina

La precisión de una aritmética de coma flotante se caracteriza por el **épsilon de la máquina**, ϵ , que se define como el número más pequeño de F que cumple $1 \oplus \epsilon > 1$. Siguiendo con el ejemplo anterior, $\epsilon = 1/4$ y se puede decir que el épsilon de la máquina no es, en general, el número más pequeño representable, pero da una medida relativa de hasta dónde dos números muy próximos serán diferentes: si $h > 0$, $x > 0$ y h es el número más pequeño representable tal que $x + h > x$, entonces $1 + \frac{h}{x} > 1$ y, por tanto, $\frac{h}{x} = \epsilon$.

En el conjunto $F(\beta, t, L, U)$, si t es par, entonces $\epsilon = \frac{1}{2}\beta^{-t+1}$; así, cuando $\beta = 2$, $\epsilon = 2^{-t}$. Cuando t es impar, $\epsilon = \beta^{-t+1}$.

Se pueden encontrar aproximaciones al épsilon de la máquina si se ejecuta un programa del tipo siguiente:

```

EPS = 1.

10      EPS = 0.5 * EPS

EPSP1 = EPS + 1.

IF (EPSP1.GT.1.)  GO TO 10

WRITE (2,*) EPS

```

Así, en un VAX 8600, si se ejecuta este programa, se obtiene:

$$\epsilon \approx 2.9802322 \cdot 10^{-8} \text{ con precisión simple,}$$

$$\epsilon \approx 6.9388939 \cdot 10^{-18} \text{ con precisión doble y}$$

$$\epsilon \approx 4.8148249 \cdot 10^{-35} \text{ con precisión cuádruple.}$$

La precisión de un ordenador dependerá del fabricante y del tipo de variables que se definan; la unidad de información viene dada por el número de dígitos binarios o longitud de la palabra (word):

16 para muchos microprocesadores,

32 para IBM 3090, VAX y

64 para superordenadores.

El 80i87 Intel, en su representación interna temporal, utiliza 80 bits: 64 mantisa + 15 exponente + 1 signo.

El VAX 8600 para una variable real de doble precisión ocupa 8 bytes consecutivos: 56 bits para la mantisa + 8 bits para el exponente + 1 bit para el signo. En realidad, como es un elemento de $F(2, 53, -127, 127)$ normalizado, no se representa $d_1 = 1$; el exponente se almacena con un exceso de 128 y, en la práctica $e \in [1, 255]$, en lugar de $[-127, 127]$ (tipo D_floating).

Cuando se habla de doble precisión para variables en coma flotante, y se programa en Fortran o C, generalmente se quiere decir que estos valores numéricos se almacenan ocupando el lugar de dos palabras.

Ejemplo de error de redondeo

Se quiere calcular $e^{-5.5}$ a partir del desarrollo $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$.

Si se trabaja con una aritmética de 5 cifras ([For77]), se obtiene

$$\begin{aligned}
 e^{-5.5} = & \quad 1. \quad 0000 \\
 & - \quad 5. \quad 5000 \\
 & + \quad 15. \quad 125 \\
 & - \quad 27. \quad 730 \\
 & + \quad 38. \quad 129 \\
 & - \quad 41. \quad 942 \\
 & + \quad 38. \quad 446 \\
 & - \quad 30. \quad 208 \\
 & + \quad 20. \quad 768 \\
 & - \quad 12. \quad 692 \\
 & + \quad 6. \quad 9803 \\
 & - \quad 3. \quad 4902 \\
 & + \quad 1. \quad 5997 \\
 & \vdots \\
 \hline
 & + \quad 0. \quad 0026363
 \end{aligned}$$

cálculo realizado con 25 términos de la serie. Debido a los términos que ocupan posiciones enteras, la representación de los decimales que darían realmente una contribución efectiva no se realiza, ya que se trabaja con sólo 5 cifras.

El resultado exacto es $e^{-5.5} = 0.00408677$ que se puede calcular, en este caso, considerando

$$e^{-5.5} = \frac{1}{1.0000 + 5.0000 + 15.125 + \dots} = 0.0040865$$

El error en el resultado se ha reducido en un 0.007 %. Destaca en este ejemplo la importancia de la longitud de los registros, si no se quiere caer en errores de redondeo que dan lugar a la situación anterior, que se llama **cancelación catastrófica**.

1.3 Análisis del error

1.3.1 Definiciones

Sea x un valor exacto y \bar{x} una aproximación de x .

Se define el **error absoluto** de \bar{x} por $\Delta x = x - \bar{x}$

el **error relativo** de \bar{x} por $\varepsilon_x = \frac{\Delta x}{x}$, ($x \neq 0$)

Ejemplo. Si $x = \sqrt{2}$, $\bar{x} = 1.414$, $\left\{ \begin{array}{l} \Delta x = 1.414 - \sqrt{2} = -0.0002135\dots \\ \frac{\Delta x}{x} = \frac{-0.0002135\dots}{\sqrt{2}} \end{array} \right.$

En muchos casos sólo se podrá conocer una cota superior del error absoluto de la aproximación; así,

$$|\Delta x| \leq 0.22 \cdot 10^{-3}, \quad x = 1.414 \pm 0.22 \cdot 10^{-3} \quad \text{y} \quad 1.41378 \leq x \leq 1.41422$$

son tres formas equivalentes de expresar una misma cota.

Cuando un número se redondea a t decimales, se produce un error de $0.5 \cdot 10^{-t}$ como máximo, mientras que el error de truncamiento puede llegar a 10^{-t} . Los errores de truncamiento son sistemáticos: el cálculo aproximado siempre es más pequeño que el exacto. Cuando un valor aproximado se redondea o trunca, el error que se produce se ha de añadir a la cota de error que ya tenía.

Ejemplo. Sea $b = 2.3514 \pm 0.2$. Si se redondea a un decimal $b_a = 2.4$. El error

$$|b_a - \bar{b}| = |2.4 - 2.3514| = 0.0486 < 0.05$$

se tendrá que añadir a la cota del error: $b = 2.4 \pm 0.25$.

Se definirán ahora conceptos relacionados con los errores absoluto y relativo:

Se dice que \bar{a} tiene (o presenta) **t decimales correctos** si $|\Delta a| \leq 0.5 \cdot 10^{-t}$.

Si \bar{a} tiene t decimales correctos, todos los dígitos de la mantisa en base 10 de \bar{a} (que no sean ceros que necesita el punto decimal) de exponente mayor o igual que 10^{-t} , se llaman **cifras significativas**.

Ejemplo.

Aproximaciones con cota de error	Decimales correctos	Cifras significativas
$0.001234 \pm 0.5 \cdot 10^{-5}$	5	3
$56.789 \pm 0.5 \cdot 10^{-3}$	3	5
210000 ± 5000		2

1.3.2 Errores de representación

Sea el conjunto $F(\beta, t, L, U)$ y $x \in \mathbf{R}$; si se redondea la expresión exacta $x = f \beta^e$, $\beta^{-1} \leq |f| < 1$, $e \in [L, U]$, se transforma en $x_a = f_a \beta^e$, donde f_a es el redondeo de f a t dígitos; se tiene

$$|f_a - f| \leq \frac{1}{2} \beta^{-t}$$

y una cota del error absoluto es

$$|x_a - x| \leq \frac{1}{2} \beta^{-t} \beta^e$$

Una estimación del error relativo viene dada por

$$\frac{|x_a - x|}{|x|} \leq \frac{\frac{1}{2} \beta^{-t} \beta^e}{|f| \beta^e} = \frac{\frac{1}{2} \beta^{-t}}{|f|}$$

y de la condición de normalización $|f| \geq \beta^{-1}$, se obtiene

$$\frac{|x_a - x|}{|x|} \leq \frac{1}{2} \beta^{-t+1} = \mu$$

La cantidad μ recibe el nombre de **unidad de redondeo** y es independiente de la magnitud del número que se está representando. Una notación equivalente en la expresión del error relativo es $\exists \delta > 0, |\delta| \leq \mu$, tal que

$$x_a = x(1 + \delta)$$

Ejemplo. El procesador de Intel tiene una unidad de redondeo que vale

$$\mu = \frac{1}{2} 2^{-64+1} = 2^{-64} \approx 0.5421 \cdot 10^{-21}$$

De forma análoga se puede deducir la **unidad de truncamiento** y vale $\mu_T = 2\mu = \beta^{-t+1}$.

1.3.3 Errores en las operaciones en coma flotante

Si se multiplican dos números con t dígitos, se obtienen resultados con $2t$ o $2t - 1$ dígitos. La aritmética de coma flotante que se utiliza se define en el conjunto $F(\beta, t, L, U)$, y se supone que la longitud de los registros donde se efectúan las operaciones aritméticas tienen longitud $2t + 2$. Entonces, es posible soportar una aritmética con la misma precisión por lo menos (con registros más cortos también se puede hacer, pero los algoritmos son más complicados y largos). En los ejemplos se utiliza $F(10, 4, -9, 9)$.

Se presenta, en primer lugar, la suma (la diferencia es lo mismo considerando: $x - y = x + (-y)$).

```

SUMA           z := x + y ;

acción suma (mx, ex, my, ey, mz, ez)
    enter: ex, ey, ez;
    enter_t_dígitos: mx, my, mz;
    enter_(2t + 2)_dígitos: a, b;
    ez := ex;          (suponemos x ≥ y)
    si ex - ey ≥ t + 2 entonces mz := mx;
    si no
        a := mx/βt+1;
        b := my/βt+1-(e_x - e_y);
        a := a + b;
    finsi;
    Normalizar (a, ez)
    Redondear (a, ez)
    mz := a * β-(t+1);
finacción

```

Si se considera $x = f_x \beta^{e_x}$, $y = f_y \beta^{e_y}$ y $x \geq y$; sea $z = fl(x+y)$. Para hacer el cálculo se ha de desplazar y $e_x - e_y$ posiciones a la derecha (es decir, igualar exponentes con x normalizado) y sumar:

$$0.1234 \cdot 10^1 + 0.4567 \cdot 10^{-1} = (0.1234 + 0.004567) \cdot 10^1 = 0.127967 \cdot 10^1$$

Pero si se da la circunstancia que $e_x - e_y \geq t+2$, entonces $fl(x+y) = x$. Así, por ejemplo,

$$0.1234 \cdot 10^1 + 0.3258 \cdot 10^{-5} = 0.1234003258 \cdot 10^1 \approx 0.1234 \cdot 10^1$$

Todas las operaciones se realizan con números enteros: $m_x = f_x \beta^{2t+1}$; $m_y = f_y \beta^{2t+1}$.

Si $e_x - e_y < t+2$, entonces m_y puede ser almacenado sin errores después del desplazamiento, ya que se ha considerado que los registros aritméticos admiten $2t+2$ dígitos; entonces la suma $m_x + m_y$ no contiene error. En general, el resultado no está normalizado y será necesario que $\beta^{-1} \leq |f| < 1$. Lo que sucede es que $|f| \geq 1$ ($0.5678 \cdot 10^2 + 0.4794 \cdot 10^2 = 1.0472 \cdot 10^2$) o $|f| < \beta^{-1}$ ($0.5678 \cdot 10^2 + (-0.5600 \cdot 10^2) = 0.0078 \cdot 10^2$). Será, por tanto, necesario un algoritmo de normalización, seguido del de redondeo:

NORMALIZACIÓN	REDONDEO
<pre> acción normalizar (m, e) enter_(2t+2)_dígitos: m ; enter_(t+1)_dígitos: e ; si m / β^{2t+1} ≠ 0 entonces m := m / β ; (desplazamiento a la e := e+1; derecha una posición) si e > U entonces 'overflow+ si no mientras m / β^{2t} = 0 hacer m := m * β; (desplazamiento a la iz- e := e - 1; quierda una posición) si e < L entonces 'underflow+ x := 0 ; finsi finmientras finsi finacción </pre>	<pre> acción redondear (m, e) enter_(2t+2)_dígitos: m ; enter_(t+1)_dígitos: e ; c := m / β^{t+1} ; m := m - c ; si (β es impar) entonces si c > (β^{t+1} - 1) / 2 entonces m := m + β^{t+1} ; finsi si no si c > β^{t+1} / 2 entonces m := m + β^{t+1} ; finsi si (c = β^{t+1} / 2 i t impar) entonces m := m + β^{t+1} ; finsi finsi si (m / β^{2t+1} ≠ 0) entonces m := m / β ; e := e + 1 ; finsi finacción </pre>

Redondear a t dígitos puede dar números no normalizados; por ejemplo:

$$0.99995 \cdot 10^4 \doteq 1.0000 \cdot 10^4 = 0.1000 \cdot 10^5$$

y, por lo tanto, es necesario volver a normalizar.

Los algoritmos de multiplicación y división son fáciles:

PRODUCTO $z := x * y ;$ acción producto $(m_x, e_x, m_y, e_y, m_z, e_z) ;$ enter: $e_x, e_y, e_z ;$ enter t -dígitos: $m_x, m_y, m_z ;$ enter $2t + 2$ -dígitos: $a ;$ $e_z := e_x + e_y ;$ $a := m_x * m_y ;$ Normalizar $(a, e_z) ;$ Redondear $(a, e_z) ;$ $m_z := a / \beta^{t+1}$ finacción	COCIENTE $z := x/y ;$ acción cociente $(m_x, e_x, m_y, e_y, m_z, e_z) ;$ enter: $e_x, e_y, e_z ;$ enter t -dígitos: $m_x, m_y, m_z ;$ enter $2t + 2$ -dígitos: $a ;$ si $m_y := 0$ entonces ‘división por cero+’; si no $e_z := e_x - e_y ;$ $a := m_x/m_y ;$ Normalizar $(a, e_z) ;$ Redondear $(a, e_z) ;$ $m_z := a / \beta^{t+1}$ finsi finacción
---	---

Se supone que los registros aritméticos admiten $2t + 2$ dígitos. De este modo los resultados de sumar y multiplicar antes de normalizar y redondear son exactos. Entonces, se tiene la misma estimación del error para las operaciones aritméticas que para la representación en coma flotante:

Teorema 1.1 Si se representa cualquiera de las operaciones $+, -, *, /$ y \bowtie , con la condición $x \bowtie y \neq 0$ y los registros tienen longitud $2t + 2$, entonces

$$\left| \frac{x \bowtie y - fl(x \bowtie y)}{x \bowtie y} \right| \leq \mu$$

o, equivalentemente, $fl(x \bowtie y) = (x \bowtie y)(1 + \delta)$, para algún $\delta > 0$ tal que $|\delta| \leq \mu$ (μ : unidad de redondeo).

Orden de convergencia

Una de las técnicas que más se utilizan en cálculo numérico es la construcción de sucesiones que tienden a la solución del problema (métodos iterativos). Se introduce una nueva terminología:

Si $\{a_n\}_{n \in \mathbb{N}}$ es una sucesión convergente hacia α , se dice que esta sucesión es de orden b_n , $O(b_n)$, si $\{b_n\}_{n \in \mathbb{N}}$ es otra sucesión convergente hacia cero con $b_n \neq 0$, $\forall n$ y

$$\frac{|a_n - \alpha|}{b_n} \leq C, \quad \text{para } n \text{ suficientemente grande,}$$

donde C es una constante independiente de n . A veces se escribe $a_n = \alpha + O(b_n)$.

Ejemplos.

$$\bullet \frac{\sin n}{n} = 0 + O\left(\frac{1}{n}\right) \quad \text{y} \quad \bullet \frac{n+3}{n^3} = 0 + O\left(\frac{1}{n^2}\right).$$

El concepto de orden se generaliza a funciones de la siguiente manera:

si $\lim_{x \rightarrow 0} F(x) = L$, se dice que la convergencia es de orden $G(x)$, $O(G(x))$,

si $\exists C > 0$, constante independiente de x , tal que

$$\frac{|F(x) - L|}{|G(x)|} \leq C, \quad \text{para } x \text{ suficientemente pequeño.}$$

También se denota por $F(x) = L + O(G(x))$.

El orden de convergencia está relacionado con la velocidad (número de iteraciones, en realidad) de convergencia de un método iterativo, y a mayor orden se consigue el resultado con un número menor de iteraciones.

Ejemplos.

$$\frac{\sin x}{x} = 1 + O(x^2), \quad \text{cuando } x \rightarrow 0$$

$$\frac{\ln(1+x)}{x} = 1 + O(x), \quad \text{cuando } x \rightarrow 0$$

$$\text{y} \quad \frac{1 - \cos x}{x} = 0 + O(x), \quad \text{cuando } x \rightarrow 0$$

1.3.4 Propagación del error

Sea $g : \mathbf{R} \rightarrow \mathbf{R}$ una función derivable y sea \bar{x} una aproximación conocida de x con una cota de error ϵ , $x = \bar{x} \pm \epsilon$. Si se aplica el teorema del valor medio, se tiene $\Delta g = g(x) - g(\bar{x}) = g'(\xi)(x - \bar{x})$. Por razones prácticas, se evalúa la derivada de la función g en \bar{x} , y se obtiene la fórmula de propagación del error:

$$|\Delta g| = |g'(\xi)| |\Delta x| \lesssim |g'(\bar{x})| \epsilon$$

El símbolo \lesssim quiere decir “más pequeño o aproximadamente igual a”.

Ejemplo. Se quiere calcular la expresión $\ln(1 - 1/e)$ y se duda entre calcularla tal como está escrita, o bien utilizando la expresión equivalente $\ln(e - 1) - 1$ con una aproximación de e igual a $\bar{e} = 2.72$.

Si se considera la fórmula de propagación del error con $g_1(x) = \ln(1 - 1/x)$ y $g_2(x) = \ln(x - 1) - 1$, como se tiene el mismo $\Delta e = e - \bar{e}$, será suficiente calcular las derivadas de g_1 y g_2 :

$$g'_1(x) = \frac{1}{x(x-1)} \quad g'_2(x) = \frac{1}{x-1}$$

y, por lo tanto, como $g'_1(x) < g'_2(x)$ para x en un entorno de e , es más conveniente emplear la primera expresión. Si se supone que las operaciones son exactas, se tiene

$$|\Delta g_1| \lesssim |g'_1(\bar{e})| |\Delta e| = 0.213748 \cdot 0.001718 \doteq 0.367 \cdot 10^{-3}$$

Análogamente, si se considera una función real con dos variables, es igualmente posible calcular el error propagado de las operaciones elementales; sea $g : \mathbf{R}^2 \rightarrow \mathbf{R}$ una función diferenciable y \bar{x}_1 y \bar{x}_2 aproximaciones de x_1 y de x_2 con cotas de error ϵ_1 y ϵ_2 , $x_i = \bar{x}_i \pm \epsilon_i$, $i = 1, 2$. Si desarrollamos por la fórmula de Taylor, se tiene

$$g(x_1, x_2) = g(\bar{x}_1, \bar{x}_2) + \left(\frac{\partial g(\bar{x})}{\partial x_1} \right) \epsilon_1 + \left(\frac{\partial g(\bar{x})}{\partial x_2} \right) \epsilon_2 + O(\epsilon^2)$$

con $\epsilon = \|(\epsilon_1, \epsilon_2)\|$. La fórmula de la propagación del error para dos variables es

$$|\Delta g| \lesssim \left| \frac{\partial g(\bar{x})}{\partial x_1} \right| |\epsilon_1| + \left| \frac{\partial g(\bar{x})}{\partial x_2} \right| |\epsilon_2|$$

Ejercicio. Generalícese este resultado a n variables.

Si $x_i = \bar{x}_i \pm \Delta x_i$, $y = 1, 2$, entonces

* cuando $g(x_1, x_2) = x_1 \pm x_2$, $|\Delta g| \lesssim |\Delta x_1| + |\Delta x_2|$.

* cuando $g(x_1, x_2) = x_1 \cdot x_2$, $|\Delta g| \lesssim |\bar{x}_2||\Delta x_1| + |\bar{x}_1||\Delta x_2|$, y una cota del error relativo es

$$\left| \frac{\Delta g}{g} \right| \lesssim \left| \frac{\Delta x_1}{\bar{x}_1} \right| + \left| \frac{\Delta x_2}{\bar{x}_2} \right|$$

* cuando $g(x_1, x_2) = x_1/x_2$, $|\Delta g| \lesssim |1/\bar{x}_2||\Delta x_1| + |\bar{x}_1/\bar{x}_2^2||\Delta x_2|$, y se tiene la misma cota superior del error relativo que en el producto.

Algoritmos con cancelación

Cuando se usan aproximaciones en los cálculos es importante no perder información a causa de algoritmos inadecuados.

Ejemplo. La ecuación de segundo grado $x^2 - 18x + 1 = 0$ tiene las siguientes soluciones $x_{1,2} = 9 \pm \sqrt{80}$.

Si $\sqrt{80}$ se calcula con 4 decimales correctos:

$$\begin{aligned} x_1 &= 9 + 8.9443 \pm 0.5 \cdot 10^{-4} = 17.9443 \pm 0.5 \cdot 10^{-4} \\ x_2 &= 9 - 8.9443 \pm 0.5 \cdot 10^{-4} = 0.0557 \pm 0.5 \cdot 10^{-4} \end{aligned}$$

La primera raíz, x_1 , presenta 6 cifras significativas, mientras que la segunda, x_2 , sólo tiene 3. En este caso la cancelación puede evitarse tomando

$$x_2 = \frac{(9 - \sqrt{80})(9 + \sqrt{80})}{9 + \sqrt{80}} = \frac{1}{9 + \sqrt{80}} = \frac{1}{17.9443 \pm 0.5 \cdot 10^{-4}}$$

De este modo se consigue $1/17.9443 = 0.055728002$ y el error relativo de propagación del cociente para x_2 es como máximo $0.5 \cdot 10^{-4} / 17.9443 < 0.3 \cdot 10^{-5}$. El error absoluto es menor que $0.3 \cdot 10^{-5} \cdot 0.05573 < 0.17 \cdot 10^{-6}$. Redondeando a 7 decimales, $x_2 = 0.0557280 \pm 0.2 \cdot 10^{-6}$, que presenta 5 cifras significativas.

Distintas situaciones donde se presentan problemas de cancelaciones se pueden resumir en

$$g(x + \delta) - g(x) \quad \text{con } |\delta| \ll 1$$

si g admite derivadas de orden suficientemente grande en un entorno de x , se puede desarrollar la función g en el punto x en potencias de δ .

Ejercicios.

1. Discutir las raíces del polinomio $p(x) = x^2 - 2ax + \eta$, si $a > 0$ y $\eta \ll a$.
2. Calcular la raíz más pequeña de la ecuación $x^2 - 40x + 1 = 0$ de dos formas: (a) $20 - \sqrt{399}$ y (b) $(20 + \sqrt{399})^{-1}$, donde se ha tomado $\sqrt{399} \approx 19.97498$. Comparar los errores.
3. Arreglad la expresión $\sqrt{4.12345} - \sqrt{4.12335}$ de manera que con una aritmética de 5 dígitos de mantisa no se anule.
4. Expresad de forma más conveniente para el cálculo

$$\cos(x + \delta) - \cos x, \quad \text{con } |\delta| \ll 1.$$

Números de condición

Sea $g : D \rightarrow \mathbf{R}^m$, donde D es una región de \mathbf{R}^n , i g es diferenciable en un entorno del vector aproximado $\bar{\mathbf{x}}$, donde $\mathbf{x} = \bar{\mathbf{x}} \pm \Delta\mathbf{x}$, con $\Delta\mathbf{x} = (\Delta x_1, \dots, \Delta x_n)^T$. Si se aplica la fórmula de Taylor a la función $\mathbf{y} = g(\bar{\mathbf{x}} + \Delta\mathbf{x})$ y se toman los términos hasta orden 1, se tiene

$$\Delta y_i = y_i - \bar{y}_i \approx \sum_{j=1}^n \frac{\partial g^i(\bar{\mathbf{x}})}{\partial x_j} \Delta x_j \quad i = 1 \div m$$

que es la fórmula del error absoluto propagado.

Si $\bar{x}_j \neq 0$, $j = 1 \div n$, y $\bar{y}_i \neq 0$, $i = 1 \div m$, se tiene

$$\frac{\Delta y_i}{\bar{y}_i} \approx \frac{1}{\bar{y}_i} \sum_{j=1}^n \frac{\partial g^i(\bar{\mathbf{x}})}{\partial x_j} \Delta x_j = \sum_{j=1}^n \frac{\bar{x}_j}{\bar{y}_i} \frac{\Delta x_j}{\bar{x}_j} \frac{\partial g^i(\bar{\mathbf{x}})}{\partial x_j}$$

Si se approxima el error relativo en la componente x_j por $\varepsilon_{x_j} \approx \Delta x_j / \bar{x}_j$, y en la componente y_i por $\varepsilon_{y_i} \approx \Delta y_i / \bar{y}_i$ (error relativo propagado), se tiene

$$\varepsilon_{y_i} \approx \sum_{j=1}^n \frac{\bar{x}_j}{g^i(\bar{\mathbf{x}})} \frac{\partial g^i(\bar{\mathbf{x}})}{\partial x_j} \varepsilon_{x_j} \quad i = 1 \div m$$

Estos $m \cdot n$ valores

$$\frac{\bar{x}_j}{g^i(\bar{x})} \frac{\partial g^i(\bar{x})}{\partial x_j} \quad i = 1 \div m \quad j = 1 \div n$$

dan una medida de que un problema está mal condicionado: se llaman **números de condición**.

Ejemplo. Sea $y = g(p, q) = -p + \sqrt{p^2 + q}$ que es una solución de la ecuación $x^2 + 2px - q = 0$; entonces, el error relativo propagado viene dado por

$$\varepsilon_y \approx \frac{-p}{\sqrt{p^2 + q}} \varepsilon_p + \frac{p + \sqrt{p^2 + q}}{2 \sqrt{p^2 + q}} \varepsilon_q = A \varepsilon_p + B \varepsilon_q$$

Si $q > 0$, como $|A| \leq 1$ y $|B| \leq 1$, g está bien condicionada; pero si $q \approx -p^2$, el problema está mal condicionado, ya que el error puede llegar a ser muy grande a pesar de que ε_p y ε_q sean muy pequeños.

1.3.5 Acumulación del error

Si se quiere calcular con aritmética de coma flotante la suma

$$s_n = \sum_{k=1}^n x_k$$

A menudo se escribe $fl(a+b) = (a+b)(1+\varepsilon)$ con $|\varepsilon| \leq \mu$. Si se calculan las sumas parciales:

$$\begin{aligned} \tilde{s}_1 &= x_1 \\ \tilde{s}_i &= fl(\tilde{s}_{i-1} + x_i) \quad i = 2 \div n \end{aligned}$$

Entonces $\tilde{s}_i = (\tilde{s}_{i-1} + x_i)(1 + \varepsilon_i)$, $y = 2 \div n$, o también $\tilde{s}_n = \tilde{x}_1 + \dots + \tilde{x}_n$, donde $\tilde{x}_1 = x_1(1 + \varepsilon_2) \dots (1 + \varepsilon_n)$, y, en general,

$$\tilde{x}_i = x_i(1 + \varepsilon_i)(1 + \varepsilon_{i+1}) \dots (1 + \varepsilon_n) \quad i = 2 \div n$$

Para tener una estimación del error es necesario el siguiente lema, que será demostrado más adelante:

Lema 1.1 Sean $\{\varepsilon_i\}$, $i = 1 \div n$, números reales con $|\varepsilon_i| \leq \mu \forall i = 1 \div n$ y se supone que $n\mu < 0.1$. Entonces, $\exists \delta_n > 0$ tal que $(1 + \varepsilon_1) \dots (1 + \varepsilon_n) = 1 + \delta_n$, y $|\delta_n| \leq 1.06n\mu$.

El análisis del error hacia delante viene dado por el siguiente

Teorema 1.2 Si $n\mu < 0.1$, entonces

$$|\tilde{s}_n - s_n| \leq |x_1| |\delta_{n-1}| + |x_2| |\delta_{n-1}| + |x_3| |\delta_{n-3}| + \cdots + |x_n| |\delta_1|,$$

donde $|\delta_k| \leq k \cdot 1.06\mu$, $k = 1 \div n-1$.

Demostración: Del lema anterior,

$$\tilde{s}_n = x_1(1 + \delta_{n-1}) + x_2(1 + \delta_{n-1}) + x_3(1 + \delta_{n-2}) + \cdots + x_n(1 + \delta_1)$$

donde los δ_k satisfacen la desigualdad del teorema. \square

El análisis del error hacia delante es mucho más complicado para otros problemas y situaciones. Wilkinson ([Wil63], [Wil65]) introdujo para estos problemas el error hacia atrás (ver la sección Estabilidad de algoritmos) y su análisis. Aplicándolo al ejemplo de la suma ya presentado, se puede establecer el siguiente

Teorema 1.3 Si $n\mu < 0.1$, entonces se tiene

$$\tilde{s}_n = \tilde{x}_1 + \tilde{x}_2 + \cdots + \tilde{x}_n,$$

$$\begin{aligned} \tilde{x}_1 &= x_1(1 + \delta_{n-1}), \\ \tilde{x}_i &= x_i(1 + \delta_{n-i+1}), \quad i = 2 \div n, \\ |\delta_k| &\leq k \cdot 1.06\mu, \quad k = 2 \div n. \end{aligned}$$

Un resultado importante al sumar diferentes elementos es

$$|\tilde{s}_n - s_n| \leq \{(n-1)|x_1| + (n-1)|x_2| + (n-2)|x_3| + \cdots + 2|x_{n-1}| + |x_n|\} \cdot 1.06\mu$$

que asegura que para minimizar la cota del error se ha de sumar en orden creciente en valor absoluto, es decir, de pequeño a grande.

Ejercicio. Sumar de la mejor manera posible 9999, 999, 99 y 9 con una aritmética decimal de coma flotante con 4 dígitos de mantisa si (a) se redondea, (b) se trunca la representación de todos los resultados, ya intermedios, ya finales.

De este modo, se ha llegado a un punto donde lo que se quiere calcular es una cota del error que se pueda escribir de la siguiente forma

$$\frac{(1 + \varepsilon_1) \cdots (1 + \varepsilon_k)}{(1 + \varepsilon_{k+1}) \cdots (1 + \varepsilon_n)} = 1 + \delta_n$$

y hacer una estimación de δ_n .

Ejemplo. Se quiere calcular

$$fl\left(\frac{x_1x_2}{x_3x_4}\right) = \frac{fl(x_1x_2)}{fl(x_3x_4)}(1+\varepsilon_1) = \frac{(x_1x_2)(1+\varepsilon_2)(1+\varepsilon_1)}{(x_3x_4)(1+\varepsilon_3)} = \frac{x_1x_2}{x_3x_4} \frac{(1+\varepsilon_1)(1+\varepsilon_2)}{(1+\varepsilon_3)}$$

donde ε_1 es el error relativo en la división, mientras que ε_2 y ε_3 son los errores relativos en las dos multiplicaciones. En este ejemplo se obtiene

$$1 + \delta_3 = \frac{(1+\varepsilon_1)(1+\varepsilon_2)}{(1+\varepsilon_3)}$$

En general, se tiene el teorema siguiente:

Teorema 1.4 Si $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ son errores de redondeo tales que $|\varepsilon_i| \leq \mu$, $i = 1 \div n$ y $n\mu \leq 0.1$, entonces $\exists \delta_n$ tal que

$$\frac{(1+\varepsilon_1) \cdots (1+\varepsilon_k)}{(1+\varepsilon_{k+1}) \cdots (1+\varepsilon_n)} = 1 + \delta_n$$

con $|\delta_n| \leq n(1.06\mu)$.

La condición $n\mu \leq 0.1$ asegura que n , el número de errores de redondeo, no puede ser ‘demasiado grande’, pero este hecho depende de la precisión de la máquina:

Si $\mu = 10^{-7}$, $n < 10^6$; en cambio, si $\mu = 10^{-15}$, $n = 10^{14}$ es aceptable.

Ejercicio. Construir la demonstración del teorema anterior siguiendo los pasos siguientes:

Lema 1. Si $0 \leq \mu < 1$, entonces $1 - n\mu \leq (1 - \mu)^n$.

(Indicación: $(1 - \mu)^n = 1 - n\mu + \dots$).

Lema 2. Si $0 \leq x \leq 0.1$, entonces $1 + x \leq e^x \leq 1 + 1.06x$.

(Indicación: Desarrollad e^x).

Lema 3. Si $0 \leq n\mu \leq 0.1$, entonces $(1 + \mu)^n \leq 1 + 1.06n\mu$.

(Indicación: Utilizar el lema 2).

Lema 4. Si $|\varepsilon_i| \leq \mu$, $i = 1 \div n$ y $n\mu \leq 0.1$, entonces

$$1 - n\mu \leq (1 + \varepsilon_1)(1 + \varepsilon_2) \cdots (1 + \varepsilon_n) \leq 1 + 1.06n\mu$$

Lema 5. Si $0 \leq x \leq 0.1$, entonces $(1 - x)^{-1} \leq 1 + 1.06x$.

Lema 6. Si $|\varepsilon_i| \leq \mu$, $i = 1 \div n$ y $n\mu \leq 0.1$, entonces

$$1 - n\mu \leq \frac{1}{(1 + \varepsilon_1) \cdots (1 + \varepsilon_n)} \leq 1 + 1.06n\mu$$

(Indicación: Utilizar los lemas 4 y 5).

El teorema queda demostrado a partir de los lemas 4 y 6. \square

Nota. Todo lo que se ha explicado hasta ahora se ha hecho con una aritmética ideal, es decir, en el caso ideal de que

$$\frac{|fl(x \bowtie y) - (x \bowtie y)|}{|x \bowtie y|} \leq \mu, \quad |x \bowtie y| \neq 0$$

pero es posible que, debido a la longitud de los registros intermedios, se tenga (ver [Van83])

$$\frac{|fl(x \bowtie y) - (x \bowtie y)|}{|x \bowtie y|} \leq r\mu, \quad |x \bowtie y| \neq 0$$

donde

- $r = 1$ en una aritmética de redondeo ideal.
- $r = 2$ en una aritmética con truncamiento o con registros de longitud $t + 1$ para el producto y el cociente.
- $r = 4$ con registros de longitud $t + 1$ para la suma y la diferencia o de longitud t para el producto y el cociente.
- $r \geq 4$ con longitud simple para la suma y la diferencia.

Entonces, en esta situación más general, tiene que substituirse, en el teorema y en todos los lemas relacionados, el valor de μ por el valor correspondiente de $r\mu$.

Ejercicio. Aplicar el análisis del error hacia delante y atrás de los siguientes cálculos :

$$\bullet fl\{fl(x^2) + y\} \quad \bullet fl\left\{\frac{(x+y)^2}{u+v}\right\}$$

1.4 Cálculo de series

Cuando se desarrolla una función $f(x)$ en potencias de x : $f(x) = \sum_{n=0}^{\infty} a_n x^n$,

en realidad se calcula $f(x) = \sum_{n=0}^N a_n x^n$

que no es más que la evaluación de un polinomio que, normalmente, se realiza mediante el método de Horner:

$$\begin{aligned} f(x) &= a_0 + a_1 x + a_2 x^2 + \dots + a_N x^N \\ &= a_0 + x \cdot (a_1 + x \cdot (a_2 + \dots + x a_N) \dots) \end{aligned}$$

En muchas situaciones, la incógnita en el momento de evaluar la serie es precisamente cuántos términos son necesarios, es decir N . Se considera la serie numérica

$$\sum_{j=0}^{\infty} a_j = S, \quad \text{con } S_n = \sum_{j=0}^N a_j \quad (\text{suma parcial } N\text{-ésima})$$

y $R_N = S - S_N = a_{N+1} + a_{N+2} + \dots$ que se llama resto N -ésimo y es, en realidad, el error absoluto.

1.4.1 Métodos de comparación

- Comparación con una serie geométrica: Si $\left| \frac{a_{j+1}}{a_j} \right| \leq \kappa < 1, \forall j \geq N$, entonces

$$|R_N| \leq \frac{\kappa}{1-\kappa} |a_N|. \text{ En particular, si } \kappa = 1/2, \text{ se tiene } |R_N| \leq |a_N|.$$

Ejemplo. Si se considera la serie $\sum_{n=1}^{\infty} \frac{\sqrt{n}}{\pi^{2n}}$, y se conoce la acotación $|a_6| < 3 \cdot 10^{-6}$, entonces

$$\left| \frac{a_{j+1}}{a_j} \right| = \sqrt{\frac{j+1}{j}} \frac{\pi^{2j}}{\pi^{2j+2}} \leq \sqrt{1 + \frac{1}{6}} \pi^{-2} < 0.11, \text{ para } j \geq 6. \text{ Se toma } \kappa = 0.11.$$

$$\text{Finalmente, } |R_6| \leq 3 \cdot 10^{-6} \frac{0.11}{1 - 0.11} < 4 \cdot 10^{-7}.$$

- Comparación con una integral impropia: Si $|a_j| \leq f(j), \forall j \geq N$, donde $f(x)$ es una función real decreciente $\forall x \geq N$, entonces

$$|R_N| \leq \int_N^{\infty} f(x) dx$$

Ejemplo. Si se tiene $a_j = (j^3 + 1)^{-1}$ y se elige $f(x) = x^{-3}$, se puede acotar el resto de la serie por

$$R_N \leq \int_N^{\infty} x^{-3} dx = \frac{N^{-2}}{2}$$

- Para series alternadas: Si $\sum_{n=1}^{\infty} (-1)^{n+1} a_n, a_n > 0$, y $\{a_n\}$ sucesión monótona decreciente que tiende a cero, entonces $|R_N| \leq |a_{N+1}|$.

Ejemplo. Para la serie $\sum_{k=1}^{\infty} \frac{(-1)^k}{k+1}$ se tiene $\left| \frac{1}{k+2} \right| \leq \frac{1}{2} \cdot 10^{-3}$, de manera que, si se quieren conseguir 3 decimales correctos, es necesario que $k \geq 1998$, que es muy costoso (son necesarios muchos términos) y no refleja la realidad, ya que la cota calculada no es nada aproximada. Con esta acotación anterior, sólo se obtiene una cota superior del error.

1.4.2 Aceleración de la convergencia

- Si $\sum_{j=0}^{\infty} b_j = S$ y $\lim_{j \rightarrow \infty} \frac{a_j}{b_j} = 1$, entonces $\sum_0^{\infty} a_j = S + \sum_0^{\infty} (a_j - b_j)$.

Ejemplo. Se quiere evaluar la serie $\sum_{j=1}^{\infty} \frac{1}{\sqrt{j^4 + 1}}$ y se sabe que $\sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6}$; entonces, $\sum_{j=1}^{\infty} a_j = \frac{\pi^2}{6} + \sum_{j=1}^{\infty} \left(\frac{1}{\sqrt{j^4 + 1}} - \frac{1}{j^2} \right) = 1.64493 - 0.30119 = 1.3437$. Sólo ha sido necesario calcular 5 términos para conseguir una precisión de $0.5 \cdot 10^{-4}$; mientras que, si se hubiera calculado directamente habrían sido necesarios, aplicando el método del apartado anterior, 20000 términos de la serie.

- Método de Aitken: se aplica a la aceleración del cálculo de límites de sucesiones; en particular, $\lim_{N \rightarrow \infty} \sum_{k=0}^N a_k x^k = \lim_{N \rightarrow \infty} S_N$.

Si la serie tiene suma S ; entonces,

$$S = S_N + \sum_{k=1}^{\infty} (S_{k+N} - S_{k+N-1})$$

Si estas diferencias forman una progresión geométrica de razón κ , $|\kappa| < 1$, se tiene

$$S = S_N + \sum_{k=1}^{\infty} \kappa (S_{k+N-1} - S_{k+N-2}) = S_N + (S_N - S_{N-1}) \sum_{k=1}^{\infty} \kappa^k = S_N + \frac{(S_N - S_{N-1})\kappa}{1 - \kappa}$$

donde

$$\kappa = \frac{S_N - S_{N-1}}{S_{N-1} - S_{N-2}}$$

Por tanto, dada una sucesión $\{S_N\}$ tal que converge hacia S , se define una nueva sucesión $\{S'_N\}$:

$$S'_N = S_N - \frac{(S_N - S_{N-1})^2}{S_N - 2S_{N-1} + S_{N-2}}$$

Si la sucesión $\{S_N\}$ es tal que las diferencias consecutivas forman exactamente una progresión geométrica, la nueva sucesión $\{S'_N\}$ es la sucesión constante: S, S, \dots

Ejemplo. Una manera de calcular el valor de π es por medio de la serie de la función

$$\arctan x = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{2k+1}, \text{ si entonces } \frac{\pi}{4} = \arctan 1 = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$$

Para que el error sea más pequeño que 10^{-4} se necesitan 4999 términos, si se evalúa directamente $(1/(2k+3)) < 10^{-4}$ (en realidad es suficiente sumar 2499 términos de pequeño a grande, en valor absoluto). Por el método de Aitken se consigue la precisión deseada con la evaluación de 9 términos.

1.5 Problemas

1. Sea el sistema:

$$\begin{cases} 3x + ay = 10 \\ 5x + by = 20 \end{cases}$$

con $a = 2.100 \pm 5 \cdot 10^{-4}$ y $b = 3.300 \pm 5 \cdot 10^{-4}$. ¿Con qué exactitud puede ser determinado el valor de $x + y$?

2. Calcular las raíces de la ecuación $x^2 - 200x + 1 = 0$, con 9 cifras exactas (utilizando la calculadora de mano). ¿Por qué no es correcto emplear la expresión $100 \pm \sqrt{100^2 - 1}$?
3. Se quiere calcular la expresión $(\sqrt{2} - 1)^6$ empleando el valor aproximado 1.4 por $\sqrt{2}$. ¿Qué expresión da mejor resultado?

1) $(\sqrt{2} - 1)^6$	2) $1/(\sqrt{2} + 1)^6$
3) $(3 - 2\sqrt{2})^3$	4) $1/(3 + 2\sqrt{2})^3$
5) $99 - 70\sqrt{2}$	6) $1/(99 + 70\sqrt{2})$

4. Resolver la ecuación

$$\frac{a}{x} = 0.2 - \frac{x}{100}$$

con $a = 1 \pm 0.001$ y estudiad los efectos de la incertidumbre de a sobre las soluciones de la ecuación.

5. Obtener expresiones equivalentes a las siguientes de manera que los errores en los cálculos sean pequeños:

a) $\frac{1}{1+2x} - \frac{1-x}{1+x}, \quad \text{si } |x| \ll 1$

b) $\sqrt{x + \frac{1}{x}} - \sqrt{x - \frac{1}{x}}, \quad \text{si } x \gg 1$

6. A qué distancia de $x = 0$ se obtienen, respectivamente, cuatro y seis decimales correctos utilizando las siguientes aproximaciones:

a) $\sin x \approx x \quad$ b) $\cos x \approx 1 - \frac{x^2}{2} \quad$ c) $(1 - x^2)^{-\frac{1}{2}} \approx 1 + \frac{x^2}{2}$

7. Sea la serie de potencias

$$\sum_{n=1}^{\infty} \frac{(x-3)^n}{n \cdot 3^n}$$

- a) Calcular el intervalo de convergencia y su suma $s(x)$.
- b) Calcular $s(2)$ con una precisión de 10^{-2} utilizando la serie y, posteriormente, con la expresión de s .

8. a) Demostrar que, si la serie $\sum_{k=0}^{\infty} (-1)^k a_k$ es convergente, también lo es

$$\frac{a_0}{2} + \frac{1}{2} \sum_{k=0}^{\infty} (-1)^k (a_k - a_{k+1})$$

y tiene la misma suma (serie transformada de Euler).

- b) Calcular la serie resultante de aplicar 3 veces esta transformación a $\sum_{k=0}^{\infty} \frac{(-1)^k}{k+1}$, y la suma con 3 decimales correctos.
c) ¿Cuántos términos es necesario sumar directamente para obtener la misma precisión que en b)?

9. Sea la familia de integrales siguiente:

$$I_j = \int_0^1 \frac{x^j}{x^2 + x + 6} dx \quad (j \geq 0)$$

- a) Encontrar una ley de recurrencia para las I_n e indicar cómo hay que utilizarla para calcularlas de forma numéricamente estable.
b) Si se conocen I_{n-1} e I_n con errores absolutos menores que ϵ y 2ϵ respectivamente, ¿con cuántas cifras decimales correctas se podrá calcular I_j por medio del método encontrado en a) (suponiendo una aritmética exacta)?
c) Explicad cómo se pueden obtener todos los I_j con t cifras decimales correctas sin calcular ninguna integral.

10. Si se trabaja con una aritmética de coma flotante de 3 dígitos de mantisa y se considera las cantidades correctamente redondeadas

$$a = 13.4 \quad b = 6.75 \quad c = 3.27 \quad d = 15.6$$

¿con qué precisión puede ser determinado el valor propio dominante de la matriz

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} ?$$

¿Cuántos dígitos de mantisa se necesitan para asegurar que el valor propio calculado tiene 3 dígitos correctos?

11. Se tiene el sistema de ecuaciones lineales

$$\begin{cases} x + ay = 5 \\ bx + 2y = d \end{cases}$$

con $a = 1.00 \pm 5 \cdot 10^{-3}$, $b = 1/a$ y $d = b - a$. ¿Con qué exactitud se puede determinar x y y ?

12. Se quiere calcular la función de Bessel $J_0(x)$ para $x = 20$ con seis decimales correctos y se ha decidido hacerlo con el siguiente desarrollo:

$$J_0(x) = \sum_{n=0}^{\infty} (-1)^n \frac{(x/2)^{2n}}{(n!)^2}$$

que converge para cualquier x .

- a) ¿Cuántos términos se han de considerar?
- b) ¿Cuánto vale el término más grande?
- c) ¿Cuántos dígitos hay que tener en cuenta en los cálculos?

13. Se tiene un algoritmo que calcula la integral:

$$I(a, b) = \int_0^1 \frac{e^{-bx}}{a+x^2} dx$$

Las cantidades físicas a y b han sido calculadas y sus valores son:

$$a = 0.400 \pm 0.003 \quad b = 0.340 \pm 0.005$$

Si se usa el algoritmo para unos cuantos valores de a y b , se ha obtenido:

a	b	I
0.39	0.34	1.425032
0.40	0.32	1.408845
0.40	0.34	1.398464
0.40	0.36	1.388198
0.41	0.34	1.372950

¿Qué grado de incertidumbre se tendrá en $I(a, b)$?

14. Calcular la expresión de la cota del error de la función:

$$y = \sqrt{2 + \sin x}$$

Hay que tener en cuenta que los errores relativos en valor absoluto son menores que ϵ para la representación de números, que 2ϵ para las cuatro operaciones elementales, que 3ϵ para las raíces cuadradas, que 4ϵ para el valor de x y que 5ϵ para la función \sin .

15. Sea

$$S = \sum_{n=1}^{\infty} a_n, \quad a_n \geq 0 \quad \forall n$$

y se define $c_r = a_r + 2a_{2r} + 4a_{4r} + \dots$

- a) Demostrar que $a_r = c_r - 2c_{2r}$ y, por tanto, $S = c_1 - c_2 + c_3 - c_4 + \dots$
- b) Utilizar el resultado del apartado a) para calcular

$$\sum_{n \geq 1} \frac{1}{n^3}$$

con 4 decimales correctos.

1.6 Prácticas

1.6.1 Práctica ejemplo

Escribir una rutina para resolver las ecuaciones de segundo grado:

$$ax^2 + bx + c = 0$$

con a, b, c reales en coma flotante. Es necesario estudiar y prever los posibles casos de overflow y underflow así como calcular las dos soluciones con la máxima precisión posible.

En primer lugar, es preciso decidir los casos triviales:

1. $a = b = c = 0$, cualquier número complejo es solución de la ecuación.
2. $a = b = 0, c \neq 0$, no existe solución de la ecuación.
3. $a = 0, b \neq 0, c \neq 0$, en este caso se tiene una ecuación de primer grado y su solución es $x_1 = -\frac{c}{b}$; se considera que la segunda raíz es $x_2 = \infty$.
4. $a \neq 0$, es una ecuación de segundo grado, se supone que $a > 0$ y se estudian los casos siguientes:
 - (a) $c = 0$, las dos raíces son $x_1 = 0, x_2 = -\frac{b}{a}$.
 - (b) $b = 0$, según el signo de c se tienen dos raíces reales $x_{1,2} = \pm\sqrt{c}$ para $c \geq 0$, y dos raíces imaginarias puras $x_{1,2} = \pm i\sqrt{-c}$ para $c < 0$.
 - (c) $b \neq 0, c \neq 0$, las dos soluciones son $x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$; este caso se estudia con mucho más detalle a continuación.

Como es sabido, si el discriminante $d = b^2 - 4ac$ es positivo se tienen dos raíces reales, y si es negativo se tienen dos raíces complejas conjugadas. Se trabajará en el primer supuesto: caso de raíces reales.

En primer lugar, para evitar cancelaciones que produzcan grandes errores en el cálculo de una de las dos raíces, se utilizará el algoritmo siguiente:

$$x_1 = -\text{sign}(b) \frac{|b| + \sqrt{b^2 - 4ac}}{2a} \quad x_2 = \frac{c}{ax_1} \quad |x_1| \geq |x_2|$$

Una segunda modificación consistirá en dividir los coeficientes de la ecuación por a ; se hará, naturalmente, si b/a y c/a son representables en el conjunto en coma flotante que se trabaja.

A continuación se estudiará bajo qué condiciones se tendrá *overflow* y *underflow*. Para ello, se supondrá que el conjunto en coma flotante es tal que:

$$2^{-M} < a, |b|, |c| < 2^M$$

Para que no haya *overflow* de la raíz mayor, se ha de cumplir que $|x_1| < 2^M$, es decir

$$\frac{|b| + \sqrt{b^2 - 4ac}}{2a} < 2^M$$

de donde, operando convenientemente,

$$0 < \sqrt{\left(\frac{b}{a}\right)^2 - 4\frac{c}{a}} < 2^{M+1} - \frac{|b|}{a}$$

y, por tanto, $\frac{|b|}{a} < 2^{M+1}$ como condición necesaria, y continuando se puede llegar a la desigualdad siguiente:

$$2^{2M} + \frac{c}{a} - 2^M \frac{|b|}{a} > 0$$

Con un razonamiento parecido al anterior se deduce que, para no tener *overflow* de la raíz más pequeña, $\frac{|b|}{a} < 2^{M+1}$ es una condición suficiente, y que para $\frac{|b|}{a} > 2^{M+1}$ es necesario que

$$2^{2M} + \frac{c}{a} - 2^M \frac{|b|}{a} < 0$$

Todas estas observaciones quedan plasmadas en la figura 1.4.

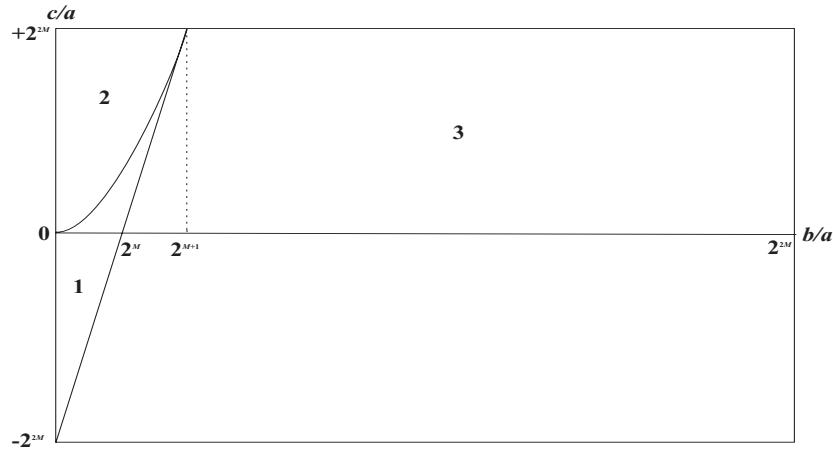


Fig. 1.4 Overflow de las raíces:

- 1. x_1, x_2 reales y calculables.
- 2. x_1, x_2 complejas y calculables.
- 3. x_1 overflow y x_2 calculable.

Si se estudia el *underflow* de las dos raíces se tiene, por un razonamiento análogo a los anteriores, pero ahora comparando con 2^{-M} , los resultados siguientes:

$\frac{|b|}{a} > 2^{-M+1}$ es una condición necesaria para que no haya *underflow* de la raíz pequeña, y además se tiene que cumplir

$$2^{-2M} + \frac{c}{a} - 2^{-M} \frac{|b|}{a} > 0$$

Una condición suficiente para que no haya *underflow* de la raíz mayor es que $\frac{|b|}{a} > 2^{-M+1}$, y cuando $\frac{|b|}{a} < 2^{-M+1}$ es necesario que

$$2^{-2M} + \frac{c}{a} - 2^{-M} \frac{|b|}{a} < 0$$

Toda esta discusión se representa en la figura 1.5.

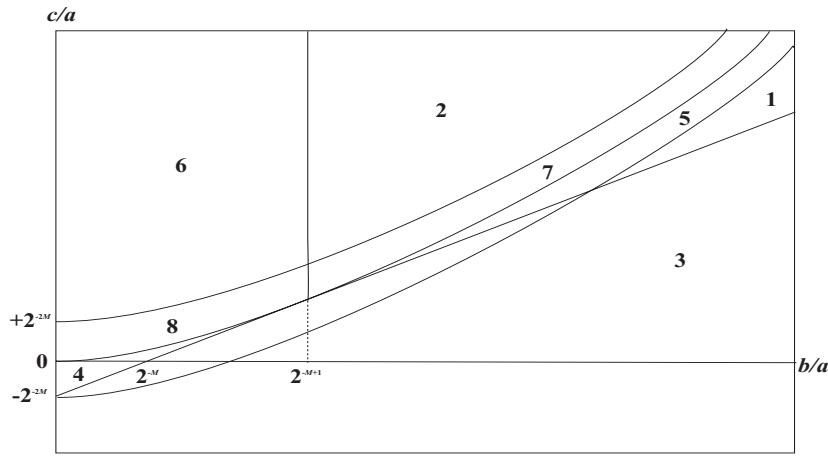


Fig. 1.5 Underflow de les raíces:

1. x_1, x_2 reales y calculables.
2. x_1, x_2 complejas y calculables.
3. x_1 calculable y x_2 underflow.
4. x_1, x_2 underflow.
5. x_1, x_2 reales y indistinguibles.
6. underflow de la parte real.
7. underflow de la parte imaginaria.
8. underflow de las partes real e imaginaria.

En el caso particular de *overflow* de la raíz mayor, pero no de la menor, es preciso calcular x_2 sin emplear x_1 . Además, se observa que, generalmente, b/a y c/a no son representables y, si se utiliza la fórmula clásica, se producen grandes errores de cancelación (p.e: $a = 10^{-30}$, $b = 10^{30}$, $c = 10^{30}$). Hay que buscar alguna fórmula alternativa: Se desarrolla la raíz del discriminante por la fórmula de Taylor

$$\sqrt{b^2 - 4ac} = |b| \left(1 - \frac{4ac}{b^2} \right)^{1/2} = |b| - \frac{2ac}{|b|} - \frac{2a^2c^2}{\xi^{3/2}}$$

con ξ punto intermedio entre b^2 y $b^2 - 4ac$; entonces

$$\frac{|b| - \sqrt{b^2 - 4ac}}{2a} = \frac{c}{|b|} + \frac{ac^2}{\xi^{3/2}}$$

por tanto, se aproxima $x_2 \sim -\text{sign}(b) \frac{c}{|b|}$ con un error del orden de $\frac{ac^2}{|b|^3}$.

La última posibilidad que se considera de las soluciones reales son las que $0 < \frac{1}{2} \sqrt{\left(\frac{b}{a}\right)^2 - 4\frac{c}{a}} < 2^{-M}$, es decir, las raíces que, para el conjunto de representación, son dobles. Estas son las que cumplen

$$\left(\frac{b}{a}\right)^2 - 4\frac{c}{a} - 2^{-2M+2} < 0$$

(esta condición también está representada en la figura 1.5).

Si se considera el caso del discriminante negativo, es decir, cuando se tienen dos soluciones complejas conjugadas e iguales a:

$$x_{1,2} = \frac{-\operatorname{sign}(b)}{2a} \left[|b| \pm i\sqrt{4ac - b^2} \right] = p \pm ir$$

se tiene que la parte real será representable si $2^{-M+1} < \frac{|b|}{a} < 2^{M+1}$, y la parte imaginaria no presentará *overflow* si $(\frac{b}{a})^2 - 4\frac{c}{a} + 2^{2M+2} > 0$, y no habrá *underflow* si $(\frac{b}{a})^2 - 4\frac{c}{a} + 2^{-2M+2} < 0$. Estas situaciones también se representan en las figuras 1.4 y 1.5.

A continuación se presenta un posible programa, que puede mejorarse si se consideran con más detalle las conclusiones anteriores. Para obtener una mayor precisión en el cálculo de las raíces simples, pero muy próximas, se calcula el discriminante con una mayor precisión que el resto de operaciones (todas las variables son de doble precisión y el discriminante es de cuádruple).

```

PROGRAM ECSEGUNDO
C      EL NUMERO MAS GRANDE REPRESENTABLE ES 10**32
PARAMETER EXPMAX=32
REAL*8 A,B,C,XR1,XR2,XC1,XC2
REAL*16 QA,QB,QC,D

PRINT *, 'Entra los valores de a,b,c: '
ACCEPT *, A,B,C

C      PRIMERAS DISCUSIONES SOBRE LOS COEFICIENTES

IF (A .EQ. 0.D0) THEN
    IF (B .EQ. 0.D0) THEN
        IF (C .EQ. 0.D0) THEN
            PRINT *, 'Cualquier valor complejo es solucion.'
            STOP
        ENDIF
        PRINT *, 'No existe solucion.'
        STOP
    ENDIF
    PRINT *, 'Es una ecuacion de primer grado: '
    PRINT *, -C/B
    STOP
ENDIF
IF (C .EQ. 0.D0) THEN
    XR1 = 0.D0
    XC1 = 0.D0
    XR2 = -B/A
    XC2 = 0.D0
    PRINT 100
    PRINT 101, XR1,XC1,XR2,XC2
    STOP
ENDIF
IF (B .EQ. 0.D0) THEN
    IF (C .GE. 0.D0) THEN
        XR1 = DSQRT(C)
        XC1 = 0.D0
        XR2 = -XR1
        XC2 = 0.D0
        PRINT 100
        PRINT 101, XR1,XC1,XR2,XC2
        STOP
    ENDIF

```

```

        ELSE
            XR1 = 0.DO
            XC1 = DSQRT(-C)
            XR2 = 0.DO
            XC2 = -XC1
            PRINT 100
            PRINT 101, XR1, XC1, XR2, XC2
            STOP
        ENDIF
    ENDIF
C     ECUACION DE SEGUNDO GRADO COMPLETA
    IF ((A.NE.1.DO).AND.(DLOG10(DABS(B))-DLOG10(A).LT.EXPMAX).AND.
*   (DLOG10(DABS(C))-DLOG10(A).LT.EXPMAX)) THEN
        QB = QEXTD(B)/QEXTD(A)
        B = B/A
        QC = QEXTD(C)/QEXTD(A)
        C = C/A
        QA = 1.DO
        A = 1.DO
    ELSE
        QA = QEXTD(A)
        QB = QEXTD(B)
        QC = QEXTD(C)
    ENDIF
C     CALCULO DEL DISCRIMINANTE EN CUADRUPLE PRECISION
    D = QB*QB-4.Q0*QA*QC
C     RAICES REALES
    IF (D .GE. 0.Q0) THEN
        IF (2*DLOG10(DABS(B))-DLOG10(A).LT.EXPMAX) THEN
            XR1 = -DSIGN(0.5DO,B)*(DABS(B)+QSQRT(D))/A
            XC1 = 0.DO
            XR2 = C/XR1/A
            XC2 = 0.DO
            PRINT 100
            PRINT 101,XR1,XC1,XR2,XC2
            STOP
        ELSE
            PRINT *, 'Overflow de la raíz mayor'
            XR2 =-DSIGN(1.DO,B)*C/DABS(B)
            XC2 = 0.DO
            PRINT 102, XR2, XC2
            STOP
        ENDIF
    C     RAICES COMPLEJAS
    ELSE
        XR1 = -DSIGN(0.5DO,B)*DABS(B)/A
        XR2 = XR1
        XC1 = QSQRT(-D)/A*0.5DO
        XC2 = -XC1
        PRINT 100
        PRINT 101, XR1, XC1, XR2, XC2
        STOP
    ENDIF
    STOP
100  FORMAT (' las dos soluciones son: ')
101  FORMAT(E25.16,'+',E25.16,'i',/,E25.16,'+',E25.16,'i')
102  FORMAT(E25.16,'+',E25.16,'i')
END

```

Algunos de los resultados obtenidos por esta rutina son los siguientes:

a	b	c	x_1	x_2
0	1	1	-1	
0	0	1	no	no
0	1	0	0	
1	2	1	-1	-1
1	1	0	0	1
1	0	1	1	-1
1	0	-1	+i	-i
1	4	3.99999999	-2.00009999	1.99990000
10^{-30}	10^{30}	10^{30}	overflow	1
10^{-25}	10^{32}	10^{30}	overflow	0.01

1.6.2 Enunciados

1. Se define la función error por

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

- (a) Calcular la serie de Taylor de la función y escribid un programa para evaluar $\operatorname{erf}(x)$; sumad términos de manera que el primero menoscipado ya no modifique el valor de la suma (es recomendable doble precisión).
- (b) Ya que se tiene una serie alternada, el error de truncamiento de la suma infinita es más pequeño que el error de redondeo. Investigar el efecto del error de redondeo comparando la suma calculada con vuestra subrutina con los valores de una tabla ([Abr72]) o a partir de alguna librería numérica (NAG, IMSL, etc.) de vuestro sistema. Aplicarlo a $x = 0.5, 10.0(0.5)$. Comprobar resultados y discutir la precisión del método, explicando las posibles causas de los problemas.
- (c) Una manera de evitar los problemas anteriores consiste en construir un desarrollo asintótico en potencias de x^{-1} y poder evaluar $\operatorname{erf}(x)$ para valores de x grandes:

$$\operatorname{erf}(x) \asymp 1 - \frac{e^{-x^2}}{\sqrt{\pi} x} \left(1 - \frac{1}{2x^2} + \frac{1 \cdot 3}{(2x^2)^2} - \frac{1 \cdot 3 \cdot 5}{(2x^2)^3} + \dots \right)$$

donde el símbolo \asymp indica que el límite del cociente entre los dos miembros del símbolo se aproxima a 1 en la medida que $x \rightarrow \infty$.

A pesar de que la serie no converge para ningún valor de x , si se toma un número conveniente de términos (N) con $x > x_0$, se puede obtener una buena aproximación. Determinar el valor de N (dejamos de sumar cuando un término es mayor, en valor absoluto, que el anterior) para diferentes valores de x y dar una estimación del error cometido.

- (d) Construir un programa que tenga en cuenta todos los apartados anteriores y poder así calcular correctamente (por lo menos 12 decimales) la función $\text{erf}(x)$ para todo x ; ¿qué x_0 es el más adecuado para empezar a considerar la aproximación asintótica?
2. Las funciones de Bessel pueden expresarse en forma de serie de potencias:

$$J_m(x) = \left(\frac{x}{2}\right)^m \sum_{k=0}^{\infty} \frac{(ix/2)^{2k}}{k! (m+k)!}$$

- (a) Escribir un programa para evaluar $J_0(x)$ y $J_1(x)$; sumar términos de manera que el primer menoscipado ya no modifique el valor de la suma (es recomendable doble precisión).
- (b) Como se tiene una serie alternada, el error de truncamiento de la suma infinita es más pequeño que el error de redondeo. Investigar el efecto del error de redondeo comparando la suma calculada por vuestra subrutina con los valores de una tabla ([Abr72]) o a partir de alguna librería numérica (NAG, IMSL, etc.) de vuestro sistema. Aplicarlo a $x = 0.0, 20.0(1.0)$. Comprobar los resultados y discutid la precisión del método, explicando las posibles causas de los problemas.
- (c) Una manera de evitar los problemas anteriores consiste en construir un desarrollo asintótico y poder evaluar $J_m(x)$ para valores de x suficientemente grandes:

$$\begin{aligned} J_m(x) &\asymp \sqrt{\frac{2}{\pi x}} \{ P(m, x) \cos \chi - Q(m, x) \operatorname{sen} \chi \} \\ \text{con } \chi &= x - \left(\frac{m}{2} + \frac{1}{4}\right) \pi \quad y \\ P(m, x) &\asymp 1 - \frac{(\mu-1)(\mu-9)}{2!(8x)^2} + \frac{(\mu-1)(\mu-9)(\mu-25)(\mu-49)}{4!(8x)^4} - \dots \\ Q(m, x) &\asymp -\frac{(\mu-1)}{8x} - \frac{(\mu-1)(\mu-9)(\mu-25)}{3!(8x)^3} + \dots \end{aligned}$$

donde el símbolo \asymp indica que el límite del cociente entre los dos miembros del símbolo se aproxima a 1 en la medida que $x \rightarrow \infty$ y $\mu = 4m^2$.

Determinar el valor de N (dejamos de sumar cuando un término es mayor, en valor absoluto, que el anterior) para diferentes valores de x ($x \in [0.0, 20.0]$), y dar una estimación del error cometido ($m = 0, 1$).

- (d) Construir un programa que tenga en cuenta todos los apartados anteriores con el objetivo de calcular correctamente (por lo menos 12 decimales) la función $J_m(x)$ para todo x ; ¿qué valor de x_0 es el más adecuado para empezar a considerar la aproximación asintótica?

3. La función hipergeométrica puede expresarse en forma de serie de potencias:

$$F(a, b; c; x) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{n! (c)_n} x^n .$$

donde $(a)_0 = 1$ y $(a)_n = a(a+1)\cdots(a+n-1)$.

- (a) Escribir un programa para evaluar $F(a, b; c; x)$ con

a	b	c
1/2	1/2	1
-1/2	1/2	1

- (b) Calcular la serie para $x = 0.1, 0.9 (0.1)$, e investigar el efecto del error de redondeo comparando la suma calculada por vuestra subrutina y los valores de una tabla ([Abr72]) o a partir de alguna librería numérica (NAG, IMSL, etc.) de vuestro sistema, teniendo en cuenta que la función hipergeométrica se relaciona con las integrales elípticas

$$K(\kappa) = \frac{\pi}{2} F\left(\frac{1}{2}, \frac{1}{2}; 1; \kappa\right) = \int_0^{\frac{\pi}{2}} \frac{d\phi}{\sqrt{1 - \kappa \sin^2 \phi}}$$

$$E(\kappa) = \frac{\pi}{2} F\left(-\frac{1}{2}, \frac{1}{2}; 1; \kappa\right) = \int_0^{\frac{\pi}{2}} \sqrt{1 - \kappa \sin^2 \phi} d\phi$$

que están tabuladas. Comprobar los resultados y discutir la precisión del método, explicando las posibles causas de los problemas.

- (c) Comprobar que

$$\lim_{n \rightarrow \infty} F\left(1, n; 1; \frac{x}{n}\right) = e^x$$

con una tabla de valores $n = 10^5, 10^7, 10^{10}$ y 10^{18} , y $x = 0.0, 0.001, 0.5, 6.5, 12.5, 18.5, 27.5$, así como

$$F\left(\frac{1}{2}, -\frac{1}{2}; \frac{1}{2}; \sin^2 x\right) = \cos x$$

con $x = k \frac{\pi}{2}$, $k = 0 \div 0.9 (0.1)$, haciendo servir el programa construido en el primer apartado.

4. Escribir una subrutina DIVIDE (A, B, C, D, E, F) con aritmética de doble precisión real donde las variables de entrada sean A, B, C y D y las de salida E y F tales que

$$e + y f = \frac{a + y b}{c + y d} \quad \text{donde } i^2 = -1.$$

Es conveniente incluir un parámetro para el caso que el denominador se anule; también se tiene que considerar que la subrutina evite *underflows* y *overflows* innecesarios. La eliminación de todos los posibles fallos es muy difícil, pero tenéis que dar cuenta de cuales son los puntos débiles de vuestro programa.

- (a) Probar la siguiente llamada:

CALL DIVIDE (1.234D38, 2.345D37, 9.876D37, -5.565D36, E, F).

- (b) Comparar el resultado con el cociente complejo de doble precisión del FORTRAN.

5. Escribir una subrutina SQROOT (A, B, E, F) con aritmética de doble precisión real donde las variables de entrada sean A y B y las de salida E y F tales que

$$e + y f = \sqrt{a + y b}$$

donde $e \geq 0$, y si $e = 0$, entonces $f \geq 0$. Evitar todos los *underflows* y *overflows* innecesarios y comparar los resultados de vuestro programa con la función incorporada CDSQRT que calcula la raíz cuadrada para números complejos en doble precisión; probar la siguiente llamada:

CALL SQRROOT (1.234D38, 2.345D37, E, F).

6. Se quiere evaluar

$$S(x) = \sum_{k=1}^{\infty} \frac{1}{\sqrt{k^3 + x}} - \sum_{k=1}^{\infty} \frac{1}{\sqrt{k^3 - x}}$$

para $|x| < 1$, con un error más pequeño que $e = 0.5 \cdot 10^{-8}$.

- (a) ¿Cuántos términos son necesarios para evaluar la primera serie con un error menor que e ?
- (b) Si cada término cuesta de evaluar $50\mu s$, ¿cuánto se tardaría en evaluar todos los términos de las dos series que se han calculado en el apartado anterior?
- (c) Hacer algunas operaciones algebraicas de manera que $S(x)$ pueda ser evaluada más rápidamente. Programar el método y evaluar $S(x)$ para $x = 0.5$ y $x = 0.999999999$. Comparar el número de términos empleados y el tiempo de ejecución respecto a la estimación anterior.

7. Evaluar la función

$$\Phi(x) = \sum_{k=1}^{\infty} \frac{1}{k(k+x)} \quad \text{para } x = 0, 1 (0.1)$$

con un error menor que $e = 0.5 \cdot 10^{-8}$.

Indicación: del hecho que

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}$$

probar que $\Phi(1) = 1$. Entonces, expresar $\Phi(x) - \Phi(1)$ en forma de serie que converja más rápidamente que la definida por $\Phi(x)$. Repetir este procedimiento para que la convergencia sea suficientemente rápida. ¿Se puede repetir indefinidamente?

Escribir un programa con las primeras aceleraciones encontradas. Sumar términos hasta obtener la precisión deseada y hacer que el programa devuelva el número de términos y el valor de la suma. Comparar el número de términos teóricos con el número práctico (sumando de pequeño a grande).

8. Se quiere evaluar

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2 + 1}$$

con un error menor que $e = 0.5 \cdot 10^{-10}$.

- (a) ¿Cuántos términos son necesarios para tener un error menor que e ?

- (b) Hacer algunas operaciones algebraicas para poder utilizar las fórmulas

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

$$\sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}$$

y acelerar el método de sumación.

Construir un programa que evalúe S a partir de los métodos comentados; calcular el número de términos necesarios para cada método, a fin de obtener la precisión deseada, y comparar el número de términos teóricos con el número real (sumando de pequeño a grande).

2 Interpolación polinómica

2.1 Introducción

Sea una familia de funciones de una variable real x , $\phi = \phi(x; a_0, \dots, a_n)$, con $n+1$ parámetros a_0, \dots, a_n que caracterizan las funciones de la familia.

El problema de interpolar consiste en determinar estos $n+1$ parámetros de manera que, para $n+1$ parejas dadas (x_i, y_i) , se cumpla

$$\phi(x_i; a_0, \dots, a_n) = y_i \quad i = 0 \div n$$

A (x_i, y_i) también se les llama nodos, nudos o puntos soporte.

Hay diferentes clases de interpolación según los tipos de funciones ϕ :

$$\begin{aligned} \phi(x; a_0, \dots, a_n) &= a_n + a_{n-1}x + \dots + a_0x^n && \text{Polinómica.} \\ \phi(x; a_0, \dots, a_n) &= a_0 + a_1e^{x_i} + \dots + a_ne^{nxi} && \text{Trigonométrica.} \\ \phi(x; a_0, \dots, a_n, b_0, \dots, b_m) &= \frac{a_0 + a_1x + \dots + a_nx^n}{b_0 + b_1x + \dots + a_mx^m} && \text{Racional.} \\ \phi(x; a_0, \dots, a_n, \lambda_0, \dots, \lambda_m) &= a_0e^{\lambda_0x} + \dots + a_ne^{\lambda_nx} && \text{Exponencial.} \end{aligned}$$

2.2 Interpolación polinómica

Se desea determinar los $n+1$ coeficientes del polinomio de grado n , $P(x) = a_n + a_{n-1}x + \dots + a_0x^n$, de tal manera que pase por los $n+1$ puntos (x_i, y_i) $i = 0 \div n$. Si se evalúa $P(x)$ en cada x_i se obtiene el sistema de ecuaciones lineal siguiente:

$$\left\{ \begin{array}{lcl} P(x_0) & = & a_n + x_0 a_{n-1} + \dots + x_0^n a_0 = y_0 \\ \vdots & & \vdots \\ P(x_n) & = & a_n + x_n a_{n-1} + \dots + x_n^n a_0 = y_n \end{array} \right.$$

Se tiene, por tanto, un sistema lineal de $n + 1$ ecuaciones con $n + 1$ incógnitas: a_n, a_{n-1}, \dots, a_1 y a_0 . El determinante del sistema se llama determinante de Vandermonde y su valor es

$$\begin{vmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ 1 & x_n & \cdots & x_n^n \end{vmatrix} = \prod_{i=1}^n \prod_{j=0}^{i-1} (x_i - x_j) \neq 0, \quad \text{si } x_i \neq x_j \text{ para } i \neq j.$$

En estas condiciones, la solución del problema existe y es única, ya que el problema de interpolación da lugar a un sistema lineal compatible y determinado. La dificultad que presenta la resolución de sistemas lineales grandes, que es costosa y con posible inestabilidad numérica, exige que se propongan otras formulaciones que darán lugar al mismo polinomio interpolador.

2.2.1 Fórmula de Lagrange

Se construye el polinomio interpolador $P(x) = \sum_{i=0}^n y_i l_i(x)$, de manera que $l_i(x_j) = \delta_{ij}$ (delta de Kronecker: $\delta_{ij} = 0$ si $i \neq j$ y $\delta_{ij} = 1$ si $i = j$). Entonces $P(x_j) = y_j$.

Los polinomios $l_i(x)$ se llaman polinomios de Lagrange y vienen definidos por

$$l_i(x) = \frac{(x - x_0)(x - x_1) \cdots (\widehat{x_i}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (\widehat{x_i}) \cdots (x_i - x_n)} = \frac{\omega(x)}{\omega'(x_i)(x - x_i)}$$

donde $\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$ y el símbolo $\widehat{x_i}$ quiere decir que aparecen los términos de la sucesión a_{i-1} i a_{i+1} , pero no el término a_i .

Ejemplo. Si se construye el polinomio interpolador para la función $y = \operatorname{sen} \pi x$, si se toman los valores $x = 0, 1/6$ i $1/2$, se obtiene $y_{0,1,2} = 0, 1/2$ i 1 . Además,

$$\begin{aligned} P_2(x) &= \frac{(x - 1/6)(x - 1/2)}{(0 - 1/6)(0 - 1/2)} \cdot 0 + \frac{(x - 0)(x - 1/2)}{(1/6 - 0)(1/6 - 1/2)} \cdot \frac{1}{2} + \\ &+ \frac{(x - 0)(x - 1/6)}{(1/2 - 0)(1/2 - 1/6)} \cdot 1 = \frac{7}{2}x - 3x^2 \end{aligned}$$

2.2.2 Métodos de Aitken y Neville

Ambos métodos se utilizan para evaluar el polinomio interpolador en un punto, construyendo una sucesión de los valores de los polinomios de grados crecientes que van interpolando cada vez más puntos. La propiedad que se hace servir en ambos métodos es la dada en el ejercicio siguiente.

Ejercicio. Sea f definida en x_0, x_1, \dots, x_n y $0 \leq m_i \leq n$ con $m_i \in \mathbb{N}$, $i = 1 \div k$ y se denota el polinomio de Lagrange de grado menor o igual que k que coincide con f en $x_{m_1}, x_{m_2}, \dots, x_{m_k}$ por $p_{m_1, \dots, m_k}(x)$. Si $x_i, x_j \in \{x_0, x_1, \dots, x_k\}$, son números diferentes y se define

$$p(x) = \frac{1}{x_i - x_j} \begin{vmatrix} x - x_j & p_{0,1,\dots,i-1,i+1,\dots,k}(x) \\ x - x_i & p_{0,1,\dots,j-1,j+1,\dots,k}(x) \end{vmatrix}$$

demostrar que $p(x)$ es el polinomio de grado $\leq k$ que interpola f en x_0, x_1, \dots, x_k .

Método de Aitken. Se construyen polinomios $p_{i,j}(x)$, $i = j \div n$, de grado menor o igual que j , que interpolen los puntos $\{x_0, x_1, \dots, x_{j-1}, x_j\}$, $j = 0 \div n$; se llega al polinomio $p_n \equiv p_{n,n}$, de grado menor o igual que n , que interpolará los puntos $\{x_0, x_1, \dots, x_n\}$:

$$\begin{aligned} p_{i,0}(x) &= f_i & i = 0 \div n \\ p_{i,j+1}(x) &= \frac{(x_j - x)p_{i,j}(x) - (x_i - x)p_{j,j}(x)}{x_j - x_i} = \frac{1}{x_j - x_i} \begin{vmatrix} x_j - x & p_{j,j}(x) \\ x_i - x & p_{i,j}(x) \end{vmatrix} \\ & & i = j + 1 \div m \quad j = 0 \div n - 1 \end{aligned}$$

El esquema de construcción para $n = 3$ es el de la tabla 2.1.

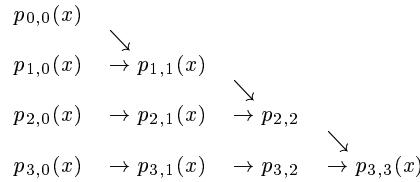


Tabla 2.1 Esquema del método de Aitken.

Método de Neville. Se construyen polinomios $p_{i,j}(x)$, $i = 0 \leq n - j$, de grado menor o igual que j , que interpolen los puntos $x_i, x_{i+1}, \dots, x_{i+j}$, $j = 0 \div n$; se llega al polinomio $p_n \equiv p_{0,n}$, de grado menor o igual que n , que interpolará los puntos $\{x_0, x_1, \dots, x_n\}$:

$$\begin{aligned} p_{i,0}(x) &= f_i & i = 0 \div n \\ p_{i,j+1}(x) &= \frac{(x_{i+j+1} - x)p_{i,j}(x) - (x_i - x)p_{i+1,j}(x)}{x_{i+j+1} - x_i} \\ &= \frac{1}{x_{i+j+1} - x_i} \begin{vmatrix} x_{i+j+1} - x & p_{i+1,j}(x) \\ x_i - x & p_{i,j}(x) \end{vmatrix} \\ & & i = 0 \div n - j - 1 \quad j = 0 \div n - 1 \end{aligned}$$

El esquema de construcción para $n = 3$ es el de la tabla 2.2.

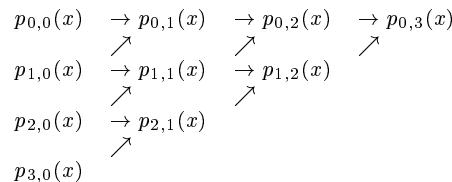


Tabla 2.2 Esquema del método de Neville.

2.2.3 Fórmulas de Newton. Diferencias divididas

El polinomio interpolador puede expresarse de la siguiente forma:

$$\begin{aligned} p(x) = & a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots \\ & + a_n(x - x_0)(x - x_1) \cdots (x - x_{n-1}) . \end{aligned}$$

El método de Newton permite calcular los coeficientes a_j , $j = 0 \div n$, por medio del cálculo de las *diferencias divididas de Newton*, definidas por:

$$\begin{aligned} f[x_i] &= f_i, \quad i = 0 \div n, \\ f[x_i, x_{i+1}, \dots, x_{i+j}, x_{i+j+1}] &= \frac{f[x_{i+1}, \dots, x_{i+j+1}] - f[x_i, \dots, x_{i+j}]}{x_{i+j+1} - x_i} \\ &\quad i = 0 \div n - j, \quad j = 0 \div n - 1. \end{aligned}$$

El esquema de construcción de las diferencias divididas se da en la tabla 2.3 para $n = 3$:

x_0	$f[x_0]$			
x_1	$f[x_1]$	$f[x_0, x_1]$		
x_2	$f[x_2]$	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$	$f[x_0, x_1, x_2, x_3]$
x_3	$f[x_3]$	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	

Tabla 2.3 Esquema de diferencias divididas.

Con $a_j = f[x_0, x_1, \dots, x_j]$, $j = 0 \div n$, se tiene

$$\begin{aligned} P_n(x) = & f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots \\ & + f[x_0, x_1, \dots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}) . \end{aligned}$$

Para demostrarlo es necesario el siguiente

Lema 2.1 Si $y = Q(x)$ es un polinomio de grado n , cualquier diferencia dividida de orden $n + 1$ construida a partir de Q es nula; es decir, para cualquier conjunto de números diferentes x, x_0, \dots, x_n , se tiene $Q[x, x_0, x_1, \dots, x_n] \equiv 0$.

Demostración: De la definición de diferencia dividida, si se considera

$$Q[x, x_0] = \frac{Q(x) - Q(x_0)}{x - x_0} \quad \text{y} \quad Q[x, x_0, x_1] = \frac{Q[x, x_0] - Q[x_0, x_1]}{x - x_1}$$

se tienen polinomios en x de grado $n - 1$ y $n - 2$, respectivamente, ya que los numeradores admiten como raíces x_0 en el primer caso, y x_1 en el segundo. Reiterando este razonamiento, se llega a $Q[x, x_0, x_1, \dots, x_{n-1}]$ que es un polinomio de grado cero (una constante independiente del valor de x). Sea $Q[x, x_0, x_1, \dots, x_{n-1}] = C$. Entonces,

$$Q[x, x_0, x_1, \dots, x_n] = \frac{Q[x, x_0, x_1, \dots, x_{n-1}] - Q[x_0, x_1, \dots, x_n]}{x - x_n} = \frac{C - C}{x - x_n} \equiv 0$$

□

Teorema 2.1 Sea $P(x)$ el polinomio interpolador de Lagrange de grado máximo n , tal que $P(x_i) = y_i$, $i = 0 \div n$. Entonces,

$$\begin{aligned} P(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \dots \\ &\quad + f[x_0, x_1, \dots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}) \end{aligned}$$

Demostración: De la definición de diferencias divididas se tiene

$$P[x, x_0, \dots, x_{m-1}] = P[x_0, \dots, x_m] + (x - x_m)P[x, x_0, \dots, x_m] \quad m = 1, 2, \dots, n$$

Aplicando esta definición, se tiene

$$\begin{aligned} P(x) &= P(x_0) + P[x, x_0](x - x_0) = P(x_0) + \{P[x_0, x_1] + (x - x_1)P[x, x_0, x_1]\}(x - x_0) \\ &= P(x_0) + P[x_0, x_1](x - x_0) + P[x, x_0, x_1](x - x_0)(x - x_1) \\ &= P(x_0) + P[x_0, x_1](x - x_0) + P[x_0, x_1, x_2](x - x_0)(x - x_1) + \dots \\ &\quad + P[x_0, x_1, \dots, x_n](x - x_0)(x - x_1) \cdots (x - x_{n-1}) + \\ &\quad + \underbrace{P[x, x_0, x_1, \dots, x_n]}_0 \underset{\text{(Lema)}}{(x - x_0)(x - x_1) \cdots (x - x_{n-1})(x - x_n)}. \end{aligned}$$

□

Ejemplo. Si se tiene la tabla de diferencias divididas siguiente

0	132.651			
0.2	148.877	81.13		
0.3	157.464	85.87	15.8	
0.4	166.375	89.11	16.2	0
0.7	195.112	95.79	16.7	0
0.9	216.000	104.44	17.3	

el polinomio interpolador es

$$P(x) = 132.651 + 81.13(x - 0) + 15.8(x - 0)(x - 0.2) + 0(x - 0)(x - 0.2)(x - 0.3)$$

2.2.4 Error en la interpolación polinómica

Sea una función que se evalúa en $n+1$ puntos, $f(x_i) = y_i$ $i = 0 \div n$, y el polinomio interpolador de grado máximo n , $P_n(x)$. ¿Cuál es el error producido en un cierto punto \bar{x} ? La respuesta la da el siguiente

Teorema 2.2 Sea f una función que admite derivadas hasta orden $n+1$ con continuidad en un entorno de \bar{x} . Sea $I(\bar{x}, x_0, \dots, x_n)$ el intervalo más pequeño que contiene los puntos \bar{x}, x_0, \dots, x_n . Entonces, $\exists c \in I(\bar{x}, x_0, \dots, x_n)$ tal que

$$f(\bar{x}) - P_n(\bar{x}) = \frac{f^{n+1}(c)}{(n+1)!} \omega(\bar{x}) \quad \text{on } \omega(x) = \prod_{i=0}^n (x - x_i)$$

Demostración: Sea $\bar{x} \neq x_i$ $i = 0 \div n$ (¿por qué?) y se construye una función auxiliar dependiente de k de manera que $F(z) = f(z) - P_n(z) - k\omega(z)$ se anule para $z = \bar{x}$; es decir: se quiere encontrar un valor de k tal que $F(\bar{x}) = 0$.

Por el teorema de Rolle, $F(z)$ tendrá $n+2$ ceros $(\bar{x}, x_0, \dots, x_n)$, $F'(z)$ tendrá $n+1$ ceros, $F''(z)$ tendrá n y, por fin, $F^{n+1}(z)$ tendrá un cero: c . Es decir, $0 = F^{n+1}(c) = f^{n+1}(c) - k(n+1)!$ y el valor de k es

$$k = \frac{f^{n+1}(c)}{(n+1)!}$$

Substituyendo este valor en la definición de F se tiene el teorema demostrado. \square

Ejemplo. Si se tiene una tabla que da el valor de la función

$$f(x) = \int_0^x e^{t^2} dt$$

para $x = 1.0, 1.1, 1.2$, calcular una cota del error cuando se interpola la función con un polinomio de grado 2 con la finalidad de aproximar $f(1.15)$.

Se tiene $f'(x) = e^{x^2}$, $f''(x) = 2xe^{x^2}$ y $f'''(x) = (4x^2 + 2)e^{x^2}$.

Entonces, $\max_{1 < x < 1.2} |f'''(x)| = [4(1.2)^2 + 2]e^{1.44} < 32.752$ y una cota del error cuando se interpola el punto 1.15 es

$$\frac{32.752}{6} |(1.15 - 1)(1.15 - 1.1)(1.15 - 1.2)| < 0.0021$$

Término de error y diferencias divididas

Si se añade un nodo: $(x_{n+1}, y_{n+1}) = (\bar{x}, f(\bar{x}))$, entonces, por la fórmula de Newton, se tiene

$$P_{n+1}(x) = \underbrace{y_0 + f[x_0, x_1](x - x_0) + \dots + f[x_0, x_1, \dots, x_n](x - x_0)(x - x_1) \dots (x - x_{n-1})}_{P_n(x)}$$

$$+ f[x_0, x_1, \dots, x_n, \bar{x}] \underbrace{(x - x_0)(x - x_1) \cdots (x - x_n)}_{\omega(x)}$$

es decir, $f(\bar{x}) = P_{n+1}(\bar{x}) = P_n(\bar{x}) + f[x_0, x_1, \dots, x_n, \bar{x}] \omega(\bar{x})$.

Ahora se tiene $f(\bar{x}) - P_n(\bar{x}) = f[x_0, x_1, \dots, x_n, \bar{x}] \omega(\bar{x})$; comparándolo con el teorema anterior se puede afirmar

$$f[x_0, x_1, \dots, x_n, \bar{x}] = \frac{f^{n+1}(c)}{(n+1)!}$$

y, por tanto, se pueden relacionar las derivadas y las diferencias divididas por la igualdad

$$f[x_0, x_1, \dots, x_n] = \frac{f^{n+1}(\bar{c})}{n!}$$

En particular, si se supone que f^{n+1} es continua y colapsando todos los puntos en x_0 , se tiene

$$\underbrace{f[x_0, x_0, \dots, x_0]}_{n+1 \text{ nodos}} = \frac{f^{n+1}(x_0)}{n!}$$

Ejemplo. Se puede calcular el polinomio de grado 3 que cumple:

$p(0) = 0$, $p'(0) = 1$, $p(1) = 3$ y $p'(1) = 6$, mediante una tabla de diferencias divididas donde se repiten nodos:

0	0		
0	0	1	
		3	2
1	3		1
1	3	6	

y el polinomio interpolador, que coincide con el polinomio incógnita, es

$$p(x) = 0 + 1x + 2x^2 + 1x^2(x-1) = x + x^2 + x^3$$

2.2.5 Elección de los nodos. Polinomios de Chebishev

Se ha visto que el error en la interpolación polinómica se puede representar por

$$f(\bar{x}) - P_n(\bar{x}) = \frac{f^{n+1}(c)}{(n+1)!} \omega(\bar{x}), \quad \text{donde } \omega(x) = \prod_{i=0}^n (x - x_i).$$

Se va a estudiar cómo elegir los nodos de manera que minimicen la función $\omega(x)$. Para ello, es necesario estudiar primero los polinomios de Chebishev, que vienen definidos por

$$T_n(x) = \cos(n \arccos x) \quad \left\{ \begin{array}{l} T_0(x) = 1 \\ T_1(x) = x \\ T_2(x) = 2x^2 - 1 \\ T_3(x) = 4x^3 - 3x \\ T_4(x) = 8x^4 - 8x^2 + 1 \\ \vdots \end{array} \right.$$

Propiedades:

- 1) El coeficiente del término de mayor grado de $T_n(x)$ es 2^{n-1} para $n \geq 1$ y T_0 es mónico.
- 2) Simetrías: $T_n(-x) = (-1)^n T_n(x)$.
- 3) Ceros: $T_n(x)$ tiene n ceros en $[-1, 1]$: $x_k = \cos\left(\frac{2k+1}{n}\frac{\pi}{2}\right)$, $k = 0 \div n-1$.
- 4) Extremos: $T_n(x)$ tiene $n+1$ extremos: $\tilde{x}_k = \cos\frac{k\pi}{n}$ y $T_n(\tilde{x}_k) = (-1)^k$, $k = 0 \div n$.

La propiedad fundamental de esta sucesión de polinomios es:

Teorema 2.3 Se consideran los polinomios mónicos de grado $n+1$ que pueden escribirse de la forma siguiente $\omega(x) = (x-x_0) \cdots (x-x_n)$ con x_i variando en el intervalo $[-1, 1]$. Entonces, la norma $\|\omega(x)\|_\infty = \max_{x \in [-1, 1]} |\omega(x)|$ es minimizada por $\omega(x) = \frac{1}{2^n} T_{n+1}(x)$.

Demostración: Se supone que existe un polinomio mónico de grado $n+1$, $v(x)$, tal que

$$\|v(x)\|_\infty < \left\| \frac{1}{2^n} T_{n+1}(x) \right\|_\infty = \frac{1}{2^n}$$

Si se consideran los cambios de signo del polinomio $Q(x) = v(x) - \frac{1}{2^n} T_{n+1}(x)$ en los $n+2$ extremos de $T_{n+1}(x)$, donde $\tilde{x}_k = \cos\frac{k\pi}{n+1}$, $k = 0 \div n+1$:

$$\left\{ \begin{array}{ll} \text{Si } k \text{ es par,} & v(\tilde{x}_k) < \frac{1}{2^n} T_{n+1}(\tilde{x}_k) = \frac{1}{2^n} \implies Q(\tilde{x}_k) < 0 \\ \text{Si } k \text{ es impar,} & v(\tilde{x}_k) > \frac{1}{2^n} T_{n+1}(\tilde{x}_k) = \frac{-1}{2^n} \implies Q(\tilde{x}_k) > 0 \end{array} \right.$$

Entonces, el polinomio de grado máximo n , $Q(x)$, tiene $n+1$ cambios de signo y, por tanto, $n+1$ ceros diferentes, lo cual es absurdo. En definitiva, no puede existir $v(x)$ con norma estrictamente inferior a la de $\frac{1}{2^n} T_{n+1}(x)$. \square

Es decir, las abscisas de los nodos a interpolar, x_i $i = 1 \div n+1$, son los ceros del polinomio de Chebishev de grado $n+1$.

Ejercicios.

1. Si se tiene una función $f : [-1, 1] \rightarrow \mathbf{R}$ y se interpola en las abscisas de Chebishev ($n+1$ ceros de $T_{n+1}(x)$), comprobar que una cota del error viene dada por

$$|f(\bar{x}) - P(\bar{x})| \leq \frac{M_n}{(n+1)! \cdot 2^n}$$

donde M_n es una cota superior de $f^{(n+1)}$ en el intervalo $[-1, 1]$.

2. Si se toma un intervalo cualquiera, $[a, b]$, y se hace el cambio

$$t = \frac{a+b}{2} + \frac{b-a}{2} x \quad (x \in [-1, 1] \iff t \in [a, b])$$

demostrar que

$$\|\tilde{T}_m(t)\|_{\infty, [a, b]} = \left(\frac{b-a}{2}\right)^m \cdot \frac{1}{2^{m-1}}$$

donde $\tilde{T}_m(t)$ es el polinomio mónico de Chebishev de grado m definido en el intervalo $[a, b]$.

3. Deducir la cota del error correspondiente al ejercicio 1 para una función definida en un intervalo cualquiera.

2.2.6 Diferencias finitas. Fórmulas de Newton

Sean las abscisas de los nodos de interpolación x_0, x_1, \dots, x_n equiespaciadas: $x_{i+1} - x_i = h$, $i = 1 \div n-1$. Si se considera $x = x_0 + sh$, entonces $x - x_i = (s-i)h$, y la definición del polinomio interpolador de Newton es

$$\begin{aligned} P_n(x) = P_n(x_0 + sh) &= y_0 + sh f[x_0, x_1] + s(s-1)h^2 f[x_0, x_1, x_2] + \dots + \\ &\quad + s(s-1) \cdots (s-n+1)h^n f[x_0, x_1, \dots, x_n] \\ &= \sum_{k=0}^n \binom{s}{k} k! h^k f[x_0, x_1, \dots, x_k] \end{aligned}$$

Se define la primera diferencia finita (progresiva) por $\Delta f(x_0) = f(x_1) - f(x_0)$, y, en general, $\Delta f(x_k) = f(x_{k+1}) - f(x_k)$, entonces se tiene una relación entre las diferencias finitas y las divididas:

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{1}{h} \Delta f(x_0)$$

$$f[x_k, x_{k+1}] = \frac{f(x_{k+1}) - f(x_k)}{x_{k+1} - x_k} = \frac{1}{h} \Delta f(x_k)$$

Análogamente,

$$f[x_0, x_1, x_2] = \frac{1}{2h^2} f[\Delta f(x_1), \Delta f(x_0)] = \frac{1}{2h^2} \Delta^2 f(x_0)$$

$$f[x_k, x_{k+1}, x_{k+2}] = \frac{1}{2h^2} f[\Delta f(x_{k+1}), \Delta f(x_k)] = \frac{1}{2h^2} \Delta^2 f(x_k)$$

donde se ha tomado $\Delta^k f(x_i) = \Delta^{k-1}(\Delta f(x_i))$.

En general, se tiene

$$f[x_0, x_1, \dots, x_k] = \frac{1}{k! h^k} \Delta^k f(x_0)$$

Substituyendo este resultado en la expresión del polinomio interpolador, se obtiene la fórmula de Newton para diferencias finitas (progresivas):

$$P_n(x) = \sum_{k=0}^n \binom{s}{k} \Delta^k f(x_0), \quad \text{con } s = \frac{x - x_0}{h}$$

Se obtiene de este modo una fórmula conveniente para interpolar en un entorno de x_0 , cuando $|s|$ es pequeño. Si se toma $n = 1$, se tiene la fórmula de interpolación lineal $P_1(x) = y_0 + s \Delta y_0$.

Ejemplo. Se interpola la función $y = f(x) = e^x$ en el intervalo $[3.5, 3.7]$ con un espaciado $h = 0.05$ para calcular en un entorno de 3.50:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
3.50	33.115	1.698	0.087	0.005
3.55	34.813	1.785	0.092	0.003
3.60	36.598	1.877	0.095	
3.65	38.475	1.972		
3.70	40.447			

Como las terceraas diferencias son prácticamente constantes, se considera $n = 3$. Con $x_0 = 3.50$ y $y_0 = 33.115$ se tiene

$$P_3(x) = 33.115 + 1.698 s + 0.087 \frac{s(s-1)}{2} + 0.005 \frac{s(s-1)(s-2)}{6}, \quad \text{donde } s = \frac{x - 3.5}{0.05}$$

Si se reordenan los nodos $x_n, x_{n-1}, \dots, x_1, x_0$, se tiene la fórmula de Newton (regresiva) para diferencias divididas:

$$\begin{aligned} P_n(x) &= f(x_n) + f[x_{n-1}, x_n](x - x_n) + f[x_{n-2}, x_{n-1}, x_n](x - x_n)(x - x_{n-1}) + \\ &\quad + \dots + f[x_0, \dots, x_n](x - x_n) \dots (x - x_1) \end{aligned}$$

Suponiendo las abscisas equiespaciadas y definiendo $x = x_n + ph$ y $x_i = (p+n-i)h$, se tiene la fórmula de Newton regresiva para diferencias finitas:

$$\begin{aligned} P_n(x) &= f(x_n) + ph f[x_{n-1}, x_n] + p(p+1)h^2 f[x_{n-2}, x_{n-1}, x_n] + \dots + \\ &\quad p(p+1) \dots (p+n-1)h^n f[x_0, \dots, x_n] = \\ &= y_n + p \Delta y_{n-1} + \frac{p(p+1)}{2!} \Delta^2 y_{n-2} + \dots + \frac{p(p+1) \dots (p+n-1)}{n!} \Delta^n y_0 \end{aligned}$$

donde $p = (x - x_n)/h$.

Ejemplo. Se calcula $\log_{10} 1044$ a partir de la tabla

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1000	3.0000000	43214	-426	8
1010	3.0043214	42788	-418	9
1020	3.0086002	42370	-409	8
1030	3.0128372	41961	-401	
1040	3.0170333	41560		
1050	3.0211893			

Se toma $x_n = 1050$ y $y_n = 3.0211893$. Entonces, $p = -0.6$ y

$$\begin{aligned}\log_{10} 1044 &= 3.0211893 + \\ &+ (-0.6) \left(0.0041560 + 0.4 \left[\frac{-0.0000401}{2} + 1.4 \frac{0.0000008}{6} \right] \right) \\ &= 3.0187005\end{aligned}$$

Ejercicio. Demostrar las siguientes propiedades del operador Δ (diferencia finita progresiva) :

1. $\Delta(f + g) = \Delta f + \Delta g$
2. $\Delta(\alpha f) = \alpha \Delta f$
3. $\Delta^m(\Delta^n f) = \Delta^{m+n} f$, donde $\Delta^0 f = f$ y $m, n \in \mathbb{N}$.
4. $f(x_k) = f(x_0 + k\Delta x) = (1 + \Delta)^k f(x_0)$
- Si f admite derivadas hasta orden n ,
5. Existe un $\xi \in (0, 1)$ tal que $\Delta^n f(x) = (\Delta x)^n \cdot f^{(n)}(x + \xi n \Delta x)$.
6. Del resultado 5, calculando el límite, se tiene

$$f^{(n)}(x) = \lim_{\Delta x \rightarrow 0} \frac{\Delta^n f(x)}{(\Delta x)^n}$$

2.2.7 Fenómeno Runge y teorema de Faber

La interpolación polinómica no es una buena aproximación de la función que se está interpolando; es decir, si $f(x)$ es una función continua cualquiera y $P_n(x)$ es el polinomio que interpola f en los nodos x_0, x_1, \dots, x_n , no es cierto, en general, que

$$\lim_{n \rightarrow \infty} P_n(x) = f(x) \text{ en } [a, b]$$

Runge demostró que si la función $f(x) = 1/(1+x^2)$ con $x \in [-5, 5]$ era interpolada tomando nodos equidistantes, P_n convergía hacia f si $|x| \leq 3.63\dots$ y divergía fuera de este intervalo (ver la práctica 8 de este capítulo).

Sea la sucesión triangular de puntos complejos o reales T :

$$\begin{array}{ccccccc} & & x_{00} & & & & \\ & & x_{10} & x_{11} & & & \\ & & x_{20} & x_{21} & x_{22} & & \\ & & \vdots & & & \ddots & \end{array}$$

Se supone que todos los $x_{ni} \in [a, b]$ y que para cada $n, n = 0, 1, \dots$ se construye el polinomio interpolador tal que $P_n(x_{ni}) = f(x_{ni})$ $i = 0 \div n$, ya que todos los puntos han sido generados en el dominio de f . Es decir, P_n interpola f en los puntos de la fila $(n+1)$ -ésima de T . Hecha esta construcción, Faber demostró (1914) que, para cada sucesión triangular de interpolación, T , de un intervalo $[a, b]$, existe una función continua f y un punto $x \in [a, b]$ tales que los correspondientes polinomios interpoladores $P_n(x)$ no convergen puntualmente hacia $f(x)$ cuando $n \rightarrow \infty$.

2.2.8 Interpolación de Hermite

Sea una función $y = f(x)$ evaluada en $n+1$ puntos x_0, x_1, \dots, x_n donde vale, respectivamente, y_0, y_1, \dots, y_n y su derivada en los mismos puntos vale y'_0, y'_1, \dots, y'_n . Entonces el polinomio interpolador de Hermite (interpola puntos no sólo de la función, sino también de su derivada) $P(x)$ de grado máximo $2n+1$ viene definido por

$$P(x) = \sum_{k=0}^n [1 - 2l'_k(x_k)(x - x_k)] l_k(x)^2 y_k + \sum_{k=0}^n (x - x_k) l_k(x)^2 y'_k$$

Es suficiente tener en cuenta que $l_k(x_i) = \delta_{ki}$; substituyendo, resulta

$$P(x_i) = y_i \quad P'(x_i) = y'_i \quad i = 0 \div n$$

Si se considera la función auxiliar $F(z) = f(z) - P(z) - [f(x) - P(x)]\{\omega(z)^2/\omega(x)^2\}$ donde x no es abscisa de ningún nodo y se sigue la demostración del error de interpolación de Lagrange, se obtiene la fórmula del error de interpolación del polinomio de Hermite:

$$\frac{f^{2n+2)}(c)}{(2n+2)!} \prod_{k=0}^n (x - x_k)^2 \quad \text{con } c \in I(x, x_0, x_1, \dots, x_n)$$

Ejercicio. Considerar los números $x_i, y_i^{(k)}$, con $i = 0 \div m, k = 0 \div n_i - 1$ y $x_0 < x_1 < \dots < x_n$. Si se define n tal que

$$n + 1 = \sum_{i=0}^m n_i$$

demostrar que existe un único polinomio de grado no superior a n tal que $P^{(k)}(x_i) = y_i^{(k)}$, con $i = 0 \div m, k = 0 \div n_i - 1$. Para más detalles, ver [Sto80].

2.3 Interpolación por splines cúbicas

Dada una colección de puntos (x_i, y_i) $i = 0 \div n$, esta interpolación consiste en hacer pasar por cada dos puntos consecutivos un polinomio de grado 3 precisando condiciones de continuidad de la primera y la segunda derivadas de $S(x)$:

Sean (x_i, y_i) y (x_{i+1}, y_{i+1}) dos puntos consecutivos; se construye un polinomio de tercer grado

$$s_i(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \quad i = 0 \div n - 1$$

tal que

1. Interpolación en los nodos y continuidad de la función

$$s_i(x_i) = y_i \quad s_i(x_{i+1}) = y_{i+1} \quad i = 0 \div n - 1$$

2. Continuidad de la primera derivada

$$s'_i(x_i+) = s'_{i-1}(x_i-) \quad i = 1 \div n - 1$$

3. Continuidad de la segunda derivada

$$s''_i(x_i+) = s''_{i-1}(x_i-) \quad i = 1 \div n - 1$$

Si se calcula el número de ecuaciones e incógnitas, se tiene para $n + 1$ puntos, n polinomios cúbicos y, por tanto, $4n$ incógnitas. Por otro lado, la condición de continuidad de $S(x)$ da lugar a $2n$ ecuaciones, la de $S'(x)$ a $n - 1$ ecuaciones y la de $S''(x)$ a $n - 1$; en total, $2n + 2(n - 1) = 4n - 2$ ecuaciones. Se tendrán que añadir dos condiciones: si se considera $s''_0(x_0) = s''_{n-1}(x_n) = 0$, se tendrá la spline natural; Forsythe ([For77]) construye la spline imponiendo condiciones sobre la tercera derivada en los extremos, como ya se verá.

2.3.1 Cálculo efectivo

La spline interpoladora está completamente determinada por las segundas derivadas, $\tau_i = s''_i(x_i)$, y podrá ser calculada resolviendo un sistema de ecuaciones lineal. Se considera el intervalo $[x_i, x_{i+1}]$ con $h_i = x_{i+1} - x_i$. Si se aplica interpolación lineal en los puntos (x_i, τ_i) y (x_{i+1}, τ_{i+1}) , se tiene

$$s''_i(x) = \tau_i \bar{w} + \tau_{i+1} w \quad \text{donde} \quad w = \frac{x - x_i}{h_i} \quad \text{y} \quad \bar{w} = \frac{x - x_{i+1}}{-h_i} = 1 - w$$

Integrando esta segunda derivada y llamando $S(x)$ a la spline general, se obtiene

$$S'(x) = \int S''(x) dx + A = \tau_i \int \bar{w} dx + \tau_{i+1} \int w dx + A$$

$$S'(x) = -h_i \tau_i / 2 \cdot \bar{w}^2 + h_i \tau_{i+1} / 2 \cdot w^2 + A$$

Integrando otra vez, se calcula $S(x)$:

$$S(x) = \int S'(x) dx + B = h_i^2 \tau_i / 6 \cdot \bar{w}^3 + h_i^2 \tau_{i+1} / 6 \cdot w^3 + Ax + B$$

$$S(x) = h_i^2 [\tau_i / 6 \cdot \bar{w}^3 + \tau_{i+1} / 6 \cdot w^3] + Cw + D\bar{w}$$

donde $C = Ax_{i+1} + B$ y $D = Ax_i + B$.

Si se denota por $\sigma_i = \tau_i/6$ $i = 0 \div n - 1$ y se impone que $S(x)$ interpole los puntos x_i y x_{i+1} , se obtienen los valores de las constantes de integración:

$$S(x_i) = h_i^2(\sigma_i 1^3 + \sigma_{i+1} 0^3) + C \cdot 0 + D \cdot 1 = y_i \implies D = y_i - h_i^2 \sigma_i$$

$$S(x_{i+1}) = h_i^2(\sigma_i 0^3 + \sigma_{i+1} 1^3) + C \cdot 1 + D \cdot 0 = y_{i+1} \implies C = y_{i+1} - h_i^2 \sigma_{i+1}$$

El resultado del que se parte para identificar la spline es el siguiente:

$$S(x) = w y_{i+1} + \bar{w} y_i + h_i^2 [(w^3 - w) \sigma_{i+1} + (\bar{w}^3 - \bar{w}) \sigma_i]$$

donde sólo se desconoce σ_i y σ_{i+1} . Para establecer el sistema lineal que determinará las σ se impone la condición 2 (la continuidad de $S'(x)$):

$$S'(x) = \frac{y_{i+1} - y_i}{h_i} + h_i [(3w^2 - 1) \sigma_{i+1} - (3\bar{w}^2 - 1) \sigma_i] \quad \text{para } x \in [x_i, x_{i+1}]$$

donde, si se llama $\Delta_i = (y_{i+1} - y_i)/h_i$, entonces la condición $s'_i(x_i+) = s'_{i-1}(x_i-)$ se puede escribir como

$$\Delta_i + h_i (-\sigma_{i+1} - 2\sigma_i) = \Delta_{i-1} + h_{i-1} (2\sigma_i + \sigma_{i-1}) \quad i = 1 \div n - 1$$

que es el sistema

$$h_{i-1} \sigma_{i-1} + 2(h_{i-1} + h_i) \sigma_i + h_i \sigma_{i+1} = \Delta_i - \Delta_{i-1} \quad i = 1 \div n - 1 \quad (2.1)$$

que matricialmente se expresa por

$$\begin{pmatrix} h_0 & 2(h_0 + h_1) & h_1 & & \\ & h_1 & 2(h_1 + h_2) & h_2 & \\ & & \ddots & & \\ & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \end{pmatrix} \begin{pmatrix} \sigma_0 \\ \sigma_1 \\ \vdots \\ \sigma_n \end{pmatrix} = \begin{pmatrix} \Delta_1 - \Delta_0 \\ \Delta_2 - \Delta_1 \\ \vdots \\ \Delta_{n-1} - \Delta_{n-2} \end{pmatrix}$$

Se ha obtenido un sistema lineal de $n-1$ ecuaciones con $n+1$ incógnitas; faltan dos condiciones que dan lugar a splines cúbicas diferentes:

- Spline natural: $\sigma_0 = \sigma_n = 0$. La matriz del sistema es de orden $n-1$, es tridiagonal, simétrica y, si $x_0 < x_1 < \dots < x_n$, es diagonalmente dominante y regular.
- Spline completa: Si se tiene los datos y'_0 y y'_n , se define
 $y'_0 = S'(x_0) = \Delta_0 + h_0 (-\sigma_1 - 2\sigma_0)$ y $y'_n = S'(x_n) = \Delta_{n-1} + h_{n-1} (2\sigma_n + \sigma_{n-1})$
de manera que se han añadido las ecuaciones

$$\begin{cases} 2 h_0 \sigma_0 + h_0 \sigma_1 = \Delta_0 - y'_0 \\ h_{n-1} \sigma_{n-1} + 2 h_{n-1} \sigma_n = y'_n - \Delta_{n-1} \end{cases}$$

- Spline Forsythe ([For77]): Aproxima las tercera derivadas de $S(x)$ en los puntos extremos x_0 y x_n mediante las diferencias divididas de los 4 primeros y últimos puntos. Si de $S''(x) = 6 w \sigma_{i+1} + 6 \bar{w} \sigma_i$ se deriva, se obtiene

$$S'''(x) = 6 \frac{\sigma_{i+1} - \sigma_i}{h_i}$$

En particular, Forsythe sugiere la aproximación

$$6 \frac{\sigma_1 - \sigma_0}{h_0} \approx 6 [x_0, x_1, x_2, x_3] \quad \text{y} \quad 6 \frac{\sigma_n - \sigma_{n-1}}{h_{n-1}} \approx 6 [x_{n-3}, x_{n-2}, x_{n-1}, x_n]$$

que da lugar a las ecuaciones siguientes:

$$\begin{cases} -h_0 \sigma_0 + h_0 \sigma_1 &= h_0^2 [x_0, x_1, x_2, x_3] \\ h_{n-1} \sigma_{n-1} - h_{n-1} \sigma_n &= -h_{n-1}^2 [x_{n-3}, x_{n-2}, x_{n-1}, x_n] \end{cases}$$

Para evaluar la spline, así como su derivada o integral, es conveniente considerar el polinomio expresado por

$$S(x) = y_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad x \in [x_i, x_{i+1}]$$

con

$$b_i = \frac{y_{i+1} - y_i}{h_i} - h_i(\sigma_{i+1} + 2\sigma_i), \quad c_i = 3\sigma_i \quad \text{y} \quad d_i = \frac{\sigma_{i+1} - \sigma_i}{h_i}.$$

2.3.2 Curvatura mínima de las splines cúbicas

Si $f \in C^2([a, b])$, se define la seminorma

$$\|f\|^2 = \int_a^b (f''(t))^2 dt$$

Nótese que, para $f(x) = cx + d$, se tiene $\|f\| = 0$.

Una propiedad geométrica de las splines cúbicas es la que expresa el siguiente

Teorema 2.4 Dados $X = \{a = x_0 < x_1 < \dots < x_n = b\}$, $Y = \{y_0, \dots, y_n\}$ y $f \in C^2([a, b])$ tal que $f(x_i) = y_i$, $i = 0 \div n$; entonces, $\|f\|^2 \geq \|S\|^2$, para cualquier función spline con una de las condiciones siguientes:

$$\begin{aligned} S''(a) &= S''(b) = 0 \\ S^{(k)}(a) &= S^{(k)}(b) \quad k = 0, 1, 2 \quad \text{si } f^{(k)}(a) = f^{(k)}(b) \\ S'(a) &= f'(a), \quad S'(b) = f'(b) \end{aligned}$$

Demostración: Este resultado es consecuencia de la siguiente igualdad:

$$\|f - S\|^2 = \|f\|^2 - \|S\|^2 - 2 \left[(f'(x) - S'(x))s''(x)|_a^b - \sum_{i=1}^n (f(x) - S(x))S'''(x)|_{x_{i-1}}^{x_i} \right]$$

que se deja para que el lector la compruebe. \square

La minimización de la integral de la segunda derivada al cuadrado se traduce en que la spline minimiza con la forma que toma su energía potencial, que es, aproximadamente, proporcional a la curvatura: si se tiene la función $y = f(x)$, la curvatura de esta función viene dada por $f''(x) \cdot (1 + f'(x)^2)^{-3/2}$ y, si $|f'(x)| \ll 1$, la curvatura se aproxima a $f''(x)$.

2.4 Problemas

1. Calcular $f(3)$ a partir de la interpolación cuadrática de la tabla siguiente:

x		1	2	4	5
$f(x)$		0	2	12	21

utilizando los puntos 1, 2 y 4 para el primer cálculo y, después, los puntos 2, 4 y 5, y comparar los resultados. Calcular $f(3)$ por interpolación cúbica.

2. Se considera la función $f(x) = e^{x/3}$. Determinar un polinomio interpolador $p(x)$ de grado tan pequeño como sea posible, tal que

$$\forall x \in [-1, 1], \quad |f(x) - p(x)| \leq 10^{-4}$$

3. Se quiere construir una tabla de valores de $\sin x$ en el intervalo $[0, \pi/4]$ en los puntos $0, h, 2h, \dots, mh = \pi/4$. Determinar el tamaño de h (y también de m) de manera que la interpolación en tres puntos consecutivos dé un error más pequeño que 10^{-6} en el intervalo definido por los tres puntos, al aproximar la función $\sin x$ por un polinomio de grado 2.

4. Se quiere calcular $\sin 1.2$ a partir de una tabla de valores que da $\sin x$ sólo en $x = .1, .2, .3, \dots, .9, 1.0$. El cálculo se basa en el método siguiente: se interpola en los puntos $x_0 = 1.0, x_1 = .9, \dots, x_n = 1 - .1n$. ¿Cuál es el valor de n para el que el error es menor que .0001?

5. Calcular una solución de la ecuación $x - e^{-x} = 0$, si se sabe que

$$e^{-0.50} = .60653, \quad e^{-0.55} = .57695 \quad y \quad e^{-0.60} = .54881.$$

6. Si $f(0) = 3, f'(0) = 1, f(1) = 2$ y $f'(1) = -2$, determinar dónde está el máximo de f entre 0 y 1. ¿Cuánto vale?

7. Sea la función de Bessel definida por:

$$J_0(x) = \frac{1}{\pi} \int_0^\pi \cos(x \operatorname{sen} t) dt$$

- a) Si se tabula en puntos equidistantes $x_i = x_0 + ih$ y se interpola linealmente, calcular una cota superior de h para obtener un error máximo inferior a 10^{-6} .

- b) Dado el error máximo por

$$\max_{x \in [x_0, x_n]} |P_n(x) - J_0(x)|$$

donde $P_n(x)$ es el polinomio interpolador de $J_0(x)$ en los nudos $x_i^{(n)} = i/n, i = 0 \div n$, encontrad una expresión de este error en función sólo de n y de x .

8. Aproximar mediante una recta la gráfica de la función $f(x) = 1/(x+a)$, con $x \in [-1, 1]$ si $a \geq 2$ utilizando interpolación

a) equidistante b) de Chebishev

Calcular cotas del error cometido y comparar los resultados.

9. Si $P_n(x)$ es el polinomio interpolador de grado máximo n de la función $f \in C^{(n)}[a, b]$ en los puntos equiespaciados $a = x_0 < x_1 < \dots < x_n = b$, $x_j - x_{j-1} = h$, se define $M_k = \max_{x \in J} |f^{(k)}(x)|$ con $J = [a, b]$; demostrar que

$$\forall x \in J, \quad |f(x) - P_n(x)| \leq \frac{h^{n+1}}{4(n+1)} M_{n+1}$$

10. Con las hipótesis del problema anterior, demostrar que

$$|f(x) - P_2(x)| \leq \frac{\sqrt{3}}{27} h^3 M_3 \quad \text{y} \quad |f(x) - P_3(x)| \leq \frac{h^4}{24} M_4 .$$

11. ¿Qué espaciado de los valores de x hay que tomar en una tabla de e^x dada con 15 decimales, si se quiere utilizar un polinomio interpolador de grado máximo 2 con una aproximación de 8 decimales correctos en el intervalo $[0, 2]$?

12. Demostrar que

$$f[x_0, x_1, \dots, x_n] = \sum_{i=0}^n \frac{f(x_i)}{\omega'(x_i)}$$

13. Se quiere interpolar la función $f(x) = (1+2x)^{-1}$ en el intervalo $[4, 5]$. Dar una estimación del número de puntos de interpolación que se necesitan para que el error de interpolación sea menor que 10^{-6} y comparar este error con el que se produciría aplicando interpolación equidistante con el mismo número de puntos.

14. Se quiere tabular la función

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

en puntos equidistantes del intervalo $[0, 1]$ con paso $h = 0.01$.

Encontrar el número t de decimales exactos con que se han de asignar a los valores de una tabla para la que el error debido a la interpolación por polinomios de grado 1, 3 y 5 (usando los puntos más próximos) no exceda al error propagado en el cálculo del polinomio interpolador a causa del error de redondeo de los datos.

15. Calcular $\operatorname{tg}(\pi/8)$ mediante interpolación hermítica en 0 y $\pi/4$. Acotar el error cometido y comparar la cota encontrada con el error exacto.

16. Calcular \bar{x} tal que $\bar{x}^3 = 0.002$, por medio de interpolación inversa en $x_0 = 0, x_1 = 0.1, x_2 = 0.2$. Aplicar también interpolación lagrangiana, suponiendo que x es un polinomio en y .

17. Encontrar una condición para el polinomio interpolador de grado máximo 2, $P_2(x_i) = f(x_i)$ $i = 0, 1, 2$, donde $x_2 - x_1 = x_1 - x_0 = h$, tal que $P_2(x) = \bar{y}$ tenga una solución en el intervalo $[x_0, x_2]$. Aplicarlo a los siguientes casos:
- $f(x_0) = 1, f(x_1) = 1, f(x_2) = 0.5, \bar{y} = 1.1$.
 - $f(x_0) = 1, f(x_1) = 1, f(x_2) = -1, \bar{y} = 1.1$.
18. Construir un polinomio de grado a lo sumo 3 tal que $P(1) = 1, P'(1) = 2, P''(1) = 4$ y $P(2) = 5$. Calcular $P(1.6)$.
19. Calcular la spline cuadrática $q(x)$ que pasa por $(0, 0), (1, 2), (2, 2)$ y $(3, 0)$ con la condición adicional $q'_0(0) = 1$.

2.5 Prácticas

2.5.1 Práctica ejemplo

Escribir una rutina que construya una spline cúbica con condiciones de contorno de Forsythe y una función que la evalúe en un punto dado.

Aplicarlo a la tabla siguiente sobre la evolución de la población catalana:

Año	Habitantes	Año	Habitantes	Año	Habitantes
1900	1984115	1940	2915757	1975	5660393
1910	2099218	1950	3218596	1981	5956414
1920	2355908	1960	3888485	1986	5978638
1930	2731627	1970	5107606		

Se construye una rutina (**SUBROUTINE SPLINE** tomada de [For77]) que tiene por entrada el número de nudos, $N > 2$, el vector de abscisas en orden creciente, X , y las ordenadas de los nudos, Y . La salida consta de los vectores B , C , D , que son los coeficientes de la spline:

$$S(X) = Y(I) + B(I)*(X-X(I)) + C(I)*(X-X(I))^2 + D(I)*(X-X(I))^3$$

para $X(I) \leq X \leq X(I+1)$ $I=1, \dots, N$

Se acompaña esta rutina de la función **SAVAL** que evalúa la spline cúbica en un punto U por el método de Horner, donde

$$\begin{aligned} \text{si } U < X(1) &\implies I = 1 \\ \text{si } U \geq X(N) &\implies I = N \end{aligned}$$

```

SUBROUTINE SPLINE (N, X, Y, B, C, D)
INTEGER*4 N, NM1, IB, I
REAL*8 X(N), Y(N), B(N), C(N), D(N), T

NM1 = N-1
IF ( N .LT. 3 ) RETURN

C      EL SISTEMA TRIDIAGONAL TIENE: B EN LA DIAGONAL,
C      D FUERA DE LA DIAGONAL Y C EN EL SEGUNDO MIEMBRO.

D(1) = X(2) - X(1)
C(2) = (Y(2) - Y(1))/D(1)
DO I = 2, NM1
    D(I) = X(I+1) - X(I)
    B(I) = 2.*(D(I-1) + D(I))
    C(I+1) = (Y(I+1) - Y(I))/D(I)
    C(I) = C(I+1) - C(I)
ENDDO

```

```

C      CONDICIONES FINALES: LA TERCERA DERIVADA EN X(1) Y X(N)
C      SE OBTIENEN DEL CALCULO DE LAS DIFERENCIAS DIVIDIDAS.
B(1) = -D(1)
B(N) = -D(N-1)
C(1) = 0.
C(N) = 0.
IF ( N .EQ. 3 ) GO TO 15
C(1) = C(3)/(X(4)-X(2)) - C(2)/(X(3)-X(1))
C(N)=C(N-1)/(X(N)-X(N-2))-C(N-2)/(X(N-1)-X(N-3))
C(1) = C(1)*D(1)**2/(X(4)-X(1))
C(N) = -C(N)*D(N-1)**2/(X(N)-X(N-3))
C      SE APLICA ELIMINACION GAUSSIANA
15  DO I = 2, N
      T = D(I-1)/B(I-1)
      B(I) = B(I) - T*D(I-1)
      C(I) = C(I) - T*C(I-1)
ENDDO
C      RESOLUCION HACIA ATRAS
C(N) = C(N)/B(N)
DO IB = 1, NM1
      I = N-IB
      C(I) = (C(I) - D(I)*C(I+1))/B(I)
ENDDO
C      C(I) ES AHORA LA SIGMA(I), SE CALCULAN LOS COEFICIENTES DEL POLINOMIO
B(N) = (Y(N) - Y(NM1))/D(NM1) + D(NM1)*(C(NM1) + 2.*C(N))
DO I = 1, NM1
      B(I) = (Y(I+1) - Y(I))/D(I) - D(I)*(C(I+1) + 2.*C(I))
      D(I) = (C(I+1) - C(I))/D(I)
      C(I) = 3.*C(I)
ENDDO
C(N) = 3.*C(N)
D(N) = D(N-1)
RETURN
END

REAL*8 FUNCTION SAVAL ( N, U, X, Y, B, C, D)
INTEGER*4 N, I, J, K
REAL*8 U, X(N), Y(N), B(N), C(N), D(N), DX
DATA I/1/
IF ( I .GE. N ) I = 1
IF ( U .LT. X(I) ) GO TO 10
IF ( U .LE. X(I+1) ) GO TO 30
C      LOCALIZACION DEL SUBINTERVALO
10  I = 1
J = N+1
20  K = (I+J)/2
IF ( U .LT. X(K) ) J = K
IF ( U .GE. X(K) ) I = K
IF ( J .GT. I+1 ) GO TO 20
C      EVALUACION DE LA SPLINE
30  DX = U - X(I)
SAVAL = Y(I) + DX*(B(I) + DX*(C(I) + DX*D(I)))
RETURN
END

```

```

PROGRAM POBLACION
REAL*8 X(50), Y(50), B(50), C(50), D(50), S, U, SAVAL
INTEGER*4 N, I
EXTERNAL SAVAL,SPLINE
ACCEPT *, N
C
N VAL 11
DO I=1,N
    ACCEPT *, X(I), Y(I)
ENDDO
CALL SPLINE ( N, X, Y, B, C, D)
DO I=1,7
    U= X(I) + (X(I+1)-X(I))/2.
    S= SAVAL ( N, U, X, Y, B, C, D)
    PRINT *, U, S
ENDDO
U=1978.0
S=SAVAL ( N, U, X, Y, B, C, D)
PRINT *, U, S
END

```

A partir de la tabla dada con $N = 11$, se ha generado la spline cúbica y se ha evaluado en los puntos medios dando lugar a la tabla ampliada 2.4 y a la figura 2.1.

Año	Habitantes	Año	Habitantes	Año	Habitantes
1900	1 984 115	1940	2 915 757	1973	5 465 520
1905	2 025 830	1945	3 035 930	1975	5 660 393
1910	2 099 218	1950	3 218 596	1978	5 856 660
1915	2 204 730	1955	3 484 070	1981	5 956 414
1920	2 355 908	1960	3 888 485	1984	5 987 380
1925	2 553 210	1965	4 463 560	1986	5 978 638
1930	2 731 627	1970	5 107 606		

Tabla 2.4 Resultados de la spline.

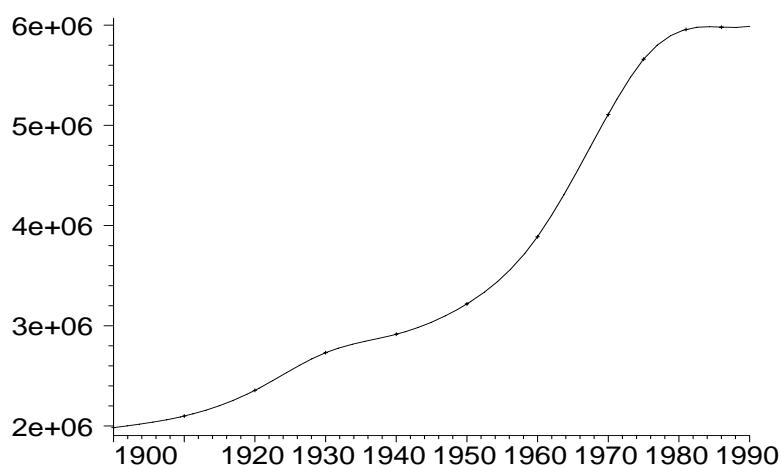


Figura 2.1 Evolución de la población catalana

2.5.2 Enunciados

1. Se generan $n = 21$ puntos tomando

$$x_i = \frac{i - 1}{20} \quad y_i = \operatorname{erf}(x_i) \quad i = 1 \div n$$

Utilizando la rutina **SPLINE**, calcular los coeficientes de la spline cúbica interpoladora donde los valores de y_i se pueden obtener a partir de alguna librería (NAG, IMSL, etc.) de vuestro sistema.

Utilizar **SAVAL** para evaluar la spline en los puntos medios de todos los subintervalos y comparar con los valores de la función $\operatorname{erf}(x)$.

Explicitar el error en cada cálculo y dar una estimación del error máximo que se presenta.

2. Se toman $n = 21$ puntos al calcular

$$x_i = \frac{i - 1}{4} \quad y_i = J_0(x_i) \quad i = 1 \div n$$

Utilizando la rutina **SPLINE**, calcular los coeficientes de la spline cúbica interpoladora donde los valores de y_i se pueden obtener a partir de alguna librería (NAG, IMSL, etc.) de vuestro sistema.

Utilizar **SAVAL** para evaluar la spline en los puntos medios de todos los subintervalos y comparar con los valores de la función $J_0(x)$.

Explicitar el error en cada cálculo y dar una estimación del error máximo que se presenta.

3. Se consideran $n = 21$ puntos tomando

$$x_i = \frac{i - 1}{21} \quad y_i = F\left(\frac{1}{2}, \frac{1}{2}; 1; x_i\right) \quad i = 1 \div n$$

Utilizando la rutina **SPLINE**, calcular los coeficientes de la spline cúbica interpoladora donde los valores de y_i se pueden obtener a partir de alguna librería (NAG, IMSL, etc.) de vuestro sistema.

Utilizar **SAVAL** para evaluar la spline en los puntos medios de todos los subintervalos y comparar con los valores de la función $F\left(\frac{1}{2}, \frac{1}{2}; 1; x\right)$.

Explicitar el error en cada cálculo y dar una estimación del error máximo que se presenta.

4. Modificar la rutina **SPLINE** para que calcule los coeficientes de la spline cúbica natural, es decir, con las condiciones: $s''(x_1) = s''(x_n) = 0$. Aplicarlo a la práctica ejemplo comparando los resultados.

5. Modificar la rutina **SPLINE** para que calcule los coeficientes de la spline completa (las condiciones de contorno están explicitadas en el apartado 2.3).

Generar un conjunto de datos x_i , $i = 1 \div n$, igualmente espaciados en el intervalo $[0, \pi]$ con $y_i = \cos x_i$. Comparar los resultados de **SPLINE** sin modificaciones con la construida al aplicarlas a la función $\cos x$ en los puntos medios de los subintervalos. Comparar los valores próximos a los extremos. Considerar $n = 11, 21$ y 31 .

6. En el caso que los nudos tengan las abscisas equiespaciadas ($h_i = h$, $i = 1, 2, \dots, n-1$), la rutina **SPLINE** puede ser simplificada. Construir dicha rutina para nudos equiespaciados. Se considera el intervalo $[-\pi/2, \pi/2]$ con x_i , $i = 1 \div n$ igualmente espaciadas con $y_i = \sin x_i$.

Comparar los resultados de **SPLINE** sin modificaciones con la construida, para la función $\sin x$, en los puntos medios de los subintervalos. Comparar los valores próximos a los extremos. Considerar $n = 11, 21$ y 31 . ¿Cuántas operaciones requiere la rutina construida?

7. Un experimento químico ficticio produce los datos siguientes, donde se sabe que los errores son despreciables:

t	-1.000	-0.960	-0.860	-0.790	0.220	0.500	0.930
y	-1.000	-0.151	0.894	0.986	0.895	0.500	-0.306

Se quiere hacer una estimación de $y(t)$ para $t \in [-1.000, 1.000]$, interpolando los datos. El resultado tendría que dar una curva muy suave.

- (a) Dibujar una curva que interpole (de forma intuitiva).
- (b) Construir el polinomio de 6º grado que interpole los datos y representarlo gráficamente.
- (c) Mediante la rutina **SPLINE**, calcular los coeficientes de la spline cúbica interpoladora.

Utilizar la función **SAVAL** para obtener los valores de puntos interiores (3 en cada subintervalo), realizar una gráfica de la spline y comparar todos los resultados obtenidos.

8. Se quiere estudiar el efecto del fenómeno Runge sobre la función

$$f(x) = \frac{1}{1 + 25x^2} \quad x \in [-1, 1]$$

Para ello:

- (a) Construir una rutina **DIFDIV** tal que, dados el número de puntos, N , y un conjunto de nudos, (x_i, y_i) , $i = 0 \div N$, calcule los coeficientes del polinomio interpolador por el método de Newton de las diferencias divididas. Programar otra rutina **HORNER** que, dado \bar{x} y los coeficientes obtenidos con **DIFDIV**, evalúe el polinomio interpolador en \bar{x} .
 - (b) Escribir una rutina **NODOS** que construya $N+1$ nodos en dos modalidades: equidistantes y de Chebishev, para suministrarlos posteriormente a **DIFDIV** y **HORNER**.
 - (c) Dibujar las gráficas de $f(x)$ y de los polinomios interpoladores de f con 4, 5, 6, 7, 8, 9 y 10 nodos, en ambos casos. Comentar los resultados obtenidos.
9. Sean n puntos (x_i, y_i) , $i = 1 \div n$; se quiere dibujar una curva suave que pase por ellos en el orden que son dados. Un método consiste en dibujar un polígono cerrado que une los puntos respetando el orden; si se llama t a la longitud de arco a lo largo del polígono, $0 \leq t \leq T$, entonces los vértices se obtienen para los valores de t siguientes:

$$0 = t_1 < t_2 < \dots < t_n = T$$

Si se considera el conjunto de datos $(t_i, x_i) \quad i = 1 \div n$, se puede considerar la spline cúbica $x(t)$ de período T . Es decir, a cada intervalo $[t_i, t_{i+1}]$, $x(t)$ es un polinomio cúbico y $x(0) = x(T)$, $x'(0) = x'(T)$ y $x''(0) = x''(T)$.

Análogamente, se puede considerar (t_i, y_i) y la spline cúbica periódica $y(t)$. Entonces, la curva deseada viene dada en forma paramétrica para $x(t)$ y $y(t)$.

Modificar la rutina **SPLINE** para poder calcular las splines cúbicas interpoladoras $x(t)$ y $y(t)$. Probar el programa utilizando como datos iniciales los vértices de un polígono regular de n vértices con $n = 3, 4, 5$ y 6 . Comprobar la proximidad de vuestras soluciones a circunferencias.

3 Aproximación de funciones

3.1 Introducción

Si planteamos el problema de la aproximación de funciones de manera muy general, podríamos decir que consiste en, dados una función f de un cierto espacio de funciones \mathcal{F} con una norma $\|\cdot\|$ y un cierto subespacio $\mathcal{G} \subset \mathcal{F}$, hallar la función $f^* \in \mathcal{G}$ tal que

$$\forall g \in \mathcal{G}, \quad \|f - g\| \leq \|f - f^*\|$$

es decir, hallar la función de \mathcal{G} que mejor aproxima a f bajo la norma dada. Naturalmente, las funciones del subespacio serán funciones sencillas o fácilmente evaluables, por lo menos más que la función que queremos aproximar; además, también necesitaremos que este subespacio tenga dimensión finita.

Las funciones aproximadoras se escogen según el tipo de función f . Por ejemplo, si tiene algún tipo de periodicidad, se escogerán funciones trigonométricas y entonces hablaremos de **aproximación trigonométrica**; si f tiene algún tipo de comportamiento polinómico, lo aproximaremos por polinomios de grado inferior o igual a n y lo llamaremos **aproximación polinómica**; también podemos escoger funciones racionales (**aproximación racional**), o funciones exponenciales (**aproximación exponencial**).

Otro aspecto importante consiste en que, si para calcular la norma de una función necesitamos solamente los valores de la función en un número finito y determinado de puntos, entonces hablaremos de **aproximación discreta**, o bien, si necesitamos evaluarla en todo un intervalo $[a, b]$, la llamaremos **aproximación continua**.

Para funciones continuas, las normas que generalmente se utilizan son la norma infinito, o del máximo, y la norma euclíadiana:

$$\begin{aligned} \|f\|_\infty &= \max_{x \in [a, b]} |f(x)| & \|f\|_\infty &= \max_{i=1 \dots N} |f(x_i)| \\ \|f\|_2 &= \left[\int_a^b |f(x)|^2 dx \right]^{1/2} & \|f\|_2 &= \left[\sum_{i=1}^N |f(x_i)|^2 \right]^{1/2} \end{aligned}$$

la primera nos da la **aproximación min-max** y la segunda la **aproximación por mínimos cuadrados**. Muchas veces, cuando se quiere dar una importancia diferenciada a las medidas, se introducen unos pesos a estas normas:

$$\begin{aligned} \|f\|_{\infty,w} &= \max_{x \in [a,b]} w(x)|f(x)| & \|f\|_{\infty,w} &= \max_{i=1 \dots N} w_i|f(x_i)| \\ \|f\|_{2,w} &= \left[\int_a^b w(x)|f(x)|^2 dx \right]^{1/2} & \|f\|_{2,w} &= \left[\sum_{i=1}^N w_i|f(x_i)|^2 \right]^{1/2} \end{aligned}$$

donde $w(x) > 0$ sobre $[a,b]$ y $w_i > 0$.

Como se puede observar, el problema que se plantea es substancialmente diferente del de interpolación, ya que éste necesita de unos puntos, muchas veces muy bien escogidos, para poder aproximar correctamente. Por otro lado, si el número de puntos que tenemos es elevado, el grado del polinomio interpolador también será elevado y se producirán, como ya sabemos, unas grandes oscilaciones entre los puntos de interpolación.

La aproximación de funciones se utiliza en estadística por hallar funciones de un tipo determinado previamente (polinómicas, racionales, exponenciales, etc.) que aproximen lo mejor posible una nube de puntos. Es muy útil para evaluar de manera sencilla y rápida muchas funciones elementales, método que se utiliza generalmente en calculadoras y ordenadores.

3.1.1 Fundamentos teóricos

El teorema más conocido sobre aproximación de funciones es el Teorema de Weierstrass:

Teorema 3.1 Sea $f \in C[a,b]$; entonces, para cada $\varepsilon > 0$ existe un polinomio $P_n(x)$ de grado suficientemente grande, tal que

$$\max_{x \in [a,b]} |f(x) - P_n(x)| \leq \varepsilon$$

Este teorema, queda bastante lejos de lo que hemos enunciado al principio, ya que se restringe a la aproximación polinómica por la norma del máximo. Además, la demostración no es constructiva y no sabemos qué polinomio es $P_n(x)$. De todas maneras nos será útil para justificar la convergencia de este tipo de aproximación.

El primer resultado teórico que necesitamos es un teorema de existencia y unicidad de la función aproximadora.

Teorema 3.2 Sea \mathcal{F} un espacio vectorial de funciones con un producto escalar, y sean $g_1, \dots, g_n \in \mathcal{F}$ un conjunto de funciones linealmente independientes; entonces, para cada $f \in \mathcal{F}$ existe una única combinación lineal $f^* = \alpha_1 g_1 + \dots + \alpha_n g_n$ tal que

$$\|f - f^*\| \leq \|f - g\|$$

para cualquier otra combinación lineal g . (La norma $\|\cdot\|$ es la definida por el producto escalar del espacio de funciones, es decir, $\|f\|^2 = \langle f, f \rangle$).

Demostración: Para demostrar la unicidad hace falta que la norma sea estricta (ver ejercicio). Para más detalles ver ([Dav75a]) \square

Ejercicios.

1. Demostrar que, en un espacio con producto escalar, si

$$\|x + y\| = \|x\| + \|y\|$$

entonces x y y son linealmente dependientes. A una norma que cumple esta propiedad se le llama **estricta**.

2. Demostrar que la norma infinito de funciones no es estricta.

Una vez vista su existencia y unicidad, necesitamos calcular esta función que aproxima mejor. Desgraciadamente la demostración del teorema no es constructiva, pero tenemos varios métodos para obtener la función aproximadora.

Conservando la misma notación del teorema anterior, consideremos la función continua y diferenciable respecto a las a_i :

$$\|f - f^*\|^2 = \|f - \sum_{i=1}^n a_i g_i\|^2 = \langle f - f^*, f - f^* \rangle$$

desarrollando el producto escalar se obtiene

$$\left\langle f - \sum_{i=1}^n a_i g_i, f - \sum_{i=1}^n a_i g_i \right\rangle = \langle f, f \rangle - 2 \sum_{i=1}^n a_i \langle f, g_i \rangle + \sum_{i=1}^n \sum_{k=1}^n a_i a_k \langle g_i, g_k \rangle$$

Entonces para hallar el mínimo de esta función es suficiente igualar a cero sus derivadas parciales

$$\frac{\partial}{\partial a_j} \langle f, f \rangle - 2 \frac{\partial}{\partial a_j} \sum_{i=1}^n a_i \langle f, g_i \rangle + \frac{\partial}{\partial a_j} \sum_{i=1}^n \sum_{k=1}^n a_i a_k \langle g_i, g_k \rangle = 0 \quad \text{para } j = 1 \div n$$

es decir, obtenemos el sistema de ecuaciones lineales, llamadas **ecuaciones normales**:

$$\sum_{i=1}^n a_i \langle g_i, g_j \rangle = \langle f, g_j \rangle \quad \text{para } j = 1 \div n \tag{3.1}$$

que, de forma matricial, es $Ga = b$, o bien

$$\begin{pmatrix} \langle g_1, g_1 \rangle & \cdots & \langle g_1, g_n \rangle \\ \vdots & & \vdots \\ \langle g_n, g_1 \rangle & \cdots & \langle g_n, g_n \rangle \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \langle f, g_1 \rangle \\ \vdots \\ \langle f, g_n \rangle \end{pmatrix}$$

Ejercicio. Demostrar que la matriz G de las ecuaciones normales es simétrica.

Estas ecuaciones no siempre son las mejores para calcular las a_i , ya que, en muchos casos, pueden dar sistemas muy mal condicionados. Así pues, será necesario utilizar bases de funciones que den matrices lo mejor condicionadas posibles. En particular, si la base es **ortogonal**

$$\langle g_i, g_j \rangle = \begin{cases} 0 & i \neq j \\ \alpha_i & i = j \end{cases}$$

o bien, **ortonormal** ($\alpha_i = 1$), obtenemos una matriz diagonal y no hace falta resolver ningún sistema de ecuaciones, resultando entonces los siguientes valores para los coeficientes:

$$a_j = \frac{\langle f, g_j \rangle}{g_j g_j} \quad \text{para } j = 1 \div n$$

Ejemplos.

- Considerar el espacio de funciones de cuadrado integrable $L^2[0, 1]$ y el producto escalar $\langle f, g \rangle = \int_0^1 f(x)g(x) dx$. Si el espacio de dimensión finita de funciones de aproximación es el de los polinomios de grado igual o inferior a N , podemos considerar como base las funciones

$$g_i(x) = x^{i-1} \quad \text{para } i = 1 \div N + 1$$

Entonces, los productos escalares de (3.1) quedan como

$$\langle g_i, g_j \rangle = \frac{1}{i+j-1} \quad \langle f, g_i \rangle = \int_0^1 x^{i-1} f(x) dx$$

y la matriz del sistema es la matriz de Hilbert que resulta ser una matriz muy mal condicionada (ver 4).

- Considerar las funciones continuas sobre el intervalo $[-1, 1]$, el producto escalar

$$\langle f, g \rangle = \int_{-1}^1 \frac{f(x)g(x)}{(1-x^2)^{1/2}} dx$$

y los polinomios

$$g_1(x) = \frac{1}{\sqrt{\pi}}, \quad g_n(x) = \sqrt{\frac{2}{\pi}} T_n(x), \quad n = 2, 3, \dots$$

donde $T_n(x) = \cos(n \cos^{-1} x)$ son los **polinomios de Chebyshev**, que son ortogonales con la norma definida por este producto escalar. Entonces, los coeficientes a_n de la ecuación 3.1 quedan inmediatamente determinados por

$$a_1 = \frac{1}{\sqrt{\pi}} \int_{-1}^1 \frac{f(x)}{(1-x^2)^{1/2}} dx \quad a_n = \sqrt{\frac{2}{\pi}} \int_{-1}^1 \frac{f(x)T_n(x)}{(1-x^2)^{1/2}} dx$$

Ejercicios.

1. Considerar las funciones de cuadrado integrable sobre el intervalo $[-\pi, \pi]$, el producto escalar $\langle f, g \rangle = \int_{-\pi}^{\pi} f(x)g(x) dx$ y la base de funciones

$$\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos x, \frac{1}{\sqrt{\pi}} \sin x, \frac{1}{\sqrt{\pi}} \cos 2x, \dots$$

Hallar la expresión general de los coeficientes a_n .

2. Considerar $f \in C[a, b]$ evaluada en m puntos $(x_1, f_1), \dots, (x_m, f_m)$, y la norma euclídea discreta

$$\|f\|_2^2 = \sum_{i=1}^m |f_i|^2$$

Hallar la expresión de las ecuaciones normales al aproximar f por los polinomios $g_i(x) = x^{i-1}$ ($i = 1 \div n$) y con esta norma euclidiana.

3.1.2 Ortogonalización de Gram-Schmidt. Familias ortogonales

Sea $\varphi_0, \dots, \varphi_n, \dots$ una sucesión de funciones tales que cualquier subconjunto finito de estas funciones es linealmente independiente. Utilizando el método de **ortogonalización de Gram-Schmidt**, podemos obtener una nueva sucesión de funciones ortogonales u ortonormales.

Estas nuevas funciones se construyen de la manera siguiente:

$$\begin{aligned} \psi_0 &= \varphi_0 & d_0 &= \langle \psi_0, \psi_0 \rangle \\ \psi_n &= \varphi_n - \sum_{j=0}^{n-1} \frac{\langle \varphi_n, \psi_j \rangle}{d_j} \psi_j & d_n &= \langle \psi_n, \psi_n \rangle \quad \text{para } n = 1, \dots \end{aligned} \quad (3.2)$$

Ejercicios.

1. Comprobar que las funciones ψ_n obtenidas por el método de Gram-Schmidt son ortogonales.
 2. Demostrar que todo conjunto de funciones ortogonales son linealmente independientes.

Si ahora definimos las funciones

$$\bar{\psi}_n = \frac{\psi_n}{\sqrt{d_n}} \quad \text{para } n = 0, 1, \dots$$

tenemos un conjunto de funciones ortonormales.

Para el caso particular de las familias de polinomios ortogonales podemos deducir una recurrencia bastante sencilla. El primer polinomio será de grado cero $\psi_0(x) = b_0 \neq 0$; suponemos, ahora, que tenemos $n+1$ polinomios ortogonales, $\psi_0(x), \dots, \psi_n(x)$, de grado igual a su subíndice

y queremos calcular el siguiente $\psi_{n+1}(x)$. Consideramos $\varphi_{n+1}(x) = \alpha_n x \psi_n(x)$ y aplicamos la fórmula 3.2 para hallar $\psi_{n+1}(x)$

$$\psi_{n+1} = \varphi_{n+1} - \sum_{j=0}^n \frac{\langle \varphi_{n+1}, \psi_j \rangle}{d_j} \psi_j$$

Estudiamos por separado los productos escalares $\langle \varphi_{n+1}, \psi_j \rangle$

$$\langle \varphi_{n+1}, \psi_j \rangle = \langle \alpha_n x \psi_n, \psi_j \rangle = \alpha_n \langle x \psi_n, \psi_j \rangle = \alpha_n \langle \psi_n, x \psi_j \rangle$$

Si expresamos $x \psi_j$, polinomio de grado $j+1 \leq n+1$, como combinación lineal de la familia de polinomios ortogonales (recordemos que son linealmente independientes) y descomponemos el producto escalar en suma de productos escalares, tenemos

$$\begin{aligned}\langle \varphi_{n+1}, \psi_j \rangle &= 0 \quad \text{para } j = 0 \div n-2 \\ \langle \varphi_{n+1}, \psi_{n-1} \rangle &= \alpha_n \langle \psi_n, x \psi_{n-1} \rangle \\ \langle \varphi_{n+1}, \psi_n \rangle &= \alpha_n \langle \psi_n, x \psi_n \rangle\end{aligned}$$

Por lo tanto, obtenemos finalmente la recurrencia

$$\psi_{n+1}(x) = \alpha_n(x - \beta_n) \psi_n(x) - \gamma_n \psi_{n-1}(x) \quad (3.3)$$

donde

$$\begin{aligned}\alpha_n &= \frac{b_{n+1}}{b_n} \quad \text{para } n \geq 0 \\ \beta_n &= \frac{\langle \psi_n, x \psi_n \rangle}{\langle \psi_n, \psi_n \rangle} \quad \text{para } n \geq 0 \\ \gamma_n &= \frac{\alpha_n \langle \psi_n, \psi_n \rangle}{\alpha_{n-1} \langle \psi_{n-1}, \psi_{n-1} \rangle} \quad \text{para } n \geq 1\end{aligned}$$

Esta fórmula es válida para todo $n \geq 0$ tomando $\psi_{-1}(x) = 0$. Si queremos que los polinomios sean mónicos, solamente hace falta hacer $\alpha_n = 1$.

Ejemplos.

1. Los polinomios de Legendre. Consideramos el intervalo $[-1, 1]$ i el producto escalar euclíadiano. Definimos los polinomios

$$P_0(x) = 1, \quad P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} [(x^2 - 1)^n] \quad \text{para } j \geq 1.$$

Estos polinomios son ortogonales; en efecto, si $i \neq j$, integrando por partes

$$\langle P_i, P_j \rangle = \int_{-1}^1 P_i(x) P_j(x) dx = 0$$

y si $i = j$, por inducción se demuestra que

$$\langle P_i, P_i \rangle = \frac{(2i)!}{(2^i i!)^2} \int_{-1}^1 (1 - t^2) dx = \frac{2}{2i + 1}$$

De la definición se deduce que

$$P_n(-x) = (-1)^n P_n(x)$$

Esta relación nos asegura que los coeficientes β_n son todos iguales a cero. También, de la definición se comprueba que el coeficiente principal del polinomio es

$$b_n = \frac{2n(2n-1)\cdots(n+1)}{2^n n!} \quad \text{para } n \geq 0$$

De todo esto, deducimos que

$$\alpha_n = \frac{2n+1}{j+1} \quad \beta_n = 0$$

$$\gamma_n = \frac{\alpha_n \langle P_n, P_n \rangle}{\alpha_{n-1} \langle P_{n-1}, P_{n-1} \rangle} = \frac{n}{n+1} \quad \text{para } j \geq 0$$

y nos queda finalmente, según la fórmula 3.3, la recurrencia

$$\begin{aligned} P_0(x) &= 1 & P_1(x) &= x \\ P_{n+1}(x) &= \frac{1}{n+1} [(2n+1)xP_n(x) - nP_{n-1}(x)] \quad \text{para } j = 1, \dots \end{aligned}$$

2. Los polinomios de Gram.

Consideramos el producto escalar

$$\langle f, g \rangle = \sum_{i=0}^m f(x_i)g(x_i) \quad \text{para } x_i = i = 1 \div m$$

Los polinomios ortogonales para este producto escalar son los polinomios de Gram:

$$P_{km}(x) = \sum_{i=0}^k (-1)^i \binom{k}{i} \binom{k+1}{i} \frac{x(x-1)\cdots(x-i+1)}{m(m-1)\cdots(m-i+1)} \quad j = 0 \div m$$

Más concretamente, su norma queda determinada por

$$\langle P_{km}, P_{km} \rangle = \frac{(m+k+1)!(m-k)!}{(2k+1)!(m!)^2} \quad k = 0 \div m$$

Se puede demostrar que su relación de recurrencia es

$$P_{0m}(x) = 1 \quad P_{1m}(x) = 1 - \frac{2x}{m}$$

$$\begin{aligned} (k+1)(m-k)P_{k+1,m}(x) &= (2k+1)(m-2x)P_{km}(x) - \\ &\quad k(m+k+1)P_{k-1,m}(x) \quad k = 1 \div m-1 \end{aligned}$$

Los polinomios ortogonales bajo un producto escalar cumplen una importante propiedad:

Teorema 3.3 El polinomio $P_n(x)$ de grado n de una familia ortogonal de polinomios respecto el producto escalar

$$\langle f, g \rangle = \int_a^b w(x) f(x)g(x) dx$$

tiene n ceros simples en el interior del intervalo $[a, b]$.

Demostración: Suponemos que $P_n(x)$ tiene solamente k cambios de signo en el interior de $[a, b]$ en los puntos t_1, \dots, t_k , con $0 \leq k < n$. Consideramos el nuevo polinomio

$$Q(x) = P_n(x) \prod_{i=1}^k (x - t_i) \quad (Q(x) = P_n(x) \text{ si } k = 0)$$

Está claro que este polinomio tiene signo constante en todo el intervalo $[a, b]$ y, por lo tanto

$$\left\langle P_n, \prod_{i=1}^k (x - t_i) \right\rangle = \int_a^b w(x) P_n(x) \prod_{i=1}^k (x - t_i) dx \neq 0$$

Por otro lado, esto está en contradicción con el hecho de que el producto escalar del polinomio ortogonal de grado n con cualquier otro polinomio de grado estrictamente inferior a n ha de ser cero. Entonces, $P_n(x)$ ha de tener n cambios de signo en el interior de el intervalo $[a, b]$, es decir, n raíces simples. \square

3.2 Aproximación mínimo-cuadrática polinómica

Dada $f \in \mathcal{C}[a, b]$, queremos hallar la mejor aproximación polinómica de grado igual o inferior a n , $P_n(x)$, con la norma $\|\cdot\|_2$. Por ser esta norma estricta, la existencia y unicidad de la aproximación queda asegurada por el teorema 3.2.

Como ya se ha visto en un ejemplo anterior, según qué base de polinomios escogemos, la matriz de productos escalares (ecuación 3.1) puede ser una matriz muy mal condicionada, a pesar de ser simétrica. Por lo tanto, este no será un buen procedimiento para hallar la aproximación mínimo-cuadrática. Como ya habíamos comentado, hace falta utilizar bases ortogonales o ortonormales.

3.2.1 Aproximación continua por polinomios ortogonales

Consideraremos un conjunto de $n + 1$ polinomios $P_j(x)$ ($j = 0 \div n$) ortogonales por la norma

$$\|f\|_{2,w}^2 = \int_a^b w(x)|f(x)|^2 dx$$

con $w(x) \geq 0$, continua a $[a, b]$ y que denominaremos función peso.

Denotamos por $P_n^*(x)$ la mejor aproximación polinómica de una función $f \in \mathcal{C}[a, b]$ por la norma $\|\cdot\|_{2,w}$ anteriormente definida. Siendo el conjunto de polinomios ortogonales linealmente independientes, podemos escribir

$$P_n^* = \sum_{j=0}^n c_j^* P_j$$

Las ecuaciones normales, en este caso, quedan

$$c_j^* \|P_j\|_{2,w}^2 = \int_a^b w(x) P_j(x) f(x) dx \quad \text{para } j = 0 \div n$$

y los coeficientes serán

$$c_j^* = \frac{1}{\|P_j\|_{2,w}^2} \int_a^b w(x) P_j(x) f(x) dx \quad \text{para } j = 0 \div n$$

Ejemplo. Consideramos la función $f(x) = e^x$ en el intervalo $[-1, 1]$ y la aproximamos por polinomios de Legendre.

Como ya hemos visto, estos polinomios son ortogonales por el producto escalar

$$\langle f, g \rangle = \int_{-1}^1 f(x) g(x) dx$$

es decir,

$$\langle P_i, P_j \rangle = \begin{cases} 0 & y \neq j \\ \frac{2}{2i+1} & i = j \end{cases}$$

Por lo tanto, la aproximación mínimo-cuadrática será $P_n^*(x) = \sum_{j=0}^n c_j^* P_j(x)$, donde

$$c_j^* = \frac{2j+1}{2} \int_{-1}^1 P_j(x) e^x dx$$

Si calculamos los cuatro primeros coeficientes, tenemos

$$\begin{aligned} c_0^* &= 1.17520119 & c_1^* &= 1.10363832 \\ c_2^* &= 0.35781435 & c_3^* &= 0.07045563 \end{aligned}$$

En la figura 3.1 podemos ver el error de aproximar por los cuatro primeros polinomios de Legendre. Se debe observar que hay cuatro cambios de signo de la función de error al aproximar por un polinomio de tercer grado.

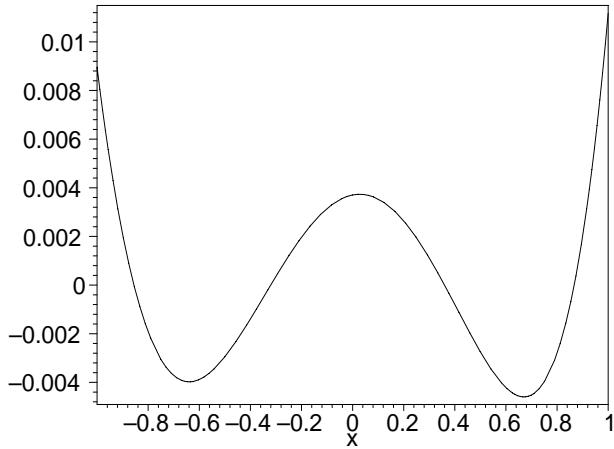


Fig. 3.1 Error de la aproximación mínimo-cuadrática de e^x por polinomios de Legendre.

3.2.2 Aproximación discreta

Para el caso de la aproximación mínimo-cuadrática discreta, se utiliza el producto escalar

$$\langle f, g \rangle = \sum_{i=1}^m w_i f(x_i)g(x_i) \quad \text{con } w_i \geq 0$$

donde los valores w_i son ciertos pesos determinados. La norma derivada de este producto escalar la podemos escribir como

$$\|f\|_{2,w}^2 = \sum_{i=1}^m w_i |f(x_i)|^2 = \|W\mathbf{f}\|_2^2$$

donde $W = \text{diag}(w_1, \dots, w_m)$ y $\mathbf{f} = (f(x_1), \dots, f(x_m))^T$.

Con esta notación, la mejor aproximación de una función f por funciones del tipo $\sum_{j=0}^n a_j g_j$, con $n \ll m$ y las g_j linealmente independientes, será tal que

$$\|W\mathbf{r}\|_2 = \|W(\sum_{j=0}^n a_j \mathbf{g}_j - \mathbf{f})\|_2 = \|W(\sum_{j=0}^n a_j \mathbf{g}_j - W\mathbf{f})\|_2 = \|WG\mathbf{a} - W\mathbf{f}\|_2$$

sea mínima, donde un elemento (i, j) de la matriz G es $g_j(x_i)$, $\mathbf{a} = (a_0, \dots, a_n)^T$ y $\mathbf{g}_j = (g_j(x_1), \dots, g_j(x_m))^T$.

Si consideramos la descomposición en valores singulares de la matriz WG (ver el apéndice A) $WG = UEV^T$, y teniendo en cuenta que el producto por una matriz ortogonal no altera la norma euclíadiana, podemos escribir

$$\|W\mathbf{r}\|_2 = \|U^T W\mathbf{r}\|_2 = \|(U^T WGV)V^T \mathbf{a} - U^T W\mathbf{f}\|_2 = \|E\mathbf{z} - \mathbf{d}\|_2$$

donde

$$\mathbf{z} = V^T \mathbf{a} \quad \mathbf{d} = U^T W\mathbf{f}$$

Esta nueva expresión de la norma del error tiene la ventaja que, siendo Σ una matriz de la forma

$$E = \begin{pmatrix} D \\ 0 \end{pmatrix}$$

con $D = \text{diag}(\sigma_1, \dots, \sigma_n)$, para hacer mínimo $\|W\mathbf{r}\|_2$, es suficiente considerar

$$\begin{aligned} z_i &= \frac{d_i}{\sigma_i} && \text{si } \sigma_i \neq 0 \\ z_i &= \text{cualquier valor, } && \text{si } \sigma_i = 0 \end{aligned}$$

Cuando hay algún $\sigma_i = 0$, es decir, cuando el rango de WG no es n , la solución no es única; entonces, podemos escoger, por ejemplo, la solución de menor módulo o, lo que es lo mismo, hacer las correspondientes $z_i = 0$. Una vez tenemos el vector \mathbf{z} , calculamos la combinación lineal que es solución del problema de mínimos cuadrados haciendo $\mathbf{a} = V\mathbf{z}$.

Naturalmente, para calcular numéricamente la descomposición en valores singulares de la matriz WG utilizaremos el algoritmo que se comenta en el capítulo 7.

Queda claro que este algoritmo no tiene ninguna utilidad en el caso de realizar la aproximación mínimos cuadráticos discretos por alguna familia de polinomios ortogonales, ya que entonces, es mucho mejor utilizar las ecuaciones normales directamente. Su interés radica en el hecho de solucionar aproximaciones mucho más generales, incluso en los casos donde haya algún problema de rango.

La práctica resuelta es un ejemplo de aplicación de este procedimiento para calcular una aproximación mínimos cuadráticos discretos.

3.3 Aproximación min-max de funciones

La aproximación **min-max** de una función es el caso particular de mejor aproximación utilizando la norma infinito o del máximo. Es decir, dada una función f continua sobre el intervalo $[a, b]$, buscamos una función h^* , también continua pero de un cierto subespacio de dimensión finita $\mathcal{H} \subset \mathcal{C}[a, b]$ (polinomios, funciones racionales, etc.), tal que

$$\forall h \in \mathcal{H}, \quad \|h^* - f\|_\infty \leq \|h - f\|_\infty$$

La existencia de esta aproximación viene dada por el teorema 3.2, ya que el primer apartado sirve para cualquier espacio normado; pero, en cambio, la demostración de la unicidad utiliza

implícitamente el producto escalar o, mejor dicho, que la norma sea estricta, una propiedad que la norma infinito no tiene.

3.3.1 Aproximación polinómica continua

Como ya se ha visto anteriormente, el problema consiste en, dada una función $f \in \mathcal{C}[a, b]$, hallar un polinomio de grado igual o inferior a n , P_n^* , tal que

$$\max_{x \in [a, b]} |P_n^*(x) - f(x)| \leq \max_{x \in [a, b]} |P_n(x) - f(x)| \quad \forall P_n \in \mathcal{P}_n \quad (3.4)$$

donde \mathcal{P}_n denota el espacio de polinomios de grado inferior o igual a n . Esta aproximación queda caracterizada por el teorema de Chebyshev:

Teorema 3.4 Un polinomio $P_n \in \mathcal{P}_n$ es una aproximación min-max de la función $f \in \mathcal{C}[a, b]$ si, y solamente si, $f(x) - P_n(x)$ toma los valores $\pm \|f - P_n\|_\infty$, como mínimo $n + 2$ veces dentro del intervalo $[a, b]$ y con signos alternados.

Demostración: ver ([Dav75a]) \square

El hecho de que $f(x) - P_n(x) = \pm \|f - P_n\|_\infty$ como a mínimo en $n + 2$ puntos y con signos alternados quiere decir que la función $f - P_n$ tiene, por lo menos, $n + 1$ ceros en el intervalo $[a, b]$ y, por lo tanto, P_n es el polinomio de grado n que interpola la función f en estos $n + 1$ puntos.

Utilizando este teorema podemos demostrar la unicidad. Ver ([Dav75a]).

También podemos enunciar un teorema de convergencia de las aproximaciones min-max polinómicas.

Teorema 3.5 Dadas una función $f \in \mathcal{C}[a, b]$ y una sucesión de aproximaciones polinómicas min-max de f , $P_n^*(x)$ para $n = 1, 2, \dots$; entonces,

$$\lim_{n \rightarrow \infty} P_n^*(x) = f(x)$$

Demostración: Este teorema es una consecuencia inmediata del teorema de Weierstrass (teorema 3.1). Sea P_n un polinomio dado por este teorema; como P_n^* es la mejor aproximación con la norma del máximo, también ha de cumplir la misma desigualdad

$$\max_{x \in [a, b]} |f(x) - P_n(x)| < \varepsilon$$

desigualdad que nos asegura la convergencia. \square

3.3.2 Aproximación polinómica discreta

Consideramos un conjunto de puntos $X_m = \{x_1 < \dots < x_m\}$ y una función f conocida en estos puntos; como ya se sabe, si $m \leq n + 1$, existe un único polinomio $P_n^* \in \mathcal{P}_n$ tal que

$$P_n^*(x_i) = f(x_i) \quad \text{para } i = 1 \div m$$

y, además, este P_n^* es la mejor aproximación discreta con la norma infinito.

Pero si $m \geq n + 2$, ya no podemos utilizar la interpolación; de todas maneras, tenemos un teorema de caracterización de la mejor aproximación min-max polinómica discreta de f . Denotaremos por $P_n^*(X_m)$ el polinomio de grado n que aproxima la función f con la norma infinito sobre los puntos X_m .

Teorema 3.6 $P_n^*(X_m)$ es la aproximación min-max discreta de f si, y solamente si, existe X_{n+2}^* , subconjunto de $n + 2$ puntos de X_m , tal que

$$f(x) - P_n^*(X_m; x) = \pm \max_{x \in X_{n+2}} |f(x) - P_n^*(x)| = \pm E_n(f; X_m)$$

en estos $n + 2$ puntos y con signos alternados.

Demostración: Es semejante a la del teorema de Chebyshev. \square

Este último teorema, tal como se puede observar, es equivalente al teorema de Chebyshev de la aproximación continua, pero ahora para la discreta. De la misma manera, se puede deducir la unicidad de la aproximación min-max discreta.

Ejemplo. Consideramos el conjunto de puntos $X_5 = \{-1, -0.5, 0, 0.5, 1\}$ y la función $f(x) = e^x$ que queremos aproximar min-max por un polinomio de segundo grado. Si denotamos el polinomio por $p(x) = a_2 x^2 + a_1 x + a_0$ y los cinco posibles subconjuntos de X_5 (uno de ellos será el citado en el teorema 3.6) por

$$\begin{array}{lll} X_{5,1} = \{-1, -0.5, 0, 0.5\} & X_{5,2} = \{-1, -0.5, 0, 1\} \\ X_{5,3} = \{-1, -0.5, 0.5, 1\} & X_{5,4} = \{-1, 0, 0.5, 1\} \\ & X_{5,5} = \{-0.5, 0, 0.5, 1\} \end{array}$$

podemos plantear cinco sistemas de dimensión cuatro de la forma

$$a_0 + x_i a_1 + x_i^2 a_2 + (-1)^i E = e^{x_i} \quad \text{para } i = 1 \div 4$$

con $x_i \in X_{5,j}$ y donde $(-1)^i E$ será un error con signo alternado sobre los cuatro puntos escogidos. Una vez efectuados todos los cálculos, podemos evaluar los cinco polinomios en los cinco puntos de X_m y evaluar el error respecto e^x en estos puntos

	$X_{5,1}$	$X_{5,2}$	$X_{5,3}$	$X_{5,4}$	$X_{5,5}$
-1.0	0.01255	0.02594	0.04434	0.03318	0.24532
-0.5	-0.01255	-0.02594	-0.04434	-0.06666	0.02070
0.0	0.01255	0.02594	-0.01086	-0.03318	-0.02070
0.5	-0.01255	0.08113	0.04434	0.03318	0.02070
1.0	-0.25347	-0.02594	-0.04434	-0.33178	-0.02070

Observamos que todas las columnas tienen signos alternados en los puntos utilizados en cada caso para al cálculo del polinomio, pero solamente una columna tiene un error inferior a los otros en el punto no utilizado: la tercera. Por lo tanto, esta es la aproximación min-max de $f(x) = e^x$ por un polinomio de segundo grado y los puntos especiales son -1, -0.5, 0.5 y 1.

Como se puede suponer, la llave de los algoritmos para calcular la aproximación min-max polinómica discreta consistirá en hallar estos $n + 2$ puntos particulares que asegura el teorema 3.6.

3.4 Problemas

1. Calcular la fórmula recurrente de los polinomios de Legendre para el intervalo $[a, b]$.
2. Los polinomios de Chebyshev se definen como

$$T_0(x) = 1 \quad T_n(x) = \cos(n \arccos x) \quad \text{para } n \geq 1, x \in [-1, 1]$$

Demostrar que

- (a) son ortogonales respecto al producto escalar continuo

$$\langle f, g \rangle = \int_{-1}^1 \frac{f(x)g(x)}{(1-x^2)^{1/2}} dx$$

- (b) T_0, \dots, T_m son ortogonales respecto el producto escalar discreto

$$\langle f, g \rangle_m = \sum_{i=0}^m f(x_i)g(x_i)$$

donde $x_i = \cos \frac{2i+1}{m+1} \frac{\pi}{2}$ son los $m+1$ ceros del polinomio T_{m+1} .

- (c) la relación de recurrencia es

$$T_0(x) = 1 \quad T_1(x) = x$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \quad n \geq 1$$

y el coeficiente principal de T_n es 2^{n-1} .

3. Considerar n funciones linealmente independientes $\varphi_1, \dots, \varphi_n$ y las correspondientes funciones ortogonales ψ_1, \dots, ψ_n . Hallar la mejor aproximación de φ_n dentro del espacio de funciones generado pero $\psi_1, \dots, \psi_{n-1}$.
4. Calcular la aproximación min-max cuadrática de la función $f(x) = \sqrt{x}$ sobre los puntos $\{0, 1/4, 1/2, 3/4, 1\}$.
5. Demostrar que la aproximación min-max de x^{n+1} sobre \mathcal{P}_n es, de hecho, un polinomio de grado $n-1$.
6. Calcular la aproximación min-max continua de x^{n+2} sobre \mathcal{P}_n .
7. Sea f^* la mejor aproximación de la función f dentro del espacio \mathcal{F} generado por las funciones ψ_0, \dots, ψ_N , ortonormales por el producto escalar $\langle \cdot, \cdot \rangle$. Demostrar que para cualquier $g \in \mathcal{F}$ la mejor aproximación de $f - g$ es $f^* - g$.
8. Demostrar que, si $P_n^* \in \mathcal{P}_n$ es la aproximación min-max de f , entonces la aproximación min-max de $f - p$ dentro de \mathcal{P}_n es $P_n^* - p$, para cualquier $p \in \mathcal{P}_n$.
9. Sea q_n^* la aproximación mínimo-cuadrática discreta de f sobre X_m ; demostrar que $f - q_n^*$ tiene signo alternado en $n+2$ puntos de X_m .

3.5 Prácticas

3.5.1 Práctica ejemplo

En la tabla siguiente tenemos el número de bacterias por unidad de volumen en función del tiempo transcurrido

Horas x	0	1	2	3	4	5	6
Bacterias y	32	47	65	92	132	190	275

Calcular una curva del tipo $y = ab^x$ que aproxime esta nube de puntos. Hacer una predicción del número de bacterias al cabo de 7 horas.

Si aplicamos logaritmos a la ecuación de la curva aproximadora, tenemos

$$\log y = \log a + x \log b$$

es decir, que tenemos una nueva curva que ahora es combinación lineal de 1 y x con coeficientes $\log a$ y $\log b$ respectivamente. Esta será la curva que aproximarán por el método de mínimos cuadrados los puntos $(x_i, \log y_i)$.

Utilizaremos lo que hemos comentado sobre aproximación mínimo-cuadrática discreta, es decir, la descomposición en valores singulares. En el programa que presentamos a continuación se utiliza la rutina de descomposición en valores singulares presentada en [For77] y según el algoritmo descrito en el Capítulo 7.

```

PROGRAM BACTERIA
C
PARAMETER NM=10
REAL A(NM,2),X(NM),Y(NM),D(2),U(NM,2),V(NM,2),WORK(2),B(2),C(2)
REAL LA,LB,AA,BB,CC
INTEGER I,J,IERR,M,N
C
READ (1,*) M
READ (1,*) (X(I),Y(I),I=1,M)
READ (1,*) CC
DO 10 I=1,M
    A(I,1) = 1.0
    A(I,2) = X(I)
    Y(I) = LOG(Y(I))
10    CONTINUE
C
N = 2
CALL SVD(NM,M,N,A,D,.TRUE.,U,.TRUE.,V,IERR,WORK)
IF (IERR.NE.0) WRITE (2,100) IERR
C
DO 20 I=1,2
    B(I) = 0.
    DO 22 J=1,M
        B(I) = B(I)+U(J,I)*Y(J)
22    CONTINUE
20

```

22

```

20      CONTINUE
DO 30 I=1,2
      IF (D(I).NE.0) THEN
          C(I) = B(I)/D(I)
      ELSE
          C(I) = 0
      ENDIF
30      CONTINUE
      LA = 0.
      LB = 0.
      DO 40 I=1,2
          LA = LA+V(1,I)*C(I)
          LB = LB+V(2,I)*C(I)
40      CONTINUE
      AA = EXP(LA)
      BB = EXP(LB)
      WRITE (2,200) AA,BB
      WRITE (2,300) AA*BB**CC
      STOP

100     FORMAT(' ERROR IERR = ',I4)
200     FORMAT(' COEFICIENTES: A = ',F15.10,'      B = ',F15.10)
300     FORMAT(' EVALUACION PEDIDA: ',F10.2)
      END

SUBROUTINE SVD(NM,M,N,A,W,MATU,U,MATV,V,IERR,RV1)
C
C      INTEGER I,J,K,K1,L1,M,N,MN,NM,ITS,IERR
C      DOUBLE PRECISION A(NM,N),W(N),U(NM,N),V(NM,N),RV1(N)
C      DOUBLE PRECISION C,F,G,H,S,X,Y,Z,SCALE,ANORM
C      LOGICAL MATU,MATV
C
C      ESTA SUBRUTINA CALCULA LA DESCOMPOSICION EN VALORES SINGULARES
C      A=U SIGMA Vt DE UNA MATRIZ REAL RECTANGULAR M X N.
C      SE UTILIZA BIDIAGONALIZACION DE HOUSEHOLDER Y UNA VARIANTE
C      DEL ALGORITMO QR.
C
C      PARAMETROS DE ENTRADA:
C
C      NM :      HA DE SER IGUAL AL NUMERO DE FILAS DE LAS MATRICES
C              DECLARADAS EN EL PROGRAMA DE LLAMADA. DEBE VALER,
C              COMO MINIMO EL MAXIMO DE M I N.
C      M :      NUMERO DE FILAS DE A (I U).
C      N :      NUMERO DE COLUMNAS DE A (I U) Y ORDEN DE V.
C      A :      CONTIENE LA MATRIZ RECTANGULAR DE ENTRADA QUE SE QUIERE
C              DESCOMPONER.
C      MATU :    DEBE VALER .TRUE. SI SE QUIERE OBTENER LA MATRIZ U
C              DE LA DESCOMPOSICION, Y .FALSE. EN CASO CONTRARIO.
C      MATV :    DEBE VALER .TRUE. SI SE QUIERE OBTENER LA MATRIZ V
C              DE LA DESCOMPOSICION, Y .FALSE. EN CASO CONTRARIO.
C
C      PARAMETROS DE SALIDA:
C
C      A :      INALTERADA (SI NO ES SOBREESCRITA POR U O V).
C      W :      CONTIENE LOS N VALORES SINGULARES DE A (NO NEGATIVOS)
C              QUE SON LOS ELEMENTOS DIAGONALES DE SIGMA. NO ESTAN
C              ORDENADOS. SI SE SALE CON IERR <> 0, LOS VALORES

```

```

C           SINGULARES CORRESPONDIENTES A LOS INDICES IERR+1, IERR+2,
C           ... , N HAN DE SER CORRECTOS.
C   U :      CONTIENE LA MATRIZ U (VECTORES COLUMNAS ORTOGONALES) DE LA
C           DESCOMPOSICION SI MATU ERA .TRUE. SI NO, U SE UTILIZA
C           COMO MATRIZ DE TRABAJO TEMPORAL. U PUEDE COINCIDIR
C           CON A. SI SE SALE CON IERR <> 0, LAS COLUMNAS
C           DE U CORRESPONDIENTES A LOS INDICES DE LOS VALORES SINGULARES
C           CORRECTOS TAMBIEN DEBEN SER CORRECTOS.
C   V :      CONTIENE LA MATRIZ V (ORTOGONAL) DE LA DESCOMPOSICION SI
C           MATV ERA .TRUE. SI NO, NO SE UTILIZA. V TAMBIEN PUEDE
C           COINCIDIR CON A SI NO SE QUIERE CALCULAR U. SI SE SALE
C           CON IERR <> 0, LAS COLUMNAS DE V CORRESPONDIENTES A LOS
C           INDICES DE LOS VALORES SINGULARES CORRECTOS TAMBIEN DEBEN
C           SER CORRECTOS.
C   IERR :    CERO:     INDICA FINAL NORMAL
C           K>0 :     INDICA QUE EL TEST DE CONVERGENCIA PARA EL VALOR
C                   SINGULAR K-ESIMO NO SE HA SUPERADO DESPUES DE
C                   30 ITERACIONES.
C   RV1 :     MATRIZ DE TRABAJO TEMPORAL.
C
C   IERR = 0
C
C       DO 100 I = 1, M
C       DO 100 J = 1, N
C           U(I,J) = A(I,J)
100   CONTINUE
C   .....REDUCCION DE HOUSEHOLDER A FORMA BIDIAGONAL.....
C   G = 0.DO
C   SCALE = 0.DO
C   ANORM = 0.DO
C
C   DO 300 I = 1, N
C       L = I + 1
C       RV1(I) = SCALE * G
C       G = 0.DO
C       S = 0.DO
C       SCALE = 0.DO
C       IF (I .GT. M) GO TO 210
C
C   DO 120 K = I, M
120   SCALE = SCALE + DABS(U(K,I))
C
C   IF (SCALE .EQ. 0.0) GO TO 210
C
C   DO 130 K = I, M
C       U(K,I) = U(K,I) / SCALE
C       S = S + U(K,I)**2
130   CONTINUE
C
C   F = U(I,I)
C   G = -DSIGN(DSQRT(S),F)
C   H = F * G - S
C   U(I,I) = F - G
C   IF (I .EQ. N) GO TO 190
C
C   DO 150 J = L, N
C       S = 0.0
C
C   DO 140 K = I, M

```

```

140      S = S + U(K,I) * U(K,J)
C
      F = S / H
C
      DO 150 K = I, M
         U(K,J) = U(K,J) + F * U(K,I)
150    CONTINUE
C
190    DO 200 K = I, M
200    U(K,I) = SCALE * U(K,I)
C
210    W(I) = SCALE * G
      G = 0.0
      S = 0.0
      SCALE = 0.0
      IF (I .GT. M .OR. I .EQ. N) GO TO 290
C
      DO 220 K = L, N
220    SCALE = SCALE + ABS(U(I,K))
C
      IF (SCALE .EQ. 0.0) GO TO 290
C
      DO 230 K = L, N
         U(I,K) = U(I,K) / SCALE
         S = S + U(I,K)**2
230    CONTINUE
C
      F = U(I,L)
      G = -DSIGN(DSQRT(S),F)
      H = F * G - S
      U(I,L) = F - G
C
      DO 240 K = L, N
240    RV1(K) = U(I,K) / H
C
      IF (I .EQ. M) GO TO 270
C
      DO 260 J = L, M
         S = 0.0
C
         DO 250 K = L, N
250    S = S + U(J,K) * U(I,K)
C
         DO 260 K = L, N
            U(J,K) = U(J,K) + S * RV1(K)
260    CONTINUE
C
270    DO 280 K = L, N
280    U(I,K) = SCALE * U(I,K)
C
290    ANORM = DMAX1(ANORM,DABS(W(I))+DABS(RV1(I)))
300 CONTINUE
C      .....ACUMULACION DE LAS TRANSFORMACIONES POR LA DERECHA.....
IF (.NOT. MATV) GO TO 410
DO 400 I = N, 1, -1
   IF (I .EQ. N) GO TO 390
   IF (G .EQ. 0.0) GO TO 360
C
   DO 320 J = L, N

```

```

C      .....DOBLE DIVISION PARA EVITAR UN POSIBLE UNDERFLOW.....
320  V(J,I) = (U(I,J) / U(I,L)) / G
C
      DO 350 J = L, N
      S = 0.0
C
      DO 340 K = L, N
340  S = S + U(I,K) * V(K,J)
C
      DO 350 K = L, N
      V(K,J) = V(K,J) + S * V(K,I)
350  CONTINUE
C
360  DO 380 J = L, N
      V(I,J) = 0.0
      V(J,I) = 0.0
380  CONTINUE
C
390  V(I,I) = 1.0
      G = RV1(I)
      L = I
400 CONTINUE
C      .....ACUMULACION DE LAS TRANSFORMACIONES POR LA IZQUIERDA.....
410 IF(.NOT. MATU) GO TO 510
C      .....PARA I=MIN(M,N) PASO -1 HASTA 1 HACER --.....
      MN = MIN(M,N)
      DO 500 I = MN, 1, -1
      L = I + 1
      G = W(I)
      IF (I .EQ. N) GO TO 430
C
      DO 420 J = L, N
420  U(I,J) = 0.0
C
430  IF (G .EQ. 0.0) GO TO 475
      IF (I .EQ. MN) GO TO 460
C
      DO 450 J = L, N
      S = 0.0
C
      DO 440 K = L, M
440  S = S + U(K,I) * U(K,J)
C      .....DOBLE DIVISION PARA EVITAR UN POSIBLE UNDERFLOW.....
      F = (S / U(I,I)) / G
C
      DO 450 K = I, M
      U(K,J) = U(K,J) + F * U(K,I)
450  CONTINUE
C
460  DO 470 J = I, M
470  U(J,I) = U(J,I) / G
C
      GO TO 490
C
475  DO 480 J = I, M
480  U(J,I) = 0.0
C
490  U(I,I) = U(I,I) + 1.0
500 CONTINUE

```

```

C
C      .....DIAGONALIZACION DE LA FORMA BIDIAGONAL.....
C
 510 DO 700 K = N, 1, -1
      K1 = K - 1
      ITS = 0
C
C      .....TEST POR SPLITTING.....
 520  DO 530 L = K, 1, -1
      L1 = L - 1
      IF (DABS(RV1(L)) + ANORM .EQ. ANORM) GO TO 565
C      .....RV1(1) ES SIEMPRE CERO, POR LO TANTO NO SE SALE
C          POR EL FINAL DEL BUCLE.....
      IF (DABS(W(L1)) + ANORM .EQ. ANORM) GO TO 540
 530  CONTINUE
C      .....CANCELACION DE RV1(L) SI L MAYOR QUE 1.....
 540  C = 0.D0
      S = 1.D0
C
      DO 560 I = L, K
          F = S * RV1(I)
          RV1(I) = C * RV1(I)
          IF (DABS(F) + ANORM .EQ. ANORM) GO TO 565
          G = W(I)
          H = DSQRT(F*F+G*G)
          W(I) = H
          C = G / H
          S = -F / H
          IF (.NOT. MATU) GO TO 560
C
      DO 550 J = 1, M
          Y = U(J,L1)
          Z = U(J,I)
          U(J,L1) = Y * C + Z * S
          U(J,I) = -Y * S + Z * C
 550  CONTINUE
C
 560  CONTINUE
C      .....TEST DE CONVERGENCIA.....
 565  Z = W(K)
      IF (L .EQ. K) GO TO 650
C      .....SHIFT DE LA CAJA 2X2 INFERIOR.....
      IF (ITS .EQ. 30) GO TO 1000
      ITS = ITS + 1
      X = W(L)
      Y = W(K1)
      G = RV1(K1)
      H = RV1(K)
      F = ((Y - Z) * (Y + Z) + (G - H) * (G + H)) / (2.0 * H * Y)
      G = DSQRT(F*F+1.0)
      F = ((X - Z) * (X + Z) + H * (Y / (F + DSIGN(G,F)) - H)) / X
C      .....PROXIMA TRANSFORMACION QR.....
      C = 1.0
      S = 1.0
C
      DO 600 I1 = L, K1
          I = I1 + 1
          G = RV1(I)
          Y = W(I)

```

```

H = S * G
G = C * G
Z = DSQRT(F*F+H*H)
RV1(I1) = Z
C = F / Z
S = H / Z
F = X * C + G * S
G = -X * S + G * C
H = Y * S
Y = Y * C
IF (.NOT. MATV) GO TO 575
C
      DO 570 J = 1, N
         X = V(J,I1)
         Z = V(J,I)
         V(J,I1) = X * C + Z * S
         V(J,I) = -X * S + Z * C
570   CONTINUE
C
575   Z = DSQRT(F*F+H*H)
W(I1) = Z
C     .....LA ROTACION PUEDE SER ARBITRARIA SI Z ES CERO.....
IF (Z .EQ. 0.0) GO TO 580
C = F / Z
S = H / Z
580   F = C * G + S * Y
X = -S * G + C * Y
IF (.NOT. MATU) GO TO 600
C
      DO 590 J = 1, M
         Y = U(J,I1)
         Z = U(J,I)
         U(J,I1) = Y * C + Z * S
         U(J,I) = -Y * S + Z * C
590   CONTINUE
C
600   CONTINUE
C
      RV1(L) = 0.0
      RV1(K) = F
      W(K) = X
      GO TO 520
C     .....CONVERGENCIA.....
650   IF (Z .GE. 0.0) GO TO 700
C     .....W(K) SE HACE NO NEGATIVO.....
      W(K) = -Z
      IF (.NOT. MATV) GO TO 700
C
      DO 690 J = 1, N
690   V(J,K) = -V(J,K)
C
700 CONTINUE
C
      GO TO 1001
C     .....ACTIVAR ERROR -- NO HA HABIDO CONVERGENCIA A UN
C           VALOR SINGULAR DESPUES DE 30 ITERACIONES.....
1000 IERR = K
1001 RETURN
END

```

Los resultados obtenidos son:

COEFICIENTES: $A = 32.1468200684$, $B = 1.4269589186$ EVALUACION PEDIDA: 387.27 .

3.5.2 Enunciados

- El tiempo total necesario para parar un automóvil después de observar un peligro se compone del tiempo de reacción del conductor más el tiempo de frenada. En la tabla siguiente se puede encontrar las distancias de parada de un automóvil, desde que se observa el peligro, en función de la su velocidad.

Velocidad km/h	30	50	65	80	95	110
Distancia m	15	30	43	62	89	120

Hacer un programa para aproximar polinómicamente una nube de puntos cualesquiera por el método de mínimos cuadrados. Aplicarlo a los datos anteriores para calcular una recta y una parábola que aproxime la distancia de parada en función de la velocidad. ¿Cuál de las dos aproxima mejor? Calcular la distancia de parada para una velocidad de $120 \text{ km}/\text{h}$.

- Considerar los datos sobre velocidades y distancias de la práctica anterior. Calcular su aproximación min-max discreta por un polinomio de segundo grado para obtener la distancia de parada a $120 \text{ km}/\text{h}$. Hacer, para este cálculo, un programa que calcule la aproximación min-max polinómica discreta sobre un conjunto dado de puntos.
- La termodinámica dice que la presión (P) de un gas en función del su volumen (V) queda establecida por la ecuación

$$PV^\gamma = C,$$

donde γ y C son constantes. Considerar la tabla siguiente de valores de presión y volumen para un cierto gas

$V cm^3$	890	1013	1186	1454	1943	3179
$Patm$	4.146	3.353	2.547	1.924	1.360	0.684

Calcular los valores de γ y C approximando esta tabla por mínimos cuadrados.

- Aproximación de Padé (1892). Consideramos una función f tal que

$$f(x) = \sum_{k=0}^N c_k x^k + O(x^{N+1})$$

Queremos hallar una función racional $R_{mn} = \frac{p_m(x)}{q_n(x)}$ ($m + n = N$) tal que

$$f(x) - R_{mn}(x) = O(x^{N+1})$$

esta función racional se llama **aproximación de Padé** (m, n). Si $q_n(0) \neq 0$, podemos escribir

$$q_n(x)f(x) - p_m(x) = O(x^{N+1})$$

Definiendo $p_m(x) = \sum_{i=0}^m a_i x^i$ y $q_n(x) = \sum_{i=0}^n b_i x^i$, tenemos

$$\left(\sum_{k=1}^N c_k x^k \right) \left(\sum_{j=1}^n b_j x^j \right) = \sum_{i=1}^m a_i x^i + O(x^m + m + 1)$$

de esta igualdad podemos deducir el sistema de $m+n+1$ ecuaciones lineales y $m+n+2$ incógnitas siguiente:

$$\begin{aligned} \sum_{j=0}^n c_{m+i-j} b_j &= 0 \quad i = 1 \div n \\ \sum_{j=0}^{\min\{y,n\}} c_{i-j} b_j &= a_i \quad i = 0 \div m \end{aligned}$$

bajo el convenio que $c_k = 0$ si $k < 0$.

Si ahora fijamos $b_0 = 1$ nos queda el sistema de ecuaciones lineales

$$\begin{pmatrix} c_m & \cdots & c_{m-n+1} \\ \vdots & & \vdots \\ c_{m+n-1} & \cdots & c_m \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = - \begin{pmatrix} c_{m+1} \\ \vdots \\ c_{m+n} \end{pmatrix}$$

de donde calculamos b_1, \dots, b_n y las fórmulas

$$\sum_{j=0}^{\min\{y,j\}} c_{i-j} b_j = a_i \quad i = 0 \div m$$

para calcular a_0, \dots, a_m .

- (a) Hacer un programa que permita calcular la aproximación de Padé (m, n) de la función e^x en el intervalo $[-1, 1]$ para cualquier valor de m y n .
 - (b) Intentar comprobar numéricamente la conjetura de Meinardus:
- $$\lim_{m+n \rightarrow \infty} \frac{\|R_{mn}(x) - e^x\|_\infty}{\frac{2^{-m-n} m! n!}{(m+n)!(m+n+1)!}} = 1$$
5. Escribir un programa que calcule la aproximación de Padé (m, n) de la función $f(x) = \sqrt{x}$ en el intervalo $[a, b] \subset (0, \infty)$ para diferentes valores de m y n .
 6. Hacer un programa que calcule la aproximación min-max polinómica de una función con un error inferior a una cota dada previamente. Calcular la aproximación polinómica con un error inferior a 10^{-6} . Os sugerimos algunas funciones:
 - (a) e^x en el intervalo $[-1, 1]$
 - (b) \sqrt{x} en el intervalo $[0, b]$ cualquiera y, en particular, para $b = 1$
 - (c) $\sin x$, $\cos x$ y $\operatorname{tg} x$ en el intervalo $[0, \pi/4]$
 - (d) $\arccos x$ y $\arcsin x$ en el intervalo $[0, 1]$

4 Sistemas Lineales

4.1 Introducción

En este capítulo se muestran las técnicas de resolución numérica de sistemas lineales, $A x = b$, teniendo en cuenta que el sistema tiene una única solución (sistema compatible y determinado) cuando se considera una matriz cuadrada A que será no singular o regular ($\det A \neq 0$). Si desde un punto de vista formal la resolución se puede resolver por la regla de Cramer o calculando la inversa de la matriz A , en el caso de sistemas lineales de tamaño grande, estos métodos son muy costosos cuando no numéricamente inestables.

Se presentan, en primer lugar, los métodos directos que dan la solución numérica del sistema en un número finito de pasos (si no se consideran los errores de redondeo); después se introducen las técnicas iterativas, donde la sucesión de vectores que se obtienen tiende a la solución y se puede tener cierta idea de dónde está la solución cuando se trunca; finalmente, se presenta el método para resolver sistemas lineales sobredeterminados que están relacionados con el tema de las aproximaciones por mínimos cuadrados.

Se considera el sistema $Ax = b$ que, en forma clásica, se escribe de la siguiente manera

$$\left. \begin{array}{l} a_{11} x_1 + a_{12} x_2 + \cdots + a_{1n} x_n = b_1 \\ a_{21} x_1 + a_{22} x_2 + \cdots + a_{2n} x_n = b_2 \\ \dots \\ a_{n1} x_1 + a_{n2} x_2 + \cdots + a_{nn} x_n = b_n \end{array} \right\}$$

A menudo interesa hacer una serie de manipulaciones para conseguir que la matriz del sistema U sea triangular superior; entonces el sistema es más fácil de resolver

$$\left. \begin{array}{l} u_{11} x_1 + u_{12} x_2 + \cdots + u_{1n} x_n = \bar{b}_1 \\ u_{22} x_2 + \cdots + u_{2n} x_n = \bar{b}_2 \\ \dots \\ u_{nn} x_n = \bar{b}_n \end{array} \right\}$$

El algoritmo de resolución, que recibe el número de **substitución hacia atrás**, consiste en:

$$x_n = \frac{\bar{b}_n}{u_{nn}} \quad x_i = \frac{1}{u_{ii}} \left(\bar{b}_i - \sum_{j=i+1}^n u_{ij} x_j \right) \quad i = n-1 \div 1$$

Nótese que $u_{ii} \neq 0$, $i = 1 \div n$, ya que se ha supuesto que $\det A = \det U \neq 0$.

Para sistemas triangulares inferiores puede aplicarse la **substitución hacia delante**:

$$\begin{aligned} l_{11} x_1 &= \tilde{b}_1 \\ l_{21} x_1 + l_{22} x_2 &= \tilde{b}_2 \\ &\vdots \\ l_{n1} x_1 + l_{n2} x_2 + \dots + l_{nn} x_n &= \tilde{b}_n \end{aligned} \quad \left. \right\}$$

De este modo, para la resolución de un sistema triangular, utilizando los algoritmos de sustitución hacia delante o hacia atrás, se necesitan n divisiones y

$$\sum_{k=1}^n (k-1) = \frac{1}{2} n(n-1) \approx \frac{1}{2} n^2$$

sumas y multiplicaciones.

4.2 Métodos directos

Los métodos directos permiten obtener la solución del sistema después de un número finito de pasos. Se aplica este método, en particular, a matrices de configuraciones especiales, como son las matrices diagonales, triangulares, y las unitarias, ya que en estos casos es muy fácil. Se inicia el tema con la introducción a la eliminación gaussiana y la factorización LU para aprovecharlo, posteriormente, en el cálculo de inversas; finalmente, se presenta un estudio del error con la introducción del número de condición, el mal condicionamiento de las matrices y la acumulación de los errores de redondeo.

4.2.1 Eliminación gaussiana

Es uno de los métodos más empleados y consiste en reducir el sistema inicial a un sistema triangular. Sea $a_{11} \neq 0$; entonces se puede eliminar x_1 de las $n-1$ filas de la matriz $A = A^{(1)} = (a_{ij}^{(1)})$, restando de la i -ésima fila el factor $m_{i1} = \frac{a_{i1}}{a_{11}}$ $i = 2 \div n$ de la primera fila.

Las $n-1$ ecuaciones quedan del siguiente modo

$$\left. \begin{aligned} a_{22}^{(2)} x_2 + \dots + a_{2n}^{(2)} x_n &= b_2^{(2)} \\ \dots &\dots \dots \\ a_{n2}^{(2)} x_2 + \dots + a_{nn}^{(2)} x_n &= b_n^{(2)} \end{aligned} \right\} \quad \text{donde} \quad \left\{ \begin{array}{l} a_{ij}^{(2)} = a_{ij} - m_{i1} a_{1j} \quad i = 2 \div n \\ b_i^{(2)} = b_i - m_{i1} b_1 \end{array} \right.$$

Si $a_{22} \neq 0$, se considera $m_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}$, $i = 3 \div n$, y se repetiría el procedimiento. Se tiene, entonces, matrices $A^{(k)} = (a_{ij}^{(k)})$, $k = 2 \div n$, con $a_{ij}^{(k)} = 0$, $i > j$, $j = 1 \div k - 1$:

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & a_{1k}^{(k)} & \cdots & a_{1n}^{(k)} \\ a_{22}^{(k)} & \cdots & a_{2k}^{(k)} & \cdots & a_{2n}^{(k)} \\ \ddots & \vdots & \cdots & \vdots & & \\ & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} & & \\ 0 & \vdots & \cdots & \vdots & & \\ & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} & & \end{pmatrix} \quad k = 2 \div n$$

de forma que $A^{(n)}$ es una matriz triangular superior. Los mismos cálculos se realizan con el vector b ; así se obtienen los vectores $b^{(k)}$ de componentes $b_i^{(k)}$, $i = 1 \div n$, $k = 2 \div n$.

El paso k -ésimo de este método se realiza calculando los valores de los elementos de las filas que van de la k -ésima hasta la n -ésima y los de los elementos correspondientes del término independiente:

$$\left. \begin{array}{lcl} m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} & b_i^{(k+1)} &= b_i^{(k)} - m_{ik} b_k^{(k)} \\ & a_{ij}^{(k+1)} &= a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)} \end{array} \right\}$$

$$j = k + 1 \div n \quad i = k + 1 \div n \quad y \quad k = 1 \div n - 1$$

En este paso k -ésimo se necesitan $n - k$ divisiones y $(n - k)(n - k + 1)$ multiplicaciones y diferencias. Para n suficientemente grande se puede no contar las divisiones y, sabiendo que $1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$, se tiene

$$\sum_{k=1}^{n-1} (n - k)(n - k + 1) = \frac{(n - 1)n(2n - 1)}{6} + \frac{1}{2}n(n - 1) \approx \frac{1}{3}n^3$$

Una vez realizada esta triangularización, sólo queda la resolución del sistema triangular superior $A^{(n)} x = b^{(n)}$ (procedimiento de sustitución hacia atrás).

Estrategias

- Se ha supuesto que $a_{kk}^{(k)} \neq 0$, pero, si se anula, siempre se puede encontrar un elemento de la misma columna no nulo (si no fuera así, la matriz sería singular). La solución consiste a intercambiar las dos ecuaciones. El método de eliminación gaussiana es aplicable directamente sin permutaciones de filas, siempre y cuando el producto $a_{11}^{(1)} \cdot a_{22}^{(2)} \cdots a_{nn}^{(n)}$ sea no nulo ($\det A \neq 0$).
- Para asegurar la estabilidad numérica, se pueden hacer intercambios de filas a pesar de que el elemento pivote ($a_{kk}^{(k)}$) no sea nulo, pero sí suficientemente pequeño para propagar errores que pueden cambiar el sentido de los resultados.

Ejercicios.

1. Dado el sistema $\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_1 + x_2 + 2x_3 = 2 \\ x_1 + 2x_2 + 2x_3 = 1 \end{cases}$, comprobar que $a_{22}^{(2)} = 0$ y resolvedlo permutando las filas 2 y 3.

2. Dado el sistema anterior donde se ha perturbado el coeficiente a_{22} en una diez milésima $\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_1 + 1.0001x_2 + 2x_3 = 2 \\ x_1 + 2x_2 + 2x_3 = 1 \end{cases}$, resolved el sistema con aritmética de 3 y 4 dígitos.

Comparar los resultados y, sin mejorar la precisión de l'aritmética (3 dígitos), encontrad la solución.

Por lo tanto, será necesario, a veces, hacer algún tipo de intercambio.

• **Pivotamiento parcial.** En el paso k -ésimo, en la columna k -ésima se elige como pivote el elemento mayor en valor absoluto: Se toma r como el entero más pequeño tal que $|a_{rk}^{(k)}| = \max |a_{ik}^{(k)}|$ $i = k \div n$, y se intercambian las filas k y r . La rutina de las prácticas es **DECOMP**, que se sirve de este tipo de eliminación gaussiana.

• **Pivotamiento total.** En este caso, de la submatriz que comienza en la fila y columna k -ésima hasta la n -ésima se elige como pivote el elemento mayor en valor absoluto: Se toman r y s como los enteros más pequeños para los que $|a_{rs}^{(k)}| = \max |a_{ij}^{(k)}|$ $i, j = k \div n$ y se permutan las filas k y r y las columnas k y s , teniendo en cuenta que cambia el orden de las incógnitas. No tiene por qué ser el método más estable de eliminación gaussiana, ya que el producto de los elementos pivotes es igual al determinante de la matriz y, si se seleccionan en los primeros pasos los mayores, inevitablemente se tendrán al final los menores con el peligro de inestabilidad numérica que este hecho comporta (ver [Wil65]).

4.2.2 Factorización LU

Si la matriz A factoriza en producto de dos matrices triangulares: $A = LU$, donde L es una matriz triangular inferior y U una triangular superior. Entonces, el sistema $Ax = b$ es equivalente a $LUx = b$, que se descompone en dos sistemas lineales triangulares

$$Ly = b \quad Ux = y$$

Teorema 4.1. Sea A una matriz $n \times n$ y si se denota por A_k las submatrices cuadradas $k \times k$ formadas por las primeras k filas y columnas de A . Si $\det A_k \neq 0$, $k = 1 \div n$, entonces existen matrices triangulares únicas $L = (l_{ij})$ y $U = (u_{ij})$, inferior y superior respectivamente, con $l_{ii} = 1$, $i = 1 \div n$, tales que $A = LU$.

Demostración: Se demuestra por inducción; si $n = 1$, la factorización es única: $a_{11} = 1 \cdot u_{11}$.

Supóngase el teorema cierto para $n = k - 1$; entonces, para $n = k$, si se presentan A_k , L_k y U_k en bloques, se tiene

$$A_k = \begin{pmatrix} A_{k-1} & b \\ c^T & a_{kk} \end{pmatrix} \quad L_k = \begin{pmatrix} L_{k-1} & 0 \\ l^T & 1 \end{pmatrix} \quad U_k = \begin{pmatrix} U_{k-1} & u \\ 0 & u_{kk} \end{pmatrix}$$

donde b , c , l y u son vectores columna con $k - 1$ componentes. Si se efectúa el producto $L_k U_k$ y se identifica con A_k , se obtiene

$$\begin{array}{ll} 1) & L_{k-1} U_{k-1} = A_{k-1} \\ 3) & l^T U_{k-1} = c^T \end{array} \quad \begin{array}{ll} 2) & L_{k-1} u = b \\ 4) & l^T u + u_{kk} = a_{kk} \end{array}$$

Como, por hipótesis de inducción, L_{k-1} y U_{k-1} están determinadas únicamente y $\det L_{k-1} \cdot \det U_{k-1} = \det A_{k-1} \neq 0$, entonces L_{k-1} y U_{k-1} son no singulares, y, por lo tanto, l y u están únicamente determinados por los sistemas 2) y 3). Finalmente, $u_{kk} = a_{kk} - l^T u$, con lo que queda probado el teorema. \square

Ejercicio. Dado el sistema $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ t \end{pmatrix} = \begin{pmatrix} 2 \\ 10 \\ 44 \\ 190 \end{pmatrix}$, comprobar que, si se hace la eliminación gaussiana sin pivotamiento, se tienen las matrices

$$U = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 6 & 12 \\ 0 & 0 & 6 & 24 \\ 0 & 0 & 0 & 24 \end{pmatrix} \quad L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 1 & 7 & 6 & 1 \end{pmatrix}$$

donde $U = A^{(4)}$ y en L están los elementos $l_{ik} = m_{ik}$, $i = k + 1 \div 4$, con $k = 1 \div 4$.

Se puede demostrar ([Dah74]) que, cuando todos los menores principales son diferentes de cero, la factorización $L U$ y la eliminación gaussiana son equivalentes, y ésta da lugar a las matrices L y U , donde $l_{ik} = m_{ik}$, $i = k + 1 \div n$, $u_{kj} = a_{kj}^{(k)}$, $j = k \div n$ y $k = 1 \div n$. Si algún menor principal es cero, puede no existir la factorización $L U$:

Ejercicio. Sea la matriz $A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$; comprobar la imposibilidad de factorizar la matriz si no se permutan las filas o las columnas.

En realidad, para cualquier matriz no singular A , las filas pueden ser reordenadas de tal manera que exista una factorización $L U$. Este resultado es una consecuencia de la equivalencia entre la eliminación gaussiana y la factorización $L U$. Para permutar dos filas, r y k por ejemplo, se introduce la matriz de permutaciones P , que es igual a la matriz identidad salvo en los elementos

$$p_{rr} = p_{kk} = 0 \quad p_{rk} = p_{kr} = 1$$

Entonces, el producto $P \cdot A$ intercambia las filas k -ésima y r -ésima de A . Si se multiplica $A \cdot P$, el resultado será el mismo que se tenía con A , pero con las columnas k -ésima y r -ésima de A permutadas. Entonces, se tiene

Teorema 4.2. Sea A una matriz $n \times n$ no singular; entonces existen matrices triangulares inferior $L = (l_{ij})$ con $l_{ii} = 1$, triangular superior $U = (u_{ij})$ con $u_{ii} \neq 0$, $i = 1 \div n$ y matriz de permutaciones P tales que $P A = L U$.

En este caso, si se tiene el sistema $A x = b$, se considera

$$P A x = P b \implies L U x = P b$$

y de aquí se resuelven los sistemas triangulares siguientes $L y = P b$ y, posteriormente, $U x = y$.

4.2.3 Métodos compactos

Cuando se resuelve un sistema por eliminación gaussiana se tiene, aproximadamente, $\frac{n^3}{3}$ productos y sumas. Puede ser más interesante considerar la factorización $L U$ directamente, ya que en el paso k -ésimo de Gauss son determinadas la columna k -ésima de L y la fila k -ésima de U , mientras que los métodos compactos se construyen de tal manera que los elementos a_{ij} , $i, j > k$, no son manipulados.

La ecuación matricial $A = L U$ es equivalente a las ecuaciones

$$a_{ij} = \sum_{k=1}^p l_{ik} u_{kj} \quad p = \min\{i, j\}$$

que se pueden pensar como un sistema de n^2 ecuaciones con $n(n + 1)$ incógnitas de L y U , donde se puede elegir igual a 1 los elementos de la diagonal de una de las matrices triangulares y dar lugar al método de Doolittle, si $l_{ii} = 1$, $i = 1 \div n$, y al método de Crout, si $u_{ii} = 1$, $i = 1 \div n$.

Método de Doolittle

Si se impone que los elementos diagonales de L sean la unidad: $l_{kk} = 1$, $k = 1 \div n$, y teniendo en cuenta que los nuevos elementos diagonales de U son diferentes de cero, $u_{kk} \neq 0$, $k = 1 \div n$, ya que la matriz A es no singular, se tiene el algoritmo de cálculo siguiente:

$$\left\{ \begin{array}{lcl} u_{1j} & = & a_{1j} \\ u_{kj} & = & a_{kj} - \sum_{r=1}^{k-1} l_{kr} u_{rj} & j = k \div n \\ l_{i1} & = & \frac{a_{i1}}{u_{11}} & k = 1 \div n \\ l_{ik} & = & \frac{1}{u_{kk}} \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir} u_{rk} \right) & i = k + 1 \div n \end{array} \right.$$

El orden en los cálculos de las l' s y u' s es el siguiente: u_{1j} , $j = 1 \div n$, l_{i1} , $i = 2 \div n$, u_{2j} , $j = 2 \div n$, l_{i2} , $i = 3 \div n, \dots, u_{n-1,n-1}$, $u_{n-1,n}$, $l_{n,n-1}$ y u_{nn} .

Ejercicio. Dada la matriz A , comprobar que el método de Doolittle da lugar a la factorización $L U$ siguiente:

$$\begin{pmatrix} 1 & -1 & 2 \\ -1 & 5 & 4 \\ 2 & 4 & 14 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & 3/2 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 2 \\ 0 & 4 & 6 \\ 0 & 0 & 1 \end{pmatrix}$$

Método de Crout

Si se exige que los elementos diagonales de U sean la unidad, $u_{kk} = 1$, $k = 1 \div n$, y teniendo en cuenta que los nuevos elementos diagonales de L son diferentes de cero, $l_{kk} \neq 0$, $k = 1 \div n$, ya que la matriz A es no singular, se tiene el algoritmo de cálculo siguiente:

$$\left\{ \begin{array}{l} l_{i1} = a_{i1} \\ l_{ik} = a_{ik} - \sum_{r=1}^{k-1} l_{ir} u_{rk} \quad i = k \div n \\ u_{1j} = \frac{a_{1j}}{l_{11}} \\ u_{kj} = \frac{1}{l_{kk}} \left(a_{kj} - \sum_{r=1}^{k-1} l_{kr} u_{rj} \right) \quad j = k+1 \div n \end{array} \right. \quad k = 1 \div n$$

El orden en los cálculos de las l' s y u' s es el siguiente: l_{i1} , $i = 1 \div n$, u_{1j} , $j = 2 \div n$, l_{i2} , $i = 2 \div n$, u_{2j} , $j = 3 \div n, \dots, l_{n-1,n-1}$, $l_{n,n-1}$, $u_{n-1,n}$, y, finalmente l_{nn} .

Ejercicio. Dada la matriz A , comprobar que el método de Crout da lugar a la factorización $L U$ siguiente:

$$\begin{pmatrix} 1 & -1 & 2 \\ -1 & 5 & 4 \\ 2 & 4 & 14 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 4 & 0 \\ 2 & 6 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 2 \\ 0 & 1 & 3/2 \\ 0 & 0 & 1 \end{pmatrix}$$

Método de Choleski

Este método se emplea cuando la matriz A es simétrica y definida positiva; una propiedad importante de este tipo de matrices es que la propiedad es conservada para la matriz inversa, A^{-1} , y para cualquier submatriz principal; además, todos los determinantes de las matrices consideradas en esta sección son positivos.

Si se aplica la factorización $L U$ (se puede pensar en el método de Doolittle), se tiene $u_{11} = a_{11} > 0$ y $u_{kk} = \frac{\det A_k}{\det A_{k-1}} > 0$, $k = 2 \div n$. Se construye, por tanto, la matriz diagonal $D =$

$\text{diag}(u_{11}, \dots, u_{nn})$ que es no singular. Entonces, la factorización se expresa por la igualdad

$$A = L U = L D D^{-1} U = L' D^{-1} U$$

y, como que A es simétrica,

$$A = A^T = U^T D^{-1} (L')^T \implies U^T = L' = L D$$

por la unicidad de la factorización ; además $U = D L^T$ y, en definitiva, se tiene $A = L D L^T$ que permite escribir el algoritmo siguiente:

$$\begin{aligned} d_{kk} &= a_{kk} - \sum_{r=1}^{k-1} l_{kr}^2 d_{rr} & k = 1 \div n \\ l_{ik} &= \frac{1}{d_{kk}} \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir} d_{rr} l_{kr} \right) & i = k+1 \div n \end{aligned}$$

El número de operaciones se reduce casi a la mitad. Si \mathcal{L} es la matriz $L D^{1/2}$, con $D^{1/2} = \text{diag}(d_{11}^{1/2}, \dots, d_{nn}^{1/2})$, entonces se puede expressar A en la forma:

$$A = \mathcal{L} \mathcal{L}^T$$

que es la llamada factorización de Choleski; para no tener que hacer el cálculo de las raíces cuadradas, se puede pensar el sistema $A x = b$, teniendo en cuenta que $A = L D L^T$, como dos sistemas lineales triangulares

$$L y = b \quad L^T x = D^{-1} y$$

4.2.4 Cálculo de inversas

Dada una matriz A , la cantidad de cálculos para obtener directamente su inversa obliga a presentar soluciones alternativas. Una de ellas es aprovechar la factorización $L U$ tal que $P A = L U$; entonces, se resuelven los N sistemas

$$L U a_k = P e_k \quad k = 1 \div n$$

donde a_k es la k -ésima columna de A^{-1} y e_k es el k -ésimo vector de la base canónica. Para resolver los n sistemas se necesitan aproximadamente $2/3 n^3$ operaciones y, si se suman las $n^3/3$ de la factorización $L U$, da del orden de n^3 operaciones necesarias para poder calcular A^{-1} .

Se puede utilizar un método de Gauss modificado de manera que también se consigan ceros en la parte triangular superior, y resolver simultáneamente los n sistemas, de manera que se pasaría de la ecuación matricial $A X = I$, donde I es la matriz identidad y $X = A^{-1}$, a la ecuación $D X = B$, y, por lo tanto, $A^{-1} = D^{-1} B$ (**Método de Gauss-Jordan**).

Un algoritmo alternativo y complementario al método de Gauss-Jordan consiste en invertir la aplicación lineal $A : x \mapsto y = A x$ de la manera siguiente: se considera el sistema lineal donde los coeficientes son precisamente los términos de la matriz A

$$\left. \begin{array}{l} a_{11} x_1 + a_{12} x_2 + \cdots + a_{1n} x_n = y_1 \\ a_{21} x_1 + a_{22} x_2 + \cdots + a_{2n} x_n = y_2 \\ \vdots \\ a_{n1} x_1 + a_{n2} x_2 + \cdots + a_{nn} x_n = y_n \end{array} \right\}$$

Primer paso: Se busca $a_{r1} \neq 0$ tal que $|a_{r1}| = \max_i |a_{i1}|$ (pivotamiento parcial).

Seguidamente, despejando x_1 y substituyéndolo en las otras ecuaciones, se obtiene un nuevo sistema

$$\left. \begin{array}{l} a'_{11} y_1 + a'_{12} x_2 + \cdots + a'_{1n} x_n = x_1 \\ a'_{21} y_1 + a'_{22} x_2 + \cdots + a'_{2n} x_n = y_2 \\ \vdots \\ a'_{n1} y_1 + a'_{n2} x_2 + \cdots + a'_{nn} x_n = y_n \end{array} \right\}$$

donde $a'_{11} = \frac{1}{a_{11}}$, y $a'_{1k} = -\frac{a_{1k}}{a_{11}}$, $k = 2 \div n$, para la primera fila y

$$a'_{i1} = \frac{a_{i1}}{a_{11}} \quad a'_{ik} = a_{ik} - \frac{a_{i1} a_{ik}}{a_{11}} \quad i, k = 2 \div n$$

Segundo paso: Se busca $a'_{r2} \neq 0$ tal que $|a'_{r2}| = \max_{i \geq 2} |a'_{i2}|$, y se repite el procedimiento intercambiando x_2 i y_2 , y así sucesivamente.

Se van obteniendo nuevas matrices: $A = A^{(0)}$, $A^{(1)}, \dots, A^{(n)} = A^{-1}$, ya que

$$(A^{(n)} P) y = x \implies A^{(n)} P = A^{-1}$$

con P matriz de permutaciones debida a los pivotamientos parciales.

Ejemplo. Si se aplica este método a la matriz $A = A^{(0)} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 3 & 6 \end{pmatrix}$, sin pivotamiento, se obtiene la sucesión

$$A^{(1)} = \begin{pmatrix} 1 & -1 & -1 \\ 1 & 1 & 2 \\ 1 & 2 & 5 \end{pmatrix} \rightarrow A^{(2)} = \begin{pmatrix} 2 & -1 & 1 \\ -1 & 1 & -2 \\ -1 & 2 & 1 \end{pmatrix} \rightarrow A^{-1} = \begin{pmatrix} 3 & -3 & 1 \\ -3 & 5 & -2 \\ 1 & -2 & 1 \end{pmatrix}$$

Ejercicio. Calcular el número de operaciones y deducid que este último algoritmo explicado en esta sección es de orden n^3 .

4.2.5 Cotas de error

Una de las fuentes de errores más frecuente en los métodos numéricos al resolver sistemas lineales proviene de la propagación de los errores en los datos donde juega un papel importante la sensibilidad de la matriz del sistema (número de condición) y que será el primero que se trata

para, posteriormente, hacer un análisis de los errores al ir perturbando los diferentes elementos del sistema. Otra fuente de error depende de la acumulación de errores de redondeo en el método empleado y se analizará en el método de eliminación gaussiana.

Se utilizarán siempre normas matriciales consistentes con las normas vectoriales prescindiendo, en principio, del subíndice de la correspondiente norma vectorial de la que depende (ver apéndice A).

Número de condición

Sea el sistema $A x = b$ y se supone que actúa sobre b cierta perturbación Δb tal que se tiene

$$A(x + \Delta x) = b + \Delta b \implies \Delta x = A^{-1}\Delta b$$

y, tomando norma consistente con alguna norma vectorial, se obtiene $\|\Delta x\| \leq \|A^{-1}\| \|\Delta b\|$. Como, por otro lado, de $\|b\| \leq \|A\| \|x\|$ se deduce $\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$, se puede acotar el error relativo por

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A^{-1}\| \frac{\|\Delta b\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\Delta b\|}{\|b\|}$$

y, si se define el número de condición de la matriz A , $\kappa(A)$, por

$$\kappa(A) = \begin{cases} \|A\| \|A^{-1}\| & \text{si } A \text{ no singular} \\ +\infty & \text{si } A \text{ singular} \end{cases}$$

se tiene

$$\frac{\|\Delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\Delta b\|}{\|b\|}$$

Ejercicio. Probar que $\kappa(A) \geq 1$, $\forall A$ matriz cuadrada.

Si $\kappa(A)$ es grande, pequeñas perturbaciones en b producen grandes perturbaciones en x y se puede asegurar que se tiene un problema muy sensible a las condiciones iniciales. Cuando el número de condición de la matriz de un sistema es grande, se dice que se tiene un **sistema mal condicionado**.

Si se introduce el **vector residual**, $r(\bar{x})$, como un criterio de comparación entre la solución exacta, x , y la solución calculada numéricamente, \bar{x} , es decir

$$r(\bar{x}) = A\bar{x} - b$$

y se considera $\Delta x = x - \bar{x}$, se tiene

$$\frac{\|x - \bar{x}\|}{\|x\|} = \frac{\|\Delta x\|}{\|x\|} \leq \kappa(A) \frac{\|r(\bar{x})\|}{\|b\|}$$

que asegura que no es suficiente la evaluación del vector residual, ya que un factor del error es el número de condición.

Ejemplo. Si se considera el sistema lineal con $A = \begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix}$ y $b = \begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix}$, una solución aproximada es $\bar{x} = (0.9911, -0.4870)^T$ con vector residual $r = (-10^{-8}, 10^{-8})^T$. Parece que el error de \bar{x} es muy pequeño, pero la solución exacta es $x = (2, -2)^T$ (comprobarlo substituyendo).

Es fácil ver dónde está mal condicionado el sistema; después de eliminar x_1 se tiene

$$a_{22}^{(2)} = 0.1441 - \frac{0.2161}{1.2969} \approx 10^{-8}$$

Un mínimo cambio en el elemento 0.1441 producirá un cambio considerable en $a_{22}^{(2)}$ y, por lo tanto, en x_2 . Como A y b vienen dados con una precisión mayor que 10^{-8} , hablar de una solución no es correcto. Como se ha visto, no es suficiente considerar el vector residual; la matriz inversa de A es $A^{-1} = 10^8 \begin{pmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{pmatrix}$ y $\kappa(A) \approx 3.3 \cdot 10^8$ con la norma $\|\cdot\|_1$.

Errores en los datos

Si se perturba la matriz A , se obtiene

$$(A + \Delta A)(x + \Delta x) = b \implies \Delta x = -A^{-1} \Delta A (x + \Delta x)$$

de donde, tomando norma consistente con alguna norma vectorial, se tiene

$\|\Delta x\| \leq \|A^{-1}\| \|\Delta A\| \|x + \Delta x\|$, e introduciendo el número de condición, se puede escribir en la siguiente forma

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \kappa(A) \frac{\|\Delta A\|}{\|A\|}$$

Ejercicio. Considerar, como en [Aub91], la perturbación en A y b a la vez: $(A + \Delta A)(x + \Delta x) = b + \Delta b$. Demostrar que si $\|A^{-1}\| \|\Delta A\| \leq 1$, entonces

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\Delta A\|}{\|A\|}} \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right)$$

Aplicarlo al caso $\Delta A = \begin{pmatrix} \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon \end{pmatrix}$ y $\Delta b = \begin{pmatrix} \varepsilon \\ \varepsilon \\ \varepsilon \end{pmatrix}$.

Errores de redondeo en la eliminación gaussiana

Ejemplo. Si se considera el sistema lineal con $A = \begin{pmatrix} 0.005 & 1 \\ 1 & 1 \end{pmatrix}$ y $b = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}$, la solución exacta es $x_1 = \frac{5000}{9950} \approx 0.503\dots$, $x_2 = \frac{4950}{9950} \approx 0.497\dots$. Si se utiliza una representación en coma flotante de 2 dígitos y se toma $a_{11} = 0.005$ como primer pivote, se obtiene

$$\left(\begin{array}{cc|c} 0.005 & 1 & 0.5 \\ 0 & -200 & -99 \end{array} \right)$$

de donde $x_2 = \frac{99}{200} = 0.5$ y $x_1 = 200 (0.5 - 0.5) = 0.0$.

Si se intercambian las filas y se toma el elemento a_{21} anterior como pivote se tiene

$$\left(\begin{array}{cc|c} 1 & 1 & 1 \\ 0 & 1 & 0.5 \end{array} \right)$$

de donde $x_2 = \frac{0.5}{1} = 0.5$ y $x_1 = 1 - 0.5 = 0.5$. Por lo tanto, la elección de unos pivotes o de otros añadida a los errores de redondeo pueden llevar a situaciones como la presente donde se tienen tres "soluciones".

Si se consideran todas las sucesivas manipulaciones que se hacen sobre un elemento de la matriz A cuando se aplica el método de eliminación gaussiana, un elemento por encima de o en la diagonal, $a_{ij}^{(k)}$, $i \leq j$, es cambiado en pasos sucesivos hasta $k = i$ y después se mantiene invariante; es decir, se tiene

$$\left. \begin{array}{lcl} a_{ij}^{(2)} & = & a_{ij}^{(1)} + m_{i1} a_{1j}^{(1)} + \varepsilon_{ij}^{(2)} \\ a_{ij}^{(3)} & = & a_{ij}^{(2)} + m_{i2} a_{2j}^{(2)} + \varepsilon_{ij}^{(3)} \\ \vdots & & \vdots \\ a_{ij}^{(i)} & = & a_{ij}^{(i-1)} + m_{i,i-1} a_{i-1,j}^{(i-1)} + \varepsilon_{ij}^{(i)} \end{array} \right\} \quad (4.1)$$

y, sumando,

$$\begin{aligned} a_{ij}^{(i)} &= a_{ij}^{(1)} + m_{i1} a_{1j}^{(1)} + \cdots + m_{i,i-1} a_{i-1,j}^{(i-1)} + \\ &\quad + \varepsilon_{ij}^{(2)} + \cdots + \varepsilon_{ij}^{(i)} \end{aligned}$$

Un elemento por debajo de la diagonal, $a_{ij}^{(k)}$, $i > j$, se cambia en pasos sucesivos hasta $k = j$, donde se emplea para calcular el pivote que anula su posición y que después se mantiene nulo; es decir, se tiene

$$\left. \begin{array}{lcl} a_{ij}^{(2)} & = & a_{ij}^{(1)} + m_{i1} a_{1j}^{(1)} + \varepsilon_{ij}^{(2)} \\ a_{ij}^{(3)} & = & a_{ij}^{(2)} + m_{i2} a_{2j}^{(2)} + \varepsilon_{ij}^{(3)} \\ \vdots & & \vdots \\ a_{ij}^{(j)} & = & a_{ij}^{(j-1)} + m_{i,j-1} a_{j-1,j}^{(j-1)} + \varepsilon_{ij}^{(j)} \\ 0 & = & a_{ij}^{(j)} + m_{i,j} a_{j,j}^{(j)} + \varepsilon_{ij}^{(j+1)} \end{array} \right\} \quad (4.2)$$

donde m_{ij} viene dado por $m_{ij} = -\frac{a_{ij}^{(j)}}{a_{jj}^{(j)}} + \eta_{ij}$ y, por lo tanto, $\varepsilon_{ij}^{(j+1)} = -a_{jj}^{(j)} \eta_{ij}$. Sumando todas las igualdades de 4.2, se obtiene

$$\begin{aligned} 0 &= a_{ij}^{(1)} + m_{i1} a_{1j}^{(1)} + \cdots + m_{i,j} a_{j,j}^{(j)} + \\ &\quad + \varepsilon_{ij}^{(2)} + \cdots + \varepsilon_{ij}^{(j)} + \varepsilon_{ij}^{(j+1)} \end{aligned}$$

Entonces, los elementos $a_{ij}^{(i)}$ con $i \leq j$ son los elementos de la matriz triangular superior $A^{(n)}$. Los ceros de la última ecuación corresponden a los elementos por debajo de la diagonal principal de la matriz. Por tanto, considerando los errores que se producen en los cálculos y expresándolo todo en términos de elementos no intermedios del método, el error que se comete sobre un elemento $a_{ij}^{(k)}$ a partir de la sucesión de términos $a_{ij}^{(1)}, a_{ij}^{(2)}, \dots$ etc., si la representación es exacta, es

$$\begin{aligned} a_{ij}^{(1)} + \Delta a_{ij}^{(1)} &= a_{ij}^{(1)} + \cdots + \varepsilon_{ij}^{(2)} + \cdots + \varepsilon_{ij}^{(i)} & i \leq j \\ a_{ij}^{(1)} + \Delta a_{ij}^{(1)} &= a_{ij}^{(1)} + \cdots + \varepsilon_{ij}^{(2)} + \cdots + \varepsilon_{ij}^{(j+1)} & i > j \end{aligned}$$

teniendo $i - 1$ errores del tipo ε_{ij} para $i \leq j$, y j errores del tipo ε_{ij} para $i > j$. La columna del término independiente se puede pensar como la última columna de la matriz.

Falta estudiar la acumulación de los errores de redondeo inducidos por la aritmética de coma flotante: se tiene

$$\begin{aligned} a_{ij}^{(k)} &= fl(a_{ij}^{(k-1)} + m_{i,k-1} a_{k-1,j}^{(k-1)}) \\ &= [a_{ij}^{(k-1)} + m_{i,k-1} a_{i,k-1}^{(k-1)} (1 + \delta_1)] (1 + \delta_2) \end{aligned}$$

con $|\delta_i| \leq 2^{-t}$ y tal que

$$\begin{aligned} \varepsilon_{ij}^{(k)} &= a_{ij}^{(k)} - (a_{ij}^{(k-1)} + m_{i,k-1} a_{k-1,j}^{(k-1)}) \\ &= a_{ij}^{(k)} - \frac{a_{ij}^{(k)}}{1 + \delta_2} + (1 + \delta_1) m_{i,k-1} a_{k-1,j}^{(k-1)} - m_{i,k-1} a_{k-1,j}^{(k-1)} \\ &= a_{ij}^{(k)} - \frac{a_{ij}^{(k)}}{1 + \delta_2} + \delta_1 m_{i,k-1} a_{k-1,j}^{(k-1)} \end{aligned}$$

y, si no se consideran los términos cuadráticos (de orden 2^{-2t}), se obtiene la cota

$$|\varepsilon_{ij}^{(k)}| \leq 2^{-t} (|a_{ij}^{(k)}| + |m_{i,k-1} \cdot a_{k-1,j}^{(k-1)}|) \quad (4.3)$$

La desigualdad en 4.3 se aplica a todas las igualdades de 4.1 y 4.2 excepto la última, donde se tiene

$$m_{ij} = fl\left(-\frac{a_{ij}^{(j)}}{a_{jj}^{(j)}}\right) = -\frac{a_{ij}^{(j)}}{a_{jj}^{(j)}} (1 + \delta)$$

tal que $\varepsilon_{ij}^{(j+1)} = -(a_{ij}^{(j)} + m_{ij} a_{jj}^{(j)}) = \delta a_{ij}^{(j)}$ y

$$|\varepsilon_{ij}^{(j+1)}| \leq 2^{-t} |a_{ij}^{(j)}| \quad (4.4)$$

Los errores, pues, pueden ser grandes en la medida que los pivotes de los sucesivos elementos calculados lo sean. La técnica del pivotamiento hace que $|m_{ij}| \leq 1$, y así se restringe el crecimiento de $|a_{ij}^{(k)}|$ ($|a_{ij}^{(k)}| < 2^{k-1} |a_{ij}^{(1)}|$).

Como $\Delta a_{ij}^{(1)}$ es la suma de $r - 1$ (s) errores de redondeo para $r \leq s$ ($r > s$), teniendo en cuenta las cotas 4.3 y 4.4, en cualquier caso, se tiene

$$|\varepsilon_{ij}^{(k)}| \leq 2 \cdot 2^{-t} |g| \quad \text{donde} \quad |g| = \max_{i,j} (|a_{ij}^{(k)}|)_{k=2 \div n}$$

Entonces, $\Delta a_{ij}^{(1)} \leq \begin{cases} 2(i-1)2^{-t}|g|, & \text{si } i \leq j, \\ 2j2^{-t}|g|, & \text{si } i > j, \end{cases}$ y, con el mismo argumento, $|\Delta b_i^{(1)}| \leq 2(i-1)2^{-t}|h|$, donde $|h|$ es el elemento máximo en valor absoluto de todas las componentes que han aparecido durante el proceso.

4.3 Métodos iterativos

Se presenta, en primer lugar, un ejemplo introductorio de los métodos iterativos de Jacobi y Gauss-Seidel. Si se considera el sistema $Ax = b$ dado por

$$\left. \begin{array}{rcl} 10x_1 - x_2 + 2x_3 & = & 6 \\ -x_1 + 11x_2 - x_3 + 3x_4 & = & 25 \\ 2x_1 - x_2 + 10x_3 - x_4 & = & -11 \\ 3x_2 - x_3 + 8x_4 & = & 15 \end{array} \right\}$$

que tiene por solución $x^* = (1, 2, -1, 1)^T$. Si se despeja x_i en cada ecuación i -ésima se tiene

$$\left. \begin{array}{l} x_1 = 1/10x_2 - 1/5x_3 + 3/5 \\ x_2 = 1/11x_1 + 1/11x_3 - 3/11x_4 + 25/11 \\ x_3 = -1/5x_1 + 1/10x_2 + 1/10x_4 - 11/10 \\ x_4 = -3/8x_2 + 1/8x_3 + 15/8 \end{array} \right\}$$

Si se elige la aproximación inicial $x^{(0)} = (0, 0, 0, 0)^T$, se genera una sucesión de vectores de la manera siguiente:

$$\left. \begin{array}{l} x_1^{(1)} = 1/10x_2^{(0)} - 1/5x_3^{(0)} + 3/5 = 0.6000 \\ x_2^{(1)} = 1/11x_1^{(0)} + 1/11x_3^{(0)} - 3/11x_4^{(0)} + 25/11 = 2.2727 \\ x_3^{(1)} = -1/5x_1^{(0)} + 1/10x_2^{(0)} + 1/10x_4^{(0)} - 11/10 = -1.1000 \\ x_4^{(1)} = -3/8x_2^{(0)} + 1/8x_3^{(0)} + 15/8 = 1.8750 \end{array} \right\}$$

La siguiente tabla proporciona las diez primeras iteraciones:

k	0	1	2	3	...	8	9	10
$x_1^{(k)}$	0.0000	0.6000	1.0473	0.9326	...	1.0006	0.9997	1.0001
$x_2^{(k)}$	0.0000	2.2727	1.7159	2.0533	...	1.9987	2.0004	1.9998
$x_3^{(k)}$	0.0000	-1.1000	-0.8052	-1.0493	...	-0.9990	-1.0004	-0.9998
$x_4^{(k)}$	0.0000	1.8750	0.8852	1.1309	...	0.9989	1.0006	0.9998

Se ha detenido el cálculo después de diez iteraciones siguiendo el criterio siguiente:

$$\frac{\|x^{(k)} - x^{(k-1)}\|_\infty}{\|x^{(k)}\|_\infty} < 10^{-3}, \text{ cuando, en realidad, el error absoluto es } \|x^{(10)} - x^*\|_\infty = 0.0002$$

Si se vuelve a considerar el sistema $Ax = b$ definido por

$$\left. \begin{array}{l} 10x_1 - x_2 + 2x_3 = 6 \\ -x_1 + 11x_2 - x_3 + 3x_4 = 25 \\ 2x_1 - x_2 + 10x_3 - x_4 = -11 \\ 3x_2 - x_3 + 8x_4 = 15 \end{array} \right\}$$

y se utiliza en cada paso toda la información calculada hasta el momento con la aproximación inicial $x^{(0)} = (0, 0, 0, 0)^T$ se genera una nueva sucesión de vectores de la manera siguiente:

$$\left. \begin{array}{l} x_1^{(k)} = 1/10x_2^{(k-1)} - 1/5x_3^{(k-1)} + 3/5 \\ x_2^{(k)} = 1/11x_1^{(k)} + 1/11x_3^{(k-1)} - 3/11x_4^{(k-1)} + 25/11 \\ x_3^{(k)} = -1/5x_1^{(k)} + 1/10x_2^{(k)} + 1/10x_4^{(k-1)} - 11/10 \\ x_4^{(k)} = -3/8x_2^{(k)} + 1/8x_3^{(k)} + 15/8 \end{array} \right\}$$

obteniéndose la tabla siguiente:

k	0	1	2	3	4	5
$x_1^{(k)}$	0.0000	0.6000	1.0300	1.0065	1.0009	1.0001
$x_2^{(k)}$	0.0000	2.3272	2.0370	2.0036	2.0003	2.0000
$x_3^{(k)}$	0.0000	-0.9873	-1.0140	-1.0025	-1.0003	-1.0000
$x_4^{(k)}$	0.0000	0.8789	0.9844	0.9983	0.9999	1.0000

Se ha detenido el cálculo después de cinco iteraciones debido a

$$\frac{\|x^{(5)} - x^{(4)}\|_\infty}{\|x^{(5)}\|_\infty} = 4 \times 10^{-4}$$

Forma general

Dado un sistema lineal $Ax = b$, se ha construido un nuevo sistema lineal equivalente $x = Bx + c$, que se resuelve de forma iterativa: se comienza con $x^{(0)}$ arbitrario y por medio de la recurrencia

$$x^{(k+1)} = Bx^{(k)} + c \quad k \geq 0$$

y se va calculando una sucesión de vectores $x^{(k)}$ ($k \geq 0$), que tiende (ese es el deseo) a la solución del sistema.

El teorema fundamental de convergencia es el siguiente:

Teorema 4.3. El método iterativo $x^{(k+1)} = Bx^{(k)} + c$ es convergente $\forall x^{(0)} \iff \rho(B) < 1$.

Demostración: Si x^* es la solución, $x^* = Bx^* + c$ y, si se resta esta ecuación de la de iteración, se obtiene

$$x^{(k+1)} - x^* = B(x^{(k)} - x^*) = \dots = B^{k+1}(x^{(0)} - x^*)$$

Si se supone que el método es convergente, se tiene $\lim_{k \rightarrow \infty} (x^{(k+1)} - x^*) = 0$, que es equivalente a $\lim_{k \rightarrow \infty} B^{(k+1)} = 0$. En definitiva, el teorema que hay que demostrar puede enunciarse de la siguiente manera

$$\lim_{k \rightarrow \infty} B^{(k)} = 0 \iff \rho(B) < 1$$

\Leftarrow] Si $\rho(B) < 1$, entonces, dado un $\varepsilon > 0$, existe una norma tal que $\|B\| \leq \rho(B) + \varepsilon$ (ver el apéndice A); por tanto, $\|B\| < 1$ y, entonces,

$$\|B^k\| \leq \|B\|^k \xrightarrow{k \rightarrow \infty} 0$$

y, finalmente, $\lim_{k \rightarrow \infty} B^k = 0$.

\Rightarrow] Si se supone que $\rho(B) \geq 1$ y sea λ un valor propio tal que $|\lambda| \geq 1$ con v vector propio asociado al valor propio λ , entonces,

$$\|B^k v\| = \|\lambda^k v\| \geq \|v\| \implies \frac{\|B^k v\|}{\|v\|} \geq 1$$

y, por lo tanto, $\|B^k\| \geq 1$ para cualquier k ; este resultado implica que $\lim_{k \rightarrow \infty} B^k \neq 0$. \square

4.3.1 Método de Jacobi

Si se considera el sistema $Ax = b$ con $a_{ii} \neq 0$, $i = 1 \div n$, y se despeja incógnita x_i de la ecuación i -ésima (ver el ejemplo introductorio), se tiene

$$x_i = \frac{1}{a_{ii}} \left(- \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j + b_i \right) \quad i = 1 \div n$$

El método de Jacobi consiste en construir una sucesión de vectores que se aproximen a la solución del sistema de la manera siguiente:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(- \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j^{(k)} + b_i \right) \quad i = 1 \div n \quad k = 0, 1, 2, \dots$$

En forma matricial se expresa del siguiente modo: sea la matriz A suma de tres matrices $A = L + D + U$, donde la matriz D es diagonal con todos los elementos no nulos, U es la parte

de A triangular superior y L la inferior. El sistema $A x = b$ es, entonces, equivalente al sistema $x = -D^{-1}(L + U)x + D^{-1}b$ que permite definir el método de Jacobi en forma matricial

$$x^{(k+1)} = -D^{-1}(L + U)x^{(k)} + D^{-1}b$$

donde la matriz de iteración es $B_J = -D^{-1}(L + U)$. Así, por ejemplo, se tiene

$$\|B_J\|_\infty = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^n \left| \frac{a_{ij}}{a_{ii}} \right|, \text{ y si } A \text{ es una matriz diagonal dominante estricta (con } |a_{ii}| >$$

$$\sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|, \quad i = 1 \div n), \text{ entonces, } \|B_J\|_\infty < 1, \text{ y el método de Jacobi converge.}$$

4.3.2 Método de Gauss-Seidel

El método de Jacobi no utiliza los valores calculados en un paso $x_i^{(k+1)}$ hasta obtener todas las componentes del vector. El método de Gauss-Seidel los aprovecha en cuanto los tiene (ver el ejemplo del principio de esta sección):

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(- \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} + b_i \right)$$

$$i = 1 \div n \quad k = 0, 1, 2, \dots$$

En lenguaje matricial se tiene el **método iterativo de Gauss-Seidel** al traducir a:

$$x^{(k+1)} = -D^{-1}Lx^{(k+1)} - D^{-1}Ux^{(k)} + D^{-1}b$$

y, en definitiva, se tiene

$$x^{(k+1)} = -(D + L)^{-1}Ux^{(k)} + (D + L)^{-1}b$$

La matriz de iteración es $B_{GS} = -(D + L)^{-1}U$.

Si se considera la norma sub-infinito, $\|B_J\|_\infty = \max_{x \neq 0} \frac{\|y\|_\infty}{\|x\|_\infty}$, con $y = B_{GS}x$. Sea k tal que $\|y\|_\infty = |y_k|$; entonces, de la k -ésima ecuación de $y = -D^{-1}L y - D^{-1}Ux$, se tiene

$$|y_k| = \|y\|_\infty \leq s_k \|y\|_\infty + r_k \|x\|_\infty$$

donde $s_k = \sum_{j=1}^{k-1} \left| \frac{a_{kj}}{a_{kk}} \right|$ y $r_k = \sum_{k+1}^n \left| \frac{a_{kj}}{a_{kk}} \right|$, y se tiene

$$\|B_{GS}\|_\infty \leq \max_{1 \leq k \leq n} \frac{r_k}{1 - s_k} < 1 \quad \text{si } r_k + s_k < 1$$

y el método de Gaus-Seidel es convergente si A es una matriz diagonal dominante.

Ejercicio. Si A es una matriz simétrica definida positiva, demostrar que el método de Gauss-Sidel es convergente.

4.3.3 Razón de convergencia y estimación del error

De la desigualdad siguiente

$$\|x^{(k)} - x^*\| \leq \|B\|^k \|x^{(0)} - x^*\|$$

parece lógico definir el **factor de convergencia asintótico**

$$\alpha = \lim_{k \rightarrow \infty} \|x^{(k)} - x^*\|^{1/k}$$

que da un criterio de convergencia del método iterativo a partir de la matriz de iteración B ; cuanto menor sea α , mayor convergencia o, también, serán necesarias menos iteraciones.

Si se tiene una norma tal que $\|B\| \leq \rho(B) + \varepsilon$, entonces

$$\begin{aligned} \|x^{(k)} - x^*\| &\leq (\rho(B) + \varepsilon)^k \|x^{(0)} - x^*\| \\ \|x^{(k)} - x^*\|^{1/k} &\leq (\rho(B) + \varepsilon) \|x^{(0)} - x^*\|^{1/k} \end{aligned}$$

y, de aquí, $\alpha = \lim_{k \rightarrow \infty} \|x^{(k)} - x^*\|^{1/k} \leq \rho(B) + \varepsilon$; por tanto, $\alpha \leq \rho(B)$. Se puede demostrar ([Ort72]) que existe una sucesión de vectores $x^{(k)}$ tal que $\alpha = \rho(B)$. Cuanto menor sea $\rho(B)$, y por tanto α , mayor será la convergencia; se puede definir la **velocidad de convergencia** por $R = -\log(\rho(B))$.

Para tener una estimación del error en $x^{(k)}$, se considera el vector error k -ésimo expresado de la siguiente forma

$$x^{(k)} - x^* = -B(x^{(k)} - x^{(k-1)}) + B(x^{(k)} - x^*)$$

Si $\|B\| = \beta < 1$, se obtiene

$$\|x^{(k)} - x^*\| \leq \frac{\beta}{1-\beta} \|x^{(k)} - x^{(k-1)}\|$$

Ejercicio. Del hecho que $B(x^{(k)} - x^{(k-1)}) = B^k(x^{(1)} - x^{(0)})$, deducid que, si $\|B\| = \beta < 1$, entonces

$$\|x^{(k)} - x^*\| \leq \frac{\beta^k}{1-\beta} \|x^{(1)} - x^{(0)}\|$$

4.3.4 Refinamiento iterativo de la solución

Si la matriz A del sistema $Ax = b$ es mal condicionada, la solución calculada, x_0^* , puede representar una mala aproximación de la solución exacta. A fin de mejorar la exactitud se considera el método siguiente:

$$r_k = b - Ax_k^* \quad Ad_k = r_k \quad x_{k+1}^* = x_k^* + d_k$$

Si previamente se ha factorizado la matriz $A = LU$, las correcciones d_k se calculan resolviendo los dos sistemas triangulares. La sucesión de vectores x_k^* convergerá, en general, hacia la solución x^* del sistema $Ax = b$.

Ejemplo. Si se considera el sistema $(A|b) = \left(\begin{array}{ccc|c} 3.3330 & 15920 & -10.333 & 15913 \\ 2.2220 & 16.710 & 9.6120 & 28.544 \\ 1.5611 & 5.1791 & 1.6852 & 8.4254 \end{array} \right)$, se

tiene la solución exacta $(1, 1, 1)^T$; pero si se realiza la eliminación gaussiana con redondeo a 5 cifras, se obtiene la solución aproximada $x_0^* = (1.2001, 0.99991, 0.92538)^T$; el vector residual es $r_0 = b - Ax_0^* = (-0.00518, 0.27413, -0.18616)^T$ con $\|r_0\|_\infty = 0.27413$. Si se denota por d_0 a la solución aproximada de $Ad = r_0$,

$$d_0 \approx A^{-1}r_0 = A^{-1}(b - Ax_0^*) = A^{-1}b - A^{-1}Ax_0^* = x^* - x_0^*$$

por tanto, d_0 es una estimación del error producido en aproximar la solución x^* original del sistema por la aproximación x_0^* . Entonces,

$$\|d_0\| \approx \|x^* - x_0^*\| \leq \|A^{-1}\| \cdot \|r_0\| \approx \|A^{-1}\| \cdot 10^{-t} \|A\| \|x_0^*\| = 10^{-t} \|x_0^*\| \kappa(A)$$

De este modo (ver [For67]) se tiene una manera de calcular el número de condición sin hacer el cálculo de la inversa de A :

$$\kappa(A) \approx \frac{\|d_0\|}{\|x_0^*\|} 10^t \quad (4.5)$$

donde aparece, en el denominador, la solución x_0^* del sistema $Ad = r_0$ mediante eliminación gaussiana y una aritmética de t dígitos.

Si se resuelve $Ad_0 = r_0$, entonces $d_0 = (-0.20008, 8.9987 \cdot 10^{-5}, 0.074607)^T$ y con la estimación 4.5

$$\kappa(A) \approx 10^5 \frac{0.2008}{1.2001} = 16672$$

El cálculo directo de A^{-1} , con el esfuerzo computacional que requiere, da lugar a $\kappa(A) = 15999$ (del mismo orden).

Ya que se conoce la solución, se puede calcular el error relativo

$$\|x^* - x_0^*\|_\infty = 0.2001 \quad \text{y} \quad \frac{\|x^* - x_0^*\|_\infty}{\|x^*\|_\infty} = 0.2001$$

y una cota del error viene dada por

$$\frac{\|x^* - x_0^*\|_\infty}{\|x^*\|_\infty} \leq \kappa(A) \frac{\|r_0\|_\infty}{\|b\|_\infty} = \frac{15999 \cdot 0.27413}{15934} = 0.27525 \quad (\text{demasiado grande})$$

Si se aplica el método de refinamiento iterativo:

$$x_1^* = x_0^* + d_0 = (1.0000, 1.0000, 0.99999)^T$$

Sea $r_1 = b - Ax_1^*$ y se resuelve el sistema $Ad_1 = r_1$. Se obtiene, entonces,

$$d_1 = (1.5002 \cdot 10^{-9}, 2.0951 \cdot 10^{-10}, 1.0 \cdot 10^{-5})^T$$

y como $\|d_1\|_\infty \leq 10^{-5}$, finalmente se llega a

$$x_2^* = x_1^* + d_1 = (1.0000, 1.0000, 1.0000)^T$$

4.3.5 Métodos de sobrerelajación

Teniendo en cuenta el método de Gauss-Seidel

$$a_{ii} \tilde{x}_i^{(k+1)} = - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} + b_i \quad i = 1 \div n \quad k = 0, 1, 2, \dots$$

si se elige como nuevo vector de la sucesión, $x_i^{(k+1)}$ entre el $\tilde{x}_i^{(k+1)}$ calculado con este método (G-S) y el vector $x_i^{(k)}$ del paso anterior:

$$x_i^{(k+1)} = (1 - \omega) x_i^{(k)} + \omega \tilde{x}_i^{(k+1)}$$

que sólo depende de ω y recibe el nombre de **factor de relajación** (sobrerelajación si $\omega > 1$, subrelajación si $\omega < 1$). Substituyendo $\tilde{x}_i^{(k+1)}$, se tiene

$$\begin{aligned} a_{ii} x_i^{(k+1)} &= a_{ii} x_i^{(k)} + \omega \left(- \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} + b_i - a_{ii} x_i^{(k)} \right) \\ i &= 1 \div n \quad k = 0, 1, 2, \dots \end{aligned}$$

En notación matricial, donde $A = L + D + U$, se obtiene

$$(D + \omega L) \mathbf{x}^{(k+1)} = \{(1 - \omega) D - \omega U\} \mathbf{x}^{(k)} + \omega \mathbf{b} \quad k \geq 0$$

con $B_\omega = (D + \omega L)^{-1} \{(1 - \omega) D - \omega U\}$ y para $\omega = 1$, se tiene $B_\omega = B_{GS}$. El teorema de Kahan sirve para tener una idea de los valores que puede tomar ω :

Teorema 4.4. Si A tiene todos los elementos diagonales no nulos, $a_{ii} \neq 0$, $i = 1 \div n$, entonces $|\omega - 1| \leq \rho(B_\omega)$.

Demostración: Como L es una matriz triangular inferior estricta, $\det D^{-1} = \det(D + \omega L)^{-1}$. Y de aquí, al considerar que U es también triangular estricta, $\det B_\omega = \det(D + \omega L)^{-1}$. $\det \{(1 - \omega) D - \omega U\} = \det[(1 - \omega) I - \omega D^{-1} U] = \det[(1 - \omega) I] = (1 - \omega)^n$. Entonces, el producto de los valores propios de B_ω no puede ser estrictamente menor que $1 - \omega$. \square

Por otro lado, para que haya convergencia es necesario que $\rho(B_\omega) < 1$. Entonces, $|\omega - 1| < 1 \implies 0 < \omega < 2$ (si ω es real). Se verá que esta condición es suficiente para la convergencia de una clase importante de matrices:

Una matriz A es **2-cíclica** si existe una matriz de permutaciones P tal que $P^T A P = \begin{pmatrix} D_1 & C_1 \\ C_2 & D_2 \end{pmatrix}$, donde D_i , $i = 1, 2$, son matrices diagonales. Además, se dirá que una matriz A es **consistentemente ordenada** si los valores propios de la matriz $B(\alpha) = \alpha^{-1} D^{-1} L + \alpha D^{-1} U$ son independientes de α . Normalmente se utiliza $S = \text{diag}(1, \alpha, \alpha^2, \dots, \alpha^{n-1})$ tal que $S B(\alpha) S^{-1}$ no depende de α .

Teorema 4.5. Si $A \in \mathcal{L}(\mathbf{R}^n)$ es 2-cíclica, consistentemente ordenada y $a_{ii} \neq 0$, $i = 1 \div n$, existen $p, r \in \mathbf{N}$, con $p + 2r = n$ y $\mu_1, \dots, \mu_r \in \mathbf{C}$ tales que

$$\begin{aligned} 1) \det(\mu I - B_J) &= \mu^p \cdot \prod_{i=1}^r (\mu^2 - \mu_i^2) \\ 2) \det(\lambda I - B_\omega) &= (\lambda + \omega - 1)^p \cdot \prod_{i=1}^r [(\lambda + \omega - 1)^2 - \lambda \omega^2 \mu_i^2] \end{aligned}$$

Por tanto, si μ_i es un valor propio de la matriz de iteración del método de Jacobi, también lo es $-\mu_i$; además, B_ω tiene tantos valores propios igual a $1 - \omega$ como valores propios igual a 0 tiene B_J . Finalmente, si $\pm\mu_i$ son valores propios de B_J , entonces los ceros de $(\lambda + \omega - 1)^2 - \lambda \omega^2 \mu_i^2 = 0$ lo son de B_ω .

Demostración: Ver [Ort72]. \square

Como consecuencia de este teorema, se puede asegurar que, en las condiciones del teorema, se tiene que $\rho(B_{GS}) = \rho(B_J)^2$, y, por lo tanto, si el método de Jacobi es convergente, también lo es el de Gauss-Seidel, y el factor de convergencia asintótica es el cuadrado del de Jacobi. Hay ejemplos donde el método de Gauss-Seidel no es convergente y Jacobi lo es (ver el problema 7).

La respuesta a la cuestión de qué valor de ω_O minimiza el radio espectral $\rho(B_\omega)$ con $0 < \omega < 2$, la da el teorema siguiente:

Teorema 4.6. Si $A \in \mathcal{L}(\mathbf{R}^n)$ es 2-cíclica, consistentemente ordenada, $a_{ii} \neq 0$, $i = 1 \div n$, $\rho(B_J) < 1$ y los valores propios de B_J son todos reales; entonces,

$$\omega_O = \frac{2}{1 + \sqrt{1 - \rho(B_J)^2}} \quad \text{y} \quad \rho(B_{\omega_O}) = \omega_O - 1$$

Demostración: Los valores propios de B_ω son soluciones de

$$\left\{ \begin{array}{l} \lambda = 1 - \omega \\ (\lambda + \omega - 1)^2 = \lambda \omega^2 \mu_i^2 \end{array} \right.$$

que son los puntos de intersección de la recta $y = g_\omega(\lambda) = \frac{\lambda + \omega - 1}{\omega}$ con la parábola $y = m_i(\lambda) = \pm\sqrt{\lambda} \mu_i$.

Si se varía la ω de 1 hasta 0 la recta $y = g_\omega(\lambda)$ que pasa por el punto $(1, 1)$, va aumentando su pendiente de manera que las raíces $\lambda_i^{(-)}$ y $\lambda_i^{(+)}$ van creciendo monótonamente. El otro valor posible, $\lambda = 1 - \omega$, también crece de 0 hacia 1. En definitiva, si $\omega \in [0, 1]$ crece, $\rho(B_\omega)$ decrece.

Si ω crece por encima de 1, la recta $y = g_\omega(\lambda)$ corta el eje y en la parte positiva. En la medida que ω crece hacia 2, habrá un momento de tangencia donde se tendrá el límite de los valores de corte con la parábola $y = m_i(\lambda)$, donde $\lambda_i^{(-)}$ crece y $\lambda_i^{(+)}$ decrece. Al hacer coincidir las pendientes y los valores de las funciones

$$\frac{1}{\omega} = \frac{1}{2} \lambda^{-1/2} \mu_i \quad \frac{\lambda + \omega - 1}{\omega} = \lambda^{1/2} \mu_i$$

y eliminando λ_i , se tiene una ecuación de segundo grado en ω : $\mu_i^2 \omega^2 - 4\omega + 4 = 0$, que tiene la solución

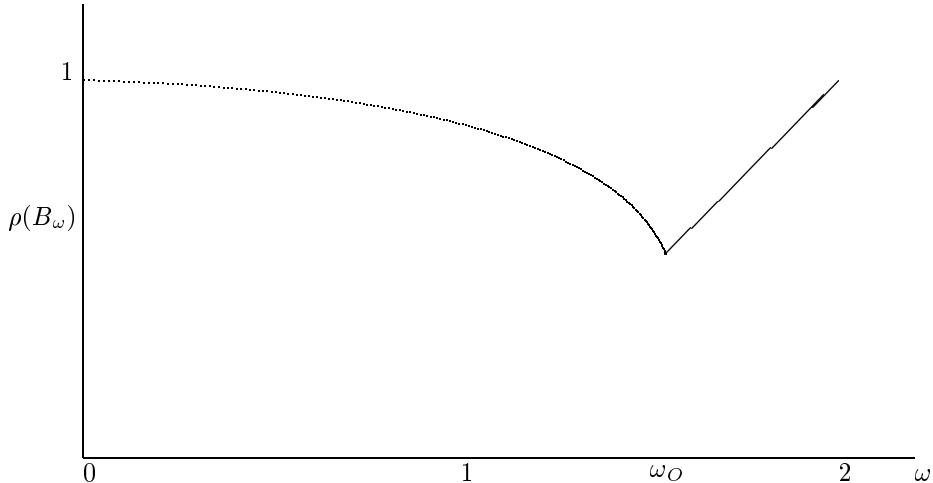
$$\omega = \frac{2}{1 + \sqrt{1 - \mu_i^2}} \quad \text{y} \quad \sqrt{\lambda_i^{(+)}} = \frac{1}{2} \left[\omega \mu_i \pm \sqrt{\omega^2 \mu_i^2 - 4\omega + 4} \right]$$

Si ω crece por encima de este valor, las raíces correspondientes a las λ_i son complejas y $|\lambda_i| = \lambda_i^{1/2} \bar{\lambda}_i^{1/2} = \omega - 1$. \square

En general, se tiene

$$\rho(B_\omega) = \begin{cases} \omega - 1 & \text{si } \omega_O \leq \omega < 2 \\ 1 - \omega + \frac{1}{2}\omega^2 \mu^2 + \omega \mu \sqrt{1 - \omega + \frac{1}{4}\omega^2 \mu^2} & \text{si } 0 < \omega \leq \omega_O \end{cases}$$

donde $\mu = \rho(B_\omega)$ según la gráfica siguiente:



Ejercicio. Dado el sistema $Ax = b$ con $A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ y $b = \begin{pmatrix} 7 \\ 1 \\ 1 \end{pmatrix}$,

comprobar que el valor de ω óptimo para aplicar el método de sobrerelajación es $4 - 2\sqrt{2}$, que da lugar a un radio espectral de $3 - 2\sqrt{2}$, mientras que con el método de Jacobi se tiene un radio igual a $\frac{1}{\sqrt{2}}$.

4.4 Sistemas lineales sobredeterminados

Se quiere dar un modelo lineal para un conjunto de datos medidos o calculados a partir de la observación; entonces, se realizan un número de medidas mayor que el número de incógnitas con la finalidad de reducir la influencia de los errores.

Dada una matriz A del tipo $m \times n$, con $m \geq n$, y el vector b (del tipo $m \times 1$), se quiere calcular un vector x tal que Ax sea ‘la mejor’ aproximación de b . Se considera la solución por el método de mínimos cuadrados, es decir, se calcula el vector x que minimiza la norma euclídea del vector residual $\|r\|_2 = \sqrt{r^T \cdot r}$, donde $r = b - Ax$.

Teorema 4.7. Si x satisface las **ecuaciones normales** $A^T(b - Ax) = 0$, entonces, $\forall y \in \mathbf{R}^n$

$$\|b - Ax\|_2 \leq \|b - Ay\|_2$$

Demostración: Sean $r_x = b - Ax$ e $r_y = b - Ay$; entonces,
 $r_y = (b - Ax) + (Ax - Ay) = r_x + A(x - y)$

de donde $r_y^T r_y = r_x^T r_x + (x - y)^T A^T A (x - y)$, ya que $A^T r_x = 0$. Por lo tanto, $\|r_y\|_2^2 = \|r_x\|_2^2 + \|A(x - y)\|_2^2 \geq \|r_x\|_2^2$. \square

Entonces, el sistema $A^T A x = A^T b$, donde $A^T A$ es una matriz cuadrada simétrica de orden n , da lugar a n ecuaciones lineales con n incógnitas.

Teorema 4.8. $A^T A$ es una matriz no singular \iff las columnas de A son linealmente independientes.

Demostración: \Leftarrow] Si $x \neq 0$, $Ax \neq 0$. Además

$$x^T (A^T A) x = (Ax)^T (Ax) = \|Ax\|_2^2 > 0$$

y $A^T A$ es definida positiva y, por lo tanto, no singular.

\Rightarrow] Si las columnas de A son linealmente dependientes, $\exists x_0 \neq 0$ tal que $Ax_0 = 0$, entonces $A^T A x_0 = 0$ y la matriz $A^T A$ es singular. \square

Ejemplo. Las observaciones en coordenadas polares del movimiento del cometa 1968 Tentax han sido

r	2.70	2.00	1.61	1.20	1.02
φ	48°	67°	83°	108°	126°

Las coordenadas satisfacen $r = \frac{p}{1 - e \cos \varphi}$, donde p es un parámetro y e la excentricidad de la órbita. Para calcular p y e se expresan los parámetros en forma lineal: $\frac{1}{p} - \frac{e}{p} \cos \varphi = \frac{1}{r}$. El sistema sobre determinado viene dado por

$$\begin{pmatrix} 1 & -0.669 \\ 1 & -0.390 \\ 1 & -0.122 \\ 1 & 0.309 \\ 1 & 0.588 \end{pmatrix} \begin{pmatrix} 1/p \\ e/p \end{pmatrix} = \begin{pmatrix} 0.370 \\ 0.500 \\ 0.621 \\ 0.833 \\ 0.980 \end{pmatrix}$$

que da lugar a las ecuaciones normales

$$\begin{pmatrix} 5 & -0.285 \\ -0.285 & 1.058 \end{pmatrix} \begin{pmatrix} 1/p \\ e/p \end{pmatrix} = \begin{pmatrix} 3.304 \\ 0.315 \end{pmatrix}$$

Las soluciones son $1/p = 0.688$ y $e/p = 0.482$ que clasifican la órbita en una elipse de excentricidad $e = 0.7$ y con el parámetro de la órbita $p = 1.45$.

4.5 Problemas

1. Resolver el sistema lineal $Ax = b$ con:

$$A = \begin{pmatrix} 0.15 & 2.11 & 30.75 \\ 0.64 & 1.21 & 2.05 \\ 3.21 & 1.53 & 1.04 \end{pmatrix} \quad y \quad b = \begin{pmatrix} -26.38 \\ 1.01 \\ 5.23 \end{pmatrix},$$

por el método de eliminación gaussiana:

- a) sin intercambiar ninguna fila ni columna.
- b) con pivotamiento parcial.
- c) con pivotamiento completo y comparando los resultados anteriores con la solución exacta.

2. Efectuar la factorización LU de la matriz:

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 3 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 5 \end{pmatrix}$$

3. Calcular la inversa de la siguiente matriz $A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 6 & 27 & 64 \\ 1 & 16 & 81 & 216 \end{pmatrix}$, en primer lugar por el método de Gauss-Jordan, y posteriormente, a partir de la factorización LU .

4. Considerar el sistema $Ax = b$ con

$$(A | b) = \left(\begin{array}{cccc|c} 10 & 7 & 8 & 7 & 32 \\ 7 & 5 & 6 & 5 & 23 \\ 8 & 6 & 10 & 9 & 33 \\ 7 & 5 & 9 & 10 & 31 \end{array} \right)$$

Resolverlo exactamente.

Comprobar, después, que $(6, -7.2, 2.9, -0.1)^T$ y $(1.50, 0.18, 1.19, 0.89)^T$ dan vectores residuales muy pequeños. Estimar el número de condición de la matriz.

5. Considerar el sistema de ecuaciones lineal siguiente:

$$\begin{cases} 1.000x_1 + 2.000x_2 = 3.000 \\ 0.499x_1 + 1.001x_2 = 1.500 \end{cases}$$

y la solución aproximada $x_1 = 2.000$ y $x_2 = 0.500$.

- a) Calcular la solución exacta.
- b) Calcular el residuo (norma del vector residual) de la solución aproximada dada.
- c) Estudiar el condicionamiento de la matriz.
- d) La solución aproximada, ¿es una buena solución?, ¿por qué?

6. Sean $A_n = \begin{pmatrix} 1 & 2 \\ 2 & 4 + \frac{1}{n^2} \end{pmatrix}$ y $b_n = \begin{pmatrix} 1 \\ 2 - \frac{1}{n^2} \end{pmatrix}$. Una solución “aproximada” (residuo pequeño) del sistema $A_{100}x = b_{100}$ es $\tilde{x} = (1, 0)^T$.

- a) ¿Es exacta la solución?
- b) Discutir el condicionamiento del problema para n cualquiera.
- c) Estimar el error relativo de la solución \tilde{x} cuando n crece.

7. Demostrar que, para el sistema siguiente:

$$\begin{cases} x_1 + 2x_2 - 2x_3 = 1 \\ x_1 + x_2 + x_3 = 1 \\ 2x_1 + 2x_2 + x_3 = 1 \end{cases}$$

el método de Jacobi converge y, en cambio, el de Gauss-Seidel no.

8. Considerar el sistema lineal $Ax = b$ donde $A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 1 \\ a & 0 & 1 \end{pmatrix}$. ¿Para qué valores de a converge el método de Jacobi?
9. Dado el sistema $Ax = b$ con $A = \begin{pmatrix} I & M \\ M^T & I \end{pmatrix}$, donde I y M son matrices del tipo $n \times n$ (I matriz identidad), probar que el método iterativo de Jacobi es convergente si $\|M\|_2 < 1$.

10. Dado el sistema de ecuaciones

$$\begin{pmatrix} 1 & -a \\ -a & 1 \end{pmatrix} \mathbf{x} = \mathbf{b}$$

donde $a \in \mathbf{R}$, bajo ciertas condiciones, puede resolverse por el método iterativo de sobre-relajación

$$\begin{pmatrix} 1 & 0 \\ -\omega a & 1 \end{pmatrix} \mathbf{x}^{(k+1)} = \begin{pmatrix} 1 - \omega & \omega a \\ 0 & 1 - \omega \end{pmatrix} \mathbf{x}^{(k)} + \omega \mathbf{b}$$

- 1. ¿Para qué valores de a es el método convergente si $\omega = 1$?
- 2. Comprobar que el valor óptimo de ω es $\frac{2}{a^2}(1 - \sqrt{1 - a^2})$.
- 3. Si se toma $a = 0.5$, calcular el valor de $\omega \in \{0.8, 0.9, 1.0, 1.1, 1.2, 1.3\}$ que minimiza el radio espectral de la matriz

$$\begin{pmatrix} 1 & 0 \\ -\omega a & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 - \omega & \omega a \\ 0 & 1 - \omega \end{pmatrix}$$

11. Se tiene el sistema lineal $x = b - 3Mx$ y se sabe que $\|M\|_\infty = 0.28$. Si se resuelve el sistema usando el método iterativo $x_{k+1} = b - 3Mx_k$. ¿Cuántas iteraciones son necesarias para disminuir el error en un factor 10^{-6} ?

12. Sea A una matriz $n \times n$ y se supone que X_0 es una aproximación de A^{-1} ; se considera la sucesión de matrices definidas por la recurrencia:

$$E_k = I - A X_k \quad X_{k+1} = X_k (I + E_k + E_k^2) \quad (k \geq 0)$$

- a) Comprobar que $E_{k+1} = E_k^3$ y que $X_k = A^{-1} (I - E_0^{3^k})$ ($k \geq 0$).
- b) Dar condiciones necesarias y suficientes de convergencia de X_k ($k \geq 0$) hacia A^{-1} .
- c) Si se parte de una matriz X_0 tal que $E_0 = \begin{pmatrix} 0.2 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.1 \\ 0.3 & 0.1 & 0.2 \end{pmatrix}$, ¿cuántas iteraciones k se tienen que ejecutar para que se pueda asegurar que $\|E_k\|_\infty < 10^{-12}$?

13. La intensidad de radiación de una fuente radioactiva viene dada por la siguiente ley

$$I = I_0 e^{-\alpha t}$$

Determinar las constantes α y I_0 sabiendo que se han realizado las siguientes medidas:

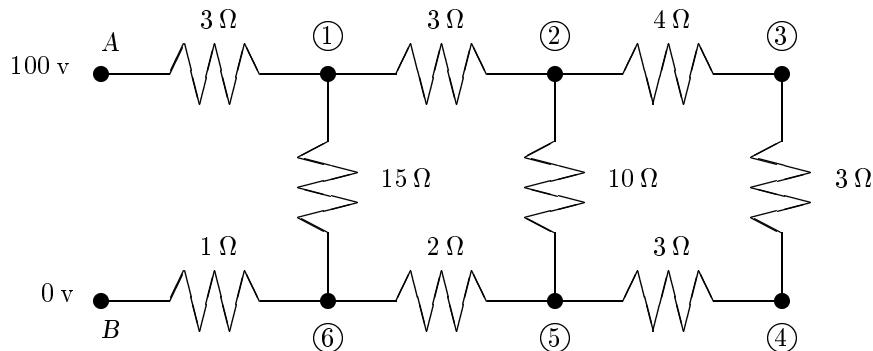
t	0.2	0.3	0.4	0.5	0.6	0.7	0.8
I	3.16	2.38	1.75	1.34	1.00	0.74	0.56

4.6 Prácticas

4.6.1 Práctica ejemplo

Escribir una rutina que utilice el método de eliminación de Gauss con pivotamiento parcial para la resolución de un sistema lineal con n ecuaciones.

Aplicarlo al esquema eléctrico siguiente, calculando las soluciones (los potenciales en los nudos) donde los valores de las resistencias se dan en ohms y el potencial entre A i B es de 100 voltios.



Se construyen las rutinas DECOMP (que realiza la factorización de Gauss de una matriz) y SOLVE (que resuelve un sistema lineal $A \cdot X = B$ del que se ha calculado previamente la factorización gaussiana), tal que mediante una llamada se encuentra la solución de un sistema de ecuaciones lineales (ver [For77]).

El sistema de ecuaciones lineales que se quiere resolver corresponde a las ecuaciones lineales de los potenciales en los nudos del esquema eléctrico presentado. Los valores de las resistencias se expresan en ohmios y el potencial aplicado entre A y B es igual a 100 voltios.

Según la ley de Ohm, la intensidad de la corriente I_{pq} que circula entre los nudos p y q en el tramo pq es igual a:

$$I_{pq} = \frac{V_p - V_q}{R_{pq}}$$

donde V_p y V_q son los potenciales en los nudos p y q respectivamente, y R_{pq} es la resistencia en el tramo pq . Para valores de R_{pq} dados en ohmios la intensidad vendrá dada en amperios. La ecuación que relaciona los potenciales en los nudos del esquema eléctrico se obtienen aplicando la ley de intensidades de Kirchoff: ‘La suma de intensidades de la corriente en un nudo es nula’.

La aplicación de estas dos leyes en cada uno de los nudos da lugar a las seis ecuaciones del sistema que

se quiere resolver: $Ax = b$, donde

$$A = \begin{pmatrix} 11.0 & -5.0 & 0.0 & 0.0 & 0.0 & -1.0 \\ -20.0 & 41.0 & -15.0 & 0.0 & -6.0 & 0.0 \\ 0.0 & -3.0 & 7.0 & -4.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & -1.0 & 2.0 & -1.0 & 0.0 \\ 0.0 & -3.0 & 0.0 & -10.0 & 28.0 & -15.0 \\ -2.0 & 0.0 & 0.0 & 0.0 & -15.0 & 47.0 \end{pmatrix} \quad y \quad b = \begin{pmatrix} 500.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{pmatrix}.$$

Cada ecuación representa el voltaje en el nudo correspondiente del sistema eléctrico original.

Se presenta a continuación el programa principal y las rutinas antes mencionadas SUBROUTINE DECOMP y SUBROUTINE SOLVE.

```

PROGRAM CIRCUITO
EXTERNAL DECOMP,SOLVE

REAL*8 A(10,10),B(10),TREB(10),COND,CONDPI
INTEGER IPVT(10),I,J,N,NDIM

NDIM= 10
N= 6
C
C      INICIALIZACION DEL SISTEMA
C
DO 1 I = 1, N
    READ(1,*) (A(I,J), J=1,N)
1 CONTINUE

DO 12 I = 1, N
    WRITE(6,2) (A(I,J), J=1,N)
12 CONTINUE
2 FORMAT(1H ,10F5.1)
    WRITE(6,8)
C
C      DESCOMPOSICION DE GAUSS
C
CALL DECOMP(NDIM,N,A,COND,IPVT,TREB)

WRITE(6,3) COND
3 FORMAT(6H COND=, E15.5)
    WRITE(6,8)
C
C      COMPROBACION DEL NUMERO DE CONDICIONES
C      POR SI LA MATRIZ ES SINGULAR
C
CONDPI = COND + 1
IF (CONDPI .EQ. COND) WRITE(6,4)
4 FORMAT(34H MATRIZ SINGULAR CON ESTA PRECISIION)
    IF (CONDPI .EQ. COND) STOP

DO 13 I=1,N,1
    READ(1,*),B(I)
13 CONTINUE
C
C      RESOLUCION DEL SISTEMA CON SOLVE
C

```

```

CALL SOLVE(NDIM,N,A,B,IPVT)
DO 6 I = 1, N
    WRITE(6,7) B(I)
6 CONTINUE
7 FORMAT(1H ,F10.5)
STOP
8 FORMAT(1H )
END

SUBROUTINE DECOMP(NDIM,N,A,COND,IPVT,TREB)
C
INTEGER NDIM,N
REAL*8 A(NDIM,N),COND,TREB(N)
INTEGER IPVT(N)

C DESCOMPOSICION DE LA MATRIZ A POR ELIMINACION GAUSSIANA Y ESTIMACION
C DEL NUMERO DE CONDICION DE LA MATRIZ A.

C UTILIZA SOLVE PARA CALCULAR LA SOLUCION DE SISTEMAS LINEALES.

C ENTRADA
C NDIM = NUMERO DE ECUACIONES DEL SISTEMA QUE CONTIENE A.
C N = ORDEN DE LA MATRIZ A.
C A = MATRIZ QUE SE QUIERE TRIANGULARIZAR

C SALIDA
C A = CONTIENE LA MATRIZ TRIANGULAR SUPERIOR U
C Y UNA VERSION PERMUTADA DE LA MATRIZ TRIANGULAR
C INFERIOR I-L DE TAL MANERA QUE (MATRIZ PERMUTADA)*A=L*U

C COND = ESTIMACION DEL NUMERO DE CONDICION DE A.

C PARA EL SISTEMA LINEAL A*X = B, CAMBIOS EN A Y B
C PUEDEN PRODUCIR GRANDES CAMBIOS EN EL NUMERO DE COND
C SI COND+1.0 .EQ. COND, A ES SINGULAR PARA LA PRECISION DADA.

C COND= 1.0E+32 SI SE DETECTA SINGULARIDAD.

C IPVT = VECTOR DE PIVOTAMIENTOS.

C IPVT(K) = INDICE DEL K-ESIMO PIVOTE
C IPVT(N) =(-1)**(NUMERO DE INTERCAMBIOS)

C TREB = ESPACIO DE TRABAJO. EL VECTOR TREB TIENE QUE SER INCLUIDO
C EN LA LLAMADA. EL CONTENIDO DE ENTRADA SE IGNORA.

C EL CONTENIDO DE SALIDA ES GENERALMENTE POCO IMPORTANTE.

C EL DETERMINANTE DE A SE PUEDE OBTENER EN LA SALIDA SI SE EJECUTA:
C DET(A) = IPVT(N) * A(1,1) * A(2,2) * ...* A(N,N).

```

```

REAL*8 EK, T, ANORM, YNORM, ZNORM
INTEGER NM1,I,J,K,KP1,KB,KM1,M
C
IPVT(N) = 1
IF (N .EQ. 1) GO TO 80
NM1 = N - 1
C
C CALCULO DE LA NORMA_1 DE A
C
ANORM = 0.0
DO 10 J = 1, N
T = 0.0
DO 5 I = 1, N
T = T + DABS(A(I,J))
5 CONTINUE
IF (T .GT. ANORM) ANORM = T
10 CONTINUE
C
C ELIMINACION GAUSSIANA CON PIVOTAMIENTO PARCIAL
C
DO 35 K = 1,NM1
KP1= K+1
C
C SE BUSCA EL PIVOTE
C
M = K
DO 15 I = KP1,N
IF (DABS(A(I,K)) .GT. DABS(A(M,K))) M = I
15 CONTINUE
IPVT(K) = M
IF (M .NE. K) IPVT(N) = -IPVT(N)
T = A(M,K)
A(M,K) = A(K,K)
A(K,K) = T
C
C SEGUIENTE PASO SI EL PIVOTE ES CERO
C
IF (T .EQ. 0.0) GO TO 35
C
C CALCULO DE LOS COEFICIENTES
C
DO 20 I = KP1,N
A(I,K) = -A(I,K)/T
20 CONTINUE
C
C INTERCAMBIO Y ELIMINACION POR COLUMNAS
C
DO 30 J = KP1,N
T = A(M,J)
A(M,J) = A(K,J)
A(K,J) = T
IF (T .EQ. 0.0) GO TO 30
DO 25 I = KP1,N
A(I,J) = A(I,J) + A(I,K)*T
25 CONTINUE
30 CONTINUE
35 CONTINUE

C COND= (NORMA_1 DE A)*(ESTIMACION DE LA NORMA_1 DE LA MATRIZ INVERSA DE A)

```

```

C   LA ESTIMACION SE OBTIENE CON UN PASO DE ITERACION INVERSA POR EL VECTOR
C   SINGULAR. LO QUE IMPLICA RESOLVER DOS SISTEMAS DE ECUACIONES

C   (A-TRASPUESTA)*Y= E C, A*Z= Y DONDE E ES UN VECTOR DE 1 0 -1 PARA CAUSAR
C   CRECIMIENTO EN Y.

C   ESTIMACION = (NORMA_1 DE Z)/(NORMA_1 DE Y)

C   RESOLUCION DEL SISTEMA (A-TRASPUESTA)*Y = E

DO 50 K = 1,N
T = 0.0
IF (K .EQ. 1) GO TO 45
KM1 = K-1
DO 40 I = 1, KM1
    T = T + A(I,K)*TREB(I)
40  CONTINUE
45  EK = 1.0
    IF (T .LT. 0.0) EK = -1.0
    IF (A(K,K) .EQ. 0.0) GO TO 90
    TREB(K) = -(EK + T)/A(K,K)
50 CONTINUE
DO 60 KB = 1, NM1
    K = N - KB
    T = 0.0
    KP1 = K+1
    DO 55 I = KP1, N
        T = T + A(I,K)*TREB(K)
55  CONTINUE
    TREB(K) = T
    M = IPVT(K)
    IF (M .EQ. K) GO TO 60
    T = TREB(M)
    TREB(M) = TREB(K)
    TREB(K) = T
60 CONTINUE
C
      YNORM = 0.0
      DO 65 I = 1, N
          YNORM = YNORM + DABS(TREB(I))
65  CONTINUE
C
C   RESOLUCION DEL SISTEMA A*Z = Y
C
      CALL SOLVE(NDIM, N, A, TREB, IPVT)
C
     ZNORM = 0.0
      DO 70 I = 1,N
          ZNORM = ZNORM + DABS(TREB(I))
70  CONTINUE
C
C   ESTIMACION DEL NUMERO DE CONDICION
C
      COND = ANORM*ZNORM/YNORM
      IF (COND .LT. 1.0) COND = 1.0
      RETURN
C
C
      80 COND = 1.0

```

```

      IF (A(1,1) .NE. 0.0) RETURN
C
C      A ES MATRIZ SINGULAR
C
 90 COND = 1.0D+32
      RETURN
      END

SUBROUTINE SOLVE(NDIM, N, A, B, IPVT)
C
      INTEGER NDIM, N, IPVT(N)
      REAL*8 A(NDIM,N),B(N)
C
C      CALCULA LA SOLUCION DEL SISTEMA LINEAL A*X = B
C
C      NO SE PUEDE UTILIZAR SI DECOMP HA CONSIDERADO A SINGULAR

C      ENTRADA
C      NDIM = NUMERO DE ECUACIONES DEL SISTEMA QUE CONTIENE A
C      N = ORDEN DE LA MATRIZ A
C      A = MATRIZ TRIANGULARIZADA POR DECOMP
C      B = VECTOR DE TERMINOS INDEPENDIENTES
C      IPVT = VECTOR DE PIVOTAMIENTOS OBTENIDO POR DECOMP

C      SALIDA
C      B = VECTOR SOLUCION: X
C
      INTEGER KB, KM1, NM1, KP1, I, K, M
      REAL*8 T
C
C      ELIMINACION HACIA DELANTE
C
      IF (N .EQ. 1) GO TO 50
      NM1 = N-1
      DO 20 K = 1, NM1
         KP1 = K+1
         M = IPVT(K)
         T = B(M)
         B(M) = B(K)
         B(K) = T
         DO 10 I = KP1, N
            B(I) = B(I) + A(I,K)*T
 10    CONTINUE
 20    CONTINUE
C
C      SUBSTITUCION HACIA ATRAS
C
      DO 40 KB = 1,NM1
         KM1 = N-KB
         K = KM1+1
         B(K) = B(K)/A(K,K)
         T = -B(K)
         DO 30 I = 1, KM1
            B(I) = B(I) + A(I,K)*T
 30    CONTINUE
 40    CONTINUE

```

```

30      CONTINUE
40      CONTINUE
50      B(1) = B(1)/A(1,1)
      RETURN
      END

```

El número de condición del sistema es: `COND= 0.25175D+03` y la solución viene dada por:

$$\begin{cases} x(1) = 70.0 \text{ V} \\ x(2) = 52.0 \text{ V} \\ x(3) = 40.0 \text{ V} \\ x(4) = 31.0 \text{ V} \\ x(5) = 22.0 \text{ V} \\ x(6) = 10.0 \text{ V} \end{cases}$$

4.6.2 Enunciados

1. Resolver el sistema

$$\begin{pmatrix} 1.00 & 0.70 & 0.49 & 0.343 \\ 1.00 & 0.80 & 0.64 & 0.512 \\ 1.00 & 0.90 & 0.81 & 0.729 \\ 1.00 & 1.10 & 1.21 & 1.331 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} \operatorname{erf}(0.7) \\ \operatorname{erf}(0.8) \\ \operatorname{erf}(0.9) \\ \operatorname{erf}(1.1) \end{pmatrix}$$

mediante `DECOMP` y `SOLVE`; calcular una estimación del número de condición de la matriz y la solución $(x_1, x_2, x_3, x_4)^T$. Determinar la suma $x_1 + x_2 + x_3 + x_4$ y compararla con $\operatorname{erf}(1.0)$. ¿Por qué son valores próximos?

2. Resolver el sistema

$$\begin{pmatrix} 1.00 & 0.80 & 0.64 & 0.512 \\ 1.00 & 0.90 & 0.81 & 0.729 \\ 1.00 & 1.10 & 1.21 & 1.331 \\ 1.00 & 1.20 & 1.44 & 1.728 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} J_0(0.8) \\ J_0(0.9) \\ J_0(1.1) \\ J_0(1.2) \end{pmatrix}$$

mediante `DECOMP` y `SOLVE`; calcular una estimación del número de condición de la matriz y la solución $(x_1, x_2, x_3, x_4)^T$. Determinar la suma $x_1 + x_2 + x_3 + x_4$ y compararla con $J_0(1.0)$. ¿Por qué son valores tan próximos?

3. Resolver el sistema

$$\begin{pmatrix} 1.00 & 0.30 & 0.09 & 0.027 \\ 1.00 & 0.40 & 0.16 & 0.064 \\ 1.00 & 0.60 & 0.36 & 0.216 \\ 1.00 & 0.70 & 0.49 & 0.343 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} K(0.3) \\ K(0.4) \\ K(0.6) \\ K(0.7) \end{pmatrix}$$

donde K es la función elíptica

$$K(x) = \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - x \sin^2 \phi}}$$

mediante **DECOMP** y **SOLVE**; calcular una estimación del número de condición de la matriz y la solución $(x_1, x_2, x_3, x_4)^T$. Determinar la suma $x_1 + \frac{x_2}{2} + \frac{x_3}{4} + \frac{x_4}{8}$ y compararla con $K(0.5)$. ¿Por qué son valores tan próximos?

4. Escribir una rutina que calcule la matriz inversa de A :

SUBROUTINE INV (NDIM, N, A, X, COND, IPVT, WORK)

donde se tiene que resolver N sistemas lineales $A \mathbf{x}_j = \mathbf{e}_j$, con \mathbf{e}_j la j -ésima columna de la matriz identidad y la inversa de A es X con columnas \mathbf{x}_j . La subrutina ha de llamar una vez a **DECOMP** y N veces a **SOLVE** (una para cada columna de X).

Para comprobar si la subrutina funciona, tomad un juego de pruebas con matrices diferentes de órdenes diversos. Aplicar el método dos veces: primero, invirtiendo A , que da X y después, invirtiendo X , que da Z ; entonces, una posible estimación del método es $\|Z - A\|$.

5. En la tabla escrita a continuación se presentan los índices de refracción, n , para longitudes de ondas diferentes, λ , del cristal de borosilicato:

λ	6563	6439	5890	5338	5086	4861	4340	3988
n	1.50883	1.50917	1.51124	1.51386	1.51534	1.51690	1.52136	1.52546

Utilizando los valores correspondientes a la segunda, cuarta y séptima columna de la tabla, determinar las constantes A , B y C de la ecuación de Cauchy

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4}.$$

Comprobar la precisión obtenida para los valores restantes y aplicad el método de resolución para sistemas sobredeterminados para dar otra solución. Hacer los cálculos mediante las rutinas **DECOMP** y **SOLVE**.

6. Resolver el sistema $A \mathbf{x} = \mathbf{b}$, donde A es la matriz de Hilbert $n \times n$ con elementos

$$a_{ij} = \frac{1}{i+j-1} \quad \text{y} \quad b_i = \frac{1}{i} + \frac{1}{i+1} + \cdots + \frac{1}{i+n-1}$$

para $n = 3 \div 10$

- a) por eliminación gaussiana sin pivotamiento
- b) por medio de **DECOMP** y **SOLVE**
- c) por el método de Gauss-Seidel

Comentar las ventajas y los inconvenientes de los tres métodos y la evolución de los resultados cuando n va tomando valores más grandes.

7. En la interpolación por splines cúbicas aparecen sistemas lineales que tienen matrices del tipo siguiente

$$\begin{pmatrix} -1 & 1 & & & & & 0 \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & 4 & 1 & \\ 0 & & & & 1 & 4 & 1 \\ & & & & & 1 & -1 \end{pmatrix}$$

- a) ¿Cómo cambia el número de condición de esta matriz cuando el orden crece? ¿Qué propiedades especiales tiene la matriz resultado de DECOMP?
- b) ¿Cómo se podría simplificar el método de eliminación gaussiana en este caso? ¿Cómo resolvérais un sistema lineal con esta matriz para órdenes elevados? Construir un programa que permita resolver el sistema lineal $A \mathbf{x} = \mathbf{b}$ para \mathbf{b} dado, de orden elevado.
8. Se quiere construir una matriz S tal que la matriz dada A se convierta en tridiagonal: $T = S^{-1}AS$; para ello se calcula una sucesión de vectores $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ (columnas de S) biortogonales a otra sucesión $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$ ($\mathbf{y}_j^T \cdot \mathbf{x}_i = 0$, si $i \neq j$) que vienen definidas por

$$\begin{aligned}\mathbf{x}_{k+1} &= A \mathbf{x}_k - b_k \mathbf{x}_k - c_{k-1} \mathbf{x}_{k-1} & k = 1 \div n-1 \\ \mathbf{y}_{k+1} &= A^T \mathbf{y}_k - b_k \mathbf{y}_k - c_{k-1} \mathbf{y}_{k-1}\end{aligned}$$

con $\mathbf{x}_0 = \mathbf{y}_0$, \mathbf{x}_1 y \mathbf{y}_1 arbitrarios y

$$b_k = \frac{\mathbf{y}_k^T A \mathbf{x}_k}{\mathbf{y}_k^T \mathbf{x}_k} \quad c_{k-1} = \frac{\mathbf{y}_{k-1}^T A \mathbf{x}_k}{\mathbf{y}_{k-1}^T \mathbf{x}_{k-1}} = \frac{\mathbf{y}_k^T \mathbf{x}_k}{\mathbf{y}_{k-1}^T \mathbf{x}_{k-1}} \quad c_0 = 0 \quad k = 1 \div n-1$$

De este modo se tiene construido el método de Lanczos; parece claro que

$$\begin{aligned}A \mathbf{x}_1 &= \mathbf{x}_2 + b_1 \mathbf{x}_1 \\ A \mathbf{x}_k &= \mathbf{x}_{k+1} + b_k \mathbf{x}_k + c_{k-1} \mathbf{x}_{k-1} & k = 2 \div n-1 \\ A \mathbf{x}_n &= b_n \mathbf{x}_n + c_{n-1} \mathbf{x}_{n-1}\end{aligned}$$

ya que $\mathbf{x}_{n+1} = 0$. Por lo tanto,

$$A S = S \begin{pmatrix} b_1 & c_1 & & & 0 \\ 1 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & b_{n-1} & c_{n-1} \\ 0 & & & 1 & b_n \end{pmatrix} = S T$$

a) Construir una rutina que aplique el método.

b) Comprobadlo numéricamente con las matrices

$$A_1 = \begin{pmatrix} 2 & -2 & 3 \\ 1 & 1 & 1 \\ 1 & 3 & -1 \end{pmatrix} \quad \text{y} \quad A_2 = \begin{pmatrix} 1.0 & 1.0 & 0.5 \\ 1.0 & 1.0 & 0.25 \\ 0.5 & 0.25 & 2.0 \end{pmatrix}.$$

c) La anulación de $\mathbf{y}_k^T \mathbf{x}_k$ puede presentarse en cualquier situación, y de aquí puede aparecer inestabilidad numérica del método a pesar que la matriz A sea bien condicionada. Estudiar el método para la matriz

$$A_3 = \begin{pmatrix} 5 & 1 & -1 \\ -5 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad \text{con} \quad \mathbf{x}_1 = \begin{pmatrix} 0.6 \\ -1.4 \\ 0.3 \end{pmatrix} \quad \text{y} \quad \mathbf{y}_1 = \begin{pmatrix} 0.6 \\ 0.3 \\ -0.1 \end{pmatrix}.$$

5 Derivación e integración numérica

5.1 Introducción

Este capítulo trata del estudio de problemas y del cálculo efectivo de derivadas e integrales de funciones en que, ya sea por su dificultad en la definición, ya por la no necesidad de una precisión infinita, o ya debido a que se tiene una tabla de valores, es más conveniente realizar el estudio numérico.

Casi todos los apartados se basan en la interpolación polinómica (capítulo 2) como medio de obtención de fórmulas adecuadas. Se introduce la extrapolación de Richardson como una técnica utilizable tanto en derivación como en integración cuando se exige que una cantidad h (incremento de la variable o distancia entre puntos consecutivos en el sumatorio que approxima la integral) tienda a cero. Finalmente, empleando los polinomios de aproximación, se define la integración gaussiana.

5.2 Derivación interpolatoria

Sea una función $f \in C^{n+1}[a, b]$ que está evaluada en $n + 1$ puntos, $f(x_i) = y_i$, con $x_i \in [a, b]$, $i = 0 \div n$, y el polinomio interpolador de grado máximo n , $P_n(x)$: $f(x) = P_n(x) + E(x)$ donde, como se ha visto en el capítulo 2,

$$P_n(x) = \sum_{i=0}^n l_i(x) f(x_i) \quad \text{y} \quad E(x) = \frac{f^{n+1}(\xi(x))}{(n+1)!} \omega(x) \quad \text{donde } \omega(x) = \prod_{i=0}^n (x - x_i),$$

$\xi(x) \in I(x, x_0, \dots, x_n)$ que es el intervalo más pequeño que contiene los puntos x, x_0, \dots, x_n y $l_i(x)$ son los polinomios de Lagrange (ver 2.2.1).

Con la notación $f_i = f(x_i)$ y derivando $f(x)$, se tiene la derivada en función de las derivadas de los polinomios de Lagrange

$$f'(x) = \sum_{i=0}^n l'_i(x) \cdot f_i + E'(x)$$

donde

$$E'(x) = \omega(x) \frac{d}{dx} \left(\frac{f^{n+1)}(\xi(x))}{(n+1)!} \right) + \omega'(x) \frac{f^{n+1)}(\xi(x))}{(n+1)!}$$

Debido al desconocimiento de $\xi(x)$, el primer sumando no es fácil de evaluar, pero si se toma $x = x_k$, $k = 0 \div n$, en los nodos se tiene $\omega(x_k) = 0$ y el error en estos puntos se reduce a

$$E'(x_k) = \omega'(x_k) \frac{f^{n+1)}(\xi(x_k))}{(n+1)!}$$

En definitiva, se tiene el siguiente

Teorema 5.1. Si $f \in C^{n+1)}[a, b]$ y x_0, \dots, x_n son puntos diferentes del intervalo $[a, b]$; entonces,

$$f'(x_k) = \sum_{i=0}^n l'_i(x_k) \cdot f(x_i) + \omega'(x_k) \frac{f^{n+1)}(\xi_k)}{(n+1)!} \quad k = 0 \div n$$

donde $\xi_k \in I(x_0, \dots, x_n)$.

Ejemplo. Si se toman 3 puntos $x_k = x_0 + k h$, $k = 0, 1, 2$ y $f \in C^3[x_0, x_2]$, se puede calcular $f'(x_k) = l'_0(x_k) f_0 + l'_1(x_k) f_1 + l'_2(x_k) f_2 + E'(x_k)$ para $k = 0, 1, 2$. Los polinomios de Lagrange derivados son

$$l'_0(x) = \frac{1}{2h^2}(2x - 2x_0 - 3h), \quad l'_1(x) = \frac{-2}{h^2}(x - x_0 - h), \quad l'_2(x) = \frac{1}{2h^2}(2x - 2x_0 - h).$$

Entonces, la expresión de la derivada de f en los puntos x_0 , x_1 y x_2 es

$$\begin{aligned} f'(x_0) &= \frac{1}{2h}(-3f_0 + 4f_1 - f_2) + \frac{h^2}{3}f^{(3)}(\xi_0) \\ f'(x_1) &= \frac{1}{2h}(-f_0 + f_2) - \frac{h^2}{6}f^{(3)}(\xi_1) \\ f'(x_2) &= \frac{1}{2h}(f_0 - 4f_1 + 3f_2) + \frac{h^2}{3}f^{(3)}(\xi_2) \end{aligned}$$

La fórmula más utilizada es la segunda; si se escribe a en lugar de x_1 , se tiene

$$f'(a) = \frac{f(a+h) - f(a-h)}{2h} - \frac{h^2}{6}f^{(3)}(\xi) \quad (5.1)$$

con $\xi \in (a-h, a+h)$.

Ejercicios.

- Desarrollar, por la fórmula de Taylor, hasta grado 2 en potencias de h , $f(a+h)$ y $f(a-h)$. Comprobar que se tiene la fórmula anterior.

2. Si se evalúa la derivada mediante el cociente $[f(a+h) - f(a)]/h$ o por $[f(a) - f(a-h)]/h$, utilizando los desarrollos de Taylor, demostrar que el error es del orden de h .
3. Calcular de qué orden es el error si se approxima $f'(a)$ por

$$\frac{1}{12h} [f(a-2h) - 8f(a-h) + 8f(a+h) - f(a+2h)]$$

4. Calcular derivadas de orden superior considerando los cocientes siguientes:

$$\frac{1}{h^2} [f(a+h) - 2f(a) + f(a-h)]$$

$$\frac{1}{2h^3} [f(a+2h) - 2f(a+h) + 2f(a-h) - f(a-2h)]$$

Se analiza el error de redondeo y de truncamiento de $f'(a)$ si se toma la derivada definida por la fórmula 5.1. Se considera

$$\left. \begin{array}{l} f(a-h) = f_0 + e_0 \\ f(a+h) = f_2 + e_2 \end{array} \right\} \Rightarrow \text{Si se toma la aproximación } \bar{f}'(a) = \frac{1}{2h} (-f_0 + f_2)$$

y el valor exacto $f'(a) = \frac{1}{2h} (-f_0 - e_0 + f_2 + e_2) - \frac{h^2}{6} f^{(3)}(\xi)$; entonces

$$|f'(a) - \bar{f}'(a)| = \left| \frac{1}{2h}(e_2 - e_0) - \frac{h^2}{6} f^{(3)}(\xi) \right| \quad \text{con } \xi \in (a-h, a+h)$$

Si se tiene una cota superior, e , de $|e_2|$ y de $|e_0|$, se puede conseguir una cota del error de redondeo

$$\left| \frac{1}{2h}(e_2 - e_0) \right| \leq \frac{e}{h}$$

Por tanto, el error de redondeo en los datos crece cuando h decrece. Pero, por otro lado, el error de truncamiento disminuye ($O(h^2)$) cuando h decrece. Si se supone que ξ no depende de h , se puede minimizar el error. Si se denota por $C = \left| \frac{f^{(3)}(\xi)}{6} \right|$, entonces el error viene dado por $E(h) = \frac{e}{h} + C h^2$ con un mínimo para $h = \left(\frac{e}{2C} \right)^{\frac{1}{3}}$.

Ejercicio. Se sabe que

$$f'(a) = \frac{1}{12h} [f(a-2h) - 8f(a-h) + 8f(a+h) - f(a+2h)] + \frac{h^4}{30} f^{(5)}(\xi)$$

y se supone que la derivada de orden n de $f(x)$ satisface la condición $|f^{(n)}(x)| \leq 10^n$, $\forall x \in \mathbf{R}$, y que los valores tabulados de $f(x)$ presentan errores de redondeo acotados por $5 \cdot 10^{-5}$. Deducir el valor de h que es necesario tomar para que la derivada de f en a presente el mínimo error posible.

5.3 Extrapolación de Richardson

En muchos cálculos se desea calcular el valor límite de cierta magnitud cuando la longitud del paso tiende a cero. A tal fin se utiliza un método de extrapolación, llamado de Richardson, si se conoce el comportamiento de la función $F(h)$ en un entorno de $h = 0$. Este es el caso cuando se quiere calcular la derivada de una función en un punto; por ejemplo, si se tiene $F(h) = \frac{f(a+h) - f(a-h)}{2h}$, para diferentes valores de h , y se quiere calcular $\lim_{h \rightarrow 0} F(h) = f'(a)$.

Si el error de truncamiento cuando $h \rightarrow 0$ es conocido, es posible calcular buenas aproximaciones de $F(0)$, evaluando $F(h)$ para diferentes valores de h y extrapolando:

Teorema 5.2. Sea $F(h) = F(0) + a_1 h^{p_1} + a_2 h^{p_2} + \dots$, donde $p_1 < p_2 < \dots$ y se define

$$F_1(h) = F(h) \quad F_{k+1}(h) = F_k(h) + \frac{F_k(h) - F_k(qh)}{q^{p_k} - 1} \quad (q > 1)$$

entonces $F_n(h)$ admite un desarrollo de la forma

$$F_n(h) = F(0) + a_n^{(n)} h^{p_n} + a_{n+1}^{(n)} h^{p_{n+1}} + \dots$$

Demostración: Por inducción; si $n = 1$, es claro. Se supone cierto para $n = k$, y se demuestra para $n = k + 1$:

$$F_{k+1}(h) = F_k(h) + \frac{F_k(h) - F_k(qh)}{q^{p_k} - 1} \quad (q > 1)$$

$F_{k+1}(h)$ tiene las mismas potencias de h que $F_k(h)$, pero el coeficiente de h^{p_k} es

$$a_k^{(k)} + \frac{a_k^{(k)} - a_k^{(k)} q^{p_k}}{q^{p_k} - 1} = a_k^{(k)} - a_k^{(k)} = 0. \square$$

Se utiliza la siguiente tabla de extrapolación:

h	F_1	$\frac{\Delta}{q^{p_1}-1}$	F_2	$\frac{\Delta}{q^{p_2}-1}$	F_3	$\frac{\Delta}{q^{p_3}-1}$	F_4
$q^3 h$	$F_1(q^3 h)$						
$q^2 h$	$F_1(q^2 h)$		$F_2(q^2 h)$				
qh	$F_1(qh)$		$F_2(qh)$		$F_3(qh)$		
h	$F_1(h)$		$F_2(h)$		$F_3(h)$		$F_4(h)$

Las entradas son calculadas fila a fila y la extrapolación se obtiene en la misma columna tomando dos valores consecutivos. El esquema se para cuando, para h suficientemente pequeño, la diferencia entre dos valores consecutivos de la misma columna da una cota superior del error:

Si $F(0) = F(h) + ch^p + O(h^m)$, $m > p$, y se toma qh en lugar de h con $q > 1$, se tiene

$F(0) = F(qh) + c(qh)^p + O(q^m h^m)$. Multiplicando la primera por q^p y restándole la segunda, se puede despejar $F(0)$ que da

$$F(0) = \frac{q^p F(h) - F(qh)}{q^p - 1} + \frac{q^p O(h^m) - O(q^m h^m)}{q^p - 1}$$

donde el segundo sumando es de orden h^m ($O(h^m)$).

Ejemplo. Se quiere aproximar $f'(1)$ a partir de la tabla de $f(x) = e^x$:

x	0.00	0.50	0.75	1.00	1.25	1.50	2.00
e^x	1.0000	1.6487	2.1170	2.7183	3.4903	4.4817	7.3891

Tomando $q = 2$ y considerando

$$F(h) = \frac{f(1+h) - f(1-h)}{2h} = F(0) + ch^2 + dh^4 + eh^6 + \dots$$

se tiene la tabla de extrapolación siguiente

h	F_1	$\Delta/3$	F_2	$\Delta/15$	F_3
1	3.1946	-0.1205			
0.5	2.8330		2.7125		
		-0.0288		0.0004	
0.25	2.7466		2.7178		<u>2.7182</u>

El error es menor que $|F_2(0.25) - F_2(0.50)| = 4 \cdot 10^{-4}$. En realidad, el error de truncamiento es del orden de $0.25^4 \approx 3.91 \cdot 10^{-3}$ para $F_2(0) = 2.7178$ y del orden de $0.25^6 \approx 2.44 \cdot 10^{-4}$ para $F_3(0) = 2.7182$.

5.4 Integración numérica

El cálculo de las integrales definidas de una función real, $I\{f\} = \int_a^b f(x) dx$, es un problema clásico. Este problema tiene interés numérico en el momento que se desconoce una primitiva de f o el cálculo es muy costoso. Entonces, la idea general consiste en discretizar convenientemente y calcular una suma finita correspondiente a una cierta partición del intervalo $[a, b]$:

$I_n\{f\} = \sum_{j=0}^n \beta_j f(x_j)$, donde x_0, x_1, \dots, x_n se llaman nodos y los $\beta_0, \beta_1, \dots, \beta_n$ coeficientes o pesos de la integración.

El problema básico consiste en elegir los nodos y los coeficientes de manera que el error de integración $E_n\{f\} = I\{f\} - I_n\{f\}$ sea ‘pequeño’.

Una manera de medir el error es mediante el **grado de precisión**, que es el mayor número natural $m \in \mathbb{N}$ tal que $E_n\{x^k\} = 0$, $k = 1 \div m$, pero $E_n\{x^{k+1}\} \neq 0$. De este modo, si una

fórmula tiene grado de precisión m , todos los polinomios de grado máximo m son integrados exactamente.

En el caso de abcisas equiespaciadas, $h = \frac{b-a}{n}$, también se dice que el error es de **orden n** cuando $E_n\{f\} = O(h^n)$.

5.4.1 Fórmulas de Newton–Côtes

Sea una partición equiespaciada $a = x_0 < x_1 < \dots < x_{n-1} < x_n = b$ y el polinomio interpolador $P_n(x) = \sum_{j=0}^n l_j(x) f(x_j)$; entonces, se considera la aproximación

$$\int_a^b f(x) dx \approx \int_a^b \sum_{j=0}^n l_j(x) f(x_j) dx = \sum_{j=0}^n \left(\int_a^b l_j(x) dx \right) f(x_j)$$

Si se hace el cambio de variables $x = a + th$, se tiene

$$l_j(x) = \varphi_j(t) = \prod_{\substack{k=0 \\ k \neq j}}^n \frac{t-k}{j-k}$$

Finalmente, se obtiene

$$\int_a^b f(x) dx \approx h \cdot \sum_{j=0}^n \left(\int_0^1 \varphi_j(t) dt \right) f(x_j) = h \cdot \sum_{j=0}^n \alpha_j f(x_j)$$

donde los números de Côtes α_j vienen definidos por $\alpha_j = \int_0^1 \varphi_j(t) dt$ y dependen únicamente del número de puntos elegidos y no dependen ni de la función a integrar ni del intervalo de integración.

Ejemplos.

1. Para $n = 1$, $\alpha_0 = \int_0^1 \frac{t-1}{0-1} dt = \frac{1}{2}$ y $\alpha_1 = \int_0^1 \frac{t-0}{1-0} dt = \frac{1}{2}$. Entonces, si se toma $h = b - a$, se tiene la fórmula de los trapecios:

$$\int_a^b f(x) dx \approx h \cdot \frac{f(a) + f(b)}{2}$$

2. Si $n = 2$, $\alpha_0 = \int_0^2 \frac{(t-1)(t-2)}{(0-1)(0-2)} dt = \frac{1}{3}$, $\alpha_1 = \int_0^2 \frac{(t-0)(t-2)}{(1-0)(1-2)} dt = \frac{4}{3}$ y $\alpha_2 = \int_0^2 \frac{(t-0)(t-1)}{(2-0)(2-1)} dt = \frac{1}{3}$. Tomando $h = \frac{b-a}{2}$, se tiene la fórmula de Simpson:

$$\int_a^b f(x) dx \approx \frac{h}{3} \cdot \left[f(a) + f\left(\frac{a+b}{2}\right) + f(b) \right]$$

Nótese que los coeficientes α_j cumplen $\sum_{j=0}^n \alpha_j = n$ y son números racionales. Si s es el denominador común de los $\alpha_0, \alpha_1, \dots, \alpha_n$, entonces $\sigma_j = s\alpha_j \in \mathbf{Z}$, $j = 0 \div n$ i, denotando $f(x_j)$ por y_j , se tiene

$$\int_a^b P_n(x) dx = h \cdot \sum_{j=0}^n \alpha_j y_j = \frac{b-a}{n \cdot s} \sum_{j=0}^n \sigma_j y_j$$

que están tabulados para diferentes valores de n (tabla 5.1). Más información para fórmulas de Newton-Côtes de orden superior se puede encontrar en [Abr72].

n	σ_i						ns	Error	Nombre
1	1	1					2	$h^3 1/12 f^2(c)$	Trapecios
2	1	4	1				6	$h^5 1/90 f^4(c)$	Simpson
3	1	3	3	1			8	$h^5 3/80 f^4(c)$	$3/8$
4	7	32	12	32	7		90	$h^7 8/945 f^6(c)$	Milne
5	19	75	50	50	75	19	288	$h^7 275/12096 f^6(c)$	
6	41	216	27	272	27	216	41	$h^9 9/1400 f^8(c)$	Weddle

Tabla 5.1

Si se impone que el error presente la forma $K f^{(m)}(\xi)$, es decir,

$$E_n\{f\} = \int_a^b f(x) dx - \sum_{j=0}^n \beta_j f(x_j) = K f^{(m)}(\xi)$$

y se aplica a las funciones potenciales, $f(x) = x^m$, da cero para $m \leq n$. Si $m > n$ y se va aumentando m hasta que $E_n(x^{m-1}) = 0$ y $E_n(x^m) \neq 0$, entonces

$$\int_a^b x^m dx = \sum_{j=0}^n \beta_j f(x_j) + K m!$$

y se obtiene K : $K = \frac{1}{m!} \left[\int_a^b x^m dx - \sum_{j=0}^n \beta_j f(x_j) \right]$, tal que $E_n\{f\} = K f^{(m)}(\xi)$.

Ejemplos.

- Si se toma $\int_0^1 F(t) dt = \frac{1}{2} [F(0) + F(1)] + E_1\{F\}$, se tiene $E_1\{1\} = E_1\{t\} = 0$ y, si se considera $F(t) = t^2$, entonces

$$\frac{1}{3} = \frac{1}{2} (0^2 + 1^2) + 2! K \implies K = \frac{-1}{12}$$

Si se hace el cambio en el intervalo $[a, b]$, teniendo en cuenta que $F(t) = f(a + th)$, con $h = b - a$ y que $F'' = h^2 f''(a + th)$, se obtiene

$$\int_a^b f(x) dx = \frac{h}{2} [f(a) + f(b)] - \frac{1}{12} h^2 f''(c)$$

2. Si se toma $\int_0^2 F(t) dt = \frac{1}{3} [F(0) + 4 F(1) + F(2)] + E_2\{F\}$, se obtiene

$$E_1\{1\} = E_1\{t\} = E_1\{t^2\} = E_1\{t^3\} = 0 \text{ y, si se considera } F(t) = t^4, \text{ entonces } K = \frac{-1}{90}$$

y $E_n\{F\} = \frac{-1}{90} F^{(4)}(\xi)$. En el intervalo $[a, b]$ se tiene

$$\begin{aligned} \int_a^b f(x) dx &= h \int_0^2 F(t) dt = \frac{h}{3} [F(0) + 4 F(1) + F(2)] - \frac{h F^{(4)}(\xi)}{90} = \\ &= \frac{h}{3} \left[f(a) + 4 f\left(\frac{a+b}{2}\right) + f(b) \right] - \frac{1}{90} h^5 f^{(4)}(c) \quad c \in (a, b) \end{aligned}$$

Por tanto, el error se puede expresar por $k h^{p+1} f^{(p)}(c)$ donde $c \in [a, b]$ y p y k dependen únicamente de n y no de la función. También se comprueba que, si se toma $n + 1$ nodos, se tiene grado de precisión n , si $n + 1$ es par y grado de precisión $n + 1$ si $n + 1$ es impar.

Fórmulas compuestas

En general, las fórmulas de Newton–Côtes no se aplican directamente sobre todo un intervalo, sino que previamente se define una partición para así poder repetir el método en cada subintervalo.

Si se divide el intervalo $[a, b]$ en n subintervalos:

$a = x_0 < x_1 < \dots < x_{n-1} < x_n = b$ donde $x_{i+1} - x_i = h$, $i = 0 \div n - 1$ y se aplica una fórmula de Newton–Côtes a cada $[x_i, x_{i+1}]$, se tiene

$$\int_a^b f(x) dx = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x) dx = \sum_{i=0}^{n-1} I_i$$

Si se aplica la regla de los trapecios a $[x_i, x_{i+1}]$, se tiene $T_i(h) = \frac{h}{2} [f(x_i) + f(x_{i+1})]$ con $x_{i+1} - x_i = h = \frac{b-a}{n}$, y, por tanto,

$$T(h) = \sum_{i=0}^{n-1} T_i(h) = h \cdot \left[\frac{f(a)}{2} + f(a+h) + \dots + f(a+(n-1)h) + \frac{f(b)}{2} \right]$$

El error en cada intervalo es $T_i(h) - I_i = \frac{h^3}{12} f^{(2)}(c_i)$ con $c_i \in (x_i, x_{i+1})$. Entonces,

$$T(h) - I = \frac{h^3}{12} \sum_{i=0}^{n-1} f^{(2)}(c_i) = \frac{h^2}{12} \frac{b-a}{n} \sum_{i=0}^{n-1} f^{(2)}(c_i)$$

Si $f \in C^2[a, b]$, existe un $c \in [\min_i c_i, \max_i c_i] \subset [a, b]$ tal que $f^{(2)}(c) = \frac{1}{n} \sum_{i=0}^{n-1} f^{(2)}(c_i)$ y, por tanto,

$$T(h) - I = \frac{b-a}{12} h^2 f^{(2)}(c), \text{ con } c \in (a, b).$$

De forma similar, si se toma partición con $n = 2m$, es decir, $a = x_0 < x_1 < \dots < x_{2m-1} < x_{2m} = b$, se obtiene la regla de Simpson compuesta

$$\begin{aligned} S(h) = \frac{h}{3} [(f(a) + f(b)) + 4(f(a+h) + f(a+3h) + \dots + f(x_{2m}-h)) + \\ + 2(f(a+2h) + f(a+4h) + \dots + f(x_{2m}-2h))], \text{ donde } h = \frac{b-a}{2m}. \end{aligned}$$

El error en cada intervalo $[x_{2k-2}, x_{2k}]$, $k = 1 \div m$, es $\frac{h^5}{90} f^{(4)}(c_k)$ donde $c_k \in (x_{2k-2}, x_{2k})$. Entonces,

$$S(h) - I = \frac{h^5}{90} \sum_{k=1}^m f^{(4)}(c_k) = \frac{h^4}{90} \frac{b-a}{2m} \sum_{k=1}^m f^{(4)}(c_k)$$

Si $f \in C^4[a, b]$, existe un $c \in [\min_i c_i, \max_i c_i] \subset [a, b]$ tal que $f^{(4)}(c) = \frac{1}{m} \sum_{k=1}^m f^{(2)}(c_i)$ y, por tanto, $S(h) - I = \frac{b-a}{180} h^4 f^{(4)}(c)$, con $c \in (a, b)$.

Ejercicios.

- Deducir la regla de los 3/8 compuesta y demostrar que el error viene dado por

$$\frac{b-a}{80} h^4 f^{(4)}(c)$$

- Deducir la regla de Milne compuesta y demostrar que el error viene dado por

$$\frac{2(b-a)}{945} h^6 f^{(6)}(c)$$

5.4.2 Método de Romberg

Se demuestra, en primer lugar, el siguiente desarrollo asintótico para la regla de los trapecios $T(h) = \int_a^b f(x) dx + a_1 h^2 + a_2 h^4 + a_3 h^6 + \dots$

Para ello se desarrolla la función $f(x)$ en un entorno del punto medio \bar{x}_i del intervalo $[x_i, x_{i+1}]$:

$$f(x) = f(\bar{x}_i) + (x - \bar{x}_i) f'(\bar{x}_i) + \frac{1}{2}(x - \bar{x}_i)^2 f''(\bar{x}_i) + \frac{1}{6}(x - \bar{x}_i)^3 f'''(\bar{x}_i) + \frac{1}{24}(x - \bar{x}_i)^4 f^{(4)}(\bar{x}_i) + \dots$$

y se integra término a término

$$I_i = \int_{x_i}^{x_{i+1}} f(x) dx = h_i f'(\bar{x}_i) + \frac{1}{24} h_i^3 f^{(3)}(\bar{x}_i) + \frac{1}{1920} h_i^5 f^{(5)}(\bar{x}_i) + \dots$$

donde $h_i = x_{i+1} - x_i$.

Si se evalúa el desarrollo anterior de $f(x)$ en los puntos $x = x_i$ y $x = x_{i+1}$, se obtiene al calcular la semisuma

$$\frac{f(x_i) + f(x_{i+1})}{2} = f(\bar{x}_i) + \frac{1}{8} h_i^2 f''(\bar{x}_i) + \frac{1}{384} h_i^4 f^{(4)}(\bar{x}_i) + \dots$$

despejando $f(\bar{x}_i)$, se obtiene

$$f(\bar{x}_i) = \frac{f(x_i) + f(x_{i+1})}{2} - \frac{1}{8} h_i^2 f''(\bar{x}_i) - \frac{1}{384} h_i^4 f^{(4)}(\bar{x}_i) - \dots$$

y se puede substituir en la integral

$$\int_{x_i}^{x_{i+1}} f(x) dx = h_i \cdot \frac{f(x_i) + f(x_{i+1})}{2} - \frac{1}{12} h_i^3 f''(\bar{x}_i) - \frac{1}{480} h_i^5 f^{(4)}(\bar{x}_i) - \dots$$

es decir, $I_i = T_i(h) - \frac{1}{12} h_i^3 f''(\bar{x}_i) - \frac{1}{480} h_i^5 f^{(4)}(\bar{x}_i) - \dots$

Si se considera todo el intervalo $[a, b]$, se tiene

$$I = \int_a^b f(x) dx = T(h) - \frac{b-a}{12} f''(c_2) h^2 - \frac{b-a}{480} f^{(4)}(c_4) h^4 - \dots$$

En definitiva, el método de Romberg consiste en aplicar extrapolación repetida de Richardson a la fórmula

$$T(h) = a_0 + a_1 h^2 + a_2 h^4 + a_3 h^6 + \dots$$

con $a_0 = \int_a^b f(x) dx$. Se calculará la fórmula de los trapecios, $T(h)$, para diferentes valores de h :

$$T(h_0) \quad T(h_1) \quad T(h_2) \quad \dots \quad T(h_k) \quad \text{con}$$

$$h_0 = h \quad h_1 = q^{-1}h \quad h_2 = q^{-2}h \quad \dots \quad h_k = q^{-k}h \quad \text{donde } q = 2,$$

y se extrapolará teniendo en cuenta que $p_1 = 2$, $p_2 = 4$, $p_3 = 6$, etc. Entonces, se tiene el esquema siguiente:

$$T_1(h) = T(h) \quad \text{i} \quad T_{k+1}(h) = T_k(h) + \frac{T_k(h) - T_k(2h)}{4^k - 1}$$

Ejemplo. Si se tiene $\int_0^{0.8} \frac{\sin x}{x} dx$, se puede considerar la tabla siguiente

h	T_1	$T_2 = T_1 + \frac{\Delta}{3}$	$T_3 = T_2 + \frac{\Delta}{15}$
0.8	0.758680		
0.4	0.768760	0.772120	
0.2	0.771262	0.772096	0.772095
0.1	0.771887	0.772095	0.772095

Por lo tanto, $\int_0^{0.8} \frac{\sin x}{x} dx = 0.772095 \pm 0.000005$.

5.4.3 Elección del paso de integración

Si se quiere un determinado error de integración, generalmente no es necesario integrar todo el intervalo con el mismo paso, sino que es suficiente integrar con el máximo posible, de manera que se obtenga la integral con una cota del error deseado. Para ello se utilizan dos fórmulas de integración de órdenes diferentes, y se obtiene así una cierta cota del error producido en función de la diferencia de los resultados de las dos fórmulas.

Sea P_i el resultado de integrar f en el intervalo $[x_i, x_{i+1}]$ mediante una fórmula con grado de precisión $m-1$:

$$I_i - P_i = k h_i^{m+1} f^{(m)}(\bar{x}_i) + \dots \quad \text{donde } \bar{x}_i = \frac{x_i + x_{i+1}}{2}.$$

Sea Q_i el resultado de integrar f en los intervalos $[x_i, \bar{x}_i]$ y $[\bar{x}_i, x_{i+1}]$ aplicando la fórmula anterior con paso $h_i/2$:

$$I_i - Q_i = k \left(\frac{h_i}{2} \right)^{m+1} \left[f^{(m)}(\bar{x}_i + \frac{h_i}{4}) + f^{(m)}(\bar{x}_i + \frac{3h_i}{4}) \right] + \dots$$

Entonces $I_i - Q_i = k \left(\frac{h_i}{2} \right)^{m+1} (2 f^{(m)}(\bar{x}_i) + \dots) = \frac{1}{2^m} (I_i - P_i) + \dots$; es decir, el error decrece, aproximadamente, en un factor de 2^m al tomar paso mitad. Si se menoscopia los términos de orden superior y se despeja $I_i = \frac{2^m}{2^m - 1} (Q_i - \frac{1}{2^m} P_i)$, al restarlo de Q_i , se obtiene

$$Q_i - I_i = \frac{1}{2^m - 1} (P_i - Q_i)$$

es decir, el error de Q_i es $(2^m - 1)^{-1}$ veces la diferencia entre las dos fórmulas.

Por tanto, si se divide el paso hasta que se cumpla la condición de bisección

$$\frac{1}{2^m - 1} |P_i - Q_i| \leq \frac{h_i}{b - a} \varepsilon$$

donde ε es el error deseado, se obtendrá una técnica de integración con cálculo automático del paso por medio de la bisección del intervalo cuando es necesario, ya que

$$|Q - I| = \left| \sum_{i=0}^{n-1} (Q_i - I_i) \right| \leq \sum_{i=0}^{n-1} |(Q_i - I_i)| \leq \sum_{i=0}^{n-1} \frac{1}{2^m - 1} |P_i - Q_i| \leq \varepsilon$$

5.4.4 Integrales impropias

Para calcular numéricamente estas integrales de las cuales previamente se habrá estudiado la existencia, es necesario utilizar diferentes técnicas según sea la función.

★ Eliminación de singularidades:

1. Cambio de variable. $\int_0^1 x^{-1/2} e^x dx = 2 \int_0^1 e^{t^2} dt$, donde se ha hecho el cambio $x = t^2$, y ya se puede integrar numéricamente.
2. Integración por partes. $\int_0^1 x^{-1/2} e^x dx = 2e - 2 \int_0^1 x^{1/2} e^x dx$.
3. Desarrollo por series. La integral $I = \int_{0.0001}^1 x^{-3} e^x dx$ no tiene ninguna singularidad, pero en uno de los extremos del intervalo toma valores muy grandes. Entonces puede ser conveniente desarrollar la función e^x y separar la integral en

$$I = \int_{0.0001}^1 x^{-3} \left(1 + x + \frac{x^2}{2} \right) dx + \int_{0.0001}^1 x^{-3} \left(e^x - 1 - x - \frac{x^2}{2} \right) dx$$

donde la primera integral se puede calcular analíticamente y la segunda numéricamente, pero se ha reducido los valores que toma la función y sus derivadas.

★ Intervalo infinito de integración:

1. Se quiere calcular $\int_{-\infty}^{+\infty} f$ pero se considera $\int_{R_1}^{R_2} f$ de manera que el valor de la integral sea menor que un ε prefijado por $x \leq R_1$ y $x \geq R_2$. Si se tiene $\int_{-\infty}^{+\infty} e^{-x^2} dx$, se calcula $\int_{-4}^{+4} e^{-x^2} dx$, ya que $e^{-x^2} < 0.5 \cdot 10^{-6}$ si $x = \pm 4$. El error producido es

$$2 \int_4^{+\infty} e^{-x^2} dx = \int_{16}^{\infty} e^{-t} t^{-1/2} dt < 16^{-1/2} \int_{16}^{\infty} e^{-t} dt = 1/4 e^{-16} < 10^{-7}.$$

2. Desarrollo en serie de potencias negativas. $\int_0^{\infty} (1+x^2)^{-4/3} dx =$
 $= \int_0^{\infty} x^{-8/3} (1+x^{-2})^{-4/3} dx =$
 $= \int_0^{\infty} x^{-8/3} \left(1 - \frac{4}{3}x^{-2} + \frac{14}{9}x^{-4} - \frac{140}{81}x^{-6} + \dots \right) dx = \int_0^R S + \int_R^{\infty} S$

donde S representa la serie y la primera integral se puede calcular numéricamente después de acotar la segunda con un R conveniente, ya que

$$\int_R^\infty S = R^{-5/3} \left(\frac{3}{5} - \frac{4}{11}R^{-2} + \frac{14}{61}R^{-4} + \dots \right)$$

3. Cambio de variable. $\int_0^\infty (1+x^2)^{-4/3} dx = - \int_1^0 \left\{ 1 + \left(\frac{1-t}{t} \right)^2 \right\}^{-4/3} t^{-2} dt =$
 $= \int_0^1 \{t^2 + (1-t)^2\}^{-4/3} t^{2/3} dt$, donde el cambio ha sido $t = \frac{1}{1+x}$. Este integrando presenta una singularidad de la primera derivada en $t = 0$, hecho que se resuelve numéricamente después del cambio $t = u^3$.

5.5 Integración gaussiana

Si se quiere calcular $I\{f\} = \int_a^b f(x) dx$ mediante una fórmula del tipo

$$I_m\{f\} = \sum_{j=1}^m \beta_j f(x_j)$$

el problema básico consiste en elegir los nodos x_1, \dots, x_m y los coeficientes β_1, \dots, β_m de manera que el error de integración $E_m\{f\} = I\{f\} - I_m\{f\}$ sea ‘pequeño’.

Si $G(x)$ es el polinomio que interpola los puntos $(x_i, f(x_i))$, $i = 1 \div m$, se tiene

$$E_m\{f\} = \int_a^b (f(x) - G(x)) dx = \int_a^b \frac{f^{(m)}(c)}{m!} p_m(x) dx$$

donde $p_m(x) = (x - x_1)(x - x_2) \cdots (x - x_m)$ es el polinomio fundamental de interpolación. Si $f(x)$ es un polinomio de grado menor o igual que $m-1$, entonces $E_m\{f\} = 0$, ya que se utilizan m nodos y, por tanto, se tiene grado de precisión $m-1$, por lo menos. Una cota del error es

$$|E_m\{f\}| \leq \frac{1}{m!} \max_{x \in [a, b]} |f^{(m)}(x)| \int_a^b |p_m(x)| dx$$

Se desea calcular los puntos x_j tales que el error se anule cuando $f(x)$ sea un polinomio de grado $m+r$ con $r = 0 \div k$ y k tan grande como sea posible. Para determinar estos nodos se recuerda que, si se deriva m veces un polinomio de grado $m+r$, se tiene un polinomio de grado a lo sumo r . Entonces una condición necesaria y suficiente de que $E_m\{f\}$ se anule para todos los polinomios de grado $m+k$ es que

$$\int_a^b p_m(x) x^r dx = 0, \quad r = 0, 1, \dots, k \tag{5.2}$$

ya que se tiene

$$\int_a^b f(x) dx = \sum_{j=1}^m \beta_j f(x_j) + E_m\{f\}$$

y $E_m\{f\} = \int_a^b \frac{f^{(m)}(d)}{m!} p_m(x) dx.$

La condición 5.2 se puede interpretar como la condición de ortogonalidad en $[a, b]$ de $p_m(x)$ a todos los polinomios de grado como máximo k . En realidad, si se toma $p_m(x)$ como el m -ésimo polinomio ortogonal, entonces 5.2 se cumple para $k = m - 1$. En resumen, se tiene el siguiente

Teorema 5.3. La fórmula de integración 5.3 puede tener grado máximo de precisión igual a $2m - 1$. La condición necesaria y suficiente es que los m nodos, x_j , sean los ceros del polinomio ortogonal en $[a, b]$ de grado m : $p_m(x)$.

Este método se llama integración gaussiana y los coeficientes β_j de 5.3 quedan determinados una vez se hayan calculado los nodos x_j , $j = 1 \div m$:

Como $\int_a^b f(x) dx = \sum_{j=1}^m \beta_j f(x_j)$, si se toma $f(x) = \frac{p_m(x)}{x - x_j}$, se tiene

$$\int_a^b \frac{p_m(x)}{x - x_j} dx = 0 + \dots + \beta_j p'_m(x_j) + \dots + 0$$

de donde

$$\beta_j = \frac{1}{p'_m(x_j)} \int_a^b \frac{p_m(x)}{x - x_j} dx = \int_a^b l_j(x) dx, \quad j = 1 \div m$$

Teorema 5.4. Los coeficientes β_j , $j = 1 \div m$, son positivos.

Demostración: La fórmula de integración gaussiana con m nodos tiene grado de precisión $2m - 1$ y $E_m\{f\} = 0$ si f es un polinomio de grado $\leq 2m - 1$; en particular, si se toma el polinomio auxiliar de grado $2m - 2$ $q_j(x) = \frac{p_m^2(x)}{(x - x_j)^2}$, la siguiente integración es exacta:

$$\int_a^b q_j(x) dx = \sum_{k=0}^n \beta_k q_j(x_k)$$

pero
$$\begin{cases} q_j(x_k) &= 0 \text{ si } k \neq j \\ q_j(x_j) &= \prod_{\substack{i=0 \\ i \neq j}}^n (x_j - x_i)^2 (p'_m(x_j))^2 > 0 \end{cases}$$

$$\text{Entonces, } \beta_j = \frac{1}{q_j(x_j)} \int_a^b q_j(x) dx = \frac{1}{(p'_m(x_j))^2} \int_a^b \left(\frac{p_m(x)}{x - x_j} \right)^2 dx > 0. \square$$

Nota: El polinomio ortogonal $p_m(x) = (x - x_1)(x - x_2) \cdots (x - x_m)$ es mónico y, por lo tanto, si $Q_m(x) = a_m x^m + a_{m-1} x^{m-1} + \dots + a_1 x + a_0$ es el polinomio ortonormal, tienen los mismos ceros $p_m(x) = \frac{1}{a_m} Q_m(x)$ y se deduce que

$$\int_a^b (p_m(x))^2 dx = \frac{1}{a_m^2} \int_a^b (Q_m(x))^2 dx = \frac{1}{a_m^2}$$

5.5.1 Gauss-Legendre

Si se toma $[a, b] = [-1, 1]$, como ya se ha visto en el capítulo 3, los polinomios ortogonales son los de Legendre y se pueden normalizar del siguiente modo: $\sqrt{\frac{2m+1}{2}} P_m(x)$, por lo que $a_m = \frac{(2m)!}{2^m (m!)^2} \sqrt{\frac{2m+1}{2}}$. Se tiene, entonces,

$$\begin{aligned} P_0(x) &= 1 & P_3(x) &= \frac{1}{2}(5x^3 - 3x) \\ P_1(x) &= x & P_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3) \\ P_2(x) &= \frac{1}{2}(3x^2 - 1) & P_5(x) &= \frac{1}{8}(63x^5 - 70x^3 + 15x) \\ P_{n+1}(x) &= \frac{2n+1}{n+1} x P_n(x) - \frac{n}{n+1} P_{n-1}(x) \end{aligned}$$

Recordatorio: Si la integral tiene otros límites, se define un cambio lineal

$$\int_a^b f(x) dx = \frac{b-a}{2} \sum_{j=1}^m \beta_j \cdot f \left(\frac{b-a}{2} x_j + \frac{b+a}{2} \right) + E_m \{ f \}$$

Teorema 5.5. Si $f \in C^{2m} [a, b]$, y $\xi_i \neq x_j$, $i, j = 1 \div m$ con $\xi_i \in [a, b]$, donde x_j son las raíces del polinomio ortogonal $p_m(x)$ en $[a, b]$, entonces

$$E_m \{ f \} = \int_a^b \frac{f^{(2m)}(d)}{(2m)!} p_m(x) (x - \xi_1)(x - \xi_2) \cdots (x - \xi_m) dx$$

Demostración: Se interpola $f(x)$ en los puntos $x_1, x_2, \dots, x_m, \xi_1, \xi_2, \dots, \xi_m$: $f(x) = P_{2m-1}(x) + R_{2m-1}$ donde $P_{2m-1}(x)$ es el polinomio interpolador y $R_{2m-1}(x)$ es el error siguiente:

$$R_{2m-1}(x) = \frac{f^{(2m)}(d)}{(2m)!} \prod_{j=1}^m (x - x_j)(x - \xi_j) \quad d \in I(x, x_1, \dots, x_m, \xi_1, \dots, \xi_m)$$

Entonces, $\int_a^b f = \int_a^b P_{2m-1} + \int_a^b R_{2m-1} =$

$= \sum_{j=1}^m \beta_j P_{2m-1} + \sum_{j=1}^m \beta_j R_{2m-1} + E_m\{f\}$, y como el grado de precisión de la fórmula es $2m-1$, se tiene

$$\begin{cases} \int_a^b P_{2m-1}(x) dx &= \sum_{j=1}^m \beta_j P_{2m-1}(x_j) \\ R_{2m-1}(x_j) &= 0 \quad j = 1 \div m \end{cases}$$

En definitiva, $E_m\{f\} = \int_a^b R_{2m-1}(x) dx$. \square

Corolario 5.1. Si $f \in C^{2m}[a, b]$, entonces $E_m\{f\} = \frac{f^{(2m)}(\xi)}{(2m)!} \int_a^b (p_m(x))^2 dx$, donde $\xi \in (a, b)$.

Demostración: Si, en el teorema anterior, se hace que las ξ_i tiendan hacia las x_i , se obtiene el corolario. Ver también [Isa66]. \square

Si se considera el intervalo $[a, b] = [-1, 1]$, se tiene el polinomio de Legendre ‘mónico’ $p_m(x)$ que se relaciona con $P_m(x)$, polinomio de Legendre de grado m , por $p_m(x) = \frac{2^m (m!)^2}{(2m)!} P_m(x)$; teniendo en cuenta que $\langle P_m(x), P_m(x) \rangle = \frac{2}{2m+1}$ (ver el capítulo 3), se obtiene

$$E_m\{f\} = \frac{2}{(2m+1)!} \left[\frac{2^m (m!)^2}{(2m)!} \right]^2 f^{(2m)}(\xi) \quad \text{con } \xi \in (-1, 1).$$

En la siguiente tabla se presentan las abscisas y los coeficientes de la fórmula de cuadratura de Gauss-Legendre, donde m es el número de puntos utilizados:

m	x_k	β_k
2	$\pm 0.57735 \ 02691 \ 89626$	1.00000 00000 00000
3	$\pm 0.77459 \ 66692 \ 41483$ 0.00000 00000 00000	0.55555 55555 55556 0.88888 88888 88889
4	$\pm 0.86113 \ 63115 \ 94053$ $\pm 0.33998 \ 10435 \ 84856$	0.34785 48451 37454 0.65214 51548 62546
5	$\pm 0.90617 \ 98459 \ 38664$ $\pm 0.53846 \ 93101 \ 05683$ 0.00000 00000 00000	0.23692 68850 56189 0.47862 86704 99366 0.56888 88888 88889
6	$\pm 0.93246 \ 95142 \ 03152$ $\pm 0.66120 \ 93864 \ 66265$ $\pm 0.23861 \ 91860 \ 83197$	0.17132 44923 79170 0.36076 15730 48139 0.46791 39345 72691
7	$\pm 0.94910 \ 79123 \ 42759$ $\pm 0.74153 \ 11855 \ 99394$ $\pm 0.40584 \ 51513 \ 77397$ 0.00000 00000 00000	0.12948 49661 68870 0.27970 53914 89277 0.38183 00505 05119 0.41795 91836 73469

Ejemplo. Si se quiere calcular $\int_1^3 \frac{dx}{x}$ con $m = 3$ (fórmula de Gauss-Legendre), se introduce el cambio $y = x - 2$, $-1 \leq y \leq 1$ (intervalo estándar) y queda

$$\int_{-1}^1 \frac{dy}{y+2} \approx \frac{5}{9} \cdot \frac{1}{1.225403} + \frac{8}{9} \cdot \frac{1}{2} + \frac{5}{9} \cdot \frac{1}{2.774598} \approx 1.098039$$

El valor exacto es $\ln 3 = 1.098612$. Una estimación del error viene dada por

$$E_3\{f\} = \frac{2}{7!} \left[\frac{2^3 (3!)^2}{6!} \right]^2 f^{(6)}(\eta) = \frac{8}{175} \frac{1}{(\eta+2)^7}$$

$$\text{y } 0.000021 = \frac{8}{175} \frac{1}{3^7} < E_3\{f\} < \frac{8}{175} = 0.045714.$$

Ejercicio. Aproximar por el método de Gauss-Legendre las integrales siguientes con $m = 2, 3, 4, 5$

$$\begin{array}{ll} \text{a) } \int_{-4}^4 \frac{dx}{1+x^2} dx & \text{b) } \int_0^1 e^{-10x} \sin x dx \\ \text{c) } \int_0^5 x e^{-3x^2} dx & \text{d) } \int_{-1}^1 \frac{\cos x}{\sqrt{1-x^2}} dx = \pi J_0(1) = 2.40394 \end{array}$$

calculando en primer lugar una estimación del error y calculando después el error exacto.

5.5.2 Integración gaussiana con peso

Si se tiene una función peso en el integrando, $W\{g\} = \int_a^b g(x) w(x) dx$, se approxima de forma análoga por

$$W_m\{g\} = \sum_{j=1}^m \gamma_j g(x_j) \quad (5.4)$$

De la misma manera que en la sección anterior, se tiene el siguiente teorema, que asegura la misma precisión pero con otras familias de polinomios ortogonales atendiendo a la función peso y al intervalo de integración:

Teorema 5.6. La fórmula de integración 5.4 tiene grado de precisión $2m - 1$ si se toma $q_m(x) = (x - x_1) \cdots (x - x_m)$ tal que los x_j son los ceros del polinomio ortogonal de grado m , $q_m(x)$, en el intervalo $[a, b]$ con la función peso $w(x)$ (respecto al producto escalar $\langle f, g \rangle = \int_a^b w(x) f(x) g(x) dx$):

$$\int_a^b q_m(x) w(x) x^r dx = 0 \quad r = 1 \div m - 1$$

Los coeficientes γ_j se pueden calcular por

$$\gamma_j = \frac{1}{q'_m(x_j)} \int_a^b \frac{q_m(x)}{x - x_j} w(x) dx = \int_a^b l_j(x) w(x) dx \quad j = 1 \div n$$

son todos positivos y se llaman números de Christoffel. Además, si $g \in C^{2m}[a, b]$, se tiene una expresión similar del error

$$E_m\{g\} = W\{g\} - W_m\{g\} = \frac{g^{(2m)}(\xi)}{(2m)!} \int_a^b (q_n(x))^2 w(x) dx, \quad \text{donde } \xi \in (a, b).$$

Se presentan a continuación diferentes fórmulas de integración según la expresión de la función peso $w(x)$.

Integración Gauss-Chebishev

Si se considera el intervalo $[-1, 1]$ y la función peso definida por $w(x) = \frac{1}{\sqrt{1-x^2}}$, se tiene la integral $W\{g\} = \int_{-1}^1 \frac{g(x)}{\sqrt{1-x^2}} dx$.

Este método presenta la ventaja de que todos los coeficientes γ_j son iguales y no es necesario calcularlos

$$W_m\{g\} = \frac{\pi}{m} \sum_{j=1}^m g(x_j)$$

donde los x_j son los ceros del polinomio de Chebishev de grado m . En particular, la expresión del error viene dada por

$$E_m\{g\} = \frac{g^{(2m)}(\xi)}{(2m)!} \frac{2\pi}{2^{2m}}, \quad \text{con } \xi \in (-1, 1).$$

Ejemplo. Si se quiere calcular $I = \int_{-1}^1 \frac{e^x dx}{\sqrt{1-x^2}}$ con seis decimales correctos, hay que encontrar m tal que el error $E_m\{g\} = \frac{2\pi}{2^{2m}(2m)!} e^\xi$ con $\xi \in (-1, 1)$. Entonces,

$$|E_m\{g\}| \leq \frac{2\pi e}{2^{2m}(2m)!} \approx 4.6 \cdot 10^{-9} \quad \text{para } m = 5.$$

Por tanto, $I \approx \frac{\pi}{5} \sum_{j=1}^5 \exp\left[\cos\left(\frac{2j-1}{5}\frac{\pi}{2}\right)\right] = 3.977463$ con todos los decimales correctos.

Ejercicios.

1. Aproximar la integral $I = \int_{-1}^1 \frac{\cos x dx}{\sqrt{1-x^2}}$ por la fórmula de integración de Gauss-Chebishev con $m = 2, 3, 4, 5$ y calcular el error a partir del valor exacto $\pi J_0(1) = 2.4039391$.
2. Aproximar la integral $I = \int_{-1}^1 |x| \frac{dx}{\sqrt{1-x^4}}$ por la fórmula de integración de Gauss-Chebishev con $m = 2, 3, 4, 5$ y calcular el error a partir del valor exacto $\pi/2$.

Integración Gauss-Laguerre

Si se toma el intervalo $[0, \infty)$ y se considera la función peso $w(x) = e^{-x}$, se tiene la integral $W\{g\} = \int_0^\infty g(x) e^{-x} dx$. Los polinomios ortogonales son los de Laguerre

$$W_m\{g\} = \sum_{j=1}^m \beta_j g(x_j) \quad \text{con } \beta_j = \frac{m!}{x_j (L'_m(x_j))^2},$$

donde los x_j son los ceros del polinomio de Laguerre de grado m , $L_m(x)$. El error al aplicar este método viene dado por $E_m\{g\} = \frac{(m!)^2}{(2m)!} g^{(2m)}(\xi)$ con $\xi \in (0, \infty)$. Más detalles sobre las abscisas y los pesos se presentan en la tabla siguiente:

m	x_j	β_j
2	0.58578 64376 27	0.85355 33905 93
	3.41421 35623 73	0.14644 66094 07
3	0.41577 45567 83	0.71109 30099 29
	2.29428 03602 79	0.27851 77335 69
	6.28994 50829 37	0.10389 25650 16 · 10 ⁻¹
4	0.32254 76896 19	0.60315 41043 42
	1.74576 11011 58	0.35741 86924 28
	4.53662 02969 21	0.38887 90851 50 · 10 ⁻¹
	9.39507 09123 01	0.53929 47055 61 · 10 ⁻³
5	0.26356 03197 18	0.52175 56105 83
	1.41340 30591 07	0.39866 68110 83
	3.59642 57710 41	0.75942 44968 17 · 10 ⁻¹
	7.08581 00058 59	0.36117 58679 92 · 10 ⁻²
	12.64080 08442 76	0.23369 97238 58 · 10 ⁻⁴
6	0.22284 66041 79	0.45896 46739 50
	1.18893 21016 73	0.41700 08307 72
	2.99273 63260 59	0.11337 33820 74
	5.77514 35691 05	0.10399 19745 31 · 10 ⁻¹
	9.83746 74183 83	0.26101 72028 15 · 10 ⁻³
	15.98287 39806 02	0.89854 79064 30 · 10 ⁻⁶

Ejercicios.

- Intentad calcular $\Gamma(8) = \int_0^\infty x^7 e^{-x} dx$ por el método de Gauss-Laguerre con $m = 3$. ¿Es una buena aproximación de $\Gamma(8) = 7!$?
- Aproximar la integral $\int_0^\infty e^{-10x} \sin x dx$ mediante la integración de Gauss-Laguerre con $m = 2, 3, 4, 5$, y comparar con el resultado exacto.

Integración Gauss-Hermite

Si se toma el intervalo $(-\infty, \infty)$ y se considera la función peso $w(x) = e^{-x^2}$, se tiene la integral $W\{g\} = \int_{-\infty}^{\infty} g(x) e^{-x^2} dx$. Los polinomios ortogonales son los de Hermite

$$W_m\{g\} = \sum_{j=1}^m \gamma_j g(x_j) \text{ con } \gamma_j = \frac{2^{m+1} m! \sqrt{\pi}}{(H'_n(x_j))^2},$$

donde los x_j son los ceros del polinomio de Hermite de grado m , $H_m(x)$. Más detalles sobre las abscisas y los pesos se presentan en la tabla siguiente:

m	x_j	γ_j
2	$\pm 0.70710\ 67811\ 86548$	0.88622 69254 53
3	0.00000 00000 00000 $\pm 1.22474\ 48713\ 91589$	1.18163 59006 04 0.29540 89751 51
4	$\pm 0.52464\ 76232\ 75290$ $\pm 1.65068\ 01238\ 85785$	0.80491 40900 06 $0.81312\ 83544\ 73 \cdot 10^{-1}$
5	0.00000 00000 00000 $\pm 0.95857\ 24646\ 13819$ $\pm 2.02018\ 28704\ 56086$	0.94530 87204 83 0.39361 93231 52 $0.19953\ 24205\ 91 \cdot 10^{-1}$
6	$\pm 0.43607\ 74119\ 27617$ $\pm 1.33584\ 90740\ 13697$ $\pm 2.35060\ 49736\ 74492$	0.72462 95952 24 0.15706 73203 23 $0.45300\ 09905\ 51 \cdot 10^{-2}$

El error que se comete al aplicar el método de Gauss–Hermite viene dado por

$$E_m \{ g \} = \frac{g^{(2m)}(\xi)}{(2m)!} \frac{m! \sqrt{\pi}}{2^m}$$

con $\xi \in (-\infty, \infty)$.

Ejercicio. Calcular la integral $\int_{-\infty}^{\infty} |x| e^{-3x^2} dx$, mediante el método de Gauss–Hermite para $m = 2, 3, 4, 5$, y comparar el resultado con su valor real. Dar una cota del error.

5.6 Problemas

1. Sea X una cantidad física que depende de la presión de un gas de acuerdo con la fórmula:

$$X = C_0 + C_1 P^2 + C_2 P^3 + C_3 P^6$$

donde C_0, C_1, C_2 y C_3 son constantes diferentes de cero. Determinar el valor de X en el vacío ($P = 0$), si se tiene la tabla siguiente:

P (mm Hg)	0.8	0.4	0.2	0.1	0.05
X unidades	740	487	475	485	489

2. Calculad, utilizando la fórmula compuesta de los trapecios, la integral

$$\int_0^1 \cos(\cos x) dx$$

con una precisión de $0.5 \cdot 10^{-2}$.

3. Calcular, por el método de Simpson, la integral $\int_0^1 e^{e^x} dx$, eligiendo el paso de integración conveniente para que el error sea menor que 10^{-2} .

4. Calcular $\int_0^{0.5} (1 - x^2)^{1/2} dx$,

- a) desarrollando por la fórmula de Taylor (con 3 decimales correctos)
 b) aplicando el método de Romberg (con 5 decimales correctos)

5. Calcular $\int_0^{0.4} (1 - 0.5 \sin^2 x)^{-1/2} dx$, usando la fórmula de Simpson con pasos $h = 0.2, 0.1, 0.05$, determinando una cota del error y extrapolando los resultados.

6. Calcular la integral $\int_0^{0.9} e^{x^2} dx$ aplicando el método de los trapecios con pasos $h = 0.9, 0.3, 0.1$; emplear el método de Romberg para extrapolar y, finalmente, comparar los resultados con el que se obtiene de aplicar la fórmula de Euler–Maclaurin con dos términos correctores:

$$\int_a^b f(x) dx = T(h) - \frac{h^2}{12} [f'(b) - f'(a)] + \frac{h^4}{720} [f^{(3)}(b) - f^{(3)}(a)] - \frac{h^6}{30240} [f^{(5)}(b) - f^{(5)}(a)] + R$$

7. Calcular la integral impropia $\int_0^\infty (x^3 + x)^{-1/2} dx$ con un error inferior a $\varepsilon = 10^{-2}$ dividiéndola en dos partes. La primera se calcula por el método de Romberg con un error inferior a $2\varepsilon/3$, definiendo previamente un cambio de variable adecuado. Para el cálculo de la segunda integral se desarrolla el integrando en una serie de potencias de x^{-1} , se integra la serie y se suma hasta que el error sea menor que $\varepsilon/3$.

8. Determinar la longitud de la elipse

$$x^2 + \frac{y^2}{4} = 1$$

con una exactitud de 6 cifras decimales.

9. Construir una fórmula de integración de dos puntos para integrales de la forma:

$$\int_{-1}^1 (1 + x^2) f(x) dx$$

con grado de precisión igual a tres. Dar la expresión del error aplicándolo al cálculo de $\int_{-1}^1 (1 + x^2) x^4 dx$.

10. Determinar los pesos de la siguiente fórmula de integración de manera que sea exacta para todos los polinomios de grado ≤ 3 :

$$\int_a^b f(x) dx = A_0 f(a) + A_1 f(b) + C_0 f^{(2)}(a) + C_1 f^{(2)}(b)$$

11. Demostrar que la fórmula de integración:

$$\int_{-1}^1 f(x) dx \approx \frac{1}{9} (5 f(-\sqrt{0.6}) + 8 f(0) + 5 f(\sqrt{0.6}))$$

es exacta para polinomios de grado ≤ 5 . Estimar el error desarrollando los dos miembros en el entorno de $x = 0$.

Generalizar la fórmula a un intervalo $[a, b]$ cualquiera y construir la fórmula de integración compuesta.

Calcular $\int_{-1}^1 e^x dx$ con pasos 1, 0.5 y 0.25, acotando los errores y aplicar, finalmente, extrapolación de Richardson.

12. Indicar como calcular numéricamente las integrales siguientes:

$$\int_0^\infty \frac{\sin x}{x} dx \quad \int_0^1 \frac{dx}{x^{1/2} + x^{1/3}}$$

13. Calcular

$$\int_{-1}^1 \frac{\cos x}{\sqrt{1-x^2}} dx$$

con un error que sea menor que 10^{-6} .

14. Sea $(x_0, y_0) \in \mathbf{R}^2$ y $f : \mathbf{R}^2 \rightarrow \mathbf{R}$. Utilizar extrapolación de Richardson para calcular la derivada parcial $\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)$ a partir de la siguiente tabla de valores:

	y_{-2}	y_{-1}	y_0	y_1	y_2
x_{-2}	0.7329	0.7705	0.8100	0.8515	0.8952
x_{-1}	0.8166	0.8585	0.9025	0.9488	0.9974
x_0	0.9048	0.9512	1.0000	1.0513	1.1052
x_1	0.9976	1.0487	1.1025	1.1590	1.2185
x_2	1.0949	1.1510	1.2100	1.2720	1.3373

donde $x_k = x_0 + k h$, $y_k = y_0 + k h$ y $h = 0.05$.

15. Calcular $\int_0^\pi e^{\cos x} dx$ por el método de Gauss-Chebishev con el mínimo número de puntos de manera que el error sea menor que 10^{-8} .

16. Deducir la fórmula de Newton-Côtes abierta de $n - 1$ abscisas:

$$\int_{x_0}^{x_n} f(x) dx \approx \sum_{k=1}^{n-1} A_k f(x_k)$$

donde las abcisas, x_k , $k = 0 \div n$, son equiespaciadas. Tomad los valores de $n = 3, 4, 5$. ¿Qué grado de precisión tienen estas fórmulas?

17. Calcular por el método de Gauss-Chebishev con dos puntos la integral

$$\int_0^\pi \cos^3 x dx$$

Dar una estimación del error. ¿Cuál es el error exacto?

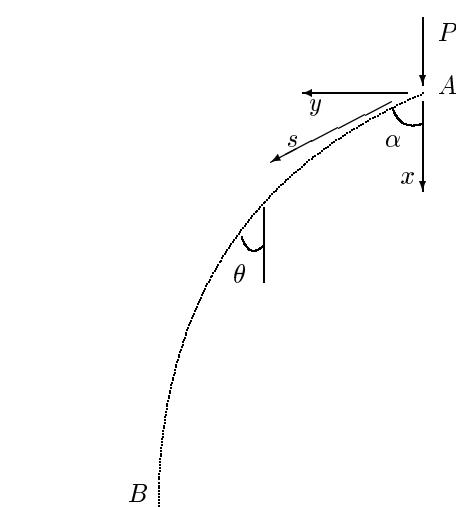
18. Sabiendo que $\int_0^\infty e^{-ax^2} \cos bx dx = \frac{1}{2} \sqrt{\frac{\pi}{a}} e^{-\frac{b^2}{4a}}$, aplicar el método de cuadratura de Gauss-Hermite a la integral $\int_{-\infty}^\infty e^{-x^2} \cos x dx$, para $m = 2, 3, 4, 5$. Dar una cota del error.

5.7 Prácticas

5.7.1 Práctica ejemplo

Escribir una rutina que utilice el método de Romberg para 8 subintervalos con adaptación del paso. Aplicarlo a la siguiente situación:

Un pilar muy ligero de longitud L tiene un módulo de Young E y un momento de inercia I . Está encastado en el extremo B e inicialmente está vertical. Una fuerza exterior P en el extremo libre A causa una flexión en el pilar. Si $\theta = \theta(s)$ es el ángulo entre el pilar y la vertical en un punto cualquiera, $x = x(s)$ la deformación vertical y s la distancia desde A , la ecuación diferencial exacta es



$$EI \frac{d\theta}{ds} = -P y(s)$$

donde y es la deformación horizontal con

$$\frac{dy}{ds} = \operatorname{sen} \theta \quad \text{que da lugar a}$$

$$ds = -\frac{d\theta}{\sqrt{\frac{2P}{EI} (\cos \theta - \cos \alpha)}}$$

Si se integra, se obtiene

$$L = \sqrt{\frac{EI}{P}} \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - \operatorname{sen}^2 \frac{\alpha}{2} \operatorname{sen}^2 \phi}}$$

donde α es el valor de θ en A .

La carga de Euler para que el pilar empiece a curvarse viene dada por $P_e = \frac{\pi^2 EI}{4L^2}$. Sean (x_A, y_A) la ordenada y la abcisa de A respecto a B . Calcular los valores de $\frac{P}{P_e}$, $\frac{x_A}{L}$ y $\frac{y_A}{L}$ para $\alpha = 20^\circ, 120^\circ (20^\circ)$, teniendo en cuenta que, si llamamos $\mathcal{I}(\lambda) = \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - \lambda^2 \operatorname{sen}^2 \phi}}$ y $\mathcal{J}(\lambda) = \int_0^{\pi/2} \frac{1 - 2\lambda^2 \operatorname{sen}^2 \phi}{\sqrt{1 - \lambda^2 \operatorname{sen}^2 \phi}} d\phi$, donde $\lambda = \operatorname{sen} \frac{\alpha}{2}$, se tiene

$$y_A = 2 \sqrt{\frac{EI}{P}} \operatorname{sen} \frac{\alpha}{2} \quad \text{y} \quad x_A = \sqrt{\frac{EI}{P}} \mathcal{J}(\lambda).$$

Dada la fórmula de los trapecios $T(h) = I + c_1 h^2 + c_2 h^4 + c_3 h^6 + \dots$ con $I = \int_a^b f(x) dx$, se realizan m cálculos sobre el intervalo $[0, h]$ con pasos $h, h/2, \dots, h/2^{m-1}$.

Sea T_0 la primera columna definida por

$$T_0^k = \frac{h}{2^k} \left(\frac{1}{2}f_0 + f_1 + \cdots + f_{2^k-1} + \frac{1}{2}f_{2^k} \right) = T_0 \left(\frac{h}{2^k} \right)$$

y, entonces, $T_0^k = I + c_1 \left(\frac{h}{2^k} \right)^2 + c_2 \left(\frac{h}{2^k} \right)^4 + c_3 \left(\frac{h}{2^k} \right)^6 + \cdots$. Aplicando la técnica de extrapolación de Richardson la columna siguiente vendrá definida por

$$T_1^k = T_0^k + \frac{T_0^k - T_0^{k-1}}{2^2 - 1} = \frac{1}{3}(4T_0^k - T_0^{k-1})$$

Este procedimiento se repite hasta conseguir la tabla siguiente:

$$\begin{array}{ccccccc} T_0^0 & & & & & & \\ T_0^1 & T_1^1 & & & & & \\ T_0^2 & T_1^2 & T_2^2 & & & & \\ \vdots & \vdots & & \ddots & & & \\ T_0^m & T_1^m & \dots & \dots & T_m^m & & \end{array}$$

Se obtiene $T_1^k = I + \frac{1}{3} \sum_{j=2}^{\infty} \left(\frac{4}{2^{2j}} - 1 \right) c_j \left(\frac{h}{2^k} \right)^{2j}$ y así, al extrapolar con espaciados $h/2^{k-1}$ y $h/2^k$, se tiene un error del orden de $(h/2^k)^4$. En general, si $T_n^k = \frac{1}{4^n - 1} (4^n T_{n-1}^k - T_{n-1}^{k-1})$, $k \geq 1$ y $k \leq n \leq m$, se puede (ver [Ral78]) expresar el término del error por $T_n^k = I + O \left(\frac{h}{2^k} \right)^{2n+2}$ y, por tanto, $T_m^m = I + O \left(\frac{h}{2^m} \right)^{2m+2}$ que, para $m = 3$, da la fórmula que se ha utilizado para construir la rutina ROMBERG8: $T_3^3 = I + O \left(\frac{h}{8} \right)^8$. La fórmula de integración es

$$\frac{h}{5670} \left[217(f(0) + f(h)) + 1024(f(\frac{h}{8}) + f(\frac{3h}{8}) + f(\frac{5h}{8}) + f(\frac{7h}{8})) + 352(f(\frac{h}{4}) + f(\frac{3h}{4})) + 436f(\frac{h}{2}) \right] + O(h^8)$$

El usuario de ROMBERG8 tendrá que construir una función FUN para calcular el integrando; los extremos de integración $A=a$, $B=b$, así como las tolerancias de los errores relativo RELERR y absoluto ABSERR son los parámetros de entrada. La rutina retorna el resultado aproximado de la integral y satisface la tolerancia mayor en RESULT, una estimación del error absoluto ERRMAX, el número de evaluaciones de la función FUN(X) en NOFUN y un indicador, FLAG, de si los cálculos se han realizado sin ninguna anomalía.

Se ha elegido una regla que divide el intervalo en dos evaluándolo en primer lugar completo $QV=P_i$, y después aplicándolo a cada mitad $QN=Q_i$, donde QE es el valor sobre la parte izquierda

del intervalo y QD el correspondiente a la derecha. Como el error es del orden de h^8 , se impone la siguiente condición de bisección:

$$\frac{1}{255} |P_i - Q_i| \leq \frac{h_i}{b-a} \varepsilon$$

y se comparan las cantidades

$$\text{ESTERR} = \text{ABS}(\text{QN}-\text{QV})/255$$

$$\text{TOLERR} = \text{AMAX1}(\text{ABSERR}, \text{RELERR} * \text{ABS}(\text{AREA})) * (\text{PAS}/\text{PES})$$

donde

PAS/PES es $h_i/(b-a)$,
 AREA es una estimación de la integral en $[a, b]$ y
 ε es el máximo entre ABSERR y $\text{RELERR} * \left| \int_a^b f \right|$.

Si el subintervalo cumple la condición de bisección, es decir, $\text{ESTERR} \leq \text{TOLERR}$, entonces QN se añade a RESULT y ESTERR a ERRMAX . Además, se suma a CER la cantidad $(\text{QN}-\text{QV})/255$, ya que $Q_i + (Q_i - P_i)/255$ es más aproximado (extrapolación Richardson) a I_i que no Q_i . El valor final de CER es acumulado en RESULT .

Cada vez que el intervalo es dividido, los nodos y los valores de la función para la mitad derecha son almacenados para su uso posterior. La variable $\text{NIVMAX} = 30$ define el límite de nivel de partición. Cuando se llega a subintervalos de longitud del orden de $(b-a)/2^{30}$, la bisección se para y el subintervalo se acepta a pesar de no cumplir ninguna condición de error; en este caso, un contador del número de intervalos en estas condiciones, parte entera de FLAG , se aumenta en 1. Otro límite es necesario: la variable NOMAX es el límite de llamadas a la función. Cuando se excede, NIVMAX se reduce a $\text{NIVSORT} = 6$, de manera que lo que falta de cálculos en el intervalo se realizan con un número menor de evaluaciones que NOMAX . Además, si PX es el punto que presenta el problema, $(B-\text{PX})/(B-A)$ se retorna en la parte fraccionaria de FLAG y es la parte del intervalo que se ha calculado con un límite de bisección más tolerante. El punto PX puede ser calculado por $B - \text{frac}(\text{FLAG}) * (B-A)$. Una alternativa a esta subrutina es QUANC8 de [For77] que utiliza una fórmula de Newton-Côtes con 8 intervalos.

Los resultados obtenidos al calcular

$$\frac{P}{P_e} = \frac{4}{\pi^2} \mathcal{I}^2(\lambda), \quad \frac{x_A}{L} = \frac{\mathcal{J}(\lambda)}{\mathcal{I}(\lambda)} \quad \text{y} \quad \frac{y_A}{L} = \frac{2\lambda}{\mathcal{I}(\lambda)}.$$

mediante ROMBERG8 se presentan en la tabla siguiente:

α	20°	40°	60°	80°	100°	120°
P / P_e	1.01540	1.06366	1.15172	1.29389	1.51839	1.88480
x_A / L	0.96973	0.88120	0.74102	0.55940	0.34899	0.12316
y_A / L	0.21941	0.42224	0.59321	0.71950	0.79154	0.80317

El programa principal que llama a la subrutina ROMBERG8 para la integración se presenta a continuación; se introduce un COMMON para poder calcular las funciones que dependen del parámetro λ , al que se llama T. Todo ello queda reflejado en el programa principal así como en las funciones:

```

REAL*8 FUNCTION FUN(X)
REAL*8 X,T
COMMON T
FUN= DSQRT(1.0-(T**2)*((DSIN(X))**2))
FUN= 1.0/FUN
RETURN
END

REAL*8 FUNCTION FUN1(X)
REAL*8 X,T,AUX
COMMON T
AUX=(DSIN(X))**2
FUN1=(1.0-2.0*(T**2)*AUX)/(DSQRT(1.0-(T**2)*AUX))
RETURN
END

PROGRAM PILAR
EXTERNAL FUN,FUN1
REAL*8 FUN,A,B,ABSERR,RELERR,RESULT,ERREST,FLAG
REAL*8 GRAD,RAD,PI,T,FUN1,IL,JL,UN,DOS,TRES
COMMON T
PI= 4.0*DATAN(1.D0)
A= 0.0
B= PI/2.0
ABSERR= 0.0
RELERR= 10.0D-14
DO 10 GRAD=20.0,120.0,20.0
RAD= GRAD*PI/180.0
T = DSIN(RAD/2.0)
CALL ROMBERG8(FUN,A,B,ABSERR,RELERR,RESULT,ERREST,NOFUN,FLAG)
IL = RESULT
CALL ROMBERG8(FUN1,A,B,ABSERR,RELERR,RESULT,ERREST,NOFUN,FLAG)
JL = RESULT
UN = (4.0*(IL**2))/(PI**2)
DOS= JL/IL
TRES= (2.0*T)/IL
WRITE(6,100),UNO,DOS,TRES
10  CONTINUE
100 FORMAT(T20,F7.5,T30,F7.5,T40,F7.5)
END

```

```

SUBROUTINE ROMBERG8(FUN,A,B,ABSERR,RELERR,RESULT,ERRMAX,NOFUN,
* FLAG)
REAL*8 FUN, A, B, ABSERR, RELERR, RESULT, ERRMAX, FLAG
INTEGER NOFUN

C
C      CALCULO DE LA INTEGRAL DE FUN(X) DESDE A HASTA B
C      CON UNA TOLERANCIA PREFIJADA.
C      RUTINA ADAPTATIVA AUTOMATICA BASADA EN
C      EL METODO DE ROMBERG PARA 8 INTERVALOS.
C
C      ENTRADA ..
C      FUN NOMBRE DEL INTEGRANDO: FUNCION SUBPROGRAMA FUN(X).
C      A LIMITE INFERIOR DE INTEGRACION.
C      B LIMITE SUPERIOR DE INTEGRACION. (B PUEDE SER MENOR QUE A)
C      RELERR TOLERANCIA DEL ERROR RELATIVO. (NO NEGATIVO)
C      ABSERR TOLERANCIA DEL ERROR ABSOLUTO. (NO NEGATIVO)
C
C      SALIDA ..
C      RESULT APROXIMACION DE LA INTEGRAL. TIENE QUE SATISFACER
C      DE LAS DOS TOLERANCIAS DEL ERROR LA MAS LIGERA.
C      ERRMAX ESTIMACION DE LA MAGNITUD DEL ERROR REAL.
C      NOFUN NUMERO DE VECES QUE SE EVALUA LA FUNCION.
C      FLAG INDICADOR DE CONFIANZA EN RESULT.
C      SI FLAG = 0, RESULT SATISFACE LA TOLERANCIA DEL ERROR.
C      SI FLAG = XXX.YYY , ENTONCES
C      XXX = EL NUMERO DE INTERVALOS QUE NO CONVERGEN Y
C      0.YYY = LA FRACCION DE INTERVALO CALCULADA APROX.
C      DEBIDO A QUE SE HA LLEGADO AL LIMITE DE NOFUN.
C
C      REAL*8 W0,W1,W2,W3,W4,AREA,X0,FO,PES,PAS,CER,TEMP
C      REAL*8 QV,QN,QDIFF,QE,ESTERR,TOLER
C      REAL*8 QD(31),F(16),X(16),FSAVE(8,30),XSAVE(8,30)
C      INTEGER NIVMIN,NIVMAX,NIVSOR,NOMAX,NOFIN,LEV,NIM,I,J

C
C      *** ETAPA 1 *** INICIALIZACION GENERAL
C      DEFINICION DE CONSTANTES.
NIVMIN = 1
NIVMAX = 30
NIVSOR = 6
NOMAX = 5000
NOFIN = NOMAX - 8*(NIVMAX-NIVSOR+2** (NIVSOR+1))

C
C      ALERTA CUANDO NOFUN LLEGA A NOFIN
W0 = 1736.D0 / 5670.0
W1 = 8192.D0 / 5670.0
W2 = 2816.D0 / 5670.0
W3 = W1
W4 = 3488.D0 / 5670.0

C
C      SE INICIALIZAN LAS SUMAS PARCIALES A CERO.
FLAG = 0.0
RESULT = 0.0
CER = 0.0
ERRMAX = 0.0
AREA = 0.0
NOFUN = 0
IF (A .EQ. B) RETURN

```

```

C      *** ETAPA 2 *** INICIALIZACION PARA EL PRIMER INTERVALO
LEV = 0
NIM = 1
X0 = A
X(16) = B
QV = 0.0
F0 = FUN(X0)
PES = (B - A) / 16.0
X(8) = (X0 + X(16)) / 2.0
X(4) = (X0 + X(8)) / 2.0
X(12) = (X(8) + X(16)) / 2.0
X(2) = (X0 + X(4)) / 2.0
X(6) = (X(4) + X(8)) / 2.0
X(10) = (X(8) + X(12)) / 2.0
X(14) = (X(12) + X(16)) / 2.0
DO 25 J = 2, 16, 2
F(J) = FUN(X(J))
25 CONTINUE
NOFUN = 9
C
C      *** ETAPA 3 *** CALCULOS CENTRALES
C      NECESA QV,X0,X2,X4,...,X16,F0,F2,F4,...,F16.
C      CALCULA X1,X3,...X15, F1,F3,...F15, QE,QD,QN,QDIFF, AREA.
30 X(1) = (X0 + X(2)) / 2.0
F(1) = FUN(X(1))
DO 35 J = 3, 15, 2
X(J) = (X(J-1) + X(J+1)) / 2.0
F(J) = FUN(X(J))
35 CONTINUE
NOFUN = NOFUN + 8
PAS = (X(16) - X0) / 16.0
QE = (W0*(F0 + F(8)) + W1*(F(1)+F(7)) + W2*(F(2)+F(6))
* + W3*(F(3)+F(5)) + W4*F(4)) * PAS
* QD(LEV+1)=(W0*(F(8)+F(16))+W1*(F(9)+F(15))+W2*(F(10)+F(14))
* + W3*(F(11)+F(13)) + W4*F(12)) * PAS
QN = QE + QD(LEV+1)
QDIFF = QN - QV
AREA = AREA + QDIFF
C
C      *** ETAPA 4 *** TEST DE CONVERGENCIA INTERVALAR
ESTERR = DABS(QDIFF) / 255.0
TOLERR = DMAX1(ABSER,RELER*DABS(AREA)) * (PAS/PES)
IF (LEV .LT. NIVMIN) GO TO 50
IF (LEV .GE. NIVMAX) GO TO 62
IF (NOFUN .GT. NOFIN) GO TO 60
IF (ESTERR .LE. TOLERR) GO TO 70
C
C      *** ETAPA 5 *** NO CONVERGENCIA
C      SE SITUA EL PROXIMO INTERVALO.
50 NIM = 2*NIM
LEV = LEV+1
C
C      SE GUARDAN LOS ELEMENTOS DE LA DERECHA PARA MAS ADELANTE.
DO 52 I = 1,8
FSAVE(I,LEV) = F(I+8)
XSAVE(I,LEV) = X(I+8)
52 CONTINUE

```

```

C      SE REUNEN LOS ELEMENTOS DE LA IZQUIERDA PARA UTILIZARLOS.
C      QV = QE
C      DO 55 I = 1,8
C      J = -I
C      F(2*I+18) = F(I+9)
C      X(2*I+18) = X(I+9)
55    CONTINUE
      GO TO 30
C
C      *** ETAPA 6 *** SECCION ALERTA
C      EL NUMERO DE LLAMADAS A LA FUNCION ESTA A PUNTO DE EXCEDER
C      EL LIMITE.
60    NOFIN = 2*NOFIN
      NIVMAX = NIVSOR
      FLAG = FLAG + (B-X0) / (B - A)
      GO TO 70
C
C      EL NIVEL NORMAL ES NIVMAX.
62    FLAG = FLAG + 1.0
C
C      *** ETAPA 7 *** INTERVALO CONVERGENTE
C      SE AÑADEN LOS RESULTADOS A LAS SUMAS PARCIALES.
70    RESULT = RESULT + QN
      ERRMAX = ERRMAX + ESTERR
      CER = CER + QDIFF / 255.0
C
C      SE SITUA EL PROXIMO INTERVALO.
72    IF (NIM .EQ. 2*(NIM/2)) GO TO 75
      NIM = NIM/2
      LEV = LEV-1
      GO TO 72
75    NIM = NIM + 1
      IF (LEV .LE. 0) GO TO 80
C      SE REUNEN LOS ELEMENTOS NECESARIOS PARA EL PROXIMO INTERVALO.
      QV = QD(LEV)
      X0 = X(16)
      F0 = F(16)
      DO 78 I = 1,8
      F(2*I) = FSAVE(I,LEV)
      X(2*I) = XSAVE(I,LEV)
78    CONTINUE
      GO TO 30
C
C      *** ETAPA 8 *** FINALIZACION Y RETURN
80    RESULT = RESULT + CER
C
C      SE ASEGURA QUE ERRMAX NO ES MENOR QUE LA PRECISION MAQUINA
      IF (ERRMAX .EQ. 0.0) RETURN
82    TEMP = DABS(RESULT) + ERRMAX
      IF (TEMP .NE. DABS(RESULT)) RETURN
      ERRMAX = 2.0*ERRMAX
      GO TO 82
      END

```

5.7.2 Enunciados

1. La función erf viene definida por la integral

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

y es muy fácil de evaluar numéricamente.

- (a) Utilizar la rutina ROMBERG8, con diferentes valores de la tolerancia del error **RELERR** $=10^{-10}, 10^{-12}, 10^{-14}$, para construir una tabla de la función $\operatorname{erf}(x)$ con $x = 0.0, 2.0$ (0.1).
- (b) Comparar esta tabla con los valores obtenidos en la práctica 1.1. ¿Se puede exigir la misma precisión? ¿Cuál de los dos procedimientos es menos costoso en tiempo de cálculo?

2. La función J_0 viene definida por la integral

$$J_0(x) = \frac{1}{\pi} \int_0^\pi \cos(x \sin t) dt$$

y es muy fácil de evaluar numéricamente.

- (a) Utilizar la rutina ROMBERG8, con diferentes valores de la tolerancia del error **RELERR** $=10^{-10}, 10^{-12}, 10^{-14}$, para construir una tabla de la función $J_0(x)$ con $x = 0.0, 4.0$ (0.25).
- (b) Comparar esta tabla con los valores obtenidos en la práctica 1.2. ¿Se puede exigir la misma precisión? ¿Cuál de los dos procedimientos es menos costoso en tiempo de cálculo?

3. La función $F(a, b; c; x)$ viene definida por la integral

$$F(a, b; c; x) = \frac{\Gamma(c)}{\Gamma(b) \Gamma(c-b)} \int_0^1 t^{b-1} (1-t)^{c-b-1} (1-tx)^{-a} dt$$

que en el caso particular de $a = 1/2, b = 1/2, c = 1$, realizando el cambio $t = \sin^2 \phi$ y teniendo en cuenta que $\Gamma(1/2) = \sqrt{\pi}$, se tiene

$$F(1/2, 1/2; 1; x) = \frac{2}{\pi} \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - x \sin^2 \phi}}$$

que es muy fácil de evaluar numéricamente.

- (a) Utilizar la rutina ROMBERG8, con diferentes valores de la tolerancia del error **RELERR** $=10^{-10}, 10^{-12}, 10^{-14}$, con el fin de construir una tabla de la función $F(1/2, 1/2; 1; x)$ para $x = 0.05, 0.95$ (0.05).
- (b) Comparar esta tabla con los valores obtenidos en la práctica 1.3. ¿Se puede exigir la misma precisión? ¿Cuál de los dos procedimientos es menos costoso en tiempo de cálculo?

4. Calcular aproximaciones de π mediante la integral

$$\pi = \int_0^1 \frac{4}{1+x^2} dx$$

- (a) Utilizar la regla de los trapecios y la de Simpson con $h = 1/2m$ y $m = 4, 16, 64$. ¿Es el error proporcional a h^p ? ¿Por qué?
- (b) Utilizar las rutinas SPLINE (cúbica natural) y SAVAL, construidas en la práctica 2.4, para implementar un método de integración que se llamará SPLINT y que se basa en la integración analítica de la spline cúbica. El número de intervalos $n = 2m$ será un parámetro de la rutina y elegir los mismos puntos del apartado anterior. ¿Es el error proporcional a alguna potencia de h ?
- (c) Construir una tabla que presente los resultados de la integración numérica efectuada en los apartados anteriores y los que devuelve la rutina ROMBERG8 con diferentes valores de la tolerancia del error RELEERR = $10^{-10}, 10^{-12}, 10^{-14}$.

5. Construir una rutina, llamada DSPLINE, con lista de argumentos N, X, Y, ARG que calcule la derivada de la spline cúbica completa en el punto $x = \text{ARG}$, si se tienen N+1 puntos de coordenadas (X(I), Y(I)), I = 0, ..., N. Si se define

$$\begin{aligned} J_0(x) &= \frac{1}{\pi} \int_0^\pi \cos(x \operatorname{sen} t) dt \\ J_1(x) &= \frac{1}{\pi} \int_0^\pi \cos(x \operatorname{sen} t - t) dt \end{aligned}$$

comprobar que $J'_0(x) = -J_1(x)$,

- a. por derivación analítica;
- b. por aproximación numérica de la derivada;
- c. con la rutina DSPLINE.

Cuando sea necesario calcular las funciones $J_0(x)$ y $J_1(x)$, se realiza mediante la rutina ROMBERG8 en los puntos $x = 0.1, 0.3, 0.5, 0.7, 0.9$, con diferentes valores de la tolerancia del error RELEERR = $10^{-10}, 10^{-12}, 10^{-14}$.

6. Evaluar la integral $\int_0^1 \frac{2}{2 + \operatorname{sen} 10\pi x} dx = 1.15470054$ por la regla de los trapecios con 2, 4, 8 y 16 intervalos. Aplicar extrapolación reiterada de Richardson.

Si se hace servir la fórmula de Euler–Maclaurin para determinar el error cuando se aplica la regla de los trapecios, ¿qué se puede asegurar?

Evaluar la integral con la ayuda de la rutina ROMBERG8 con diferentes valores de la tolerancia del error RELEERR = $10^{-10}, 10^{-12}, 10^{-14}$ y comparar los resultados.

7. Calcular los pesos de la fórmula de cuadratura siguiente:

$$\int_0^1 \sqrt{x} f(x) dx \approx \sum_{j=0}^n \varrho_j \cdot f\left(\frac{j}{n}\right)$$

para $n = 1, 2, 3$, tal que la fórmula tenga grado de precisión n . Dar una estimación del error. Calcular $\int_0^1 \sqrt{x} \cos x dx$, por el método construido y mediante la rutina ROMBERG8 con diferentes valores de la tolerancia del error RELEERR= $10^{-10}, 10^{-12}, 10^{-14}$, comparando los resultados. ¿Se puede exigir la misma precisión? ¿Cuál de los dos procedimientos es menos costoso en tiempo de cálculo?

8. Se definen las integrales de Fresnel por

$$C(x) = \int_0^x \cos \frac{\pi t^2}{2} dt \quad \text{y} \quad S(x) = \int_0^x \sin \frac{\pi t^2}{2} dt.$$

Para $10^{-8} < x \leq 1.6$ se tiene

$$\begin{aligned} C(x) &= x \left[\sum_{n=0}^{\infty} \frac{1}{(2n)! (4n+1)} \left(\frac{-\pi^2 x^4}{4} \right)^n \right] \\ S(x) &= \frac{\pi}{2} x^3 \left[\sum_{n=0}^{\infty} \frac{1}{(2n+1)! (4n+3)} \left(\frac{-\pi^2 x^4}{4} \right)^n \right] \end{aligned}$$

que pueden ser calculadas hasta que, en valor absoluto, un término sea menor que $\varepsilon = 10^{-8}$.

Si $x > 1.6$, se considera

$$\begin{aligned} C(x) &= \frac{1}{2} + \frac{1}{\pi x} \left[A(x) \sin \frac{\pi x^2}{2} - B(x) \cos \frac{\pi x^2}{2} \right] \\ S(x) &= \frac{1}{2} - \frac{1}{\pi x} \left[A(x) \cos \frac{\pi x^2}{2} - B(x) \sin \frac{\pi x^2}{2} \right] \end{aligned}$$

donde

$$\begin{aligned} A(x) &= 1 + \sum_{n=1}^{\infty} \frac{(4n-1)!}{2^{2n-1}(2n-1)!} \cdot \frac{1}{(-\pi^2 x^4)^n} \\ B(x) &= \frac{1}{\pi x^2} + \sum_{n=1}^{\infty} \frac{(4n+1)!}{2^{2n}(2n)!} \cdot \frac{1}{(-\pi^2 x^4)^n} \end{aligned}$$

con el mismo criterio de truncamiento.

Calcular $C(x)$ y $S(x)$, para $x = 0.5, 3(0.5)$, por medio de las series y comparar los resultados utilizando la rutina de integración ROMBERG8 con diferentes valores de la tolerancia del error RELEERR= $10^{-10}, 10^{-12}, 10^{-14}$.

6 Ceros de funciones no lineales

6.1 Introducción

Cuando se tiene una ecuación no lineal, $f(x) = 0$, y se quiere resolver, muchas veces se está obligado a aplicar métodos numéricos, ya sea porque no existe ninguna fórmula que dé solución, ya porque, a pesar de que exista, es demasiado costoso ponerla en práctica.

Los métodos numéricos que se introducen en este capítulo están basados en la idea de aproximaciones sucesivas; entonces, se dice que se tiene un método iterativo, es decir, una sucesión x_0, x_1, \dots que se quiere que converja hacia la solución (raíz o cero de f) de la ecuación.

Por el teorema de Bolzano, cuando la función f sea continua, se evalúa en puntos a y b tales que $f(a) \cdot f(b) < 0$, y de este modo se asegura que la ecuación $f(x) = 0$ tiene por lo menos un cero real, α , dentro del intervalo (a, b) .

Algunas veces se ha de tener cuidado con los criterios de truncamiento del método como son

$$|x_{k+1} - x_k| \leq \varepsilon \quad \text{o} \quad |f(x_{k+1})| \leq \varepsilon$$

ya que pueden dar lugar a errores cuando los elementos de la sucesión son próximos, sin que la distancia a la raíz sea pequeña, o es posible que un punto alejado de la raíz tenga una imagen próxima a cero.

Después de introducir los métodos de resolución más conocidos para ecuaciones con una variable, se presentan los sistemas de ecuaciones no lineales con métodos de resolución y métodos de continuación dependiendo de un parámetro.

Finalmente, se estudian los métodos numéricos que se utilizan en el cálculo de raíces de polinomios, de los que previamente se han encontrado las cotas de las raíces y separado los intervalos donde el polinomio presenta las raíces.

6.2 Métodos de intervalos encajados

Se supone que f es continua en el intervalo $[a_0, b_0]$ y que $f(a_0) \cdot f(b_0) < 0$. Se elige el siguiente punto de la sucesión de manera que se obtenga una sucesión de intervalos encajados

$$[a_0, b_0] \supset [a_1, b_1] \supset [a_2, b_2] \supset \dots$$

que están en las hipótesis del teorema de Bolzano y, por lo tanto, todos contienen la solución de $f(x) = 0$. La convergencia será lenta, pero está asegurada.

6.2.1 Método de la bisección

Si se considera $f(a_0) < 0$ y $f(b_0) > 0$, los intervalos $I_k = [a_k, b_k]$, $k = 1, 2, 3, \dots$, son determinados de la forma siguiente: el punto medio del intervalo I_{k-1} es $m_k = \frac{a_{k-1} + b_{k-1}}{2}$, y si se supone que $f(m_k) \neq 0$ (ya que, si no, $m_k = \alpha$ que es la raíz y se ha terminado), se tiene

$$(a_k, b_k) = \begin{cases} (m_k, b_{k-1}) & \text{si } f(m_k) < 0 \\ (a_{k-1}, m_k) & \text{si } f(m_k) > 0 \end{cases}$$

De la construcción de I_k se deduce que $f(a_k) < 0$ y $f(b_k) > 0$; después de n pasos se tiene la raíz dentro del intervalo $I_n = [a_n, b_n]$ que tiene longitud

$$b_n - a_n = \frac{1}{2} (b_{n-1} - a_{n-1}) = \dots = \frac{1}{2^n} (b_0 - a_0)$$

Si se toma m_{n+1} como una aproximación de la raíz α , se tiene

$$\alpha = m_{n+1} \pm d_n, \quad \text{donde } d_n = \frac{1}{2^{n+1}} (b_0 - a_0),$$

que es una cota del error absoluto.

Ejemplo. Si se tiene la ecuación $f(x) = x - e^{1/x} = 0$, después de evaluar la función en 0, 1 y 2, se encuentra un cambio de signo para $f(1) < 0$ y $f(2) > 0$. Por tanto, se toma $I_0 = [1, 2]$ que da lugar a la tabla siguiente:

k	a_{k-1}	b_{k-1}	m_k	$f(m_k)$
1	1.	2.	1.5	—
2	1.5	2.	1.75	—
3	1.75	2.	1.875	+
4	1.75	1.875	1.8125	+
5	1.75	1.8125	1.78125	+
6	1.75	1.78125	1.765625	+
7	1.75	1.765625	1.7578125	—
8	1.7578125	1.765625	1.7617187	

La cota del error es $\frac{1}{2^9} \approx 0.00195$. La raíz con seis decimales correctos es 1.763223; por lo tanto, el error es del orden de 0.00150 y es necesario calcular por el método de la bisección hasta $k = 21$ para tener la aproximación con la precisión deseada de $0.5 \cdot 10^{-7}$.

Ejercicios

1. Encontrar los ceros de las siguientes funciones con dos decimales correctos:

a. $1 - x - e^{-2x}$
c. $(x + 1)e^{x-1} - 1$

b. $x^4 - 4x^3 + 2x^2 - 8$
d. $3x^2 + \operatorname{tg} x$

2. Las ecuaciones siguientes tienen una raíz en el intervalo $[0.3, 1.4]$. Determinadlas con un error inferior a 0.01.

a. $x \cos x = \ln x$

b. $2x - e^{-x} = 0$

c. $e^{-2x} = 1 - x$

6.2.2 Método de la Regula-Falsi

Se construye la recta que pasa por los puntos $(a_{k-1}, f(a_{k-1}))$ y $(b_{k-1}, f(b_{k-1}))$; se toma como nuevo punto c_k , que es el punto de corte con el eje $y = 0$:

$$c_k = a_{k-1} - f(a_{k-1}) \frac{b_{k-1} - a_{k-1}}{f(b_{k-1}) - f(a_{k-1})}$$

Seguidamente, se repite lo mismo que en el método de la bisección, eligiendo entre los intervalos $[a_{k-1}, c_k]$ y $[c_k, b_{k-1}]$, según el signo de $f(c_k)$ y que da lugar al intervalo $[a_k, b_k]$.

Ejemplo. Si se tiene la función $f(x) = x - e^{1/x}$ y se quiere calcular la raíz que presenta en $[1, 2]$ por el método de la Regula-Falsi, se tiene:

k	a_{k-1}	b_{k-1}	c_k	$f(c_k)$
1	1.	2.	1.830264	—
2	1.	1.830264	1.783184	—
3	1.	1.783184	1.769252	+
4	1.	1.769252	1.765052	+
5	1.	1.765052	1.763778	+
6	1.	1.763778	1.763392	+
7	1.	1.763392	1.763274	—
8	1.	1.763274	1.763238	—
9	1.	1.763238	1.763228	—
10	1.	1.763228	1.763224	—
11	1.	1.763224	1.763223	—

Han sido necesarias 11 iteraciones para llegar a la aproximación de la solución con la precisión deseada.

Dependiendo del tipo de función, es muy posible que la longitud de los intervalos de la sucesión decrezca muy lentamente y puede no tender a cero, como en el ejemplo anterior; esta situación se debe al hecho de que, en general, en un entorno del cero, la función mantiene su concavidad y el método consiste entonces en una aproximación por un extremo mientras que el otro permanece fijo; por esta razón, la longitud del intervalo no es muy útil como cota del error de aproximación. De todos modos, es un método más rápido que el de bisección e igual de seguro.

6.3 Métodos iterativos

Se presentan a continuación los dos métodos más conocidos y empleados para encontrar ceros de funciones de una forma iterada sin tener en consideración los signos de la función en los extremos del intervalo; ambos métodos tienen un sentido geométrico claro al aproximarse mediante la tangente y la secante del punto o de los puntos anteriormente calculados.

6.3.1 Método de Newton

Se comienza por una aproximación inicial x_0 y se construye una sucesión x_1, x_2, x_3, \dots , donde x_{n+1} se calcula de la forma siguiente: la función $f(x)$ se approxima por su recta tangente en el punto $(x_n, f(x_n))$ y x_{n+1} es la abscisa del punto de intersección de la tangente con el eje $y = 0$.

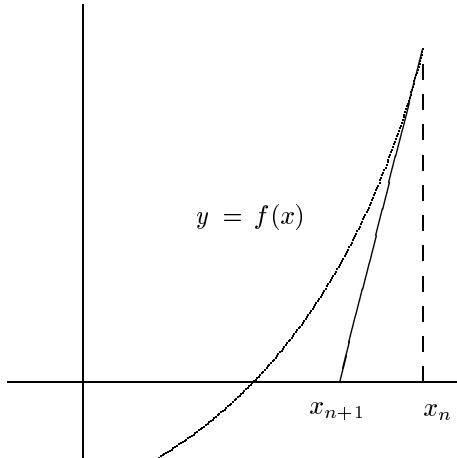
Entonces, si se quiere determinar x_{n+1} , se tiene

$$f(x_n) + (x_{n+1} - x_n) f'(x_n) = 0,$$

y despejando x_{n+1} , se obtiene la fórmula de iteración del método de Newton:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

La iteración puede pararse cuando el segundo sumando, la corrección, es menor que el error deseado en el momento de encontrar la raíz.



Ejemplo. Si se tiene la ecuación $x = e^{1/x}$, se considera la función $f(x) = x - e^{1/x}$ y se tiene $f'(x) = 1 + e^{1/x}/x^2$. Si se quiere calcular la solución con 6 decimales correctos, $|x_n - \alpha| \leq 0.5 \cdot 10^{-6}$, y se aplica el método con $x_0 = 1.5$, se tiene

n	x_n	$f(x_n)$	$x_n - \alpha$
0	1.5	-0.447734	-0.263223
1	1.739987	-0.0366406	-0.023236
2	1.763078	-0.000227662	-0.000145
3	1.763223	-0.871349D-08	0.000000
4	1.763223		

La última columna se ha podido llenar a causa de que se conocía la raíz. Nótese que sólo se necesitan 3 iteraciones para obtener la precisión deseada.

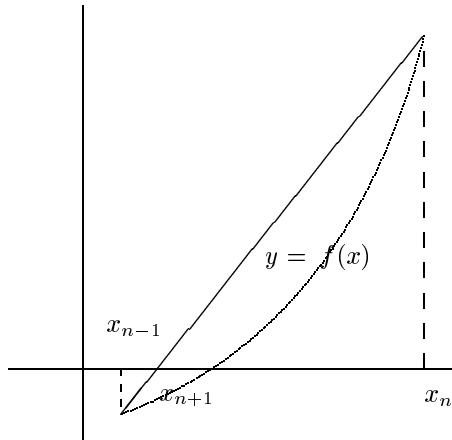
6.3.2 Método de la secante

El método de la secante se puede deducir del método de Newton approximando la derivada $f'(x_n)$ por el cociente $(f(x_n) - f(x_{n-1})) / (x_n - x_{n-1})$. La fórmula que aparece es equivalente a la del método de la Regula-Falsi, pero no se tienen en cuenta los signos en los extremos del intervalo. Sean x_0 y x_1 dos aproximaciones iniciales; se construye la sucesión x_2, x_3, \dots de forma recursiva:

$$x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, \quad \text{con } f(x_n) \neq f(x_{n-1}).$$

La interpretación geométrica es como sigue: el punto x_{n+1} está determinado por la abscisa del punto de intersección de la recta secante que pasa por los puntos: $(x_{n-1}, f(x_{n-1}))$ y $(x_n, f(x_n))$, y el eje $y = 0$.

Al inicio del método se necesitan dos aproximaciones, pero en cada paso sólo se evalúa la función en un punto.



Ejemplo. Si se toma la función $f(x) = x - e^{1/x}$, el método de la secante da lugar con aproximaciones iniciales de $x_0 = 1.0$ y $x_1 = 2.0$ a una sucesión de iterados que converge hacia la solución; para $n = 4$, se obtiene $x_4 = 1.763223$, que es la aproximación de la raíz con la precisión deseada.

Otra forma de expresar el método de iteración de la secante es

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_nf(x_{n-1})}{f(x_n) - f(x_{n-1})}$$

que tiene menos coste (menos cantidad de cálculo), pero que presenta dificultades cuando $x_n \approx x_{n-1}$ y $f(x_n) \cdot f(x_{n-1}) > 0$.

6.3.3 Métodos iterativos o del punto fijo

Dada la ecuación $f(x) = 0$, se puede expresar de la forma $x = g(x)$, donde g es una función continua. A partir de una aproximación inicial x_0 , se genera la sucesión

$$x_{n+1} = g(x_n)$$

Si $(x_n) \xrightarrow{n \rightarrow \infty} \alpha$, entonces $\alpha = \lim_{n \rightarrow \infty} x_{n+1} = \lim_{n \rightarrow \infty} g(x_n) = g(\alpha)$. Se dice que α es un cero de f o, también, un punto fijo de g .

Ejemplo. Si se tiene la ecuación $x^3 - x - 5 = 0$, se puede escribir de muchas maneras equivalentes, $x = g_i(x)$, como son:

$$g_1(x) = x^3 - 5, \quad g_2(x) = \sqrt[3]{x + 5} \quad o \quad g_3(x) = \frac{5}{x^2 - 1}.$$

La sucesión (x_n) puede no converger hacia α a pesar de que x_0 sea muy próximo a α , dependiendo de la elección de g .

Teorema 6.1. Se supone que la ecuación $x = g(x)$ tiene una solución α y en un entorno de α , J_α , existe $g'(x)$ y $|g'(x)| \leq m < 1$. Entonces, $\forall x_0 \in J_\alpha$, si se aplica el método iterativo simple $x_n = g(x_{n-1})$, se tiene:

- (a) $x_n \in J_\alpha$, $n = 0, 1, 2, \dots$
- (b) $\lim_{n \rightarrow \infty} x_n = \alpha$.
- (c) α es la única raíz de $x = g(x)$ en J_α .

Demostración: El apartado (a) se demuestra por el principio de inducción: Supóngase que $x_{n-1} \in J_\alpha$; por el teorema del valor medio

$$x_n - \alpha = g(x_{n-1}) - g(\alpha) = g'(c_n)(x_{n-1} - \alpha) \quad (6.1)$$

donde $c_n \in J_\alpha$. Entonces, $|x_n - \alpha| \leq m |x_{n-1} - \alpha|$ y, por lo tanto, $x_n \in J_\alpha$.

El apartado (b) se obtiene de considerar la desigualdad anterior

$$|x_n - \alpha| \leq m |x_{n-1} - \alpha| \leq \dots \leq m^n |x_0 - \alpha|$$

y, como que $m < 1$, $\lim_{n \rightarrow \infty} x_n = \alpha$.

Si se supone, finalmente, que $x = g(x)$ tiene otro punto fijo o raíz β , $\beta \neq \alpha$, $\beta \in J_\alpha$, entonces $\alpha - \beta = g(\alpha) - g(\beta) = g'(c)(\alpha - \beta)$ con $c \in J_\alpha$, y se tiene $|\alpha - \beta| \leq m |\alpha - \beta| < |\alpha - \beta|$ (contradicción). \square

Se tienen teoremas más generales de existencia en los teoremas 6.2 y 6.3 que se presentan más adelante.

Ejercicio. La ecuación $x^3 - x - 5 = 0$ tiene una raíz próxima a $x_0 = 1.9$. Comprobar que de las tres elecciones hechas en el ejemplo anterior, sólo cumple las hipótesis g_2 , y da lugar a una sucesión convergente.

El método de Newton se puede considerar como un método iterativo tomando $g(x) = x - \frac{f(x)}{f'(x)}$; si se considera el problema de puntos fijos en \mathbf{R}^2 , es decir, si se tiene una aplicación $G : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ y se toma

$$G \left(\begin{array}{c} x_1 \\ x_2 \end{array} \right) = \left(\begin{array}{c} x_2 \\ \frac{x_1 f(x_2) - x_2 f(x_1)}{f(x_2) - f(x_1)} \end{array} \right)$$

se tiene el método de la secante expresado en términos del método del punto fijo.

Ejemplo. Si se considera la ecuación $x - e^{1/x} = 0$, y se toma como función de iteración $g(x) = e^{1/x}$ y la aproximación inicial $x_0 = 1.5$, es necesario calcular 25 iteraciones para obtener la raíz con una precisión de $0.5 \cdot 10^{-6}$. Si se aplica el método de la secante a la misma función con $x_0 = 1.0$ y $x_1 = 2.0$, el cálculo se reduce a 5 iteraciones:

k	Iteración	Secante
1	1.947734	1.830264
2	1.670991	1.759565
3	1.819291	1.763286
4	1.732672	1.763223
5	1.780944	1.763223
:	:	
25	1.763223	

Ejercicio. Si a la ecuación $x = e^{1/x}$ se le suma x a los dos miembros y se despeja, se obtiene la función de iteración $g(x) = \frac{x + e^{1/x}}{2}$. Justificad por qué sólo se necesitan 9 iteraciones para tener la solución con la misma precisión que en el ejemplo anterior.

Estimación del error

Si se tiene en consideración los errores de redondeo, cuando se calcula $x_{n+1} = g(x_n)$, en realidad se genera una sucesión $\bar{x}_0, \bar{x}_1, \bar{x}_2, \dots$ donde se tiene $\bar{x}_{n+1} = g(\bar{x}_n) + \delta_n \quad n = 0, 1, 2, \dots$. Si se resta la ecuación anterior de $\alpha = g(\alpha)$, y se aplica el teorema del valor medio, se tiene

$$\bar{x}_{n+1} - \alpha = g'(c_n)(\bar{x}_n - \alpha) + \delta_n \quad c_n \in I(\bar{x}_n, \alpha)$$

Sumando y restando $g'(c_n) \bar{x}_{n+1}$, después de agrupar los términos se tiene

$$(1 - g'(c_n))(\bar{x}_{n+1} - \alpha) = g'(c_n)(\bar{x}_n - \bar{x}_{n+1}) + \delta_n$$

Si se supone que $|g'(c_n)| \leq m$ y que $\exists \delta$ tal que $|\delta_n| < \delta$, se tiene la desigualdad siguiente:

$$|\bar{x}_{n+1} - \alpha| < \frac{m}{1-m} |\bar{x}_{n+1} - \bar{x}_n| + \frac{1}{1-m} \delta$$

que da una estimación estricta del error de \bar{x}_{n+1} en función de cantidades conocidas. El primer sumando del segundo miembro da el error de truncamiento (aproximar un límite por un término de la sucesión) y el segundo el error de redondeo (error de cálculo que no aparece si se supone que la aritmética es exacta).

Ejemplo. El método iterativo $x_{n+1} = 1 - e^{-2x_n}$ da lugar a la sucesión siguiente

$$\bar{x}_1 = 0.798, \bar{x}_2 = 0.7973, \bar{x}_3 = 0.79701, \bar{x}_4 = 0.79689,$$

donde se ha redondeado en cada paso. Si se quiere calcular el error de considerar \bar{x}_4 como la raíz, se tiene en cuenta la desigualdad anterior para acotar el error. Como

$$g'(x) = 2e^{-2x} \quad y \quad |g'(x)| < |g'(0.795)| < 0.41,$$

se tiene $m = 0.41$ y $\delta = 0.5 \cdot 10^{-5}$; finalmente se puede calcular el error cometido

$$|\bar{x}_4 - \alpha| < \frac{0.41}{0.59} 12 \cdot 10^{-5} + \frac{1}{0.59} 0.5 \cdot 10^{-5} < 9.2 \cdot 10^{-5}$$

Ejercicio. Demostrar que, si la aritmética es exacta, se tiene la acotación siguiente:

$$|\bar{x}_n - \alpha| < \frac{m^n}{1-m} |\bar{x}_1 - \bar{x}_0|$$

Ejemplo. Buscar un cero de $h(x) = e^x - 3x$ es equivalente a calcular las x' s tales que $e^x / 3 = x$; por lo tanto, se toma $g(x) = e^x / 3$ y, en caso de que no fuera contractiva, se calcularía su inversa, ya que

$$(h^{-1})'(y) = \frac{1}{h'(x)}$$

dónde $y = h(x)$; y se tiene $|(h^{-1})'(y)| < 1$.

6.4 Orden de convergencia

Sea una sucesión de puntos $(x_n)_{n \in \mathbb{N}}$, que se puede pensar generada por $x_{n+1} = g(x_n)$. Sea α el límite de esta sucesión, es decir, un punto fijo de g : $g(\alpha) = \alpha$. Entonces, se dice que la sucesión y, en consecuencia, el método que la genera, tiene **orden de convergencia** por lo menos p , si, para cualquier punto $x_0 \in J_\alpha$, J_α entorno de α , existen $\nu \geq 0$ y $C > 0$, constantes, tales que

$$|x_{k+1} - \alpha| \leq C |x_k - \alpha|^p$$

para cualquier $k \geq \nu$; si $p = 1$, se ha de imponer $C < 1$.

En el caso que se obtenga

$$\lim_{n \rightarrow \infty} \frac{x_{n+1} - \alpha}{(x_n - \alpha)^p} = L$$

se dice que la sucesión tiene orden de convergencia por lo menos p ; si $p = 1$ es necesario que $|L| < 1$. Si $L \neq 0$, se dice que el orden es p y que L es la constante asintótica del error; además, para $p = 1, 2$ y 3 se dice que se tiene convergencia por lo menos lineal, cuadrática y cúbica, respectivamente.

Si se denota el error por $\varepsilon_k = |x_k - \alpha|$, se supone orden de convergencia por lo menos $p > 1$ y no se considera los errores de redondeo, se tiene

$$\varepsilon_{k+1} \leq C \varepsilon_k^p \leq C^{1+p} \varepsilon_{k-1}^{p^2} \leq C^{1+p+p^2} \varepsilon_{k-2}^{p^3} \leq \dots \leq C^{1+p+\dots+p^k} \varepsilon_0^{p^{k+1}}$$

De manera que, teniendo en cuenta que $1 + p + \dots + p^k = \frac{p^{k+1} - 1}{p - 1}$, se define $M = C^{\frac{1}{p-1}}$ y se puede escribir

$$\varepsilon_{k+1} \leq \frac{1}{M} (M \varepsilon_0)^{p^{k+1}}$$

tal que, cuanto mayor sea el orden de convergencia, más rápidamente convergerá la sucesión. Más adelante se verá que no es suficiente el estudio del orden; también es necesario saber la cantidad de cálculo en cada iteración y se tendrá un concepto nuevo, como es la eficiencia del método.

A continuación se estudia el orden de convergencia de los métodos presentados hasta ahora:

- Método de Newton. Se supone que $f \in \mathcal{C}^2(J_\alpha)$, J_α entorno de la raíz simple α y, por lo tanto, $f'(\alpha) \neq 0$ y $f'(x) \neq 0$, en un entorno de α .

Si se desarrolla la función f por Taylor en un entorno de x_n , se tiene

$$0 = f(\alpha) = f(x_n) + (\alpha - x_n) f'(x_n) + \frac{1}{2} (\alpha - x_n)^2 f''(c)$$

con c entre x_n y α . Dividiendo por $f'(x_n)$, se tiene

$$\frac{f(x_n)}{f'(x_n)} + \alpha - x_n = \alpha - x_{n+1} = \frac{-\frac{1}{2} (\alpha - x_n)^2 f''(c)}{f'(x_n)}$$

Tomando $\varepsilon_n = x_n - \alpha$, se obtiene

$$\varepsilon_{n+1} = \frac{1}{2} \varepsilon_n^2 \frac{f''(c)}{f'(x_n)}$$

Al calcular el límite cuando x_n tiende a α da convergencia cuadrática:

$$\frac{\varepsilon_{n+1}}{\varepsilon_n^2} \rightarrow \frac{1}{2} \frac{f''(\alpha)}{f'(\alpha)}$$

- Método de la secante. Se supone que $f \in \mathcal{C}^2(I_\alpha)$, I_α entorno de la raíz simple α y, por lo tanto, $f'(\alpha) \neq 0$ y $f'(x) \neq 0$ en un entorno de α .

A partir de la fórmula de interpolación de Newton se tiene

$$f(x) = f(x_n) + (x - x_n) [x_{n-1}, x_n] + \frac{1}{2} (x - x_n) (x - x_{n-1}) f''(c) \quad (6.2)$$

donde $[x_{n-1}, x_n]$ es la diferencia dividida de primer orden y $c \in \text{int}(x, x_{n-1}, x_n)$. Por otro lado, el método de la secante se puede escribir como

$$0 = f(x_n) + (x_{n+1} - x_n) [x_{n-1}, x_n] \quad (6.3)$$

Si se substituye $x = \alpha$ en 6.2 y se resta la ecuación 6.3, como $f(\alpha) = 0$, se tiene

$$(\alpha - x_{n+1}) [x_{n-1}, x_n] + \frac{1}{2} (\alpha - x_n) (\alpha - x_{n-1}) f''(c) = 0$$

Por el teorema del valor medio $[x_{n-1}, x_n] = f'(d)$, $d \in \text{int}(x_{n-1}, x_n)$, y se tiene

$$\varepsilon_{n+1} = \frac{f''(c)}{2 f'(d)} \cdot \varepsilon_n \cdot \varepsilon_{n-1}$$

Si se supone que el método converge, entonces, calculando el límite, c y d tienden a α . Se tiene

$$|\varepsilon_{n+1}| \approx L \cdot |\varepsilon_n| \cdot |\varepsilon_{n-1}| \quad (6.4)$$

con la misma L que en el método de Newton. Para determinar el orden de convergencia se supone que

$$|\varepsilon_{n+1}| \approx K |\varepsilon_n|^p, \quad |\varepsilon_n| \approx K |\varepsilon_{n-1}|^p$$

y, substituyendo en 6.4, se obtiene $K |\varepsilon_n|^p \approx L |\varepsilon_n| K^{-1/p} |\varepsilon_{n-1}|^{1/p}$ que se cumple si $p = 1 + \frac{1}{p}$;

es decir, si $p = \frac{1 + \sqrt{5}}{2} \approx 1.618 \dots$.

- El método iterativo simple, $x_{n+1} = g(x_n)$, suponiendo la función g contractiva, es en general un método de, por lo menos, orden 1. Para verlo, de la demostración del teorema, se tiene $x_n - \alpha = g'(c_n) (x_{n-1} - \alpha)$. Si se supone $g \in \mathcal{C}^1(J_\alpha)$, J_α entorno de la raíz α y que $g'(\alpha) \neq 0$ y $g'(x) \neq 0$ en un entorno de α , entonces

$$|\varepsilon_n| = |g'(c_n)| |\varepsilon_{n+1}| \implies \lim_{n \rightarrow \infty} \frac{|\varepsilon_{n+1}|}{|\varepsilon_n|} = L \neq 0$$

El método puede mejorar el orden si se toma $g \in \mathcal{C}^p(J_\alpha)$, J_α entorno de la raíz α y se supone que $g^{(j)}(\alpha) = 0$, $j = 1 \div p-1$, y $g^{(p)}(\alpha) \neq 0$; entonces, desarrollando por Taylor, se tiene

$$x_{n+1} = g(x_n) = \alpha + \frac{1}{p!} g^{(p)}(c_n) (x_n - \alpha)^p$$

con $c_n \in \text{int}(x_n, \alpha)$. Si $\lim_{n \rightarrow \infty} x_n = \alpha$, $\lim_{n \rightarrow \infty} \frac{|\varepsilon_{n+1}|}{|\varepsilon_n|^p} = \frac{1}{p!} |g^{(p)}(\alpha)| \neq 0$ y, entonces, se tiene que el método es de orden p para la raíz α .

Ejercicios.

1. Comprobar que el método de Newton, si $f'(\alpha) \neq 0$, tiene convergencia por lo menos cuadrática y que, si $f'(\alpha) = 0$, la tiene lineal.
2. Considerar el método de la Regula-Falsi con $f''(\alpha) \neq 0$; demostrar que tiene convergencia por lo menos lineal.
3. Sin entrar en consideraciones sobre el comportamiento de f , estudiar el orden de convergencia del método de la bisección.

Una idea que acompaña el concepto de orden es el de número de decimales correctos en cada iteración del método. Si se llama

$$u_n = -\log_{10} |\varepsilon_n| = \log_{10} |x_n - \alpha|$$

es decir, al número de decimales correctos, aproximadamente, y el orden es p ,

$$\lim_{n \rightarrow \infty} \frac{|\varepsilon_{n+1}|}{|\varepsilon_n|^p} = L \neq 0$$

entonces,

$$u_{n+1} = -\log_{10} |\varepsilon_{n+1}| \approx -p \log_{10} |\varepsilon_n| - \log_{10} L$$

En definitiva, $u_{n+1} \approx p \cdot u_n + \tilde{L}$ y, por lo tanto, en cada paso el número de decimales correctos en un método de orden p es multiplicado por p , aproximadamente.

6.5 Aceleración de la convergencia

Si la sucesión construida es convergente y se aproxima el método por $x_{k+1} - \alpha = \kappa(x_k - \alpha)$, cuando, en realidad, se tiene $x_{k+1} - \alpha = g'(c_n)(x_k - \alpha)$, repitiendo la hipótesis con el índice desplazado, se tiene $x_k - \alpha = \kappa(x_{k-1} - \alpha)$. Resolviendo en α , se obtiene

$$\alpha = \frac{x_{k-1}x_{k+1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}}$$

Como que el comportamiento de la sucesión no será, en general, exactamente lineal, se toma este cálculo como un elemento de una nueva sucesión $(x'_n)_{n \in \mathbb{N}}$, definido, como antes, por

$$x'_{k+1} = \frac{x_{k-1}x_{k+1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}} = x_{k+1} - \frac{(\Delta x_k)^2}{\Delta^2 x_{k-1}}$$

El valor de x'_{k+1} será más próximo a α que el término x_{k+1} . Este método se llama método Δ^2 de aceleración de Aitken y ha sido introducido también en la aceleración de las sucesiones de sumas parciales de una serie numérica convergente en el capítulo 1. Además, se tiene

$$\lim_{k \rightarrow \infty} \frac{x'_k - \alpha}{x_k - \alpha} = 0$$

que quiere decir que la nueva sucesión calculada converge más rápidamente y acelera la convergencia de métodos que presentan convergencia lineal. Una posible estrategia sería: a partir de un proceso $x_{k+1} = g(x_k)$ de primer orden y de unas iteraciones $x_1 = g(x_0)$, $x_2 = g(x_1)$, se introduce el cálculo de

$$x'_2 = x_2 - \frac{(\Delta x_1)^2}{\Delta^2 x_0}$$

Después se continúa con $x_3 = g(x'_2)$, $x_4 = g(x_3)$, y se volvería a aplicar el proceso Δ^2 de Aitken a la terna x'_2 , x_3 y x_4 .

Ejemplo. Si se toma la ecuación $x = 2 - e^x$, y se intenta resolver por el método de iteración simple con $g(x) = 2 - e^x$, g no es contractiva y el método no converge. Tomando la función inversa $g^{-1}(y) = \ln(2 - y)$, y volviendo a denotar por x la variable, se tiene el método iterativo simple $x_{k+1} = \ln(2 - x_k)$, con el que se obtiene la siguiente tabla de iteración, a partir de la aproximación inicial $x_0 = 0$:

$x_1 = 0.69315$	$x_{15} = 0.44336$
$x_2 = 0.26767$	$x_{16} = 0.44252$
$x_3 = 0.54948$	$x_{17} = 0.44306$
$x_4 = 0.37196$	$x_{18} = 0.44272$
$x_5 = 0.48755$	$x_{19} = 0.44296$
\vdots	$x_{20} = 0.44279$
$x_{12} = 0.44079$	$x_{21} = 0.44286$
$x_{13} = 0.44414$	$x_{22} = 0.44285$
$x_{14} = 0.44208$	$x_{23} = 0.44286$

Si se acelera desde x_{12} , x_{13} y x_{14} , se obtiene, con un único cálculo del método de aceleración de Aitken, el mismo resultado que se obtiene en x_{23} .

6.6 Métodos de interpolación y de Taylor

1. Método de interpolación directa. Sean $x_n, x_{n-1}, \dots, x_{n-k}$ los puntos de una sucesión que se approxima a α y se considera el polinomio interpolador de grado máximo k tal que

$$P_k(x_{n-j}) = f(x_{n-j}) = f_{n-j} \quad j = 0 \div k$$

El método consiste en definir x_{n+1} mediante $P_k(x_{n+1}) = 0$; es decir, como el cero de $P_k(x)$ más próximo a x_n . Y el proceso se repite tomando $x_{n+1}, x_n, \dots, x_{n-k+1}$ hasta que x_{n+r} sea suficientemente próximo a la solución.

Para $k = 1$, se tiene el método de la secante, y para $k = 2$, el método de Muller, que tiene orden 1.84. Para valores de k mayores, el cálculo de los ceros del polinomio P_k es costoso y normalmente no se utiliza.

2. Método de interpolación inversa. Sean $x_n, x_{n-1}, \dots, x_{n-k}$ los puntos de una sucesión que se approxima a α y se considera el polinomio interpolador de grado máximo k tal que

$$Q_k(f_{n-j}) = x_{n-j} \quad j = 0 \div k$$

Entonces, se define x_{n+1} por $x_{n+1} = Q_k(0)$, que será una mejor aproximación de α que x_n . El proceso se repite hasta donde se considere conveniente.

Para $k = 1$, se tiene el método de la secante, y para $k = 2$, el método de Brent, que tiene orden 1.84. Para valores de k mayores, el método es más sencillo de implementar que no el directo, pero la expresión de $Q_k(y)$ es pesada y no permite un cálculo rápido de x_{n+1} .

3. Método de Taylor directo. Sea $f : \mathbf{R} \rightarrow \mathbf{R}$ suficientemente derivable en un entorno J de la solución α . Entonces, si se desarrolla por Taylor alrededor del punto $x_k \in J$, se tiene

$$0 = f(\alpha) = f(x_k) + f'(x_k)(\alpha - x_k) + \frac{f''(x_k)}{2!}(\alpha - x_k)^2 + \dots$$

Si se trunca la serie en cierta potencia r de $(\alpha - x_k)$ se tiene

r = 1 $0 \approx f(x_k) + f'(x_k)(\alpha - x_k)$ y se tiene el método de Newton al despejar α :

$$\alpha \approx x_k - \frac{f(x_k)}{f'(x_k)}$$

r = 2 $0 \approx f(x_k) + f'(x_k)(\alpha - x_k) + \frac{f''(x_k)}{2!}(\alpha - x_k)^2$, de donde, despejando α , se obtiene

$$\alpha \approx x_k - \frac{f'(x_k) \pm \sqrt{f'(x_k)^2 - 2f(x_k)f''(x_k)}}{f''(x_k)}$$

Para $r > 2$, la ecuación en α no permite una solución sencilla. Se puede tomar el segundo miembro como una aproximación de α y si se denota por x_{k+1} se puede continuar el proceso hasta la precisión deseada.

4. Método de Taylor inverso. Sea $f : \mathbf{R} \rightarrow \mathbf{R}$ tal que admite derivadas hasta orden $m+2$ en un entorno J de la solución α y que α es una raíz simple de f . Entonces, f admite inversa en un entorno de α : $x = g(y)$ y, si se desarrolla por Taylor la función g alrededor del punto $y_k = f(x_k)$, se tiene

$$x = g(y) = x_k + \sum_{j=1}^{m+1} \frac{(y - y_k)^j}{j!} g^{(j)}(y_k) + \frac{(y - y_k)^{m+2}}{(m+2)!} g^{(m+2)}(d)$$

donde $d \in I(y, y_k)$ (mínimo intervalo que contiene a y y a y_k). Para $y = 0$, se tiene

$$\alpha = x_k + \sum_{j=1}^{m+1} \frac{(-1)^j}{j!} f_k^j g_k^{(j)} + \frac{(-1)^{m+2}}{(m+2)!} f_k^{m+2} g^{(m+2)}(d) \quad (6.5)$$

con $y_k = f(x_k) = f_k$ y $g^{(j)}(y_k) = g_k^{(j)}$. Si se menosprecia el término complementario, se tiene la siguiente familia de métodos iterativos:

$$x_{k+1} = x_k + \sum_{j=1}^{m+1} \frac{(-1)^j}{j!} f_k^j g_k^{(j)} \quad (6.6)$$

que depende de la facilidad de cálculo de $g_k^{(j)}$.

Si se resta 6.5 de 6.6, se tiene

$$x_{k+1} - \alpha = \frac{(-1)^{m+2}}{(m+2)!} f_k^{m+2} g^{(m+2)}(d)$$

y como, por el teorema del valor medio puede afirmar que $f_k = f(x_k) = f(x_k) - f(\alpha) = f'(c)(x_k - \alpha)$, con $c \in I(x_k, \alpha)$, se tiene

$$|\varepsilon_{k+1}| = |x_{k+1} - \alpha| = \frac{1}{(m+2)!} |[f'(c)]^{m+2} g^{(m+2)}(d)| \cdot |\varepsilon_k|^{m+2}$$

Tenemos así que 6.6 es un método de orden por lo menos $m+2$.

Para $m = 0$ se tiene el método de Newton, que es por lo menos de orden 2, como ya se había visto.

Para $m = 1$ se tiene el método de Chebyshev:

$$x_{k+1} = x_k - \frac{f_k}{f'_k} \left(1 + \frac{f_k f''_k}{2(f'_k)^2} \right)$$

que es por lo menos de orden 3.

Para $m = 2$ se tiene el método inverso de Taylor, de orden 4:

$$x_{k+1} = x_k - \frac{f_k}{f'_k} \left(1 + \frac{f_k f''_k}{2(f'_k)^2} - \frac{f_k^2}{6} \cdot \frac{f'_k f'''_k - 3(f''_k)^2}{(f'_k)^4} \right)$$

A pesar de ser un orden alto, se tiene el cálculo de muchas funciones y la eficiencia no es buena.

Ejemplo. Se va a calcular un cero de la ecuación $x - 3 \ln x = 0$ que pertenece al intervalo $(1, e)$. Para $x_0 = e/2$, los resultados se presentan en la tabla siguiente:

$m = 0$	$m = 1$	$m = 2$
$x_1 = 1.7224\dots$	$x_1 = 1.8111\dots$	$x_1 = 1.8585\dots$
$x_2 = 1.8454\dots$	$x_2 = 1.857116\dots$	$x_2 = 1.8571838\dots$

Con dos iteraciones el método de orden 4 da lugar a 7 decimales correctos, mientras que el método de Chebyshev da 4 y el de Newton 2. Se necesitan dos iteraciones más en el método de Newton y una más en el método de Chebyshev para obtener la misma precisión. Si se trabaja con un VAX 8600, el tiempo necesario para obtener la misma precisión de $0.5 \cdot 10^{-7}$ en μs es $42.73\mu s$ para $m = 0$, donde se realizan 4 iteraciones, $52.36\mu s$ con el método de Chebyshev y $56.15\mu s$, a pesar que sólo se han ejecutado dos iteraciones.

6.7 Eficiencia de un método iterativo

La elección entre la utilización de un método o de otro depende, en gran medida, del orden del método y del coste de la evaluación de cada paso de iteración. Por tanto, se tiene que considerar el número de evaluaciones de la función f y de sus derivadas en cada paso (suponiendo que la cantidad de computación necesaria para combinarlas es menoscable).

Si se supone que se tienen dos métodos, a y b , de órdenes p_a y p_b , donde los errores vienen dados por

$$\varepsilon_{k+1} \approx C_a \varepsilon_k^{p_a} \quad \eta_{k+1} \approx C_b \eta_k^{p_b}$$

donde $\varepsilon_k = x_k - \alpha$ y $\eta_k = y_k - \alpha$ y si se define $S_k = -\log_{10} |\varepsilon_k|$ y $T_k = -\log_{10} |\eta_k|$, se tiene

$$S_{k+1} = -\log_{10} C_a + p_a S_k \quad T_{k+1} = -\log_{10} C_b + p_b T_k$$

que son ecuaciones en diferencias con solución:

$$S_k = S_0 p_a^k - \log_{10} M_{a,k} \quad T_k = T_0 p_b^k - \log_{10} M_{b,k}$$

donde $M_{c,k} = C_c^{\frac{p_c^k - 1}{p_c - 1}}$, $c = a, b$. Si se toma la misma aproximación inicial x_0 , entonces $S_0 = T_0$ y, suponiendo que se ha llegado a la precisión deseada cuando $S_I = T_J$, se tiene

$$S_0 (p_a^I - p_b^J) + \log_{10} \left(\frac{M_{b,J}}{M_{a,I}} \right) = 0 \quad (6.7)$$

que relaciona el número de iteraciones y el orden de los dos métodos.

Sean θ_a y θ_b los costes per iteración (Fröberg y Traub los toman igual a 1 para f y para cualquier de las derivadas: [Fr 77], [Tra64]). Los costes totales vendrán dados por $\Theta_a = I \theta_a$ y $\Theta_b = J \theta_b$, y se tiene $\Theta_a = \frac{I}{J} \frac{\theta_a}{\theta_b} \Theta_b$.

Una estimación de I/J a partir de 6.7 no siempre es posible. Si se supone que el segundo sumando de 6.7 es muy pequeño respecto al primero y se puede eliminar (esto ocurre cuando C_a y C_b son próximos a la unidad), se tiene

$$p_a^I = p_b^J \implies \frac{I}{J} = \frac{\log p_b}{\log p_a}$$

Entonces,

$$\Theta_a = \left(\frac{\theta_a \log p_b}{\theta_b \log p_a} \right) \Theta_b$$

que sugiere el índice de eficiencia siguiente:

$$E = p^{1/\theta}$$

donde la unidad de θ se llama Horner.

Si se considera que el coste de evaluar f , f' , f'' , etc. es 1, se puede escribir la tabla siguiente:

Método	θ	p	E
Secante	1	1.62	1.62
Muller	1	1.84	1.84
Newton	2	2	1.41
Chebyshev	3	3	1.44

Si se tiene en cuenta la cantidad de trabajo necesario para el cálculo de $f'(x_k)$, c , y se quiere comparar el método de Newton con el método de la secante, donde se supone que $f''(\alpha) \neq 0$, se tiene

$$I \log 2 \approx J \log q \quad q = \frac{1 + \sqrt{5}}{2}$$

Si N iteraciones del método de Newton tienen un coste de $N(1+c)$ y S iteraciones del método de la secante tienen un coste de S , entonces el método de Newton será más eficiente que el de la secante, si

$$N(1+c) \leq S \approx N \frac{\log 2}{\log q}$$

y, por lo tanto, $c \leq 0.44$. Se tiene que el método de Newton será más eficiente cuando el coste de la derivada de f sea menor que 0.44 veces el coste de f .

6.8 Ceros múltiples

Se dice que α tiene multiplicidad q en la ecuación $f(x) = 0$, si $f(x) = (x - \alpha)^q g(x)$, con $0 \neq |g(\alpha)| < \infty$.

Si $f \in \mathcal{C}^q(U_\alpha)$, y α es una raíz de $f(x) = 0$ con multiplicidad q , entonces $f^{(j)}(\alpha) = 0$, $j < q$, y desarrollando por Taylor,

$$f(x) = \frac{1}{q!} (x - \alpha)^q f^{(q)}(c) \quad f'(x) = \frac{1}{(q-1)!} (x - \alpha)^{q-1} f^{(q)}(d)$$

con c y d entre x y α . Si se introduce la función

$$u(x) = \frac{f(x)}{f'(x)}$$

se tiene $\lim_{x \rightarrow \alpha} \frac{u(x)}{x - \alpha} = \frac{1}{q}$ y la ecuación $u(x) = 0$ tiene una raíz simple en $x = \alpha$, y todos los métodos anteriores pueden ser aplicados a esta ecuación. En particular, Newton es

$$x_{n+1} = x_n - \frac{u(x_n)}{u'(x_n)}, \quad \text{con } u'(x_n) = 1 - \frac{f''(x_n)}{f'(x_n)} \cdot u(x_n).$$

El método de la secante se escribe

$$x_{n+1} = x_n - u(x_n) \frac{x_n - x_{n-1}}{u(x_n) - u(x_{n-1})}$$

Queda claro que, para raíces simples, estos métodos son menos eficientes, ya que evalúan más derivadas y de órdenes mayores.

Ejemplo. Si se quiere calcular las raíces positivas del polinomio

$$f(x) = x^4 - 8.6 x^3 - 35.51 x^2 + 464.4 x - 998.46,$$

después de evaluar una muestra de puntos, se tiene un cero próximo a 7 y un posible cero doble en el intervalo [4, 5]. Si se toma $x_0 = 7.0$, después de 4 iteraciones ya se tiene $x_4 = 7.34847$. Para el cálculo de la otra raíz se toma el método de Newton para la función $u(x) = \frac{f(x)}{f'(x)}$, y con $x_0 = 4.0$, después de 3 iteraciones, se tiene la raíz $x_3 = 4.300000$; mientras que, con la misma aproximación inicial, el método de Newton sin modificar necesita 18 iteraciones para llegar a la misma precisión.

Otra técnica consiste en considerar un método de Newton modificado de la manera siguiente: se supone que α es una raíz de f con multiplicidad q y se define

$$x_{n+1} = x_n - q \frac{f(x_n)}{f'(x_n)}$$

entonces, $\alpha - x_{k+1} = \alpha - x_k + q u(x_k)$, y de aquí,

$$(\alpha - x_{k+1}) f'(x_k) = (\alpha - x_k) f'(x_k) + q f(x_k) \equiv G(x_k)$$

Tal como se ha definido la función G , se tiene

$$G^{(j)}(\alpha) = 0 \quad j = 0 \div q \quad G^{(q+1)}(\alpha) \neq 0$$

Desarrollando por Taylor $G(x)$ y f' en un entorno de α se tiene

$$G(x) = \frac{(x - \alpha)^{q+1}}{(q + 1)!} G^{(q+1)}(c_1) \quad f'(x) = \frac{(x - \alpha)^{q-1}}{(q - 1)!} f^{(q)}(c_2)$$

ya que α es una raíz de f con multiplicidad q , y c_1 y c_2 entre x y α . Substituyendo en la definición de G , se tiene

$$(\alpha - x_{k+1}) \frac{(x_k - \alpha)^{q-1}}{(q - 1)!} f^{(q)}(c_2) = \frac{(x_k - \alpha)^{q+1}}{(q + 1)!} G^{(q+1)}(c_1)$$

En definitiva, se tiene un método de por lo menos orden 2, siempre y cuando $f^{(q)}(x) \neq 0$ en un entorno de α :

$$\varepsilon_{k+1} = \frac{1}{q(q + 1)} \frac{G^{(q+1)}(c_1)}{f^{(q)}(c_2)} \varepsilon_k^2$$

Ejercicio. Demostrar que una forma de aproximar la multiplicidad desconocida de una raíz es la siguiente. De la fórmula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},$$

deducir que $\varepsilon_{n+2} = \left(1 - \frac{1}{q}\right) \varepsilon_{n+1} + O(\varepsilon_n^2)$. Por el método de aceleración de Aitken

$$\alpha - x_{n+2} = -\frac{(\Delta x_{n+1})^2}{\Delta^2 x_n}, \quad \alpha - x_{n+1} = \frac{-(\Delta x_n)^2}{\Delta^2 x_{n-1}}.$$

Así, una vez evaluados ε_{n+2} y ε_{n+1} , se puede aproximar q por $q \approx \frac{1}{1 - \frac{\varepsilon_{n+2}}{\varepsilon_{n+1}}}$.

6.9 Sistemas no lineales

6.9.1 Método de iteración simple

Sea una función $G : E \subset \mathbf{R}^n \rightarrow \mathbf{R}^n$ definida por

$$G(\mathbf{x}) = (G_1(x_1, \dots, x_n), \dots, G_n(x_1, \dots, x_n))^T, \quad \text{con } \mathbf{x} = (x_1, \dots, x_n)^T,$$

de la que se quiere estudiar los iterados sucesivos $\mathbf{x}^{(k+1)} = G(\mathbf{x}^{(k)})$ y ver si converge hacia un cierto $\mathbf{x}^* \in \mathbf{R}^n$ tal que $G(\mathbf{x}^*) = \mathbf{x}^*$, y que se llama punto fijo de la función G .

El concepto de orden de convergencia de una sucesión en \mathbf{R}^n es el mismo si se cambia el valor absoluto por alguna norma. Un teorema análogo al teorema 6.1 es el siguiente:

Teorema 6.2. Sea \mathbf{x}^* un punto fijo de $G : \mathbf{R}^n \rightarrow \mathbf{R}^n$, y se supone que G es contractiva en un entorno de radio r de \mathbf{x}^* : $S_r(\mathbf{x}^*) = \{\mathbf{x} \in \mathbf{R}^n / \|\mathbf{x} - \mathbf{x}^*\| < r\}$, es decir

$$\|G(\mathbf{x}) - G(\mathbf{y})\| \leq m \|\mathbf{x} - \mathbf{y}\|, \quad \text{con } 0 \leq m < 1 \quad \forall \mathbf{x}, \mathbf{y} \in S_r(\mathbf{x}^*);$$

entonces, $\forall \mathbf{x}_0 \in S_r(\mathbf{x}^*)$, la sucesión generada tiene las propiedades siguientes:

- (a) $\mathbf{x}^{(n)} \in S_r(\mathbf{x}^*) \quad \forall n \in \mathbf{N}$.
- (b) $\|\mathbf{x}^{(n+1)} - \mathbf{x}^*\| \leq m \|\mathbf{x}^{(n)} - \mathbf{x}^*\|$.
- (c) $\lim_{n \rightarrow \infty} \mathbf{x}^{(n)} = \mathbf{x}^*$.

Un teorema que asegura la existencia de un punto fijo es el siguiente:

Teorema 6.3. Sea $G : \mathbf{R}^n \rightarrow \mathbf{R}^n$, $S_r(\mathbf{x}^{(0)})$ un entorno centrado en $x_0 \in \mathbf{R}^n$ de radio r y $0 < m < 1$ una constante tal que

1. G es contractiva con constante $m \quad \forall \mathbf{x}, \mathbf{y} \in \overline{S_r(\mathbf{x}^{(0)})}$.
2. $\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \leq (1 - k) r < r$; entonces,
 - (a) $\mathbf{x}^{(n)} \in S_r(\mathbf{x}^{(0)})$, $\forall n \in \mathbf{N}$.
 - (b) G tiene un único punto fijo $\mathbf{x}^* \in \overline{S_r(\mathbf{x}_0)}$.
 - (c) $\lim_{n \rightarrow \infty} \mathbf{x}^{(n)} = \mathbf{x}^*$.
- (d) $\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| \leq m \|\mathbf{x}^{(n)} - \mathbf{x}^*\|$. Además, $\|\mathbf{x}^{(n)} - \mathbf{x}^*\| \leq \frac{m^n}{1-m} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$.

Demostración: Para probar (a) se hace por inducción: se sabe que $\mathbf{x}^{(1)} \in S_r(\mathbf{x}^{(0)})$ y, suponiendo que $\mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)} \in S_r(\mathbf{x}^{(0)})$, se tiene

$$\|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| = \|G(\mathbf{x}^{(n)}) - G(\mathbf{x}^{(n-1)})\| \leq m \|\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}\| \leq m^n \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$$

y, por lo tanto,

$$\begin{aligned} \|\mathbf{x}^{(n+1)} - \mathbf{x}^{(0)}\| &\leq \|\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}\| + \|\mathbf{x}^{(n)} - \mathbf{x}^{(n-1)}\| + \cdots + \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \\ &\leq (m^n + m^{n-1} + \cdots + 1) \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \\ &\leq \sum_{i=0}^n m^i (1-m) r = (1 - m^{n+1}) r < r \end{aligned}$$

Para demostrar (b) y (c) se verá, en primer lugar, que la sucesión $(\mathbf{x}^{(n)})_{n \in \mathbf{N}}$ es de Cauchy:

$$\begin{aligned} \|\mathbf{x}^{(k)} - \mathbf{x}^{(l)}\| &\leq \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| + \|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)}\| + \cdots + \|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\| \\ &\leq m^l (1 + m + \cdots + m^{k-l-1}) \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \\ &< \frac{m^l}{1-m} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| < m^l r \end{aligned}$$

Por tanto, es convergente hacia un $\mathbf{x}^* \in \overline{S_r(\mathbf{x}^{(0)})}$, que es punto fijo de G :

$$\begin{aligned}\|G(\mathbf{x}^*) - \mathbf{x}^*\| &\leq \|G(\mathbf{x}^*) - G(\mathbf{x}^{(n)})\| + \|G(\mathbf{x}^{(n)}) - \mathbf{x}^*\| \\ &\leq m \|\mathbf{x}^* - \mathbf{x}^{(n)}\| + \|\mathbf{x}^{(n+1)} - \mathbf{x}^*\| < \varepsilon\end{aligned}$$

Así, se tiene $\|G(\mathbf{x}^*) - \mathbf{x}^*\| \leq 0$, que implica $G(\mathbf{x}^*) = \mathbf{x}^*$. Además, este punto fijo es único ya que, si existiera otro $\bar{\mathbf{x}}^*$, se tendría

$$\|\mathbf{x}^* - \bar{\mathbf{x}}^*\| = \|G(\mathbf{x}^*) - G(\bar{\mathbf{x}}^*)\| \leq m \|\mathbf{x}^* - \bar{\mathbf{x}}^*\|$$

y, como $m > 0$, se obtiene $\mathbf{x}^* = \bar{\mathbf{x}}^*$.

La demostración de (d) son dos líneas:

$$\|\mathbf{x}^* - \mathbf{x}^{(l)}\| \leq \lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)} - \mathbf{x}^{(l)}\| \leq \frac{m^l}{1-m} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|$$

$$\|\mathbf{x}^{(n+1)} - \mathbf{x}^*\| = \|G(\mathbf{x}^{(n)}) - G(\mathbf{x}^*)\| \leq m \|\mathbf{x}^{(n)} - \mathbf{x}^*\| \quad \square$$

Cuando la función G es diferenciable se tiene un resultado debido a Ostrowski ([Ort70]):

Teorema 6.4. Si $G : \mathbf{R}^n \rightarrow \mathbf{R}^n$ es diferenciable en un entorno centrado en $\mathbf{x}^* \in \mathbf{R}^n$, punto fijo de G , y el radio espectral de la diferencial $\rho(DG(\mathbf{x}^*)) < 1$, entonces existe un entorno del punto fijo donde para cualquier aproximación inicial $\mathbf{x}^{(0)}$ perteneciente a este entorno, sus iterados sucesivos convergen hacia \mathbf{x}^* .

Ejemplo. Si se utiliza el método iterativo para resolver el siguiente sistema no lineal

$$\begin{aligned}x_1 &= \operatorname{sen}(x_1 + x_2) \\ x_2 &= \cos(x_1 - x_2)\end{aligned}$$

se tiene

$$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{pmatrix} = \begin{pmatrix} \operatorname{sen}(x_1^{(k)} + x_2^{(k)}) \\ \cos(x_1^{(k)} - x_2^{(k)}) \end{pmatrix}$$

y da lugar a la tabla siguiente

k	x_1	x_2	$d_k = \ \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\ _\infty$	d_k / d_{k-1}
0	1.	1.		
1	0.9092974268	1.	$9.1 \cdot 10^{-2}$	
2	0.9932534701	0.9958893410	$3.4 \cdot 10^{-2}$	0.37
3	0.9329240076	0.9986150524	$1.0 \cdot 10^{-2}$	0.24
4	0.9356349198	0.9978431191	$3.0 \cdot 10^{-3}$	0.30
5	0.9349487614	0.9980656939	$6.9 \cdot 10^{-4}$	0.23
6	0.9351131325	0.9980087876	$1.6 \cdot 10^{-4}$	0.23
7	0.9350750470	0.9980227202	$3.8 \cdot 10^{-5}$	0.24
8	0.9350836077	0.9980194493	$8.6 \cdot 10^{-6}$	0.23
9	0.9350817328	0.9980201935	$1.9 \cdot 10^{-6}$	0.22

Se puede hacer una estimación de la constante asintótica del error L tomando

$$L = \lim_{k \rightarrow \infty} \frac{\|\varepsilon_k\|}{\|\varepsilon_{k-1}\|} \quad \text{como} \quad \frac{d_k}{d_{k-1}},$$

que da $L \approx 0.24$. Si se supone una aritmética exacta,

$$\|\varepsilon_k\| \leq d_{k+1} \frac{1}{1-L} \leq \frac{L}{1-L} d_k = 0.32 d_k$$

Entonces, $\|\varepsilon_9\| \leq 0.32 \cdot 1.9 \cdot 10^{-6} < 10^{-6}$ y $\mathbf{x}^{(9)} = (0.935082, 0.998020)^T$ será la solución con un error menor que 10^{-6} .

Una condición suficiente que puede substituir la del radio espectral menor que 1 en el teorema 6.4 es que G tenga una matriz jacobiana $DG(x)$ tal que $\|DG(x)\| \leq m < 1$ en un entorno del punto fijo. Esta norma matricial tiene que ser consistente con la norma vectorial con la que se trabaja (ver el apéndice A).

6.9.2 Método de Newton

Sea $F : E \subset \mathbf{R}^n \rightarrow \mathbf{R}^n$ que da lugar al siguiente sistema de ecuaciones no lineales: $F(\mathbf{x}) = 0$. También se puede escribir

$$\begin{cases} F_1(x_1, \dots, x_n) = 0 \\ F_2(x_1, \dots, x_n) = 0 \\ \vdots \\ F_n(x_1, \dots, x_n) = 0 \end{cases}$$

Si se supone que F es diferenciable con continuidad y se desarrolla por Taylor en un punto próximo, $\mathbf{x}^{(k)}$ de la solución \mathbf{x}^* , se tiene

$$F(\mathbf{x}) = F(\mathbf{x}^{(k)}) + DF(\mathbf{x}^{(k)}) \cdot (\mathbf{x} - \mathbf{x}^{(k)}) + R_1$$

y, por tanto, si se evalúa F en \mathbf{x}^* , y no se considera el término complementario, se obtiene

$$0 = F(\mathbf{x}^*) \approx F(\mathbf{x}^{(k)}) + DF(\mathbf{x}^{(k)}) \cdot (\mathbf{x}^* - \mathbf{x}^{(k)})$$

es decir,

$$\mathbf{x}^* \approx \mathbf{x}^{(k)} - \left(DF(\mathbf{x}^{(k)})\right)^{-1} \cdot F(\mathbf{x}^{(k)})$$

donde $DF(\mathbf{x}^{(k)})$ es la matriz jacobiana de F en $\mathbf{x}^{(k)}$:

$$DF(x) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \dots & \frac{\partial F_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial F_n}{\partial x_1} & \dots & \frac{\partial F_n}{\partial x_n} \end{pmatrix}$$

Para que tenga sentido el método de Newton es necesario que el jacobiano sea no nulo en un entorno de la solución que se calcula y, en definitiva, se tiene

$$\mathbf{x}^{(k+1)} \approx \mathbf{x}^{(k)} - \left(DF(\mathbf{x}^{(k)})\right)^{-1} \cdot F(\mathbf{x}^{(k)})$$

Normalmente se siguen los pasos siguientes:

1. Se resuelve el sistema lineal $DF(\mathbf{x}^{(k)}) \cdot \mathbf{y}^{(k)} = -F(\mathbf{x}^{(k)})$
2. Finalmente, se calcula $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{y}^{(k)}$

Dificultades del método:

a) Se tiene que calcular la matriz jacobiana en cada paso de iteración, es decir, hay n^2 derivadas parciales. Si n es grande y/o las derivadas son complicadas, el cálculo “a mano” es imposible; entonces se utilizan técnicas de diferenciación simbólica o de aproximación por cocientes incrementales, que sólo necesitan evaluar las funciones $F_i \ i = 1 \div n$ (ver el capítulo 5):

$$\frac{\partial F_i}{\partial x_j}(\mathbf{x}) \approx \frac{1}{h} [F_i(x_1, \dots, x_j + h, \dots, x_n) - F_i(\mathbf{x})]$$

El coste se puede reducir si se calcula la matriz jacobiana cada l pasos y no en cada iteración:

1. Evaluación de $DF(\mathbf{x}^0)$
 2. Cálculo de $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - (DF(\mathbf{x}^0))^{-1} F(\mathbf{x}^{(k)}) \quad k = 0 \div l$
 3. Evaluación de $DF(\mathbf{x}^{(l+1)})$
 4. Cálculo de $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - (DF(\mathbf{x}^{(l+1)}))^{-1} F(\mathbf{x}^{(k)}) \quad k = l + 1 \div 2l + 1$
- \vdots

b) Esta modificación del método también aligera la segunda desventaja, que consiste en el hecho de tener que resolver un sistema de ecuaciones lineales en cada paso. Así, en el transcurso de l pasos se tiene una misma matriz del sistema que puede ser factorizada ($DF = LU$) una sola vez.

c) Si en la fórmula de Taylor

$$DF(\mathbf{x}^{(k)})(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^{(k)}) + O(\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2)$$

se considera

$$\mathbf{s}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}, \quad \text{y} \quad \mathbf{y}^{(k)} = F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^{(k)})$$

se tiene la ecuación llamada de la secante $A \mathbf{s}^{(k)} = \mathbf{y}^{(k)}$ que se cumple, en primer orden, para $A = DF(\mathbf{x}^{(k)})$. Entonces, una formulación tipo Newton es

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - A_k^{-1} F(\mathbf{x}^{(k)})$$

donde se ha calculado una aproximación de la matriz jacobiana $A_k \approx DF(\mathbf{x}^{(k)})$. El método de Broyden (1965) (consultar [Sto80], [All90]) consiste en imponer la aproximación siguiente de la matriz jacobiana A_{k+1} que cumpla $A_{k+1} \mathbf{s}^{(k)} = \mathbf{y}^{(k)}$, donde $\mathbf{s}^{(k)}$ y $\mathbf{y}^{(k)}$ ya son conocidos. Esta ecuación no determina únicamente la matriz y se impone que esté a distancia mínima de A_k :

$$A_{k+1} = \min\{\|A - A_k\|_F \mid A \mathbf{s}^{(k)} = \mathbf{y}^{(k)}\}$$

donde $\|\cdot\|_F$ es la norma definida por $\|A\|_F = \left(\sum_{i,j=1}^n a_{ij}^2 \right)^{1/2}$. Entonces, el cálculo de A viene dado por

$$A_{k+1} = A_k + \frac{\mathbf{y}^{(k)} - A_k \mathbf{s}^{(k)}}{\|\mathbf{s}^{(k)}\|^2} (\mathbf{s}^{(k)})^H$$

La convergencia hacia $\mathbf{x}^* \in \mathbf{R}^n$ de la sucesión generada de este modo es superlineal:

$$\lim_{n \rightarrow \infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|} = 0$$

- d) La convergencia del método depende de la aproximación inicial \mathbf{x}^0 . En general, no es posible encontrar un aproximación inicial suficientemente “buena”; entonces se usan técnicas de continuación.

6.9.3 Métodos de continuación

Si se quieren encontrar las soluciones de $G(\mathbf{x}) = \mathbf{0}$ con $G : \mathcal{U} \subset \mathbf{R}^n \rightarrow \mathbf{R}^n$, donde \mathcal{U} es un abierto de \mathbf{R}^n y no se conoce ninguna proximación inicial para aplicar un método iterativo, se puede construir

$$F(\mathbf{x}, \mu) = G(\mathbf{x}) \mu + H(\mathbf{x}) (1 - \mu) \quad \mu \in [0, 1]$$

donde $H(\mathbf{x})$ es una función cuyos ceros son conocidos o pueden ser calculados fácilmente ($G(\mathbf{x}) - G(\mathbf{x}^0)$, por ejemplo), y se tiene un problema donde se quiere encontrar $\mathbf{x}(\mu)$, con $0 \leq \mu \leq 1$, a partir de $F(\mathbf{x}, \mu) = \mathbf{0}$, donde $F : \mathcal{U} \times I \rightarrow \mathbf{R}^n$, y I es el intervalo de variación de μ .

Cuando μ varía de 0 a 1, entonces $F(\mathbf{x}, \mu)$ varía de $H(\mathbf{x})$, donde se conocen los ceros, hasta $G(\mathbf{x})$, que es la función de la cual se quiere encontrar los ceros. Si se conoce la solución \mathbf{x}_k^* tal que $F(\mathbf{x}_k^*, \mu_k) = \mathbf{0}$, se intenta resolver $F(\mathbf{x}, \mu_{k+1}) = \mathbf{0}$ por un método iterativo tomando como aproximación inicial del sistema precisamente $\mathbf{x}_{k+1}^{(0)} = \mathbf{x}_k^*$; para $\mu_{k+1} - \mu_k$ pequeño, es de esperar que $\mathbf{x}_{k+1}^* - \mathbf{x}_k^*$ sea pequeño y el método sea convergente.

Se supone que $F \in \mathcal{C}^1(\mathcal{U} \times I)$ y se amplía el vector con una nueva componente $x_0 = \mu$; por lo tanto, se tiene

$$\mathbf{y} = (x_0, x_1, \dots, x_n)^T \quad \text{y} \quad F(\mathbf{y}) = \mathbf{0}.$$

Si se diferencia, se tiene el sistema lineal homogéneo $DF(\mathbf{y}) \cdot \Delta\mathbf{y} = \mathbf{0}$, donde $DF(\mathbf{y})$ es una aplicación lineal de \mathbf{R}^{n+1} en \mathbf{R}^n . Sea A_j el menor obtenido de la matriz jacobiana asociada a $DF(\mathbf{y})$ habiendo eliminado la j -ésima columna y multiplicado por $(-1)^j$ con $j = 0 \div n$.

$$DF(\mathbf{y}) \leftrightarrow \begin{pmatrix} \frac{\partial F^1}{\partial x_0} & \frac{\partial F^1}{\partial x_1} & \cdots & \frac{\partial F^1}{\partial x_n} \\ \frac{\partial F^2}{\partial x_0} & \frac{\partial F^2}{\partial x_1} & \cdots & \frac{\partial F^2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial F^n}{\partial x_0} & \frac{\partial F^n}{\partial x_1} & \cdots & \frac{\partial F^n}{\partial x_n} \end{pmatrix}$$

Entonces, la solución del sistema homogéneo viene dada por

$$\frac{\Delta x_0}{A_0} = \frac{\Delta x_1}{A_1} = \cdots = \frac{\Delta x_n}{A_n}$$

que, introduciendo el arco como parámetro, al definir Δs como el elemento de longitud de arco en \mathbf{R}^{n+1} tal que

$$(\Delta s)^2 = \sum_{j=0}^n (\Delta x_j)^2$$

y si se pasa al límite, se tiene el sistema diferencial

$$\frac{dx_j}{ds} = \frac{\pm A_j}{\sqrt{\sum_{k=0}^n A_k^2}} \quad j = 0 \div n$$

El signo da el sentido que se quiere seguir a partir de los puntos fijos. El método falla si todos los A_j son nulos ($\text{rang}(DF(\mathbf{y})) < n$). En particular, si $A_0 = 0$, pero hay otro A_k no nulo, se puede intercambiar la función del parámetro y continuar el cálculo.

Para la integración del sistema puede utilizarse el método de Euler o algún método explícito multipaso lineal de orden bajo (≤ 4) como es el de Adams-Bashforth (ver capítulo 8). Se dirá que se tiene una predicción de la solución y se toma como una aproximación inicial para resolver $F(\mathbf{y}) = \mathbf{0}$, que se resolverá con un método de Newton modificado, ya que el número de variables es mayor que el de ecuaciones:

Sea $F : \mathcal{U} \subset \mathbf{R}^m \rightarrow \mathbf{R}^n$ con $m > n$; se quiere encontrar una solución de $F(\mathbf{z}) = \mathbf{0}$ y se tiene una aproximación $\mathbf{z}^{(0)}$. Si se desarrolla por Taylor hasta orden 1 en un entorno de $\mathbf{z}^{(k)}$ y se toma $\mathbf{z} = \mathbf{z}^{(k+1)}$, se tiene

$$F(\mathbf{z}^{(k)}) + DF(\mathbf{z}^{(k)}) \cdot \Delta\mathbf{z}^{(k)} = \mathbf{0}$$

añadiendo la condición de que la norma euclíadiana, $\|\Delta\mathbf{z}^{(k)}\|_2$, sea mínima, se tiene un problema de extremos condicionados con función de Lagrange

$$L(\Delta\mathbf{z}, \lambda) = \langle \Delta\mathbf{z}, \Delta\mathbf{z} \rangle + \lambda^T (F(\mathbf{z}) + DF(\mathbf{z}) \cdot \Delta\mathbf{z})$$

donde λ es un multiplicador de Lagrange n -dimensional. Entonces se tiene la recurrencia

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} - DF(\mathbf{z}^{(k)})^T \left(DF(\mathbf{z}^{(k)}) \cdot \left(DF(\mathbf{z}^{(k)}) \right)^T \right)^{-1} \cdot F(\mathbf{z}^{(k)})$$

que no presenta ningún problema si $\text{rang}(DF(\mathbf{z}^{(k)})) = n$.

Davidenko propone derivar la ecuación $F(\mathbf{x}, \mu) = 0$ respecto al parámetro μ , obteniéndose el sistema diferencial

$$\frac{d\mathbf{x}}{d\mu} = -(D_{\mathbf{x}}F)^{-1} D_{\mu}F$$

con condiciones iniciales $\mu = 0$ y $\mathbf{x} = \mathbf{0}$. No es el método más eficiente debido a la posibilidad de catástrofe, es decir, de que no exista la matriz inversa de $D_{\mathbf{x}}F$.

Ejemplo. Se quiere calcular un punto fijo de la función $F(\mathbf{z}) = (F^1(z), \dots, F^{10}(z))^T$, definida por

$$F^j(\mathbf{z}) = \exp\{\cos[j * (z_1 + \dots + z_{10})]\} \quad j = 1 \div 10$$

Se considera la función auxiliar $H : \mathbf{R}^{11} \rightarrow \mathbf{R}^{10}$ definida por $H(\mathbf{z}, \lambda) = \mathbf{z} - \lambda F(\mathbf{z})$; para $(\mathbf{z}, \lambda) = (0, 0)$, se tiene una solución trivial, y se quiere continuar esta solución al variar $\lambda = 0 \div 1$, ya que, para $\lambda = 1$, se tiene la ecuación del punto fijo $F(\mathbf{z}) = \mathbf{z}$ que se quiere resolver. El proceso se entenderá cuando $(\mathbf{z}, \lambda) = (\mathbf{z}^*, 1)$, donde \mathbf{z}^* es el punto fijo.

El método de continuación considerado consiste en la resolución por Newton del sistema de ecuaciones $H(\mathbf{z}, \lambda_k) = 0$ (método corrector), la predicción de un punto aproximado nuevo a partir de los calculados por medio del método de Adams–Bashforth de orden máximo 4 y una atención especial al cambio de variable independiente cuando el rango de la matriz $\left(\frac{\partial H^j}{\partial z_i} \right)$, ($i, j = 1 \div 10$) no sea 10.

Los resultados obtenidos para $\lambda = 1$ son

z_1	1.4919 134	z_6	2.1869 661
z_2	0.5066 654	z_7	0.7729 181
z_3	0.3890 434	z_8	0.3720 929
z_4	0.9273 171	z_9	0.5865 923
z_5	2.4198 068	z_{10}	1.75384 404

6.10 Cálculo de las raíces de polinomios

A pesar de que todo lo que se ha dicho hasta ahora puede ser aplicado al cálculo de ceros de polinomios, se dedica este apartado a métodos mucho más específicos, como también a estudiar la relación entre los coeficientes del polinomio y sus raíces, a acotar las raíces y separarlas y, finalmente, se introducen técnicas de resolución de ecuaciones polinómicas, así como un método de Newton modificado.

6.10.1 Relación entre raíces y coeficientes

Sea $P(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n$ un polinomio a coeficientes reales con raíces reales x_1, x_2, \dots, x_n ; entonces $P(x)$ se puede descomponer $P(x) = a_0(x-x_1)(x-x_2) \cdots (x-x_n)$ y, agrupando términos, se tiene

$$P(x) = a_0 \left[x^n - x^{n-1} \sum_{i=1}^n x_i + x^{n-2} \sum_{1 \leq i < j \leq n} x_i \cdot x_j - \cdots + (-1)^n x_1 \cdot x_2 \cdots x_n \right]$$

Si se toma $\sum_{i=1}^n x_i^2 = \left(\sum_{i=1}^n x_i \right)^2 - 2 \sum_{i < j} x_i \cdot x_j = \left(\frac{a_1}{a_0} \right)^2 - 2 \left(\frac{a_2}{a_0} \right)$ y si x_m es la raíz mayor en valor absoluto, se puede escribir

$$|x_m| \leq \frac{\sqrt{a_1^2 - 2 a_0 a_2}}{a_0}$$

Ejemplo. Si se considera el polinomio $128 x^4 - 256 x^3 + 160 x^2 - 32 x + 1$, y se aplica la acotación anterior, sabiendo que sus todas sus raíces son reales, se tiene una cota superior $|x_m| \leq 1.225$ (en realidad, la raíz mayor es $\alpha \approx 0.9619$).

6.10.2 Acotación de las raíces

Las dos reglas que se presentan son aplicables cuando los ceros son reales.

1. **Regla de Laguerre:** Si, cuando se divide $P(x)$ por $x-L$, con $L > 0$, todos los coeficientes del polinomio cociente y el resto son positivos, entonces L es cota superior de las raíces de $P(x)$.

Demostración: Si se divide $P(x) = (x - L) (c_0 x^{n-1} + c_1 x^{n-2} + \cdots + c_{n-2} x + c_{n-1}) + r$ con $c_i > 0$, $i = 0 \div n-1$ y $r > 0$, entonces, para cualquier $x > L$, se tiene $P(x) > 0$ y, por lo tanto, no hay raíces mayores que L . \square

Ejemplo. Dado el polinomio $P(x) = 5x^4 - 8x^3 - x^2 + x - 6$, se toman valores enteros de L y crecientes; entonces, por el método de Ruffini

$\begin{array}{r rrrr r} 5 & -8 & -1 & 1 & & -6 \\ \hline 2 & 10 & 4 & 6 & & 14 \\ \hline & 5 & 2 & 3 & & 8 \end{array}$	$\begin{array}{r rrrr r} 6 & -1 & 1 & 8 & & -5 \\ \hline 1 & 6 & 5 & 6 & & 14 \\ \hline & 6 & 5 & 6 & & 9 \end{array}$
--	--

Después de comprobar que 2 es una cota superior de las raíces del polinomio, se hace el cambio $y = \frac{1}{x}$ y se aplica la regla de Laguerre al polinomio $Q(y) = 6y^4 - y^3 + y^2 + 8y - 5$, donde se han cambiado todos los signos del polinomio resultante. En definitiva, las raíces positivas pertenecen al intervalo $(1, 2)$. Para calcular un intervalo para las raíces negativas, el cambio que se realiza es $z = -x$, que da lugar al polinomio $R(z) = 5z^4 + 8z^3 - z^2 - z - 6$.

$$\begin{array}{c|cccc|c} & 5 & 8 & -1 & -1 & -6 \\ \hline 1 & & 5 & 13 & 12 & 11 \\ & 5 & 13 & 12 & 11 & 5 \end{array}$$

$$\begin{array}{c|cccc|c} & 6 & 1 & 1 & -8 & -5 \\ \hline 2 & & 12 & 26 & 54 & 92 \\ & 6 & 13 & 27 & 46 & 87 \end{array}$$

Se obtiene así, después del cambio $w = \frac{1}{z}$, el intervalo $(-1, -0.5)$.

2. Regla de Newton: Si $L > 0$ es tal que $P(L), P'(L), \dots, P^{(n)}(L)$ son positivos, entonces L es cota superior de las raíces de $P(x)$.

Demostración: Si se desarrolla $P(x)$ en potencias de $x - L$, se tiene

$$P(x) = P(L) + (x - L) P'(L) + \frac{1}{2} (x - L)^2 P''(L) + \dots + \frac{1}{n!} (x - L)^n P^{(n)}(L)$$

y, dando a x valores mayores que L , se tiene $P(x) > 0$; por lo tanto, no hay raíces de $P(x)$ superiores a L . \square

Ejemplo. Dado el polinomio $P(x) = 5x^4 - 8x^3 - x^2 + x - 6$, se calcula la sucesión de polinomios derivados

$$\begin{aligned} P(x) &= 5x^4 - 8x^3 - x^2 + x - 6 \\ P'(x) &= 20x^3 - 24x^2 - 2x + 1 \\ P''(x) &= 60x^2 - 48x - 2 \\ P^{(3)}(x) &= 120x - 48 \\ P^{(4)}(x) &= 120 \end{aligned}$$

Para $x = 1$, se tiene $P^{(4)}(1) > 0$, $P^{(3)}(1) > 0$, y $P''(1) > 0$, pero $P'(1) < 0$. Entonces se aumenta la variable x hasta $x = 2$, y se tiene $P'(2) > 0$ y también $P(2) > 0$. La cota superior de las raíces de $P(x)$ calculada per la regla de Newton es 2.

Una consecuencia de la regla de Newton es que, cuando se considera un valor de x y se substituye en los polinomios $P^{(n)}(x), P^{(n-1)}(x), \dots, P^{(n-k)}(x)$, son todos positivos; entonces para cualquier valor mayor que x , también son positivos (se aplica la regla de Newton a $P^{(n-k)}(x)$). En la práctica, cuando se tiene que aumentar los valores de x , no hay que volver hacia atrás, ya que se mantiene el signo.

6.10.3 Separación de las raíces

Sucesiones de Sturm

Definición: Se dice que una sucesión de polinomios $P_0(x), P_1(x), \dots, P_n(x)$ es una sucesión de Sturm para P_0 en el intervalo $[a, b]$, si satisface:

1. Si para $x = \alpha$, $P_k(\alpha) = 0$, entonces $P_{k-1}(\alpha) \cdot P_{k+1}(\alpha) < 0$, $k = 1 \div n - 1$.
2. Si $P_0(\alpha) = 0$, entonces $P_0 \cdot P_1$ cambia de signo en α .
3. $P_n(x)$ no tiene raíces reales.

Teorema 6.5. Dada una sucesión de Sturm para $P_0(x)$ en el intervalo $[a, b]$, donde a y b no son raíces de $P_0(x)$, el número de raíces diferentes del polinomio $P_0(x)$ en el intervalo $[a, b]$ es igual a $V(a) - V(b)$, donde $V(c)$ es el número de cambios de signo de la sucesión $P_0(c), P_1(c), \dots, P_n(c)$.

Demostración: Sea α una raíz de $P_k(x)$, $k = 1 \div n-1$. Los polinomios son funciones continuas tal que mantienen el signo en un entorno: $\exists \delta > 0$ tal que los polinomios que no se anulan en α presentan signo constante en el intervalo $[\alpha-\delta, \alpha+\delta]$; por la condición 1., $V(\alpha-\delta) = V(\alpha+\delta)$.

Si en el intervalo $[a, b]$ sólo presentan ceros los polinomios $P_k(x)$, $k = 1 \div n-1$, entonces se tiene $V(a) = V(b)$. Además, como $P_n(x)$ no tiene raíces reales en el intervalo, respecto al número de cambios de signo no habrá ninguna variación cuando x pase por una raíz de $P_0(x)$, ya que se mantiene constante.

Por la condición 2., cuando x pase por una raíz de $P_0(x)$, $V(x)$ siempre crece o decrece en una unidad, de donde la diferencia $V(a) - V(b)$, en valor absoluto, da el número de raíces diferentes de $P_0(x)$ en el intervalo (a, b) . \square

Construcción: Se toma $P_0(x) = P(x)$ y $P_1(x) = P'(x)$; entonces se aplica el algoritmo de la división entera de Euclides

	$Q_1(x)$	$Q_2(x)$	
$P_0(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$
$-\lambda_2 P_2(x)$	$-\lambda_3 P_3(x)$		

Así se obtiene la sucesión a partir de los cocientes $P_{k-1} = Q_k(x)P_k(x) - \lambda_k P_{k+1}(x)$, donde $k = 1 \div n-1$, $Q_k(x)$ es el cociente y $\lambda_{k+1} P_{k+1}(x)$ es el resto. Después de cambiar el signo al resto y suprimir una constante positiva λ_k , se toma como nuevo divisor $P_{k+1}(x)$, y se repite el proceso. Si todos los ceros del polinomio $P_0(x)$ son simples, se tiene una sucesión de Sturm, donde $P_n = \text{m.c.d.}(P_0(x), P_1(x))$ y, por lo tanto, una constante. Cuando hay raíces múltiples, pueden eliminarse redefiniendo $P_0(x)$ y $P_1(x)$: Si $\text{m.c.d.}(P_0(x), P_1(x)) = P_r(x)$, se toma

$$P_0(x) = \frac{P_0(x)}{P_r(x)} \quad \text{y} \quad P_1(x) = P'_0(x)$$

Ejemplo. Sea el polinomio $x^3 - 3x^2 + x - 2$; entonces

$x^3 - 3x^2 + x - 2$	$\frac{1}{3}x - \frac{1}{3}$	$\frac{3}{4}x - \frac{39}{16}$
$-x^3 + 2x^2 - \frac{1}{3}x$	$3x^2 - 6x + 1$	$4x + 5$
$-x^2 + \frac{2}{3}x - 2$	$-3x^2 - \frac{15}{4}x$	
$x^2 - 2x + \frac{1}{3}$	$-\frac{39}{4}x + 1$	
$-\frac{4}{3}x - \frac{5}{3}$	$\frac{39}{4}x + \frac{195}{16}$	
$\lambda = 1/3$	$\frac{211}{16}$	

Se obtiene la sucesión de Sturm siguiente:

Valors de x	$-\infty$	0	$+\infty$	1	5
$P_0(x) = x^3 - 3x^2 + x - 2$	-	-	+	-	+
$P_1(x) = 3x^2 - 6x + 1$	+	+	+	-	+
$P_2(x) = 4x + 5$	-	+	+	+	+
$P_3(x) = -1$	-	-	-	-	-

Para saber cuantas raíces negativas tiene el polinomio, se calcula $V(-\infty) - V(0) = 0$. Las positivas son $V(0) - V(\infty) = 1$. Además, como que $V(1) - V(5) = 1$, $P(x)$ tiene una raíz en el intervalo $(1, 5)$.

Regla de Descartes

El número de raíces reales positivas de un polinomio con coeficientes reales, contada cada una tantas veces como multiplicidad tenga, no es mayor y de la misma paridad que el número de cambios de signo que presenta la sucesión de los coeficientes del polinomio.

La demostración de esta regla se basa en el teorema de Budan-Fourier, que afirma: El número de raíces real pertenecientes al intervalo (a, b) , contada cada una tantas veces como multiplicidad tenga, no es mayor y de la misma paridad que el número de variaciones de signo en la sucesión $P(x), P'(x), P''(x), \dots, P^{(n)}(x)$ al pasar de $x = a$ a $x = b$.

Ejercicio. Demostrar la regla de Descartes a partir del teorema de Budan-Fourier tomando el intervalo $(0, b)$, con b cota superior de las raíces positivas.

Ejemplos.

- Si se tiene el polinomio $P(x) = 16x^3 + 12x^2 - 8x - 1$, la sucesión de signos es $++--$; por lo tanto, hay un único cambio de signo, lo que implica una raíz real positiva. Si se hace el cambio $y = -x$, la sucesión es $-++-$, y se tienen dos cambios de signo; en consecuencia, hay dos o ninguna raíz real negativa.
- Si se sabe que todas las raíces son reales, entonces se puede afinar más: si se tiene el polinomio característico de una matriz simétrica $P(x) = x^3 - 2x^2 - 5x + 6$ que presenta 2 cambios de signo, el polinomio $P(x)$ tiene 2 o 0 raíces positivas; como $P(-x)$ tiene un cambio de signo, $P(x)$ tiene una raíz real negativa y, como todas son reales, dos han de ser positivas.

6.10.4 Método de Newton modificado y deflación

Sea el polinomio $P(x) = a_0 x^n + a_1 x^{n-1} + \dots + a_{n-1} x + a_n$ con todos los coeficientes reales y $a_0 > 0$.

Teorema 6.6. Si todas las raíces $\xi_1 \geq \dots \geq \xi_n$ son reales, el método de Newton da lugar a una sucesión (x_k) estrictamente decreciente, $\forall x^0 > \xi_1$.

Demostración: Como $a_0 > 0$, $P(x^0) > 0$. Si todas las raíces son diferentes, por el teorema de Rolle, el polinomio derivado, $P'(x)$, tiene $n - 1$ raíces α_i 's:

$$\xi_n \leq \alpha_{n-1} \leq \xi_{n-1} \leq \cdots \leq \xi_2 \leq \alpha_1 \leq \xi_1$$

y $P'(x) > 0$ para $x > \alpha_1$ (ya que $a_0 > 0$) y, por lo tanto, la función es creciente a partir de α_1 ; si se vuelve a aplicar Rolle, $P''(x) > 0$ para $x > \alpha_1$, y se tiene que el polinomio es una función convexa a partir de α_1 . Si se toma $x^k > \xi_1$, se tiene

$$x^{k+1} = x^k - \frac{P(x^k)}{P'(x^k)} < x^k$$

Además, del hecho que $x^k > \xi_k \geq \alpha_1$, por Taylor

$$0 = P(\xi_1) = P(x^k) + (\xi_1 - x^k) P'(x^k) + \frac{1}{2} (\xi_1 - x^k)^2 P''(\zeta) \quad \xi_1 < \zeta < x^k$$

Como $\frac{1}{2} (\xi_1 - x^k)^2 P''(\zeta) < 0$, se tiene $P(x^k) + (\xi_1 - x^k) P'(x^k) < 0$, y de la igualdad siguiente $P(x^k) = P'(x^k) (x^k - x^{k+1})$, se obtiene

$$P'(x^k) (x^k - x^{k+1}) + (\xi_1 - x^k) P'(x^k) = P'(x^k) (\xi_1 - x^{k+1}) < 0$$

De $P'(x^k) > 0$, se deduce que $x^{k+1} > \xi_1$. \square

El método de Newton para polinomios recibe el nombre de método de Birge-Vieta; en lugar de calcular la derivada del polinomio en un punto, $P'(x^k)$, se puede aplicar Ruffini teniendo en cuenta que

$$\text{Si } P(x) = (x - x^k) Q(x) + R \implies P(x^k) = R,$$

$$\text{y si } P'(x) = Q(x) + (x - x^k) Q'(x) \implies P'(x^k) = Q(x^k).$$

De este modo, concatenando dos divisiones por Ruffini del polinomio dividido por el factor $x - x^k$, se obtienen los restos $P(x^k)$ y $P'(x^k)$.

Si se toma la aproximación inicial, x^0 , muy alejada de la raíz mayor, ξ_1 , del polinomio $P(x)$, la convergencia del método de Newton es lenta:

$$x^{k+1} = x^k - \frac{a_0 (x^k)^n + \cdots}{n a_0 (x^k)^{n-1} + \cdots} \approx x^k \left(1 - \frac{1}{n}\right)$$

que puede ser un cambio infinitesimal. Un método adecuado en estos casos consiste en definir la iteración

$$x^{k+1} = x^k - 2 \frac{P(x^k)}{P'(x^k)}, \quad k = 0, 1, \dots$$

Se puede demostrar que, tomando un $x^0 > \xi_1$, son posibles dos situaciones:

- (a) La sucesión decrece a la derecha de ξ_1 : $x^0 \geq x^1 \geq \cdots \geq x^k \geq x^{k+1} \geq \cdots \geq \xi_1$, tal que $\lim_{k \rightarrow \infty} x^k = \xi_1$ y se tiene convergencia monótona y más rápida que la de Newton, o

(b) $\exists x^{k_0} = y$, tal que $P(x^0) \cdot P(x^k) > 0$ para $0 \leq k < k_0$, y $P(x^0) \cdot P(x^{k_0}) < 0$. Entonces, tomando $y^0 = y$ y aplicando el método de Newton

$$y^{k+1} = y^k - \frac{P(y^k)}{P'(y^k)} \quad k = 0, 1, \dots$$

se obtiene $x^{k_0-1} \geq \xi_1 \geq x^{k_0} = y \geq \alpha_1$ y, también, $x^{k_0-1} \geq y^1 \geq y^2 \geq \dots \geq \xi_1$ y

$$\lim_{k \rightarrow \infty} y^k = \xi_1$$

6.10.5 Método de Laguerre

Sea $P(x)$ un polinomio de grado > 2 con todas las raíces reales: $\xi_1 \leq \xi_2 \leq \dots \leq \xi_n$ con $\xi_1 < \xi_n$. Se denotan por I_i el intervalo $[\xi_i, \xi_{i+1}]$, $i = 0 \div n$, $\xi_0 = -\infty$ y $\xi_{n+1} = +\infty$.

Sea x^0 una aproximación a un cero de $P(x)$, con $x^0 \in I_i$. El método consiste en construir una parábola con dos ceros reales en I_i , donde uno de ellos es más próximo al cero de $P(x)$ que x^0 .

Como hay muchas paráolas dependiendo de un parámetro real λ que cumplan la condición anterior, se añade que la elección de λ ha de ser tal que uno de los ceros de la parábola tenga distancia mínima al cero de $P(x)$. Para ello se define

$$S(\lambda) = \sum_{i=1}^n \left(\frac{\lambda - \xi_i}{x^0 - \xi_i} \right)^2 > 0$$

y se tiene que la parábola $\Psi(y) = (x^0 - y)^2 S(\lambda) - (\lambda - y)^2$ tiene dos raíces $y_{1,2}$ diferentes si $\lambda \neq x^0$ (lo que se supone a partir de ahora). La función Ψ cumple las hipótesis del teorema de Bolzano, ya que, además de continua, si $P(x^0) \neq 0$,

$$\begin{aligned} \Psi(x^0) &< 0 \\ \Psi(\xi_i) &> 0 \quad i = 0 \div n + 1 \end{aligned}$$

Entonces, se tiene en el intervalo I_i : $\xi_i \leq y_1 \leq x^0 \leq y_2 \leq \xi_{i+1}$, y se ha de encontrar λ tal que un cero de $\Psi(\lambda)$ esté tan próximo como sea posible a un cero de $P(x)$. Por tanto, se ha de maximizar $|x^0 - y|$ como función de λ o, con el cambio de variables $\mu = \lambda - x^0$, como función de μ , ya que $\frac{d}{d\lambda} = \frac{d}{d\mu}$.

Ejercicios.

1. Comprobar que $\frac{P'(x^0)}{P(x^0)} = \sum_{k=1}^n \frac{1}{x^0 - \xi_k} \equiv S_1$.

2. Derivando $\frac{P'(x)}{P(x)}$, obtener

$$\frac{P'(x^0)^2 - P(x^0) P''(x^0)}{P(x^0)^2} = \sum_{k=1}^n \frac{1}{(x^0 - \xi_k)^2} \equiv S_2.$$

3. Mediante el cambio $\mu = \lambda - x^0$, comprobar que

$$\left(\frac{\lambda - \xi_k}{x^0 - \xi_k} \right)^2 = \frac{\mu^2}{(x^0 - \xi_k)^2} + \frac{2\mu}{x^0 - \xi_k} + 1.$$

A partir de los resultados de los ejercicios e introduciendo el cambio $\eta = x^0 - y$, entonces se tiene $\lambda - y = \mu + \eta$ y

$$\begin{aligned} S(\lambda) &= \sum_{k=1}^n \left(\frac{\mu^2}{(x^0 - \xi_k)^2} + \frac{2\mu}{x^0 - \xi_k} + 1 \right) \\ &= \mu^2 S_2 + 2\mu S_1 + n \end{aligned}$$

Si se ordena $\Psi(\eta)$ en potencias de μ , resulta

$$\begin{aligned} \Psi(\eta) &= \eta^2 (\mu^2 S_2 + 2\mu S_1 + n) - (\mu + \eta)^2 \\ &= \mu^2 (\eta^2 S_2 - 1) + 2\mu \eta (\eta S_1 - 1) + (n-1)\eta^2 \end{aligned}$$

Para encontrar el máximo valor de $|\eta|$ con μ real, se impone que el discriminante de la ecuación de segundo grado, $\Psi(\eta) = 0$, sea nulo

$$D = 4\eta^2 [\{S_1^2 - S_2(n-1)\}\eta^2 - 2\eta S_1 + n] = 0$$

Nota. Si se tiene la ecuación $ax^2 + 2bx + c = 0$, una forma de expresar la solución es

$$x_{1,2} = \frac{c}{-b \pm \sqrt{b^2 - ac}}$$

De la nota se deduce $\eta = \frac{n}{S_1 \pm \sqrt{(n-1)(nS_2 - S_1^2)}}$. Como $\eta = x^0 - y$, se obtiene

$$y = x^0 - \frac{n P(x^0)}{P'(x^0) \pm \sqrt{H(x^0)}}$$

donde $H(x) = (n-1)[(n-1)P'(x)^2 - nP(x)P''(x)]$.

Ejercicio. Demostrar que $H(x^0)$ es no negativa si todos los ceros de $P(x)$ son reales. Para ello, considerar que

$$H(x^0) = P(x^0)^2 (n-1)(nS_2 - S_1^2)$$

y es suficiente demostrar, por inducción o mediante multiplicadores de Lagrange, que la expresión

$$n \sum_{k=1}^n a_i^2 - \left(\sum_{k=1}^n a_i \right)^2,$$

es no negativa para cualquier conjunto de números reales a_i , $i = 1 \div n$.

En definitiva, se tiene el método iterativo

$$x^{k+1} = x^k - \frac{n P(x^k)}{P'(x^k) \pm \sqrt{H(x^k)}}$$

Ejercicio. Si ξ es una raíz simple de $P(x)$ y se define la función de iteración

$$g(x) = x - \frac{n P(x)}{P'(x) \pm \sqrt{H(x)}}$$

comprobar que $g'(\xi) = g''(\xi) = 0$, si se toma el signo igual al de $P'(\xi)$. En estas condiciones, el método de Laguerre es de orden por lo menos 3.

Si se supone que todas las raíces son reales y que se tiene una aproximación inicial x^0 que es la menor (mayor) de todas las raíces, $x^0 < \xi_1$ ($x^0 > \xi_n$), entonces se tiene $x^k < x^{k+1} < \xi_1$ ($x^k > x^{k+1} > \xi_n$), si se toma el signo de acuerdo con $P(x)$. Por tanto, el método es interesante porque asegura la convergencia para cualquier condición inicial x^0 .

Si todos los ceros son reales, pero hay algunos no simples, el método es convergente pero de primer orden en un entorno de los ceros múltiples. Para polinomios que presenten raíces complejas, no hay convergencia global; si se toma una aproximación inicial real, puede converger en el plano complejo, ya que en esta situación H puede tomar valores negativos. Si el método converge hacia una raíz compleja simple, entonces se puede demostrar que el método sigue siendo de tercer orden, por lo menos (Parlett 1964; referenciado en [Ral78]).

Ejemplos.

1. Si se tiene el polinomio $P(x) = x^4 + 8.1x^3 - 19.8x^2 - 5.9x + 21$ y se toma como aproximación inicial $x^0 = 10^6$, después de 6 iteraciones se obtiene $x_6 = 1.5$, que es la raíz.
2. Con la aproximación inicial $x^0 = 10^6$, la ecuación

$$x^4 + 2x^3 + 3x^2 + 4x + 5 = 0$$

tiene una solución aproximada, después de 6 iteraciones: $x_6 = -1.2878155 + 0.8578968 i$ que presenta 7 decimales correctos (la otra pareja de soluciones es $0.2878155 \pm 1.4160932 i$).

6.10.6 Método de Bairstow

Este método consiste en el cálculo de raíces complejas conjugadas de un polinomio con coeficientes reales $P(x)$. Este método aproxima el polinomio cuadrático $x^2 + px + q$ que tiene las raíces complejas de $P(x)$. Si se divide $P(x)$ por este polinomio, se tiene

$$P(x) = P_1(x)(x^2 + px + q) + Ax + B, \quad \text{con } \operatorname{gr} P_1(x) = n - 2. \quad (6.8)$$

El método consiste en resolver el sistema $\begin{cases} A(p, q) = 0 \\ B(p, q) = 0 \end{cases}$ Si se aplica el método de Newton,

$$\begin{pmatrix} p_{k+1} \\ q_{k+1} \end{pmatrix} = \begin{pmatrix} p_k \\ q_k \end{pmatrix} - \begin{pmatrix} A_p & A_q \\ B_p & B_q \end{pmatrix}^{-1} \begin{pmatrix} A(p_k, q_k) \\ B(p_k, q_k) \end{pmatrix}$$

Se ha reducido el problema al cálculo de las derivadas parciales de la matriz jacobiana en los puntos (p_k, q_k) . Derivando 6.8

$$0 = \frac{\partial}{\partial p} P(x) = (x^2 + px + q) \frac{\partial P_1}{\partial p} + x P_1(x) + A_p x + B_p \quad (6.9)$$

$$0 = \frac{\partial}{\partial q} P(x) = (x^2 + px + q) \frac{\partial P_1}{\partial q} + P_1(x) + A_q x + B_q$$

Si se divide $P_1(x)$ por $x^2 + px + q$, se tiene $P_1(x) = P_2(x)(x^2 + px + q) + A_1 x + B_1$ y, si z_i , $i = 1, 2$, son las dos raíces diferentes y conjugadas de $x^2 + px + q$, se tiene $P_1(z_i) = A_1 z_i + B_1$. Evaluando 6.9 en z_1 y z_2 , se obtienen 4 ecuaciones

$$0 = z_i (A_1 z_i + B_1) + A_p z_i + B_p \quad (6.10)$$

$$0 = A_1 z_i + B_1 + A_q z_i + B_q \quad (6.11)$$

Restando en 6.11 $i = 1$ de $i = 2$, y teniendo en cuenta que $z_1 \neq z_2$, se obtiene $A_q = -A_1$ y $B_q = -B_1$.

Substituyendo en 6.10 los valores obtenidos se tiene

$$0 = -z_i^2 A_q + z_i (A_p - B_q) + B_p \quad (6.12)$$

Como $z_i^2 + p z_i + q = 0$, $z_i^2 = -(p z_i + q)$ y substituyendo en 6.12, se tiene

$$0 = -z_i (A_p - B_q + A_q p) + B_p + A_q q$$

Restando $i = 1$ de $i = 2$, y como $z_1 \neq z_2$, se obtiene

$$\begin{aligned} A_p - B_q + A_q p &= 0 \\ B_p + A_q q &= 0 \end{aligned}$$

En definitiva,

$$\begin{aligned} A_p &= -B_1 + A_1 p \\ B_p &= A_1 q \end{aligned}$$

Conocidos A_1 y B_1 , se puede calcular la matriz jacobiana.

Nota. Para calcular A_1 y B_1 , se ha de conocer $P_1(x)$, A y B que se calculan de la manera siguiente (división sintética):

Si $P(x) = a_0 x^n + a_1 x^{n-1} + \dots + a_{n-1} x + a_n$, se realiza el cálculo siguiente

$$\begin{array}{c|cccccc|cc} & a_0 & a_1 & a_2 & a_3 & \cdots & a_{n-2} & a_{n-1} & a_n \\ \hline -p & & -p b_0 & -p b_1 & -p b_2 & \cdots & -p b_{n-3} & -p b_{n-2} & \\ -q & & -q b_0 & -q b_1 & -q b_2 & \cdots & -q b_{n-4} & -q b_{n-3} & -q b_{n-3} \\ \hline b_0 & b_1 & b_2 & b_3 & \cdots & b_{n-2} & b_{n-1} & b_n \end{array}$$

donde $A = b_{n-1}$ y $B = b_n$. Análogamente, se obtienen los coeficientes de $P_2(x)$, A_1 y B_1 .

Por ejemplo, si se quiere dividir $x^5 - 4x^4 - x^3 + 2x^2 + x - 10$ por $x^2 - 2x + 3$, se escribe

$$\begin{array}{c|ccccc|cc} & 1 & -4 & -1 & 2 & & 1 & -10 \\ \hline 2 & & 2 & -4 & -16 & & -16 & \\ -3 & & & -3 & 6 & & 24 & 24 \\ \hline & 1 & -2 & -8 & -8 & & 9 & 14 \end{array}$$

El cociente es $Q(x) = x^3 - 2x^2 - 8x - 8$ y el resto da $R(x) = 9x + 14$.

6.11 Problemas

1. Utilizando los métodos de bisección, secante y Newton, calcular los ceros de las funciones siguientes con ocho cifras significativas:
 - 1) $4 \operatorname{sen} x + 1 - x$
 - 2) $1 - x - e^{-2x}$
 - 3) $(x + 1)e^{x-1} - 1$
 - 4) $x^4 - 4x^3 + 2x^2 - 8$
 - 5) $e^x + x^2 + x$
 - 6) $e^x - x^2 - 2x - 2$
 - 7) $3x^2 + \operatorname{tg} x$
2. Se quiere resolver la ecuación $x + \ln x = 0$ y se sabe que una raíz es próxima a 0.5; se quiere aplicar el método de iteración simple y se puede elegir entre las fórmulas siguientes:

$$x_{n+1} = -\ln x_n \quad x_{n+1} = e^{-x_n} \quad x_{n+1} = (x_n + e^{-x_n})/2$$

¿Qué fórmulas se pueden usar?, ¿cuál es la recomendable? Dar una fórmula mejor.
3. Sea α una raíz simple de $f(x) = 0$ (con $f \in C^4(I)$, donde I es un entorno de α) y se considera el método iterativo

$$x_{n+1} = \frac{1}{2}(x'_{n+1} + x''_{n+1})$$

donde

$$x'_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad \text{y} \quad x''_{n+1} = x_n - \frac{u(x_n)}{u'(x_n)}, \quad \text{con} \quad u(x) = \frac{f(x)}{f'(x)}.$$

Demostrar que, si la sucesión $\{x_n\}_{n \in \mathbb{N}}$ converge hacia α , el orden de convergencia del método es por lo menos 3.

4. Se quiere determinar una raíz de la ecuación $x = g(x)$ con un error menor que $0.5 \cdot 10^{-4}$. Se ha calculado $x_4 = 0.43789$ y $x_5 = 0.43814$, y se sabe que $|g'(x)| \leq 0.4$; ¿cuántas iteraciones son necesarias, teniendo en cuenta el error de redondeo?
5. Se considera $f : \mathbf{R} \rightarrow \mathbf{R}$ de clase C^2 , con un único cero α tal que $\forall x, f'(x) \neq 0$ y $|4f(x)f''(x)| \leq [f'(x)]^2$.
 - a) Demostrar que el método de Newton $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ convergen hacia α , para cualquier elección de x_0 .
 - Si $|x_{16} - x_{17}| \leq 10^{-4}$,
 - acotar $|x_{16} - \alpha|$,
 - encontrar k tal que $|x_k - \alpha| \leq 10^{-7}$.
6. a) Probar que, si se aplica interpolación cuadrática para aproximar $f'(x)$, el método de Newton se transforma en

$$x_{n+1} = x_n - \frac{f(x_n)}{r}$$

donde

$$r = [x_n, x_{n-1}] + (x_n - x_{n-1}) \cdot [x_n, x_{n-1}, x_{n-2}]$$

- b) Demostrar que, si se toma x_{n+1} como cero de la misma parábola de interpolación, entonces

$$x_{n+1} = x_n - \frac{2f(x_n)}{r \pm \sqrt{r^2 - 4f(x_n) \cdot [x_n, x_{n-1}, x_{n-2}]}}$$

NOTA: El símbolo $[\cdot, \cdot, \cdot]$ representa las diferencias divididas de f . Este método se llama de Muller-Traub.

7. Sean f y g funciones reales derivables en $[a, b]$ tal que la ecuación $f(x) = g(x)$ tiene una única solución en $[a, b]$.
- Estudiar bajo qué condiciones el método iterativo $f(x_{n+1}) = g(x_n)$ es convergente localmente.
 - Aplicarlo a la ecuación $(1+x) \operatorname{sen} x = 1$ en el intervalo $[0.5, 1]$ con $x_0 = 0.5$ para obtener la solución con un error inferior a $0.5 \cdot 10^{-2}$.
8. Se quiere encontrar hacia dónde convergen los iterados por el método de Newton para la función $p(x) = x^3 - x$ al variar el punto inicial x_0 .
- Demostrar que, si $x_0 \in (a_0, +\infty)$, siendo $a_0 = 1/\sqrt{3}$, entonces el método converge hacia $+1$, y si $x_0 \in (-\infty, -a_0)$, hacia -1 .
 - Probar que existe un a_1 tal que, si $x_0 \in (a_1, a_0)$, el método converge hacia -1 y si $x_0 \in (-a_0, -a_1)$, hacia $+1$. Encontrar la ecuación que determina a_1 y obtener una aproximación de a_1 .
 - Probar que \exists una sucesión a_0, a_1, a_2, \dots decreciente tal que, si $x_0 \in (a_{i+1}, a_i)$, el método de Newton converge hacia $+1$ si i es impar, hacia -1 si i es par y no converge si x_0 es igual a algún a_i . Escribir la ecuación que permite encontrar a_{i+1} en función de a_i .
 - Probar que la sucesión $\{a_i\}_{i \in \mathbb{N}}$ tiene límite $a > 0$.
 - Estudiar la convergencia del método para $x_0 \in (-a, a)$.
9. Aplicar el método de Newton para resolver el sistema no lineal

$$\begin{cases} x &= \operatorname{sen}(x+y) \\ y &= \cos(x-y) \end{cases}$$

próximo a $(1, 1)$, con una precisión tal que $\|\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}\|_\infty \leq 10^{-11}$, con $\mathbf{z} = (x, y)^T$.

10. Encontrar todas las soluciones del sistema:

$$\begin{cases} \operatorname{sen}(xy) &= 1/2 \\ \cos x &= e^y \end{cases}$$

con una precisión tal que $\|\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}\|_\infty \leq 10^{-6}$, con $\mathbf{z} = (x, y)^T$.

11. Calcular todas las raíces de los polinomios siguientes:

- $6x^3 + x^2 - 29x - 14$
- $10x^3 - 101x^2 + 8x + 20$

12. Aplicar el teorema de Sturm a los polinomios

- a) $x^6 + 4x^5 + 4x^4 - x^2 - 4x - 4$
- b) $x^6 + 4x^5 + 4x^4 - x^2 - 4x + 4$

13. Dado el polinomio $P(x) = x^4 + 4x^3 - 4x^2 - 16x - 8$, comprobar que 17 es una cota superior de las raíces y mediante el método de Newton modificado calcular la raíz positiva mayor (2.14628). Calcular con una precisión de 4 decimales correctos las otras raíces, aplicando el método de Maehly.

14. Se considera la función de iteración

$$x_{k+1} = x_k - \frac{u_k}{Q(u_k)}, \quad \text{donde } u_k = \frac{f_i}{f'_i},$$

para resolver la ecuación $f(x) = 0$.

- a) Si se toma $\text{gr } Q(x) = 1$, construir un método de orden 3 (método de Halley).
- b) Aplicarlo al cálculo de $\sqrt{2}$ y compararlo con el método de Newton.

15. a) Demostrar que el método de Newton aplicado a la función

$$f(x) = \frac{1}{x} - a$$

permite calcular $\frac{1}{a}$ sin realizar ninguna división.

- b) ¿Qué relación exacta hay entre ε_{k+1} y ε_k , si $\varepsilon_k = |x_k - \frac{1}{a}|$?
- c) Si $a = 0.4$ y $\varepsilon_0 = 0.2$, ¿para qué valores de k se tiene $\varepsilon_k \leq 10^{-20}$?
- d) Estudiar, en función de a , los intervalos de convergencia del método según se elija la aproximación inicial x_0 .

16. Si $f \in \mathcal{C}^3$ y α es un cero simple de f , se construye un método iterativo del siguiente modo: se calcula el polinomio cuadrático interpolador de f en x_{k-1} y x_k con

$$P_2(x_{k-1}) = f(x_{k-1}), \quad P_2(x_k) = f(x_k), \quad P'_2(x_k) = f'(x_k).$$

Entonces, se calcula x_{k+1} como el cero de $P_2(x)$ más próximo a x_k .

- a) Explicitar el método iterativo; es decir, encontrad el cálculo que da x_{k+1} en función de x_k y x_{k-1} .
- b) Si se define $\varepsilon_k = |x_k - \alpha|$ y se supone que $f^{(3)}(\alpha) \neq 0$, demostrar que

$$\lim_{k \rightarrow \infty} \frac{\varepsilon_{k+1}}{\varepsilon_k^2 \varepsilon_{k-1}} = \left| \frac{f^{(3)}(\alpha)}{6 f'(\alpha)} \right|$$

- c) Encontrar el orden y su constante asintótica del error.

6.12 Prácticas

6.12.1 Práctica ejemplo

Escribir una rutina que utilice el método de interpolación cuadrática inversa para calcular ceros de una función. Aplicarlo al problema siguiente (ver [Car79]):

Una de las más conocidas ecuaciones de estado en P - V - T (presión, volumen, temperatura), es la llamada de Beattie-Bridgeman

$$P = \frac{RT}{V} + \frac{\beta}{V^2} + \frac{\gamma}{V^3} + \frac{\delta}{V^4}$$

donde R es la constante universal de los gases, los sumandos segundo, tercero y cuarto son correcciones de la ley de los gases perfectos $P = \frac{RT}{V}$ para gases reales y los parámetros β , γ y δ vienen definidos por

$$\beta = RTB_0 - A_0 - \frac{Rc}{T^2}, \quad \gamma = -RTB_0 b + A_0 a - \frac{RcB_0}{T^2} \quad \text{y} \quad \delta = \frac{RB_0 bc}{T^2}.$$

A_0 , B_0 , a , b y c son constantes determinadas en forma empírica y tabuladas para cada gas.

A partir de la presión del gas P , la temperatura T y las constantes R , A_0 , B_0 , a , b y c , calcular el volumen molar V y el coeficiente de compresibilidad $Z = \frac{PV}{RT}$, que es un índice muy útil para analizar la discrepancia entre el comportamiento de un gas real y uno de perfecto (donde $z = 1$).

Calcular los resultados del coeficiente de compresibilidad del gas metano para temperaturas de 0°C y 200°C para presiones de 1, 2, 5, 20, 40, 60, 80, 100, 120, 160, 180 y 200 atmósferas, dibujando una gráfica de la compresibilidad en función de la presión. Usad los valores de las constantes características del metano:

$$\begin{array}{llll} A_0 & = & 2.27690 & B_0 = 0.05587 \\ a & = & 0.01855 & b = -0.01587 \\ c & = & 12.83 \cdot 10^4 & \end{array}$$

Se construye una función donde una llamada es de la forma `ZZ = ZEROAP (A, B, F, TOL)`; las variables de entrada son: los extremos del intervalo donde hay un cero, $A < B$, la función de la que se quiere encontrar un cero aproximado, F , y el error tolerado en este cálculo, TOL . El programa supone que $F(A)$ y $F(B)$ tienen signos opuestos y se basa en un algoritmo de Dekker (1969) que mejoró Brent (1973) (ver `ZEROIN` de [For77]).

En cada llamada a `ZEROAP` hay tres abscisas presentes:

1. B es el último punto encontrado (iterado) y el más próximo al cero.
2. A es el iterado anterior.
3. C es el iterado anterior o uno de anterior tal que $F(B) \cdot F(C) < 0$.

El cero siempre está entre B y C ; además, $|F(B)| \leq |F(C)|$. El criterio que retorna el valor aproximado del cero es el siguiente: si

$$|B - C| \leq TOL + 4. * EPS * DABS(B)$$

entonces la función `ZEROAP` fa $ZZ = B$.

En cada paso `ZEROAP` elige el próximo iterado entre dos candidatos: uno obtenido por bisección del intervalo y el otro calculado por interpolación cuadrática inversa (si A , B y C son diferentes) o interpolación lineal (método de la secante) cuando hay alguno repetido.

El número de evaluaciones de la función no puede exceder al valor

$$\left[\log_2 \left(\frac{B - A}{TOL1} \right) \right]^2$$

donde $TOL1 = 0.5 * TOL + 2.0 * EPS * DABS(B)$. El cero calculado, R , es tal que F cambia el signo en el intervalo $[R - 2 \cdot TOL1, R + 2 \cdot TOL1]$.

```
REAL*8 FUNCTION F(T,P,V)
REAL*8 T,P,V,R
REAL*8 BETA,GAMMMA,DELTA,A0,B0,A,B,C
R= 0.08205
A0= 2.2769
B0= 0.05587
A= 0.01855
B= -0.01587
C= 12.83D04
BETA= R*T*B0-A0-((R*C)/(T**2))
GAMMA= -R*T*B0*B+A0*A-((R*C*B0)/(T**2))
DELTA= (R*B0*B*C)/(T**2)
F= (R*T/V)+(BETA/(V**2))+(GAMMA/(V**3))
F= F+(DELTA/(V**4))-P
RETURN
END
```

```
PROGRAM VOLUMEN
EXTERNAL F,ZEROAP
REAL*8 T,V,P,Z,F,FTPV,AX,BX,TOL,ZZ,R
INTEGER I,NIQ,NIB,NIS
PRINT*, 'TOLERANCIA'
READ*,TOL
R= 0.08205
DO 10 I=1,28,1
READ(1,*),T,P
T= T+273.15
IF (T.EQ.273.15) THEN
  AX= 0.19950041
  BX= 50.0
ELSE
  AX= 0.005
  BX= 50.0
ENDIF
```

```

      ZZ= ZEROAP(AX,BX,F ,TOL,T,P,V,NIS,NIQ,NIB)
      Z= (P*ZZ)/(R*T)
      T= T-273.15
      WRITE(2,100),T,P,ZZ,Z,NIB,NIS,NIQ
10   CONTINUE
100  FORMAT(T2,F5.1,T10,F5.1,T18,F16.10,T40,F12.10,T56,I6
*           T64,I4,T72,I6)
      END

      REAL*8 FUNCTION ZEROAP (A, B, F , TOL)
      REAL*8 AX,BX,F,TOL
C      SE CALCULA UN CERO DE LA FUNCION F(X) EN EL INTERVALO A,B
C
C      ENTRADA..
C      A EXTREMO INFERIOR DEL INTERVALO INICIAL
C      B EXTREMO SUPERIOR DEL INTERVALO INICIAL
C      F FUNCION SUBPROGRAMA QUE EVALUA F(X) PARA QUALQUIER X EN
C          EL INTERVALO A,B
C      TOL LONGITUD DESEADA DEL INTERVALO DEL RESULTADO FINAL (.GE. 0.0)
C
C      SALIDA..
C      ZEROAP APROXIMA LA ABSISA DE UN CERO DE F EN EL INTERVALO A,B
C      SE SUPONE QUE F(A)Y F(B)TIENEN SIGNOS OPUESTOS, SIN COMPROBARLO.
C      ZEROAP DEVUELVE UN CERO X EN EL INTERVALO DADO A,B CON UNA TOLERANCIA
C          4 * MACHEPS * DABS(X) + TOL,
C      DONDE MACHEPS ES EL EPSILON DE LA MAQUINA.
C
      REAL A,B,C,D,E,EPS,FA,FB,FC,TOL1,XM,P,Q,R,S
C
C      CALCULO DE EPS, PRECISION RELATIVA DE LA MAQUINA
      EPS = 1.0
10   EPS = EPS/2.0
      TOL1 = 1.0 + EPS
      IF(TOL1.GT. 1.0) GO TO 10
C
C      INICIALIZACION
      A = AX
      B = BX
      FA = F(A)
      FB = F(B)
C
C      COMIENZA EL METODO
20   C = A
      FC = FA
      D = B - A
      E = D
30   IF(ABS(FC) .GE. ABS(FB)) GO TO 40
      A = B
      B = C
      C = A
      FA = FB
      FB = FC
      FC = FA
C
C      TEST DE CONVERGENCIA
40   TOL1 = 2.0*EPS*ABS(B) + 0.5*TOL

```

```

XM = .5*(C - B)
IF (ABS(XM) .LE. TOL1) GO TO 90
IF (FB .EQ. 0.0) GO TO 90
C
C   ES NECESARIA UNA BISECCION
      IF (ABS(E) .LT. TOL1) GO TO 70
      IF (ABS(FA) .LE. ABS(FB)) GO TO 70
C
C   LA INTERPOLACION CUADRATICA ES POSIBLE
      IF (A .NE. C) GO TO 50
C
C   INTERPOLACION LINEAL
      S = FB/FA
      P = 2.0*XM*S
      Q = 1.0 - S
      GO TO 60
C
C   INTERPOLACION CUADRATICA INVERSA
50   Q = FA/FC
      R = FB/FC
      S = FB/FA
      P = S*(2.0*XM*Q*(Q - R) - (B - A)*(R - 1.0))
      Q = (Q - 1.0)*(R - 1.0)*(S - 1.0)
C
C   AJUSTE DE LOS SIGNOS
60   IF (P .GT. 0.0) Q = -Q
      P = ABS(P)
C
C   LA INTERPOLACION ES ACEPTABLE
      IF ((2.0*P) .GE. (3.0*XM*Q - ABS(TOL1*Q))) GO TO 70
      IF (P .GE. ABS(0.5*E*Q)) GO TO 70
      E = D
      D = P/Q
      GO TO 80
C
C   BISECCION
70   D = XM
      E = D
C
C   METODO COMPLETADO
80   A = B
      FA = FB
      IF (ABS(D) .GT. TOL1) B = B + D
      IF (ABS(D) .LE. TOL1) B = B + SIGN(TOL1,XM)
      FB = F(B)
      IF((FB*(FC/ABS(FC))) .GT. 0.0) GO TO 20
      GO TO 30
C
C   HECHO
90   ZEROAP = B
      RETURN
      END

```

Los resultados obtenidos se presentan en la figura 6.1 y en la tabla siguiente:

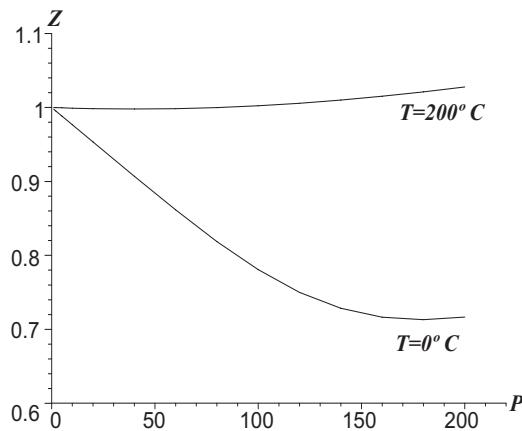


Figura 6.1 Compresibilidad en función de la presión

T	P	V	Z	[1]	[2]	[3]
0.0	1.0	22.3599262238	0.9976784145	1	4	4
0.0	2.0	11.1539258957	0.9953549034	3	8	8
0.0	5.0	4.4303135872	0.9883816680	7	12	14
0.0	10.0	2.1890854836	0.9767489000	12	16	21
0.0	20.0	1.0684106350	0.9534291103	18	20	29
0.0	40.0	0.5081635118	0.9069506967	24	24	37
0.0	60.0	0.3218188584	0.8615548866	24	32	39
0.0	100.0	0.1995003968	0.7806868901	43	57	40
0.0	80.0	0.2293661833	0.8187278895	24	38	40
0.0	100.0	0.1995003968	0.7806868901	43	57	40
0.0	120.0	0.1995003968	0.7499068181	62	76	40
0.0	140.0	0.1995003968	0.7283031204	81	95	40
0.0	160.0	0.1995003968	0.7163651424	100	114	40
0.0	180.0	0.1995003968	0.7130638765	119	133	40
0.0	200.0	0.1995003968	0.7166361781	138	152	40
200.0	1.0	38.8180122375	0.9998983755	138	157	42
200.0	2.0	19.4070835114	0.9997993281	140	162	46
200.0	5.0	7.7606439590	0.9995173426	143	167	51
200.0	10.0	3.8786890507	0.9990967227	147	172	56
200.0	20.0	1.9381077290	0.9984495594	152	177	62
200.0	40.0	0.9685317278	0.9979115786	158	182	69
200.0	60.0	0.6459901929	0.9983879033	164	186	78
200.0	80.0	0.4852141738	0.9998657508	171	191	86
200.0	100.0	0.3891199529	1.0023182491	178	194	95
200.0	120.0	0.3253661096	1.0057157862	185	198	105
200.0	140.0	0.2800856829	1.0100568427	193	202	115
200.0	160.0	0.2463186532	1.0151755223	201	207	124
200.0	180.0	0.2202132046	1.0210294343	209	211	133
200.0	200.0	0.1995036900	1.0277877156	217	213	143

donde

T : temperatura (en grados centígrados)
 V : volumen (en litro/gr.molar)

P : presión (en atmósferas)
 Z : coeficiente de compresibilidad

y

- [1] es el número de iteraciones por bisección
- [2] el número de iteraciones en que se ha utilizado el método de la secante y
- [3] el número de iteraciones por interpolación cuadrática inversa

6.12.2 Enunciados

- Escribir una rutina, llamada INVERF, que calcule la función inversa de la función $\text{erf}(x)$ donde $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. Los argumentos de esta rutina serán $\text{ERFX} = x_0$ que es el valor que corresponde a $\text{erf}(x)$, $0 \leq \text{erf}(x) < 1$, para el que se busca la función inversa: $\text{erf}(x) = x_0$ y TOL , que es el error máximo permitido en el cálculo de x .

Cuando sea necesario calcular la función erf , utilizad la rutina construida en el capítulo 5, ROMBERG8. El cálculo de x se debe realizar de dos maneras:

- Por el método de Newton, contando el número de iteraciones necesarias.
- Utilizando la función ZEROAP, donde tenéis que introducir la siguiente modificación: hay que contar cuántas iteraciones se han realizado mediante bisección, secante o interpolación cuadrática inversa.

Una tabla de comprobación es la siguiente; la podéis mejorar haciendo una llamada a alguna librería numérica de vuestro sistema (NAG, IMSL u otra):

ERFX	x	ERFX	x
0.00000 00000	0.00	0.84270 07929	1.00
0.11246 29160	0.10	0.96610 51464	1.50
0.71115 56337	0.75	0.99532 22650	2.00

Resolver la ecuación $\text{erf}(x) = 0.75$ con la máxima precisión posible.

- Escribir una rutina, llamada INVBES, que calcule la función inversa de la función $J_0(x)$ donde $J_0(x) = \frac{1}{\pi} \int_0^\pi \cos(x \sen t) dt$. Los argumentos de esta rutina serán $BES0 = x_0$, que es el valor que corresponde a $J_0(x)$, $|J_0(x)| \leq 1$, para el que se busca la función inversa: $J_0(x) = x_0$ y TOL , que es el error máximo permitido en el cálculo de x .

Cuando sea necesario calcular la función $J_0(x)$, utilizad la rutina construida en el capítulo 5, ROMBERG8. El cálculo de x se debe realizar de dos maneras:

- Por el método de Newton, contando el número de iteraciones necesarias teniendo en cuenta que $J'_0(x) = -J_1(x) = -\frac{1}{\pi} \int_0^\pi \cos(x \sen t - t) dt$.

- (b) Utilizando la función **ZEROAP**, donde tenéis que introducir la siguiente modificación: hay que contar cuántas iteraciones se han realizado mediante bisección, secante o interpolación cuadrática inversa.

Una tabla de comprobación es la siguiente; la podéis mejorar haciendo una llamada a alguna librería numérica de vuestro sistema (NAG, IMSL u otra):

BES0	x	BES0	x
0.93846 98072	0.50	0.00250 76833	2.40
0.76519 76866	1.00	-0.04838 37765	2.50
0.51182 76717	1.50	-0.32054 25090	4.50

Resolver la ecuación $J_0(x) = 0$ con la máxima precisión posible.

3. Escribir una rutina, llamada **INVELL**, que calcule la función inversa de la función $K(x)$ donde

$$K(x) = \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - x \sin^2 \phi}}$$

Los argumentos de esta rutina serán **ELLX**= x_0 , que es el valor que corresponde a $K(x)$, $|K(x)| \geq \pi/2$, para el que se busca la función inversa: $K(x) = x_0$ y **TOL**, que es el error máximo permitido en el cálculo de x .

Cuando sea necesario calcular la función $K(x)$, utilizar la rutina construida en el capítulo 5, **ROMBERG8**. El cálculo de x se debe realizar de dos maneras:

- (a) Por el método de Newton, usando una aproximación numérica de la derivada de la función y contando el número de iteraciones necesarias.
- (b) Utilizando la función **ZEROAP**, donde tenéis que introducir la siguiente modificación: hay que contar cuántas iteraciones se han realizado mediante bisección, secante o interpolación cuadrática inversa.

Una tabla de comprobación es la siguiente; la podéis mejorar haciendo una llamada a alguna librería numérica de vuestro sistema (NAG, IMSL u otra):

ELLX	x	ELLX	x
1.59100 34538	0.05	2.15651 56474	0.75
1.68575 03548	0.25	2.90833 72484	0.95
1.85407 46773	0.50	3.35414 14457	0.98

Resolver la ecuación $K(x) = 2.0$ con la máxima precisión posible.

4. Construir una rutina que calcule los ceros de los polinomios de Legendre para cualquier grado n y tolerancia **TOL**, así como los pesos necesarios para la integración Gauss-Legendre.

Comprobar los resultados en alguna librería numérica de vuestro sistema (NAG, IMSL u otra) y dar un listado para $n = 2 \div 8, 16, 32$ y 64 con $\text{TOL} \leq 10^{-15}$. Los cálculos se deben realizar de dos maneras:

- (a) Por el método de Newton, contando el número de iteraciones necesarias.
- (b) Utilizando la función **ZEROAP**, donde tenéis que introducir la siguiente modificación: hay que contar cuántas iteraciones se han realizado mediante bisección, secante o interpolación cuadrática inversa.

5. Una fórmula útil de integración numérica, atribuida a Chebyshev, es

$$\int_{-1}^1 f(x) dx \approx \frac{2}{n+1} \sum_{i=0}^n f(x_i)$$

donde las x_i son las $n + 1$ raíces del polinomio que corresponde a la sucesión

$$\left. \begin{array}{rcl} C_0(x) & = & 1 \\ C_1(x) & = & x \\ C_2(x) & = & x^2 - 1/3 \\ C_3(x) & = & x^3 - 1/2 x \\ C_4(x) & = & x^4 - 2/3 x^2 + 1/45 \end{array} \right\| \left. \begin{array}{rcl} C_5(x) & = & x^5 - 5/6 x^3 + 7/72 x \\ C_6(x) & = & x^6 - x^4 + 1/5 x^2 - 1/105 \\ C_7(x) & = & x^7 - 7/6 x^5 + 119/360 x^3 - 149/6480 x \\ C_8(x) & = & x^8 - 3/2 x^7 + 27/40 x^5 - 57/560 x^3 \\ & & + 53/22400 x \end{array} \right\}$$

Todas las raíces de estos polinomios son reales, pertenecen al intervalo $(-1, 1)$, mientras que $C_n(x)$, para $n = 8$ y $n > 10$, son complejas.

Escribir una función llamada CPOL con argumentos: N , que es el grado del polinomio $C_n(x)$, $n = 2, 3, 4, 5, 6, 7$ y 9 y X , que es el valor de la variable; la función evalúa el polinomio en X después de haber introducido los coeficientes con una orden del tipo DATA.

Localizar las raíces de cada uno de los polinomios $C_n(x)$. Los cálculos se deben realizar de dos maneras:

- (a) Por el método de Newton, contando el número de iteraciones necesarias.
- (b) Utilizando la función ZEROAP, donde tenéis que introducir la siguiente modificación: hay que contar cuántas iteraciones se han realizado mediante bisección, secante o interpolación cuadrática inversa.

Imprimir los resultados en forma tabular. Tener en cuenta que n y $C_n(x)$ tienen la misma paridad y que ningún polinomio tiene dos raíces consecutivas dentro de un intervalo de longitud menor que 0.05 .

6. Un flujo turbulento de un fluido atraviesa un conducto cilíndrico estrecho de coeficiente de rozamiento c_f y número de Reynolds R_e ; la relación entre ellos es la siguiente

$$\sqrt{\frac{1}{c_f}} = -0.4 + 1.74 \ln(R_e \sqrt{c_f})$$

Calcular c_f para $R_e = 10^4, 10^5$ y 10^6 , de dos maneras:

- (a) Por el método de Newton, contando el número de iteraciones necesarias.
- (b) Utilizando la función ZEROAP, donde tenéis que introducir la siguiente modificación: hay que contar cuántas iteraciones se han realizado mediante bisección, secante o interpolación cuadrática inversa.

7. Escribir una rutina, llamada INVFRS, que calcule la función inversa de la función $S(x)$, donde

$$S(x) = \int_0^x \sin \frac{\pi t^2}{2} dt$$

Los argumentos de esta rutina serán FRSX= x_0 , que es el valor que corresponde a $S(x)$, $0 \leq S(x) < 0.72$, para el que se busca la función inversa: $S(x) = x_0$ y TOL, que es el error máximo permitido en el cálculo de x .

Cuando sea necesario calcular la función $S(x)$, utilizar la rutina construida en el capítulo 5, ROMBERG8. El cálculo de x se debe realizar de dos maneras:

- (a) Por el método de Newton, contando el número de iteraciones necesarias.
- (b) Utilizando la función ZEROAP, donde tenéis que introducir la siguiente modificación: hay que contar cuántas iteraciones se han realizado mediante bisección, secante o interpolación cuadrática inversa.

Una tabla de comprobación es la siguiente; la podéis mejorar haciendo una llamada a alguna librería numérica de vuestro sistema (NAG, IMSL u otra):

FRSX	x	FRSX	x
0.00052 36	0.1	0.69750 50	1.5
0.03335 94	0.4	0.49631 30	3.0
0.24934 14	0.8		

Encontrar las cinco primeras soluciones de la ecuación $S(x) = 0.5$, con la máxima precisión posible.

8. Escribir una rutina, llamada INVFRC, que calcule la función inversa de la función $C(x)$, donde

$$C(x) = \int_0^x \cos \frac{\pi t^2}{2} dt$$

Los argumentos de esta rutina serán FRCX= x_0 , que es el valor que corresponde a $C(x)$, $0 \leq C(x) < 0.78$, para el que se busca la función inversa: $C(x) = x_0$ y TOL, que es el error máximo permitido en el cálculo de x .

Cuando sea necesario calcular la función $C(x)$, utilizar la rutina construida en el capítulo 5, ROMBERG8. El cálculo de x se debe realizar de dos maneras:

- (a) Por el método de Newton, contando el número de iteraciones necesarias.
- (b) Utilizando la función ZEROAP, donde tenéis que introducir la siguiente modificación: hay que contar cuántas iteraciones se han realizado mediante bisección, secante o interpolación cuadrática inversa.

Una tabla de comprobación es la siguiente; la podéis mejorar haciendo una llamada a alguna librería numérica de vuestro sistema (NAG, IMSL u otra):

FRCX	x	FRCX	x
0.09999 75	0.1	0.44526 12	1.5
0.39748 08	0.4	0.60572 08	3.0
0.72284 42	0.8		

Encontrar las cinco primeras soluciones de la ecuación $S(x) = 0.5$ con la máxima precisión posible.

9. Un método clásico para resolver la ecuación cúbica es el de Cardan: Si se tiene

$$x^3 + a x^2 + b x + c = 0$$

y se considera la sustitución $x = y - a/3$, se reduce la ecuación a

$$y^3 + p y + q = 0, \text{ donde } p = b - \frac{a^2}{3} \text{ y } q = c - \frac{ab}{3} + \left(\frac{a}{3}\right)^3.$$

Una raíz de la ecuación transformada se encuentra de la manera siguiente:

$$s = \left[\left(\frac{p}{3} \right)^3 + \left(\frac{q}{2} \right)^2 \right]^{1/2} \text{ y } y_1 = \left[-\frac{q}{2} + s \right]^{1/3} + \left[-\frac{q}{2} - s \right]^{1/3}$$

y la raíz real de la ecuación original vendrá dada por $x_1 = y_1 - a/3$. Entonces, dividiendo por $x - x_1$, se tiene una ecuación cuadrática.

- (a) Aplicar este método para encontrar la raíz real de

$$x^3 + 3x^2 + \alpha^2 x + 3\alpha^2 = 0$$

para diferentes valores de α . Estudiar la pérdida de precisión para valores de α próximos al inverso del épsilon de la máquina.

- (b) Aplicar el método de Newton a la misma ecuación para el mismo conjunto de valores de α . Estudiar los efectos del redondeo y la elección de la aproximación inicial.

7 Valores y vectores propios

7.1 Introducción

En este capítulo pretendemos estudiar los métodos numéricos más importantes para calcular los valores y vectores propios de una matriz cualquiera. Es decir, dada una matriz $A \in \mathcal{L}(\mathbf{R}^n)$, queremos hallar las $\lambda \in \mathbf{C}$ para las cuales $\exists x \in \mathbf{C}^n$ con $x \neq 0$ tales que $Ax = \lambda x$, donde λ es un **valor propio** y x es el **vector propio** asociado a este valor propio.

Como los valores propios de una matriz son, también, las raíces de su polinomio característico $p(\lambda) = \det(A - \lambda I)$, se podría reducir el problema a calcular los ceros de este polinomio por alguno de los métodos ya vistos en el capítulo 6; pero, en general el cálculo del polinomio característico de una matriz es excesivamente costoso y, además, pequeños errores en los coeficientes pueden dar graves errores en sus raíces. Este tipo de resolución solamente se utilizará en matrices muy sencillas como, por ejemplo, las tridiagonales y simétricas. Para matrices cualesquiera tenemos, básicamente, dos tipos de métodos:

- Los métodos de tipo puramente iterativo, a través de los cuales, utilizando de forma reiterada un mismo tipo de transformación a la matriz inicial, se obtiene una sucesión de la cual se calculan uno o más valores propios. El más conocido es el método de la potencia, que, asociado con métodos de deflación, nos permite ir hallando los distintos valores propios de la matriz. Otro es el método de Jacobi para matrices simétricas.
- Los métodos basados en la factorización de alguna manera particular de la matriz A , para obtener iterativamente una sucesión de matrices con los mismos valores propios que converge a una matriz triangular superior. Como veremos, estos algoritmos no se aplican directamente sobre la matriz inicial, sino que previamente se transforma la matriz a una forma reducida: Hessenberg superior, o tridiagonal si la matriz inicial es simétrica.

El cálculo de valores y vectores propios aparece, por ejemplo, en algunos problemas de mecánica, en las vibraciones de estructuras, en la optimización y estudio de la estabilidad de otros métodos numéricos iterativos, etc.

7.2 Cotas de los valores propios

Hay algunas propiedades de matrices que nos permiten de hallar cotas de los valores propios; generalmente, estas cotas son bastante burdas, pero en algunos casos pueden dar una idea bastante exacta.

Uno de los resultados más elementales es el que nos relaciona el módulo de los valores propios de una matriz con la norma de la matriz; recordémoslo:

Teorema 7.1 Si λ es un valor propio de la matriz A ; entonces,

$$|\lambda| \leq \|A\|$$

este teorema es de fácil aplicación para las normas $\|\cdot\|_1$ y $\|\cdot\|_\infty$.

Ejemplo. Consideremos la matriz

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 2 & 1 & 1 \end{pmatrix}.$$

los valores de distintas normas de esta matriz son

$$\|A\|_1 = 5, \quad \|A\|_\infty = 4, \quad \|A\|_2 = 4.110728.$$

Los valores propios de la matriz son

$$\lambda_1 = 4, \quad \lambda_{1,2} = \frac{-1 \pm \sqrt{7}i}{2}.$$

Está claro, pues, que la mejor cota la da, en este caso, la norma infinito.

Un otro teorema muy útil es el de **Gerschgorin**:

Teorema 7.2 Sea $A \in \mathcal{L}(\mathbf{R}^n)$ con valores propios $\lambda_1, \dots, \lambda_n$. Entonces, cada λ_i pertenece a la unión de los discos

$$C_i = \{z \in \mathbf{C} / |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|\}$$

Además, si S es la unión de m discos, tal que S es disjunto con los otros discos restantes, entonces S contiene exactamente m valores propios contados con su multiplicidad.

Demostración: Por el hecho de ser λ valor propio de la matriz A , tenemos que

$$Ax = \lambda x \quad \text{para } x \neq 0,$$

en particular, para cada $i = 1 \div n$ se cumple

$$(\lambda - a_{ii})x_i = \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} x_j$$

Escogiendo i de manera que $|x_i| = \|x\|_\infty$ podemos escribir que

$$|\lambda - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^n \frac{|a_{ij}| |x_j|}{|x_i|} \leq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|$$

cosa que nos demuestra que λ está contenido en el disco C_i ; por lo tanto, cualquier valor propio estará contenido en la unión de todos los discos C_i .

para la segunda parte, definimos la familia de matrices

$$A_t = D + tB, \quad \text{con } D = \text{diag}(a_{11}, \dots, a_{nn}) \quad \text{y } B = A - D.$$

$C_i(t)$ denota el disco con centro a_{ii} y radio $t \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|$. Si suponemos que S es la unión de los m primeros discos, definimos

$$S(t) = \bigcup_{i=1}^m C_i(t) \quad \bar{S}(t) = \bigcup_{i=m+1}^n C_i(t)$$

Como $\bar{S}(1)$ es disjunto con S , $S(t)$ es disjunto con $\bar{S}(t)$ para todo $t \in [0, 1]$. En particular, $S(0)$ contiene m valores propios de la matriz $A_0 = \text{diag}(a_{11}, \dots, a_{nn})$, concretamente a_{11}, \dots, a_{mm} .

Por otro lado, la primera parte del teorema nos dice que $S(t) \cup \bar{S}(t)$ contiene todos los valores propios de la matriz A_t ; pero, como las dos uniones de discos son disjuntas, la continuidad de las raíces de un polinomio respecto a sus coeficientes, nos demuestra que los m primeros valores propios de A_t siguen estando en $S(t)$ y, en particular, $S(1) = S$ debe contener a m de ellos. \square

Ejercicio. Demostrar que el teorema también es válido si consideramos los discos

$$D_i = \{z \in \mathbf{C} / |z - a_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ji}|\}$$

Ejemplos.

1. Considerar la matriz

$$A = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 4 & 1 \\ 1 & 1 & 10 \end{pmatrix}$$

Si aplicamos el teorema 7.1 por la norma $\|\cdot\|_\infty$ tenemos que $\rho(A) \leq 12$, pero si aplicamos el teorema de Gershgorin podemos afirmar que los valores propios deben estar en la unión de los siguientes discos:

$$\begin{aligned} C_1 &= \{z \in \mathbf{C} / |z - 2| \leq 4\} \\ C_2 &= \{z \in \mathbf{C} / |z - 4| \leq 3\} \\ C_3 &= \{z \in \mathbf{C} / |z - 10| \leq 2\} \end{aligned}$$

Además, como el último disco es disjunto con los otros dos, podemos afirmar que C_3 contiene exactamente un valor propio y los otros dos deben estar en $C_1 \cup C_2$.

2. Considerar la matriz

$$A = \begin{pmatrix} 0.1 & 0.0 & 1.0 \\ 0.2 & 0.3 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix}$$

Si aplicamos el teorema de Gershgorin, los valores propios deben estar en la unión de los siguientes discos:

$$\begin{aligned} C_1 &= \{z \in \mathbf{C} / |z - 0.1| \leq 1.0\} \\ C_2 &= \{z \in \mathbf{C} / |z - 0.3| \leq 0.3\} \\ C_3 &= \{z \in \mathbf{C} / |z - 0.1| \leq 0.2\} \end{aligned}$$

En este caso no sabemos si $\rho(A) < 1$, tampoco lo podemos ver si utilizamos las cotas dadas por las normas $\|\cdot\|_1$ y $\|\cdot\|_\infty$, pero si multiplicamos la matriz A por una matriz diagonal $D = \text{diag}(2, 1, 1)$, tenemos

$$D^{-1}AD = \begin{pmatrix} 0.1 & 0.0 & 0.5 \\ 0.4 & 0.3 & 0.1 \\ 0.2 & 0.1 & 0.1 \end{pmatrix}$$

entonces, los valores propios están contenidos en los discos

$$\begin{aligned} C_1 &= \{z \in \mathbf{C} / |z - 0.1| \leq 0.5\} \\ C_2 &= \{z \in \mathbf{C} / |z - 0.3| \leq 0.5\} \\ C_3 &= \{z \in \mathbf{C} / |z - 0.1| \leq 0.3\} \end{aligned}$$

Ahora sí que podemos asegurar que $\rho(A) < 1$.

7.3 Transformación de matrices a forma reducida

Antes de entrar en los métodos específicos para el cálculo de valores propios, veremos primero las técnicas utilizadas para reducir una matriz real cualquiera a una matriz Hessenberg superior, utilizando un número finito de transformaciones ortogonales. Si la matriz inicial es simétrica,

entonces se obtiene una matriz tridiagonal simétrica. Recordemos que una matriz **Hessenberg superior** es tal que sus elementos a_{ij} son nulos para $i = 3 \div n$ y $j = 1 \div i - 2$.

Como las transformaciones realizadas son ortogonales, los valores propios de la matriz final serán los mismos y los vectores propios estarán modificados por la transformación ortogonal aplicada.

7.3.1 Método de Givens

El **método de Givens** (1954) consiste en utilizar rotaciones adecuadas para ir obteniendo ceros en los elementos por debajo de la primera subdiagonal de la matriz. Es decir, si lo hacemos por columnas, pondremos ceros a las posiciones

$$(3, 1), \dots, (n, 1), (4, 2), \dots, (n, 2), \dots, (n, n-2)$$

Consideremos, en general, un giro $G_\phi(p, q)$ de ángulo ϕ en el plano (p, q) sobre la matriz B :

$$D = G_\phi^T(p, q) B G_\phi(p, q)$$

donde la matriz del giro es

$$\begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & \cos \phi & & -\sin \phi & & 0 \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix}$$

con los $\sin \phi$ y $\cos \phi$ colocados en las intersecciones de las filas p, q con las columnas p, q .

La matriz resultante, D , solamente tiene modificadas las columnas y las filas p y q de la siguiente manera:

$$\begin{aligned} d_{pi} &= b_{pi} \cos \phi + b_{qi} \sin \phi \\ d_{qi} &= b_{qi} \cos \phi - b_{pi} \sin \phi \\ d_{ip} &= b_{ip} \cos \phi + b_{iq} \sin \phi \\ d_{iq} &= b_{iq} \cos \phi - b_{ip} \sin \phi \quad \text{para } i \neq p, q \end{aligned} \tag{7.1}$$

$$d_{pp} = b_{pp} \cos^2 \phi + (b_{pq} + b_{qp}) \cos \phi \sin \phi + b_{qq} \sin^2 \phi$$

$$d_{qq} = b_{qq} \cos^2 \phi - (b_{pq} + b_{qp}) \cos \phi \sin \phi + b_{pp} \sin^2 \phi$$

$$d_{pq} = b_{pq} \cos^2 \phi + (b_{qq} - b_{pp}) \cos \phi \sin \phi - b_{qp} \sin^2 \phi$$

$$d_{qp} = b_{qp} \cos^2 \phi + (b_{qq} - b_{pp}) \cos \phi \sin \phi - b_{pq} \sin^2 \phi$$

Entonces, el método de Givens consiste en realizar una rotación en el plano p, q para obtener un cero en el lugar $(q, p - 1)$. Si, como habíamos dicho al principio, operamos por columnas, tenemos las siguientes condiciones (ver la figura 7.1):

$$\begin{aligned} q &\geq p + 1 \\ b_{ij} &= 0 \quad \text{para } j = 1 \div p - 2, i = j + 2 \div n \\ b_{i,p-1} &= 0 \quad \text{para } i = p + 1 \div q - 1 \text{ (si } q \neq p + 1) \end{aligned}$$

	*	*	*	⊗	*	⊗	*
	*	*	*	⊗	*	⊗	*
	0	*	*	⊗	*	⊗	*
$p \rightarrow$	0	0	⊗	⊗	⊗	⊗	⊗
	0	0	0	⊗	*	⊗	*
$q \rightarrow$	0	0	⊗	⊗	⊗	⊗	⊗
	0	0	*	⊗	*	⊗	*
				↑		↑	
				p		q	

Fig. 7.1 Método de Givens.

- * Elemento no obligatoriamente nulo.
- Elemento a colocar a cero.
- ⊗ Elementos que quedan modificados por el giro.

y lo que queremos es $d_{q,p-1} = 0$, por lo tanto, de (7.1) deducimos

$$\tan \phi = \frac{b_{q,p-1}}{b_{p,p-1}}$$

o bien que

$$\sin \phi = \frac{b_{q,p-1}}{\sqrt{b_{p,p-1}^2 + b_{q,p-1}^2}} \quad \cos \phi = \frac{b_{p,p-1}}{\sqrt{b_{p,p-1}^2 + b_{q,p-1}^2}}$$

Además, debido a la forma como escogemos el giro, los elementos que ya eran cero lo siguen siendo.

Si hacemos un recuento de operaciones necesarias, se puede comprobar ([Wil65]) que se necesitan del orden de $\frac{10}{3}n^3$ multiplicaciones para el caso no simétrico y $\frac{4}{3}n^3$ para el caso simétrico.

Ejemplo. Aplicar el método de Givens a la matriz

$$A = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 2 & 2 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 2 & 1 & 1 & 1 \end{pmatrix}$$

Para poner a cero el elemento $a_{3,1}$ hacemos un giro en el plano 2, 3 con un ángulo ϕ tal que $\sin \phi = 1/\sqrt{5}$ y $\cos \phi = 2/\sqrt{5}$, y obtenemos la primera transformación de la matriz:

$$A_1 = \begin{pmatrix} 1.0000 & 2.2361 & 0 & 2.0000 \\ 2.2361 & 2.0000 & -1.0000 & 1.3416 \\ 0 & -1.0000 & 1.0000 & 0.4472 \\ 2.0000 & 1.3416 & 0.4472 & 1.0000 \end{pmatrix}$$

El siguiente elemento es el $a_{4,1}$, el seno y coseno adecuados son, respectivamente, $s = 2/3$ y $c = 0.7454$ y se obtiene la nueva matriz:

$$A_2 = \begin{pmatrix} 1.0000 & 3.0000 & 0 & 0 \\ 3.0000 & 2.3333 & -0.4472 & 0.1491 \\ 0 & -0.4472 & 2.0000 & 1.0000 \\ 0 & 0.1491 & 1.0000 & -0.3333 \end{pmatrix}$$

Finalmente, solamente queda poner a cero el elemento $a_{4,2}$ utilizando $s = 0.3163$ y $c = -0.9487$ para obtener la matriz:

$$A_3 = \begin{pmatrix} 1.0000 & 3.0000 & 0 & 0 \\ 3.0000 & 2.3333 & -0.4714 & 0 \\ 0 & -0.4714 & 1.1667 & 1.5000 \\ 0 & 0 & 1.5000 & 0.5000 \end{pmatrix}$$

7.3.2 Método de Householder

El **método de Householder** (1958) consiste en aplicar transformaciones ortogonales mediante matrices de Householder que ponen ceros en toda una columna cada vez.

Una **matriz de Householder** asociada al vector w es

$$P(w) = I - 2ww^T \quad \text{con} \quad w^Tw = 1$$

estas matrices son simétricas:

$$P^T(w) = I - 2(ww^T)^T = I - 2ww^T = P(w)$$

y ortogonales:

$$P^T(w)P(w) = P^2(w) = (I - 2ww^T)^2 = I - 4ww^T + 4w(w^Tw)w^T = I$$

Consideremos, ahora, un vector cualquiera v ; sea

$$s^2 = v^Tv, \quad u = v - se_1yw = \frac{1}{(u^Tu)^{1/2}}u,$$

de manera que $u \neq 0$ y $e_1 = (1, 0, \dots, 0)^T$. Si aplicamos la matriz de Householder asociada al vector unitario w al vector v y hacemos operaciones, obtenemos

$$P(w)v = v - \frac{2(s^2 - sv_1)}{u^Tu}(v - se_1) = v - (v - se_1) = se_1 \quad (7.2)$$

donde v_1 denota la primera componente del vector v .

Es decir, obtenemos un vector con todas sus componentes nulas, excepto la primera, que vale $s = (v^T v)^{1/2}$. Esto es lo que utilizaremos para poner ceros a cada columna de la matriz.

Consideremos ahora $A_0 = A$ y supongamos que hemos llegado hasta

$$A_k = \left(\begin{array}{c|c} H_k & B_k \\ \hline 0 & \bar{a}^{(k)} \end{array} \right) C_k$$

donde H_k es ya una matriz Hessenberg superior de dimensión $k + 1$, B_k es una matriz de dimensión $(k + 1) \times (n - k - 1)$, C_k es de dimensión $(n - k - 1) \times (n - k - 1)$ y $\bar{a}^{(k)}$ son las $n - k - 1$, últimas componentes de la columna $k + 1$ de la matriz A_k .

Entonces consideramos la matriz de Householder Q_k de dimensión $n - k - 1$ asociada al vector $w^{(k)}$ tal que

$$\begin{aligned} w^{(k)} &= \frac{1}{(2s_k)(1/2)(s_k - a_{k+2,k+1}^{(k)})^{1/2}} u^{(k)} \quad \text{con } u^{(k)} = \bar{a}^{(k)} - s_k e_1 \\ s_k^2 &= \|\bar{a}^{(k)}\|_2^2 = \sum_{i=k+2}^n (a_{i,k+1}^{(k)})^2 \end{aligned}$$

que nos define la matriz

$$P_k = \left(\begin{array}{c|c} I & 0 \\ \hline 0 & Q_k \end{array} \right) = I - \frac{1}{s_k(s_k - a_{k+2,k+1}^{(k)})} v^{(k)} v^{(k)T}$$

con $v^{(k)} = (0, \dots, 0, u^{(k)T})^T$.

Entonces, de 7.2, la nueva matriz A_{k+1} es

$$A_{k+1} = P_k A_k P_k = \left(\begin{array}{c|c} H_k & B_k Q_k \\ \hline 0 & s_k e_1 \end{array} \right) Q_k C_k Q_k$$

y se obtiene una columna más en la forma Hessenberg superior, y así sucesivamente hasta llegar a la columna $n - 2$, es decir, la matriz A_{n-2} .

El número total de multiplicaciones necesarias para la reducción completa de una matriz no simétrica es del orden $\frac{5}{3}n^3$ y para matrices simétricas $\frac{2}{3}n^3$ ([Wil65]), es decir, aproximadamente la mitad de las del método de Givens.

Ejemplo. Aplicar el método de Householder a la matriz

$$A = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 2 & 2 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 2 & 1 & 1 & 1 \end{pmatrix}$$

En primer lugar hacemos $A_0 = A$; seguidamente, para hallar $A_1 = P_0 A_0 P_0$ nos hacen falta los valores de

$$s_0 = 3.0000 \quad u^{(0)} = \begin{pmatrix} -1.0000 \\ 1.0000 \\ 2.0000 \end{pmatrix}$$

$$Q_0 = I - 0.3333 u^{(0)} u^{(0)T} = \begin{pmatrix} 0.6667 & 0.3333 & 0.6667 \\ 0.3333 & 0.6667 & -0.6667 \\ 0.6667 & -0.6667 & -0.3333 \end{pmatrix}$$

y se obtiene la nueva matriz

$$A_1 = \begin{pmatrix} 1.0000 & 3.0000 & 0 & 0 \\ 3.0000 & 2.3333 & -0.3333 & 0.3333 \\ 0 & -0.3333 & -0.6667 & -0.3333 \\ 0 & 0.3333 & -0.3333 & 2.3333 \end{pmatrix}$$

La segunda y última transformación vendrá definida por

$$s_1 = 0.4714, \quad u^{(1)} = \begin{pmatrix} -0.8047 \\ 0.3333 \end{pmatrix}$$

$$Q_1 = I - 2.6360 u^{(1)} u^{(1)T} = \begin{pmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{pmatrix}$$

que nos da la matriz final

$$A_2 = P_1 A_1 P_1 = \begin{pmatrix} 1.0000 & 3.0000 & 0 & 0 \\ 3.0000 & 2.3333 & 0.4714 & 0 \\ 0 & 0.4714 & 1.1667 & 1.5000 \\ 0 & 0 & 1.5000 & 0.5000 \end{pmatrix}$$

la cual, exceptuando los signos, es igual a la obtenida por el método de Givens.

Ejercicios.

1. Aplicar los métodos de Givens y de Householder a la matriz

$$\begin{pmatrix} 0 & 12 & 16 & -15 \\ 12 & 388 & 309 & 185 \\ 16 & 309 & 312 & 80 \\ -15 & 185 & 80 & -600 \end{pmatrix}$$

2. Estudiar cómo se aplicarían los métodos de Givens y de Householder para hallar la factorización de una matriz cualquiera como producto de una matriz ortogonal por una matriz triangular superior. Esta factorización se llama **factorización QR**.

7.3.3 Comparación de los dos métodos

Los dos métodos utilizan transformaciones ortogonales para hallar una matriz Hessenberg superior. Un teorema de ([Wil65]) nos demuestra la equivalencia de los dos métodos:

Teorema 7.3 Sean $AQ_1 = Q_1H_1$ y $AQ_2 = Q_2H_2$, donde las Q_i son unitarias y las H_i Hessenberg superiores; si las matrices Q_1 y Q_2 tienen la misma primera columna, entonces

$$Q_2 = Q_1D \quad H_2 = D^H H_1 D$$

donde D es una matriz diagonal con los elementos de módulo igual a la unidad.

Este teorema nos justifica la equivalencia de las matrices Hessenberg obtenidas por los dos métodos, pero las matrices utilizadas en las distintas transformaciones para obtener una columna de ceros pueden ser totalmente distintas. Por lo tanto, los resultados numéricos pueden ser muy contradictorios debido al comportamiento distinto de los errores.

7.4 Métodos basados en el polinomio característico

Cuando tenemos la matriz reducida a una forma más simple, resulta un poco más sencillo calcular su polinomio característico y, de éste, calcular sus raíces, utilizando alguno de los métodos del capítulo 6.

7.4.1 Valores y vectores propios para matrices tridiagonales simétricas

Consideremos la matriz tridiagonal simétrica

$$A = \begin{pmatrix} a_1 & b_2 & & & & \\ b_2 & a_2 & b_3 & & & \\ & b_3 & a_3 & b_4 & & \\ & & \ddots & \ddots & \ddots & \\ & & & b_{n-1} & a_{n-1} & b_n \\ & & & & b_n & a_n \end{pmatrix}$$

con todos los elementos b_i no nulos, ya que en este caso dividiríamos A en submatrices de orden inferior.

En lugar de calcular el polinomio $\det(A - \lambda I)$, calcularemos $\det(\lambda I - A)$, que tiene las mismas raíces. Esto lo haremos de forma recurrente:

$$\begin{aligned} p_0(\lambda) &= 1 & p_1(\lambda) &= \lambda - a_1 \\ p_i(\lambda) &= (\lambda - a_i)p_{i-1}(\lambda) - b_i^2 p_{i-2}(\lambda) & \text{para } i = 2 \div n \end{aligned}$$

Hay que destacar que el polinomio $p_{j+1}(\lambda)$ corresponde al determinante de la matriz formada por las j primeras filas y columnas de $\lambda I - A$. Está claro que el polinomio $p_j(\lambda)$ es de grado j .

Una de las propiedades más importantes de esta sucesión de polinomios es que es una sucesión de Sturm ([Gou73]) y entonces le podemos aplicar lo que ya sabemos del capítulo 6 para enunciar el siguiente teorema:

Teorema 7.4 Los valores propios de la matriz A son todos reales y distintos (bajo la condición citada anteriormente de $b_i \neq 0$). Además, el número de valores propios de A entre a y b es igual a la diferencia de cambios de signo de la sucesión de polinomios $\{p_0(\lambda), \dots, p_n(\lambda)\}$, evaluada en b y en a .

Finalmente, si para calcular las raíces del polinomio $p_n(\lambda)$ utilizamos el método de Newton, su derivada también se puede calcular de forma recurrente:

$$\begin{aligned} p'_0(\lambda) &= 0 & p'_1(\lambda) &= 1 \\ p'_i(\lambda) &= p_{i-1}(\lambda) + (\lambda - a_i)p'_{i-1}(\lambda) - b_i^2 p'_{i-2}(\lambda) & \text{para } i = 2 \div n \end{aligned}$$

Para calcular los vectores propios asociados a los valores propios ya calculados, hará falta resolver el sistema

$$\begin{aligned} (\lambda - a_1)v_1 - b_2v_2 &= 0 \\ &\vdots \\ -b_ib_{i-1} + (\lambda - a_i)v_i - b_{i+1}v_{i+1} &= 0 \\ &\vdots \\ -b_nv_{n-1} + (\lambda - a_n)v_n &= 0 \end{aligned}$$

con la condición $v_1 \neq 0$ (se puede tomar $v_1 = 1$) para tal que v no sea nulo.

Para matrices Hessenberg superiores se pueden construir recurrencias semejantes un poco más complicadas, pero entonces no tenemos la seguridad que las raíces del polinomio sean todas reales y será necesario, por lo tanto, utilizar algún método de búsqueda de raíces complejas.

Ejemplo. Considerar la matriz de Hilbert de orden cinco

$$A = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 \end{pmatrix}$$

En primer lugar, por el método de Householder, la transformamos en una matriz tridiagonal (todos los cálculos se han realizado con 16 cifras, a pesar de estar escritas con menos):

$$H = \begin{pmatrix} 1.046 \cdot 10^{-4} & 2.292 \cdot 10^{-4} & 0 & 0 & 0 \\ 2.292 \cdot 10^{-4} & 7.923 \cdot 10^{-3} & 1.429 \cdot 10^{-2} & 0 & 0 \\ 0 & 1.429 \cdot 10^{-2} & 2.556 \cdot 10^{-1} & -3.302 \cdot 10^{-1} & 0 \\ 0 & 0 & -3.302 \cdot 10^{-1} & 1.413 \cdot 10^0 & -3.222 \cdot 10^{-1} \\ 0 & 0 & 0 & -3.222 \cdot 10^{-1} & 1.111 \cdot 10^{-1} \end{pmatrix}$$

La sucesión de Sturm correspondiente al polinomio $\det(\lambda I - H)$ es

$$\begin{aligned} p_0(\lambda) &= 1 & p_1(\lambda) &= \lambda - 1.04566 \times 10^{-4} \\ p_2(\lambda) &= \lambda^2 - 0.00802753\lambda + 7.75939 \times 10^{-7} \\ p_3(\lambda) &= \lambda^3 - 0.263580\lambda^2 + 0.00184798\lambda - 1.76935 \times 10^{-7} \\ p_4(\lambda) &= \lambda^4 - 1.67619\lambda^3 + 0.265159\lambda^2 - 0.00173545\lambda + 1.65344 \times 10^{-7} \\ p_5(\lambda) &= \lambda^5 - 1.78730\lambda^4 + 0.347591\lambda^3 - 0.00383508\lambda^2 + 0.00000115293\lambda - 3.74930 \times 10^{-12} \end{aligned}$$

El polinomio característico que se ha obtenido (con las 16 cifras) tiene un error máximo del orden de 10^{-16} respecto al polinomio característico exacto. De hecho, el error oscila desde este 10^{-16} en el coeficiente de grado cuatro hasta 10^{-23} en el término independiente. Es decir, hemos obtenido el polinomio característico de la matriz de Hilbert de orden cinco con once cifras exactas en el peor de los casos.

La sucesión de derivadas es

$$\begin{aligned} p'_0(\lambda) &= 0 & p'_1(\lambda) &= 1 \\ p'_2(\lambda) &= 2\lambda + 0.00781840 \\ p'_3(\lambda) &= 3\lambda^2 + 0.510895\lambda + 0.00179453 \\ p'_4(\lambda) &= 4\lambda^3 + 4.48515\lambda^2 + 0.507289\lambda + 0.00168240 \\ p'_5(\lambda) &= 5\lambda^4 + 3.25340\lambda^3 + 0.9.59365\lambda^2 + 0.00327591 + 8.06670 \times 10^{-7} \end{aligned}$$

Seguidamente, podemos evaluar la sucesión de Sturm en distintos puntos para poder aislar las raíces

	0.0	0.0001	0.0003	0.0004	0.01	0.02	0.2	0.3	1.5	1.6
p_0	+	+	+	+	+	+	+	+	+	+
p_1	-	-	+	+	+	+	+	+	+	+
p_2	+	-	-	-	+	+	+	+	+	+
p_3	-	+	+	+	-	-	-	+	+	+
p_4	+	-	-	-	+	+	-	-	-	+
p_5	-	+	+	-	-	+	+	-	-	+
cambios	5	4	4	3	3	2	2	1	1	0

Finalmente, si calculamos las raíces de este polinomio con cinco cifras exactas, utilizando el método de Newton y de la bisección, tenemos

$$\begin{aligned} \lambda_1 &= 1.5670 & \lambda_2 &= 0.20853 & \lambda_3 &= 0.011407 \\ \lambda_4 &= 0.00030589 & \lambda_5 &= 0.0000032879 \end{aligned}$$

Una vez más, todos estos cálculos han sido realizados con el software numérico MATLAB.

7.5 Métodos iterativos

Dentro de este grupo de métodos, veremos los dos más clásicos: el método de la potencia con sus variantes y el método de Jacobi. Los dos métodos no utilizan ningún tipo de factorización de la matriz. El primero va multiplicando por la misma matriz un vector inicial para obtener, de esta manera, una sucesión de vectores convergentes a un vector propio asociado al valor propio más grande. El segundo, el método de Jacobi, que sirve únicamente para matrices simétricas, consiste en multiplicar a la derecha y a la izquierda de la matriz inicial, por giros adecuados, y conseguir, de esta forma, una sucesión de matrices simétricas convergentes a una matriz diagonal.

7.5.1 Métodos de la potencia

Como ya se ha comentado anteriormente, este método consiste en generar una sucesión de vectores que converge a un vector propio asociado al valor propio de mayor módulo. Solamente nos da este valor propio. La sucesión es la siguiente:

$$t_k = At_{k-1}, \quad \text{es decir} \quad t_k = A^k t_0.$$

Tal como está escrita, el módulo de los elementos de esta sucesión tienden a cero o a infinito según como sea $\|A\|$. Entonces, es mejor normalizar el vector a cada iteración:

$$\begin{aligned} z_{k+1} &= At_k \\ t_{k+1} &= \frac{z_{k+1}}{\|z_{k+1}\|} \quad (\text{por ejemplo } \|\cdot\|_\infty) \end{aligned}$$

Para estudiar la convergencia de este método supongamos que la matriz A es diagonalizable y sus valores propios son

$$|\lambda_1| = |\lambda_2| = \cdots = |\lambda_r| > |\lambda_{r+1}| \geq \cdots \geq |\lambda_n|$$

con $\lambda_1 = \lambda_2 = \cdots = \lambda_r$, es decir: hay un valor propio de módulo máximo y de multiplicidad r . Denotaremos por x_j los vectores propios linealmente independientes asociados a los correspondientes valores propios. Entonces, expresando t_0 en función de estos vectores,

$$t_0 = \sum_{j=1}^n \alpha_j x_j$$

podemos escribir

$$\begin{aligned} t_k &= \frac{A^k t_0}{\|A^k t_0\|} = \frac{1}{\|A^k t_0\|} \sum_{j=1}^n \alpha_j \lambda_j^k x_j \\ &= \frac{\lambda_1^k}{\|A^k t_0\|} \left[\sum_{j=1}^r \alpha_j x_j + \sum_{j=r+1}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k x_j \right] \end{aligned}$$

Tomando límites, es claro que

$$\lim_{k \rightarrow \infty} t_k = \pm \frac{\sum_{j=1}^r \alpha_j x_j}{\left\| \sum_{j=1}^r \alpha_j x_j \right\|}$$

es decir, el vector t_k converge a una cierta combinación lineal de los vectores propios asociados a los valores propios dominantes. Si consideramos las sucesiones de cocientes

$$q_i^{(k)} = \frac{z_i^{(k+1)}}{t_i^{(k)}} \quad i = 1 \div n$$

donde $v_i^{(k)}$ denota la i -ésima componente del vector v_k y, para ciertas i buenas, su límite es

$$\lim_{k \rightarrow \infty} q_i^{(k)} = \lambda_1$$

Todo esto es cierto si existe una j entre 1 y r , tal que $\alpha_j \neq 0$. En caso contrario, es decir, si el vector inicial t_0 no tuviese ninguna componente en el subespacio generado por los vectores propios dominantes, analíticamente, esto convergería al valor propio dominante inmediatamente inferior, λ_2 ; pero, a causa de los errores de redondeo que se acumulan en los cálculos, enseguida aparece una componente no nula en el subespacio de los valores propios dominantes y, por lo tanto, la sucesión convergerá hacia el valor propio dominante λ_1 . La convergencia de estas sucesiones viene dada por el cociente $\left| \frac{\lambda_{r+1}}{\lambda_1} \right|$; entonces, si los valores propios λ_1 y λ_{r+1} son próximos en módulo, la convergencia será lenta.

Finalmente, en el caso de matrices simétricas, para calcular el valor propio dominante es más aconsejable utilizar la sucesión definida por el **cociente de Rayleigh**:

$$\sigma_k = \frac{z_k^T A z_k}{z_k^T z_k}$$

que haciendo operaciones se reduce a

$$\sigma_k = t_k^T A t_k = t_k^T z_{k+1}$$

De esta manera la convergencia de σ_k hacia λ_1 viene dada por el cociente $\left| \frac{\lambda_2}{\lambda_1} \right|^2$. Este cociente también se puede utilizar en cualquier tipo de matriz, a pesar de que su convergencia no sea tan grande.

Ejemplo. Considerar la matriz

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{pmatrix}$$

aplicar el método de las potencias normalizando por la norma infinito, es decir $\|x\|_\infty = \max_{i=1 \div n} |x_i|$. Comenzar por el vector $t_0 = (1, 1, 1)^T$ y después de veinticinco iteraciones obtenemos

$$t_{25} = \begin{pmatrix} -0.50000 \\ 0.50000 \\ 1.00000 \end{pmatrix} \quad z_{26} = \begin{pmatrix} -1.49998 \\ 1.50002 \\ 3.00003 \end{pmatrix}$$

Los correspondientes $q_i^{(k)}$ son

$$q_1^{(25)} = 3.00006 \quad q_2^{(25)} = 3.00003 \quad q_3^{(25)} = 3.00003$$

Así pues, obtenemos el valor propio dominante $\lambda_1 = 3$.

Ejercicio. Estudiar la convergencia de la sucesión σ_k para cualquier matriz.

Desplazamiento del origen

Si, en lugar de aplicar el método de la potencia a la matriz A , lo aplicamos a $B = A - pI$, encontraremos el valor propio dominante de B . Denotamos por λ_i y por μ_i con $i = 1 \div n$ los valores propios de las matrices A y B , respectivamente. Es claro que

$$\mu_i = \lambda_i - p$$

Si escogemos p de manera que $\mu_1 = \lambda_1 - p$ aún sea dominante y

$$\left| \frac{\lambda_2}{\lambda_1} \right| > \left| \frac{\lambda_2 - p}{\lambda_1 - p} \right|$$

la convergencia del método sobre la matriz B será más rápida que sobre A .

Ejemplo. Considerar la misma matriz de antes

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{pmatrix}$$

con valores propios 1, 2 y 3. Aplicar una traslación igual a 1, es decir,

$$B = A - I = \begin{pmatrix} 0 & 0 & -1 \\ 1 & 1 & 1 \\ 2 & 2 & 2 \end{pmatrix}$$

Después de catorce iteraciones obtenemos

$$t_{14} = \begin{pmatrix} -0.50000 \\ 0.50000 \\ 1.00000 \end{pmatrix} \quad z_{15} = \begin{pmatrix} -1.00000 \\ 1.00002 \\ 2.00004 \end{pmatrix}$$

y las correspondientes $q_i^{(k)}$ son

$$q_1^{(14)} = 2.00010 \quad q_2^{(14)} = 2.00005 \quad q_3^{(14)} = 2.00005$$

De esta forma hemos obtenido el valor propio dominante $\lambda_1 = 2 - (-1) = 3$ de la matriz A con catorce iteraciones del método de la potencia. Recordemos que antes habíamos necesitado veinticinco para obtener una aproximación semejante.

Cálculo del valor propio más pequeño

Supongamos que tenemos una matriz A inversible; entonces, si denotamos por λ_i a los valores propios de A , los de A^{-1} serán $\mu_i = \frac{1}{\lambda_i}$. Si el valor propio de A , λ_n , de módulo mínimo es único, el método de la potencia sobre A^{-1} nos dará su valor propio dominante, es decir $\mu_n = \frac{1}{\lambda_n}$. Este método también se conoce como el **método de la potencia inversa**.

Ejemplo. Calcularemos el valor propio de menor módulo de la matriz

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{pmatrix}$$

Para esto es necesario hallar primero su inversa

$$B = A^{-1} = \begin{pmatrix} 0.66667 & -0.33333 & 0.33333 \\ -0.16667 & 0.83333 & -0.33333 \\ -0.33333 & -0.33333 & 0.33333 \end{pmatrix}$$

a la cual aplicaremos el método de la potencia, y obtendremos, después de dieciséis iteraciones

$$t_{16} = \begin{pmatrix} -0.99994 \\ 1.00000 \\ -0.00012 \end{pmatrix} \quad z_{17} = \begin{pmatrix} -0.10000 \\ 1.00003 \\ -0.00006 \end{pmatrix}$$

y hallamos, por lo tanto, el valor propio de menor módulo $\lambda_3 = 1$.

Método de Wielandt

El método de Wielandt (1944) o de la **potencia inversa desplazada** es una combinación de los dos anteriores. Es muy útil para refinar valores propios y consiste en aplicar el método de la potencia sobre la matriz

$$B_\lambda = (A - \lambda I)^{-1}$$

que tiene por valores propios $\mu_i = \frac{1}{\lambda_i - \lambda}$.

Si el valor de λ es una buena aproximación del valor propio λ_i , resulta que μ_i se convierte en el valor propio dominante de la matriz B_λ . Si, además, la aproximación es suficientemente buena, la convergencia es muy rápida.

Ejemplo. Consideramos, una vez más, la misma matriz de los ejemplos anteriores:

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{pmatrix}$$

Le aplicamos cuatro iteraciones del método de la potencia para obtener los vectores

$$t_4 = \begin{pmatrix} -0.40762 \\ 0.50147 \\ 1.00000 \end{pmatrix} \quad z_5 = \begin{pmatrix} -1.40762 \\ 1.59531 \\ 3.18768 \end{pmatrix}$$

Esto nos permite dar una primera aproximación del valor propio de $\lambda = 3.2$. Seguidamente, la nueva matriz es

$$B_\lambda = (A - \lambda I)^{-1} = \begin{pmatrix} 3.33333 & 3.78788 & 2.27273 \\ -4.16667 & -4.62121 & -2.27273 \\ -8.33333 & -8.33333 & -5.00000 \end{pmatrix}$$

la cual, después de aplicar siete iteraciones del método de la potencia, nos da

$$t_7 = \begin{pmatrix} 0.50000 \\ -0.50000 \\ -1.00000 \end{pmatrix} \quad z_8 = \begin{pmatrix} -2.50000 \\ 2.50001 \\ 5.00001 \end{pmatrix}$$

Esto da un valor propio $\mu_1 = 5$ para la matriz B_λ , es decir, $\lambda_1 = \lambda - \frac{1}{\mu_1} = 3$ para la matriz A .

7.5.2 Métodos de deflación

Un método de deflación consiste en, dada una matriz A , un valor propio λ y su vector propio asociado v , hallar una nueva matriz \bar{A} de dimensión inferior, tal que sus valores propios sean los mismos que los de la matriz inicial A , excepto, como es natural, el ya conocido λ , o bien se puedan calcular de manera sencilla.

Utilizando conjuntamente el método de la potencia y deflaciones sucesivas, podemos ir hallando todos los valores propios de una matriz.

Deflación de Wielandt

Sea λ_1 un valor propio de la matriz A de dimensión n y x_1 un vector propio asociado tal que su primera componente es igual a 1. Consideremos la matriz siguiente:

$$B = A - x_1 a_1$$

donde a_1 denota la primera fila de la matriz A . Esta nueva matriz tiene toda su primera fila igual a cero. Si λ_2 y x_2 son, respectivamente, valor y vector propios de A con la primera componente de x_2 igual a 1, podemos escribir

$$B(x_1 - x_2) = \lambda_1 x_1 - \lambda_2 x_2 - \lambda_1 x_1 + \lambda_2 x_1 = \lambda_2(x_1 - x_2)$$

Por lo tanto, λ_2 es, también, valor propio de B con vector propio $x_1 - x_2$. Por otro lado, debido al hecho de que $x_1 - x_2$ tiene la primera componente nula, es suficiente considerar la submatriz de dimensión $(n-1) \times (n-1)$ obtenida de suprimir la primera fila y columna de B . Es decir,

una vez calculada la matriz B , eliminamos su primera fila y columna y obtenemos una nueva matriz \bar{B} . De ésta, calcularemos un nuevo valor propio λ_2 y un vector propio \bar{z} ; entonces, λ_2 será un nuevo valor propio de A y su vector propio lo obtendremos añadiendo un cero de primera componente a \bar{z} , que denotaremos por z , y haciendo

$$x_2 = x_1 + \frac{\lambda_2 - \lambda_1}{a_1 z} z \quad (7.3)$$

Ejercicios.

1. Demostrar que la deflación de Wielandt no conserva las matrices simétricas.
2. Demostrar que el vector x_2 definido en 7.3 es un vector propio de la matriz inicial A .

Ejemplo. Considerar la matriz

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{pmatrix}$$

que tiene un valor y vector propios

$$\lambda_1 = 3 \quad x_1 = \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix}$$

La matriz con la primera fila de ceros y la correspondiente submatriz de dimensión 2 son

$$B = A - x_1 a_1 = \begin{pmatrix} 0 & 0 & 0 \\ 2 & 2 & 0 \\ 4 & 2 & 1 \end{pmatrix} \quad \bar{B} = \begin{pmatrix} 2 & 0 \\ 2 & 1 \end{pmatrix}$$

Esta nueva matriz tiene un valor y vector propios

$$\lambda_2 = 2 \quad \bar{z} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

y, por lo tanto, $\lambda_2 = 2$ también es valor propio de A con el vector propio

$$x_2 = x_1 + \frac{\lambda_2 - \lambda_1}{a_1 z} z = \begin{pmatrix} 1.0 \\ -0.5 \\ -1.0 \end{pmatrix}$$

Finalmente, si hacemos una segunda deflación a la matriz B , obtenemos

$$C = \bar{B} - \bar{z} \bar{b}_1 = \begin{pmatrix} 0 & 0 \\ -2 & 1 \end{pmatrix}$$

que nos indica claramente la existencia de un tercer valor propio $\lambda_3 = 1$.

Deflación de Householder

El último método de deflación que comentamos utiliza transformaciones ortogonales para obtener la nueva matriz. Esto hace que la estabilidad numérica sea muy buena y, además, conserve la simetría de las matrices.

Como ya se ha visto anteriormente, podemos hallar una matriz de Householder P tal que

$$Pv = se_1 \quad \text{con} \quad s = \|v\|_2$$

donde $e_1 = (1, 0, \dots, 0)^T$.

Si v es un vector propio asociado al valor propio λ de la matriz A , podemos escribir

$$PAPe_1 = \frac{1}{s} PAP(Pv) = \frac{1}{s} PAv = \frac{\lambda}{s} Pv = \lambda e_1$$

Esta expresión nos dice que la primera columna de la matriz $B = PAP$ es precisamente el vector λe_1 , es decir,

$$B = \left(\begin{array}{c|c} \lambda & b^T \\ \hline 0 & \bar{B} \end{array} \right)$$

Entonces, solamente hay que considerar los valores propios de la submatriz \bar{B} que también lo son de A .

Ejercicios.

1. Demostrar que, si x es un vector propio de $B = P^{-1}AP$, Px lo es de A .
2. Demostrar que, si \bar{x} es un vector propio de la matriz \bar{B} asociado al valor propio $\bar{\lambda}$, entonces $\bar{\lambda}$ también es un valor propio de B con vector propio x tal que

$$x = \begin{pmatrix} \alpha \\ \bar{x} \end{pmatrix} \quad \text{para} \quad \alpha = \frac{b^T \bar{x}}{\bar{\lambda} - \lambda}.$$

3. Hacer lo mismo que en el ejercicio anterior, pero ahora suponiendo que la matriz A inicial es simétrica.

Ejemplo. Considerar la matriz

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{pmatrix}$$

que, como ya sabemos, tiene un valor y vector propios

$$\lambda_1 = 3 \quad x_1 = \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix}$$

La correspondiente matriz de Householder P_1 que hace $P_1 x_1 = s_1 e_1$ es

$$P_1 = \begin{pmatrix} 0.40825 & -0.40825 & -0.81650 \\ -0.40825 & 0.71835 & -0.56330 \\ -0.81650 & -0.56330 & -0.12660 \end{pmatrix}$$

que aplicada sobre la matriz A nos da

$$B = P_1 A P_1 = \begin{pmatrix} 3.00000 & 0.74680 & 3.12660 \\ 0 & 0.87340 & 0.56330 \\ 0 & -0.25320 & 2.12660 \end{pmatrix}$$

en donde consideramos la submatriz derecha inferior \bar{B} de dimensión 2. Un valor y vector propios de esta nueva matriz son

$$\lambda_2 = 2 \quad \bar{x}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

naturalmente λ_2 también es un valor propio de B con vector propio

$$x'_2 = \begin{pmatrix} -7 \\ 1 \\ 2 \end{pmatrix}$$

que respecto a A el valor propio es

$$x_2 = P_1 x'_2 = \begin{pmatrix} -4.89898 \\ 2.44949 \\ 4.89898 \end{pmatrix} \equiv \begin{pmatrix} -2 \\ 1 \\ 2 \end{pmatrix}$$

Finalmente, la última deflación estará definida por la matriz de Householder

$$P_2 = \begin{pmatrix} 0.44721 & 0.89443 \\ 0.89443 & -0.44721 \end{pmatrix}$$

que nos da la última matriz

$$C = P_2 \bar{B} P_2 = \begin{pmatrix} 2.00000 & -0.81650 \\ 0 & 1.00000 \end{pmatrix}$$

donde claramente obtenemos el tercer valor propio $\lambda_3 = 1$.

7.6 Método de Jacobi

Consideramos una matriz simétrica y real A ; como es sabido sus valores propios son reales y existe una matriz ortogonal P tal que $D = P^{-1}AP$ es una matriz diagonal que contiene los valores propios de A y, por lo tanto, las columnas de P son vectores propios asociados a estos valores propios. El método de Jacobi consiste en calcular unas sucesiones de matrices que convergen a P y a D , esto se realiza utilizando giros adecuados a cada paso.

En primer lugar definimos $A_0 = A$ y $P_0 = I$; seguidamente, dados A_k y P_k , escogemos un elemento no diagonal de A_k $a_{pq}^{(k)} \neq 0$ con $p < q$, que hay que hacer cero con un giro $G_\phi(p, q)$ del tipo 7.1; naturalmente, siendo la matriz simétrica, las expresiones son mucho más sencillas

$$\begin{aligned} a_{pp}^{(k+1)} &= a_{pp}^{(k)} \cos^2 \phi + 2a_{pq}^{(k)} \sin \phi \cos \phi + a_{qq}^{(k)} \sin^2 \phi \\ a_{qq}^{(k+1)} &= a_{pp}^{(k)} \sin^2 \phi - 2a_{pq}^{(k)} \sin \phi \cos \phi + a_{qq}^{(k)} \cos^2 \phi \\ a_{pq}^{(k+1)} &= a_{qp}^{(k+1)} = a_{pq}^{(k)} (\cos^2 \phi - \sin^2 \phi) + (a_{qq}^{(k)} - a_{pp}^{(k)}) \cos \phi \sin \phi \\ a_{pi}^{(k+1)} &= a_{ip}^{(k+1)} = a_{pi}^{(k)} \cos \phi + a_{qi}^{(k)} \sin \phi \quad \text{para } i = 1 \div n, i \neq p, q \\ a_{qi}^{(k+1)} &= a_{iq}^{(k+1)} = a_{iq}^{(k)} \cos \phi - a_{ip}^{(k)} \sin \phi \quad \text{para } i = 1 \div n, i \neq p, q \end{aligned} \quad (7.4)$$

y el resto de elementos no se modifican. El ángulo es tal que

$$\tan 2\phi = \frac{2a_{pq}^{(k)}}{a_{pp}^{(k)} - a_{qq}^{(k)}}$$

y

$$\begin{aligned} a_{pp}^{(k+1)} &= a_{pp}^{(k)} + a_{pq}^{(k)} \tan \phi \\ a_{qq}^{(k+1)} &= a_{qq}^{(k)} + a_{pq}^{(k)} \tan \phi \end{aligned}$$

Así pues, las nuevas matrices serán

$$A_{k+1} = G_\phi^T(p, q) A_k G_\phi(p, q) \quad P_{k+1} = P_k G_\phi(p, q)$$

y se obtendrá finalmente que

$$\lim_{k \rightarrow \infty} A_k = D \quad \lim_{k \rightarrow \infty} P_k = P$$

El **método clásico de Jacobi** es el que, como a $a_{pq}^{(k)}$, escoge el elemento no diagonal de mayor valor absoluto de toda la matriz. También se puede escoger consecutivamente por filas: este es el **método de Jacobi cíclico**.

Para justificar su convergencia consideramos

$$\tau^2(A_k) = \sum_{\substack{i,j=1 \\ i \neq j}}^n a_{ij}^{(k)2}$$

Como la transformación ortogonal solamente afecta a las filas y columnas p y q , según indica 7.4, se puede deducir

$$a_{pj}^{(k+1)2} + a_{jq}^{(k+1)2} = a_{pj}^{(k)2} + a_{jq}^{(k)2}$$

por lo tanto, $\tau^2(A_k)$ solamente cambia debido a la anulación de los elementos $a_{pq}^{(k)}$ y $a_{qp}^{(k)}$, es decir

$$\tau^2(A_{k+1}) = \tau^2(A_k) - 2a_{pq}^{(k)2}$$

Como el elemento escogido era el de mayor módulo de entre todos los elementos de debajo de la diagonales, tenemos que

$$\tau^2(A_{k+1}) \leq \tau^2(A_k) \left(1 - \frac{2}{n(n-1)}\right) < \tau^2(A_k) \exp\left(\frac{-1}{n(n-1)}\right)$$

Esto nos dice que, después de k iteraciones, $\tau(A)^2$ ha disminuido en un factor

$$\exp\left(\frac{-k}{n(n-1)}\right)$$

y, para k suficientemente grande, la matriz A_k será tan próxima a una matriz diagonal como queramos.

Ejemplo. Considerar la matriz

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix}$$

el elemento no diagonal de mayor valor absoluto es, por ejemplo, $a_{1,3}^{(0)} = 2$; esto nos da un ángulo $\phi_0 = \pi/4$ y la matriz del giro será

$$G_{\pi/4}(1,3) = \begin{pmatrix} 0.7071 & 0 & -0.7071 \\ 0 & 1 & 0 \\ 0.7071 & 0 & 0.7071 \end{pmatrix}$$

Con esta matriz el primer iterado es

$$A_1 = \begin{pmatrix} 3.0000 & 0.7071 & 0 \\ 0.7071 & 2.0000 & 0.7071 \\ 0 & 0.7071 & -1.0000 \end{pmatrix}$$

Si seguimos el proceso, obtenemos las matrices siguientes

$$\begin{aligned} G_{0.478}(1,2) &= \begin{pmatrix} 0.8881 & -0.9547 & 0 \\ 0.9547 & 0.8881 & 0 \\ 0 & 0 & 1 \end{pmatrix} & A_2 &= \begin{pmatrix} 3.3659 & 0 & 0.3250 \\ 0 & 1.6339 & 0.6280 \\ 0.3250 & 0.6280 & -1.0000 \end{pmatrix} \\ G_{0.223}(2,3) &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.9754 & -0.2206 \\ 0 & 0.2206 & 0.9754 \end{pmatrix} & A_3 &= \begin{pmatrix} 3.3659 & 0.0735 & 0.3170 \\ 0.0735 & 1.7801 & 0 \\ 0.3170 & 0 & -1.1447 \end{pmatrix} \\ G_{0.070}(1,3) &= \begin{pmatrix} 0.9976 & 0 & -0.0698 \\ 0 & 1 & 0 \\ 0.0698 & 0 & 0.9996 \end{pmatrix} & A_4 &= \begin{pmatrix} 3.3883 & 0.0733 & 0 \\ 0.0733 & 1.7801 & 0.0051 \\ 0 & 0.0051 & -1.1670 \end{pmatrix} \\ G_{0.046}(1,2) &= \begin{pmatrix} 0.9990 & -0.0454 & 0 \\ 0.0454 & 0.9990 & 0 \\ 0 & 0 & 1 \end{pmatrix} & A_5 &= \begin{pmatrix} 3.3915 & 0 & 0.0002 \\ 0 & 1.7767 & 0.0051 \\ 0.0002 & 0.0051 & -1.1670 \end{pmatrix} \end{aligned}$$

Como se puede ver en la última iteración, el valor absoluto mayor del elemento no diagonal es ya de orden 10^{-3} .

7.7 Métodos de factorización

Como ya se ha comentado al principio, los métodos de factorización están basados en aplicar repetidamente algún tipo de factorización de la matriz. Básicamente hay dos métodos:

- El LR, basado en la descomposición en dos matrices triangulares, una superior y otra inferior.
- El QR, que utiliza la factorización en una matriz ortogonal y en otra de triangular superior.

Los dos métodos se aplican, generalmente, sobre matrices ya reducidas a forma Hessenberg superior, que se conservan durante toda la iteración. El QR es el más potente y de aplicación más general (si no este método, alguna de sus variantes).

Desde un punto de vista general, podemos decir que estos métodos empiezan por $A_1 = A$ y, conocida la matriz A_k , se calcula una factorización de ésta

$$A_k = F_k G_k \quad \text{con} \quad F_k \text{ no singular,} \quad (7.5)$$

para expresar la matriz siguiente como

$$A_{k+1} = G_k F_k \quad (7.6)$$

Teorema 7.5 Suponiendo que existen todas las factorizaciones 7.5, la sucesión de matrices generada por la ecuación 7.6 tiene las propiedades siguientes:

1. Todas las matrices A_k tienen los mismos valores propios.
 2. Si $E_k = F_1 \cdots F_k$, entonces,
- $$A_{k+1} = E_k^{-1} A E_k$$
3. Si $H_k = G_k \cdots G_1$, entonces,
- $$A^k = E_k H_k$$

Demostración: Siendo F_k no singular,

$$A_{k+1} = G_k F_k = F_k^{-1} A_k F_k = \cdots = (F_1 \cdots F_k)^{-1} A_1 (F_1 \cdots F_k) = E_k^{-1} A E_k$$

que es lo que dice la segunda propiedad. Además, también nos demuestra que todas las matrices de la sucesión tienen los mismos valores propios.

Finalmente,

$$\begin{aligned} E_k H_k &= F_1 \cdots F_k G_k \cdots G_1 = E_{k-1} A_k H_{k-1} = A_1 E_{k-1} H_{k-1} \\ &\vdots \\ &= A_1^k = A^k \end{aligned}$$

que es la tercera propiedad. \square

Bajo ciertas condiciones, también tenemos un teorema de convergencia de este tipo de métodos.

Teorema 7.6 En una sucesión de matrices definida por 7.5 y 7.6, si $\{E_k\}$ converge hacia una matriz E_∞ no singular y cada G_k es una matriz triangular superior, entonces A_k converge hacia una matriz A_∞ triangular superior.

Demostración: Si $\{E_k\}$ converge, los límites siguientes existen

$$\begin{aligned}\lim_{k \rightarrow \infty} F_k &= \lim_{k \rightarrow \infty} (E_{k-1}^{-1} E_k) = I \\ G_\infty &= \lim_{k \rightarrow \infty} G_k = \lim_{k \rightarrow \infty} (E_k^{-1} A_1 E_{k-1}) = E_\infty^{-1} A_1 E_\infty\end{aligned}$$

Además, G_∞ es triangular superior. Finalmente,

$$A_\infty = \lim_{k \rightarrow \infty} A_k = \lim_{k \rightarrow \infty} (F_k G_k) = G_\infty$$

que también es triangular superior, y queda probado el teorema. \square

Seguidamente, estudiaremos con más detalle cada uno de los dos métodos comentados anteriormente.

7.7.1 Método LR

El método **LR** de Rutishauser (1958) utiliza la descomposición LU , que no siempre existe, de la matriz

$$A_k = L_k R_k$$

con L_k matriz triangular inferior con la diagonal de unos y R_k , triangular superior. Seguidamente se calcula el producto

$$A_{k+1} = R_k L_k$$

para obtener la matriz del paso siguiente.

Naturalmente, esta sucesión, suponiendo que siempre existe la descomposición LU de las A_k , cumple las propiedades del teorema 7.5. También podemos decir que, si $E_k = L_1 \cdots L_k$ converge hacia una matriz E_∞ no singular, entonces existe A_∞ y es una matriz triangular superior.

De todas maneras, este resultado, obtenido del teorema general de convergencia visto anteriormente, queda superado por un nuevo teorema formulado particularmente para el método LR que da, además, una razón de convergencia del método.

Teorema 7.7 Sea $A \in \mathcal{L}(\mathbf{R}^n)$, que cumple las condiciones:

1. El método LR es aplicable, es decir, se pueden realizar las descomposiciones LU de todos los iterados.
2. Los valores propios de A son de la forma

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|$$

3. A es diagonalizable, es decir, $\exists X, Y = X^{-1}$ tales que

$$D = YAX$$

es diagonal y, además, existe la factorización LU de las dos matrices X y Y .

Entonces,

$$\lim_{k \rightarrow \infty} A_k = \lim_{k \rightarrow \infty} R_k = R \quad \lim_{k \rightarrow \infty} L_k = I$$

donde R es una matriz triangular superior. además, la convergencia hacia cero de los elementos de debajo de la diagonal es

$$l_{i,j}^{(k)} = O\left(\left|\frac{\lambda_i}{\lambda_j}\right|^k\right) \quad \text{para } i > j \quad \text{y } k \rightarrow \infty$$

Demostración: Ver ([Wil65]) \square

Cabe destacar que, si la matriz A es hermitiana y definida positiva, sus valores propios serán reales y positivos, y el método siempre converge (ver [Wil65]).

Ejemplos.

1. Considerar la matriz

$$A = \begin{pmatrix} 7 & 6 \\ 3 & 4 \end{pmatrix}$$

Sus valores propios exactos son 10 y 1. Hacemos una primera factorización LR:

$$L_1 = \begin{pmatrix} 1.000000 & 0 \\ 0.428571 & 1.000000 \end{pmatrix} \quad R_1 = \begin{pmatrix} 7.000000 & 6.000000 \\ 0 & 1.428571 \end{pmatrix}$$

De estas dos matrices obtenemos la primera iteración haciendo

$$A_1 = R_1 L_1 = \begin{pmatrix} 9.571429 & 6.000000 \\ 0.612245 & 1.428571 \end{pmatrix}$$

Finalmente, después de ocho iteraciones obtenemos la matriz siguiente:

$$A_8 = \begin{pmatrix} 10.0 & 6.0 \\ 6.7 \times 10^{-9} & 1.0 \end{pmatrix}$$

2. Considerar la matriz

$$A = \begin{pmatrix} 1 & -1 & 1 \\ 4 & 6 & -1 \\ 4 & 4 & 1 \end{pmatrix}$$

sus valores propios son $\lambda_1 = 5$, $\lambda_2 = 2$ y $\lambda_3 = 1$. Los tres primeros pasos del método LR son

$$A_2 = \begin{pmatrix} 1.0 & -0.2 & 1.0 \\ 20.0 & 6.0 & -5.0 \\ 4.0 & 0.8 & 1.0 \end{pmatrix} \quad A_3 = \begin{pmatrix} 1.00 & -0.04 & 1.00 \\ 100.00 & 6.00 & -25.00 \\ 4.00 & 0.16 & 1.00 \end{pmatrix}$$

$$A_4 = \begin{pmatrix} 1.000 & -0.008 & 1.000 \\ 500.000 & 6.000 & -125.000 \\ 4.000 & 0.032 & 1.000 \end{pmatrix}$$

con estos tres pasos, ya se puede intuir que el elemento $a_{2,1}^{(k)}$ no converge hacia 0 y, por lo tanto, el método LR no es convergente, a pesar de existir las factorizaciones de las distintas iteraciones. Esta no convergencia es debida al hecho de que la matriz X de la diagonalización de A no tiene descomposición LU .

Para el caso de matrices reales con valores propios conjugados, tenemos el teorema siguiente:

Teorema 7.8 Sea A una matriz real con los valores propios

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$$

Si A cumple las propiedades

1. A cada paso, existe la descomposición LU d' A_k .
2. $|\lambda_i| > |\lambda_{i+1}|$, excepto para las parejas conjugadas, que denotaremos por

$$\lambda_c = \bar{\lambda}_{c-1}$$

3. A diagonalizable ($D = YAX$) y, además, $\det(X_{ii}) \det(Y_{ii}) \neq 0$ ($i = 1 \div n$), donde X_{ii} denota el menor principal de dimensión i .

Entonces

1. Los elementos de debajo de la diagonal $a_{ij}^{(k)}$ ($i > j$) de la matriz A_k convergen hacia 0, excepto para los elementos $a_{c,c-1}^{(k)}$.
2. $a_{rr}^{(k)} \rightarrow \lambda_r$ para los valores propios reales y simples.
3. Los elementos de encima de la diagonal $a_{ij}^{(k)}$ ($j > i$) tienen límite excepto los de las filas c y columnas $(c-1)$.
4. Las matrices 2×2

$$\begin{pmatrix} a_{c-1,c-1}^{(k)} & a_{c-1,c}^{(k)} \\ a_{c,c-1}^{(k)} & a_{c,c}^{(k)} \end{pmatrix}$$

no convergen, pero sus valores propios tienden hacia λ_{c-1} y λ_c .

Demostración: ver ([Wil65]) \square

Este método es muy costoso si se aplica a matrices llenas cualesquiera; solamente se debe recordar que, a cada paso, además del producto de dos matrices, es necesario hacer una eliminación gaussiana para hallar la descomposición LU . Por lo tanto, es importante aplicarlo a matrices Hessenberg superior y adaptar los algoritmos a matrices de este tipos.

El problema más grave, sin embargo, es que en alguno de los pasos no existe la descomposición LU , ya que, entonces, el algoritmo no puede continuar y el teorema de convergencia presupone la existencia de todas las descomposiciones LU . Por otro lado, cuando existe esta descomposición es, muchas veces, numéricamente inestable, y se pierde la precisión en el cálculo de los valores propios.

Ejercicios.

1. Demostrar que el método LR conserva las matrices Hessenberg superior.
2. Demostrar que el método LR también conserva las matrices hermitianas.

7.7.2 Método QR

El método QR de Francis (1961) utiliza la factorización de la matriz en

$$A_k = Q_k R_k$$

con Q_k matriz ortogonal (o unitaria) y R_k matriz triangular superior. La primera diferencia respecto al método LR es que esta factorización siempre existe. Si la matriz no es singular, exigiendo que los elementos diagonales de R_k sean positivos, la descomposición es siempre única.

Ejemplo. Considerar la matriz que no convergía por el método LR

$$A_0 = A = \begin{pmatrix} 1 & -1 & 1 \\ 4 & 6 & -1 \\ 4 & 4 & 1 \end{pmatrix}$$

La factorización QR de la matriz A_0 es

$$Q_1 = \begin{pmatrix} -0.1741 & 0.8301 & -0.5298 \\ -0.6963 & -0.4842 & -0.5298 \\ -0.6963 & 0.2767 & 0.6623 \end{pmatrix}$$

$$R_1 = \begin{pmatrix} -5.7445 & -6.7890 & -0.1740 \\ 0 & -2.6285 & 1.5909 \\ 0 & 0 & 0.6622 \end{pmatrix}$$

La matriz de la primera iteración es

$$A_1 = R_1 Q_1 = \begin{pmatrix} 5.8484 & -1.5292 & 6.5251 \\ 0.7224 & 1.7129 & 2.4462 \\ -0.4611 & 0.1832 & 0.4385 \end{pmatrix}$$

Finalmente, después de ocho iteraciones,

$$A_8 = \begin{pmatrix} 5.0000 & -1.7321 & -7.3485 \\ 0.0000 & 2.0000 & -1.4142 \\ 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$$

donde los ceros de la parte triangular inferior son de orden 10^{-6} .

De la tercera propiedad del teorema 7.5 se deduce que

$$A^k = E_k H_k$$

es la factorización QR de la matriz A^k , ya que E_k es ortogonal y H_k triangular superior.

Ejercicio.

Demostrar que el método QR conserva las matrices Hessenberg superior.

Teorema de convergencia

El primer teorema de convergencia del método QR es semejante al enunciado para el método LR.

Teorema 7.9 Sea $A \in \mathcal{L}(\mathbf{R}^n)$ que cumple las condiciones:

1. Los valores propios de A son de la forma

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|$$

2. A es diagonalizable, es decir, $\exists X, Y = X^{-1}$ tales que

$$D = YAX$$

es diagonal y, además, existe la factorización LU de la matriz Y (la demostración del teorema utiliza, también, la factorización QR de X , pero esta siempre existe).

Entonces, la matriz A_k tiende a una matriz triangular superior de la siguiente forma :

$$\begin{aligned}\lim_{k \rightarrow \infty} a_{ij}^{(k)} &= 0 \quad \text{para } i > j \\ \lim_{k \rightarrow \infty} a_{ii}^{(k)} &= \lambda_i \quad \text{para } i = 1 \div n\end{aligned}$$

Además, la convergencia de los elementos de debajo de la diagonal es

$$a_{i,j}^{(k)} = O\left(\left|\frac{\lambda_i}{\lambda_j}\right|^k\right) \quad \text{para } i > j \quad \text{y} \quad k \rightarrow \infty$$

Demostración: ver ([Wil65]). \square

Si la matriz Y , de las hipótesis del teorema, no tiene descomposición LU , no es grave, ya que se puede asegurar la descomposición con permutaciones, es decir,

$$PY = LU$$

entonces, A_k sigue convergiendo a una matriz triangular superior, y los elementos diagonales convergen a los valores propios, pero desordenados en función de la matriz de permutación P .

Ejemplo. Considerar la matriz

$$A_0 = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{pmatrix}$$

con valores propios 1, 2 y 3. La primera descomposición QR nos da las matrices

$$\begin{aligned} Q_1 &= \begin{pmatrix} -0.40825 & 0.70711 & -0.57735 \\ -0.40825 & -0.70711 & -0.57735 \\ -0.81650 & 0.00000 & 0.57735 \end{pmatrix} \\ R_1 &= \begin{pmatrix} -2.44948 & -2.44948 & -2.44948 \\ 0 & -1.41421 & -1.41421 \\ 0 & 0 & 1.73205 \end{pmatrix} \end{aligned}$$

Haciendo el producto $R_1 Q_1$, obtenemos la primera iteración del método

$$A_1 = \begin{pmatrix} 4.00000 & 0.00000 & 1.41421 \\ 1.73205 & 1.00000 & 0.00000 \\ -1.41421 & 0.00000 & 1.00000 \end{pmatrix}$$

Después de veintiuna iteraciones el resultado es el siguiente:

$$A_{21} = \begin{pmatrix} 3.0000205 & -1.4142157 & 2.8867446 \\ 0.0000041 & 0.9999971 & -0.8164882 \\ -0.0000051 & 0.0000036 & 1.9999824 \end{pmatrix}$$

donde se puede observar que los elementos de debajo de la diagonal son ceros con un error de orden 10^{-6} y los valores propios tienen un error de orden 10^{-5} . Además, los dos últimos valores propios, en orden decreciente, han aparecido permutados tal como dice el teorema, ya que las matrices X , $Y = X^{-1}$ y D del teorema son

$$\begin{aligned} X &= \begin{pmatrix} 0.40825 & 0.66667 & 0.70711 \\ -0.40825 & -0.33333 & -0.70711 \\ -0.81640 & -0.66667 & 0.00000 \end{pmatrix} & D &= \begin{pmatrix} 3.0 & 0 & 0 \\ 0 & 1.0 & 0 \\ 0 & 0 & 1.0 \end{pmatrix} \\ Y &= \begin{pmatrix} -2.44949 & -2.44949 & -1.22475 \\ 3.00000 & 3.00000 & 0.00000 \\ 0.00000 & -1.41421 & 0.70711 \end{pmatrix} \end{aligned}$$

y la descomposición LU de la matriz Y existe con permutaciones: $PY = LU$, donde

$$\begin{aligned} P &= \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} & L &= \begin{pmatrix} 1.00000 & 0 & 0 \\ -0.81650 & 1.00000 & 0 \\ 0.00000 & 0.00000 & 1.00000 \end{pmatrix} \\ U &= \begin{pmatrix} 3.00000 & 3.00000 & 0.00000 \\ 0 & -1.41421 & -0.70711 \\ 0 & 0 & -1.22475 \end{pmatrix} \end{aligned}$$

Si hay valores propios iguales

$$\lambda_r = \lambda_{r+1} = \cdots = \lambda_{r+s-1}$$

entonces el método converge a una matriz casi triangular, es decir, los elementos de debajo de la diagonal tenderán a 0, excepto los correspondientes a una submatriz de dimensión s y centrada a la diagonal principal. Éstos, sin embargo, también convergen.

Ejemplo. Considerar la matriz

$$A = \begin{pmatrix} 1 & -4 & -1 & -4 \\ 2 & 0 & 5 & -4 \\ -1 & 1 & -2 & 3 \\ -1 & 4 & -1 & 6 \end{pmatrix}$$

con valores propios $\lambda_{1,2,3,4} = 2, 1, 1, 1$. La primera iteración nos da

$$A_1 = \begin{pmatrix} -0.285715 & -8.374922 & -3.384735 & 4.523877 \\ 1.096438 & 4.419048 & 2.752026 & 3.672421 \\ -0.003552 & 0.925075 & 0.886534 & -1.552966 \\ -0.043499 & 0.067482 & 0.080027 & -0.019868 \end{pmatrix}$$

Si seguimos iterando, las iteraciones 25 y 50 son

$$A_{25} = \begin{pmatrix} 2.000019 & -8.912374 & 4.110696 & 3.825349 \\ 0.000002 & 1.091867 & 3.279258 & -4.345455 \\ 0.000000 & -0.001360 & 1.003565 & -0.619932 \\ 0.000000 & -0.000007 & 0.007025 & 0.904548 \end{pmatrix}$$

$$A_{50} = \begin{pmatrix} 2.000000 & -8.980473 & 4.261086 & 3.486514 \\ 0.000000 & 1.041820 & 2.919275 & -4.585316 \\ 0.000000 & -0.000307 & 1.000891 & -0.698284 \\ 0.000000 & 0.000000 & 0.001280 & 0.957290 \end{pmatrix}$$

donde los ceros de la primera columna son valores de orden inferior o igual a 10^{-10} y 10^{-12} para A_{25} y A_{50} , respectivamente. Según el teorema, los elementos de debajo de la diagonal de las columnas segunda y tercera no tendrían que converger a cero; esto, numéricamente, se traduce en una convergencia a cero muy lenta de estos elementos.

Finalmente, si hay valores propios del mismo módulo, pero diferentes, como por ejemplo el caso de valores propios conjugados,

$$|\lambda_r| = |\lambda_{r+1}| = \cdots = |\lambda_{r+s-1}|$$

entonces la convergencia es semejante a la del caso anterior; pero, ahora, la submatriz de dimensión s no converge a pesar de mantener su norma constante.

Ejemplo. Considerar la matriz

$$A = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

con valores propios $\lambda_1 = 2$ y $\lambda_{2,3} = 1 \pm i$. La primera iteración nos da

$$A_1 = \begin{pmatrix} 1.500000 & 0.150756 & 0.852803 \\ -1.658312 & 1.409091 & 0.899954 \\ 0.000000 & -0.514259 & 1.090909 \end{pmatrix}$$

Finalmente, a la trigésima iteración obtenemos

$$A_{30} = \begin{pmatrix} 2.000000 & -0.999985 & -0.707074 \\ -0.000030 & 1.000015 & 0.707118 \\ 0.000000 & -1.414235 & 0.999985 \end{pmatrix}$$

donde ya aparece claramente un valor propio igual a 2 en la primera columna. La submatriz cuadrada, formada por las dos últimas filas y columnas no es una matriz triangular superior, pero sus valores propios son los complejos conjugados: $1.000015 \pm 1.000015i$.

7.7.3 Traslación respecto del origen

Para acelerar la convergencia del método QR, una de las técnicas es la llamada traslación respecto del origen. Ésta consiste en factorizar

$$A_s - k_s I = Q_s R_s$$

en lugar de A_s y para cierto $k_s \in \mathbf{R}$ determinado. Seguidamente, se define la nueva matriz

$$A_{s+1} = R_s Q_s + k_s I$$

La sucesión de matrices obtenida sigue siendo de matrices semejantes, ya que

$$A_{s+1} = R_s Q_s + k_s I = Q_s^T (A_s - k_s I) Q_s + k_s I = Q_s^T A_s Q_s$$

De esta manera, la convergencia hacia cero de los elementos de debajo de la diagonal se obtiene a partir de los valores propios de la matriz $A_s - k_s I$ y, si k_s es una aproximación de un valor propio, por ejemplo $a_{nn}^{(s)}$, se obtiene una mayor convergencia hacia cero de los elementos asociados al valor propio, $a_{n,n-1}^{(s)}$ en el caso de una matriz Hessenberg.

Ejemplo. Considerar la matriz ya utilizada en un ejemplo anterior

$$A = \begin{pmatrix} 1 & -4 & -1 & -4 \\ 2 & 0 & 5 & -4 \\ -1 & 1 & -2 & 3 \\ -1 & 4 & -1 & 6 \end{pmatrix}$$

sus valores propios son $\lambda_1 = 2$ y $\lambda_{2,3,4} = 1$. En primer lugar, la convertimos en una matriz Hessenberg superior

$$H_1 = \begin{pmatrix} 1.000000 & 1.224745 & 5.366563 & 1.643168 \\ -2.449490 & -1.000000 & -7.120393 & 5.366563 \\ 0 & 0.912871 & 4.800000 & 1.877942 \\ 0 & 0 & 0.244949 & 0.200000 \end{pmatrix}$$

Escogemos, por ejemplo, $k_1 = h_{1,1}^{(1)} = 1$ y hacemos $H_1 - k_1 I = Q_1 R_1$; entonces, la nueva matriz es

$$H_2 = R_1 Q_1 + k_1 I = \begin{pmatrix} -1.000000 & 6.219210 & 6.815162 & -0.612372 \\ -1.527525 & 4.928571 & 0.845657 & -5.746117 \\ 0 & -0.174963 & 0.071429 & -0.327327 \\ 0 & 0 & 0.000000 & 1.000000 \end{pmatrix}$$

donde el cero de $h_{4,3}^{(2)}$ es de orden 10^{-16} . Aislamos un primer valor propio 1 y continuamos con la submatriz formada pero las tres primeras filas y columnas. Escogemos, también, $k_2 = h_{1,1}^{(2)} = -1$ y hacemos $H_2 - k_2 I = Q_2 R_2$; la matriz siguiente es

$$H_3 = R_2 Q_2 + k_2 I = \begin{pmatrix} 3.447368 & 8.634932 & -3.824392 \\ -0.410223 & -0.447368 & 3.325539 \\ 0 & 0.000000 & 1.000000 \end{pmatrix}$$

aquí el valor de $h_{3,2}^{(3)}$ es también de orden 10^{-16} . Así pues, obtenemos un nuevo valor propio igual a 1 y eliminamos la última fila y columna. Consideramos, finalmente, $k_3 = h_{1,1}^{(3)} \approx 3.45$ y factorizamos $H_3 - k_3 I = Q_3 R_3$, para obtener

$$H_4 = R_3 Q_3 + k_3 I = \begin{pmatrix} 2.000000 & -9.045154 \\ 0.000000 & 1.000000 \end{pmatrix}$$

donde el cero es del mismo orden que los anteriores, 10^{-16} , y obtenemos, por lo tanto, los dos últimos valores propios 1 y 2.

Doble desplazamiento conjugado

Para acelerar la convergencia del método para matrices reales y Hessenberg superior, se utiliza el que se llama **doble desplazamiento conjugado**. Éste consiste en realizar dos desplazamientos complejos (en general) respecto al origen, pero sin utilizar aritmética compleja.

Consideremos, por ejemplo, la submatriz del extremo inferior derecho de dimensión 2×2 , esta submatriz tiene dos valores propios k_1, k_2 que serán reales o complejos conjugados. Supongamos que estamos en la iteración 1 y estudiamos las transformaciones siguientes:

$$\begin{aligned} A_1 - k_1 I &= Q_1 R_1, & A_2 &= R_1 Q_1 + k_1 I \\ A_2 - k_2 I &= Q_2 R_2, & A_3 &= R_2 Q_2 + k_2 I \end{aligned}$$

Un pequeño cálculo nos demuestra que

$$\begin{aligned} A_3 &= Q^T A_1 Q \\ QR &= (A_1 - k_1 I)(A_1 - k_2 I) \end{aligned}$$

para a $Q = Q_1 Q_2$ y $R = R_2 R_1$.

Como k_1 y k_2 , o bien son reales o bien son complejos conjugados, el producto de matrices $(A_1 - k_1 I)(A_1 - k_2 I)$ es real y, por lo tanto, QR será la factorización QR de una matriz real. Entonces, esto nos dice que la matriz A_3 también debe ser real. Entonces solamente hace falta

ver cómo podemos calcular directamente esta matriz con aritmética real y sin pasar por los pasos intermedios. Para esto, se utiliza el resultado sobre matrices Hessenberg del teorema 7.3. Así pues, trataremos de calcular una matriz C Hessenberg, según las hipótesis del teorema, que sustituirá a A_3 .

La primera columna de Q , debido al hecho de que es la matriz ortogonal que triangulariza el producto $(A_1 - k_1 I)(A_1 - k_2 I)$ que es una matriz Hessenberg, será $q_1 = \frac{a_1}{\|a_1\|_2}$ donde

$$a_1^T = (x_1, y_1, z_1, 0, \dots, 0)$$

con

$$\begin{aligned} x_1 &= a_{11}^2 + a_{12}a_{21} - a_{11}(k_1 + k_2) + k_1k_2 \\ y_1 &= a_{21}(a_{11} + a_{22} - (k_1 + k_2)) \\ z_1 &= a_{32}a_{21} \end{aligned}$$

Se debe destacar que los valores de x_1 , y_1 y z_1 se pueden calcular con aritmética real, por ser k_1 y k_2 complejos conjugados. Además, $k_1 + k_2$ y k_1k_2 se obtienen fácilmente de los valores de la submatriz de dimensión 2×2 , de la cual son valores propios.

Seguidamente, consideraremos las transformaciones de Householder del tipo

$$P_r = I - 2w_r w_r^T \quad \text{con} \quad \|w_r\|_2 = 1,$$

donde w_r es un vector con las $r - 1$ primeras componentes nulas. Escogemos P_1 de manera que $P_1 q_1 = e_1$; esta matriz tiene la primera columna igual a q_1 . El producto

$$C_1 = P_1 A_1 P_1$$

es una matriz Hessenberg superior, excepto por dos elementos de la primera columna y uno de la segunda (c_{31} , c_{41} y c_{42}). Seguidamente, esta matriz se transforma en una matriz Hessenberg aplicando las correspondientes transformaciones de Householder reales, y se obtiene finalmente una nueva matriz

$$C = P_{n-2} \cdots P_2 (P_1 A_1 P_1) P_2 \cdots P_{n-2}$$

Finalmente, si definimos $\bar{Q} = P_1 \cdots P_{n-2}$, C y \bar{Q} son dos matrices que cumplen las condiciones enunciadas por el teorema 7.3.

Con este método se consigue una mayor convergencia de el elemento $a_{n-1,n-2}$ hacia cero. Se puede demostrar ([Wil65]) que, para pasar de A_1 a A_3 , el número de operaciones es de orden $5n^2$ (recordar que A_1 es una matriz Hessenberg).

Ejemplo. Considerar la matriz con dos valores propios conjugados del ejemplo anterior

$$A_0 = \begin{pmatrix} 1 & 1 & -1 \\ -1 & 2 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

Como ya es una matriz Hessenberg superior, podemos aplicar la doble traslación directamente. Si escogemos la submatriz 2×2 inferior derecha con valores propios k_1 y k_2 tenemos que $k_1 + k_2 = 3$ y $k_1 * k_2 = 2$ y por lo tanto, las componentes de la primera columna a_1 de $(A_0 - k_1 I)(A_0 - k_2 I)$ son

$$x_1 = -1 \quad y_1 = 0 \quad z_1 = -1$$

y la primera columna, q_1 , de la matriz Q que trianguliza el producto $(A_0 - k_1 I)(A_0 - k_2 I)$ es

$$q_1 = \frac{a_1}{\|a_1\|_2} = (-0.707107, 0, -0.707107)^T$$

Entonces, w_1 será

$$w_1 = \frac{e_1 - q_1}{\|e_1 - q_1\|_2} = (0.923880, 0, 0.382683)^T$$

Haciendo ahora $P_1 = I - 2w_1 w_1^T$, obtenemos las matrices

$$\begin{aligned} P_1 &= \begin{pmatrix} -0.707107 & 0 & -0.707107 \\ 0 & 1.000000 & 0 \\ -0.707107 & 0 & 0.707107 \end{pmatrix} \\ C_1 &= P_1 A_0 P_1 = \begin{pmatrix} 0.500000 & 1.414214 & 0.500000 \\ -0.707107 & 2.000000 & -0.707107 \\ -0.500000 & 0.000000 & 1.500000 \end{pmatrix} \end{aligned}$$

Finalmente, pasamos a forma Hessenberg la matriz C_1 y obtenemos

$$A_1 = \begin{pmatrix} 0.50000000 & -1.44337573 & -0.40824829 \\ 0.86602540 & 1.50000000 & 0.70710678 \\ 0 & 0.00000000 & 2.00000000 \end{pmatrix}$$

donde el cero del lugar $(3, 2)$ y el error del valor propio 2 son de orden 10^{-16} . Recordemos que anteriormente habíamos necesitado llegar hasta la iteración número treinta (!) para obtener un resultado bastante inferior. Si calculamos los valores propios de la submatriz superior izquierda restante, obtenemos $1 \pm i$ con un error también de orden 10^{-16} .

Implementación numérica del método QR

En la implementación numérica se deben considerar fundamentalmente los siguientes aspectos:

- Las matrices sobre las cuales aplicaremos un método QR serán matrices Hessenberg o tridiagonales simétricas. Recordemos que este método las conserva.
- La factorización QR de la matriz A_k no se realizará por el método de ortonormalización de Schmidt (método que se utiliza muchas veces en la demostración de la existencia de la factorización), ya que numéricamente se obtiene una matriz Q muy lejana de una matriz ortogonal. Hay que utilizar transformaciones ortogonales como, por ejemplo, transformaciones de Givens o de Householder, para obtener una estabilidad numérica del algoritmo.

Más concretamente, se calculará la matriz $Q_k^T = P_k \cdots P_1$ tal que $R_k = Q_k^T A_k$. Seguidamente, se obtendrá el iterado siguiente, haciendo

$$A_{k+1} = R_k P_1^T \cdots P_k^T = R_k Q_k$$

Se debe destacar también que, siendo A_k una matriz Hessenberg o tridiagonal, solamente hay que colocar a cero un solo elemento de cada columna y, por lo tanto, sirven perfectamente los giros del método de Givens. El paso de A_k a A_{k+1} se puede hacer con $4n^2$ operaciones ([Wil65]).

- Al trabajar con matrices Hessenberg, cuando un elemento de debajo de la diagonal está ya muy próximo a cero, es decir, cuando su valor absoluto está por debajo de una cierta tolerancia fijada inicialmente, éste se considera nulo y, entonces, podemos dividir la matriz en dos submatrices de dimensión más pequeña, y reducir considerablemente el coste computacional.
- Finalmente, utilizaremos métodos de aceleración de la convergencia como, por ejemplo, la traslación respecto del origen o la doble traslación conjugada que, además, nos ayudará a poder aplicar la consideración anterior.

Cálculo de valores singulares

Dada una matriz A de dimensión $m \times n$, la matriz $A^T A$ es simétrica y definida positiva; así pues, sus valores propios $\lambda_1, \dots, \lambda_n$ son reales y no negativos y, por lo tanto, los **valores singulares** de la matriz A se definen como

$$\sigma_i = +\sqrt{\lambda_i} \quad i = 1 \div n$$

Inicialmente, podríamos plantear su cálculo como un problema de valores propios de la matriz $A^T A$; pero, si se hace así, pueden aparecer graves problemas de precisión, tal como se puede observar en el ejemplo siguiente:

Ejemplo. Considerar la matriz

$$A = \begin{pmatrix} 1 & 1 \\ \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$$

La matriz $A^T A$ es

$$\begin{pmatrix} 1 + \varepsilon^2 & 1 \\ 1 & 1 + \varepsilon^2 \end{pmatrix}$$

sus valores propios son $\lambda_1 = 2 + \varepsilon^2$ y $\lambda_2 = \varepsilon^2$ y los valores singulares de A son

$$\sigma_1 = +\sqrt{2 + \varepsilon^2} \quad \sigma_2 = |\varepsilon|$$

Si el valor de $|\varepsilon|$ es suficientemente pequeño para que ε^2 sea inferior al épsilon de la máquina, resulta que, numéricamente, la matriz $A^T A$ sería

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

con valores propios 2 y 0; de donde tendríamos que los valores singulares calculados serían

$$\sigma_1 = +\sqrt{2} \quad \sigma_2 = 0$$

El algoritmo de Golub y Reinsch (1970) para el cálculo de valores singulares no calcula en ningún momento la matriz $A^T A$, sino que trabaja directamente sobre la matriz A . Básicamente, consiste en dos grandes pasos: primero se transforma la matriz inicial en una de más sencilla y después se aplica una variante del método QR para obtener una sucesión de matrices convergentes a una matriz diagonal que contiene los valores singulares. Veámoslo con más detalle; para esto supondremos que $m \geq n$.

Transformación de la matriz inicial:

Multiplicando a la derecha y a la izquierda por un número finito de matrices de Householder, colocamos los ceros necesarios a las filas y columnas correspondientes, para obtener una matriz de la forma

$$\bar{A} = PAQ = \begin{pmatrix} A_0 \\ 0 \end{pmatrix}$$

donde A_0 es una matriz cuadrada bidiagonal (diagonal principal más una sobrediagonal), P y Q son las matrices ortogonales resultantes del producto de las matrices de Householder utilizadas (notar que no son iguales).

Esta nueva matriz conserva los valores singulares, ya que

$$\bar{A}^T \bar{A} = Q^T A^T P^T P A Q = Q^T A Q$$

además, si $A_0 = GDH^T$ es la descomposición en valores singulares de la matriz A_0 , entonces

$$A = P^T \begin{pmatrix} G & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} D \\ 0 \end{pmatrix} H^T Q$$

lo es de A ; luego consideraremos solamente la matriz cuadrada A_0 .

Aplicación del método QR:

Siendo A_0 bidiagonal, $M_0 = A_0^T A_0$ es tridiagonal. Si aplicamos el método QR sobre esta última matriz, tenemos

$$\begin{aligned} M_0 &= Q_0 R_0, \\ M_1 &= R_0 Q_0 = Q_0^T M_0 Q_0 = Q_0^T A_0^T S_0^T S_0 A_0 Q_0 = A_1^T A_1, \end{aligned}$$

donde hemos definido $A_1 = S_0 A_0 Q_0$ para cierta matriz ortogonal S_0 . Se puede escoger S_0 de manera que la nueva matriz A_1 sea también bidiagonal; de hecho, el cálculo efectivo de Q_0 y S_0 lo comentamos a continuación.

Consideremos las matrices

$$S_{ij} = Q_{ij} = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & \cos \phi & & & & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \\ & & & & & & & \ddots \\ 0 & & & & & & & & 1 \end{pmatrix}$$

llamadas reflexiones de Givens. Los subíndices i, j denotan las filas y columnas donde están colocados los senos y cosenos; si multiplicamos a la derecha de A , lo denotaremos por Q_{ij} y, si es a la izquierda, por S_{ij} .

Comencemos multiplicando A por una cierta Q_{12} ; esto hace aparecer un elemento subdiagonal no nulo (\oplus)

$$A_0 Q_{12} = \begin{pmatrix} * & * & & \\ \oplus & * & * & \\ & * & * & \\ & & * & * \\ & & & * \end{pmatrix}$$

Este nuevo elemento subdiagonal lo hacemos cero multiplicando por una S_{12} adecuada, pero esto hace aparecer un nuevo elemento no nulo, esta vez en la segunda sobrediagonal

$$S_{12} A_0 Q_{12} = \begin{pmatrix} * & * & \oplus & & \\ 0 & * & * & & \\ & * & * & & \\ & & * & * & \\ & & & * & \end{pmatrix}$$

Seguidamente, multiplicamos por la Q_{23} adecuada para poner a cero el nuevo elemento no nulo, y así sucesivamente hasta al final, donde obtenemos una nueva matriz bidiagonal

$$A_1 = S_{n-1,n} S_{n-2,n-1} \cdots S_{12} A_0 Q_{12} \cdots Q_{n-2,n-1} Q_{n-1,n} = S_0 A_0 Q_0$$

Queda por ver aún cuál es exactamente la matriz Q_{12} . Estudiemos cuál es la primera columna q_1 de Q_0 como matriz de la descomposición QR de M_0 . Es fácil ver que

$$q_1 = \frac{m_1}{\|m_1\|_2}$$

donde m_1 denota la primera columna de M_0 . Además, si denotamos los elementos de A_0 de la forma siguiente:

$$A_0 = \begin{pmatrix} a_1 & b_2 \\ & a_2 & b_3 \\ & & \ddots & \ddots \\ & & & a_{n-1} & b_n \\ & & & & a_n \end{pmatrix}$$

tenemos que

$$m_1 = \begin{pmatrix} a_1^2 \\ a_1 b_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Por otro lado, q_1 como primera columna del producto de matrices Q_{ij} es

$$q_1 = \begin{pmatrix} c \\ s \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

donde s y c son el senos y cosenos, respectivamente, de la matriz Q_{12} . Entonces, Q_{12} se escoge de manera que

$$\begin{aligned} c &= \alpha a_1^2 \\ s &= \alpha a_1 b_2 \end{aligned}$$

con α tal que $c^2 + s^2 = 1$. Si lo que estamos haciendo es el método QR con una traslación k respecto al origen, tendremos

$$\begin{aligned} c &= \alpha a_1^2 - k \\ s &= \alpha a_1 b_2 \end{aligned}$$

Finalmente, si aplicamos el teorema 7.3 sobre matrices Hessenberg, tenemos que la matriz A_1 así obtenida será equivalente a la de aplicar un paso del método QR directamente sobre M_0 .

La sucesión de matrices generada de esta forma, convergerá a una matriz diagonal que contendrá los valores singulares de la matriz inicial A_0 , ya que la correspondiente sucesión $M_i = A_i^T A_i$ ha de converger a una matriz diagonal que contiene los valores propios de M_0 .

Ejemplo. Considerar la matriz

$$A = \begin{pmatrix} 1 & 6 & 11 \\ 2 & 7 & 12 \\ 3 & 8 & 13 \\ 4 & 9 & 14 \\ 5 & 10 & 15 \end{pmatrix}$$

El primer paso consiste en obtener una nueva matriz \bar{A} bidiagonal utilizando matrices de Householder

$$\bar{A} = \begin{pmatrix} -7.41620 & 32.73169 & 0 \\ 0 & 10.60517 & 1.08016 \\ 0 & 0 & 0.00000 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

por lo tanto, la matriz cuadrada sobre la cual aplicaremos el método QR será

$$A_0 = \begin{pmatrix} -7.41620 & 32.73169 & 0 \\ 0 & 10.60517 & 1.08016 \\ 0 & 0 & 0.00000 \end{pmatrix}$$

La primera reflexión de Givens $Q_{1,2}$ nos dará

$$A_0 Q_{1,2} = \begin{pmatrix} 33.56105 & 0.13897 & 0 \\ 10.35262 & -2.30062 & 1.08016 \\ 0 & 0 & 0.00000 \end{pmatrix}$$

donde observamos la aparición de un nuevo elemento no nulo, el de la posición (2, 1). Si multiplicamos por la $S_{1,2}$ adecuada, obtenemos

$$S_{1,2} A_0 Q_{1,2} = \begin{pmatrix} 35.12152 & -0.54535 & 0.31839 \\ 0 & -2.23937 & 1.03216 \\ 0 & 0 & 0.00000 \end{pmatrix}$$

Esta vez el elemento no nulo nuevo es el (1, 3), una vez eliminado

$$S_{1,2} A_0 Q_{1,2} Q_{2,3} = \begin{pmatrix} 35.12152 & 0.63149 & 0 \\ 0 & 2.45431 & 0.23771 \\ 0 & 0.00000 & 0.00000 \end{pmatrix}$$

Si los dos últimos elementos de la última fila no fuesen cero, multiplicaríamos por $S_{2,3}$ para poner a cero el término (3, 2). Así pues, tenemos la nueva matriz bidiagonal

$$A_1 = \begin{pmatrix} 35.12152 & 0.63149 & 0 \\ 0 & 2.45431 & 0.23771 \\ 0 & 0 & 0.00000 \end{pmatrix}$$

Si repetimos este proceso tres veces más, obtenemos la matriz

$$A_4 = \begin{pmatrix} 35.12722 & 0.00000 & 0 \\ 0 & 2.46540 & 0.00000 \\ 0 & 0 & 0.00000 \end{pmatrix}$$

de donde se afirma que los valores singulares de la matriz A son $\sigma_1 = 35.12722$, $\sigma_2 = 2.46540$ y $\sigma_3 = 0.00000$.

7.8 Problemas

1. Hallar una matriz 3×3 con valores propios $6, 2$ y -1 y vectores propios $(2, 3, -2)^T, (9, 5, 4)^T$ y $(4, 4, -1)^T$ respectivamente.
2. Hallar el valor propio dominante de la matriz

$$A = \begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}$$

con tres cifras exactas.

3. Calcular el valor propio de módulo mínimo y su vector propio asociado de la matriz

$$A = \begin{pmatrix} 1 & 2 & -2 & 4 \\ 2 & 12 & 3 & 5 \\ 3 & 13 & 0 & 7 \\ 2 & 11 & 2 & 2 \end{pmatrix}$$

con un error inferior a 10^{-6} , a partir del hecho de que, si λ ($\neq 0$) es valor propio de A (A inversible), λ^{-1} lo es de A^{-1} .

4. Usando el hecho de que si λ es valor propio de A , entonces $\lambda - p$ es valor propio de $A - pI$, hallar los valores propios máximo y mínimo de la matriz

$$A = \begin{pmatrix} 9 & 10 & 8 \\ 10 & 5 & -1 \\ 8 & -1 & 3 \end{pmatrix}$$

con dos cifras decimales de precisión. Utilizar $p = 12$ y los vectores iniciales $(1, 1, 1)^T$ y $(-1, 1, 1)^T$.

5. La matriz

$$A = \begin{pmatrix} 14 & 7 & 6 & 9 \\ 7 & 9 & 4 & 6 \\ 6 & 4 & 9 & 7 \\ 9 & 6 & 7 & 15 \end{pmatrix}$$

tiene un valor propio próximo a 4. Hallar este valor propio con un error inferior a 10^{-6} usando la matriz $(A - 4I)^{-1}$.

6. Considerar la matriz siguiente:

$$\begin{pmatrix} 12 & -8 & 8 \\ 3 & 1 & -3 \\ 11 & -11 & 9 \end{pmatrix}$$

Calcular sus los valores propios.

7. Sea

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix}$$

Aplicar el método de Jacobi y la iteración inversa de Wielandt.

8. Dada la matriz

$$A = \begin{pmatrix} 5 & 0 & 0 & -4 \\ 0 & -2 & 10 & 0 \\ 0 & -20 & 28 & 0 \\ 2 & 0 & 0 & -1 \end{pmatrix}$$

- (a) Calcular el valor propio dominante y un vector propio asociado por el método de la potencia.
 (b) Hacer una estimación de los valores y vectores propios restantes.

9. Aplicar el método QR y alguna de sus variantes a la matriz

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 4 \end{pmatrix}$$

10. Triangularizar por el método de Householder las matrices siguientes:

$$A = \begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix} \quad \text{y} \quad B = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

11. Aplicar el método de Givens a la matriz

$$\begin{pmatrix} 2 & 3 & -4 \\ 3 & 6 & 2 \\ -4 & 2 & 10 \end{pmatrix}$$

para obtener una de tridiagonal. Aplicar un paso del método de Jacobi y obtener, también, una matriz tridiagonal. Comparar las dos matrices obtenidas. ¿Por qué son diferentes?

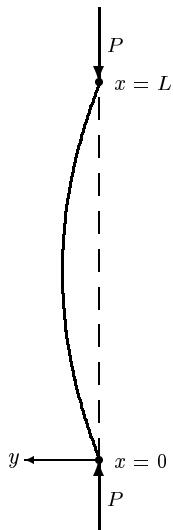
12. Sea H una matriz Hessenberg de dimensión n . Demostrar que los elementos diagonales de R cumplen

$$|r_{ii}| \geq |h_{i+1,i}| \quad i = 1 \div n-1$$

donde $H = QR$ (factorización QR de la matriz H). Como consecuencia de este hecho, demostrar que, si $\bar{\lambda}$ es una aproximación de un valor propio de H y $(H - \bar{\lambda}I) = QR$, entonces solamente el elemento r_{nn} puede ser arbitrariamente pequeño en valor absoluto.

7.9 Prácticas

7.9.1 Práctica ejemplo



La deformación de una columna de longitud L , articulada en sus extremos y sometida a una carga axial P , se puede describir mediante la ecuación

$$\frac{d^2y}{dx^2} = -\frac{Py}{EI}$$

donde E es el módulo de Young, I es el momento de inercia de la sección de la columna, x es la altura de un punto de la columna y $y(x)$ es el desplazamiento en el punto x (ver la figura 2).

Demostrar que la ecuación diferencial solamente tiene soluciones no triviales para ciertos valores característicos de la variable P :

$$P = \frac{n^2 \pi^2 EI}{L^2}, \quad n = 1, 2, \dots$$

y, entonces la solución es $A \operatorname{sen}(n\pi x/L)$, donde A es una constante. Observar que, para $n = 1$, tenemos $P = \pi^2 EI / L^2$ (carga de Euler) y es el valor por el cual la columna empieza a torcerse.

Aproximar la ecuación diferencial por diferencias finitas dividiendo L en $n+1$ partes iguales de longitud h y demostrar que, entonces, podemos escribir la ecuación diferencial como

$$Ay = \frac{Ph^2}{EI}y,$$

que es un problema de valores y vectores propios. Calcular los primeros valores característicos de P como resultado de este problema de valores propios.

La solución general de la ecuación diferencial es

$$y(x) = c_1 \cos(\alpha x) + c_2 \operatorname{sen}(\alpha x), \quad \text{con } \alpha = \left(\frac{P}{EI}\right)^{1/2}.$$

Como las condiciones frontera del problema son $y(0) = y(L) = 0$, es claro que, para que exista la solución no trivial, es necesario que

$$P = \frac{m^2 \pi^2 EI}{L^2} \quad m = 1, 2, \dots$$

y la solución es $c_2 \operatorname{sen}(m\pi x/L)$.

Si dividimos la columna en $n+1$ puntos a distancia h y aproximamos la segunda derivada por

$$y''(x) \approx \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}$$

podemos escribir la ecuación

$$Ay = \frac{Ph^2}{EI}y$$

donde

$$A = \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix}$$

Es claro, pues, que si λ es un valor propio de la matriz A

$$\lambda = \frac{Ph^2}{EI}$$

y esto nos dará unos valores concretos de P . Seguidamente, calcularemos los valores propios de la matriz para distintos valores de n , y hallaremos unos valores de P . Supongamos que la longitud de la columna es de 48 m , su módulo de Young vale $3 \times 10^6\text{ N/m}^2$ y $I = 1 \times 10^{-3}\text{ m}^4$ y calculamos P para $n = 10, 50$ y 100 . A continuación, presentamos el programa realizado para los cálculos donde se utilizan dos rutinas, ELMHES y HQR, de ([55])

```

PROGRAM COLUMNA
C
PARAMETER NP=100
DIMENSION A(NP,NP),WR(NP),WI(NP)
REAL L,II,PI,E,PI2,DX,PP
INTEGER y,N
C
WRITE(5,*) 'N = '
READ(6,*) N
L = 48.0
E = 3.0E+6
II = 1.0E-3
DX = L/N
C
C CONSTRUCCION DE LA MATRIZ
A(1,1) = 2.0
A(1,2) = -1.0
DO 20 I=2,N-1
    A(I,I) = 2.0
    A(I,I+1) = -1.0
    A(I,I-1) = -1.0
20   CONTINUE
A(N,N) = 2.0
A(N,N-1) = -1.0
C
C CALCULO DE LOS VALORES PROPIOS
CALL ELMHES(A,N,NP)
CALL HQR(A,N,NP,WR,WI)
C
DO 30 I=1,N
    WR(I) = E*II*WR(I)/DX/DX
    WRITE (2,100) WR(I)
30   CONTINUE
STOP
C

```

```

100  FORMAT(F30.3)
END

C      SUBRUTINA ELMHES
C      SE TRANSFORMA LA MATRIZ A HESSENBERG
C
C      SUBROUTINE ELMHES(A,N,NP)
C
C      DIMENSION A(NP,NP)
IF(N.GT.2)THEN
DO 17 M=2,N-1
X=0.
DO 11 J=M,N
IF(ABS(A(J,M-1)).GT.ABS(X))THEN
X=A(J,M-1)
I=J
ENDIF
CONTINUE
IF(I.NE.M)THEN
DO 12 J=M-1,N
Y=A(I,J)
A(I,J)=A(M,J)
A(M,J)=Y
CONTINUE
DO 13 J=1,N
Y=A(J,I)
A(J,I)=A(J,M)
A(J,M)=Y
CONTINUE
ENDIF
IF(X.NE.0.)THEN
DO 16 I=M+1,N
Y=A(I,M-1)
IF(Y.NE.0.)THEN
Y=Y/X
A(I,M-1)=Y
DO 14 J=M,N
A(I,J)=A(I,J)-Y*A(M,J)
CONTINUE
DO 15 J=1,N
A(J,M)=A(J,M)+Y*A(J,I)
CONTINUE
ENDIF
CONTINUE
ENDIF
CONTINUE
ENDIF
RETURN
END

C      SUBRUTINA HQR
C      METODO QR SOBRE UNA MATRIZ HESSENBERG
C
C      SUBROUTINE HQR(A,N,NP,WR,WI)
C
C      DIMENSION A(NP,NP),WR(NP),WI(NP)
ANORM=ABS(A(1,1))

```

```

DO 12 I=2,N
   DO 11 J=I-1,N
      ANORM=ANORM+ABS(A(I,J))
11   CONTINUE
12   CONTINUE
NN=N
T=0.
1   IF(NN.GE.1)THEN
   ITS=0
2   DO 13 L=NN,2,-1
      S=ABS(A(L-1,L-1))+ABS(A(L,L))
      IF(S.EQ.0.)S=ANORM
      IF(ABS(A(L,L-1))+S.EQ.S)GO TO 3
13   CONTINUE
L=1
3   X=A(NN,NN)
   IF(L.EQ.NN)THEN
      WR(NN)=X+T
      WI(NN)=0.
      NN=NN-1
   ELSE
      Y=A(NN-1,NN-1)
      W=A(NN,NN-1)*A(NN-1,NN)
      IF(L.EQ.NN-1)THEN
         P=0.5*(Y-X)
         Q=P**2+W
         Z=SQRT(ABS(Q))
         X=X+T
         IF(Q.GE.0.)THEN
            Z=P+SIGN(Z,P)
            WR(NN)=X+Z
            WR(NN-1)=WR(NN)
            IF(Z.NE.0.)WR(NN)=X-W/Z
            WI(NN)=0.
            WI(NN-1)=0.
         ELSE
            WR(NN)=X+P
            WR(NN-1)=WR(NN)
            WI(NN)=Z
            WI(NN-1)=-Z
         ENDIF
         NN=NN-2
      ELSE
         IF(ITS.EQ.30)PAUSE 'EXCESIVAS ITERACIONES'
         IF(ITS.EQ.10.OR.ITS.EQ.20)THEN
            T=T+X
            DO 14 I=1,NN
               A(I,I)=A(I,I)-X
14         CONTINUE
            S=ABS(A(NN,NN-1))+ABS(A(NN-1,NN-2))
            X=0.75*S
            Y=X
            W=-0.4375*S**2
         ENDIF
         ITS=ITS+1
         DO 15 M=NN-2,L,-1
            Z=A(M,M)
            R=X-Z
            S=Y-Z

```

```

P=(R*S-W)/A(M+1,M)+A(M,M+1)
Q=A(M+1,M+1)-Z-R-S
R=A(M+2,M+1)
S=ABS(P)+ABS(Q)+ABS(R)
P=P/S
Q=Q/S
R=R/S
IF(M.EQ.L)GO TO 4
U=ABS(A(M,M-1))*(ABS(Q)+ABS(R))
V=ABS(P)*(ABS(A(M-1,M-1))+ABS(Z)+ABS(A(M+1,M+1)))
IF(U+V.EQ.V)GO TO 4
15    CONTINUE
4      DO 16 I=M+2,NN
          A(I,I-2)=0.
          IF (I.NE.M+2) A(I,I-3)=0.
16    CONTINUE
DO 19 K=M,NN-1
  IF(K.NE.M)THEN
    P=A(K,K-1)
    Q=A(K+1,K-1)
    R=0.
    IF(K.NE.NN-1)R=A(K+2,K-1)
    X=ABS(P)+ABS(Q)+ABS(R)
    IF(X.NE.0.)THEN
      P=P/X
      Q=Q/X
      R=R/X
    ENDIF
  ENDIF
  S=SIGN(SQRT(P**2+Q**2+R**2),P)
  IF(S.NE.0.)THEN
    IF(K.EQ.M)THEN
      IF(L.NE.M)A(K,K-1)=-A(K,K-1)
    ELSE
      A(K,K-1)=-S*X
    ENDIF
    P=P+S
    X=P/S
    Y=Q/S
    Z=R/S
    Q=Q/P
    R=R/P
    DO 17 J=K,NN
      P=A(K,J)+Q*A(K+1,J)
      IF(K.NE.NN-1)THEN
        P=P+R*A(K+2,J)
        A(K+2,J)=A(K+2,J)-P*Z
      ENDIF
      A(K+1,J)=A(K+1,J)-P*Y
      A(K,J)=A(K,J)-P*X
    CONTINUE
17    DO 18 I=L,MIN(NN,K+3)
      P=X*A(I,K)+Y*A(I,K+1)
      IF(K.NE.NN-1)THEN
        P=P+Z*A(I,K+2)
        A(I,K+2)=A(I,K+2)-P*R
      ENDIF
      A(I,K+1)=A(I,K+1)-P*Q
      A(I,K)=A(I,K)-P
18    CONTINUE

```

```

18      CONTINUE
      END IF
19      CONTINUE
      GO TO 2
      ENDIF
      ENDIF
      GO TO 1
      ENDIF
      RETURN
END

```

Los resultados obtenidos para los cinco primeros valores propios para distintas discretizaciones y los valores exactos de P (que se denota por $n = \infty$) se pueden ver en la tabla siguiente:

n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
10	10.549	41.340	89.880	152.236	223.356
50	12.351	49.349	110.860	196.641	306.369
100	12.614	50.390	113.316	201.324	314.324
∞	12.851	51.404	115.660	205.617	321.276

7.9.2 Enunciados

1. Generar 11 puntos de la forma

$$t_i = \frac{i-1}{10}, \quad y_i = \text{erf}(t_i), \quad \text{para } i = 1 \div 11,$$

utilizando, por ejemplo, la rutina S15AEF de la librería NAG.

- (a) Aproximar estos puntos por polinomios de grado 1 hasta 10 por el criterio de mínimos cuadrados. Estudiar el error que se produce evaluando los polinomios en puntos intermedios y comparando el resultado con su valor exacto.
- (b) Como $\text{erf}(t)$ es una función impar, parece razonable aproximar los mismos datos por una combinación lineal de potencias impares de t

$$\text{erf}(t) \approx c_1 t + c_2 t^3 + \cdots + c_n t^{2n-1}.$$

Realizar el mismo estudio del error del apartado anterior.

- (c) Los polinomios no son una buena base para aproximar la función de error, ya que no están acotados para valores grandes de t y, en cambio, $\text{erf}(t)$ tiende a 1 para valores grandes de t . Utilizando el mismo conjunto de puntos, aproximar de la forma siguiente:

$$\text{erf}(t) \approx c_1 + e^{-t^2} (c_2 + c_3 z + c_4 z^2 + c_5 z^3)$$

donde

$$z = \frac{1}{1+t}$$

Estudiar como se comporta el error para esta nueva aproximación.

2. Generar veinte puntos (x_i, y_i) tales que $y_i = J_0(x_i)$, donde

$$J_0(x) = \frac{1}{\pi} \int_0^{\pi} \cos(x \sin t) dt$$

es la función de Bessel de orden cero y las x_i son del intervalo $[-3, 3]$ y equiespaciados.

- (a) Calcular la aproximación por mínimos cuadrados de este conjunto de puntos utilizando polinomios de grado desde 1 hasta 10. Hacer una estimación del error que se produce en los puntos intermedios comparando el resultado de evaluar los polinomios aproximadores con los resultados exactos que se pueden obtener, por ejemplo, utilizando la librería NAG.
- (b) Como la función de Bessel J_0 es par, calcular la aproximación por polinomios solamente con términos pares, es decir, de la forma

$$J_0(x) \approx c_0 + c_1 x^2 + c_2 x^4 + \cdots + c_n x^{2n}$$

Hacer, también, una estimación del error producido.

- (c) Finalmente, calcular la aproximación del conjunto de puntos bajo el criterio de mínimos cuadrados y por un polinomio del tipo

$$J_0(x) \approx c_0 + c_1(x/3)^2 + c_2(x/3)^4 + \cdots + c_n(x/3)^{2n}$$

Comparar este resultado con el anterior.

3. Considerar la tabla siguiente de la población catalana desde el año 1900 hasta el año 1986:

Año	Habitantes	Año	Habitantes	Año	Habitantes
1900	1984115	1940	2915757	1975	5660393
1910	2099218	1950	3218596	1981	5956414
1920	2355908	1960	3888485	1986	5978638
1930	2731627	1970	5107606		

- (a) Aproximar este conjunto de datos por polinomios de distintos grados y bajo el criterio de mínimos cuadrados. Hacer una predicción de la población para el año 2000. ¿Cómo afecta a la predicción el hecho de utilizar distintas bases de polinomios?
- (b) Aproximar el censo de la población por

$$y(t) \approx \bar{y}(t) = c_1 + c_2(t - 1900) + c_3 e^{\lambda(t - 1900)}$$

Probar para distintos valores de λ . Hacer, también, una predicción de la población para el año 2000. Intentar hallar un valor de λ que haga mínimo el error cuadrático.

4. El objetivo de esta práctica consiste en separar las vocales de las consonantes de un mensaje codificado.

Suponer que de un mensaje eliminamos los espacios en blanco y los signos de puntuación y seguidamente lo reescribimos sustituyendo cada letra por la anterior según la ordenación alfabética; es decir, la frase

EN UN LUGAR DE LA MANCHA, DE CUYO NOMBRE NO QUIERO ACORDARME
quedaría

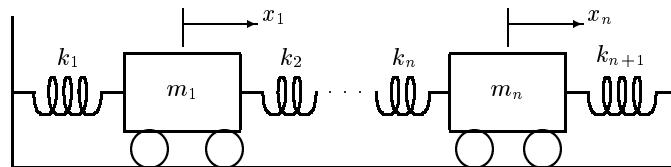
DMTMKTFZQCDKZLZMBGZCDBTXÑMÑLAQDMÑPHTDQÑZBÑQCZQLD

Escribir un programa que procese el texto y forme una matriz cuadrada de dimensión 26, de tal manera que el valor de el elemento (i, j) sea el número de veces que la pareja formada por la letra i y la letra j aparecen en el texto codificado. Esta matriz se llama **matriz de frecuencia de dígrafos**. A continuación, calcular la descomposición en valores singulares de la matriz y escoger las columnas de U y V correspondientes al valor singular de mayor módulo. Las componentes de estos vectores son todas del mismo signo y son proporcionales a las frecuencias de cada letra; por lo tanto, según la lengua habrá unas frecuencias más o menos típicas para cada letra.

Considerar a continuación las columnas de U y V correspondientes al segundo valor singular más grande en módulo. Todas sus componentes son del intervalo $[-1, 1]$ y, si dibujamos en el plano los puntos determinados por las paralelas (u_i, v_i) , obtenemos 26 puntos en el interior del cuadrado de lado 2 y centrado en el origen. La mayor parte de estos puntos están en el segundo y cuarto cuadrante; es más, la mayoría de consonantes están en un cuadrante y las vocales en el otro. Esto se debe a la mayor frecuencia de parejas del tipo vocal-consonante y consonante-vocal frente a las del tipo consonante-consonante o vocal-vocal. Naturalmente, hay excepciones, ya que, según la lengua, podemos tener, por ejemplo, una frecuencia elevada de algunas parejas de consonantes.

Coger un texto ligeramente largo, codificarlo y aplicar lo que se ha comentado anteriormente. Hacerlo también para distintos idiomas.

5. Considerar el sistema de masas y muelles de la figura siguiente. Los desplazamientos horizontales x_j se miden desde la posición de equilibrio estático del sistema. Las rigideces k_j de los muelles son las fuerzas necesarias para estirar o comprimir cada uno de ellos una unidad de longitud.



Las ecuaciones del movimiento son

$$\begin{aligned} m_1 \ddot{x}_1 &= -k_1 x_1 + k_2(x_2 - x_1) \\ &\vdots \quad \vdots \\ m_j \ddot{x}_j &= -k_j(x_j - x_{j-1}) + k_{j+1}(x_{j+1} - x_j) \quad j = 2 \div n-1 \\ &\vdots \quad \vdots \\ m_n \ddot{x}_n &= -k_n(x_n - x_{n-1}) - k_{n+1}x_n \end{aligned}$$

Demostrar que, haciendo la substitución $\mathbf{x} = \mathbf{b}e^{i\omega t}$, donde \mathbf{x} denota al vector de todos los desplazamientos, podemos escribir la ecuación

$$A\mathbf{b} = \lambda\mathbf{b}$$

donde $\lambda = -\omega^2$. Los valores del parámetro ω son las frecuencias naturales de vibración del sistema.

Escribir un programa que, dadas las masas m_j y las rigideces k_j , calcule las frecuencias naturales ω . Os sugerimos los datos siguientes: $k_j = 1 \text{ kg/seg}^2$ y $m_j = 1 \text{ kg}$. Estudiar, también, el caso de valores distintos de las masas y de las rigideces, especialmente el caso de dar valores muy grandes o muy pequeños a una de las masas o a uno de los muelles.

6. La deformación de una columna de longitud L , articulada en sus extremos y sometida a una carga axial P , se puede describir por la ecuación

$$\frac{d^2y}{dx^2} = -\frac{Py}{EI}$$

donde E es el módulo de Young, I es el momento de inercia de la sección de la columna, x es la altura de un punto de la columna y $y(x)$ es el desplazamiento en el punto x (ver la figura 2). Suponer que el momento de inercia varía linealmente entre I_0 en su extremo inferior ($x = 0$) i I_L en el extremo superior ($x = L$). Aproximando la ecuación diferencial por diferencias finitas, calcular la carga crítica por la cual la columna empieza a torcerse (ver la práctica resuelta). Os sugerimos los datos siguientes: $E = 7 \times 10^5 \text{ N/cm}^2$, $L = 120 \text{ cm}$, $I_0 = 0.05 \text{ cm}^4$ i $I_L = 0.001 \text{ cm}^4$

7. Una viga de longitud L está articulada por sus dos extremos. La deformación y es función del tiempo t y de su abscisa x , según la ecuación

$$\frac{\partial^2}{\partial x^2} \left(EI \frac{\partial^2 y}{\partial x^2} \right) = -A\rho \frac{\partial^2 y}{\partial t^2}$$

donde E es el módulo de Young, I el momento de inercia, A el área de la sección i ρ el peso específico. El cambio de variable

$$y(t, x) = X(x)e^{i\omega t}$$

donde ω es la frecuencia de vibración natural, nos conduce a la ecuación

$$\frac{d^2}{dx^2} \left(EI \frac{d^2 X}{dx^2} \right) = A\rho\omega^2 X$$

y en los extremos se cumple que $X = 0$ y $\frac{d^2 X}{dx^2} = 0$.

Dividir la viga en $n + 1$ partes iguales y aproximar la ecuación por diferencias finitas para obtener un problema de valores propios. Calcular la frecuencia ω para una viga de aleación ligera con $E = 7 \times 10^6 \text{ N/cm}^2$, $L = 25 \text{ cm}$, $\rho = 24.5 \times 10^{-3} \text{ N/cm}^3$, $A = 1.25z \text{ cm}^2$, $I = 1.25z^3/12 \text{ cm}^4$ y $z = z_0 + (z_L - z_0)(x/L)$, donde consideramos dos casos: (a) $t_0 = t_L = 1 \text{ cm}$ y (b) $t_0 = 1 \text{ cm}$, $t_L = 1.5 \text{ cm}$.

8. Los desplazamientos z de los puntos de una membrana bidimensional vibrante, tal que su perímetro está fijo, quedan definidos por la ecuación

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = \frac{m}{T} \frac{\partial^2 z}{\partial t^2}$$

donde T es la tensión de la membrana por unidad de longitud y m es la masa por unidad de superficie. Hacemos el cambio de variable $z = u(x, y)e^{i\omega t}$ y obtenemos

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{m\omega^2}{T} u$$

Si aproximamos esta ecuación diferencial por diferencias finitas, llegamos a un problema de valores propios que nos dará las distintas frecuencias ω de vibración.

Calcular las frecuencias naturales de vibración de una membrana de forma trapezoidal de altura 15 cm y bases 25 cm y 10 cm. Considerar que $\sqrt{T/m} = 5000 \text{ cm/seg.}$

8 Ecuaciones diferenciales ordinarias

8.1 Introducción

Una gran parte de las ecuaciones que nos describen el comportamiento de la naturaleza son ecuaciones diferenciales ordinarias, muchas veces nada sencillas; esto nos obligará a utilizar métodos numéricos para su estudio. Como ejemplos de distintos campos de la ciencia donde aparecen estos tipos de ecuaciones podemos citar la evolución de las especies dentro de un ecosistema, el movimiento de los cuerpos estelares, la variación de las concentraciones de los diferentes compuestos de una reacción química, la deformación de una viga, la variación del voltaje de un circuito electrónico, el comportamiento del amortiguador de un automóvil, el movimiento de un fluido, etc.

Por otro lado, en este caso se trata de resolver ecuaciones funcionales y, por lo tanto, los métodos serán un poco más complejos; además, utilizaremos otros métodos ya conocidos como, por ejemplo, el de Newton para hallar ceros de funciones no lineales, o bien los métodos clásicos de resolución de sistemas lineales.

8.2 Ecuaciones en diferencias

Antes de empezar a tratar los métodos de resolución numérica de ecuaciones diferenciales nos hacen falta unas cuantas herramientas y conceptos de ecuaciones en diferencias para poder trabajar con más facilidad.

8.2.1 Definiciones y conceptos básicos

Una **ecuación vectorial en diferencias** de primer orden es:

$$\mathbf{y}^{k+1} = G(\mathbf{y}^k, k) \quad \text{para } k = 0, 1, \dots$$

con \mathbf{y}^0 conocido y $G : \mathbf{R}^n \times \mathbf{R} \rightarrow \mathbf{R}^n$. Por lo tanto, una solución de la ecuación en diferencias será una sucesión de vectores $\{\mathbf{y}^{k+1} = G(\mathbf{y}^k, k)\}$ para \mathbf{y}^0 dado.

El concepto que básicamente utilizaremos es el de **estabilidad** de una solución de la ecuación en diferencias: diremos que una solución \mathbf{y}^k es estable si dado $\varepsilon > 0 \exists \delta > 0$ tal que si $\hat{\mathbf{y}}^k$ es otra solución de la misma ecuación en diferencias que cumple $\|\mathbf{y}^0 - \hat{\mathbf{y}}^0\| < \delta$, entonces $\|\mathbf{y}^k - \hat{\mathbf{y}}^k\| < \varepsilon$ para $k = 1, 2, \dots$.

Un segundo concepto más fuerte de estabilidad es el de **estabilidad asintótica**: diremos que una solución \mathbf{y}^k es asintóticamente estable si, además de ser estable, $\lim_{k \rightarrow \infty} \|\mathbf{y}^k - \hat{\mathbf{y}}^k\| = 0$.

Cabe observar la similitud entre los conceptos de estabilidad de las ecuaciones en diferencias y las ecuaciones diferenciales.

Seguidamente, veremos unos cuantos resultados sobre la estabilidad de distintos tipos de ecuaciones en diferencias.

8.2.2 Ecuaciones en diferencias lineales con coeficientes constantes

Como es natural, las ecuaciones en diferencias más sencillas son las lineales. Una **ecuación en diferencias lineal homogénea de primer orden** es:

$$\mathbf{y}^{k+1} = B\mathbf{y}^k \quad \text{con } B \in \mathcal{L}(\mathbf{R}^n) \quad \mathbf{y}^k \in \mathbf{R}^n.$$

La estabilidad de este tipo de ecuaciones viene dada por el teorema siguiente:

Teorema 8.1 Toda solución de la ecuación lineal $\mathbf{y}^{k+1} = B\mathbf{y}^k$ es:

1. Estable si, y solamente si, $\rho(B) \leq 1$, y si $\rho(B) = 1$; entonces, B es de clase M (una matriz es de clase M, cuando los bloques de Jordan correspondientes a los valores propios de módulo igual a $\rho(B)$ son de dimensión 1).
2. Asintóticamente estable si, y solamente si, $\rho(B) < 1$.

Demostración: 1. Sea $\{\hat{\mathbf{y}}^k\}$ otra solución de la misma ecuación, definimos los vectores siguientes: $\mathbf{w}^k = \hat{\mathbf{y}}^k - \mathbf{y}^k$ para $k = 0, 1, \dots$. Está claro que:

$$\mathbf{w}^k = B\mathbf{w}^{k-1} = B^2\mathbf{w}^{k-2} = \dots = B^k\mathbf{w}^0$$

Sabemos que $\|B^k\|$ está acotado si, y solamente si, $\rho(B) \leq 1$ y, en caso de $\rho(B) = 1$, entonces B es de clase M. Por lo tanto, sea $\sigma > 0$ una cota de $\|B^k\|$. Entonces, dado $\varepsilon > 0 \exists \delta > 0$ tal que si $\|\mathbf{w}^0\| < \delta$, tenemos que

$$\|\mathbf{w}^k\| \leq \|B^k\| \|\mathbf{w}^0\| < \sigma \delta = \varepsilon \quad \text{para } \delta = \frac{\varepsilon}{\sigma}$$

Por otro lado, si $\|B^k\|$ no está acotado, entonces $\exists \mathbf{x} \in \mathbf{R}^n$ tal que $\lim_{k \rightarrow \infty} \|B^k \mathbf{x}\| = \infty$ y, por lo tanto, es suficiente coger $\hat{\mathbf{y}}^0$ tal que $\mathbf{w}^0 = \alpha \mathbf{x}$ para ver que no es estable.

2. Para demostrar la estabilidad asintótica se razona de manera análoga, pero utilizando el resultado siguiente:

$$\lim_{k \rightarrow \infty} \|B^k\| = 0 \Leftrightarrow \rho(B) < 1$$

que fácilmente nos conduce a lo que queríamos demostrar. \square

Un segundo tipo de ecuaciones en diferencias lineales son las de orden n :

$$y_k - \alpha_{n-1}y_{k-1} - \cdots - \alpha_0y_{k-n} = 0 \quad \text{para } k = n, n+1, \dots \quad (8.1)$$

y con n condiciones iniciales: y_0, y_1, \dots, y_{n-1} . Una ecuación de este tipo se puede transformar en una ecuación vectorial de primer orden de la manera siguiente (observar, una vez más, el paralelismo con las ecuaciones diferenciales):

$$\mathbf{y}^{k+1} = B\mathbf{y}^k \quad \text{con} \quad \mathbf{y}^k = \begin{pmatrix} y_{k+n-1} \\ \vdots \\ y_k \end{pmatrix} \quad B = \begin{pmatrix} \alpha_{n-1} & \alpha_{n-2} & \cdots & \alpha_0 \\ 1 & 0 & \cdots & 0 \\ \ddots & \ddots & \ddots & \vdots \\ 0 & & 1 & 0 \end{pmatrix}$$

Un resultado inmediato del teorema anterior sobre la estabilidad de las ecuaciones en diferencias de orden n es el siguiente:

Teorema 8.2 Una ecuación en diferencias del tipo 8.1 tiene todas sus soluciones estables si, y solamente si, todas las raíces λ_i del polinomio

$$p(\lambda) = \lambda^n - \alpha_{n-1}\lambda^{n-1} - \cdots - \alpha_0$$

llamado **polinomio característico** de la ecuación en diferencias, cumplen que $|\lambda_i| \leq 1$ para $i = 1 \div n$ y, si $|\lambda_i| = 1$, entonces, es una raíz simple. Serán asintóticamente estables si, y solamente si, $|\lambda_i| < 1$ para $i = 1 \div n$.

Solución general de una ecuación en diferencias lineal de orden n

Consideraremos, en primer lugar, el caso más sencillo, es decir, cuando las raíces del polinomio característico $p(\lambda)$ de la ecuación en diferencias 8.1 son todas diferentes. En este caso tenemos que la matriz de Jordan J asociada a B es una matriz diagonal compuesta por las raíces λ_i ; entonces, si $B = P^{-1}JP$, podemos escribir que

$$\mathbf{y}^k = B^k \mathbf{y}^0 = P^{-1} J^k P \mathbf{y}^0$$

y, estudiando cómo es la matriz J^k y el correspondiente producto de matrices de la expresión final anterior, tenemos que

$$y_k = \sum_{i=1}^n c_i \lambda_i^k \quad \text{para } k = 0, 1, \dots$$

El resultado general nos da el teorema siguiente:

Teorema 8.3 La solución general de una ecuación en diferencias lineal de orden n es una cierta combinación lineal de

$$\lambda_1^k, k\lambda_1^{k-1}, \dots, \binom{k}{n_1-1}\lambda_1^{k-n_1+1}, \lambda_2^k, \dots, \lambda_m^k, k\lambda_m^{k-1}, \dots, \binom{k}{n_m-1}\lambda_m^{k-n_m+1},$$

suponiendo que las raíces diferentes del polinomio característico $p(\lambda)$ son $\lambda_1, \dots, \lambda_m$ con multiplicidades n_1, \dots, n_m respectivamente.

Demostración: De manera semejante al caso visto anteriormente, es necesario estudiar de qué forma queda la matriz J^k y, a continuación, se deduce la forma general de las soluciones. Para más detalles, ver ([Ort 72]) \square

Ejemplo. Considerar la ecuación en diferencias lineal de orden 4

$$y_{n+4} - 4y_{n+3} + 5y_{n+2} - 4y_{n+1} + 4y_n = 0$$

con condiciones iniciales $y_0 = 1, y_1 = 1, y_2 = -2, y_3 = 2$. Su polinomio característico es

$$p(\lambda) = \lambda^4 - 4\lambda^3 + 5\lambda^2 - 4\lambda + 4 = (\lambda^2 + 1)(\lambda - 2)^2$$

y las sus raíces son $\lambda_{1,2} = \pm i, \lambda_{3,4} = 2$; por lo tanto, las soluciones de la ecuación no son estables. Según el teorema anterior, su solución general es

$$y_n = c_1 i^n + c_2 (-i)^n + c_3 2^n + c_4 n 2^{n-1}$$

donde, por las condiciones iniciales dadas, obtenemos la solución particular

$$y_n = (1-i)i^n + (1+i)(-i)^n - 2^n + n2^{n-1}$$

Finalmente, si escribimos $i = e^{i\pi/2}$ nos queda

$$y_n = 2\left(\cos \frac{n\pi}{2} + \operatorname{sen} \frac{n\pi}{2}\right) + (n-2)2^{n-1}$$

Ejercicio. Considerar la sucesión de Fibonacci $\varphi_n = \varphi_{n-1} + \varphi_{n-2}$ con $\varphi_0 = 0$ y $\varphi_1 = 1$. Calcular la solución de la correspondiente ecuación en diferencias y demostrar que

$$\lim_{n \rightarrow \infty} \frac{\varphi_{n+1}}{\varphi_n} = \frac{1 + \sqrt{5}}{2}$$

Ecuaciones en diferencias lineales no homogéneas

Estas ecuaciones son las que tienen un término independiente, es decir:

$$\mathbf{y}^{k+1} = B\mathbf{y}^k + \mathbf{d}^k \quad \text{para } k = 0, 1, \dots \quad B \in \mathcal{L}(\mathbf{R}^n) \text{ y } \mathbf{d}^k \in \mathbf{R}^n. \quad (8.2)$$

Es claro que las soluciones son de la forma

$$\mathbf{y}^k = B^k \mathbf{y}^0 + \sum_{j=0}^{k-1} B^j \mathbf{d}^{k-j-1}$$

Ejercicios.

1. Demostrar que la solución general de una ecuación en diferencias lineal no homogénea es igual a la solución general de la ecuación homogénea más una solución particular de la no homogénea.
2. Demostrar que una solución particular de la ecuación

$$y_k - \alpha_{k-1}y_{k-1} - \cdots - \alpha_0y_{k-n} = d \quad \text{para } k = n, n+1, \dots$$

viene dada por la solución constante

$$y_k = \frac{d}{1 - \alpha_0 - \cdots - \alpha_{k-1}}$$

Ejemplo. Considerar, ahora, la misma ecuación del ejemplo anterior, pero con un término independiente

$$y_{n+4} - 4y_{n+3} + 5y_{n+2} - 4y_{n+1} + 4y_n = 4$$

y con condiciones iniciales $y_0 = 5, y_1 = 1, y_2 = 0, y_3 = 0$. Según el ejercicio anterior, una solución particular es $y_n = 2$ y, por lo tanto, la solución general de la ecuación no homogénea es

$$y_n = c_1 i^n + c_2 (-i)^n + c_3 2^n + c_4 n 2^{n-1} + 2$$

donde, por las condiciones iniciales dadas, obtenemos la solución

$$y_n = (1+i)i^n + (1-i)(-i)^n + 2^n - n2^{n-1} + 2$$

Finalmente, si escribimos $i = e^{i\pi/2}$ nos queda

$$y_n = 2(\cos \frac{n\pi}{2} - \operatorname{sen} \frac{n\pi}{2}) + (2-n)2^{n-1} + 2$$

El mismo resultado del teorema 8.1 se puede demostrar para estas ecuaciones, pero, además, tenemos algún otro resultado importante.

Teorema 8.4 Dada una ecuación lineal no homogénea del tipo 8.2, si la solución trivial idénticamente igual a cero de la correspondiente ecuación homogénea $\mathbf{y}^{k+1} = B\mathbf{y}^k$ es estable, entonces existe alguna norma de vectores tal que las soluciones de 8.2 cumplen:

$$\|\mathbf{y}^k\| \leq \|\mathbf{y}^0\| + \sum_{j=0}^{k-1} \|\mathbf{d}^{k-j-1}\| \quad \text{para } k = 1, \dots$$

Si la solución cero es asintóticamente estable, entonces existen una norma y un real $\alpha < 1$ tales que

$$\|\mathbf{y}^k\| \leq \alpha^k \|\mathbf{y}^0\| + \sum_{j=0}^{k-1} \alpha^j \|\mathbf{d}^{k-j-1}\| \quad \text{para } k = 1, \dots$$

Demostración: Para ser la solución cero de la ecuación homogénea estable, el teorema 8.1 dice que $\rho(B) \leq 1$ y, si $\rho(B) = 1$, entonces B es de clase M. Si es así, sabemos que $\exists \|\cdot\|$ tal que $\|B\| \leq 1$, por lo tanto, con esta norma tenemos que

$$\|\mathbf{y}^k\| \leq \|B\|^k \|\mathbf{y}^0\| + \sum_{j=0}^{k-1} \|B\|^j \|\mathbf{d}^{k-j-1}\| \leq \|\mathbf{y}^0\| + \sum_{j=0}^{k-1} \|\mathbf{d}^{k-j-1}\|$$

Por otro lado, si la solución cero de la ecuación homogénea es asintóticamente estable, entonces tenemos que $\rho(B) < 1$ y, por lo tanto, existen $\|\cdot\|$ y α tales que $\|B\| = \alpha < 1$ y, razonando de manera análoga al caso anterior, resulta

$$\|\mathbf{y}^k\| \leq \alpha^k \|\mathbf{y}^0\| + \sum_{j=0}^{k-1} \alpha^j \|\mathbf{d}^{k-j-1}\|$$

que es exactamente la expresión que queríamos obtener. \square

8.3 Problema de valores iniciales

El problema de valores iniciales de una ecuación diferencial ordinaria consiste en hallar la solución de la ecuación diferencial

$$\mathbf{y}' = f(x, \mathbf{y}) \quad \text{con } x \in [a, b] \text{ y } \mathbf{y}, \mathbf{y}' \in \mathbf{R}^n,$$

tal que $\mathbf{y}(a) = \mathbf{y}_0$ para \mathbf{y}_0 dado.

Como ya es sabido, plantear este problema tiene sentido gracias al teorema de existencia y unicidad de soluciones de las ecuaciones diferenciales ordinarias.

8.3.1 Familias de métodos

Estudiaremos los métodos llamados de variable discreta, que consisten en obtener una colección de aproximaciones $y_n \sim y(x_n)$ para $x_n = x_{n-1} + h_n$ con $h_n > 0$ y $x_0 = a$. El método numérico nos define cómo obtener la aproximación y_{n+1} en función de las anteriores aproximaciones. Básicamente, tenemos tres grandes familias de métodos:

- Métodos derivados de la serie de Taylor
- Métodos lineales multipaso
- Métodos Runge-Kutta

Métodos derivados de la serie de Taylor

Este tipo de métodos consisten en aproximar el valor de $y(x_{n+1})$ por el truncamiento, en un cierto término, de la serie de Taylor de $y(x)$ desarrollada alrededor de x_n . Es decir:

$$y_{n+1} = y_n + h_n \phi(x_n, y_n, h_n)$$

$$\phi(x, y, h) = \sum_{k=1}^N \frac{h^{k-1}}{k!} f^{(k-1)}(x, y)$$

Ejemplos.

1. Para $N = 1$ tenemos el método más sencillo, conocido como método de Euler adelante:
 $y_{n+1} = y_n + h_n f(x_n, y_n).$
2. Para $N = 2$:

$$y_{n+1} = y_n + h_n f(x_n, y_n) + \frac{h_n^2}{2} [f_x(x_n, y_n) + f_y(x_n, y_n) f(x_n, y_n)]$$

Como se puede observar en los ejemplos anteriores y también en la expresión general, estos métodos tienen el grave inconveniente de necesitar conocer y calcular las derivadas sucesivas de la función $f(x, y)$. Esto hace que la aplicación del método esté muy atada a la ecuación y que, además, el cálculo sea posiblemente bastante costoso, ya que, generalmente, estas derivadas tienen expresiones muy complejas. Por lo tanto, este tipo de métodos no son muy utilizados, pero es necesario conocer su existencia. Así pues, nos centraremos totalmente en el estudio de los otros dos tipos.

Métodos lineales multipaso

Consideramos la ecuación diferencial

$$y' = f(x, y) \quad \text{con} \quad x \in [a, b] \quad \text{y} \quad y, y' \in \mathbf{R},$$

que integramos

$$y(x + \xi) - y(x) = \int_x^{x+\xi} f(t, y(t)) dt \quad \text{para} \quad x, x + \xi \in [a, b].$$

Seguidamente, se aproxima f por un cierto polinomio interpolador $P(t)$ en unos puntos $(x_n, y_n \sim y(x_n))$ conocidos y, finalmente, se integra el polinomio en lugar de la función f . De esta manera se obtienen fórmulas del tipo:

$$\sum_{j=0}^k \alpha_j y_{n+j} = h \sum_{j=0}^k \beta_j f_{n+j} \quad \text{con} \quad f_n = f(x_n, y_n) \quad \text{y} \quad \alpha_k \neq 0. \quad (8.3)$$

Si $\beta_k = 0$ es un método lineal de k pasos **explícito**, y en caso contrario, **implícito**.

Esta idea fue propuesta por primera vez por Bashforth y Adams el año 1883 con el método que lleva su nombre. Otros métodos multipaso importantes fueron propuestos por Nyström (1925) y Milne (1926, 1953). La teoría moderna de estos métodos fue desarrollada por Dahlquist (1956) y divulgada por Henrici (1962, 1963).

Ejemplos.

1. El método de Euler adelante visto anteriormente es, también, un método de un paso explícito.
2. El método de Euler atrás $y_{n+1} = y_n + hf_{n+1}$ es un método de un paso implícito.
3. El método del trapecio $y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n)$ es un método de dos pasos implícito.

Ejercicios.

1. Hallar la expresión de un método integrando entre x_n y x_{n+1} el polinomio que interpola los puntos $(x_n, f_n), (x_{n+1}, f_{n+1})$.
2. Hallar la expresión de un método integrando entre x_n y x_{n+1} el polinomio que interpola los puntos $(x_n, f_n), (x_{n-1}, f_{n-1})$.

Métodos Runge-Kutta

Estos son unos métodos que consisten en construir fórmulas semejantes a las de los métodos derivados de la serie de Taylor, pero sin utilizar las derivadas sucesivas de $f(x, y)$:

$$\begin{aligned} y_{n+1} &= y_n + h_n \phi(x_n, y_n, h_n) \quad \text{donde} \quad \phi(x, y, h) = \sum_{r=1}^R c_r k_r, \\ \text{con } k_r &= f(x + ha_r, y + h \sum_{s=1}^M b_{rs} k_s) \quad \text{y} \quad a_r = \sum_{s=1}^M b_{rs}, \quad r = 1 \div R. \end{aligned}$$

Si $M = r - 1$, diremos que es un método Runge-Kutta **explícito** y, si $M = R$, entonces será **implícito**. Solamente estudiaremos en detalle los métodos explícitos que son de la forma:

$$y_{n+1} = y_n + h_n \sum_{r=1}^R c_r k_r$$

donde

$$\begin{aligned} k_1 &= f(x, y) \\ k_r &= f(x + ha_r, y + h \sum_{s=1}^{r-1} b_{rs} k_s) \quad \text{con} \quad a_r = \sum_{s=1}^{r-1} b_{rs}, \quad r = 2 \div R. \end{aligned}$$

La primera idea fue propuesta por Runge el año 1895; posteriores contribuciones de Heun (1900) y Kutta (1901) caracterizaron los métodos de orden cuatro y propusieron el primer método de orden cinco. Con la gran utilización de los ordenadores para el cálculo científico y técnico, este tipos de métodos han tenido mucha difusión. Actualmente, aún es uno de los grandes campos de investigación del análisis numérico.

Ejemplos.

1. El método de Heun, que es un método Runge-Kutta de orden 2:

$$y_{n+1} = y_n + \frac{h_n}{2}[f_n + f(x_{n+1}, y_n + h_n f_n)]$$

2. El método Runge-Kutta de orden 4:

$$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$\begin{aligned} k_1 &= f(x, y) & k_3 &= f\left(x + \frac{h_n}{2}, y + \frac{h_n}{2}k_2\right) \\ k_2 &= f\left(x + \frac{h_n}{2}, y + \frac{h_n}{2}k_1\right) & k_4 &= f(x + h_n, y + h_n k_3) \end{aligned}$$

8.3.2 Errores, convergencia, consistencia, orden y estabilidad

Cualquiera de los métodos vistos anteriormente se puede escribir como:

$$\sum_{i=0}^k \alpha_i y_{n+i} = h\phi(x_{n+k}, \dots, x_n; y_{n+k}, \dots, y_n; h). \quad (8.4)$$

Esta expresión es claramente una ecuación en diferencias, que estudiaremos como tal. Su **polinomio característico** es $p(\theta) = \sum_{i=0}^k \alpha_i \theta^i$.

Desde ahora, si no se dice lo contrario, consideraremos que $y(x)$ es la **solución exacta** del problema de valores iniciales y, por lo tanto, $y(x_n)$ es el valor de esta solución en el punto x_n ; y_n será la aproximación de este valor, es decir, la **aproximación de la solución** en el punto x_n , calculada para algún método numérico. Con esta notación podemos definir diferentes tipos de errores.

- **Error global de discretización:**

$$\max_{0 \leq n \leq N} \|y_n - y(x_n)\| \quad \text{con } x_0 = a \text{ y } x_N = b.$$

- **Error local:**

$$y_{n+k} - y(x_{n+k}; x_{n+k-1}, y_{n+k-1}) \quad \text{con } x_{n+k} \in [a, b],$$

donde $y(x; x_n, y_n)$ denota la solución de la ecuación diferencial $y' = f(x, y)$ con condiciones iniciales $y(x_n; x_n, y_n) = y_n$ y $x \in [x_n, b]$.

- **Error local de truncamiento:**

$$\tau(x, h) = h^{-1} \sum_{i=0}^k \alpha_i y(x + ih) - \phi(x + kh, \dots, x; y(x + kh), \dots, y(x); h)$$

con $x \in [a, b - kh]$.

Convergencia

El concepto de convergencia es el que cualquier método numérico pretende cumplir. Un método del tipo 8.4 es **convergente** si, aplicado a cualquier problema de condiciones iniciales, el error global tiende a cero para h tendiendo a cero. Es decir:

$$\lim_{h \rightarrow 0} \max_{0 \leq n \leq N} \|y_n - y(x_n)\| = 0$$

Naturalmente, este es un concepto difícil de demostrar directamente y, por lo tanto, necesitamos otros conceptos que nos ayuden.

Consistencia

El método 8.4 es consistente con la ecuación diferencial $y' = f(x, y)$ si

$$\lim_{h \rightarrow 0} \tau(h) = 0 \quad \text{con} \quad \tau(h) = \max_{x \in [a, b-kh]} \|\tau(x, h)\|.$$

Un resultado que simplifica enormemente saber si un método es consistente o no con una ecuación diferencial es el siguiente:

Teorema 8.5 Sea $y(x) \in C^1[a, b]$ la solución de la ecuación diferencial y ϕ continua; si

$$p(1) = 0 \quad \text{y} \quad \phi(x, \dots, x; y(x), \dots, y(x); 0) = p'(1)f(x, y(x)),$$

entonces el método es consistente con la ecuación. Si, además, $y \neq 0$, las dos condiciones anteriores son necesarias para la consistencia.

Demostración: Es necesario considerar las funciones accesorias siguientes:

$$\begin{aligned} g(x, h) &= \frac{y(x+h) - y(x)}{h} - y'(x) \\ G(x, h) &= ng(x, nh) - \sum_{i=1}^{k-1} i\alpha_i g(x, ih) \end{aligned}$$

que cumplen

$$\tau(x, h) = h^{-1}p(1)y(x) + p'(1)y'(x) - G(x, h) - \phi(x + kh, \dots, x; y(x+kh), \dots, y(x); h)$$

entonces es fácil ver que

$$\lim_{h \rightarrow 0} \tau(x, h) = y'(x)p'(1) - \phi(x, \dots, x; y(x), \dots, y(x); 0) = 0$$

y, como esta convergencia es uniforme, queda demostrada la consistencia.

Inversamente, la definición del error local de truncamiento $\tau(x, h)$ y de la consistencia demuestra que $p(1)y(x) = 0 \quad \forall x \in [a, b]$, y se obtienen, por lo tanto, las dos condiciones del teorema. \square

Ejemplo. Considerar el método $y_{n+2} - y_{n+1} = \frac{h}{3}(3f_{n+1} - 2f_n)$ para resolver el problema de valores iniciales $y' = 4xy^{1/2}$ y $y(0) = 1$ en el intervalo $[0, 2]$.

El polinomio característico es $p(\lambda) = \lambda^2 - \lambda$ y cumple la primera condición del teorema 8.5: $p(1) = 0$, pero no la segunda, ya que $\phi(x, x; y, y; 0) = \frac{4}{3}xy^{1/2}$ y, por otro lado, $p'(1)f(x, y) = 4xy^{1/2}$. Por lo tanto, este método no es consistente con la ecuación diferencial. En el apartado 8.3.5 están los resultados obtenidos numéricamente comparados con la solución exacta.

Ejercicios.

1. Cualquier método lineal multipaso de la forma 8.3 es consistente si

$$\sum_{i=0}^k \alpha_i = 0 \quad \text{y} \quad \sum_{i=0}^k \beta_i = k + (k-1)\alpha_{k-1} + \cdots + \alpha_1.$$

2. Considerar el método lineal de dos pasos siguiente, $y_{k+2} = 4y_{k+1} - 3y_k - 2hf_k$.

1. Demostrar que es consistente con la ecuación $y' = y$.
2. Hallar la solución general de la ecuación en diferencias resultante de aplicar el método a la ecuación anterior y con condiciones iniciales $y(0) = 1$.
3. Estudiar el error que se produce para distintos y_1 y $y_0 = 1$.

Estabilidad

El último concepto que necesitamos, de momento, es el de estabilidad de un método. Diremos que el método 8.4 es **estable** si, y solamente si, las soluciones de la ecuación en diferencias $\sum_{i=0}^k \alpha_i y_{n+i}$ son estables.

Por lo tanto, según el corolario del teorema 8.1, el método 8.4 es estable si, y solamente si, las raíces λ_i del polinomio característico $p(\lambda)$ cumplen que $|\lambda_i| \leq 1$ y las que $|\lambda_i| = 1$ son simples.

Teorema 8.6 Suponemos que un método cumple las propiedades siguientes:

- Es estable.
- $\exists L > 0$ tal que $\|\phi(\mathbf{t}; \mathbf{u}; h) - \phi(\mathbf{t}; \mathbf{v}; h)\| \leq L\|\mathbf{u} - \mathbf{v}\|$ con $h > 0$, $\mathbf{t} = (t_0, \dots, t_k)^T$, $\mathbf{u} = (u_0, \dots, u_k)^T$, $\mathbf{v} = (v_0, \dots, v_k)^T$, $t_i \in [a, b]$ y $u_i, v_i \in \mathbf{R}$.

Sea $y(x)$ la solución exacta y $y_i(h)$ la solución calculada. Entonces, $\exists c_1, c_2 \in \mathbf{R}$ independientes de h , tales que

$$\|y(a + nh) - y_n(h)\| \leq c_1 r(h) + c_2 \tau(h)$$

para $n = k, k+1, \dots, \frac{b-a}{h}$ con $r(h) = \max_{0 \leq n \leq k-1} \|y(a + nh) - y_n(h)\|$.

Demostración: [Ort72] \square

Está claro que este teorema nos demuestra la convergencia de un método si las funciones $r(h)$ y $\tau(h)$ convergen a 0 para $h \rightarrow 0$. Cabe observar que $\lim_{h \rightarrow 0} \tau(h) = 0$ es precisamente la definición de consistencia.

Orden

Un método del tipo 8.4 es al **menos de orden p** si, para cualquier ecuación diferencial $y' = f(x, y)$ con soluciones p veces diferenciables con continuidad, se cumple que $\tau(h) = O(h^p)$. El método es **exactamente de orden p** si es al menos de orden p y, además, $\tau(h) \neq O(h^{p+1})$.

Como consecuencia del teorema anterior 8.6 se deduce que, si un método es de orden p y $r(h) = O(h^p)$, entonces $\|y(a + nh) - y_n(h)\| = O(h^p)$.

Ejemplo. El método de Euler $y_{n+1} = y_n + hf_n$ es de orden 1. Estudiamos, primero, el error local de truncamiento

$$\tau(x, h) = h^{-1}(y(x + h) - y(x)) - f(x, y(x))$$

donde, suponiendo $f \in \mathcal{C}^1$ y desarrollando $y(x + h)$ alrededor del punto x , tenemos

$$\tau(x, h) = h^{-1}(hf + O(h^2)) - f(x, y(x)) = O(h)$$

Finalmente, si $\tau(x, h) = O(h)$, entonces $\tau(h)$ también lo será.

Ejercicios.

1. El método de Heun es de orden 2.
2. Todo método lineal multipaso (8.3), con $\alpha_k = 1$, es al menos de orden p si

$$i) \quad \sum_{i=0}^k \alpha_i = 0 \quad \sum_{i=0}^k \beta_i = k + (k-1)\alpha_{k-1} + \cdots + \alpha_1 \quad (8.5)$$

$$ii) \quad \beta_k k^{j-1} + \beta_{k-1} (k-1)^{j-1} + \cdots + \beta_1 = \frac{1}{j} [k^j + \alpha_{k-1} (k-1)^j + \cdots + \alpha_1] \\ \text{para } j = 2, \dots, p \quad (8.6)$$

y será exactamente de orden p si ii) no es cierto para $j = p + 1$.

3. Demostrar que el método

$$y_{n+2} - 4y_{n+1} + 3y_n = -2hf_n$$

es de orden 2. En 8.3.5 hay una aplicación numérica de este método.

El segundo ejercicio nos da una manera sencilla de calcular el orden de un método lineal multipaso cualquiera. Observar, también, que la primera condición 8.5 no es nada más que las propiedades que tiene que cumplir un método lineal consistente y, por lo tanto, todo método lineal multipaso consistente será, como mínimo, de orden 1, y viceversa, todo método lineal multipaso de orden igual o superior a 1 será consistente.

8.3.3 Métodos lineales multipaso. Teorema de Dahlquist

La ecuación general de un método lineal de k pasos es, como ya se ha visto,

$$\sum_{i=0}^k \alpha_i y_{n+i} = h \sum_{i=0}^k \beta_i f_{n+i} \quad \text{con } \alpha_k \neq 0 \quad (8.7)$$

Este método tiene dos polinomios asociados de gran importancia

$$\rho(\zeta) = \sum_{i=0}^k \alpha_i \zeta^i \quad \sigma(\zeta) = \sum_{i=0}^k \beta_i \zeta^i$$

Anteriormente, se ha estudiado cuándo un método lineal multipaso es estable y de qué manera se puede saber el orden. Como el orden de un método está totalmente relacionado con el orden de convergencia, es necesario preguntarse cuál puede ser el orden más grande posible de un método de este tipo bajo la condición de estabilidad. Esto queda resuelto por el **teorema de Dahlquist**:

Teorema 8.7 El orden más grande posible de un método lineal de k pasos estable es:

- $k + 2$ si k es par.
- $k + 1$ si k es impar.

Tal como indica el teorema, los métodos lineales multipaso estables tienen un orden máximo en función del número de pasos; entonces, cuando esto es así se dice que es un **método óptimo**.

La demostración del teorema es bastante extensa y a ([Hen62]) está claramente detallada. Algunos resultados intermedios son muy útiles para la construcción de métodos óptimos, y los resumimos en los dos lemas siguientes:

Lema 8.1 El orden p de un método lineal estable y consistente de k pasos es, como máximo $k + 2$. Una condición necesaria para que $p = k + 2$ es que:

1. k sea par.
2. todas las raíces del polinomio característico del método $\rho(\zeta)$ tengan módulo igual a 1.

Lema 8.2 Dado un método lineal de k pasos del tipo 8.7, con k par, y de orden máximo $k+2$, los polinomios $\rho(\zeta)$ y $\sigma(\zeta)$ cumplen las relaciones siguientes:

- $\rho(\zeta) = \zeta^{k/2}(\zeta - \zeta^{-1})P(\zeta + \zeta^{-1} - 2)$ con P polinomio de grado $k/2 - 1$.
- $\sigma(\zeta) = \zeta^{k/2}Q(\zeta + \zeta^{-1} - 2)$ con Q polinomio de grado $k/2$.
- donde los polinomios P y Q cumplen

$$\frac{t\sqrt{1+\frac{t^2}{4}}}{\log(\frac{t}{2}+\sqrt{1+\frac{t^2}{4}})} P(t^2) - Q(t^2) = c_{k+2} t^{k+2} + O(t^{k+4})$$

El primer lema caracteriza el polinomio $\rho(\zeta)$ y el segundo lema lo relaciona con $\sigma(\zeta)$. Por otro lado, la función $\frac{t\sqrt{1+\frac{t^2}{4}}}{\log(\frac{t}{2}+\sqrt{1+\frac{t^2}{4}})}$ es par y los coeficientes de la su serie de potencias son

j	0	1	2	3	4	5	\dots
k_{2j}	2	$\frac{1}{3}$	$-\frac{1}{90}$	$\frac{1}{756}$	$-\frac{23}{113400}$	$\frac{263}{7484400}$	

El hecho de utilizar la variable $\zeta + \zeta^{-1} - 2$ en los polinomios P y Q facilita la expresión del método en función de la tabla de diferencias finitas, ya que el término $\zeta^{k/2}(\zeta + \zeta^{-1} - 2)^\mu$ está asociado a la diferencia $\nabla^{2\mu} f_{n+k/2+\mu}$.

Ejemplo. Para construir el método óptimo de dos pasos ($k = 2$) y, por lo tanto, de orden $p = 4$ es necesario, en primer lugar, que estudiemos el polinomio $\rho(\zeta)$. Por el lema 8.1 tenemos que sus dos raíces han de ser de módulo igual a 1 si, además, le añadimos que sea consistente, propiedad que debe ser cierta para ser el orden superior a 1; entonces, una de ellas es $\zeta_1 = 1$; por lo tanto, la otra será forzosamente $\zeta_2 = -1$, ya que el método ha de ser estable. Así pues, el polinomio será

$$\rho(\zeta) = \zeta^2 - 1 = \zeta(\zeta - \zeta^{-1})$$

esto quiere decir que $P(t^2) = 1$.

El polinomio $Q(t^2)$ será de la forma $Q(t^2) = q_0 + q_2 t^2$ donde, utilizando la ecuación del lema 8.2, obtenemos

$$Q(t^2) = 2 + \frac{1}{3}t^2$$

y, por lo tanto, $\sigma(\zeta) = \zeta[2 + \frac{1}{3}(\zeta + \zeta^{-1} - 2)]$, y el método será

$$y_{n+2} - y_n = h(2f_{n+1} + \frac{1}{3} \nabla^2 f_{n+2}) = \frac{h}{3}(f_n + 4f_{n+1} + f_{n+2})$$

8.3.4 Estabilidad absoluta

Otra manera de estudiar el comportamiento del error de los distintos iterados de la solución numérica de la ecuación diferencial consiste en ver cuál es su comportamiento en una ecuación test $y' = \lambda y$ con $\lambda \in \mathbf{C}$ y $y, y' \in \mathbf{R}$, es decir, sobre la parte lineal de una ecuación general $y' = f(x, y)$. La función $\phi(x_{n+k}, \dots, x_n; y_{n+k}, \dots, y_n; h)$ de la mayoría de métodos aplicada a la ecuación test se puede expresar como una combinación lineal de las y_{n+i} :

$$\phi(x_{n+k}, \dots, x_n; y_{n+k}, \dots, y_n; h) = \lambda \sum_{i=0}^k g_i(h\lambda) y_{n+i} \quad (8.8)$$

Ejercicios.

1. Para un método lineal multipaso las funciones g_i son de la forma $g_i(h\lambda) = \beta_i$.
2. Los métodos del tipo Runge-Kutta, por ser métodos de un solo paso, solamente tienen dos funciones g_i :

$$g_1(h\lambda) = 0 \text{ y } g_0(h\lambda) = \frac{P_{R-1}(h\lambda)}{Q_R(h\lambda)}$$

con P_{R-1} y Q_R polinomios de grado $R - 1$ y R respectivamente y $Q_R \equiv 1$ si el método es explícito.

Seguidamente, denotamos por \bar{h} el valor de $h\lambda$, por $y(x_n)$ el valor exacto de la solución de la ecuación diferencial $y' = \lambda y$ en el punto x_n y por y_n los valores numéricos de los cálculos efectuados. Estos últimos tendrán un cierto error y, por lo tanto, cumplirán la ecuación siguiente:

$$\sum_{i=0}^k \alpha_i y_{n+i} = \bar{h} \sum_{i=0}^k g_i(\bar{h}) y_{n+i} + \delta_{n+k}$$

Definimos, también, la medida de error siguiente:

$$\tau_{n+k} = \tau(x_n, h) = h^{-1} \sum_{i=0}^k \alpha_i y(x_{n+i}) - \phi(x_{n+k}, \dots, x_n; y(x_{n+k}), \dots, y(x_n); h)$$

por lo tanto, considerando que ϕ es de la forma 8.8, podemos escribir

$$\sum_{i=0}^k \alpha_i y(x_{n+i}) = \bar{h} \sum_{i=0}^k g_i(\bar{h}) y(x_{n+i}) + h \tau_{n+k}$$

Entonces, el error total en el punto x_n que es $\varepsilon_n = y_n - y(x_n)$ cumple la ecuación en diferencias no homogénea siguiente:

$$\sum_{i=0}^k [\alpha_i - \bar{h} g_i(\bar{h})] \varepsilon_{n+i} = \delta_{n+k} - h \tau_{n+k}$$

Esta ecuación tiene su correspondiente polinomio característico que llamaremos **polinomio de estabilidad absoluta**

$$\Pi(r; \bar{h}) = \sum_{i=0}^k [\alpha_i - \bar{h}g_i(\bar{h})]r^i$$

Está claro que la estabilidad de la ecuación en diferencias del error vendrá dada por las raíces de este polinomio y, entonces, se define que un método es **absolutamente estable** para un valor \bar{h} dado, si todas las raíces r_i de su polinomio de estabilidad absoluta cumplen que $|r_i| < 1$. Se llama **región de estabilidad absoluta** de un método al conjunto $R_A \subset \mathbf{C}$ tal que este método es absolutamente estable $\forall \bar{h} \in R_A$.

Ejemplo. El método de Euler adelante $y_{n+1} = y_n + hf_n$ tiene las dos funciones g_i siguientes:

$$g_0(\bar{h}) = 1 \text{ y } g_1(\bar{h}) = 0$$

por lo tanto, el polinomio de estabilidad absoluta es

$$\Pi(r; \bar{h}) = r - 1 - \bar{h}$$

entonces, la raíz de este polinomio es $r_1 = 1 + \bar{h}$, y si exigimos que $|r_1| < 1$, la región de estabilidad absoluta del método de Euler es la que se puede ver a la figura 8.1.

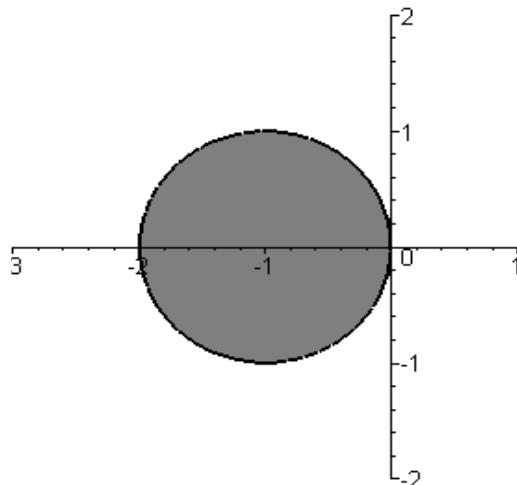


Fig. 8.1 Región de estabilidad absoluta del método de Euler.

Ejercicios.

1. El polinomio de estabilidad absoluta de un método lineal de k pasos es

$$\Pi(r; \bar{h}) = \rho(r) - \bar{h}\sigma(r)$$

2. Hallar la región de estabilidad absoluta del método de Euler atrás

$$y_{n+1} = y_n + hf_{n+1}$$

Se ha visto anteriormente que una de las condiciones necesarias para que un método sea consistente es que 1 sea una raíz de su polinomio característico $\rho(r)$; por otro lado, si añadimos la condición de estabilidad del método, hace falta que esta raíz sea simple. Además, se observa que, si $\bar{h} = 0$, los polinomios de estabilidad absoluta y el polinomio característico de un método lineal multipaso coinciden. Debido a la continuidad de las raíces de un polinomio respecto a sus coeficientes, denotamos por r_1 la raíz de $\Pi(r; \bar{h})$ tal que $\lim_{\bar{h} \rightarrow 0} r_1 = 1$. Todo esto, nos permite enunciar el teorema siguiente:

Teorema 8.8 Bajo las condiciones y notación enunciadas anteriormente, si el método lineal multipaso es de orden $p \geq 1$, entonces

$$r_1 = e^{\bar{h}} + O(\bar{h}^{p+1}) \quad \text{para } \bar{h} \rightarrow 0$$

Demostración: Como el método es de orden p tenemos que $\tau(x, h) = O(h^p)$. Si consideramos, en particular, la función $y(x) = \exp(\lambda x)$, que es solución de la ecuación diferencial $y' = \lambda y$, podemos escribir

$$\begin{aligned} \tau(x, h) &= h^{-1} \sum_{i=0}^k \alpha_i \exp(\lambda(x_n + ih)) - \sum_{i=0}^k \beta_i \lambda \exp(\lambda(x_n + ih)) = O(\bar{h}^p) \\ h\tau(x, h) &= \exp(\lambda x_n) \left[\sum_{i=0}^k \alpha_i [\exp \bar{h}]^i - \bar{h} \sum_{i=0}^k \beta_i [\exp \bar{h}]^i \right] = O(\bar{h}^{p+1}) \end{aligned}$$

por lo tanto, $\Pi(\exp(\bar{h}); \bar{h}) = \rho(\exp(\bar{h})) - \bar{h}\sigma(\exp(\bar{h})) = O(\bar{h}^{p+1})$. Si denotamos para r_1, \dots, r_k las k raíces del polinomio Π , resulta que

$$\Pi(\exp(\bar{h}); \bar{h}) = (\exp(\bar{h}) - r_1)(\exp(\bar{h}) - r_2) \cdots (\exp(\bar{h}) - r_k) = O(\bar{h}^{p+1})$$

Si hacemos $\bar{h} \rightarrow 0$, tenemos que $\exp(\bar{h}) \rightarrow 1$, $r_1 \rightarrow 1$, y no las otras raíces, ya que r_1 es simple por ser el método estable. Así pues, como el primer factor es el único que tiende a cero, es claro que

$$\exp(\bar{h}) - r_1 = O(\bar{h}^{p+1})$$

como se pretendía demostrar. \square

Este teorema nos demuestra que todo método multipaso lineal consistente y estable no es absolutamente estable para \bar{h} con parte real positiva y suficientemente pequeña.

8.3.5 Ejemplos numéricos

Una vez vistos los conceptos de consistencia, estabilidad, estabilidad absoluta y orden, estudiaremos unos cuantos casos numéricos por observar que efecto tienen sobre una ecuación en particular.

Consideramos el problema de condiciones iniciales:

$$y' = 4xy^{1/2} \quad \text{con } y(0) = 1, \quad x \in [0, 2]$$

* El primer método que estudiaremos es el siguiente:

$$y_{n+2} - (1-a)y_{n+1} = \frac{h}{2}[(3-a)f_{n+1} - (1-a)f_n]$$

Fácilmente se comprueba que, para $a = 0$, el método es estable, consistente y de orden 2, es decir, es convergente con un error de orden h^2 . Para $a = -5$, no es ni estable ni consistente.

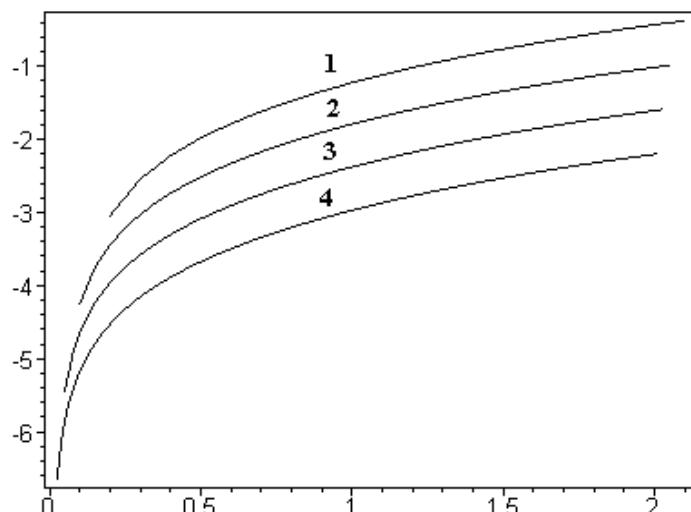


Fig. 8.2 Error de un método estable, consistente y de orden 2.
(la ordenada representa potencias de 10)

- 1. $h = 0.1$
- 2. $h = 0.05$
- 3. $h = 0.025$
- 4. $h = 0.0125$

Si aplicamos este método a la ecuación anterior con $a = 0$ y pasos $h = 0.1, 0.05, 0.025, 0.0125$ y dibujamos en unos ejes semilogarítmicos los correspondientes errores, observamos en la figura

8.2 que el error crece de forma exponencial y con la misma exponencial para todos los pasos. Si ahora estudiamos el error en el punto $x = 2$ y el cociente de estos, tal como se representa en la tabla siguiente:

h	error	cociente de errores
0.1	$3.675 \cdot 10^{-1}$	
0.05	$9.633 \cdot 10^{-2}$	0.2621
0.025	$2.462 \cdot 10^{-2}$	0.2556
0.0125	$6.222 \cdot 10^{-3}$	0.2527
0.00625	$1.564 \cdot 10^{-3}$	0.2513

observamos que, en dividir el paso por 2, el error se divide por $2^2 = 4$, tal como indican los cocientes de los errores, que tienden a 0.25, como corresponde a un método de orden 2. Resumiendo, tenemos convergencia.

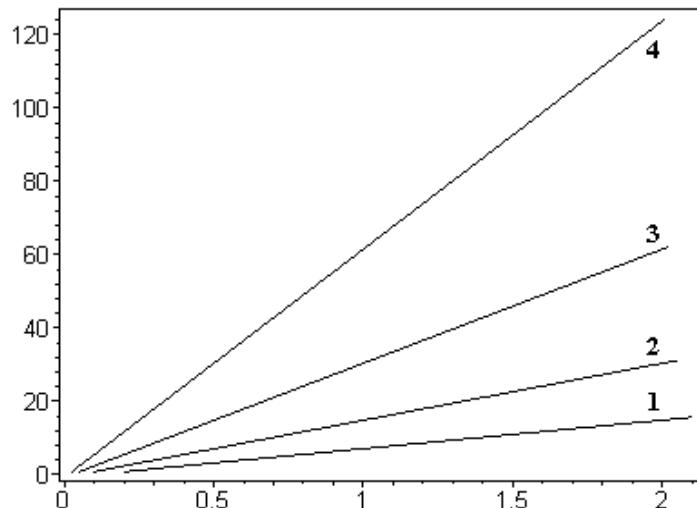


Fig. 8.3 Error de un método ni estable ni consistente.
(la ordenada representa potencias de 10)

- 1. $h = 0.1$
- 2. $h = 0.05$
- 3. $h = 0.025$
- 4. $h = 0.0125$

* Si ahora hacemos lo mismo pero con $a = -5$, se observa, en la figura 8.3, un crecimiento exponencial del error, pero esta vez la exponencial es más grande como más pequeño es el paso. No tenemos convergencia.

- ★ El segundo método es consistente, de orden 2, pero no es estable:

$$y_{n+2} - 4y_{n+1} + 3y_n = -2hf_n$$

También lo aplicamos al mismo problema de condiciones iniciales y con los mismos pasos, y obtenemos unos resultados muy semejantes a los del último caso anterior, tal como se ve en la figura 8.4. Es decir, crecimiento exponencial del error con la exponencial cada vez más grande como más pequeño es el paso y, por lo tanto, ausencia de convergencia.

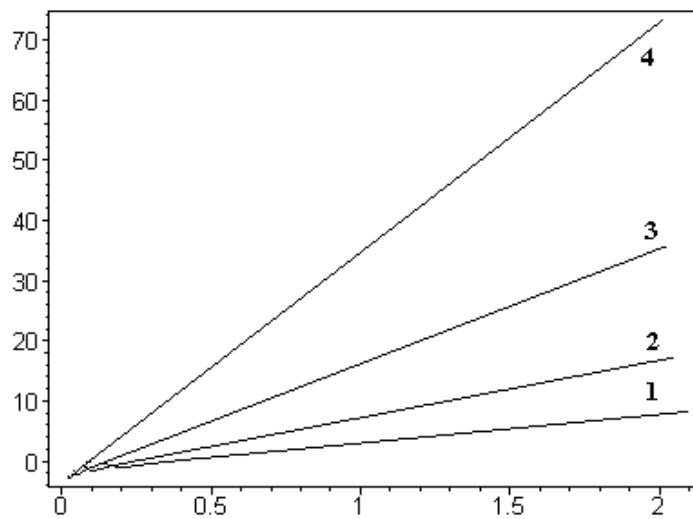


Fig. 8.4 Error de un método consistente, de orden 2, pero no estable.
(la ordenada representa potencias de 10)

1. $h = 0.1$
2. $h = 0.05$
3. $h = 0.025$
4. $h = 0.0125$

- ★ El tercer método es estable, pero no es consistente:

$$y_{n+2} - y_{n+1} = \frac{h}{3}[3f_{n+1} - 2f_n]$$

Como en los casos anteriores, lo aplicamos al mismo problema de condiciones iniciales y con los mismos pasos. Si dibujamos la solución exacta y las diferentes soluciones calculadas, parece que estas últimas converjan a alguna solución particular, pero que claramente no es la buena.

Es más, si dibujamos las diferencias entre dos soluciones calculadas de pasos consecutivos, éstas son cada vez más pequeñas, tal como se puede observar en la figura 8.5.

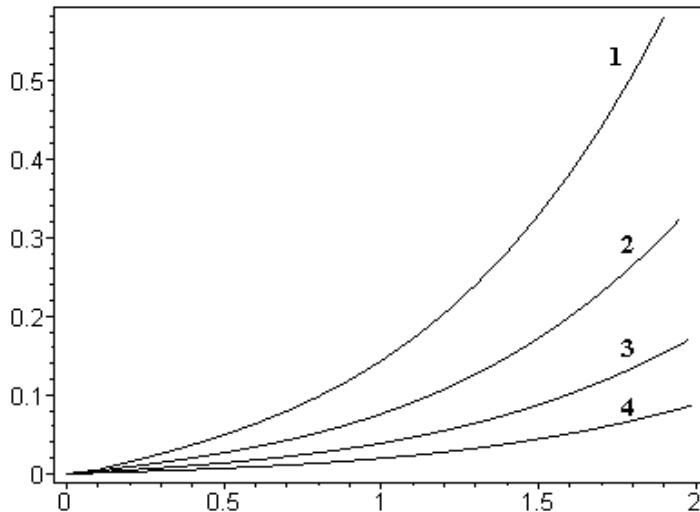


Fig. 8.5 Método estable, pero no consistente.

1. $(h = 0.05) - (h = 0.1)$
2. $(h = 0.025) - (h = 0.05)$
3. $(h = 0.0125) - (h = 0.025)$
4. $(h = 0.00625) - (h = 0.0125)$

* Finalmente, el último método que estudiaremos es el método d'Euler aplicado al problema de condiciones iniciales:

$$y' = -20y \quad \text{con} \quad y(0) = 1 \quad y \in [0, 2]$$

Como ya sabemos, el intervalo de estabilidad absoluta del método de Euler es $(-2, 0)$ y, por lo tanto, con una $\lambda = -20$ por tener estabilidad absoluta, es necesario que $h < 0.1$. Así pues, aplicamos el método con pasos $h = 0.05, 0.1, 0.2$ y observemos en la figura 8.6 que el gráfico semilogarítmico de los errores tiene tres comportamientos bien diferentes y claros. Si estamos dentro del intervalo de estabilidad absoluta, hay decrecimiento exponencial del error; si estamos fuera, tenemos crecimiento exponencial del error y si estamos en la frontera, el error se mantiene más o menos constante.

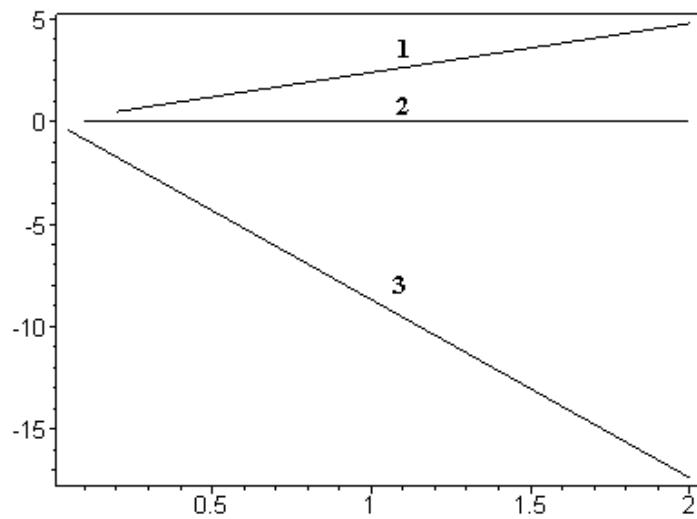


Fig. 8.6 Método de Euler y estabilidad absoluta.
(la ordenada representa potencias de 10)

- 1. $h = 0.2$
- 2. $h = 0.1$
- 3. $h = 0.05$

8.3.6 Métodos predictor-corrector

Consideramos un método lineal multipaso implícito de la forma

$$y_{n+k} + \sum_{i=0}^{k-1} \alpha_i y_{n+i} = h \beta_k f(x_{n+k}, y_{n+k}) + h \sum_{i=0}^{k-1} \beta_i f_{n+i}$$

En general, esta ecuación de variable y_{n+k} no es lineal y para hallar la solución es necesario utilizar algún método de cálculo de ceros de funciones no lineales. En particular, se utiliza la iteración siguiente:

$$y_{n+k}^{[s+1]} + \sum_{i=0}^{k-1} \alpha_i y_{n+i} = h \beta_k f(x_{n+k}, y_{n+k}^{[s]}) + h \sum_{i=0}^{k-1} \beta_i f_{n+i}$$

La convergencia de este método queda asegurada por el teorema del punto fijo que exige una constante de Lipschitz $L < 1$ de la iteración anterior. Por lo tanto, si M es la constante de Lipschitz de la función f , tendremos convergencia si

$$h < \frac{1}{M|\beta_k|} \quad (8.9)$$

Ahora solamente hace falta saber cuál es el primer elemento y_0 de la iteración; para esto se utiliza un método lineal multipaso explícito. Este tipo de algoritmo es el conocido como un **método predictor-corrector**. Con más detalle, la mecánica del algoritmo es la siguiente:

aplicamos el predictor \longrightarrow	$y_{n+k}^{[0]}$	P
evaluamos $f(y_{n+k}^{[0]})$	$f_{n+k}^{[0]}$	E
aplicamos el corrector \longrightarrow	$y_{n+k}^{[1]}$	C $P(EC)$
evaluamos $f(y_{n+k}^{[1]})$	$f_{n+k}^{[1]}$	E
aplicamos el corrector \longrightarrow	$y_{n+k}^{[2]}$	C $P(EC)^2$
⋮	⋮	⋮
evaluamos $f(y_{n+k}^{[m-1]})$	$f_{n+k}^{[m-1]}$	E
aplicamos el corrector \longrightarrow	$y_{n+k}^{[m]}$	C $P(EC)^m$

Para el siguiente paso y_{n+k+1} podemos utilizar $f_{n+k}^{[m-1]}$ o bien $f_{n+k}^{[m]}$, y se obtienen, finalmente, dos esquemas distintos

$$P(EC)^m \quad \text{o bien} \quad P(EC)^m E$$

Ejemplo. El método de Adams-Bashforth de orden 4, empieza con el método explícito de orden 2

$$y_{k+2} = y_{k+1} + \frac{h}{2}(3f_{k+1} - f_k)$$

que consiste en integrar el polinomio de grado 1 que interpola los puntos (x_{k-1}, y_{k-1}) y (x_k, y_k) . Continúa con un otro método explícito de orden 3

$$y_{k+3} = y_{k+2} + \frac{h}{12}(23f_{k+2} - 16f_{k+1} + 5f_k)$$

que es la integración del polinomio de grado 2 que interpola los puntos (x_{k-2}, y_{k-2}) , (x_{k-1}, y_{k-1}) y (x_k, y_k) . Para terminar con el método de orden 4

$$y_{k+4} = y_{k+3} + \frac{h}{24}(55f_{k+3} - 59f_{k+2} + 37f_{k+1} - 9f_k)$$

que es el predictor a utilizar.

El corrector es el método de Adams-Moulton, también de orden 4, pero implícito

$$y_{k+4} = y_{k+3} + \frac{h}{24}(9f_{k+4} + 19f_{k+3} - 5f_{k+2} + f_{k+1})$$

y que se utiliza con un esquema del tipo *PCEC*. Es decir, una vez generados los primeros cuatro puntos iniciales, el algoritmo es

$$\begin{aligned} y_{k+4}^{[0]} &= y_{k+3} + \frac{h}{24}(55f_{k+3} - 59f_{k+2} + 37f_{k+1} - 9f_k) \\ f_{k+4}^{[0]} &= f(x_{k+4}, y_{k+4}^{[0]}) \\ y_{k+4}^{[1]} &= y_{k+3} + \frac{h}{24}(9f_{k+4}^{[0]} + 19f_{k+3} - 5f_{k+2} + f_{k+1}) \\ f_{k+4} &= f(x_{k+4}, y_{k+4}^{[1]}) \end{aligned}$$

Con este tipo de esquema, el intervalo de estabilidad absoluta es $(-1.25, 0)$ y una estimación del error se puede calcular por $c_5 h^5 y^{(5)}(x_k) \sim -\frac{19}{270} (y_{k+4}^{[1]} - y_{k+4}^{[0]})$. Estos resultados y el esquema utilizado quedan justificados por los teoremas que veremos a continuación.

Una vez visto en qué consisten los métodos predictor-corrector, es necesario estudiar cuál es su error local de truncamiento y su estabilidad absoluta.

Error local de truncamiento

Consideramos un predictor definido por los polinomios

$$\rho^*(r) = \sum_{i=0}^k \alpha_i^* r^i \quad \text{con } \alpha_k^* = 1 \quad \text{y} \quad \sigma^*(r) = \sum_{i=0}^{k-1} \beta_i^* r^i$$

con error local de truncamiento $\tau^*(x, h)$ y orden q^* y un corrector definido de manera semejante por

$$\rho(r) = \sum_{i=0}^k \alpha_i r^i \quad \text{con } \alpha_k = 1 \quad \text{y} \quad \sigma(r) = \sum_{i=0}^k \beta_i r^i$$

con error local de truncamiento $\tau(x, h)$ y orden q . Además, suponemos que todas las soluciones numéricas en los puntos previos x_{n+j} para $j = 0 \div k-1$ son exactas y que la solución de la ecuación diferencial es suficientemente diferenciable.

Siendo q^* el orden del predictor, podemos escribir que

$$y(x_{n+k}) - y_{n+k}^{[0]} = h\tau(x_n, h) = c_{q^*+1}^* h^{q^*+1} y^{(q^*+1)}(x) + O(h^{q^*+2})$$

De la misma manera, siendo q el orden del corrector, tenemos que

$$\begin{aligned} y(x_{n+k}) - y_{n+k}^{[s+1]} &= h\beta_k [f(x_{n+k}, y(x_{n+k})) - f(x_{n+k}, y_{n+k}^{[s]})] + h\tau(x_n, h) = \\ &= h\beta_k \left[y(x_{n+k}) - y_{n+k}^{[s]} \right] \frac{\partial f(x_{n+k}, \eta_{n+k,s})}{\partial y} + h\tau(x_n, h) \end{aligned}$$

donde $\eta_{n+k,s}$ es un punto intermedio entre $y_{n+k}^{[s]}$ y $y(x_{n+k})$.

Si ahora suponemos que $q^* \geq q$, es decir, que el orden del predictor es superior o igual al orden del corrector, y estudiamos la expresión anterior para $s = 0, s = 1$, etc., obtenemos la expresión final siguiente:

$$y(x_{n+k}) - y_{n+k}^{[m]} = c_{q+1} h^{q+1} y^{q+1)}(x_n) + O(h^{q+2})$$

Esto nos dice que el error final es semejante al del corrector, independientemente del número de correcciones que se hagan.

Haciendo operaciones semejantes para $q^* < q$, podemos enunciar el resultado siguiente:

Teorema 8.9 Consideramos un método predictor-corrector de ordenes q^* y q respectivamente, con esquemas $P(EC)^m E$ o $P(EC)^m$ y tales que $q^* \geq 0, q \geq 1, m \geq 1$. Entonces, el error local de truncamiento es:

1. Si $q^* \geq q$, entonces $y(x_{n+k}) - y_{n+k}^{[m]} = c_{q+1} h^{q+1} y^{q+1)}(x_n) + O(h^{q+2})$.
2. Si $q^* < q$ y $r = q - q^*$, entonces, según los valores de r , tenemos:
 - $m < r$: $y(x_{n+k}) - y_{n+k}^{[m]} = kh^{q^*+m+1} + O(h^{q^*+m+2})$
 - $m = r$: $y(x_{n+k}) - y_{n+k}^{[m]} = kh^{q+1} + O(h^{q+2})$
 - $m > r$: $y(x_{n+k}) - y_{n+k}^{[m]} = c_{q+1} h^{q+1} y^{q+1)}(x_n) + O(h^{q+2})$

Estabilidad absoluta

En primer lugar, es necesario ver cuáles son los polinomios de estabilidad absoluta de los métodos predictor-corrector.

Teorema 8.10 El polinomio de estabilidad absoluta de un método predictor corrector de esquema $P(EC)^m E$ es

$$\Pi_{P(EC)^m E}(r, \bar{h}) = \rho(r) - \bar{h}\sigma(r) + M_m(\bar{h})[\rho^*(r) - \bar{h}\sigma^*(r)],$$

y, si el esquema es del tipo $P(EC)^m$, el polinomio es:

$$\Pi_{P(EC)^m}(r, \bar{h}) = \beta_k r^k [\rho(r) - \bar{h}\sigma(r)] + M_m(\bar{h})[\rho^*(r)\sigma(r) - \rho(r)\sigma^*(r)],$$

donde

$$M_m(\bar{h}) = (\bar{h}\beta_k)^m \frac{1 - \bar{h}\beta_k}{1 - (\bar{h}\beta_k)^m} \quad \text{para } m = 1, 2, \dots$$

Demostración: ([Lam79]) \square

Si consideramos la condición vista anteriormente sobre la convergencia del corrector, tenemos que $|\bar{h}\beta_k| < 1$ y, por lo tanto, $M_m(\bar{h}) \rightarrow 0$ para $m \rightarrow \infty$; los polinomios de estabilidad absoluta tienden, pues, hacia el polinomio de estabilidad absoluta del corrector para $m \rightarrow \infty$. Finalmente, utilizando la continuidad de las raíces de un polinomio respecto de sus coeficientes, podemos enunciar el resultado siguiente:

Teorema 8.11 El polinomio de estabilidad absoluta de cualquier método predictor-corrector tiene siempre una raíz del tipo $r_1 = \exp(\bar{h}) + O(\bar{h}^{q+1})$, donde q es el orden del corrector.

Por lo tanto, de este teorema se deduce fácilmente que cualquier método predictor corrector es absolutamente inestable para \bar{h} con parte real positiva y suficientemente pequeña.

Control del paso

Al utilizar un método predictor-corrector se debe tener en cuenta diferentes aspectos para determinar el paso conveniente de integración:

1. **Control del error.** Se debe obtener a cada paso alguna estimación del error, para saber si estamos integrando según la tolerancia pedida.
2. **Control de la estabilidad absoluta.** Es necesario que \bar{h} esté dentro de la región o intervalo de estabilidad absoluta.
3. **Control de la convergencia.** Hace falta que el paso sea el adecuado, según la cota 8.9, para que el corrector converja.

8.3.7 Métodos Runge-Kutta

Estudiaremos únicamente los métodos Runge-Kutta explícitos, que, tal como habíamos visto anteriormente, son de la forma:

$$\begin{aligned} y_{n+1} &= y_n + h \sum_{r=1}^R c_r k_r \\ k_1 &= f(x, y) \\ k_r &= f\left(x + ha_r, y + h \sum_{s=1}^{r-1} b_{rs} k_s\right) \quad \text{con} \quad a_r = \sum_{s=1}^{r-1} b_{rs} \quad r = 2 \div R \end{aligned}$$

Como su polinomio característico es $p(\lambda) = \lambda - 1$, estos métodos son siempre estables. Además, si $\sum_{r=1}^R c_r = 1$, entonces también son consistentes. Muchas veces, esta última condición sobre las c_r se incluye en la misma definición del método.

El siguiente concepto a estudiar de un método es el orden. Si definimos la función $q(R)$ como el orden más grande posible con una fórmula de R sumandos, ([Lam79]) nos da para un método explícito la tabla siguiente (para órdenes elevados a ([But85]) hay cotas más precisas):

$$\left| \begin{array}{ll} q(R) = R & \text{para } R = 1, 2, 3, 4 \\ q(5) = 4 & \\ q(6) = 5 & \\ q(7) = 6 & \end{array} \right| \left| \begin{array}{l} q(8) = 6 \\ q(9) = 7 \\ q(R) \leq R - 2 \quad \text{para } R \geq 10 \end{array} \right.$$

Ejemplos.

1. El método de Euler modificado de segundo orden:

$$y_{n+1} = y_n + h f(x_n + \frac{h}{2}, y_n + \frac{h}{2} f(x_n, y_n))$$

2. La fórmula de Heun de tercer orden:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{4}[k_1 + 3k_3] \\ k_1 &= f(x_n, y_n) \\ k_2 &= f(x_n + \frac{h}{3}, y_n + \frac{h}{3}k_1) \\ k_3 &= f(x_n + \frac{2h}{3}, y_n + \frac{2h}{3}k_2) \end{aligned}$$

3. Regla de Kutta de tercer orden:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}[k_1 + 4k_2 + k_3] \\ k_1 &= f(x_n, y_n) \\ k_2 &= f(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_1) \\ k_3 &= f(x_n + h, y_n - hk_1 + 2hk_2) \end{aligned}$$

4. Método de Runge-Kutta de cuarto orden:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{6}[k_1 + 2k_2 + 2k_3 + k_4] \\ k_1 &= f(x_n, y_n) & k_3 &= f(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_2) \\ k_2 &= f(x_n + \frac{h}{2}, y_n + \frac{h}{2}k_1) & k_4 &= f(x_n + h, y_n + hk_3) \end{aligned}$$

Estimaciones del error

Algunos métodos Runge-Kutta concretos ya tienen una expresión en función de las k 's calculadas para hallar valores de τ_{n+1} de manera sencilla. Pero, en general estas fórmulas no existen y es necesario utilizar métodos más universales y, por lo tanto, un poco más costosos como, por ejemplo, la extrapolación de Richardson, donde es necesario calcular el valor y_{n+1} con paso h y el y_{n+1}^* con paso $2h$, y se obtiene finalmente la estimación del error siguiente:

$$y(x_{n+1}) - y_{n+1} \sim \frac{y_{n+1} - y_{n+1}^*}{2^{q+1} - 1}$$

donde q es el orden del método. Como se puede ver, este método produce un considerable aumento del número de operaciones a efectuar a cada paso.

Otro método por obtener estimaciones del error con un coste inferior es el conocido como el método **Runge-Kutta-Fehlberg**.

Métodos Runge-Kutta-Fehlberg

Estos métodos consisten en obtener la estimación del error utilizando la diferencia entre dos valores de y_{n+1} , pero obtenidos de dos métodos de orden diferente. Consideraremos los dos métodos Runge-Kutta siguientes:

$$\begin{aligned} y_{n+1} &= y_n + h \sum_{i=1}^R c_i k_i, \quad \text{de orden } q \quad \text{y} \quad \tau(x, h) = d(x)h^q + O(h^{q+1}) \\ y_{n+1}^* &= y_n + h \sum_{i=1}^{R^*} c_i^* k_i, \quad \text{de orden } q+1 \quad q \quad \text{y} \quad \tau^*(x, h) = d^*(x)h^{q+1} + O(h^{q+2}) \end{aligned}$$

Entonces, si suponemos que $y(x_n) = y_n$, como

$$\begin{aligned} y(x_{n+1}) - y_{n+1} &= h\tau(x_n, h) = hd(x_n)h^q + O(h^{q+2}) \\ y(x_{n+1}) - y_{n+1}^* &= h\tau^*(x_n, h) = hd^*(x_n)h^{q+1} + O(h^{q+3}) \end{aligned}$$

podemos escribir que

$$y_{n+1} - y_{n+1}^* = d(x_n)h^{q+1} + O(h^{q+2})$$

y, por lo tanto,

$$y(x_{n+1}) - y_{n+1} = d(x_n)h^{q+1} + O(h^{q+2}) \sim y_{n+1} - y_{n+1}^*$$

es decir, que se puede estimar el error que se ha producido por la diferencia entre los valores calculados por los dos métodos.

Entonces, el método Runge-Kutta-Fehlberg utiliza esta estimación del error para hallar un nuevo paso de integración conservando una cota determinada del error. Sea $\hat{h} = x_{n+2} - x_{n+1}$ el

nuevo paso, y ε la cota de error que queremos obtener, entonces es necesario que $|d(x_{n+1})\hat{h}^{q+1}| < \varepsilon$. Si solamente consideramos los términos de primer orden

$$|d(x_{n+1})| \sim |d(x_n)| \sim \frac{|y_{n+1} - y_{n+1}^*|}{h^{q+1}}$$

el nuevo paso ha de ser, por lo tanto,

$$\hat{h} < h \left[\frac{\varepsilon}{|y_{n+1} - y_{n+1}^*|} \right]^{\frac{1}{q+1}}$$

De todas maneras, normalmente se aplica una cierto factor de corrección $\alpha < 1$:

$$\hat{h} = ah \left[\frac{\varepsilon}{|y_{n+1} - y_{n+1}^*|} \right]^{\frac{1}{q+1}} \quad \text{con } \alpha = 0.9$$

Estabilidad absoluta de un método Runge-Kutta

Según el ejercicio 2 de 8.3.4, es fácil ver que, para un método Runge-Kutta explícito, el polinomio de estabilidad absoluta es la forma

$$\Pi(r; \bar{h}) = r - (1 + \bar{h}P_{R-1}(\bar{h}))$$

donde $P_{R-1}(\bar{h})$ es un polinomio de grado $R - 1$. Por lo tanto, la única raíz del polinomio es $r = 1 + \bar{h}P_{R-1}(\bar{h})$. En general, ([Lam79]) demuestra que, para un método explícito de orden q si $R = q$, ($q \leq 4$), entonces

$$r = 1 + \bar{h} + \dots + \frac{1}{q!}\bar{h}^q$$

es decir, que para $R = 1, 2, 3, 4$ dado, todos los métodos Runge-Kutta explícitos de R sumandos tienen el mismo intervalo de estabilidad absoluta, en particular

R	intervalo
1	(-2, 0)
2	(-2, 0)
3	(-2.51, 0)
4	(-2.78, 0)

Si $q < R$, ($q > 4$),

$$r = \sum_{i=1}^q \frac{1}{i!} \bar{h}^i + \sum_{i=q+1}^R \gamma_i \bar{h}^i$$

Por lo tanto, en los métodos Runge-Kutta explícitos, tenemos el mismo fenómeno de inestabilidad absoluta que en los métodos lineales multipaso.

8.3.8 Comparación entre los métodos predictor-corrector y Runge-Kutta

Para la comparación consideraremos los aspectos siguientes:

- Precisión local o orden del método.
- Estabilidad absoluta.
- Coste computacional medido por el número de evaluaciones de la función de la ecuación diferencial por cada paso de integración.
- Coste o dificultad de programación.

La primera observación que se debe hacer es que no es una comparación sencilla, ya que son dos tipos de métodos muy diferentes. Por ejemplo, el orden máximo de un método Runge-Kutta está en función del número de evaluaciones de la función de la ecuación diferencial; en cambio, para el método predictor-corrector el orden máximo está en función del número de pasos del método y el número de evaluaciones de la función queda fijado por el número de correcciones que se realizan.

El error local también es de difícil comparación, ya que para un método predictor-corrector tiene una expresión muy sencilla, pero para un método Runge-Kutta es mucho más complicada y difícil de evaluar para cualquier ecuación diferencial en general. A pesar de todo, la experiencia demuestra, por ejemplo, que para métodos de cuarto orden los Runge-Kutta dan mejor precisión que los métodos predictor-corrector. Pero esta relación se invierte si comparamos métodos del mismo orden sujetos al mismo número de evaluaciones de la función de la ecuación diferencial.

La variación del intervalo de estabilidad absoluta en función del orden es muy diferente para cada uno de los dos tipos de método. Los Runge-Kutta aumentan ligeramente su intervalo de estabilidad absoluta al aumentar el orden, pero los métodos multipaso lineales tienen una importante disminución del intervalo al aumentar su orden.

Finalmente, desde el punto de vista de la programación, los métodos Runge-Kutta son mucho más fáciles de escribir y de implementar; en cambio, los predictor-corrector tienen el grave inconveniente de necesitar algún tipo de inicialización, ya que suelen ser métodos de un paso. Si, además, queremos hacer un control y cambio del paso de integración, para los Runge-Kutta es suficiente utilizar, por ejemplo, un Runge-Kutta-Fehlberg; pero para los predictor-corrector necesitamos utilizar valores anteriores que, o bien no han estado calculados, o bien habrá sido necesario guardar.

8.4 Problema de valores frontera

Dada la ecuación diferencial $y' = f(x, y)$ con $y, y' \in \mathbf{R}^n$ el **problema de valores frontera** consiste en hallar una solución $y(x)$ de esta ecuación diferencial tal que cumpla la condición $r(y(a), y(b)) = 0$ con $r : \mathbf{R}^n \times \mathbf{R}^n \rightarrow \mathbf{R}^n$. Para este tipo de problemas no hay, en general, seguridad de existencia y unicidad de solución.

Ejemplos.

1. Todo problema de condiciones iniciales se puede ver como un problema de valores frontera.

Sea $y' = f(x, y)$ con $y(a) = y_0$, la correspondiente condición para el problema de valores frontera es $r(u, v) = u - y_0$.

2. Consideramos la ecuación diferencial de segundo orden $y'' + y = 0$: su solución general es $y(x) = c_1 \operatorname{sen} x + c_2 \cos x$. Entonces, consideramos diferentes condiciones frontera:

$y(0) = 0 \quad y(\pi/2) = 1$ solución única: $y(x) = \operatorname{sen}(x)$.

$y(0) = 0 \quad y(\pi) = 0$ infinitas soluciones: $y(x) = c_1 \operatorname{sen}(x)$.

$y(0) = 0 \quad y(\pi) = 1$ no hay ninguna solución.

Hay otros problemas que son equivalentes a un problema de valores frontera, tal como se ha enunciado al principio. Veremos dos muy importantes y conocidos.

Uno de los más típicos es el **problema de valores singulares**, que consiste en una ecuación diferencial de primer orden n dimensional dependiendo de un parámetro $y' = f(x, y, \lambda)$ a la cual exigimos que cumpla $n+1$ condiciones frontera. Este problema está sobre determinado y, por lo tanto, en general no tiene solución para cualquier λ . Entonces es necesario determinar para qué valores de λ , **valores singulares**, existe solución de la ecuación diferencial. Este es un problema muy fácil de formular adecuadamente. Definimos una nueva variable $y_{n+1} = \lambda$ y añadimos una nueva ecuación diferencial $y'_{n+1} = 0 = f_{n+1}$, a las n ecuaciones anteriores, y obtenemos, así, un sistema de ecuaciones diferenciales de primer orden de dimensión $n+1$ que, conjuntamente con las $n+1$ condiciones frontera, nos da un problema de valores frontera de dimensión $n+1$.

El otro problema es de valores frontera, pero con la **frontera libre**, que consiste en una ecuación diferencial $y' = f(x, y)$ de dimensión n , con $n+1$ condiciones frontera $r(y(a), y(b)) = 0$, pero con b desconocido. En este caso, para obtener el correspondiente problema de valores frontera es necesario definir unas nuevas variables $z_1 = y_1, \dots, z_n = y_n, z_{n+1} = b - a$ y un nuevo tiempo $t = \frac{x-a}{b-a}$. Entonces el nuevo sistema de ecuaciones diferenciales es

$$\begin{aligned}\dot{z}_i &= \frac{dz_i}{dt} = z_{n+1} f_i \quad \text{para } i = 1 \div n \\ \dot{z}_{n+1} &= \frac{dz_{n+1}}{dt} = 0\end{aligned}$$

que, conjuntamente con las $n+1$ condiciones frontera $r(z(0), z(1)) = 0$, nos queda formulado de la manera adecuada.

8.4.1 Método del tiro simple

Si miramos el problema de valores frontera como un problema de condiciones iniciales $y' = f(x, y)$ con $y(a) = s \in \mathbf{R}^n$, la condición frontera es

$$r(y(a; s), y(b; s)) = r(s, y(b; s)) = F(s) = 0$$

Por lo tanto, podemos reducir la solución del problema de valores frontera a hallar el cero de la función $F(s)$, en general no lineal. Para esto, podemos utilizar el método de Newton, que consiste en resolver el sistema lineal

$$DF(s^{(i)})d^{(i)} = -F(s^{(i)}) \quad \text{con} \quad d^{(i)} = s^{(i+1)} - s^{(i)}$$

Para el cálculo de $DF(s)$ se opera de la manera siguiente:

$$DF(s) = D_u r(s, y(b; s)) + D_v r(s, y(b; s))D_s y(s; b)$$

y, finalmente, la diferencial de $y(b; s)$ respecto a s se aproxima por alguna fórmula sencilla de diferencias finitas.

El principal problema de este método es su convergencia, ya que se está utilizando el método de Newton y, por lo tanto, es necesario que el punto $s^{(0)}$ esté suficientemente próximo a la solución exacta. Además, es necesario que a cada iteración la matriz $DF(s^{(i)})$ no sea nunca singular.

Ejemplo. Consideramos el problema de valores frontera

$$yy'' = -1 - y'^2 \quad \text{con} \quad y(0) = 1 \quad y \quad y(1) = 2.$$

Su solución es $y = \sqrt{5 - (x - 2)^2}$. Si lo queremos resolver por el método del tiro simple, primero es necesario plantear el sistema de ecuaciones diferenciales de primer orden:

$$\begin{aligned} \dot{z}_1 &= z_2 \\ \dot{z}_2 &= \frac{1}{z_1}(-1 - z_2^2) \end{aligned}$$

con la condición

$$r(z(0), z(1)) = \begin{pmatrix} z_1(0) - 1 \\ z_1(1) - 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Entonces, la función de la cual buscaremos su cero es

$$F(s) = r(s, z(1; s)) = \begin{pmatrix} s_1 - 1 \\ z_1(1; s) - 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

esta función, debido a la su sencillez, se puede simplificar, ya que es evidente que $s_1 = 1$ y, por lo tanto, nos queda solamente $z_1(1; s) - 2 = 0$, que es una función dependiente de una sola variable s_2 . Así pues, la nueva función, que denotaremos por el mismo nombre, es

$$F(s_2) = z_1(1; (1, s_2)) - 2 = 0$$

de la cual buscaremos su cero aplicando el método de Newton:

$$s_2^{(i+1)} = s_2^{(i)} - \frac{F(s_2^{(i)})}{F'(s_2^{(i)})}$$

donde aproximamos la derivada de F por la fórmula siguiente:

$$F'(x) \sim \frac{F(x + \varepsilon) - F(x - \varepsilon)}{2\varepsilon}$$

Los resultados obtenidos utilizando como integrador un Runge-Kutta-Fehlberg 4,5 con tolerancia de 10^{-9} y un $\varepsilon = 10^{-5}$ para aproximar la derivada están resumidos en la tabla siguiente:

iteración i	$s_2^{(i)}$	error
0	0.100000000000000	$1.9000 \cdot 10^0$
1	0.79442718961219	$1.2065 \cdot 10^0$
2	1.72656815254329	$2.7343 \cdot 10^{-1}$
3	1.98995511990323	$1.0045 \cdot 10^{-2}$
4	1.99998735761303	$1.2642 \cdot 10^{-5}$
5	2.00000000182309	$1.8231 \cdot 10^{-9}$

Tal como se puede observar a la tabla, se han necesitado cinco iteraciones del método de Newton para llegar a obtener un error del mismo orden que la tolerancia del integrador. Por otro lado, se ha comprobado experimentalmente que, si $\hat{s}_2^{(0)} = 0$, el Runge-Kutta tenía graves problemas de funcionamiento, a causa de necesitar un número excesivo de divisiones del paso, y no integraba correctamente. Finalmente, el software numérico MATLAB con el cual se han efectuado los cálculos ha realizado un total de 8750 operaciones básicas para obtener el resultado.

8.4.2 Método del tiro paralelo

Este método es muy semejante al anterior, pero trabajando con puntos intermedios. Es decir, es necesario hallar la solución $s_k \sim y(x_k)$ en los puntos $a = x_1 < x_2 < \dots < x_m = b$ simultáneamente. Sea $y(x; x_k, s_k)$ la solución de la ecuación diferencial con condiciones iniciales $y(x_k) = s_k$; entonces, el problema se reduce a hallar los s_k tales que la función definida por $y(x) = y(x; x_k, s_k)$ para $x \in [x_k, x_{k+1}]$ y $y(b) = s_m$ sea continua, solución de la ecuación diferencial y cumpla la condición frontera. Si hacemos un recuento de incógnitas observamos que tenemos nm y las ecuaciones son:

$$\begin{aligned} y(x_{k+1}; x_k, s_k) &= s_{k+1} \quad \text{para } k = 1 \div m - 1 \\ r(s_1, s_m) &= 0 \end{aligned}$$

Si definimos $\mathbf{s} = (s_1, \dots, s_m)^T$ el problema se reduce, una vez más, a hallar el cero de la función siguiente:

$$F(\mathbf{s}) = \begin{pmatrix} F_1(s_1, s_2) \\ F_2(s_2, s_3) \\ \vdots \\ F_{m-1}(s_{m-1}, s_m) \\ F_m(s_1, s_m) \end{pmatrix} = \begin{pmatrix} y(x_2; x_1, s_1) - s_2 \\ y(x_3; x_2, s_2) - s_3 \\ \vdots \\ y(x_m; x_{m-1}, s_{m-1}) - s_m \\ r(s_1, s_m) \end{pmatrix} = \mathbf{0}$$

Naturalmente lo resolvemos, también, por el método de Newton y, por lo tanto, hace falta resolver el sistema de ecuaciones lineal siguiente:

$$DF(\mathbf{s}^{(r)})\mathbf{d}^{(r)} = -F(\mathbf{s}^{(r)}) \quad \text{con } \mathbf{d}^{(r)} = \mathbf{s}^{(r+1)} - \mathbf{s}^{(r)}$$

Como se puede observar, la dimensión de este sistema puede ser bastante grande y, por lo tanto, el coste de su resolución importante; pero, haciendo unas cuantas operaciones, se puede reducir a resolver el sistema de ecuaciones de dimensión n siguiente:

$$(A + BG_{m-1}G_{m-2} \cdots G_1)d_1 = -[F_m + B(F_{m-1} + G_{m-1}F_{m-2} + \cdots + G_{m-1}G_{m-2} \cdots G_2F_1)] \quad (8.10)$$

y las otras d_i se calculan utilizando la fórmula

$$d_{i+1} = G_i d_i + F_i \quad \text{para } i = 1 \div m-1$$

Las G_i , la A y la B son las correspondientes submatrices de dimensión n de la matriz:

$$DF(\mathbf{s}) = \begin{pmatrix} G_1 & -I & & & 0 \\ & G_2 & -I & & \\ & & \ddots & \ddots & \\ & 0 & & G_{m-1} & -I \\ A & 0 & \dots & 0 & B \end{pmatrix} \quad (8.11)$$

En este caso, solamente necesitamos que la matriz $(A + BG_{m-1}G_{m-2}\cdots G_1)$ no sea singular. Se puede demostrar que las condiciones necesarias sobre $\mathbf{s}^{(0)}$ para la convergencia del método son más modestas para el tiro paralelo que para el tiro simple.

Ejemplo. Consideramos el mismo problema de valores frontera que el ejemplo mostrado en el tiro simple:

$$\begin{aligned} \dot{z}_1 &= z_2 \\ \dot{z}_2 &= \frac{1}{z_1}(-1 - z_2^2) \end{aligned}$$

con la condición

$$r(z(0), z(1)) = \begin{pmatrix} z_1(0) - 1 \\ z_1(1) - 2 \end{pmatrix} = \mathbf{0}$$

Si utilizamos el método del tiro paralelo dividiendo el intervalo de integración en 10 partes, $x_1 = 0, x_2 = 0.1, \dots, x_{11} = 1$, la función de la cual buscamos el cero es

$$F(\mathbf{s}) = \begin{pmatrix} z(x_2; x_1, s_1) - s_2 \\ \vdots \\ z(x_{11}; x_{10}, s_{10}) - s_{11} \\ r(s_1, s_m) \end{pmatrix} = \mathbf{0}$$

Las submatrices A y B de la matriz $DF(s)$ de 8.11 son muy sencillas

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

y el sistema de ecuaciones a resolver 8.10 queda reducido a

$$d_{11} = 0 \quad d_{12} = -\frac{h_1}{g_{12}}$$

donde

$$G_{m-1} \cdots G_1 = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}$$

$$F_{m-1} + G_{m-1}F_{m-2} + \cdots + G_{m-1}\cdots G_2F_1 = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

Si utilizamos como integrador los mismos Runge-Kutta, tolerancia y aproximación de la derivada del ejemplo del tiro simple; definimos, como primera aproximación de s , la recta que interpola los puntos $(0, 1)$ y $(1, 2)$ para la primera componente y la constante 0 para todas las segundas componentes, entonces obtenemos los resultados de la tabla siguiente:

iteración i	$z_2^{(i)}(0)$	error
0	0.000000000000000	$2.0000 \cdot 10^0$
1	1.33248750160296	$6.6751 \cdot 10^{-1}$
2	1.96170472949764	$3.8295 \cdot 10^{-2}$
3	1.99990917854788	$9.0821 \cdot 10^{-5}$
4	2.00000000028333	$2.8333 \cdot 10^{-10}$

Como se puede observar en los datos de la tabla, para obtener un error de orden semejante al de la tolerancia del integrador ha sido suficiente con cuatro iteraciones del método de Newton. Además, se ha podido empezar con $z_2(0) = 0$ sin ningún tipo de problema por parte del Runge-Kutta y para obtener estos resultados el software MATLAB ha necesitado hacer unas 25500 operaciones elementales.

8.5 Problemas

1. Considerar $B \in \mathcal{L}(\mathbf{R}^n)$ tal que $\rho(B) < 1$, y la ecuación en diferencias $y^{k+1} = By^k + d^k$ donde y^0 es dado y $\{d^k\}$ es una sucesión que satisface $\|d^k\| \leq \delta$ para $k = 0, 1, \dots$. Demostrar que existen constantes c_1, c_2, α con $\alpha < 1$, tales que

$$\|y^k\| \leq c_1 \alpha^k + c_2 \quad k = 0, 1, \dots$$

2. Definimos que una solución de la ecuación $y_k - \alpha_{n-1}y_{k-1} - \dots - \alpha_0y_{k-n} = 0$ es estable si, dado $\epsilon > 0$, existe $\delta > 0$ tal que, si $|\hat{y}_i - y_i| \leq \delta$, $i = 0 \div n-1$, implica que $|y_i - \hat{y}_i| \leq \epsilon$ $i = n, n+1, \dots$. Demostrar que esta definición es equivalente a la dada.
3. Estudiar bajo qué condiciones de las raíces del polinomio característico de una ecuación en diferencias del tipo $y_k - \alpha_{n-1}y_{k-1} - \dots - \alpha_0y_{k-n} = 0$ la correspondiente solución general tiende hacia cero para $n \rightarrow \infty$.
4. Estudiar la estabilidad y hallar la solución general de la ecuación en diferencias $y_{n+2} - 2\mu y_{n+1} + \mu y_n = 1$, con $0 < \mu < 1$. Demostrar que

$$\lim_{n \rightarrow \infty} y_n = \frac{1}{1-\mu}$$

5. Estudiar la estabilidad y hallar la solución general de la ecuación en diferencias

$$y_n = py_{n+1} + qy_{n-1} \quad \text{con } p+q=1$$

6. Resolver el sistema de ecuaciones en diferencias siguiente:

$$\begin{cases} x_{n+1} &= 7x_n + 10y_n \\ y_{n+1} &= x_n + 4y_n \end{cases}$$

con $x_0 = 3, y_0 = 2$.

7. Calcular un método lineal implícito de dos pasos que contenga un parámetro libre y que tenga el orden más alto posible.
8. Calcular la orden, la estabilidad y el error del **método de Quade**:

$$y_{n+4} - \frac{8}{19}(y_{n+3} - y_{n+1}) - y_n = \frac{6h}{19}(f_{n+4} + 4f_{n+3} + 4f_{n+1} + f_n)$$

9. Estudiar la estabilidad del método lineal multipaso

$$y_{n+2} - (1-a)y_{n+1} + ay_n = \frac{h}{2} [(3-a)f_{n+1} - (1+a)f_n]$$

para distintos valores de a .

10. Estudiar la consistencia del método lineal multipaso

$$y_{n+2} - y_{n+1} = \frac{h}{3} (3f_{n+1} - 2f_n)$$

11. Demostrar que el orden del método lineal multipaso

$$y_{n+2} + (b-1)y_{n+1} - by_n = \frac{h}{4} [(b+3)f_{n+2} + (3b+1)f_n]$$

es igual a 2 si $b \neq -1$ y a 3 si $b = -1$. Demostrar que este método no es estable si $b = -1$.

12. Demostrar que un método lineal multipaso de orden q integra exactamente toda ecuación diferencial, tal que sus soluciones sean polinomios de grado inferior o igual a q .

13. Un método lineal multipaso está definido por sus dos polinomios $\rho(\zeta), \sigma(\zeta)$. Considerar las sucesiones de polinomios $\{\rho_j(\zeta)\}, \{\sigma_j(\zeta)\}$ para $j = 1, 2, 3, \dots$ definidas de la manera siguiente:

$$\begin{aligned} \rho_1(\zeta) &= \rho(\zeta) & \sigma_1(\zeta) &= \sigma(\zeta) \\ \rho_{j+1}(\zeta) &= \zeta\rho'_j(\zeta) & \sigma_{j+1}(\zeta) &= \zeta\sigma'_j(\zeta) & j = 1, 2, 3, \dots \end{aligned}$$

Demostrar que este método es de orden igual a p si y solamente si

$$\rho_1(1) = 0 \quad \rho_{j+1}(1) = j\sigma_j(1) \quad \text{para } j = 1, 2, \dots, p$$

y, además, $\rho_{p+2}(1) \neq (p+1)\sigma_{p+1}(1)$.

14. Hallar todos los métodos lineales de cuatro pasos óptimos.

15. Hallar para qué valores de α el método lineal multipaso

$$y_{n+3} + \alpha(y_{n+2} - y_{n+1}) - y_n = \frac{1}{2}(3 + \alpha)h(f_{n+2} + f_{n+1})$$

es estable. Demostrar que

- (a) existe un valor de α por el cual el método es de orden 4, pero, para que el método sea estable el orden no puede ser superior a 2.
- (b) para los valores de α en que el método es estable no hay intervalo de estabilidad absoluta.

16. Considerar el método

$$y_{n+2} - (1+a)y_{n+1} + ay_n = \frac{h}{12} [(5+a)f_{n+2} + 8(1-a)f_{n+1} - (1+5a)f_n]$$

para $a \in [-1, 1)$.

- (a) Demostrar que el intervalo de estabilidad absoluta es $(6(a+1)/(a-1), 0)$.
- (b) Ilustrar el caso $a = -0.9$, calculando la solución del problema de condiciones iniciales $y' = -20y$ con $y(0) = 1$.

17. Hacer el mismo análisis realizado en el ejercicio anterior, pero para el método

$$y_{n+2} - (1+a)y_{n+1} + ay_n = \frac{1}{2}h[(3-a)f_{n+1} - (1+a)f_n] \quad -1 \leq a < 1$$

18. Calcular el intervalo de estabilidad absoluta de los métodos

- (a) Adams-Bashforth $y_{n+2} - y_{n+1} = \frac{1}{2}h(3f_{n+1} - f_n)$

(b) Adams-Moulton $y_{n+3} - y_{n+2} = \frac{1}{24}h(9f_{n+3} + 19f_{n+2} - 5f_{n+1} + f_n)$

19. Considerar el predictor P , y dos correctores $C^{(1)}, C^{(2)}$, definidos por los polinomios:

$$\begin{aligned} P : \quad \rho^*(\zeta) &= \zeta^4 - 1 & \sigma^*(\zeta) &= \frac{4}{3}(2\zeta^3 - \zeta^2 + 2\zeta) \\ C^{(1)} : \quad \rho_1(\zeta) &= \zeta^2 - 1 & \sigma_1(\zeta) &= \frac{1}{3}(\zeta^2 + 4\zeta + 1) \\ C^{(2)} : \quad \rho_2(\zeta) &= \zeta^3 - \frac{9}{8}\zeta^2 + \frac{1}{8} & \sigma_2(\zeta) &= \frac{3}{8}(\zeta^3 + 2\zeta^2 - \zeta) \end{aligned}$$

Estudiar una estimación del error de los métodos predictor-corrector utilizando el predictor P y cada uno de los correctores $C^{(1)}, C^{(2)}$.

20. Demostrar que el método

$$\begin{aligned} y_{n+1} - y_n &= \frac{h}{10}(k_1 + 5k_2 + 4k_3) \\ k_1 &= f(x_n, y_n) \\ k_2 &= f(x_n + \frac{1}{3}h, y_n + \frac{1}{3}hk_1) \\ k_3 &= f(x_n + \frac{5}{6}h, y_n - \frac{5}{12}hk_1 + \frac{5}{4}hk_2) \end{aligned}$$

tiene orden igual a tres. Hallar una cota del error local de truncamiento. Demostrar que, si $f(x, y)$ es independiente de y , entonces este método se reduce a una fórmula de integración (cuadratura) de orden cuatro.

21. Calcular el intervalo de estabilidad absoluta del método

$$y_{n+2} - y_n = \frac{h}{2}(f_{n+1} + 3f_n)$$

22. Considerar el problema de condiciones frontera

$$y'' = \lambda \sinh \lambda y, \quad \text{con } y(0) = 0, \quad y(1) = 1, \quad (\lambda \text{ fijo}).$$

- (a) Aplicar el método del tiro simple para resolverlo, con $\lambda = 5$.
- (b) Aplicar el método del tiro paralelo, escogiendo como a trayectoria inicial $\eta(x) = x$.
- (c) Aplicar el método del tiro paralelo, escogiendo como a trayectoria inicial la solución del problema linealizado.
- (d) Aplicar un método en diferencias.

Comparar y estudiar los resultados obtenidos.

23. Considerar el problema de condiciones frontera

$$-y'' + 400y = -400 \cos^2 \pi x - 2\pi^2 \cos 2\pi x \quad \text{con } y(0) = y(1) = 0$$

Sabemos que su solución exacta es

$$y(x) = \frac{e^{-20}}{1 + e^{-20}} e^{20x} + \frac{1}{1 + e^{-20}} e^{-20x} - \cos^2 \pi x$$

Resolverlo aplicando:

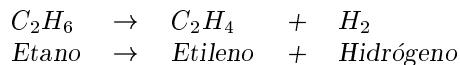
- (a) el método del tiro simple
- (b) el método del tiro paralelo

y comparar los resultados obtenidos con el resultado exacto.

8.6 Prácticas

8.6.1 Práctica Ejemplo

La pirólisis del etano en el intervalo de temperaturas 920 a 1200°K se puede representar, en esencia, por la reacción química irreversible de primer orden siguiente:



En un tubo de acero de diámetro interior 10.226 cm se introduce un caudal de etano puro de 816 kg/h a una temperatura de 920°K. El tubo está situado dentro de un horno que le suministra 1350 cal/h cm² (de superficie interior del tubo). Suponemos que el tubo no tiene obstrucciones internas (p. e. catalíticas) y que se pueden despreciar las diferencias de presión a lo largo del eje; la presión media de los gases en el interior se pueden tomar como 20.68 N/cm². Se quiere calcular qué longitud debe tener el tubo del reactor para obtener una descomposición del 75% del etano en etileno y hidrógeno. El programa ha de dar las temperaturas y los perfiles de conversión a lo largo del tubo.

Datos de la reacción de pirólisis del etano

	ΔH_f a 298°K (cal/mol)	c_p (cal/mol °K)
C_2H_6 (gas)	-20236	$3.75 + 35.7 \times 10^{-3}T - 10.12 \times 10^{-6}T^2$
C_2H_4 (gas)	12496	$5.25 + 24.2 \times 10^{-3}T - 6.88 \times 10^{-6}T^2$
H_2 (gas)	0	$7.00 - 0.385 \times 10^{-3}T + 0.6 \times 10^{-6}T^2$

$$\text{Constante de la reacción: } k = 2.075 \times 10^{20} e^{-41310/T} h^{-1}$$

$$\text{Pesos atómicos: } C = 12, H = 1.$$

$$\text{Constante de los gases: } R = 831.31 N\text{cm}/mol^\circ K$$

Descripción de las variables

- A Sección del tubo, en cm²
- c Concentración de etano, en mol/cm³
- c_p Capacidad calorífica, en cal/mol °K
- di Diámetro interior del tubo, en cm
- k Constante de la reacción, en h⁻¹
- L Longitud media desde la entrada del tubo, en cm
- n_0 Caudal molar de entrada de etano, en mol/h
- $n_{C_2H_4}$ Caudal molar de etileno en un punto, en mol/h
- $n_{C_2H_6}$ Caudal molar de etano en un punto, en mol/h

n_{H_2}	Caudal molar de hidrógeno en un punto, en mol/h
P	Presión total, en N/cm^2
q	Caudal de calor suministrado por el horno por unidad de longitud del tubo, en $cal/h\ cm$
qe_0	Caudal de entrada de etano, en kg/h
r	Velocidad específica de la reacción, en $mol/cm^3\ h$
T	Temperatura absoluta, en $^{\circ}K$
V	Volumen del reactor, en cm^3
x	Fracción molar de etano
z	Fracción de etano convertido en etileno y hidrógeno
ΔH_f	Calor de formación, en cal/mol
ΔH_R	Calor de reacción, en cal/mol

Planteamiento del problema

Continuidad de la masa. Considerando la continuidad del etano en un elemento de volumen $dV = A dL$, podemos afirmar que el etano que sale es el que entra más el que reacciona en este elemento dV , es decir

$$n_{C_2H_6} + rdV = n_{C_2H_6} + dn_{C_2H_6}$$

por lo tanto,

$$r = \frac{dn_{C_2H_6}}{dV} = \frac{d[n_0(1-z)]}{AdL} = -\frac{n_0}{A} \frac{dz}{dL}$$

Por otro lado, para la reacción irreversible en cuestión, y considerando que debido a la baja presión es válida la ley de los gases perfectos, la velocidad de reacción se puede expresar como

$$r = -kc = -k \frac{xP}{RT} = -k \frac{P}{RT} \frac{1-z}{1+z}$$

Sustituyendo este valor de r en la ecuación anterior, obtenemos la ecuación diferencial

$$\frac{dz}{dL} = k \frac{AP}{n_0 RT} \frac{1-z}{1+z}$$

Para unos parámetros dados, la ecuación solamente depende de T y de z , es decir, la temperatura y la fracción de etano convertido en etileno y hidrógeno.

Continuidad de la energía. En un elemento de volumen dV , el calor producida por la reacción ($n_0 dz(-\Delta H_R)$) más la introducida por el horno ($q dL$), han de ser iguales al aumento de la entalpía de la corriente de gases:

$$n_0 dz(-\Delta H_R) + q dL = n_0 [(1-z)c_{pC_2H_6} + z(c_{pC_2H_4} + c_{pH_2})]dT$$

de donde

$$\frac{dT}{dL} = \frac{\frac{q}{n_0} + (-\Delta H_R) \frac{dz}{dL}}{(1-z)c_{pC_2H_6} + z(c_{pC_2H_4} + c_{pH_2})}$$

El calor de reacción, ΔH_R , varía con la temperatura de acuerdo con la ley

$$\frac{d(\Delta H_R)}{dT} = c_{pC_2H_4} + c_{pH_2} - c_{pC_2H_6}$$

Primero es necesario calcular el calor de reacción a $298^{\circ}K$ a partir de los calores de formación de los distintos componentes.

$$\Delta H_R(298) = 20236 \text{ cal/mol} + 12496 \text{ cal/mol} = 32732 \text{ cal/mol}$$

Seguidamente, el calor de reacción a una temperatura T se obtiene integrando la ecuación anterior

$$\Delta H_R(T) = \Delta H_R(298) + \int_{298}^T \frac{d(\Delta H_R)}{dT}$$

$$\Delta H_R(T) = 32732 + 1.28 \times 10^{-6}(T^3 - 298^3) - 5.9425 \times 10^{-3}(T^2 - 298^2) + 8.5 \times (T - 298)$$

Así pues, nos queda el sistema de ecuaciones diferenciales de primer orden

$$\begin{aligned}\dot{z} &= k \frac{AP}{n_0 RT} \frac{1-z}{1+z} \\ \dot{T} &= \frac{\frac{q}{n_0} + (-\Delta H_R)\dot{z}}{(1-z)c_{pC_2H_6} + z(c_{pC_2H_4} + c_{pH_2})}\end{aligned}$$

donde las derivadas son respecto al parámetro longitud del tubo L . Las condiciones iniciales son $z(0) = 0$ y $T(0) = 920$ y integraremos hasta que $z = 0.75$.

Programa utilizado para los cálculos

Para integrar numéricamente la ecuación diferencial se utiliza un método Runge-Kutta-Fehlberg de órdenes 7 y 8. Una primera versión de la rutina RK78 nos fue facilitada por Carles Simó.

```

PROGRAM PIROLISIS
C
EXTERNAL REACTOR
DOUBLE PRECISION H,HMIN,HMAX,T,Y(2),YP(2),TOUT,RELERR,
1 TFINAL,TPRINT,W1(13,2),W2(2),W3(2),DI,A,QPERSUP,Q,QEO,
1 NO,TINI,P,R,PI,PERC
INTEGER IFLAG, NINT, NEQN, NP
LOGICAL STEP
COMMON /CONST/ R,/GLOB/ A,P,NO,Q
C   DI :      DIAMETRO INTERIOR DEL TUBO REACTOR
C   A :      AREA INTERIOR DE LA SECCION TRANSVERSAL DEL TUBO
C   QPERSUP : CAUDAL DE CALOR QUE APORTA EL HORNO POR UNIDAD
C             DE SUPERFICIE INTERIOR
C   Q :      CAUDAL DE CALOR POR UNIDAD DE LONGITUD
C   QEO :      CAUDAL DE ETANO A LA ENTRADA
C   NO :      CAUDAL MOLAR D ETANO A LA ENTRADA
C   TINI :     TEMPERATURA INICIAL DEL ETANO
C   P :      PRESION MEDIA DE LOS GASES A L'INTERIOR DEL TUBO
C   R :      CONSTANTE DE LOS GASES
C
READ (0,101) STEP
READ (0,151) H,HMIN,HMAX,RELERR,TPRINT
READ (0,151) DI,QPERSUP,QEO,TINI,P
READ (0,111) PERC
C
C   INICIALIZACION

```

```

NEQN = 2
T = 0.D0
TOUT = T
IFLAG = 1
PI = 4.D0*Datan(1.D0)
R = 831.31D0
NO = QEO/30.D-03
A = .25D0*PI*DI*DI
Q = QPERSUP*PI*DI
C
C      CALCULO CONDICIONES INICIALES
Y(1) = 0.D0
Y(2) = TINI
WRITE(1,81)
WRITE(1,21)
10 CALL RKF78(reactor,NEQN,H,Y,T,TOUT,RELERR,HMIN,HMAX,
  1 IFLAG,NP,NINT,W1,W2,W3,STEP)
  WRITE(1,11) T, Y(1), Y(2)
  GO TO (50,20,30,40,50,60), IFLAG
20 TOUT = T + TPRINT
  IF (Y(1) .LT. PERC) GO TO 10
  STOP
30 WRITE(1,31) RELERR
  STOP
40 WRITE(1,41)
  WRITE(1,42) NINT
  GO TO 20
50 WRITE(1,51)
  STOP
60 WRITE(1,61) RELERR
  GO TO 20

11 FORMAT(3F12.5)
21 FORMAT(11X,'L',11X,'Z',11X,'T',/)
31 FORMAT(51H SUPERADO EL NUMERO MAXIMO DE EVALUACIONES DE LA FUNCION )
41 FORMAT(43H OJO!! ES PROBABLE QUE EL ERROR COMETIDO EN )
42 FORMAT(25H EL INTERVALO SUPERE RELERR , 15)
51 FORMAT(43H PARAMETROS INCORRECTOS EN LA LLAMADA A RKF78)
61 FORMAT(' OJO!! RELERR MODIFICADO', E12.3)
81 FORMAT(' PIROLISIS ETANO ',/)

101 FORMAT(1L6)
111 FORMAT(1D13.5)
121 FORMAT(2D13.5)
131 FORMAT(3D13.5)
141 FORMAT(4D13.5)
151 FORMAT(5D13.5)
161 FORMAT(1I2)
  END

C      SUBRUTINA REACTOR C
SUBROUTINE reactor (T, Y, YP)
DOUBLE PRECISION T, Y(2), YP(2),A,P,NO,R,Q,K,DHR,CPC2H6,CPC2H4,
  1 CPH2,Z,T1,T2,T3
COMMON /CONST/ R,/GLOB/ A,P,NO,Q
C
Z = Y(1)
T1 = Y(2)
T2 = T1*Y(2)

```

```

T3 = T2*Y(2)
K = 2.075D20*DEXP(-4.1310D04/T1)
DHR = 3.2732D04 + 1.28D-06*(T3-298.D0**3.)
1 - 5.9425D-03*(T2-298.D0**2.) + 8.5D0*(T1-298.)
CPC2H6 = 3.75D0 + 35.7D-03*T1 - 10.12D-06*T2
CPC2H4 = 5.25D0 + 24.2D-03*T1 - 6.88D-06*T2
CPH2 = 7.00D0 - .385D-03*T1 + .60D-06*T2
YP(1) = K*A*P*(1.D0-Z)/(NO*R*T1*(1.D0+Z))
YP(2) = (Q/NO - DHR*YP(1))/((1.D0-Z)*CPC2H6 + Z*(CPC2H4 + CPH2))
RETURN
END

C SUBRUTINA RKF78
C
C SUBROUTINE RKF78(F,NEQN,H,Y,X,XOUT,RELERR,HMIN,HMAX,IFLAG,NPASSOS,
1 NINT,W1,W2,W3,STEP)
C
C EFECTUA LAS LLAMADAS A RK78 PARA RESOLVER LA ECUACION
C DIFERENCIAL EN EL INTERVALO [X,XOUT]
C ACTUALIZA IFLAG I NINT PARA SEÑALAR POSIBLES ANOMALIAS

C PARAMETROS DE LA SUBRUTINA:
C F SUBRUTINA DONDE SE DEFINE EL SISTEMA DE ECUACIONES
C F(X,Y1,YP)
C X = VARIABLE INDEPENDIENTE
C Y1 = VECTOR DE DIMENSION N DE ENTRADA
C YP = VECTOR DE DIMENSION N DE SALIDA
C NEQN NUMERO DE ECUACIONES DEL SISTEMA
C H INCREMENTO INICIAL/FINAL DE LA VARIABLE INDEPENDIENTE
C Y(*) VECTOR DE DIMENSION N, SOLUCION DE LAS ECUACIONES PARA X
C X VARIABLE INDEPENDIENTE
C XOUT PUNTO DONDE SE DESEA LA SOLUCION
C RELERR ERROR RELATIVO MAXIMO
C HMIN INCREMENTO MINIMO DE LA VARIABLE INDEPENDIENTE
C HMAX INCREMENTO MAXIMO DE LA VARIABLE INDEPENDIENTE
C IFLAG = 1 INDICA A RKF78 QUE ES LA PRIMERA LLAMADA
C = 2 NO HA HABIDO PROBLEMAS EN EL INTERVALO
C = 3 SUPERADO EL NUMERO MAXIMO DE EVALUACIONES DE LA
C FUNCION PERMITIDAS PARA UN INTERVALO
C = 4 EN ALGUNOS DE LOS SUBINTERVALOS EL ERROR COMETIDO PUEDE
C SUPERAR AL ERROR RELATIVO MAXIMO DESEADO
C = 5 PARAMETROS INCORRECTOS
C = 6 MODIFICADO EL ERROR RELATIVO
C NPASSOS INDICA EL NUMERO DE PASOS EN QUE SE HA SUBDIVIDIDO
C EL INTERVALO [X,XOUT]
C NINT EN EL CASO DE IFLAG = 4 INDICA EL NUMERO DE INTERVALOS EN
C QUE HA HABIDO PROBLEMAS
C W1,W2,W3 VECTORES DE TRABAJO DE DIMENSIONES 13xNEQN, NEQN Y NEQN
C RESPECTIVAMENTE
C STEP INDICA SI SE QUIERE EL RESULTADO A CADA PASO DE INTEGRACION
C
C DOUBLE PRECISION Y(NEQN),X,XOUT,DX,RELERR,REMIN,H,HMIN,HMAX,
1 W1(13,NEQN),W2(NEQN),W3(NEQN),EPS,EPSP1,NORMY,NEXTH
INTEGER NEQN,IFLAG,NFE,MAXNFE,NPASSOS,NINT
LOGICAL STEP
DATA MAXNFE/3000/
C
IF (IFLAG .EQ. 1) THEN

```

```

C      PRIMERA LLAMADA, CALCULA EL EPSILON DE LA MAQUINA
EPS = 1.0
10 EPS = EPS/2.0
EPS1 = EPS + 1.0
IF (EPSP1 .GT. 1.0) GO TO 10
NORMY = 0.
DO 20 I=1,NEQN
  NORMY = DABS(Y(I))+NORMY
20 CONTINUE
REMIN = EPS * (DMAX1(100.0*NORMY,1.0))
ENDIF
C      INICIALIZACION
NPASSOS = 0
NFE=0
NINT = 0
NEXTH = H
C
C      CONTROL PARAMETROS DE ENTRADA
IF (NEQN .LT. 1) GO TO 70
IF (RELERR .LT. REMIN) GO TO 80
IF (STEP) THEN
  DX=XOUT-X
  H = DMIN1(DX,H)
ENDIF
30 CALL RK78(X,Y,F,NEQN,H,HMIN,HMAX,RELERR,LOCERR,W1,W2,W3,EP
S,IFLAG,NFE)
IF (IFLAG.EQ.1) IFLAG = 2
NPASSOS = NPASSOS+1
IF (STEP) THEN
  IF (X .EQ. XOUT) THEN
    GO TO 40
  ELSE
    NEXTH = H
    DX = XOUT-X
    H = DMIN1(DX,H)
    GOTO 30
  ENDIF
ENDIF
IF (H.EQ.HMIN) GO TO 60
IF (NFE.GE.MAXNFE) GO TO 50
GO TO 40
C
C      X = XOUT
40 IF (STEP) H = NEXTH
RETURN
C
C      SUPERADO EL NUMERO MAXIMO DE EVALUACIONES DE LA FUNCION
50 IFLAG=3
RETURN
C
C      NO SE PUEDE LLEGAR A LA PRECISION PEDIDA CON EL PASO
C      MINIMO DADO
60 IFLAG=4
NINT = NINT+1
GO TO 30
C
C      PARAMETROS DE ENTRADA INCORRECTOS
70 IFLAG=5
RETURN

```

```

C      MODIFICACION DEL ERROR RELATIVO
80 IFLAG=6
      RELERR=REMIN
      RETURN
C
      END

C      SUBRUTINA RK78
C
      SUBROUTINE RK78(X,Y,DERIV,N,H,HMI,HMA,E1,E2,R,B,F,EPS,IFLAG,NFE)
C
C      RUNGE-KUTTA FELHBERG ORDEN 7-8 CON DETERMINACION AUTOMATICA
C      DEL PASO
C
C
C      ABSTRACT
C      DADO UN SISTEMA DE N ECUACIONES DIFERENCIALES ORDINARIAS DE PRIMER
C      ORDEN DE LA FORMA
C
C      DY(I)/DX = DERIV (X,Y(1),Y(2),...,Y(N))
C
C      Y EL VALOR DE Y(I) EN EL PUNTO X, EFECTUA UN PASO DE INTEGRACION Y
C      SE OBTIENE UNA ESTIMACION DE Y(I) EN EL PUNTO X+H Y UN AJUSTE AUTOMATICO
C      DEL PASO DE INTEGRACION H
C
C      PARAMETROS DE LA SUBRUTINA:
C
C      X      VARIABLE INDEPENDIENTE
C      Y(*)    VECTOR DE DIMENSION N, SOLUCION DE LAS ECUACIONES PARA X
C      DERIV   SUBRUTINA DONDE SE DEFINE EL SISTEMA DE ECUACIONES DIFERENCIALES
C              DERIV(A,B,F)
C                      A = VARIABLE INDEPENDIENTE
C                      B = VECTOR DE DIMENSION N DE ENTRADA
C                      F = VECTOR DE DIMENSION N DE SALIDA
C      N      NUMERO D'ECUACIONES DEL SISTEMA
C      H      INCREMENTO INICIAL/FINAL DE LA VARIABLE INDEPENDIENTE
C      HMI    INCREMENTO MINIMO DE LA VARIABLE INDEPENDIENTE
C      HMA    INCREMENTO MAXIMO DE LA VARIABLE INDEPENDIENTE
C      E1     ERROR RELATIVO MAXIMO
C      E2     ESTIMACION DEL ERROR LOCAL COMETIDO EN EL PASO DE INTEGRACION
C      R      MATRIZ DE 13xN DONDE SE GUARDAN LAS 13 EVALUACIONES DEL
C              SISTEMA DE ECUACIONES REALIZADAS PARA CADA PASO DE INTEGRACION
C      B      VECTOR DE ENTRADA A LA SUBRUTINA DERIV
C      F      VECTOR DE SALIDA DE LA SUBRUTINA DERIV
C      EPS    EPSILON DE LA MAQUINA
C      IFLAG   = 1    INDICA A RKF78 QUE ES LA PRIMERA LLAMADA
C                  = 2    NO HA HABIDO PROBLEMAS EN L'INTERVALO
C                  = 3    SUPERADO EL NUMERO MAXIMO DE EVALUACIONES DE LA
C                          FUNCION PERMITIDAS PARA UN INTERVALO
C                  = 4    EN ALGUNOS DE LOS SUBINTERVALOS EL ERROR COMETIDO PUEDE
C                          SUPERAR AL ERROR RELATIVO MAXIMO DESEADO
C                  = 5    PARAMETROS INCORRECTOS
C                  = 6    MODIFICADO EL ERROR RELATIVO
C      NFE     INDICA EL NUMERO DE EVALUACIONES DE LA FUNCION DERIV
C
C      DOUBLE PRECISION X,Y(N),HMI,HMA,E1,E2,H,H1,R(13,N),B(N),F(N),A,
1 BET,ALFA(13),BETA(79),C(11),CP(13),E3,D,DD,FACT,EPS

```

```

INTEGER IFLAG
C
C      IF (IFLAG.EQ.1) THEN
C      INICIALIZACION COEFICIENTES ALFA
      ALFA(1)=0.D0
      ALFA(2)=2.D0/27.D0
      ALFA(3)=1.D0/9.D0
      ALFA(4)=1.D0/6.D0
      ALFA(5)=5.D0/12.D0
      ALFA(6)=.5D0
      ALFA(7)=5.D0/6.D0
      ALFA(8)=1.D0/6.D0
      ALFA(9)=2.D0/3.D0
      ALFA(10)=1.D0/3.D0
      ALFA(11)=1.D0
      ALFA(12)=0.D0
      ALFA(13)=1.D0
C      INICIALIZACION COEFICIENTES BETA
      BETA(1)=0.D0
      BETA(2)=2.D0/27.D0
      BETA(3)=1.D0/36.D0
      BETA(4)=1.D0/12.D0
      BETA(5)=1.D0/24.D0
      BETA(6)=0.D0
      BETA(7)=1.D0/8.D0
      BETA(8)=5.D0/12.D0
      BETA(9)=0.D0
      BETA(10)=-25.D0/16.D0
      BETA(11)=-BETA(10)
      BETA(12)=.5D-1
      BETA(13)=0.D0
      BETA(14)=0.D0
      BETA(15)=.25D0
      BETA(16)=.2D0
      BETA(17)=-25.D0/108.D0
      BETA(18)=0.D0
      BETA(19)=0.D0
      BETA(20)=125.D0/108.D0
      BETA(21)=-65.D0/27.D0
      BETA(22)=2.D0*BETA(20)
      BETA(23)=31.D0/300.D0
      BETA(24)=0.D0
      BETA(25)=0.D0
      BETA(26)=0.D0
      BETA(27)=61.D0/225.D0
      BETA(28)=-2.D0/9.D0
      BETA(29)=13.D0/900.D0
      BETA(30)=2.D0
      BETA(31)=0.D0
      BETA(32)=0.D0
      BETA(33)=-53.D0/6.D0
      BETA(34)=704.D0/45.D0
      BETA(35)=-107.D0/9.D0
      BETA(36)=67.D0/90.D0
      BETA(37)=3.D0
      BETA(38)=-91.D0/108.D0
      BETA(39)=0.D0
      BETA(40)=0.D0
      BETA(41)=23.D0/108.D0

```

```
BETA(42)=-976.D0/135.D0
BETA(43)=311.D0/54.D0
BETA(44)=-19.D0/60.D0
BETA(45)=17.D0/6.D0
BETA(46)=-1.D0/12.D0
BETA(47)=2383.D0/4100.D0
BETA(48)=0.D0
BETA(49)=0.D0
BETA(50)=-341.D0/164.D0
BETA(51)=4496.D0/1025.D0
BETA(52)=-301.D0/82.D0
BETA(53)=2133.D0/4100.D0
BETA(54)=45.D0/82.D0
BETA(55)=45.D0/164.D0
BETA(56)=18.D0/41.D0
BETA(57)=3.D0/205.D0
BETA(58)=0.D0
BETA(59)=0.D0
BETA(60)=0.D0
BETA(61)=0.D0
BETA(62)=-6.D0/41.D0
BETA(63)=-3.D0/205.D0
BETA(64)=-3.D0/41.D0
BETA(65)=-BETA(64)
BETA(66)=-BETA(62)
BETA(67)=0.D0
BETA(68)=-1777.D0/4100.D0
BETA(69)=0.D0
BETA(70)=0.D0
BETA(71)=BETA(50)
BETA(72)=BETA(51)
BETA(73)=-289.D0/82.D0
BETA(74)=2193.D0/4100.D0
BETA(75)=51.D0/82.D0
BETA(76)=33.D0/164.D0
BETA(77)=12.D0/41.D0
BETA(78)=0.D0
BETA(79)=1.D0
C      INICIALIZACION COEFICIENTES GAMMA RK ORDEN 7
C(1)=41.D0/840.D0
C(2)=0.D0
C(3)=0.D0
C(4)=0.D0
C(5)=0.D0
C(6)=34.D0/105.D0
C(7)=9.D0/35.D0
C(8)=C(7)
C(9)=9.D0/280.D0
C(10)=C(9)
C(11)=C(1)
C      INICIALIZACION COEFICIENTES GAMMA RK ORDEN 8
CP(1)=0.D0
CP(2)=0.D0
CP(3)=0.D0
CP(4)=0.D0
CP(5)=0.D0
CP(6)=C(6)
CP(7)=C(7)
CP(8)=C(8)
```

```

CP(9)=C(9)
CP(10)=C(10)
CP(11)=0.D0
CP(12)=C(1)
CP(13)=C(1)
GO TO 50
ENDIF
C
C      SE EFECTUAN LAS 13 EVALUACIONES DE DERIV Y SE GUARDAN LOS RESULTADOS
C      A LA MATRIZ R
5 JK=1
DO 25 J=1,13
DO 10 L=1,N
B(L)=Y(L)
10  CONTINUE
A=X+ALFA(J)*H
IF(J.EQ.1)GO TO 20
J1=J-1
DO 15 K=1,J1,1
JK=JK+1
BET=BETA(JK)*H
DO 16 L=1,N
B(L)=B(L)+BET*R(K,L)
16  CONTINUE
15  CONTINUE
20  CALL DERIV (A,B,F)
DO 26 L=1,N
R(J,L)=F(L)
26  CONTINUE
25 CONTINUE
NFE=NFE+15
C
C      CALCULO DE LAS ESTIMACIONES DE Y(*) EN EL PUNTO X+H OBTENIDAS
C      CON LAS FORMULES DE ORDEN 7 I DE ORDEN 8
C
C      B(*) <-- ESTIMACION A Y(X+H) ORDEN 7
C      F(*) <-- ESTIMACION A Y(X+H) ORDEN 8
C      D    <-- NORM1(F(*)-B(*))
C      ESTIMACION DEL ERROR LOCAL COMETIDO
C      DD    <-- NORM1(F(*))
C
D=0
DD=0
DO 35 L=1,N
B(L)=Y(L)
F(L)=Y(L)
DO 30 K=1,11
BET=H*R(K,L)
B(L)=B(L)+BET*C(K)
F(L)=F(L)+BET*CP(K)
30  CONTINUE
F(L)=F(L)+H*(CP(12)*R(12,L)+CP(13)*R(13,L))
D=D+DABS(F(L)-B(L))
DD=DD+DABS(F(L))
35 CONTINUE
C
C      CONTROL DEL ERROR
FACT = 1+DD*1.D-2
E3=E1*FACT

```

```

C
C      TERMINA SI EL ERROR NO SUPERA EL MAXIMO PERMITIDO O YA NO SE PUEDE REDUCIR
C      MAS EL PASO DE INTEGRACION (H<=HMIN)
C      IF (DABS(H).LE.HMI.OR.D.LT.E3) GO TO 40
C
C      AJUSTE AUTOMATICO DEL PASO D INTEGRACION Y SE REPITE EL CALCULO
C      CON EL NUEVO PASO
C      H = H*0.9D0*(E3/D)**0.125D0
C      IF (DABS(H).LT.HMI) H=HMI
C      IF (DABS(H).LT.DABS(EPS*X)) THEN
C          H1 = DSIGN(1.D0,H)
C          H = H1*EPS*X
C      ENDIF
C      GO TO 5
C
C      SALIDA DE LA SUBRUTINA
C      ACTUALIZACION DE X,Y I PREDICCION DEL PASO DE INTEGRACION PARA
C          LA PROXIMA LLAMADA
C      SE DEVUELVE A E2 LA ESTIMACION DEL ERROR LOCAL COMETIDO EN
C          EL PASO DE INTEGRACION
C
40 X = X+H
    IF(D.EQ.0.) D=E3/256
    H = H*0.9D0*(E3/D)**0.125D0
    IF (DABS(H).GT.HMA) H=HMA*H/DABS(H)
    DO 45 L=1,N
        Y(L)=F(L)
45 CONTINUE
    E2=D
50 RETURN
END

```

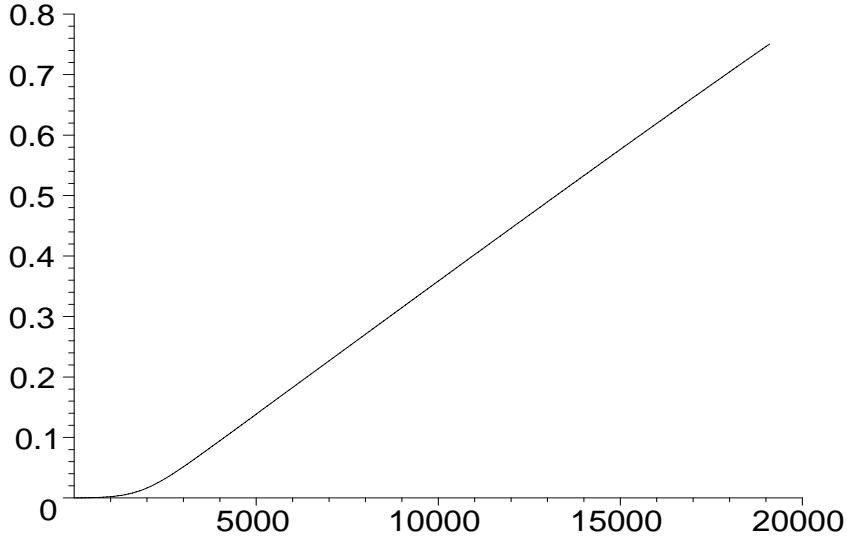


Fig. 8.7 Longitud del tubo - % de descomposición del etano.

Resultados obtenidos

Con los datos de entrada siguientes:

```
STEP = .TRUE.      H = 1.00000D+01      HMIN = 1.00000D+00
HMAX = 1.00000D+03  RELERR = 1.00000D-12  TPRINT = 2.00000D+02
DI = 1.022600D+01  QPERSUP = 1.35000D+03  QEO = 8.16000D+02
TINI = 9.20000D+02  P = 2.068000D+01      PERC = 7.50000D-01
```

Los resultados obtenidos se pueden observar en las figuras 8.7 y 8.8.

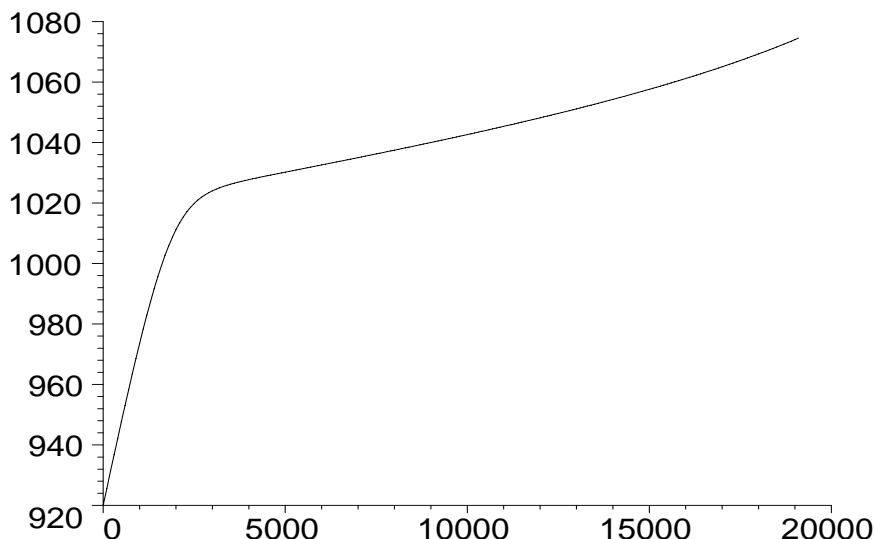


Fig. 8.8 Longitud del tubo – temperatura en el interior del tubo.

8.6.2 Enunciados

1. La función error se acostumbra a definir mediante una integral,

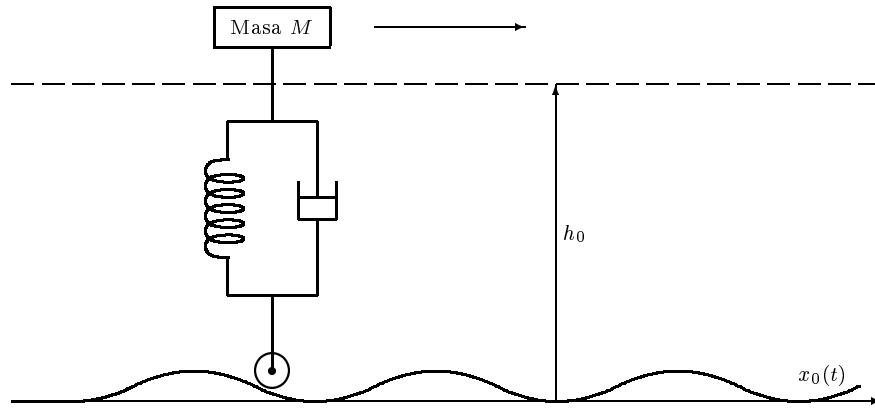
$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

pero también se puede definir como solución de la ecuación diferencial

$$\dot{y}(x) = \frac{2}{\sqrt{\pi}} e^{-x^2} \quad y(0) = 0$$

¿Por qué? Escribir un programa utilizando RKF78 que genere una tabla de $\text{erf}(x)$ para $x = 0.0, 0.1, 0.2, \dots, 1.9, 2.0$. Comparar los valores de vuestra tabla con los obtenidos, por ejemplo, con la rutina S15AEF de la librería NAG.

2. Un vehículo de masa M con un sistema de suspensión formado por un muelle y un amortiguador como los de la figura se mueve con velocidad horizontal constante. En el instante $t = 0$ el vehículo se desplaza con el centro de gravedad a una distancia h_0 del nivel de tierra y sin velocidad vertical. Para $t > 0$ el desplazamiento vertical del suelo por encima del nivel de referencia viene dado por una función $x_0(t)$.



Suponer que la respuesta del muelle es lineal con una constante de proporcionalidad k y que el coeficiente de amortiguamiento del amortiguador r es una función no lineal de las velocidades relativas de los dos extremos del amortiguador:

$$r = r_0 \left(1 + c \left| \frac{dx}{dt} - \frac{dx_0}{dt} \right| \right)$$

Se puede comprobar fácilmente que el desplazamiento del centro de gravedad del vehículo $x(t)$ es la solución de la ecuación diferencial ordinaria de segundo orden

$$M \frac{d^2x}{dt^2} = -k(x - x_0) - r \left(\frac{dx}{dt} - \frac{dx_0}{dt} \right)$$

con las condiciones iniciales

$$x(0) = 0 \quad \left. \frac{dx}{dt} \right|_{t=0} = 0$$

Escribir un programa en Fortran para calcular $x(t)$, $0 \leq t \leq t_{max}$, para un contorno de carretera dado por

$$x_0(t) = A(1 - \cos \omega t)$$

donde $2A$ es el desplazamiento máximo de la carretera sobre el nivel de referencia.

- (a) Calcular el desplazamiento vertical del vehículo sin amortiguador durante un intervalo de 14 s .

- (b) Observar que, para el caso lineal ($c = 0$), los amortiguamientos subcrítico, crítico, y supercrítico corresponden a

$$\xi = \frac{r}{2\sqrt{kM}}$$

menor, igual y mayor que la unidad, respectivamente.

- (c) Investigar el comportamiento del sistema para diversos valores de r_0 y c en los intervalos

$$\begin{aligned} r_0 &\in [140, 420] \quad (N \cdot s/cm) \\ c &\in [0, 4] \quad (s/cm) \end{aligned}$$

- (d) Comparar el desplazamiento, la velocidad y la aceleración verticales del vehículo en un caso lineal y en un caso no lineal.

Os sugerimos los datos siguientes:

$$\begin{aligned} A &= 5 \text{ cm}, \quad \omega = 7 \text{ rad/s}, \\ t_{max} &= 2.5 \text{ s}, \quad k = 1120 \text{ N/cm}, \\ M &= 18 \text{ N} \cdot \text{s}^2/\text{cm} \quad (1800 \text{ kg}). \end{aligned}$$

El programa ha de leer los valores de los parámetros M , r_0 , c , k , A , ω , t_{max} y $freq$. Este último se utiliza para controlar la frecuencia de la salida; o sea, obtendremos los valores de t , x_0 , x , dx/dt , d^2x/dt^2 , r , y $\xi(t)$ cada $freq$ segundos.

3. La ecuación fundamental de la teoría de vigas es

$$\frac{d^2y}{dx^2} = -\frac{M}{EI}$$

donde x es la distancia horizontal a lo largo de la viga, y es la deflexión vertical hacia abajo, M es el momento de torsión, E es el módulo de Young, y I es el momento del área (a veces llamado momento de inercia) de la sección transversal según el eje neutro. El momento del área I no tiene por qué ser constante, dado que la forma de la sección transversal de la viga puede cambiar con la largada (dimensión x) de la viga.

Se puede ver fácilmente que

$$\frac{dM}{dx} = V$$

donde V es la fuerza cortante y

$$\frac{dV}{dx} = -w$$

donde $w(x)$ es la carga sobre la viga.

Por lo tanto, diferenciando la ecuación fundamental dos veces y sustituyendo, obtenemos

$$\frac{d^4y}{dx^4} = \frac{-2}{I} \frac{dI}{dx} \frac{d^3y}{dx^3} - \frac{1}{I} \frac{d^2I}{dx^2} \frac{d^2y}{dx^2} + \frac{w}{EI}$$

La mayoría de aplicaciones de esta ecuación dan lugar a problemas de condiciones frontera. A pesar de esto, el problema de condiciones iniciales siguiente también tiene interés práctico.

Considerar concretamente una viga de sección transversal variable que sale de una pared. La largada de la viga es l , y soporta una carga P en el extremo. Se cumple que

$$\begin{aligned} V(x) &= P \\ M(x) &= -P(l-x) \end{aligned}$$

Sean

$$\begin{aligned} I(x) &= 200(1 + 4e^{-6x/l}) \text{ cm}^4, & l &= 250 \text{ cm}, \\ E &= 20 \times 10^6 \text{ N/cm}^2, & P &= 2225 \text{ N}. \end{aligned}$$

- (a) Calcular $y(l)$, suponiendo que la viga no se rompe ni se deforma permanentemente. El eje “fijo” de la viga en $x = 0$ determina $y(0) = 0$ y $\dot{y}(0) = 0$. Las otras dos condiciones iniciales en $x = 0$ se pueden hallar a partir de las ecuaciones por $V(0)$ i $M(0)$.
- (b) Calcular $y(l)$ para una viga de sección constante $I(x) = 300$ y para una viga colocada al revés ($I(x) = 200(1 + 4e^{-6(l-x)/l})$). Compararlas con la anterior.

- 4. Considerar un ecosistema simple consistente en un número determinado de conejos, que disponen de unos recursos de comida infinitos, y en un número de zorros que se comen los conejos. Un modelo matemático clásico debido a Volterra describe este sistema mediante un sistema de dos ecuaciones diferenciales de primer orden no lineales:

$$\begin{aligned} \dot{r} &= 2r - \alpha rf & r(0) &= r_0 \\ \dot{f} &= -f + \alpha rf & f(0) &= f_0 \end{aligned}$$

donde t es el tiempo, $r = r(t)$ es el número de conejos, $f = f(t)$ es el número de zorros, y α una constante positiva. Cuando $\alpha > 0$, los zorros se encuentran con los conejos con una probabilidad proporcional al producto del número de individuos de cada especie. Un encuentro tiene como resultado una disminución en el número de conejos y (por razones no tan obvias) un aumento del número de zorros.

- (a) Investigar el comportamiento de este sistema por $\alpha = 0.01$ y diversos valores de r_0 y f_0 desde 2 o 3 hasta unos cuantos miles. Dibujar gráficamente las soluciones interesantes y hacer una gráfica con r en un eje y f en el otro (no hace falta restringir r y f a valores enteros).
- (b) Calcular la solución para $r_0 = 300$ y $f_0 = 150$. Comprobar que el comportamiento del sistema es periódico con un periodo muy próximo a 5 unidades de tiempo.
- (c) Calcular la solución para $r_0 = 15$ y $f_0 = 22$. Hay que ver que el número de conejos, al cabo de un cierto tiempo, toma valores inferiores a 1. Esto se puede interpretar como que los conejos se extinguen. Hallar condiciones iniciales con $r_0 = f_0$ que causen la extinción de las dos especies.

- (d) Modificar el programa de forma que, cuando $r(t)$ o bien $f(t)$ tomen valores inferiores a 1, la especie se considere extinguida y no pueda ‘resucitar’.
- (e) ¿Es posible que alguno de las componentes de la solución exacta resulte negativa? ¿Es posible que la solución numérica resulte negativa? ¿Qué pasa si lo hace? (En la práctica, la respuesta de las dos últimas preguntas puede depender de los valores que se hayan dado a las tolerancias del error).
- (f) Se han propuesto distintas modificaciones de este modelo simple para intentar reflejar más de cerca qué pasa en la naturaleza. Por ejemplo, se puede evitar que el número de conejos crezca indefinidamente, cambiando la primera ecuación por

$$\dot{r} = 2 \left(1 - \frac{r}{R}\right) r - \alpha r f$$

Ahora bien, incluso en el caso $\alpha = 0$, el número de conejos no podrá superar nunca R . Escoger un valor razonable para R y replantearse otra vez algunas de las preguntas anteriores. En particular, ¿cómo se ve afectada la periodicidad de las soluciones?

- (g) Buscar los puntos de equilibrio del último modelo. La solución, cuando $t \rightarrow \infty$, ¿tiende a algún punto de equilibrio?
5. Un problema famoso de la mecánica no lineal es conocido como el del **péndulo invertido**. El péndulo es una barra rígida de longitud L sostenida en un de los extremos por un perno sin rozamiento. Mediante un motor eléctrico se da al perno un movimiento rápido s hacia arriba y hacia abajo:

$$s = A \operatorname{sen} \omega t$$

Aplicando la segunda ley de Newton, obtenemos la ecuación del movimiento

$$\ddot{\theta} = \frac{3}{2L}(g - A\omega^2 \operatorname{sen} \omega t) \operatorname{sen} \theta$$

donde g es la aceleración de la gravedad. Para valores pequeños de θ , $\operatorname{sen} \theta \simeq \theta$, y esta ecuación se convierte en la ecuación de Mathieu, de la cual se sabe que es estable para determinados valores de A y ω y determinadas condiciones iniciales. Cuando $A = 0$ tenemos la familiar ecuación del péndulo

$$\ddot{\theta} = \frac{3g}{2L} \operatorname{sen} \theta$$

que se puede aproximar linealmente por valores de θ próximos a π .

El aspecto más interesante de este problema es que hay regiones en las cuales la ecuación del movimiento es estable para valores iniciales correspondientes a una configuración invertida, y que se ha comprobado físicamente.

- (a) Escribir un programa en Fortran utilizando RKF78 para calcular el movimiento $\theta(t)$ para diversos valores de L, A, ω y condiciones iniciales $\theta(0)$ y $\dot{\theta}(0)$. Comprobar vuestro programa con los valores $L = 25 \text{ cm}$, $A = 0 \text{ cm}$, $\omega = 0 \text{ rad/s}$, $\theta(0) = 3.1 \text{ rad}$,

y $\dot{\theta}(0) = 0 \text{ rad/s}$, tomando $g = 980.69 \text{ cm/s}^2$. Comparar la solución obtenida con la solución analítica de la versión linealizada. Utilizar la solución analítica para determinar los valores correctos de los parámetros de la llamada a la rutina RKF78. Obtener los valores de θ y $\dot{\theta}$ para dos o tres oscilaciones del péndulo.

- (b) Para el caso de configuración no invertida $A = 0$, calcular θ y $\dot{\theta}$ para diferentes condiciones iniciales y representar el mapa de fases correspondiente.
- (c) Cuando el programa funcione de forma satisfactoria, probar con los casos más interesantes:

L	A	ω	$\theta(0)$	$\dot{\theta}(0)$
25	1.25	5.3	3.10	0
25	25.00	100	3.10	0
25	25.00	100	0.10	0
25	5.00	100	0.10	0
25	1.25	200	0.05	0

Interpretar físicamente las soluciones obtenidas.

6. Las ecuaciones diferenciales siguientes describen el movimiento de un cuerpo en órbita alrededor de dos cuerpos mucho más pesados. Un ejemplo podría ser la cápsula del Apolo en órbita alrededor de la Tierra y de la Luna. El sistema de coordenadas es un poco complicado. Los tres cuerpos determinan un plano en el espacio y unas coordenadas cartesianas bidimensionales en este plano. El origen está en el centro de masas de los dos cuerpos más pesados, el eje x es la línea que atraviesa estos dos cuerpos y la distancia entre ellos se toma como la unidad. Por lo tanto, si μ es la relación entre la masa de la Luna y la de la Tierra, entonces la Luna y la Tierra están situadas en las coordenadas $(1 - \mu, 0)$ y $(-\mu, 0)$ respectivamente, y el sistema de coordenadas se mueve de acuerdo con el movimiento de rotación de la Luna alrededor de la Tierra. El tercer cuerpo, el Apolo, se supone que tiene una masa despreciable en comparación con la de los otros dos, y su posición en función del tiempo es $(x(t), y(t))$. Las ecuaciones se deducen de la ley del movimiento de Newton y de la ley de atracción gravitatoria. Las derivadas de primer orden aparecen debido a la rotación del sistema de coordenadas.

$$\ddot{x} = 2\dot{y} + x - \frac{\mu^*(x + \mu)}{r_1^3} - \frac{\mu(x - \mu^*)}{r_2^3}$$

$$\ddot{y} = -2\dot{x} + y - \frac{\mu^*y}{r_1^3} - \frac{\mu y}{r_2^3}$$

$$\mu = \frac{1}{82.45} \quad \mu^* = 1 - \mu$$

$$r_1 = ((x + \mu)^2 + y^2)^{1/2} \quad r_2 = ((x - \mu^*)^2 + y^2)^{1/2}$$

Un tipo de problemas muy interesante es el estudio de soluciones periódicas. Se sabe que las condiciones iniciales

$$\begin{aligned}x(0) &= 1.2, & \dot{x}(0) &= 0, \\y(0) &= 0, & \dot{y}(0) &= -1.04935751\end{aligned}$$

determinan una solución que es periódica con periodo $T = 6.19216933$. Esto significa que el Apolo empieza su trayectoria más allá de la Luna a una altura de unas 0.2 veces la distancia entre la Tierra y la Luna y a una determinada velocidad inicial. La órbita resultante hace que el Apolo se acerque bastante a la Tierra, se aleje haciendo una gran curva en la cara de la Tierra opuesta a la Luna, vuelva a acercarse a la Tierra y, finalmente, vuelva a su posición y velocidad iniciales.

La constante de Jacobi del sistema vale

$$C = 2\Omega - (\dot{x}^2 + \dot{y}^2)$$

donde

$$\Omega = \frac{1}{2}(\mu r_2^2 + \mu^* r_1^2) + \frac{\mu}{r_2} + \frac{\mu^*}{r_1}$$

y es una medida de la energía total del sistema, que se ha de mantener constante dado que no hay ninguna pérdida ni aportación exterior.

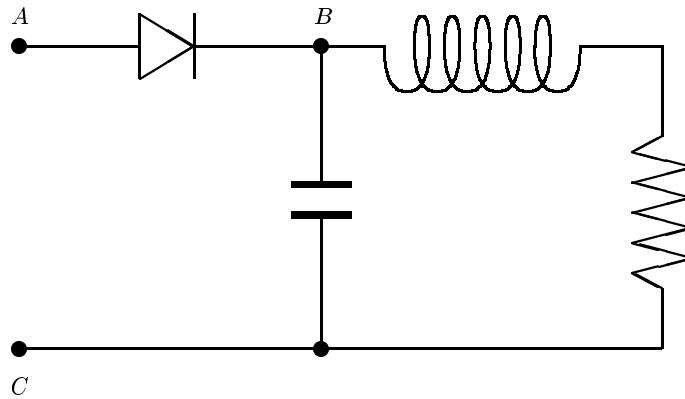
- (a) Utilizar para calcular la solución con las condiciones iniciales dadas. Comprobar que la solución es periódica y que el periodo coincide con el dado.
 - (b) ¿Cuánto se aproxima el Apolo a la superficie de la Tierra en esta órbita? En la ecuación, las distancias se suponen medidas desde los centros de la Tierra y de la Luna. Suponer que la Luna se encuentra a 383,000 km de la Tierra y que la Tierra es una esfera de radio 6400 km. Observar que el centro del sistema de coordenadas cae dentro de esta esfera, pero no en el centro.
 - (c) Calcular la solución en una órbita completa y observar el ajuste automático del paso. Se debe notar que los pasos son pequeños cuando el Apolo está cerca de la Tierra y su órbita está cambiando rápidamente y va aumentando a medida que se aleja en el espacio. Probar de asignar varios valores al error relativo. Hacer un dibujo de la solución que permita observar las variaciones en el paso. Calcular la constante de Jacobi del sistema: las oscilaciones dan idea del error cometido en el cálculo.
 - (d) Estudiar la estabilidad de la órbita modificando ligeramente las condiciones iniciales (por ejemplo, variando $x(0)$ del orden de 1.0E-3) y comprobar si el Apolo tiende a la órbita anterior o no.
 - (e) Calcular la órbita del Apolo en el mismo sistema de coordenadas fijo que teníamos en el instante inicial.
7. Al terminal A del circuito de la figura, se aplica un voltaje periódico de onda rectangular de amplitud V_m y periodo T ; el terminal C está conectado a masa. La corriente que atraviesa el diodo vale

$$\begin{aligned} I_{AB} &= 0, && \text{cuando } V_A < V_B \\ I_{AB} &= k(V_A - V_B)^{3/2} && \text{cuando } V_A \geq V_B \end{aligned}$$

donde k es una constante.

Las leyes de Ohm de los tres componentes del circuito son

$$\begin{aligned} I(t) &= CdV/dt && \text{para el condensador} \\ V(t) &= LdI/dt && \text{para la bobina} \\ V(t) &= I \cdot R && \text{para la resistencia.} \end{aligned}$$



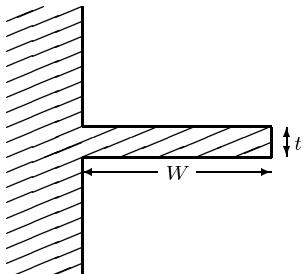
- (a) Plantear el problema aplicando las leyes de Ohm y de Kirchoff.
- (b) Escribir un programa que calcule la variación del voltaje en función del tiempo a través del condensador y de la resistencia. El cálculo ha de terminar cuando t supera un determinado t_{max} , o bien cuando la variación de tensión entre dos ciclos consecutivos es menor que un ϵ predeterminado. Calcular la tensión de rizado en cada uno de los casos. Investigar el comportamiento del sistema para diversos valores de T y k en los intervalos

$$\begin{aligned} T &\in [0.1, 0.001] \quad (s) \\ k &\in [0.003, 0.0003] \quad (A/V^{3/2}) \end{aligned}$$

Os sugerimos los datos siguientes:

$$\begin{aligned} V_m &= 100 \text{ V} & C &= 10 \mu\text{F} \\ L &= 2.5 \text{ H} & R &= 500 \Omega \\ \epsilon &= 0.001 & t_{max} &= 10T \end{aligned}$$

8. La figura siguiente representa la sección transversal de una larga aleta de refrigeración de anchura W , grueso t y conductividad térmica k , unida a una pared caliente, de forma que su base, en el punto $x = 0$, se mantiene a una temperatura constante T_w .



El calor se transmite de forma constante a través de la aleta, y se pierde por las dos caras por convección en el aire que la envuelve, con un coeficiente de transmisión h (la radiación se desprecia a temperaturas suficientemente bajas).

La ley de Fourier expresa la densidad del flujo calorífico por conducción permanente en una dirección como $q = -k dT/dx$.

La temperatura de la aleta T , que suponemos que depende solamente de la distancia x , cumple la ecuación diferencial

$$kt \frac{d^2 T}{dx^2} = 2h(T - T_a)$$

donde T_a es la temperatura media del aire que envuelve la aleta. La temperatura en el extremo de la aleta debe valer T_a .

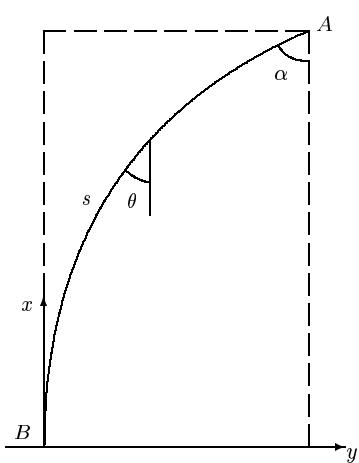
Si la superficie de la aleta es vertical, h viene dado por la relación dimensional $h = 1.192(T - T_a)^{1/3}W/m^2 \cdot ^\circ C$, con T y T_a expresadas en $^\circ C$.

Escribir un programa que calcule la distribución de la temperatura a lo largo de la aleta y el flujo de calor perdido por la aleta por unidad de longitud.

Os sugerimos los datos siguientes: $T_w = 90^\circ C$, $T_a = 21^\circ C$, $t = 0.635 cm$.

Estudiar diversos casos por $k = 45$ (acero) y 380 (cobre) $W/m \cdot ^\circ C$ y W entre 1.25 y 12.5 cm .

9. Un pilar de longitud L , encastado en la base B , tiene un módulo de Young E y un momento de inercia I . Si el pilar es suficientemente largo se torcerá debido a su propio peso, que es w por unidad de longitud.

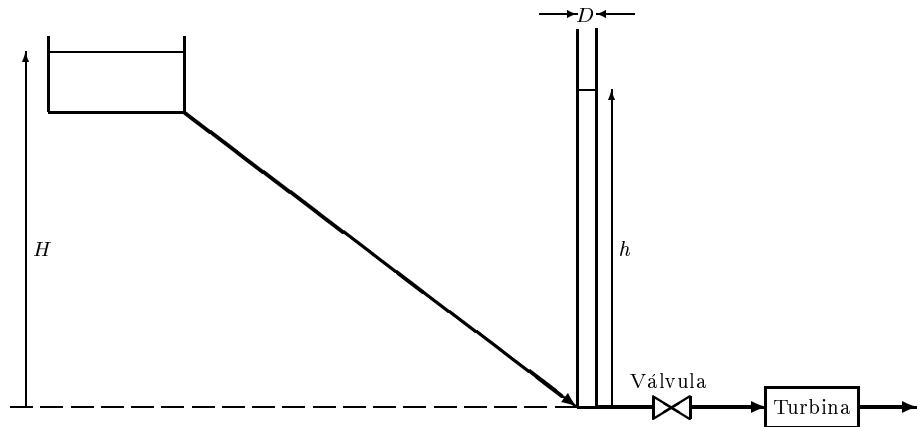


Sea s la distancia a lo largo de la directriz medida a partir de la base B , $y = y(s)$ la deformación horizontal en un punto cualquiera, $x = x(s)$ la altura en un punto cualquiera y $\theta = \theta(s)$ el ángulo entre el pilar y la vertical. Observar que $dy/ds = \operatorname{sen}(\theta)$. La ecuación que define el problema es

$$EI \frac{d\theta}{ds} = \int_s^L w[y(\zeta) - y(s)]d\zeta$$

que se puede diferenciar respecto de s para obtener una ecuación diferencial ordinaria de segundo orden. Las condiciones de contorno son $\theta = 0$ para $s = 0$ y $d\theta/dx = 0$ para $s = L$.

- (a) Calcular los valores de la variable adimensional $\gamma = wL^3/EI$ para los cuales x_A/L vale 0.99, 0.95, 0.9, 0.5 y 0. Calcular los valores correspondientes de y_A/L y α (valor de θ en el extremo libre A). Para introducir la variable adimensional γ en la ecuación hemos de efectuar el cambio $u = s/L$. El problema se puede resolver de la forma que se indica a continuación. Por un valor de γ dado hay que aplicar el método del tiro simple para resolver el problema de condiciones de contorno ($d\theta/dx = 0$ para $u = 1$), lo que cosa determinará $d\theta/ds$ en $u = 0$. El parámetro γ está inversamente relacionado con la rigidez del pilar y se debe ir variando para obtener el valor de x_A/L deseado. O sea: hay que hallar un cero de una función que depende de γ y para determinar esta función hay que resolver un problema de contorno (hallar un cero de una función dependiente de $d\theta/ds$). Previamente habrá que calcular la función θ para diferentes valores de γ y $d\theta/ds$ para escoger los intervalos adecuados.
 - (b) Calcular el valor de γ para el cual el pilar empieza a torcerse.
 - (c) Estudiar el caso concreto de un árbol de nave de aleación ligera de 38.825 m , y por el cual $w = 1.05\text{ kg/m}$, $E = 7 \times 10^{11}\text{ kg/m}^2$ y $I = 1.12 \times 10^{-8}\text{ m}^4$.
10. La instalación representada en la figura alimenta una turbina mediante un embalse de altura constante H . La chimenea de equilibrio de diámetro D está prevista para evitar presiones excesivas en la tubería si se debe cerrar la válvula rápidamente en un caso de emergencia.



Suponiendo la densidad constante y despreciando el efecto de la aceleración del agua en la chimenea, de la igualdad de la cantidad de movimiento del agua en la tubería resulta la expresión

$$g(H - h) - \frac{1}{2} f_M \frac{L}{d} u^2 = L \frac{du}{dt}$$

donde h es la altura del agua en la chimenea (h_0 en condiciones de equilibrio), f_M el factor de rozamiento de Moody, u la velocidad media en la tubería, g la aceleración de la gravedad y t el tiempo. Análogamente, por la ecuación de continuidad tenemos que

$$\frac{\pi d^2}{4} u = \frac{\pi D^2}{4} \frac{dh}{dt} + Q_v$$

El caudal de agua Q_v a través de la válvula en el periodo $0 \leq t \leq t_c$ durante el cual se está cerrando se puede aproximar por

$$Q_v = k \left(1 - \frac{t}{t_c}\right) (h - h^*)^{1/2}$$

donde la constante k depende del tipo de válvula y la altura h^* agua abajo depende de la turbina.

Calcular la variación de nivel en la chimenea de equilibrio en función del tiempo durante y después de un cierre de emergencia de la válvula.

Os sugerimos los datos siguientes:

$$\begin{aligned} g &= 9.8 \text{ m/s}^2 & H &= 30.5 \text{ m} \\ h_0 &= 26.8 \text{ m} & f_M &= 0.024 \\ L &= 610 \text{ m} & d &= 0.61 \text{ m} \\ t_c &= 6 \text{ s} & k &= 1.10 \text{ m}^{2.5}/\text{s} \\ D &= 1.22, 1.83, 3.05 \text{ y } 4.57 \text{ m.} \end{aligned}$$

Considerar dos valores extremos para h^* : (a) su valor original en condiciones normales ($h_0 - Q_{v_0}^2/k^2$), donde Q_{v_0} es el caudal original en condiciones normales, y (b) cero.

A Álgebra matricial

Si bien el lector puede desconocer algunos de los resultados que se utilizan en este libro sobre álgebra lineal y matrices, al ser suficientemente básicos para encontrarlos en cualquier texto de álgebra lineal, se presentan sin demostración.

A.1 Tipos de matrices

Dada una matriz A , se denota por A^T su **transpuesta**, que se obtiene al cambiar las filas por las columnas; por \bar{A} , la **conjugada**, que es la misma matriz pero con sus elementos conjugados; y por A^H la **adjunta**, que consiste en transponer y conjugar sus elementos. Las matrices con el mismo número de filas y columnas se llaman **matrices cuadradas**. El determinante de la matriz A se denota por $\det A$ y la matriz inversa por A^{-1} .

Una matriz cuadrada A puede ser de varios tipos; los más frecuentes son:

- **singular**: $\det A = 0$
- **simétrica**: $A^T = A$
- **hermítica**: $A^H = A$
- **ortogonal**: $A^{-1} = A^T$
- **unitaria**: $A^{-1} = A^H$
- **definida positiva**: $x^T A x > 0 \quad \forall x \neq 0$
- **Hessenberg superior**: $a_{ij} = 0$ para $j \leq i + 2$
- **triangular superior**: $a_{ij} = 0$ para $j \leq i + 1$
- **diagonal**: $a_{ij} = 0$ para $j \neq i$
- **tridiagonal**: $a_{ij} = 0$ para $j \neq i, i + 1, i - 1$

Se dice que dos matrices A y B son **semejantes** si existe una matriz no singular P tal que $B = P^{-1}AP$. Una matriz es **diagonalizable** si es semejante a una matriz diagonal.

A.2 La forma normal de Jordan

Sea $A \in \mathcal{L}(\mathbf{C}^n)$ y $\lambda_1, \lambda_2, \dots, \lambda_k$ sus valores propios diferentes con multiplicidad $\sigma_i = \sigma(\lambda_i)$. Existen $\rho(\lambda_i)$ números naturales $\nu_j^{(i)}$, para $j = 1 \div \rho(\lambda_i)$, tales que $\sigma(\lambda_i) = \nu_1^{(i)} + \dots + \nu_{\rho(\lambda_i)}^{(i)}$ y existe, también, una matriz $T \in \mathcal{L}(\mathbf{C}^n)$ no singular tal que $J = T^{-1}AT$, donde J es de la forma

$$J = \begin{pmatrix} C_{\nu_1^{(1)}}(\lambda_1) & & & & 0 \\ & \ddots & & & \\ & & C_{\nu_{\rho(\lambda_1)}^{(1)}}(\lambda_1) & & \\ & & & \ddots & \\ & & & & C_{\nu_1^{(k)}}(\lambda_k) \\ 0 & & & & & \ddots \\ & & & & & & C_{\nu_{\rho(\lambda_k)}^{(k)}}(\lambda_k) \end{pmatrix}$$

y las $C_\nu(\lambda)$ son matrices cuadradas de dimensión ν , de la forma

$$C_\nu(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & & \\ & \ddots & \ddots & & \\ & & \ddots & & \\ 0 & & & \ddots & 1 \end{pmatrix}$$

La matriz J se llama **forma normal de Jordan** de A y a las $C_\nu(\lambda)$ **bloques de Jordan**. Naturalmente, la matriz T no es única, pero los números $\nu_j^{(i)}$ están unívocamente determinados (se tienen tantos bloques de Jordan como vectores propios linealmente independientes).

Si se agrupan las columnas de T de acuerdo con las dimensiones de los bloques de Jordan, la primera columna de cada uno de estos grupos es un vector propio asociado al valor propio del correspondiente bloque de Jordan λ_i , y las otras columnas x_m para $m = 2 \div \nu$ son vectores tales que cumplen

$$\begin{aligned} (A - \lambda_i I)^j t_m &\neq 0 \quad \text{para } j = 1 \div m-1 \\ (A - \lambda_i I)^m t_m &= 0 \end{aligned}$$

estos vectores se llaman **vectores principales de grado m** (los vectores propios son vectores principales de primer grado).

Ejemplo. Si se considera la matriz

$$A = \begin{pmatrix} 3 & -5 & 0 & 5 \\ 5 & -6 & -5 & 9 \\ 5 & -4 & -2 & 4 \\ 5 & -4 & -5 & 7 \end{pmatrix}$$

su polinomio característico es

$$p(\lambda) = \lambda^4 - 2\lambda^3 - 11\lambda^2 + 12\lambda + 36 = (\lambda - 3)^2(\lambda + 2)^2$$

Está claro que sus raíces son -2 y 3 , ambas dobles. Si se estudia el rango de $A - \lambda I$ se observa que para $\lambda = -2$ es 3 y para $\lambda = 3$ es 2 ; por tanto, para el valor propio -2 se tiene un solo vector propio linealmente independiente y para el valor propio 3 se tienen dos. Así, la forma normal de Jordan será

$$J = \begin{pmatrix} -2 & 1 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

y la descomposición $J = V^{-1}AV$ vendrá dada, por ejemplo, por la matriz

$$V = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

donde la primera columna es un vector propio asociado al valor propio -2 , la segunda columna es un vector principal de grado 2 asociado al mismo valor propio y las dos últimas columnas corresponden a dos vectores propios linealmente independientes asociados al valor propio 3 .

A.3 Factorización de matrices

Dada una matriz $A \in \mathcal{L}(\mathbf{R}^n)$ existe una matriz de permutaciones P tal que

$$PA = LU$$

con L matriz triangular inferior con la diagonal igual a 1 y, U triangular superior (para más detalles ver el capítulo 4).

Dada una matriz $A \in \mathcal{L}(\mathbf{R}^n)$, existe una matriz ortogonal Q tal que

$$A = QR$$

donde R es una matriz triangular superior. Si se exige que los elementos de la diagonal principal de R sean positivos y A no es singular, esta descomposición es única (para más detalles ver el capítulo 7).

Dada una matriz $A \in \mathcal{L}(\mathbf{R}^n)$, existe una matriz ortogonal $Q \in \mathcal{L}(\mathbf{R}^n)$ tal que

$$Q^T A Q = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ & \ddots & & \vdots \\ & & \ddots & * \\ 0 & & & \lambda_n \end{pmatrix}$$

donde los λ_i son los valores propios de A . Si la matriz es simétrica, entonces la matriz resultante del producto es diagonal.

Una matriz $A \in \mathcal{L}(\mathbf{C}^n)$ **diagonaliza** si, y solamente si, existe una matriz unitaria $U \in \mathcal{L}(\mathbf{C}^n)$ tal que

$$U^H A U = D = \text{diag}(\lambda_1, \dots, \lambda_n)$$

A.3.1 Descomposición en valores singulares

Sea $A \in \mathcal{L}(\mathbf{R}^n, \mathbf{R}^m)$; entonces la matriz cuadrada de dimensión n $A^T A$ es simétrica y semidefinida positiva; por tanto, sus valores propios $\lambda_1, \dots, \lambda_n$ son no negativos y existen los valores $\sigma_i = +\sqrt{\lambda_i}$, que se llaman **valores singulares** de la matriz A . Además, existen dos matrices ortogonales U y V de dimensión m y n , respectivamente, tales que

$$U^T A V = U^{-1} A V = \Sigma$$

donde la matriz $\Sigma \in \mathcal{L}(\mathbf{R}^n, \mathbf{R}^m)$ es de la forma

$$\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \quad \text{con} \quad D = \text{diag}(\sigma_1, \dots, \sigma_r)$$

con $\sigma_1 \geq \dots \geq \sigma_r > 0$ valores singulares de la matriz A no nulos y, por tanto, r es el rango de A (para más detalles ver el capítulo 7).

A.4 Normas matriciales

Sea una familia de normas definidas para vectores reales, $x \in \mathbf{R}^n$, de la manera siguiente:

$\|x\|_p = (|x_1|^p + \dots + |x_n|^p)^{1/p}$. Para $p = 1$, se tiene $\|x\|_1 = \sum_{i=1}^n |x_i|$, y para $p \rightarrow \infty$, se obtiene $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$.

Las propiedades de una norma se cumplen en estos casos:

- 1) $\|x\| > 0 \wedge \|x\| = 0 \Leftrightarrow x = 0$
- 2) $\|r x\| = |r| \|x\| \quad \forall r \in \mathbf{R}$
- 3) $\|x + y\| \leq \|x\| + \|y\|$

Sea A una matriz cuadrada; se define una norma sobre la matriz A como una aplicación en \mathbf{R} que cumple las tres propiedades anteriores de las normas vectoriales más una multiplicativa:

- 1) $\|A\| > 0 \wedge \|A\| = 0 \Leftrightarrow x = 0$
- 2) $\|r A\| = |r| \|A\| \quad \forall r \in \mathbf{R}$
- 3) $\|A + B\| \leq \|A\| + \|B\|$
- 4) $\|A \cdot B\| \leq \|A\| \cdot \|B\|$

Si se tiene que $\|A x\| \leq \|A\| \|x\|$, se dice que la norma matricial es **consistente** con la norma vectorial (o, para abreviar, las dos normas son consistentes). Para cada norma vectorial se puede definir una norma matricial que sea consistente del siguiente modo:

$$\|A\| = \max_{\|x\| \neq 0} \frac{\|A x\|}{\|x\|} = \max_{\|x\|=1} \|A x\|$$

y se llama **norma matricial subordinada** la norma vectorial a partir de la que se define. Así, se tiene

- 1) $\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$
- 2) $\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$
- 3) $\|A\|_2 = (\rho(A^T A))^{1/2}$

donde $\rho(B) = \max_{i=1 \dots n} |\lambda_i|$ se llama **radio espectral** y es el valor absoluto (módulo, en caso que haya valores propios complejos) mayor de todos los valores propios ($\text{Spec}(B)$ es el conjunto de los valores propios de B y se llama **espectro de la matriz B**).

Propiedades:

1. $\rho(A) \leq \|A\|$ para cualquier norma matricial, ya que, si $\lambda \in \text{Spec}(A)$ y x es un vector propio asociado a λ , entonces

$$\|A\| \geq \frac{\|A x\|}{\|x\|} = \frac{\|\lambda x\|}{\|x\|} = \lambda$$

2. Si A es simétrica, entonces $\rho(A) = \|A\|_2$, ya que

$$\|A\|_2 = [\rho(A^T A)]^{1/2} = [\rho(A^2)]^{1/2} = \rho(A)$$

3. Para cualquier matriz $A \in \mathcal{L}(\mathbf{R}^n)$, dado un $\varepsilon > 0$, existe una norma matricial subordinada a alguna norma vectorial tal que

$$\|A\| \leq \rho(A) + \varepsilon$$

Solucionario

Capítulo 1

1. $x + y = -\frac{5}{3} \pm 0.036$.
2. $x_1 = 199.99\ 49\ 99\ 74\ 99$, $x_2 = 0.0050001$.
3. $1/(99 + 70\sqrt{2})$.
4. $x = 10 \pm 10\sqrt{1-a}$.
5. a) $\frac{2x^2}{1+3x} \approx 2x^2 - 6x^3$.
b) $\sqrt{x} \left(\frac{1}{x^2} + \frac{1}{8x^6} \right)$.
6. a) 0.067 y 0.014; b) 0.186 y 0.059; c) 0.107 y 0.034.
7. a) $[0, 6]$ y $s(x) = -\ln\left(2 - \frac{x}{3}\right)$; b) -0.29.
8. b) $\frac{2}{3} + \frac{3}{4} \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)(k+2)(k+3)(k+4)}$ y $0.693 \pm 0.5 \cdot 10^{-3}$ con 3 términos.
c) Directamente $k \geq 1998$ términos.
9. a) $I_{j-2} = \frac{1}{6} \left(\frac{1}{j-1} - I_{j-1} - I_j \right)$.
b) $t < \log_{10} \frac{2^{n-j-2}}{\varepsilon}$.
c) Tomando 3 n $2^{n-j-1} > 10^t$.
10. $|\Delta\lambda| \lesssim 0.067$; con 4 dígitos de mantisa.
11. $x \cdot y = -50 \pm 1.4$.
12. a) 33 términos; b) $7.6 \cdot 10^6$; c) $7 + 7 = 14$ dígitos.
13. $\Delta I(a, b) = 10^{-2}$.
14. $|y - \bar{y}| < \varepsilon \frac{4|x| + 31}{2\sqrt{2}}$.
15. b) Son necesarios 29 términos y $\zeta(3) = 1.202057$.

Capítulo 2

1. 6 ; 5.6667 ; 5.833 .
2. $p_3(x) = 0.73494 + (x + 0.9239)(0.2685 + (x + 0.38268)(0.05034 + (x - 0.38268)))$.
3. $h \lesssim 0.025$.
4. $n = 4$.
5. 0.56714 .
6. $p(0.1835) = 3.0886$.
7. a) $h \leq 2\sqrt{2} \cdot 10^{-3}$; b) $\frac{1}{2(n+1)!} (x-0)(x-\frac{1}{n}) \cdots (x-1)$.
8. a) $p(x) = \frac{a-x}{a^2-1}$; cota del error $\frac{1}{(a-1)^3}$.
b) $p(x) = \frac{2(a-x)}{2a^2-1}$; cota del error $\frac{1}{2(a-1)^3}$.
11. $h < 0.001741$.
13. Error de interpolación Chebyshev $\frac{4}{18^{n+2}}$; es necesario un polinomio de grado 4.
Para la interpolación equidistante el error es $2.1 \cdot 10^{-7}$.
14. $t = 4, 8, 12$.
15. $\operatorname{tg}(\pi/8) \approx 0.4018$; error cometido = 0.0124 ; cota del error = 0.074 .
16. 0.1785 .
17. $(2\Delta y_0 - \Delta^2 y_0)^2 - 8\Delta^2 y_0(y_0 - \bar{y}) > 0$.
18. $P(x) = 1 + 2(x-1) + 2(x-1)^2$; $P(1.6) = 2.92$.
19. $q_0(x) = x + x^2$; $q_1(x) = -4 + 9x - 3x^2$; $q_2(x) = 12 - 7x + x^2$.

Capítulo 3

1. $\bar{P}_0(x) = 1$; $\bar{P}_1(x) = \frac{2}{b-a} (x - \frac{a+b}{2})$;
 $\bar{P}_{j+1}(x) = \frac{2j+1}{j+1} \frac{2}{b-a} (x - \frac{a+b}{2}) \bar{P}_j(x) - \frac{j}{j-1} \bar{P}_{j-1}(x)$.
3. $\sum_{i=1}^{n-1} \frac{\langle \phi_n, \psi_i \rangle}{\langle \psi_i, \psi_i \rangle} \psi_i$.

4. $P_2^*(x) = p_3(x) - \lambda^{(2)} l_3(x)$, con $p_3(x)$ polinomio interpolador en los puntos $\{0, 0.25, 0.5, 1\}$

$$\text{y } \lambda^{(2)} = \frac{13}{8} (5 - 3\sqrt{2}) .$$

6. $P_n^*(x) = x^{n+2} - 2^{-n-1} T_{n+2}(x)$.

Capítulo 4

1. a) $(1.0013, 2.007, -1.0005)$; b) $(0.99994, 2.0001, -1.00)$; c) $(0.99994, 2.0002, -1.00)$.

$$2. U = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 7/2 & 1 \\ 0 & 0 & 0 & 0 & 33/7 \end{pmatrix} \text{ i } L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1/2 & 1 & 0 \\ 0 & 0 & 0 & 2/7 & 1 \end{pmatrix} .$$

$$3. A^{-1} = \begin{pmatrix} 21/10 & -29/20 & 2/5 & -1/20 \\ -3/10 & 17/20 & -7/10 & 3/20 \\ -11/30 & -1/60 & 8/15 & -3/20 \\ 3/20 & -1/20 & -3/20 & 1/20 \end{pmatrix} .$$

4. Solución $(1, 1, 1, 1)^T$; $\|r_1\|_\infty = 0.1$, $\|r_2\|_\infty = 0.01$ y $\kappa(A) = 4488$.

5. a) $(1, 1)^T$; b) $r = (0, -0.0015)^T$; c) $\kappa_1(A) = 3001$.

6. b) $\kappa_\infty(A_n) = 36n^2 + 12 + \frac{1}{n^2}$; c) Una cota superior es 18.

8. $|a| < 1$.

10. a) $|a| < 1$; c) $\omega = 1.1$.

11. $n = 80$.

12. c) $k = 4$.

13. $\alpha = 2.8882$ y $I = 5.6306$.

Capítulo 5

1. 491 mm.

2. 0.662 ± 0.005 .

3. 6.317 ± 0.005 .

4. a) 0.47839 ± 0.0001 ; b) 0.478305 .

5. Para $h = 0.2, 0.1, 0.05$, la regla de Simpson da $0.40535180, 0.40535166, 0.4053516637$. La extrapolación tiene como resultado 0.4053516640 .

6. Por trapecios 1.46156, 1.24544 y 1.21886. Por Romberg 1.21550 y por Euler-Maclaurin 1.2154985.
7. $2.7139 + 0.9937 = 3.7076$.
8. 9.68844822 .
9. $\int_{-1}^1 (1+x^2) f(x) dx = \frac{4}{3} \left(f\left(-\sqrt{\frac{2}{5}}\right) + f\left(\sqrt{\frac{2}{5}}\right) \right) + E(f)$, con $E(x^4) \approx 0.0108 f^{(4)}(\xi)$.
10. $A_0 = A_1 = (b-a)/2$ y $C_0 = C_1 = -(b-a)^3/24$.
11. El error es $\frac{f^{(6)}(\xi)}{15750}$. Además, si $h = \frac{b-a}{2}$, se obtiene
- $$\int_a^b f(x) dx = \frac{h}{9} \left(5f(a+h(1-\sqrt{0.6})) + 8f\left(\frac{a+b}{2}\right) + 5f(a+h(1+\sqrt{0.6})) \right).$$
- Por integración compuesta se obtiene 2.350336929, 2.350401261 y 2.350402369; por extrapolación 2.350402387 .
12. $\int_0^\infty \frac{\sin x}{x} dx = \int_0^\varepsilon \frac{\sin x}{x} dx + \int_\varepsilon^b \frac{\sin x}{x} dx + \int_b^\infty \frac{\sin x}{x} dx$, on $\varepsilon \approx 0.1$ i $b \approx 100$.
- $$\int_0^1 \frac{dx}{x^{1/2} + x^{1/3}} = \int_0^\varepsilon \frac{dx}{x^{1/2} + x^{1/3}} + \int_\varepsilon^1 \frac{dx}{x^{1/2} + x^{1/3}}.$$
13. $2.40393884 \pm 6 \cdot 10^{-7}$.
14. 1.9977 .
15. Son necesarios $n = 5$ términos y la integral vale 3.977463 .
16. $\int_{x_0}^{x_3} f(x) dx = \frac{3h}{2} (f_1 + f_2) + \frac{3h^3}{4} f''(\xi)$,
- $$\int_{x_0}^{x_4} f(x) dx = \frac{4h}{3} (2f_1 - f_2 + 2f_3) + \frac{28h^5}{90} f^{(4)}(\xi), \text{ etc.}$$
17. 0 ; 0 ; 0 .
18. 1.3475, 1.3820, 1.38033, 1.3803904 .

Capítulo 6

1. 1) -2.210083944 , -0.342185053 y 2.702061373 ;
- 2) 0.79681213002 ; 3) 0.557145599 ; 4) -1.0476827331 y 3.616105685 ;
- 6) 2.674060314 ; 7) $1.687342883 \in [\frac{\pi}{2}, \pi]$.
2. La segunda y la tercera; Newton ; 0.56714329 .
4. $k = 2$ iteraciones.
5. b) $\frac{4}{3} 10^{-4}$; c) $k = 22$.
7. a) $|g'(x^*)| < |f'(x^*)|$, donde x^* es la raíz; b) 0.651 ± 0.005 .

9. $(0.9350820641, 0.9980200582)$.

10. $(1.54665435, -3.72390026)$, $(1.57032922, -7.66894719)$, etc.

En general, para $k \leq -5$ se puede tomar la aproximación inicial $\left(\frac{\pi}{2}, \frac{12k+1}{3}\right)$.

11. a) $2.33333333, -0.5, -2.0$; b) $10, 0.5, -0.4$.

13. $-0.68212, -1.31784, -4.14628$.

14. a) $Q(u_k) = 1 - u_k \frac{f_k''}{2f_k'}$. b) Si $x_0 = 2$, entonces por Halley $x_3 = 1.4142139$, y por Newton $x_3 = 1.4142157$.

15. a) $x_{k+1} = x_k(2 - ax_k)$; b) $\varepsilon_{k+1} = a\varepsilon_k^2$; c) $k = 5$.

16. a) $x_{k+1} = x_k - \frac{2 f_k}{f_k' \pm \sqrt{(f_k')^2 - 4 f_k \frac{f_k' - f[x_{k-1}, x_k]}{x_k - x_{k-1}}}}$;

c) $1 + \sqrt{2}$.

Capítulo 7

1. $A = P D P^{-1}$, donde $P = \begin{pmatrix} 2 & 9 & 4 \\ 3 & 5 & 4 \\ -2 & 4 & -1 \end{pmatrix}$ y $D = \text{diag}(6, 2, 1)$.

2. 98.522.

3. 0.0122056 y $(-110.595, 24.957, -27.665, 1)^T$.

4. 19.29 y -7.08.

5. 4.040129.

6. 20.000, 4.000 y -2.000.

7. $A_6 = \begin{pmatrix} 3.3915 & 0.0000 & 0.0002 \\ 0.0000 & 1.7767 & 0.0051 \\ 0.0002 & 0.0051 & -1.1670 \end{pmatrix}$; $\lambda \approx 3.3915 \pm 0.0002$.

$3.39138238, 1.77286556$ i -1.16424794 .

8. a) 18 y $(0, 1, 2, 0)^T$.

b) 1 con $(1, 0, 0, 1)^T$, 8 con $(0, 1, 1, 0)^T$ y 3 con $(2, 0, 0, 1)^T$.

9. 1.267949, 3.0 y 4.7320508, donde $A_{10} = \begin{pmatrix} 4.7285 & 0.0781 & 0.0000 \\ 0.0781 & 3.0035 & -0.0020 \\ 0.0000 & -0.0020 & 1.2680 \end{pmatrix}$.

10. $HA = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 1.414213 \\ 0 & 0 & 1.414213 & 4 \end{pmatrix}$ y $HB = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 1 & 1.41421 \\ 0 & 1.41421 & 2 \end{pmatrix}$.

11.
$$\begin{pmatrix} 7.6 & 0.2 & 0 \\ 0.2 & 0.4 & -4.472136 \\ 0 & -4.472136 & 10 \end{pmatrix}.$$

Capítulo 8

3. $|\lambda_i| < 1$.

4. $y_k = \mu^{k/2} [c_1 \cos(k\theta) + c_2 \sin(k\theta)] + \frac{1}{1-\mu}$; estable.

5. Estable si $p \geq 1/2$; $y_k = c_1(1/p - 1)^k + c_2(-1)^k$.

6. $x_n = 5 \cdot 9^n - 2 \cdot 2^n$ y $y_n = 9^n + 2^n$.

7. $y_{n+2} - (1+a)y_{n+1} + ay_n = h/12 [(5+a)f_{n+2} + 8(1-a)f_{n+1} - (1+5a)f_n]$.

8. Orden 6; estable y $|h \tau(x, h)| = \frac{864}{133} h^7 |y^{(7)}(x)| + O(h^8)$.

9. Estable para $a \in [0, 1]$.

10. No consistente.

14. $y_{n+4} - 2\alpha y_{n+3} + 2\alpha y_{n+1} - y_n = h \left[4(1-\alpha)f_{n+2} + \frac{8-2\alpha}{3}\nabla^2 f_{n+3} + \frac{14+\alpha}{45}\nabla^4 f_{n+4} \right]$.

15. Estable en $(-32, 1)$.

18. a) $(-1, 0)$; b) $(-14.8328, 0)$.

21. $\left(-\frac{4}{3}, 0\right)$.

Bibliografía

- [Abr72] Abramowitz, M. ; Stegun, I. (editors). *Handbook of Mathematical Functions*. Dover Publications, New York, 1972.
- [Act70] Acton, F. S. *Numerical Methods that Work*. Harper & Row, New York, 1970.
- [All90] Allgower, E. L. ; Georg, K. *Numerical Continuations Methods*. Springer-Verlag, Berlin, 1990.
- [Aub91] Aubanell, A. ; Benseny, A. ; Delshams, A. *Eines bàsiques del càlcul numèric*. Publicacions de la Universitat Autònoma de Barcelona, Barcelona, 1991.
- [Azi75] Aziz, A. K. *Numerical Solutions of Boundary Value Problems*. Academic Press, New York, 1975.
- [Bra65] Bracewell, R. *The Fourier Transform and Its Applications*. McGraw-Hill, New York, 1965.
- [Bre77] Brezinski, C. *Accélération de la convergence en analyse numérique*. Springer-Verlag, Berlin, 1977.
- [Bri74] Brigham, E. O. *The Fast Fourier Transform*. Prentice Hall, Englewood Cliffs, New Jersey, 1974.
- [Bur85] Burden, R. L. ; Faires, J. D. *Análisis numérico*. Grupo Editorial Iberoamérica, México, 1985.
- [But85] Butcher, J. C. *The Numerical Analysis of Ordinary Differential Equations*. John Wiley, New York, 1985.
- [Car79] Carnahan, B. ; Luther, H. A. ; Wilkes, J. O. *Cálculo numérico. Métodos, aplicaciones*. Rueda, 1979.
- [Cia82a] Ciarlet, P. G. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Masson, Paris, 1982.
- [Cia82b] Ciarlet, P. G. ; Thomas, J. M. *Exercices d'analyse numérique matricielle et d'optimisation*. Masson, Paris, 1982.

- [Cia90] Ciarlet, P. G. ; Lions, J. L. (editors). *Handbook of Numerical Analysis*, volume I i II. North-Holland, Amsterdam, 1990.
- [Coh77] Cohen, A. M. *Análisis numérico*. Reverté, Barcelona, 1977.
- [Con74] Conte, S. D. ; de Boor, C. *Análisis numérico*. McGraw-Hill, Colombia, 1974.
- [Dah74] Dahlquist, G. ; Björck, Å. *Numerical Methods*. Prentice Hall, Englewood Cliffs, New Jersey, 1974.
- [Dav75a] Davis, P. J. *Interpolation and Approximation*. Dover Publications, New York, 1975.
- [Dav75b] Davis, P. J. ; Rabinowitz, P. *Methods of Numerical Integration*. Academic Press, New York, 1975.
- [de 78] de Boor, C. *A Practical Guide to Splines*. Springer-Verlag, New York, 1978.
- [Dem77] Demidovich, B. P. ; Maron, I. A. *Cálculo numérico fundamental*. Paraninfo, Madrid, 1977.
- [Dem80] Demidovich, B. P. ; Maron, I. A. ; Schuwalowa, E. S. *Métodos numéricos de análisis*. Paraninfo, Madrid, 1980.
- [Eld90] Elden, L. ; Wittmeyer-Koch, L. *Numerical Analysis*. Academic Press, New York, 1990.
- [For67] Forsythe, G. E. ; Moler, C. B. *Computer Solution of Linear Algebraic Systems*. Prentice Hall, Englewood Cliffs, New Jersey, 1967.
- [For77] Forsythe, G. E. ; Malcom, M. A. ; Moler, C. B. *Computer Methods for Mathematical Computations*. Prentice Hall, Englewood Cliffs, New Jersey, 1977.
- [Fr 77] Fröberg, C. E. *Introducción al análisis numérico*. Vicens Vives, Barcelona, 1977.
- [Fr 85] Fröberg, C. E. *Numerical Mathematics. Theory and Computer Applications*. Benjamin/Cummings, Menlo Park, 1985.
- [Gar86] García-Merayo, F. *Programación en FORTRAN 77*. Paraninfo, Madrid, 1986.
- [Gea71] Gear, G. W. *Numerical Initial Value Problem in Ordinary Differential Equations*. Prentice Hall, Englewood Cliffs, New Jersey, 1971.
- [Ger84] Gerald, C. F. ; Wheatley, P. O. *Applied Numerical Analysis*. Addison Wesley, Reading, Massachusetts, 1984.
- [Gol73] Golub, G. H. ; van Loan, C. F. *Matrix Computations*. John Wiley, New York, 1973.
- [Gou73] Gourlay, A. R. ; Watson, G. A. *Computational Methods for Matrix Eigenproblems*. John Wiley, Chichester, 1973.
- [Hag81] Hageman, L. A. ; Young, D. M. *Applied Iterative Methods*. Academic Press, New York, 1981.

- [Hen62] Henrici, P. *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley, New York, 1962.
- [Hen64] Henrici, P. *Elements of Numerical Analysis*. John Wiley, New York, 1964.
- [Hil74] Hildebrand, F. B. *Introduction to Numerical Analysis*. McGraw-Hill, New York, 1974.
- [Hou75] Householder, A. S. *The Theory of Matrices in Numerical Analysis*. Dover, New York, 1975.
- [Isa66] Isaacson, E. ; Keller, H. B. *Analysis of Numerical Methods*. John Wiley, New York, 1966.
- [Joh82] Johnson, L. W. ; Riess, D. R. *Numerical Analysis*. Addison Wesley, Reading, Massachusetts, 1982.
- [Kel68] Keller, H. B. *Numerical Methods for Two-Point Boundary Value Problems*. Blaisdell, London, 1968.
- [Ker78] Kernighan, B. W. ; Ritchie, D. M. *The C Programming Language*. Prentice Hall, Englewood Cliffs, New Jersey, 1978.
- [Knu69] Knuth, D. E. *The Art of Computer Programming*. Addison Wesley, Reading, Massachusetts, 1969.
- [Lam79] Lambert, J. D. *Computational Methods in Ordinary Differential Equations*. John Wiley, New York, 1979.
- [Lin79] Linz, P. *Theoretical Numerical Analysis*. John Wiley, New York, 1979.
- [Mil70] Milne, W. E. *Numerical Solution of Differential Equations*. Dover Publications, New York, 1970.
- [Ort70] Ortega, J. M. ; Rheinboldt, W. C. *Iterative Solution on Nonlinear Equations in Several Variables*. Academic Press, New York, 1970.
- [Ort72] Ortega, J. M. *Numerical Analysis. A Second Course*. Academic Press, New York, 1972.
- [Ort81] Ortega, J. M. ; Poole, W. G. *An Introduction to Numerical Methods for Differential Equations*. Pitman, Marshfield, Massachusetts, 1981.
- [Pap78] Papoulis, A. *Sistemas digitales y analógicos, transformada de Fourier, estimación espectral*. Marcombo, México, 1978.
- [Par80] Parlett, B. *The Symmetric Eigenvalue Problem*. Prentice Hall, Englewood Cliffs, New Jersey, 1980.
- [Pow80] Powell, M. J. D. *Approximation Theory and Methods*. Cambridge University Press, Cambridge, 1980.
- [Pre86] Press, W. H. ; Flannery, B. P. ; Teukolsky, S. A. ; Vetterling, W. T. *Numerical Recipes*. Cambridge University Press, New York, 1986.

- [Qui85] Quinney, D. *An Introduction to Numerical Solution of Differential Equations*. John Wiley, New York, 1985.
- [Ral67] Ralston, A. ; Wilf, M. S. (editors). *Mathematical Methods for Digital Computers*, volume I i II. John Wiley, New York, 1960, 1967.
- [Ral70] Ralston, A. *Introducción al análisis numérico*. Limusa-Wiley, México, 1970.
- [Ral78] Ralston, A. ; Rabinowitz, P. *A First Course in Numerical Analysis*. McGraw-Hill, Auckland, 1978.
- [Ric83] Rice, J. R. *Matrix Computations and Mathematical Software*. McGraw-Hill, Japan, 1983.
- [Riv69] Rivlin, T. J. *An Introduction to the Approximation of Functions*. Dover Publications, New York, 1969.
- [Sch68] Scheid, F. *Análisis numérico*. Schaum McGraw-Hill, México, 1968.
- [Sch73] Schultz, M. *Spline Analysis*. Prentice Hall, Englewood Cliffs, New Jersey, 1973.
- [Sch89] Schwarz, H. R. *Numerical Analysis. A Comprehensive Introduction*. John Wiley, New York, 1989.
- [Sha75] Shampine, L. F. ; Gordon, M. K. *Computer Solution of Ordinary Differential Equations*. W. H. Freeman, San Francisco, 1975.
- [Smi76] Smith, B. T. ; Boyle, J. M. ; Dongarra, J. J. ; Garbow, B. S. ; Ikebe, Y. ; Klema, V. C. ; Moler, C. B. *Matrix Eigensystem Routines EISPACK Guide*. Springer-Verlag, New York, 1976.
- [Sny66] Snyder, M. A. *Chebyshev Methods in Numerical Approximation*. Prentice Hall, Englewood Cliffs, New Jersey, 1966.
- [Spi70] Spiegel, M. R. *Manual de fórmulas y tablas matemáticas*. McGraw-Hill, Colombia, 1970.
- [Ste73] Stewart, G. W. *Introduction to Matrix Computations*. Academic Press, New York, 1973.
- [Sto80] Stoer, J. ; Bulirsch, R. *Introduction to Numerical Analysis*. Springer-Verlag, New York, 1980.
- [Str66] Stroud, A. H. ; Secrest, D. *Gaussian Quadrature Formulas*. Prentice Hall, Englewood Cliffs, New Jersey, 1966.
- [Str71] Stroud, A. H. *Approximate Calculation of Multiple Integrals*. Prentice Hall, Englewood Cliffs, New Jersey, 1971.
- [Str80] Strang, W. G. *Linear Algebra and its Applications*. Academic Press, New York, 1980.
- [Szi78] Szidarovsky, F. *Principles and Procedures of Numerical Analysis*. Plenum Press, 1978.

- [Tod79] Todd, J. *Basic Numerical Mathematics*, volume I i II. Academic Press, Birkhäuser Verlag, New York, 1977, 1979.
- [Tra64] Traub, J. F. *Iterative Methods for the Solution of Equations*. Prentice Hall, Englewood Cliffs, New Jersey, 1964.
- [Van83] Vandergraft, J. S. *Introduction to Numerical Computations*. Academic Press, New York, 1983.
- [Var62] Varga, R. S. *Matrix Iterative Analysis*. Prentice Hall, Englewood Cliffs, New Jersey, 1962.
- [Wil63] Wilkinson, J. H. *Rounding Errors in Algebraic Processes*. Prentice Hall, Englewood Cliffs, New Jersey, 1963.
- [Wil65] Wilkinson, J. H. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, 1965.
- [Wil71] Wilkinson, J. H. ; Reinsch, C. *Handbook for Automatic Computation*, volume 2: Linear Algebra. Springer-Verlag, Berlin, 1971.
- [You71] Young, D. M. *Iterative Solution of Large Linear Systems*. Academic Press, New York, 1971.
- [You88] Young, D. M. ; Gregory, R. T. *A Survey of Numerical Mathematics*. Dover Publications, New York, 1988.

Listado de rutinas

A continuación presentamos un listado de todas las rutinas que aparecen en este libro. Se especifica si son un programa principal (P), una subrutina (S) o una función (F). También se detalla, para cada una de ellas, a qué otras rutinas se llaman y por cuáles son llamadas.

Nombre	Comentario	Pag.
BACTERIA (P)	Ejemplo de aproximación por mínimos cuadrados, utiliza la rutina SVD.	96
CIRCUITO (P)	Ejemplo de resolución de sistemas lineales, utiliza las rutinas DECOMP y SOLVE.	133
COLUMNA (P)	Ejemplo de cálculo de valores propios, utiliza las rutinas ELMHES y HQR.	265
DECOMP (S)	Descomposición LU por eliminación gaussiana, es utilizada por CIRCUITO.	134
ELMHES (S)	Transforma una matriz a matriz Hessenberg, es utilizada por COLUMNA.	266
ECSEGUNDO (P)	Cálculo de la solución de la ecuación de segundo grado.	47
FUN (F)	Función necesaria para ROMBERG.	168
F (F)	Función necesaria para ZEROAP.	213
FUN1 (F)	Función necesaria para ROMBERG.	168
HQR (S)	Método QR sobre una matriz Hessenberg, es utilizada por COLUMNA.	266
PILAR (P)	Ejemplo de integración numérica, utiliza las rutinas ROMBERG8, FUN y FUN1.	168
PIROLISIS (P)	Ejemplo de la resolución de una ecuación diferencial, utiliza las rutinas REACTOR y RKF78.	315
POBLACION (P)	Ejemplo de interpolación por spline cúbica, utiliza las rutinas SPLINE y SAVAL.	76
REACTOR (S)	Rutina utilizada por RKF78.	316
RKF78 (S)	Método Runge-Kutta-Fehlberg de órdenes 7 y 8 para ecuaciones diferenciales, es utilizada por PIROLISIS.	317
ROMBERG8 (S)	Cálculo de la integral de una función por un método de Romberg de orden ocho, utiliza las rutinas FUN y FUN1, es utilizada por ROMBERG.	169
SAVAL (S)	Avaluación de la spline cúbica generada por SPLINE, es utilizada por POBLACION.	75

Nombre	Comentario	Pag.
SOLVE (S)	Resolución del sistema de ecuaciones lineales descompuesto por DECOMP, es utilizada por CIRCUITO.	137
SPLINE (S)	Cálculo de la spline cúbica Forsythe, es utilizada por POBLACION.	74
SVD (S)	Cálculo de los valores singulares de una matriz utilizando una variante del método QR, es utilizada por BACTERIA.	97
VOLUMEN (P)	Ejemplo del cálculo del cero de una función, utiliza las rutinas ZEROAP y FUN.	213
ZEROAP (F)	Cálculo del cero de una función combinando los métodos de la biseción, interpolación cuadrática inversa y secante, utiliza la rutina FUN, es utilizada por VOLUM.	214

Glosario de símbolos

- $\|f\|_2$ norma euclíadiana, 79
 $\|f\|_{2,w}$ norma euclíadiana con peso, 79
 $\|f\|_\infty$ norma infinito, 79
 $\|f\|_{\infty,w}$ norma infinito con peso, 79
 Δx error absoluto, 26
 $\Delta^k f(x_i)$ k -ésima diferencia finita, 64
 ϵ épsilon de la máquina, 24
 ε_x error relativo, 26
 $\kappa(A)$ número de condición de la matriz A , 114
 μ_A unidad de redondeo, 27
 μ_T unidad de truncamiento, 27
 $\Pi(r; \bar{h})$ polinomio de estabilidad absoluta, 288
 $\rho(A)$ radio espectral de la matriz A , 337
 $\rho(\zeta), \sigma(\zeta)$ polinomios asociados a un método lineal multipaso, 285
 $\tau(x, h)$ error local de truncamiento, 281
 $\tau(h) = \max_{x \in [a, b-kh]} \|\tau(x, h)\|$, 282
 $\tau_{n+k} = \tau(x_n, h)$, 287
 B_J matriz de iteración del método de Jacobi, 120
 B_{GS} matriz de iteración del método de Gauss-Seidel, 121
 B_ω matriz de iteración de los métodos de relajación, 124

$fl(x)$ flotante de x , 23

$f[x_i, x_{i+1}, \dots, x_{i+j}, x_{i+j+1}]$ diferencias divididas de Newton, 58

$l_i(x)$ polinomios de Lagrange, 56

$O(a_n)$ orden de una sucesión, 30

$\text{Spec}(A)$ espectro de la matriz A , 337

$V(a)$ número de cambios de signo en el punto a de una sucesión de polinomios, 230

$E(\kappa)$ integral elíptica de segunda especie completa ([Abr72] cap. 17)

$\text{erf}(x)$ función error ([Abr72] cap. 7)

$F(a, b; c; x)$ función hipergeométrica ([Abr72] cap. 15)

J_n función de Bessel ([Abr72] cap. 9)

$K(\kappa)$ integral elíptica de primera especie completa ([Abr72] cap. 17)

$T_n(x)$ polinomios de Chebishef ([Abr72] cap. 22)

\ll mucho más pequeño que

\lesssim más pequeño o aproximadamente igual a

$i = 1 \div n$ i es un número entero desde 1 hasta n

$\langle f, g \rangle$ producto escalar de las funciones f y g

δ_{ij} delta de Kronecker

$\text{diag}(a_1, \dots, a_n)$ matriz diagonal de elementos a_1, \dots, a_n

$I(x_0, x_1, x_2)$ intervalo más pequeño que contiene los puntos x_0, x_1, x_2

\mathcal{P}_n espacio de polinomios de grado inferior o igual a n

\mathcal{R}_{mn} espacio de funciones racionales de grado como máximo m y n

$\text{sign}(x)$ función signo de x

$y(x; x_0, y_0)$ solución de la ecuación diferencial $y' = f(x, y)$ con condiciones iniciales $y(x_0) = y_0$

Índice de Materias

A

- Adams-Bashforth, método de, 297
- Adams-Bashforth, método de , 312
- Adams-Moulton, método de, 297
- Adams-Moulton, método de , 312
- Aitken, método de, 56
- Algoritmo estable, 21
- Aproximación
 - continua, 81
 - discreta, 81
 - exponencial, 81
 - mínimo-cuadrática polinómica, 88
 - min-max, 82, 91
 - existencia de la , 91
 - polinómica, 81
 - por mínimos cuadrados, 82
 - racional, 81
 - trigonométrica, 81

B

- Base
 - ortogonal, 84
 - ortonormal, 84
- Bessel, función de, 270
- Bisección, método de la, 176, 234

C

- Cancelación catastrófica, 26
- Chebychev
 - polinomios de, 84
 - Teorema de, 92
- Chebyshef
 - polinomios de, 95
- Choleski, método de, 111
- Cifras significativas, 27
- Consistencia, 284, 286

Control del paso, 300

Convergencia, 284

Crout, método de, 111

D

- Dahlquist, teorema de, 287
- Decimales correctos, 27
- Deflación, 239
 - de Householder, 241
 - de Wielandt, 239
- Descomposición
 - en valores singulares, 91
 - LU, 246
- Desplazamiento
 - del origen, 237, 253, 257
 - doble conjugado, 254, 257
- Diferencias divididas, 58
- Doble desplazamiento conjugado, 257
- Doolittle, método de, 110

E

- Ecuación
 - en diferencias lineal, 276
 - normal, 83, 85
 - vectorial en diferencias, 275
- Eliminación gaussiana, 106
- Épsilon de la máquina, 24
- Error
 - absoluto, 26
 - de convergencia, 14
 - de discretización, 14, 283
 - de redondeo, 14, 18
 - de truncamiento, 14, 18
 - hacia atrás, 21
 - inicial o de medida, 18
 - local, 283

de truncamiento, 283, 298
 relativo, 26
 Errores del problema, 17
 Estabilidad, 285, 286
 absoluta, 289
 de un predictor-corrector, 299
 de un Runge-Kutta, 303
 asintótica, 276
 de una ecuación en diferencias, 276
 Estimación del error, 302
 Euler, método de, 286
 adelante, 281, 290, 295
 atrás, 282, 291
 modificado, 301
 Extrapolación de Richardson, 144

F

Factorización
 $L U$, 108
 QR, 232, 249
 Fibonacci, sucesión, 278
 Fórmulas de Newton-Côtes, 146
 Francis, método de, 249
 Función
 aproximadora, 81, 82
 error, 269, 324
 ortogonal, 85
 ortonormal, 85

G

Gauss
 -*Chebishev*, integración de, 158
 -*Hermite*, integración de, 160
 -*Jordan*, método de, 112
 -*Laguerre*, integración de, 159
 -*Legendre*, integración de, 155
 -*Seidel*, método de, 121
 Gauss,
 integración de, 153
 Gershgorin, teorema, 224
 Givens,
 método de, 227, 231
 reflexiones de, 259
 transformaciones de, 256
 Golub-Reinsch, método de, 258
 Gram, polinomios de, 87

Gram-Schmidt, método de ortogonalización de, 85

H

Hermite, interpolación de, 66
 Hessenberg superior, matriz, 226, 232, 233, 245
 Heun,
 fórmula de, 301
 método de, 283, 286
 Hilbert, matriz de, 84, 233
 Householder,
 deflación de, 241
 método de, 229
 matriz de, 229, 241
 transformaciones de, 255, 256

I

Integración
 Gauss-Chebishev, 158
 Gauss-Hermite, 160
 Gauss-Laguerre, 159
 Gauss-Legendre, 155
 gaussiana, 153
 Interpolación, 55
 polinómica, 55

J

Jacobi
 cíclico, método de, 243
 clásico, método de, 243
 Jacobi,
 método de, 120, 242

K

Kirchoff, ley de, 331
 Kutta, regla de, 301

L

Lagrange, fórmula de, 56
 Laguerre, método de, 204
 Lanczos, método de, 140
 Legendre, polinomios de, 86, 89, 95
 Lineal multipaso, método, 281, 285, 286, 290
 explícito, 281
 implícito, 281
 LR,

- método, 246
- M**
- Mal condicionado, 114
- Meinardus, conjetura de, 104
- Método
- óptimo, 287
 - compacto, 110
 - de Romberg, 149
 - directo, 106
 - iterativo, 118
 - lineal multipaso, 281
 - LR, 246, 249
 - QR, 249, 258
 - implementación, 256
- N**
- Neville, método de, 56
- Newton
- método de, 307
- Newton,
- fórmula de, 58
 - método de, 178, 233, 234, 306
- Newton-Côtes, fórmulas de, 146
- Norma
- estricta, 83
 - euclidiana, 85
- Número de condición, 34, 114
- O**
- Ohm, ley de, 331
- Orden, 286
- de convergencia, 182
- Ortogonalización de Gram-Schmidt, método de, 85
- P**
- Padé, aproximación de, 103
- Péndulo invertido, 328
- Pirólisis del etano, 313
- Pivotamiento
- parcial, 108
 - total, 108
- Polinomio
- característico, 277, 283
 - de estabilidad absoluta, 290
 - ortogonal, raíces, 88
- Polinomios
- ortogonales, 85
- Potencia,
- inversa desplazada, método de la, 238
 - inversa, método de la, 238
 - método de la, 235, 239
- Predictor-corrector, método, 296, 297, 304
- Problema
- de tres cuerpos, 329
 - de valores
 - frontera, 304
 - iniciales, 280
 - singulares, 305 - inestable, 20
- Q**
- QR, método, 249, 258
- implementación numérica, 256
- Quade, método de, 310
- R**
- Rayleigh, cociente de, 236
- Redondear, 24
- Reflexiones de Givens, 259
- Región de estabilidad absoluta, 290
- Regula-Falsi, método de la, 177
- Richardson, extrapolación de, 144
- Romberg, método de, 149
- Runge-Kutta, método, 282, 289, 300, 301, 304
- de orden 4, 283
 - explícito, 282
 - Fehlberg, 302
 - implícito, 282
- Runge-Kutta
- Rutishauser, método de, 246
- S**
- Schmidt, método de ortonormalización de, 256
- Secante, método de la, 179
- Sistema
- lineal sobredeterminado, 127
- Sobrerelajación, método de, 124
- Spline
- cúbica, 66
 - completa, 68
 - Forsythe, 68

natural, 68
Sturm, sucesión de, 233

Substitución
 hacia atrás, 106
 hacia delante, 106

T

Tiro, método del
 paralelo, 307
 simple, 305
Trapezio, método del, 282
Traslación del origen, 253, 257
Truncamiento, 14
 error local de, 283
Truncar, 24

U

Unidad
 de redondeo, 28
 de truncamiento, 28

V

Valor propio
 más pequeño, cálculo del, 238
Valores
 frontera, 304
 singulares, 257
 cálculo de, 257
 descomposición en, 91
Vector residual, 114
Volterra, 327

W

Weierstrass, teorema, 82
Wielandt,
 deflación de, 239
 método de, 238