ÚTILES BÁSICOS DE CÁLCULO NUMÉRICO

A. Aubanell A. Benseny A. Delshams

PRÓLOGO

En los diversos campos de la ciencia, la tecnología, la medicina, la economía, las ciencias sociales, etc., se describen a menudo fenómenos reales mediante modelos matemáticos. El estudio de estos modelos aporta un conocimiento más profundo de aquellos fenómenos y, con frecuencia, permite predecir su evolución futura. Buscar y aplicar los instrumentos más adecuados para encontrar soluciones a problemas basados en estos modelos constituye el objetivo principal de la matemática aplicada. Se trata de un arte apasionante que, desgraciadamente, no siempre puede recurrir a los métodos analíticos clásicos por diversas razones: no se adecuan al modelo concreto, su aplicación resulta excesivamente laboriosa, su solución formal es tan compleja que hace imposible cualquier interpretación posterior, etc. En tales casos, son útiles las técnicas numéricas que, mediante una labor de cálculo más o menos intensa, conducen a soluciones aproximadas. La notable tarea calculística que comporta la aplicación de la mayoría de estos métodos hace que su uso esté fuertemente ligado a la utilización de sistemas de cálculo automático. Así, sin el desarrollo que se ha producido en el campo de la informática, resultaría difícilmente imaginable el nivel actual de utilización de técnicas numéricas en ámbitos cada vez más diversos.

Este libro es fruto de una amplia tarea docente en la Facultat de Matemůtiques de la Universitat de Barcelona. Su contenido representa un primer curso de cálculo numérico y va dirigido a estudiantes de carreras científicas, técnicas o sociales que quieran conocer, de manera tan práctica como sea posible, útiles básicos que les permitan afrontar cuestiones numéricas con comodidad y rigor.

La aplicación práctica de muchos de los métodos que se presentan en este texto requiere el uso de calculadoras y de ordenadores. Muchos de los problemas numéricos de este libro pueden ser abordados utilizando calculadoras de bolsillo. La complejidad de algunos métodos o su escala de aplicación hace imprescindible el uso de ordenadores. Por esto, es altamente aconsejable que el lector conozca algún lenguaje de programación que le permita implementar métodos numéricos en programas.

En este libro se desarrollan, agrupados en cinco capítulos, los temas más básicos del cálculo numérico:

- El concepto de error es consubstancial con el cálculo numérico. En cada problema, es de gran importancia hacer un seguimiento de los errores cometidos con el fin de poder estimar el grado de aproximación de la solución que se obtiene del mismo. Al aplicar técnicas numéricas a un problema determinado, es necesario estudiar los diversos tipos de error que afectan a sus soluciones: los errores propios del método numérico usado complementados con los errores provenientes de los datos y de las operaciones. En el primer capítulo sólo se pueden dar técnicas para estimar o acotar los errores

de los dos últimos tipos, teniendo también en cuenta las limitaciones del sistema de cálculo empleado (calculadora, ordenador ...). En los capítulos siguientes, se ha procurado acompañar la presentación de cada método numérico con la información necesaria para poder analizar su error.

- Los modelos matemáticos más básicos son los modelos lineales, que provienen a menudo de una reducción de otros más complejos. El análisis de los problemas asociados a los modelos lineales comporta el cálculo de la solución de sistemas lineales o de sus valores y vectores propios asociados. El capítulo 2 contempla una gran variedad de métodos generales para la realización de estos cálculos. En cada situación, la elección concreta de un método depende de diversos factores, como la estructura del sistema (simetría, escasez de coeficientes no nulos, ...), la eficiencia, la estabilidad numérica, etc. Se pueden así escoger métodos directos, métodos iterativos, métodos más estables a costa de más operaciones, métodos específicos para sistemas con estructura especial, etc.

- Las técnicas de aproximación e interpolación ofrecen procedimientos para extraer un modelo sencillo de una realidad más compleja, expresada sea por un conjunto de datos experimentales, sea por una función de evaluación complicada. Estas reducciones, explicadas en el capítulo 3, permiten dar formas alternativas más simples de cálculo de funciones y son la base de la mayoría de los métodos numéricos para la obtención de derivadas, integrales y ceros de funciones, presentados en los capítulos posteriores. Las funciones interpoladoras y de aproximación que se usan son primordialmente polinomiales y trigonométricas. Los métodos de aproximación tienen como objetivo minimizar la función error según algún criterio. Las aproximaciones por mínimos cuadrados y minimax que se presentan corresponden a dos criterios diferentes de medida de dicha función error.

- Los cálculos con funciones a menudo no se pueden llevar a cabo analíticamente o no resulta eficiente hacerlo. En el capítulo 4 se describen diferentes fórmulas de derivación, integración y sumación numérica de funciones, acompañadas de su error. La existencia de expresiones asintóticas de estos errores permite diseñar estrategias de extrapolación para obtener una utilización más eficaz de dichas fórmulas. Se concede especial importancia al cálculo con operadores por la posibilidad que ofrece de una reformulación más ágil de alguno de los métodos presentados en este capítulo y en el anterior.

- Una de las tareas más genuinas del cálculo numérico es la resolución de ecuaciones no lineales. Los métodos de resolución se basan en la construcción de una sucesión obtenida iterativamente que converja a la solución. Su elección depende de la función y de la solución buscada. En cada caso, los diversos métodos se diferencian por la velocidad de convergencia de las sucesiones cerca de la solución y por el esfuerzo de cálculo que requieren. En el capítulo 5 se expone un resumen de los procedimientos más usuales, tanto para ecuaciones generales como, específicamente, para ecuaciones polinomiales. Se concluye el capítulo con una breve introducción a la resolución de sistemas de ecuaciones no lineales.

En todo momento se ha procurado huir de un texto reducido a un catálogo de métodos,

recetas y trucos. Se ha insistido en las ideas generales asociadas a los temas presentados: planteamiento del problema, conceptos que intervienen, estrategias de resolución, estudio de errores, cuestiones de convergencia, ... En la exposición de métodos concretos se ha buscado siempre un marco globalizador (tópicos comunes, características generales, etc.) que evite una casuística excesivamente dispersa.

Cada capítulo consta, además de las secciones dedicadas a la exposición teórica, de una sección de comentarios bibliográficos, de una sección de problemas resueltos y de una de problemas propuestos.

- En las secciones teóricas, se ha buscado un equilibrio entre la claridad del desarrollo y la profundidad conceptual. Sin embargo, el cálculo numérico es más pragmático que fundamentalista, más herramienta que materia prima. Esta idea está presente en estas secciones, de manera que nos ha preocupado más la presentación del marco general y la exposición concreta de los útiles de cálculo que los detalles de fundamentación de algunos conceptos.

- En los comentarios bibliográficos se suministran referencias complementarias al material teórico que, en algunos casos, contemplan demostraciones no presentadas explícitamente. Se aportan también citas que contienen programas implementando los métodos presentados en diferentes lenguajes de programación para que el lector pueda utilizarlos directamente o bien compararlos con los elaborados por él mismo.

- Las secciones de problemas resueltos constituyen el núcleo de cada capítulo. Tanto la concepción inicial como el objetivo final de la elaboración de este libro han estado basados en el principio de que la mejor forma de familiarizarse con el uso de cualquier instrumento es utilizarlo debidamente. Por esto, se ha cuidado mucho la exposición del proceso de resolución de los problemas. La inclusión de este material tiene dos objetivos:

  - Simplificar las exposiciones teóricas, remitiendo a la sección de problemas aquellos detalles que las harían demasiado pesadas.
  - Ofrecer ejemplos concretos de aplicación de los métodos expuestos. En este sentido, se ha intentado que los problemas resueltos abarquen esencialmente todas las herramientas teóricas presentadas y, a través de la práctica, permitan una mejor comprensión de las mismas.

- Los problemas propuestos, muy diversos en cuanto a su dificultad, son argumentos de refuerzo que permiten la ejercitación del lector.

No queremos acabar esta introducción sin expresar nuestro agradecimiento a los compañeros de docencia de la asignatura de Càlcul Num+eric en la Facultat de Matemàtiques. De todos ellos hay algo en este texto, pero nos complace mencionar a Joan Llopart, colaborador en unos apuntes previos, a Joaquim Font, paciente calculador de ceros de polinomios y a Gerard Gómez, consejero en las últimas etapas.

Deseamos hacer una mención especial a nuestro maestro común, Carles Simó. Él nos introdujo en el arte del cálculo numérico y, haciendo camino a su lado, nos descubrió paisajes de insospechada belleza. A él se deben los aciertos que este libro pueda aportar,

nosotros nos hacemos responsables de sus errores. Sean estas palabras un testimonio de gratitud hacia su cordial magisterio.

Maribel, Fanny y Rosalia han demostrado tener una paciencia tendente al infinito. Su labor también está presente en estas páginas.

Los autores, julio de 1991.

ÍNDICE DE CONTENIDOS

CAPÍTULO 1

ERRORES

Cualquier proceso de cálculo numérico debe prestar especial atención al control de los errores para poder estimar en qué medida afectan al resultado. En el presente capítulo se exponen técnicas de acotación de los errores atribuibles a los datos de entrada y a los generados en el proceso de resolución debido al necesario redondeo de los resultados de los cálculos intermedios. El estudio de los errores inherentes a cada método numérico se abordará, de manera específica, en el momento de su descripción.

1.1 CONCEPTOS GENERALES

1.1.1 Definiciones

Con gran frecuencia no conocemos el *valor exacto* $\bar{x}$ de una magnitud y hemos de conformarnos con un *valor aproximado* x. En tal caso, definimos:

- *Error absoluto de* x: $e_a(x) = x - \bar{x}$.
- *Error relativo de* x:

$$e_r(x) = \frac{e_a(x)}{\bar{x}} \ ,$$

si $\bar{x} \neq 0$.

En la práctica, si $x \neq 0$, se suele estimar

$$e_r(x) \simeq \frac{e_a(x)}{x} \ .$$

En general, no conocemos estos errores exactamente, sino sólo una *cota* de ellos; es decir, un número $\epsilon_a(x)$ tal que $\mid e_a(x) \mid \leq \epsilon_a(x)$, para el error absoluto; o bien, $\epsilon_r(x)$ tal que $\mid e_r(x) \mid \leq \epsilon_r(x)$, para el error relativo. Utilizaremos también la notación: $\bar{x} = x \pm \epsilon_a(x)$ o $x = \bar{x} \pm \epsilon_a(x)$, para el error absoluto, y $\bar{x} = x(1 \pm \epsilon_r(x))$ o $x = \bar{x}(1 \pm \epsilon_r(x))$, para el relativo.

1.1.2 Fuentes de error

Desde un punto de vista numérico, nos interesan tres *fuentes de error*: errores en los datos de entrada, errores de redondeo durante el cálculo y error de truncamiento del método empleado.

Error en los datos de entrada

Los errores en los datos pueden ser debidos a dos causas:

MEDICIONES INCORRECTAS

Los instrumentos de medida aportados por la tecnología no presentan una precisión indefinidamente fina. Debido a ello, los valores medidos están afectados por errores que dependen básicamente de la precisión de los instrumentos.

FINITUD DE LA REPRESENTACIÓN DIGITAL DE UN DATO

Elegida una base natural $b \geq 2$, cualquier número real x no negativo puede ser representado en la forma

$$x = a_n b^n + a_{n-1} b^{n-1} + \cdots + a_1 b + a_0 + a_{-1} b^{-1} + a_{-2} b^{-2} + \cdots \ ,$$

con $a_j \in \mathbb{Z}$, $0 \leq a_j < b$ $(j \leq n)$, que habitualmente se escribe

$$x = a_n a_{n-1} \ldots a_1 a_0 . a_{-1} a_{-2} \cdots {}_b \ ,$$

y se llama *representación digital de x en la base b*. Los coeficientes asociados a_j $(j \leq n)$ reciben el nombre de *dígitos* o *cifras*. Esta representación es única, excepto para los números racionales x de la forma

$$x = \frac{k}{b^n} \quad (k, n \in \mathbb{Z}) \ ,$$

que tienen dos. Así, por ejemplo,

$$\frac{132}{100} = 1.319999 \ldots {}_{10} = 1.32 {}_{10} \ ,$$

donde $1.32 {}_{10}$ es la representación finita de $\frac{132}{100}$, formada sólo por 3 cifras significativas: $a_0 = 1$, $a_{-1} = 3$, $a_{-2} = 2$. Los problemas 1.1–1.4 ofrecen más ejemplos de representación digital.

Supongamos ahora que efectuamos nuestros cálculos con una calculadora que puede representar números con t dígitos en base b. En este caso la representación de un número x con más de t dígitos no nulos no será exactamente igual a x, y la llamaremos $\mathrm{fl}(x)$, *representación en punto flotante de x*. En general,

$$\mathrm{fl}(x) = m \cdot b^q \ ,$$

donde $q \in \mathbb{Z}$ y $m = \pm 0.a_1 a_2 \ldots a_t {}_b$, con $a_j \in \mathbb{Z}$, $0 \leq a_j < b$ $(j = 1 \div t)$ y $a_1 \neq 0$; q recibe el nombre de *exponente* y m, el de *mantisa*.

El paso de x a $\mathrm{fl}(x)$ se puede hacer por *corte*, simplemente suprimiendo los dígitos de x a partir de a_t, o por *redondeo*, escogiendo $\mathrm{fl}(x)$ de manera que el error $\mathrm{fl}(x) - x$ sea mínimo. Cuando esta condición da dos posibles redondeos, normalmente se escoge el que tiene mayor valor absoluto (si estos casos apareciesen dentro de una cadena larga de cálculos, para evitar cualquier tipo de desviación del error de redondeo hacia alguna dirección fija, sería preferible redondear de manera que la última cifra de la mantisa fuese, por ejemplo, siempre par).

Las cotas del error relativo producido, en cada caso, son:

$$\epsilon_C = b^{1-t} \ , \quad \epsilon_R = \frac{1}{2} b^{1-t} \ .$$

EJEMPLOS

1. -99.962 (o $-0.99962 \cdot 10^2$) cortado a 3 cifras queda -99.9 (o $-0.999 \cdot 10^2$) y, redondeado a 3 cifras, -100. (o $-0.100 \cdot 10^3$).

2. 35.47846 redondeado a 6 cifras queda 35.4785, y 35.4785 redondeado a 5 cifras se transforma en 35.479. En cambio, redondeando (directamente) 35.47846 a 5 cifras queda 35.478.

3. El número π es $3.14159265358979323846264 3\ldots$. Si π entra como dato en una máquina que solamente admite 10 cifras en base 10, será representado por

$$\begin{aligned} \mathrm{fl}_C(\pi) &= 0.3141592653 \cdot 10^1 \text{ (si corta)} , \\ \mathrm{fl}_R(\pi) &= 0.3141592654 \cdot 10^1 \text{ (si redondea)} ; \end{aligned}$$

   donde $\mathrm{fl}_C(x)$ indica la *representación en punto flotante de x por corte* y $\mathrm{fl}_R(x)$, la *representación en punto flotante de x por redondeo*. Obsérvese que el error de la primera representación es mayor que el de la segunda.

Error de redondeo durante el cálculo

El error en un resultado no solamente puede provenir de los errores de los datos de entrada, sino también de los errores de redondeo en los resultados de los cálculos intermedios.

EJEMPLO

Consideremos una máquina que trabaja con 4 cifras decimales y corta o redondea. Tenemos los dos datos:

$$a = 0.3425 \cdot 10^5 , \quad b = 0.2517 \cdot 10^{-2} .$$

Entonces,

$$\begin{aligned} ab &= 0.8620725 \cdot 10^2 , \\ \mathrm{fl}_C(ab) &= 0.8620 \cdot 10^2 \text{ (se ha cometido un error de } 0.725 \cdot 10^2) , \\ \mathrm{fl}_R(ab) &= 0.8621 \cdot 10^2 \text{ (se ha cometido un error de } 0.275 \cdot 10^2) . \end{aligned}$$

Si no se dice lo contrario supondremos que, en cada *operación aritmética* ($+$, $-$, $\cdot$, $/$), las cotas del error relativo serán las mismas que en la representación de los datos de partida: ϵ_C o ϵ_R.

Error de truncamiento del método empleado

Cuando resolvemos un problema matemático por métodos numéricos, aunque efectuemos las operaciones exactamente, obtenemos sólo una *aproximación numérica* del resultado exacto (por ejemplo, cuando aproximamos una integral por una suma finita o una derivada por un cociente incremental, etc.).

El error producido depende del método numérico empleado y recibe el nombre de *error de truncamiento*. Para algunos métodos, se dispone de expresiones de este error, tal como veremos en los capítulos siguientes.

Al decidir sobre la conveniencia de la utilización de un método determinado, se ha de tener en cuenta no solamente su error de truncamiento, sino también los errores de redondeo producidos por las operaciones que el método comporta.

Per ejemplo, hay que huir siempre de los métodos o algoritmos que comporten *cancelaciones* de cifras al restar dos cantidadas próximas. Dado que en estas cancelaciones se producen errores relativos considerablemente grandes, conviene usar fórmulas matemáticamente equivalentes que las eviten. Con el fin de resaltar la importancia de este fenómeno, en los problemas 1.8–1.11 se ofrecen diversos ejemplos.

1.2 ESTIMACIÓN Y ACOTACIÓN DE ERRORES

El objetivo de cualquier estudio de errores es tratar de conocer el efecto que, sobre el resultado final de un problema numérico, produce cada uno de los diferentes tipos de error que pueden tener lugar.

Distinguiremos los tres tipos básicos de error: los de los datos, los de los cálculos intermedios y los errores de truncamiento del método numérico empleado. El error total sobre el resultado final será la suma de las contribuciones de los tres tipos de error.

1.2.1 Propagación de los errores de los datos

Para los problemas numéricos consistentes en efectuar operaciones aritméticas $(+, -, \cdot, /)$ con dos datos x_1 y x_2 afectados de error, tenemos las siguientes cotas del error propagado:

$$\epsilon_a(x_1 + x_2) = \epsilon_a(x_1) + \epsilon_a(x_2) \ ,$$
$$\epsilon_a(x_1 - x_2) = \epsilon_a(x_1) + \epsilon_a(x_2) \ ;$$

y las siguientes cotas aproximadas, si los errores de x_1 y x_2 son pequeños:

$$\epsilon_r(x_1 x_2) \simeq \epsilon_r(x_1) + \epsilon_r(x_2) \ ,$$
$$\epsilon_r(x_1/x_2) \simeq \epsilon_r(x_1) + \epsilon_r(x_2) \ .$$

Para un problema numérico consistente en calcular el resultado $y = f(x)$ a partir de un único dato x, obtenemos la siguiente *fórmula aproximada de propagación del error*

$$e_a(y) \simeq f'(x) \cdot e_a(x) \ , \tag{1.1}$$

como consecuencia directa del teorema del valor medio, para funciones f de una variable, derivables con continuidad. De esta fórmula se deduce una cota aproximada para el error absoluto de y, en función de una cota del error absoluto de x, dando lugar a la *fórmula aproximada de propagación del error maximal*

$$\epsilon_a(y) \simeq | f'(x) | \, \epsilon_a(x) \ . \tag{1.2}$$

Finalmente, para el problema numérico más general, que consiste en calcular un resultado $y = f(x_1, \ldots, x_n)$ a partir de unos datos $x_1, \ \ldots, \ x_n$, disponemos de la *fórmula aproximada de propagación del error*

$$e_a(y) \simeq \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) e_a(x_i) \ ; \tag{1.3}$$

a partir de la cual, conocidas cotas de $e_a(x_i)$, podemos acotar $e_a(y)$, obteniendo la *fórmula aproximada de propagación del error maximal*

$$\epsilon_a(y) \simeq \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}(x_1, \ldots, x_n)\right| \epsilon_a(x_i) \ ; \tag{1.4}$$

especialmente adecuada cuando n, el número de datos afectados de error, no es grande.
EJEMPLO

Si calculamos la suma $y = x_1 + \cdots + x_n$ de n datos x_1, ..., x_n, entonces

$$e_a(y) \simeq \sum_{i=1}^{n} e_a(x_i) \ .$$

Si cada variable tiene una cota de error ϵ, la fórmula del error maximal nos da $\epsilon_a(y) \simeq n\epsilon$, cota que sólo puede ser alcanzada cuando todos los errores $e_a(x_i)$ tengan el mismo signo y la máxima magnitud, situación altamente improbable cuando n es grande.

1.2.2 Propagación de los errores en los cálculos

La propagación de los errores en los cálculos es estudiada en dos fases, tal como se describe a continuación. Pueden encontrarse diversos ejemplos de su aplicación en los problemas resueltos 1.14–1.18.

Análisis del error hacia atrás

Partiendo de datos iniciales exactos, debido a la acumulación de los errores en las operaciones, obtenemos un resultado afectado de error. La idea básica del *análisis del error hacia atrás* consiste en estudiar las modificaciones que tendríamos que hacer sobre los datos de entrada, de forma que, suponiendo que no hubiesen errores en las operaciones, se obtuviese el mismo error en el resultado.

Este estudio se lleva a cabo aplicando sucesivamente la fórmula

$$\mathrm{fl}(a * b) = (a * b)(1 + \delta_*) \ ,$$

con $|\,\delta_*\,| \leq \epsilon_*$, a cada una de las operaciones aritméticas $* = (+, -, \cdot, /)$ que componen el proceso de cálculo, donde ϵ_* indica una cota conocida del error relativo en la operación $*$; además, para todas las funciones g que intervienen en los cálculos, se escribe

$$\mathrm{fl}(g(x)) = g(x)(1 + \delta_g) \ ,$$

con $|\,\delta_g\,| \leq \epsilon_g$, donde ϵ_g indica una cota conocida del error relativo en la evaluación de g.

A continuación, se escribe una expresión del resultado final que permita imputar los errores de los cálculos a los datos. Con este procedimiento se reduce el análisis del error en los cálculos a un análisis de propagación de errores de los datos sin errores en los cálculos.

Propagación de los errores imputados a los datos

Una vez hecha la reducción anterior, se aplica la fórmula de propagación del error maximal a las cotas de los errores imputados a los datos, considerando que los cálculos ya se hacen sin error. En el cuadro de la figura 1.1 se muestra gráficamente todo el proceso.

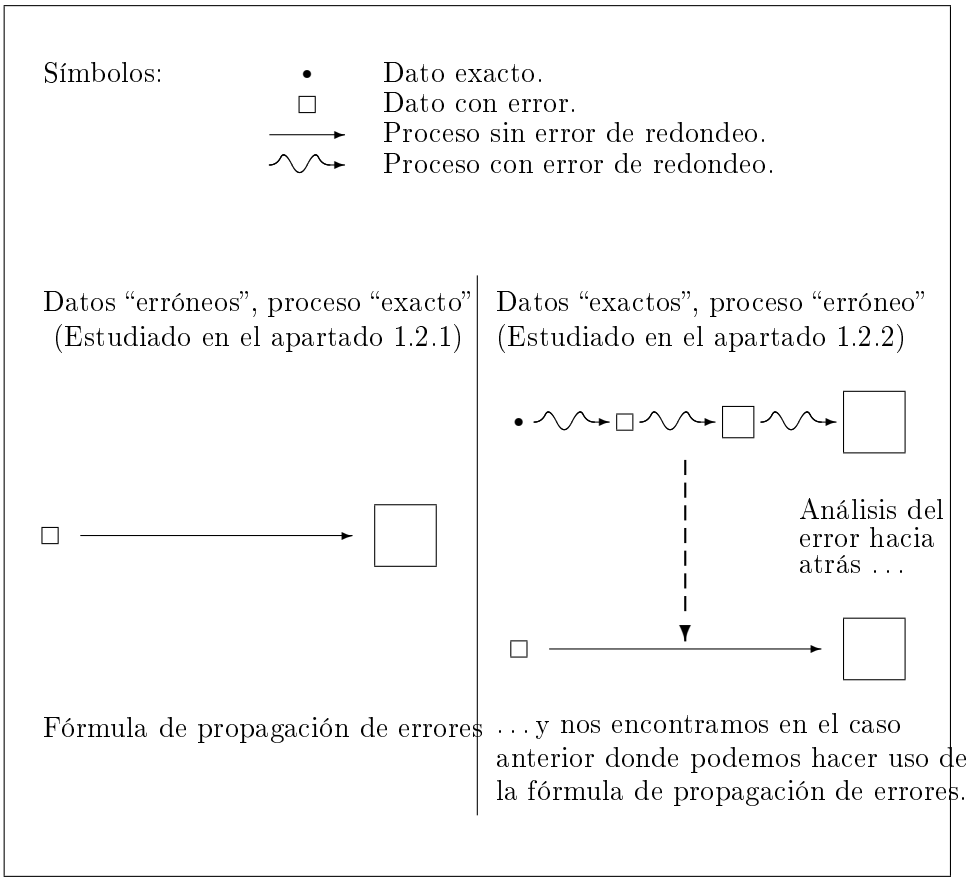


Figura 1.1: Cuadro resumen del tratamiento de errores.

1.2.3 Errores de truncamiento

El error de truncamiento depende de cada método numérico y su estudio se hará de manera específica para los diferentes métodos que se vayan presentando en los capítulos que siguen. De hecho, la estimación del error de truncamiento asociado a cada método conforma una parte fundamental de su exposición.

1.3 TRATAMIENTO INTERVALAR Y ESTADÍSTICO

La fórmula de propagación de errores (1.3) es sólo una fórmula aproximada (de primer orden) para la acotación del error. El cálculo del error usando intervalos permite obtener acotaciones rigurosas; desafortunadamente, estas acotaciones no son realistas cuando el número de cálculos es grande; es decir, son acotaciones desproporcionadas y poco probables de los errores reales. Para obtener estimaciones más cercanas a la realidad resulta muchas veces conveniente analizar el error desde un punto de vista estadístico. A continuación se presentan estos dos tratamientos.

1.3.1 Análisis de intervalos

El análisis de intervalos hace uso de las operaciones entre intervalos con el fin de determinar un intervalo en el cual, con toda seguridad, se encuentre el resultado final.

Un *intervalo (cerrado y acotado)* I en $\mathbb{R}$ es un conjunto I de números reales de la forma $[a,b] = \{x \in \mathbb{R} : a \leq x \leq b\}$, con $a, b \in \mathbb{R}$. Llamaremos $I(\mathbb{R})$ al conjunto de los intervalos en $\mathbb{R}$.

Se definen a continuación las operaciones más sencillas con intervalos. Definimos primero las *operaciones aritméticas entre intervalos*: si $I_1, I_2 \in I(\mathbb{R})$,

$$I_1 * I_2 = \{y \in \mathbb{R} : y = x_1 * x_2,\ x_1 \in I_1,\ x_2 \in I_2\} \ ,$$

para cualquier operación aritmética $*$ entre números reales.

Para un gran número de funciones entre números reales (por ejemplo, las continuas) podemos definir las correspondientes *funciones entre intervalos* de la siguiente manera: dada la función $f : \mathbb{R} \longrightarrow \mathbb{R}$, definimos

$$f : I(\mathbb{R}) \longrightarrow I(\mathbb{R}) \ , \quad f(I) = \{y : y = f(x), x \in I\} \ .$$

Usando operaciones y funciones entre intervalos, en lugar de operaciones y funciones entre números, el análisis de intervalos permite llevar a cabo un seguimiento, a lo largo de la cadena de cálculos, de la región de la recta real donde se encuentran los sucesivos resultados parciales y, en último término, del resultado final del proceso de cálculo.

1.3.2 Análisis estadístico

Para disponer de unas estimaciones más realistas del error del resultado, cuando éste se obtiene a través de un proceso numérico que comporta muchas operaciones, es conveniente introducir un *tratamiento estadístico del error*.

Para ello, se supone que los errores en los datos de entrada son *variables aleatorias independientes* con una cierta *función de distribución* dada, como se muestra en los ejemplos siguientes.

EJEMPLOS

a) Sea e el error al redondear un número a d cifras decimales. Entonces e toma valores no nulos sólo sobre $[-\epsilon, \epsilon]$, donde $\epsilon = \frac{1}{2}10^{-d}$; si suponemos que cada valor sobre este intervalo es igualmente probable, la *función de densidad* ρ de e corresponde a una *distribución uniforme* (es decir, a una función que vale $\frac{1}{2\epsilon}$ sobre $[-\epsilon, \epsilon]$ y se anula fuera de este intervalo); la *función de distribución*

$$F(x) = \int_{-\infty}^{x} \rho(t)dt$$

que nos da la probabilidad de que e tome valores más pequeños o iguales que x, es una recta de pendiente $\frac{1}{2\epsilon}$ en aquel intervalo. En la figura 1.2 se ven representadas ambas funciones.

Damos a continuación diferentes magnitudes estadísticas de la variable aleatoria e:

$$\begin{aligned}
\mu &= \int_{-\infty}^{\infty} x\rho(x)dx = 0 \quad (\textit{media de } e) \ , \\
\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2\rho(x)dx = \epsilon^2/3 \quad (\textit{varianza de } e) \ , \\
\sigma &= \frac{\epsilon}{\sqrt{3}} = \frac{10^{-d}}{\sqrt{12}} \simeq 0.2887 \cdot 10^{-d} \quad (\textit{desviación típica o estándar de } e) \ .
\end{aligned}$$

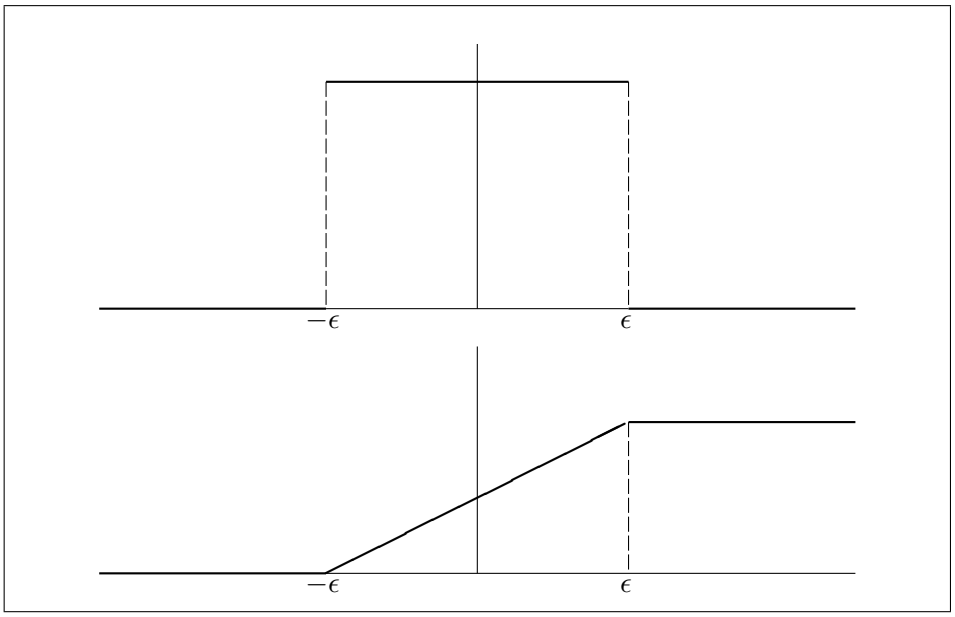


Figura 1.2: Funciones de densidad y de distribución de la distribución uniforme.

b) La mayoría de las variables aleatorias e que se manejan en la práctica tienen aproximadamente una *distribución normal*; esto es, tienen una función de densidad de probabilidad aproximadamente igual a

$$\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \ ,$$

donde μ es la media y σ^2 la varianza. En la figura 1.3 se presenta una representación gráfica de dicha función de densidad.

La probabilidad $P(\delta)$ de que e tome valores entre $\mu - \delta$ y $\mu + \delta$ viene dada por

$$P(\delta) = \int_{\mu-\delta}^{\mu+\delta} \rho(t)dt = \frac{2}{\sqrt{\pi}} \int_0^{\frac{\delta}{\sqrt{2}\sigma}} e^{-t^2} dt = \mathrm{erf}(\frac{\delta}{\sqrt{2}\sigma}) \ ,$$

donde

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

recibe el nombre de *función de error*.

Como ejemplos, $Q(\delta) = 1 - P(\delta)$ vale 0.317, 0.045 y 0.0027, cuando δ vale σ, 2σ y 3σ, respectivamente. Esto quiere decir que, si e representa un error de un dato con media $\mu = 0$ (por ejemplo) y desviación estándar σ, la afirmación de que la magnitud del error sea menor que σ, 2σ y 3σ es falsa en aproximadamente el 32%, el 5% y el 0.3% de los casos, respectivamente.

Volviendo a la teoría general de tratamiento estadístico de los errores, consideremos e_1, ..., e_n variables aleatorias con medias μ_1, ..., μ_n y desviaciones típicas σ_1, ..., σ_n, y sean d_1, ..., d_n constantes arbitrarias. Se cumplen las propiedades siguientes:

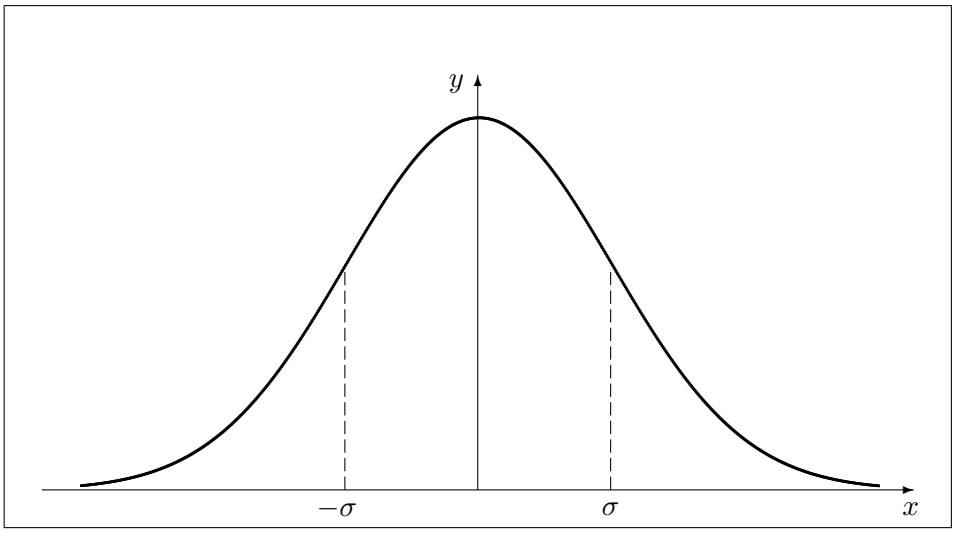


Figura 1.3: Función de densidad de la distribución normal.

1. La combinación lineal $e = \sum_{i=1}^{n} d_i e_i$ es una variable aleatoria con media

$$\mu = \sum_{i=1}^{n} d_i \mu_i \ .$$

2. Si $e_1, \ldots, e_n$ son *independientes* (es decir, el valor que toma una variable no condiciona el valor de cualquiera de las restantes), e tiene por desviación típica σ, con

$$\sigma^2 = \sum_{i=1}^{n} d_i^2 \sigma_i^2 \ .$$

   Si además, $e_1, \ldots, e_n$ son normales, también lo es e.

3. Si $e_1, \ldots, e_n$ son independientes y tienen la misma función de distribución, la función de distribución de e se aproxima a una función de distribución normal (cuando $n \to \infty$) (*Teorema central del límite*).

Si queremos calcular $y = f(x_1, \ldots, x_n)$ y suponemos que $e_a(x_i) = x_i - \bar{x}_i$ son variables aleatorias independientes con media nula y desviación típica $\sigma(e_a(x_i))$, que escribiremos simplemente $\sigma(x_i)$; $e_a(y) = y - \bar{y}$ viene dada entonces por la fórmula aproximada de propagación del error (1.3). Por tanto, teniendo en cuenta las propiedades anteriores, es una variable aleatoria con media nula y desviación estándar $\sigma(e_a(y))$, que también escribiremos simplemente $\sigma(y)$, dada por la *fórmula aproximada de propagación del error estándar*

$$\sigma(y) \simeq \left(\sum_{i=1}^{n} \left[\frac{\partial f}{\partial x_i}(x_1, \ldots, x_n) \right]^2 \sigma(x_i)^2 \right)^{1/2} \ . \tag{1.5}$$

Notemos que la aproximación de $e_a(y)$ dada en la fórmula (1.3) tiene una distribución normal, si las variables $e_a(x_i)$ la tienen, y aproximadamente normal, si éstas tienen la misma función de distribución y n es grande.

En el ejemplo de la suma $y = x_1 + x_2 + \cdots + x_n$, tratando los errores de redondeo a t cifras decimals $e_a(x_i)$ como variables aleatorias independientes con distribución uniforme de media nula y desviación típica $\epsilon = \frac{1}{2}10^{-t}$, la fórmula anterior da la desviación típica siguiente para y

$$\sigma(y) \simeq \sqrt{n}\sigma = \sqrt{\frac{n}{3}}\epsilon \ .$$

Por el teorema central del límite, $e_a(y)$ tiene distribución aproximadamente normal con media nula, si n es grande. Así, por ejemplo, se puede afirmar que la magnitud de $e_a(y)$ es menor que

$$2\sqrt{\frac{n}{3}}\epsilon \ ,$$

en aproximadamente el 95% de los casos en que se hagan sumas de n variables independientes.

COMENTARIOS BIBLIOGRÁFICOS

La mayoría de los libros de cálculo numérico tienen algún tipo de introducción al estudio de errores. Para un punto de vista similar al de este capítulo, remitimos al lector a [DB74], [RR78], [SB80], [YG72].

Unas buenas introducciones a la representación de los números se encuentran en [Hen64], [Ham73], [Knu69], y principalmente en [Wil64]. El teorema central del límite es uno de los más importantes de la teoría de probabilidades, teoría que sobrepasa el nivel de este libro y que se encuentra en los tratados clásicos [Cra46], [Fel50]. La referencia básica sobre el análisis de intervalos es [Moo66].

PROBLEMAS RESUELTOS

Problema 1.1 *Convertir*

a) 26.32_{10} *a base 2,*

b) 1314.96_{10} *a base 16,*

c) $AF.3C_{16}$ *a base 10.*

SOLUCIÓN:

Estudiemos primero cómo cambian las cifras de un número r al cambiarlo de la base b a la base B. Si $r_b = a_n \dots a_0.a_{-1} \dots {}_{b)}$ y $r_B = A_N \dots A_0.A_{-1} \dots {}_{B)}$, tenemos

$$\sum_{i=-\infty}^{n} a_i b^i = \sum_{j=-\infty}^{N} A_j B^j.$$

Igualando las partes enteras de ambos miembros

$$\sum_{i=0}^{n} a_i b^i = \sum_{j=0}^{N} A_j B^j$$

y dividiéndolas por B, resulta que A_0 es el resto de la división de la parte entera de r entre B. Tomando ahora el cociente de esta división y repitiendo la división por B, obtenemos A_1 como nuevo resto; la repetición de este proceso permite conocer todas las cifras de la parte entera de r en la base B: A_0, A_1, ..., A_N.

Análogamente, igualamos las partes fraccionarias

$$\sum_{i=-\infty}^{-1} a_i b^i = \sum_{j=-\infty}^{-1} A_j B^j \ ,$$

y, si ahora las multiplicamos por B, obtenemos como parte entera A_{-1}. Repetiendo este proceso de nuevo, con la parte fraccionaria del producto, tenemos como nueva parte entera A_{-2} y, así sucesivamente, encontraríamos todas las cifras de la parte fraccionaria de r en la base B: A_{-1}, A_{-2}, ... Para los casos pedidos:

a)

$$26 = 13 \cdot 2 + \underline{0} \quad 13 = 6 \cdot 2 + \underline{1} \quad 6 = 3 \cdot 2 + \underline{0} \quad 3 = \underline{1} \cdot 2 + \underline{1} \quad .$$

Así, $26_{10} = 11010_2$.

$$\begin{array}{llll} 0.32 \cdot 2 = \underline{0}.64 & 0.64 \cdot 2 = \underline{1}.28 & 0.28 \cdot 2 = \underline{0}.56 & 0.56 \cdot 2 = \underline{1}.12 \\ 0.12 \cdot 2 = \underline{0}.24 & 0.24 \cdot 2 = \underline{0}.48 & 0.48 \cdot 2 = \underline{0}.96 & 0.96 \cdot 2 = \underline{1}.92 \\ 0.92 \cdot 2 = \underline{1}.84 & 0.84 \cdot 2 = \underline{1}.68 & 0.68 \cdot 2 = \underline{1}.36 & 0.36 \cdot 2 = \underline{0}.72 \\ 0.72 \cdot 2 = \underline{1}.44 & 0.44 \cdot 2 = \underline{0}.88 & 0.88 \cdot 2 = \underline{1}.76 & 0.76 \cdot 2 = \underline{1}.52 \\ 0.52 \cdot 2 = \underline{1}.04 & 0.04 \cdot 2 = \underline{0}.08 & 0.08 \cdot 2 = \underline{0}.16 & 0.16 \cdot 2 = \underline{0}.32 \\ \dots & & & \end{array}$$

Por tanto, la expresión requerida es

$$\boxed{26.32_{10} = 11010.\overline{01010001111010111000}_2} \ ,$$

donde las líneas superiores indican el período del número.

b)

$$1314 = 82 \cdot 16 + \underline{2} \quad 82 = \underline{5} \cdot 16 + \underline{2} \quad .$$

Tenemos así, $1314_{10} = 522_{16}$.

$$\begin{array}{lll} 0.96 \cdot 16 = \underline{15}.36 & 0.36 \cdot 16 = \underline{5}.76 & 0.76 \cdot 16 = \underline{12}.16 \\ 0.16 \cdot 16 = \underline{2}.56 & 0.56 \cdot 16 = \underline{8}.96 & \ldots \end{array} \ .$$

Usando como dígitos en base 16: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F, el resultado es

$$\boxed{1314.96_{10} = 522.\overline{F5C28}_{16}} \ .$$

c) Simplemente hay que desarrollar

$$AF.3C_{16} = 10 \cdot 16 + 15 + 3 \cdot 16^{-1} + 12 \cdot 16^{-2} \ .$$

Finalmente,

$$\boxed{AF.3C_{16} = 175.234375_{10}} \ .$$

Problema 1.2 *a) Convertir los números siguientes a base 10:*

$$100010001.100010001_2 \ , \quad 42356.66_7 \ , \quad 1F2D.124_{16} \ .$$

b) Convertir los números siguientes, en base 10*, a las bases 2 , 7 y 16:*

$$4235.66 \ , \quad \pi \ , \quad e \ .$$

c) Efectuar las siguientes operaciones:

$$10001.1001_2 + 1001.1111_2 \ , \quad 11.1101_2 \cdot 11_2 \ , \quad 45.67_9 \cdot 44_9 \ ,$$

$$11.11_2/0.1111_2 \ , \quad 1221.2112_3/22_3 \ , \quad A1_{16}/F_{16} \ .$$

SOLUCIÓN:

a) De la misma definición de la representación digital de un número en una base cualquiera, tenemos

$$\begin{aligned} 100010001.100010001_2 &= 1\cdot 2^8+0\cdot 2^7+0\cdot 2^6+0\cdot 2^5 \\ &\quad +1\cdot 2^4+0\cdot 2^3+0\cdot 2^2+0\cdot 2^1 \\ &\quad +1\cdot 2^0 \\ &\quad +1\cdot 2^{-1}+0\cdot 2^{-2}+0\cdot 2^{-3}+0\cdot 2^{-4} \\ &\quad +1\cdot 2^{-5}+0\cdot 2^{-6}+0\cdot 2^{-7}+0\cdot 2^{-8} \\ &\quad +1\cdot 2^{-9} \\ &= 2^8+2^5+2^0+2^{-1}+2^{-5}+2^{-9} \\ &= 256+32+1+0.5+0.03125+0.001953125 \\ &= \underline{289.533203125}\ , \end{aligned}$$

$$\boxed{100010001.100010001_2 = 289.533203125}\ .$$

La parte entera de 42356.66_7 es

$$\begin{aligned} 42356_7 &= 4\cdot 7^4+2\cdot 7^3+3\cdot 7^2+5\cdot 7^1+6 \\ &= (((4\cdot 7+2)\cdot 7+3)\cdot 7+5)\cdot 7+6 \\ &= 10478\ ; \end{aligned}$$

y la parte fraccionaria

$$\begin{aligned} 0.66_7 &= 6\cdot 7^{-1}+6\cdot 7^{-2}=\frac{6\cdot 7+6}{7^2}=\frac{48}{49} \\ &= 0.\overline{979591836734693877551020408163265306122448}\ . \end{aligned}$$

Finalmente,

$$\boxed{42356.66_7 = 10478.\overline{979591836734693877551020408163265306122448}}\ .$$

Recordemos que las cifras en base 16 son: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A(=10), B(=11), C(=12), D(=13), E(=14) y F(=15). Así, de forma análoga,

$$\boxed{1F2D.124_{16} = 7981.712890625}\ .$$

b) Análogamente al problema anterior,

$$\begin{array}{lll} 4235 = 2117\cdot 2+\underline{1} & 2117 = 1058\cdot 2+\underline{1} & 1058 = 529\cdot 2+\underline{0} \\ 529 = 264\cdot 2+\underline{1} & 264 = 132\cdot 2+\underline{0} & 132 = 66\cdot 2+\underline{0} \\ 66 = 33\cdot 2+\underline{0} & 33 = 16\cdot 2+\underline{1} & 16 = 8\cdot 2+\underline{0} \\ 8 = 4\cdot 2+\underline{0} & 4 = 2\cdot 2+\underline{0} & 2 = \underline{1}\cdot 2+\underline{0} \end{array}$$

$$\begin{array}{lll} 0.66\cdot 2=\underline{1}.32 & 0.32\cdot 2=\underline{0}.64 & 0.64\cdot 2=\underline{1}.28\\ 0.28\cdot 2=\underline{0}.56 & 0.56\cdot 2=\underline{1}.12 & 1.12\cdot 2=\underline{0}.24\\ 0.24\cdot 2=\underline{0}.48 & 0.48\cdot 2=\underline{0}.96 & 0.96\cdot 2=\underline{1}.92\\ 0.92\cdot 2=\underline{1}.84 & 0.84\cdot 2=\underline{1}.68 & 0.68\cdot 2=\underline{1}.36\\ 0.36\cdot 2=\underline{0}.72 & 0.72\cdot 2=\underline{1}.44 & 0.44\cdot 2=\underline{0}.88\\ 0.88\cdot 2=\underline{1}.76 & 0.76\cdot 2=\underline{1}.52 & 0.52\cdot 2=\underline{1}.04\\ 0.04\cdot 2=\underline{0}.08 & 0.08\cdot 2=\underline{0}.16 & 0.16\cdot 2=\underline{0}.32\\ \multicolumn{3}{c}{\cdots} \end{array}$$

Resulta así,

$$\boxed{4235.66=1000010001011.1\overline{0101000111101011000}_2}\ .$$

$$\begin{array}{lll} 4235=605\cdot 7+\underline{0} & 605=86\cdot 7+\underline{3} & 86=12\cdot 7+\underline{2}\\ 12=\underline{1}\cdot 7+\underline{5} & & \end{array}\ ,$$

$$\begin{array}{lll} 0.66\cdot 7=\underline{4}.62 & 0.62\cdot 7=\underline{4}.34 & 0.34\cdot 7=\underline{2}.38\\ 0.38\cdot 7=\underline{2}.66 & \cdots & \end{array}\ .$$

Por tanto,

$$\boxed{4235.66=15230.\overline{4422}_7}\ .$$

Agrupando de cuatro en cuatro las cifras binarias, se obtiene

$$\boxed{4235.66=108B.A\overline{8F5C2}_{16}}\ .$$

Las 25 primeras cifras significativas de los números π y e son:

$$\begin{aligned} \pi &= 3.14159265358979323846264\ldots\ ,\\ e &= 2.71828182845904523536028\ldots\ , \end{aligned}$$

respectivamente.

Consideremos el paso del número π a base 7. Llamando $\pi^{(j)}$ al número π redondeado a j cifras decimales encontramos

$$\begin{aligned} \pi^{(4)} &= 3.1416=3.0664_7\pm\frac{1}{2}7^{-4}\ ,\\ \pi^{(5)} &= 3.14159=3.06637_7\pm\frac{1}{2}7^{-5}\ ; \end{aligned}$$

y, por tanto,

$$\begin{aligned} \pi &= 3.0664_7\pm(\frac{1}{2}7^{-4}+\frac{1}{2}10^{-4})\ ,\\ \pi &= 3.06637_7\pm(\frac{1}{2}7^{-5}+\frac{1}{2}10^{-5})\ . \end{aligned}$$

Tendremos también, por ejemplo, que

$$\boxed{\pi=3.141593\pm\tfrac{1}{2}10^{-6}=3.0663652_7\pm 7^{-7}}\ ,$$
$$\boxed{\pi=3.141592654\pm\tfrac{1}{2}10^{-9}=3.0663651432_7\pm 7^{-10}}\ .$$

El resto del apartado b) se deja para los lectores más pacientes.

c) Realizamos las operaciones aritméticas pedidas de forma análoga a como se llevan a cabo en base 10:

$$\begin{array}{r} 10001.1001_2 \\ 1001.1111_2 \\ \hline 11011.1000_2 \end{array}$$

$$\boxed{10001.1001_2 + 1001.1111_2 = 11011.1000_2} \ .$$

$$\begin{array}{r} 111101_2 \\ 11_2 \\ \hline 111101 \\ 111101 \\ \hline 10110111_2 \end{array}$$

$$\boxed{11.1101_2 \cdot 11_2 = 1011.0111_2} \ .$$

$$\begin{array}{r} 4567_9 \\ 44_9 \\ \hline 20501 \\ 20501 \\ \hline 225511_9 \end{array}$$

$$\boxed{45.67_9 \cdot 44_9 = 2255.11_9} \ .$$

$$\boxed{11.11_2/0.1111_2 = 100_2} \ .$$

$$\begin{array}{rl} 12212112_3 & \mid 22_3 \\ 00112 & \overline{20.1211_3} \\ \ \ 0201 & \\ \ \ 0101 & \\ \ \ 0022 & \\ \ \ 00 & \end{array}$$

$$\boxed{1221.2112_3/22_3 = 20.1211_3} \ .$$

$$\begin{array}{rl} A1_{16} & \mid F_{16} \\ 0B0 & \overline{A.B\ldots_{16}} \\ \ \ 0B0 & \\ \ \ \ \ \cdot\cdot & \end{array}$$

$$\boxed{A.1_{16}/F_{16} = A.\overline{B}_{16}} \ .$$

Problema 1.3 *En un ordenador IBM, cada celda de memoria está formada por 32 posiciones binarias (bits). Sabemos que todo número p se almacena en punto flotante hexadecimal (en base 16) de la forma siguiente: una vez escrito p en la forma $\pm p' \cdot 16^{(p"-64)}$ con $p" \geq 0$, $\frac{1}{16} \leq p' < 1$, se almacenan en los 32 bits y sucesivamente : el signo (0 si + , 1 si −) (1 bit), el exponente modificado $p"$ en base 2 (7 bits) y las primeras cifras no enteras de la mantisa p' en base 2 (24 bits, en precisión simple, y 56 bits, en precisión doble). Notamos que un número almacenado en precisión doble ocupa 2 celdas de memoria.*

a) Indicar cómo se almacenan en precisión simple 1.0, 1.1 *y* -17.0.

b) Expresar, en forma decimal, el máximo y mínimo número positivo que puede ser almacenado.

c) Calcular el error relativo máximo, en notación decimal, con que puede ser almacenado un número cualquiera en precisión simple y doble.

SOLUCIÓN:

a) El número $p = 1$ en la representación de punto flotante hexadecimal queda

$$1_{10} = 1_2 = \frac{1}{16} \cdot 16^1 = 2^{-4} \cdot 16^{65-64} \ ;$$

así,

$$p' = \frac{1}{16} = 2^{-4} = 0.0001_2 \ , \quad p" = 65 = 1000001_2 \ ,$$

y se almacena

$$\boxed{0 \mid 1000001 \mid 00010000000000000000000} \ .$$

Para $p = 1.1$, tenemos

$$1.1_{10} = 1.0\overline{0011}_2 \ , \quad 1.1 = \frac{1.1}{16} \cdot 16^1 = 0.0001\overline{0011}_2 \cdot 16^{65-64} \ ,$$

y se almacena

$$\boxed{0 \mid 1000001 \mid 00010001100110011001101} \ ,$$

redondeando adecuadamente.

Para $p = -17$:

$$-17_{10} = -10001_2, \ -17 = \frac{-17}{16^2} \cdot 16^2 = -0.00010001_2 \cdot 16^{66-64} \ .$$

Así, -17 se representa por

$$\boxed{1 \mid 1000010 \mid 00010001000000000000000} \ .$$

b) La representación del máximo número positivo corresponde a

$$\boxed{0 \mid 1111111 \mid 11111111111111111111111} \ ;$$

se trata, así, de

$$(1 - 2^{-24}) \cdot 16^{127-64} \simeq 16^{63} \simeq 7 \cdot 10^{75} \ .$$

La representación del mínimo es

$$\boxed{0 \mid 0000000 \mid 00010000000000000000000} \ ,$$

ya que $p' \geq \frac{1}{16}$. Por tanto, este número es

$$\frac{1}{16} \cdot 16^{-64} = 16^{-65} \simeq 6 \cdot 10^{-79} \ .$$

c) Suponiendo que se redondea al efectuar la representación finita de la mantisa de p, una cota del error absoluto en esta representación es $\frac{1}{2}2^{-24} \cdot 16^{p"-64}$, en precisión simple, y $\frac{1}{2}2^{-56} \cdot 16^{p"-64}$, en precisión doble.

Como $p" \geq \frac{1}{16}$, tenemos las siguientes cotas del error relativo:

$$\epsilon_r(p) = \frac{2^{-25} \cdot 16^{p"-64}}{16^{-1} \cdot 16^{p"-64}} = 2^{-21} \simeq 0.5 \cdot 10^{-6} \ ,$$

en precisión simple, y

$$\epsilon_r(p) = \frac{2^{-57} \cdot 16^{p"-64}}{16^{-1} \cdot 16^{p"-64}} = 2^{-53} \simeq 1.1 \cdot 10^{-16} \ ,$$

en precisión doble.

Problema 1.4 *En el ordenador CDC 3200, los números se almacenan en punto flotante, usando 48 posiciones binarias de la manera siguiente:*

1	11	36
s	*exponente*	*mantisa*

El número p se escribe en la forma $p = \pm p' \cdot 2^{p"}$ con $\frac{1}{2} \leq p' < 1$. La mantisa p' se almacena aproximada por sus 36 primeras cifras binarias no enteras, redondeadas adecuadamente. El exponente $p"$ ha de cumplir la acotación $\mid p" \mid < 2^{10}$ y se representa por

$$p" + 2^{10} \ (si \ \ p" \geq 0) \ , \quad p" + 2^{10} - 1 \ (si \ \ p" < 0) \ .$$

El bit correspondiente al signo (s) es 0 cuando $p \geq 0$; si $p < 0$, se escoge el complemento a 1 de toda la representación anterior (es decir, se cambian los unos por ceros y los ceros por unos; en particular, el bit del signo será 1).

a) Indicar cómo se representan los números siguientes:

$$1.0 \ , \quad 0.5 \ , \quad -0.0625 \ , \quad 0.1 \ .$$

b) Expresar, en forma decimal, el máximo y mínimo número positivo que puede ser almacenado.

c) ¿Cuántos números reales se pueden almacenar exactamente?

d) ¿Cuál es el error relativo maximal en el almacenamiento de los números?

SOLUCIÓN:

a) El número $p = 1$ en la representación de punto flotante binaria queda

$$1_{10} = 1_2 = \frac{1}{2} \cdot 2^1 \ ;$$

así,

$$p' = \frac{1}{2} = 2^{-1} = 0.1_2 \ , \quad p\text{"} = 1 = 1_2 \ ,$$

y se almacena

$$\boxed{0 \mid 10000000001 \mid 10000000000000000000000000000000000000} \ .$$

Para $p = 0.5$, tenemos

$$0.5 = \frac{1}{2} \cdot 2^0$$

y se almacena

$$\boxed{0 \mid 10000000000 \mid 10000000000000000000000000000000000000} \ .$$

Para $p = -0.0625$, consideramos primero su valor absoluto

$$0.0625_{10} = \frac{1}{16} = 0.1_2 \cdot 2^{-3} \ .$$

Así, 0.0625 se representa por

$$\boxed{0 \mid 01111111100 \mid 10000000000000000000000000000000000000} \ ;$$

por tanto, $p = -0.0625$ tendrá la representación complementaria a 1

$$\boxed{1 \mid 10000000011 \mid 01111111111111111111111111111111111111} \ .$$

Finalmente, para $p = 0.1$, tenemos

$$0.1 = 0.8 \cdot 2^{-3} = 0.\overline{1100}_2 \cdot 2^{-3}$$

y, redondeando convenientemente, se almacena

$$\boxed{0 \mid 01111111100 \mid 11001100110011001100110011001100110101} \ .$$

b) La representación del número positivo mayor corresponde a

$$\boxed{0 \mid 11111111111 \mid 11111111111111111111111111111111111111} \ ;$$

se trata, así, de

$$(1 - 2^{-36}) \cdot 2^{2^{10}-1} = (1 - 2^{-36}) \cdot 2^{1023} \simeq 9 \cdot 10^{307} \ .$$

La representación del menor es

$$\boxed{0 \mid 00000000000 \mid 10000000000000000000000000000000000000} \ ,$$

dado que $p' \geq \frac{1}{2}$. Por tanto, este número es

$$\frac{1}{2} \cdot 2^{1-2^{10}} = 2^{-1024} \simeq 6 \cdot 10^{-309} \ .$$

c) La cantidad de números que se pueden almacenar exactamente será como máximo 2^{47}, ya que cada uno de los 48 bits puede tomar los valores 0 y 1, excepto el primer bit correspondiente a la representación de la parte fraccionaria de la mantisa en base 2 que es forzosamente 1 para los números positivos y 0 para los negativos. Concretando, hay que considerar un factor 2 para el signo, un factor 2^{35} para las diferentes mantisas almacenadas exactamente y un factor $2^{11} - 1$ correspondiente a las posibilidades de almacenamiento de p" que son 2^{10} para los $p" \geq 0$ y $2^{10} - 1$ para los $p" < 0$.

Resulta, por tanto, la siguiente cantidad de números almacenados exactamente

$$2 \cdot 2^{35} \cdot (2^{11} - 1) \simeq 1.4 \cdot 10^{14} \ .$$

d) Suponiendo que se redondea al llevar a cabo la representación finita de la mantisa de p, una cota del error absoluto en esta representación es $\frac{1}{2} 2^{-36} \cdot 2^{p"}$.

Como $p' \geq \frac{1}{2}$, tenemos la cota del error relativo

$$\epsilon_r(p) = \frac{2^{-37} \cdot 2^{p"}}{2^{-1} \cdot 2^{p"}} = 2^{-36} \simeq 0.3 \cdot 10^{-10} \ .$$

Problema 1.5 *a) Determinar el error máximo para* $y = x_1 x_2^2$, *si*

$$x_1 = 2.0 \pm 0.1 \ , \quad x_2 = 3.0 \pm 0.2 \ .$$

i) Exactamente, operando con intervalos.

ii) Utilizando las fórmulas (aproximadas) del error maximal en las operaciones aritméticas.

iii) Calculando primeramente el error relativo, usando que el error relativo es aproximadamente el error absoluto del logaritmo.

b) Calcular el error estándar con los mismos datos de a), suponiendo que las cotas para los errores de x_1, x_2 *son, de hecho, desviaciones estándar.*

SOLUCIÓN:

a) i) $x_1 = 2.0 \pm 0.1$ y $x_2 = 3.0 \pm 0.2$ equivalen a

$$\begin{aligned} \bar{x}_1 - \epsilon_1 \leq x_1 \leq \bar{x}_1 + \epsilon_1 \qquad & (\bar{x}_1 = 2.0 \ , \ \epsilon_1 = 0.1) \ , \\ \bar{x}_2 - \epsilon_2 \leq x_2 \leq \bar{x}_2 + \epsilon_2 \qquad & (\bar{x}_2 = 3.0, \ \epsilon_2 = 0.2) \ . \end{aligned}$$

Como $\bar{x}_1 \pm \epsilon_1$ y $\bar{x}_2 \pm \epsilon_2$ son positivos,

$$(\bar{x}_1 - \epsilon_1)(\bar{x}_2 - \epsilon_2)^2 \leq x_1 x_2^2 \leq (\bar{x}_1 + \epsilon_1)(\bar{x}_2 + \epsilon_2)^2 \ . \quad (*)$$

Sustituyendo,

$$14.896 = 1.8 \cdot 2.8^2 \leq x_1 x_2^2 \leq 2.1 \cdot 3.2^2 = 21.504 \ ;$$

por tanto,

$$-3.104 \leq x_1 x_2^2 - \bar{x}_1 \bar{x}_2^2 \leq 3.504 \ .$$

Escogiendo la cota mayor,

$$\boxed{x_1 x_2^2 = 18 \pm 3.504} \ .$$

ii) Usaremos $\epsilon_a(y_1 + y_2) = \epsilon_a(y_1) + \epsilon_a(y_2)$ y $\epsilon_r(y_1 \cdot y_2) \simeq \epsilon_r(y_1) + \epsilon_r(y_2)$.
En nuestro caso,

$$\epsilon_r(x_1 x_2^2) = \epsilon_r(x_1 x_2 x_2) \simeq \epsilon_r(x_1) + 2\epsilon_r(x_2) \quad (**) \ ,$$

$$\epsilon_r(x_1) = \frac{\epsilon_a(x_1)}{x_1} \simeq \frac{\epsilon_a(x_1)}{\bar{x}_1} = \frac{1}{20} \ , \quad \epsilon_r(x_2) = \frac{\epsilon_a(x_2)}{x_2} \simeq \frac{\epsilon_a(x_2)}{\bar{x}_2} = \frac{2}{30} \ .$$

Así,

$$\epsilon_r(x_1 x_2^2) \simeq \frac{1}{20} + \frac{4}{30} = \frac{11}{60} \ , \quad \epsilon_a(x_1 x_2^2) \simeq 18\frac{11}{60} = 3.3$$

y, finalmente,

$$\boxed{x_1 x_2^2 = 18 \pm 3.3} \ .$$

Otra manera de llegar a este resultado es operando en $(*)$

$$\bar{x}_1 \bar{x}_2^2 - \bar{x}_2^2 \epsilon_1 - 2\bar{x}_1 \bar{x}_2 \epsilon_2 + 2\bar{x}_2 \epsilon_1 \epsilon_2 + \bar{x}_1 \epsilon_2^2 - \epsilon_1 \epsilon_2^2 \leq x_1 x_2^2 \leq$$
$$\bar{x}_1 \bar{x}_2^2 + \bar{x}_2^2 \epsilon_1 + 2\bar{x}_1 \bar{x}_2 \epsilon_2 + 2\bar{x}_2 \epsilon_1 \epsilon_2 + \bar{x}_1 \epsilon_2^2 + \epsilon_1 \epsilon_2^2$$

y despreciando los términos $\bar{x}_2 \epsilon_1 \epsilon_2$, $\bar{x}_1 \epsilon_2^2$ y $\epsilon_1 \epsilon_2^2$ en los cuales aparecen productos de errores de los datos (es decir, que contienen más de un ϵ),

$$\boxed{x_1 x_2^2 = \bar{x}_1 \bar{x}_2^2 \pm (\bar{x}_2^2 \epsilon_1 + 2\bar{x}_1 \bar{x}_2 \epsilon_2) = 18 \pm 3.3} \ .$$

iii) Usando que $\epsilon_r(y) \simeq \epsilon_a(\ln y)$, tenemos

$$\epsilon_r(x_1 x_2^2) \simeq \epsilon_a(\ln x_1 + 2 \ln x_2) \simeq \epsilon_r(x_1) + 2\epsilon_r(x_2) \ ,$$

como en $(**)$.

b) Usando la fórmula de propagación del error estándar, con $\sigma_1 = 0.1$, $\sigma_2 = 0.2$, tenemos

$$\begin{aligned} \sigma(y) &\simeq [(x_2^2)^2 \sigma_1^2 + (2x_1 x_2)^2 \sigma_2^2]^{\frac{1}{2}} \\ &= [81\sigma_1^2 + 144\sigma_2^2]^{\frac{1}{2}} \simeq 2.6 \ . \end{aligned}$$

Problema 1.6 *Dado el sistema de ecuaciones lineales*

$$\left.\begin{array}{rcrcl} 3x & + & ay & = & 10 \\ 5x & + & by & = & 20 \end{array}\right\} ,$$

donde $a = 2.100 \pm 5 \cdot 10^{-4}$ *y* $b = 3.300 \pm 5 \cdot 10^{-4}$*, ¿con qué exactitud puede ser determinado* $x + y$*?*

SOLUCIÓN:

Las componentes de la solución del sistema serán

$$x = \frac{10b - 20a}{3b - 5a} \ , \quad y = \frac{10}{3b - 5a} \ ;$$

y la suma

$$x + y = 10\frac{b - 2a + 1}{3b - 5a} \ .$$

Damos a continuación los valores aproximados de diversos resultados intermedios en el cálculo de $x + y$, así como las cotas (aproximadas) de los errores respectivos:

$$\begin{array}{rl} a = 2.1 & \epsilon_a(a) = 5 \cdot 10^{-4} \equiv \epsilon \ , \\ b = 3.3 & \epsilon_a(b) = 5 \cdot 10^{-4} \equiv \epsilon \ , \\ N = b - 2a + 1 = 0.1 & \epsilon_a(N) = \epsilon_a(b) + 2\epsilon_a(a) = 3\epsilon \ , \\ D = 3b - 5a = -0.6 & \epsilon_a(D) = 3\epsilon_a(b) + 5\epsilon_a(a) = 8\epsilon \ , \\ x + y = 10\dfrac{N}{D} = -\dfrac{10}{6} & \epsilon_a(x + y) = 10\dfrac{N\epsilon_a(D) + |D|\,\epsilon_a(N)}{D^2} \\ & \quad = \dfrac{650}{9}\epsilon \simeq 4 \cdot 10^{-2} \ . \end{array}$$

Resulta así,

$$\boxed{x + y = -\tfrac{5}{3} \pm 4 \cdot 10^{-2}} \ .$$

Problema 1.7 *Tenemos el sistema de ecuaciones lineales*

$$\left.\begin{array}{rcrcl} x & + & ay & = & 5 \\ bx & + & 2y & = & d \end{array}\right\} ,$$

donde $a = 1.00 \pm 5 \cdot 10^{-3}$*,* $b = 1/a$ *y* $d = b - a$*. ¿Con qué exactitud podemos determinar* xy*?*

SOLUCIÓN:

Si interpretamos el enunciado de forma que b se calcula a partir de a y d a partir de b y a, tendremos que expresar primero el producto xy de las componentes de la solución del sistema en función de a, b y d y encontrar después los errores de aquél a partir de los errores de a, b y d (entendiendo que los errores de b y d son los propagados del error de a).

Las componentes de la solución del sistema serán

$$x = \frac{10 - ad}{2 - ab} \ , \quad y = \frac{d - 5b}{2 - ab} \ ;$$

y el producto,

$$xy = \frac{(10 - ad)(d - 5b)}{(2 - ab)^2} \ .$$

Damos a continuación los valores aproximados de diversos resultados intermedios en el cálculo de xy, así como las cotas (aproximadas) de los errores de las respectivas expresiones:

$$\begin{aligned}
a = 1 \qquad & \epsilon_a(a) = \epsilon \equiv 5 \cdot 10^{-3} \ , \\
b = 1 \qquad & \epsilon_a(b) = \frac{1}{a} = \frac{1}{a^2}\epsilon_a(a) = \epsilon \ , \\
d = 0 \qquad & \epsilon_a(d = b - a) = \epsilon_a(a) + \epsilon_a(b) = 2\epsilon \ , \\
2 - ab = 1 \qquad & \epsilon(2 - ab) = a\epsilon_a(b) + b\epsilon_a(a) = 2\epsilon \ , \\
D = (2 - ab)^2 = 1 \qquad & \epsilon_a(D) = 2(2 - ab)\epsilon_a(2 - ab) = 4\epsilon \ , \\
N_1 = 10 - ad = 10 \qquad & \epsilon_a(N_1) = a\epsilon_a(d) + d\epsilon_a(a) = 2\epsilon \ , \\
N_2 = d - 5b = -5 \qquad & \epsilon_a(N_2) = \epsilon_a(d) + 5\epsilon_a(b) = 7\epsilon \ , \\
N = N_1 N_2 = -50 \qquad & \epsilon_a(N) = |N_2|\epsilon_a(N_1) + N_1\epsilon_a(N_2) = 80\epsilon \ , \\
xy = \frac{N}{D} = -50 \qquad & \epsilon_a(xy) = \frac{D\epsilon_a(N) + |N|\,\epsilon_a(D)}{D^2} = 280\epsilon \ .
\end{aligned}$$

Así pues, en el caso planteado, sustituyendo el valor de $\epsilon = 5 \cdot 10^{-3}$, se obtiene

$$\boxed{xy = -50 \pm 1.4} \ .$$

Problema 1.8 *Queremos calcular* $a = (7 - 4\sqrt{3})^4$*, utilizando el valor aproximado* 1.73205 *para* $\sqrt{3}$*. Escoger, entre las fórmulas equivalentes siguientes, la más adecuada desde un punto de vista numérico:*

$$\frac{1}{(7 + 4\sqrt{3})} \ , \quad (97 - 56\sqrt{3})^2 \ , \quad \frac{1}{(97 + 56\sqrt{3})^2} \ ,$$

$$18817 - 10864\sqrt{3} \ , \quad \frac{1}{18817 + 10864\sqrt{3}} \ .$$

SOLUCIÓN:

Consideramos las siguientes funciones:

$$\begin{aligned} f_1(x) &= (7-4x)^4 \ , \\ f_2(x) &= (7+4x)^{-4} \ , \\ f_3(x) &= (97-56x)^2 \ , \\ f_4(x) &= (97+56x)^{-2} \ , \\ f_5(x) &= 18817-10864x \ , \\ f_6(x) &= (18817+10864x)^{-1} \ . \end{aligned}$$

Es trivial comprobar que todas las funciones valen a en $x=\sqrt{3}$,

$$a = f_i(\sqrt{3}) \ (i = 1 \div 6) \ .$$

Cuando procedemos a evaluar $f_i(\sqrt{3})$ no disponemos del valor exacto $\sqrt{3}$, sino sólo de la aproximación $x = 1.73205$; tenemos que encontrar aquella f_i que produzca el mínimo error en el cálculo de a. Con este fin, utilizaremos la fórmula aproximada de propagación del error maximal en cada caso; así tenemos:

$$\begin{aligned} \epsilon_a(f_1(x)) &\simeq 16(7-4x)^3\epsilon_a(x) \simeq 0.005922\epsilon_a(x) \ , \\ \epsilon_a(f_2(x)) &\simeq 16(7+4x)^{-5}\epsilon_a(x) \simeq 0.00003052\epsilon_a(x) \ , \\ \epsilon_a(f_3(x)) &\simeq 112(97-56x)\epsilon_a(x) \simeq 0.5824\epsilon_a(x) \ , \\ \epsilon_a(f_4(x)) &\simeq 112(97+56x)^{-3}\epsilon_a(x) \simeq 0.00001534\epsilon_a(x) \ , \\ \epsilon_a(f_5(x)) &\simeq 10864\epsilon_a(x) \ , \\ \epsilon_a(f_6(x)) &\simeq 10864(18817+10864x)^{-2}\epsilon_a(x) \simeq 0.00000767\epsilon_a(x) \ . \end{aligned}$$

Observamos que la mejor fórmula desde un punto de vista numérico es la última, y que la peor es la dada por f_5. Calculando a con ambas, tenemos:

$$f_6(\sqrt{3}) \simeq 0.0000265717 \ , \quad f_5(\sqrt{3}) \simeq 0.000976562(!) \ .$$

El estudio anterior asegura que las 5 primeras cifras significativas del primer valor son correctas; en cambio, no asegura ninguna cifra significativa correcta del segundo valor obtenido, tal como ocurre.

Hay que hacer notar que, en operaciones similares a la efectuada en la evaluación de $f_5(\sqrt{3})$ (donde se restan números muy próximos), el error relativo puede ser muy grande. En estos casos es preciso sustituir la fórmula utilizada por otra equivalente pero mejor desde un punto de vista numérico.

Problema 1.9 *Si, en el cálculo de* $\ln(x-\sqrt{x^2-1})$ *para* $x = 30$, *la raíz cuadrada se obtiene de una tabla que da 6 cifras decimales correctas con redondeo, ¿cuál será el error absoluto en el resultado? Obtener una expresión equivalente numéricamente mejor , y acotar el error absoluto al usarla en las mismas condiciones.*

SOLUCIÓN:

Llamaremos z a la cantidad afectada de error en la expresión; tenemos así que calcular el error en $f(z) = \ln(30 - z)$, donde $z = \sqrt{899}$ tiene una cota de error $\epsilon_a(z) = \frac{1}{2}10^{-6}$. La fórmula de propagación del error da la siguiente cota:

$$\epsilon_a(f(z)) \simeq |\, f'(z)\,|\, \epsilon_a(z) = \frac{1}{30 - \sqrt{899}}\epsilon_a(z) \simeq \frac{\frac{1}{2}10^{-6}}{0.016671} \simeq 3 \cdot 10^{-5} \ .$$

Una manera de obtener una expresión equivalente, en la cual no se produce el problema de cancelación observado, consiste en volver a escribir el argumento del logaritmo en la forma

$$(x - \sqrt{x^2 - 1})\frac{x + \sqrt{x^2 - 1}}{x + \sqrt{x^2 - 1}} = \frac{1}{x + \sqrt{x^2 - 1}} \ ;$$

entonces, la expresión equivalente será

$$\ln(x - \sqrt{x^2 - 1}) = \ln(\frac{1}{x + \sqrt{x^2 - 1}}) = -\ln(x + \sqrt{x^2 - 1}) \ .$$

Para $x = 30$, calculamos $g(z) = -\ln(30 + z)$ y el error estará acotado por

$$\epsilon_a(g(z)) \simeq |\, g'(z)\,|\, \epsilon_a(z) = \frac{1}{30 + \sqrt{899}}\epsilon_a(z) \simeq \frac{\frac{1}{2}10^{-6}}{59.983329} \simeq 8.4 \cdot 10^{-9} \ .$$

Destaquemos que el factor de expansión del error de z se reduce en un factor próximo a 3600 al usar la segunda fórmula y no la primera, y que esta reducción será aún mayor si x aumenta su valor.

Problema 1.10 *Trabajando con 5 cifras decimales, calcular*

$$\sqrt[k]{2.15283} - \sqrt[k]{2.15263} \ (k = 2, 3, 4) \ .$$

a) Directamente.

b) Usando fórmulas equivalentes, mejores desde un punto de vista numérico. (Indicación: Considérese la división de polinomios $(a^k - b^k)/(a - b)$).

c) Comparar los resultados y comentarlos.

SOLUCIÓN:

a) En la tabla siguiente se dan los resultados de los cálculos hechos directamente, usando sólo 6 decimales:

Hay que destacar que sólo se obtiene una cifra significativa en todos los casos; hay, por tanto, un problema de cancelación de términos.

	$k=2$	$k=3$	$k=4$
$\sqrt[k]{2.15283}$	1.46725	1.29123	1.21130
$\sqrt[k]{2.15263}$	1.46718	1.29119	1.21127
$\sqrt[k]{2.15283}-\sqrt[k]{2.15263}$	0.00007	0.00004	0.00003

b) Para encontrar fórmulas equivalentes, consideramos la división del polinomio a^k-b^k entre $a-b$, hecha según la regla de Ruffini,

$$\begin{array}{cccccc|c} & 1 & 0 & 0 & \cdots & 0 & -b^k \\ b & & b & b^2 & \cdots & b^{k-1} & b^k \\ \hline & 1 & b & b^2 & \cdots & b^{k-1} & 0 \end{array}$$

así,

$$a^k-b^k=(a^{k-1}+a^{k-2}b+a^{k-3}b^2+\cdots+ab^{k-2}+b^{k-1})(a-b)\ .$$

Por tanto, escribiendo en cada caso $a=\sqrt[k]{A}$ y $b=\sqrt[k]{B}$, tenemos

$$\begin{aligned} \sqrt{A}-\sqrt{B} &= \frac{A-B}{\sqrt{A}+\sqrt{B}}\ , \\ \sqrt[3]{A}-\sqrt[3]{B} &= \frac{A-B}{(\sqrt[3]{A})^2+\sqrt[3]{A}\sqrt[3]{B}+(\sqrt[3]{B})^2}\ , \\ \sqrt[4]{A}-\sqrt[4]{B} &= \frac{A-B}{(\sqrt[4]{A})^3+(\sqrt[4]{A})^2\sqrt[4]{B}+\sqrt[4]{A}(\sqrt[4]{B})^2+(\sqrt[4]{B})^3}\ ; \end{aligned}$$

usando estas fórmulas con el numerador exacto $A-B=2\cdot10^{-4}$, se encuentra

$$\begin{aligned} \sqrt[k]{2.15283}-\sqrt[k]{2.15263} &= 6.81563\cdot10^{-5}\ \ (k=2)\ , \\ & \quad\ 3.99866\cdot10^{-5}\ \ (k=3)\ , \\ & \quad\ 2.81340\cdot10^{-5}\ \ (k=4)\ . \end{aligned}$$

c) $\sqrt[k]{2.15283}$ y $\sqrt[k]{2.15263}$ están calculados con 5 cifras decimales; por tanto, una cota del error al efectuar la resta en a) es $\frac{1}{2}10^{-5}+\frac{1}{2}10^{-5}=10^{-5}$. Así, los resultados de a) con las cotas de error correspondientes son:

$$(7\pm1)10^{-5}\ ,(4\pm1)10^{-5}\ ,(3\pm1)10^{-5}\ .$$

Si usamos, en cambio, las fórmulas encontradas en b), por ejemplo con $k=2$, y llamamos $D=A-B$, $a=\sqrt{A}$ y $b=\sqrt{B}$, obtenemos la cota

$$\epsilon_a(\frac{D}{a+b})\simeq\frac{D}{(a+b)^2}(\epsilon_a(a)+\epsilon_a(b))=\frac{2\cdot10^{-4}}{2.93^2}10^{-5}\simeq0.2\cdot10^{-9}\ .$$

Esta acotación asegura que podemos considerar correctas las 9 primeras cifras decimales del resultado del cálculo pedido

$$\boxed{\sqrt{2.15283}-\sqrt{2.15263}=6.8156\cdot10^{-5}}\ .$$

Llevando a cabo cálculos análogos, se ve que también se dispone de la misma exactitud para los otros valores de k:

$$\boxed{\sqrt[3]{2.15283} - \sqrt[3]{2.15263} = 3.9987 \cdot 10^{-5}} \ ,$$

$$\boxed{\sqrt[4]{2.15283} - \sqrt[4]{2.15263} = 2.8134 \cdot 10^{-5}} \ .$$

Obsérvese el considerable aumento de 4 cifras significativas que se obtiene con el uso de las fórmulas de b) en lugar de las de a).

Problema 1.11 *Dar expresiones equivalentes para las fórmulas siguientes que sean mejores desde un punto de vista numérico, cuando ϵ es mucho menor que x:*

$$\begin{aligned}
&a) \quad \sqrt[k]{x+\epsilon} - \sqrt[k]{x} \ , \\
&b) \quad \frac{1}{\sqrt{x-\epsilon}} - \frac{1}{\sqrt{x}} \ , \\
&c) \quad \frac{1}{x+\epsilon} + \frac{1}{x-\epsilon} - \frac{2}{x} \ , \\
&d) \quad \operatorname{sen}(x+\epsilon) - \operatorname{sen}(x) \ , \\
&e) \quad \cos(x+\epsilon) - \cos(x) \ , \\
&f) \quad \tan(x+\epsilon) - \tan(x) \ , \\
&g) \quad \int_x^{x+\epsilon} \frac{dt}{t} \ .
\end{aligned}$$

SOLUCIÓN:

a) Usando la técnica del ejercicio anterior, tenemos

$$\sqrt[k]{x+\epsilon} - \sqrt[k]{x} = \frac{\epsilon}{(\sqrt[k]{x+\epsilon})^{k-1} + (\sqrt[k]{x+\epsilon})^{k-2}\sqrt[k]{x} + \cdots + \sqrt[k]{x+\epsilon}(\sqrt[k]{x})^{k-2} + (\sqrt[k]{x})^{k-1}} \ .$$

b) Pasando a común denominador, tenemos

$$\begin{aligned}
\frac{1}{\sqrt{x-\epsilon}} - \frac{1}{\sqrt{x}} &= \frac{\sqrt{x} - \sqrt{x-\epsilon}}{\sqrt{x-\epsilon}\sqrt{x}} = \frac{(\sqrt{x} - \sqrt{x-\epsilon})(\sqrt{x-\epsilon} + \sqrt{x})}{(\sqrt{x}\sqrt{x-\epsilon})(\sqrt{x-\epsilon} + \sqrt{x})} \\
&= \frac{\epsilon}{(\sqrt{x}\sqrt{x-\epsilon})(\sqrt{x-\epsilon} + \sqrt{x})} \ .
\end{aligned}$$

c) Pasando también a común denominador, resulta

$$\begin{aligned}
\frac{1}{x+\epsilon} + \frac{1}{x-\epsilon} - \frac{2}{x} &= \frac{(x-\epsilon)x + (x+\epsilon)x - 2(x+\epsilon)(x-\epsilon)}{(x+\epsilon)(x-\epsilon)x} \\
&= \frac{2\epsilon^2}{(x^2-\epsilon^2)x} \ .
\end{aligned}$$

d)

$$\begin{aligned}\operatorname{sen}(x+\epsilon)-\operatorname{sen}(x) &= 2\operatorname{sen}(\frac{x+\epsilon-x}{2})\cos(\frac{x+\epsilon+x}{2}) \\ &= 2\operatorname{sen}(\frac{\epsilon}{2})\cos(x+\frac{\epsilon}{2})\ .\end{aligned}$$

e)

$$\begin{aligned}\cos(x+\epsilon)-\cos(x) &= -2\operatorname{sen}(\frac{x+\epsilon-x}{2})\operatorname{sen}(\frac{x+\epsilon+x}{2}) \\ &= -2\operatorname{sen}(\frac{\epsilon}{2})\operatorname{sen}(x+\frac{\epsilon}{2})\ .\end{aligned}$$

f)

$$\begin{aligned}\tan(x+\epsilon)-\tan(x) &= \frac{\operatorname{sen}(x+\epsilon)}{\cos(x+\epsilon)}-\frac{\operatorname{sen}(x)}{\cos(x)} \\ &= \frac{\operatorname{sen}(x+\epsilon)\cos(x)-\operatorname{sen}x\cos(x+\epsilon)}{\cos(x+\epsilon)\cos(x)}=\frac{\operatorname{sen}(\epsilon)}{\cos(x+\epsilon)\cos(x)}\ .\end{aligned}$$

g)

$$\int_x^{x+\epsilon}\frac{dt}{t}=\ln(x+\epsilon)-\ln(x)=\ln(\frac{x+\epsilon}{x})\ .$$

Problema 1.12 *Calcular, acotando el error propagado, la aceleración debida a la gravedad en la superficie del Sol (g_S), dada por la ley de la gravitación universal de Newton y partiendo de los siguientes datos aproximados, que supondremos redondeados: masa del Sol, $m_S = 1.90 \cdot 10^{30}\,\mathrm{Kg}$; masa de la Tierra, $m_T = 5.98 \cdot 10^{24}\,\mathrm{Kg}$; radio solar ($r_S$) igual a 112 radios terrestres (r_T), y la gravedad en la superficie terrestre, $g_T = 9.81\,\mathrm{m/s^2}$.*

SOLUCIÓN:

Por la ley de la gravitación universal de Newton, se cumple que

$$g_S = K\frac{m_S}{r_S^2}\ ,\quad g_T = K\frac{m_T}{r_T^2}\ ;$$

donde K es la constante de la gravitación universal, que no la tenemos como dato del problema; así que, despejándola de la segunda igualdad y sustituyéndola en la primera, tenemos la fórmula para el cálculo de g_S

$$g_S = \frac{m_S \cdot r_T^2}{m_T \cdot r_S^2}\ g_T\ .$$

Considerando las variables x_i $(i = 1 \div 4)$ tales que

$$m_S = 10^{30}x_1 \ , \quad m_T = 10^{24}x_2 \ , \quad r_S = x_3 r_T \ , \quad g_T = x_4 \ ;$$

entonces

$$g_S = \frac{x_1 x_4}{x_2 x_3^2}\ 10^6 \ .$$

Usando los valores aproximados

$$x_1 = 1.90\,\mathrm{Kg} \ , \quad x_2 = 5.98\,\mathrm{Kg} \ , \quad x_3 = 112 \ , \quad x_4 = 9.81\,\mathrm{m/s^2} \ ,$$

se obtiene la aproximación

$$g_S \simeq 248.47\,\mathrm{m/s^2} \ .$$

La fórmula de propagación del error maximal nos proporciona de forma aproximada una cota del error absoluto de g_S

$$\epsilon_a(g_S) \simeq \sum_{i=1}^{4} \left| \frac{\partial g_S}{\partial x_i}(x_1, x_2, x_3, x_4) \right| \epsilon_a(x_i)$$

$$= 10^6 \left(\frac{x_4}{x_2 x_3^2}\epsilon_a(x_1) + \frac{x_4 x_1}{x_2^2 x_3^2}\epsilon_a(x_2) + 2\frac{x_4 x_1}{x_2 x_3^3}\epsilon_a(x_3) + \frac{x_1}{x_2 x_3^2}\epsilon_a(x_4) \right) 10^6 \ .$$

Las cotas $\epsilon_a(x_i)$ son las acotaciones en el redondeo a las cifras dadas de los valores correspondientes de x_i: $\epsilon_a(x_1) = \epsilon_a(x_2) = \epsilon_a(x_4) = \frac{1}{2}10^{-2}$, $\epsilon_a(x_3) = \frac{1}{2}$. Sustituyéndolas en la fórmula anterior, resulta

$$\epsilon_a(g_S) \simeq 3.2\,\mathrm{m/s^2} \ .$$

El trabajo de cálculo es mucho más sencillo si se utilizan los errores relativos en vez de los absolutos. Los errores relativos de los datos están acotados por:

$$\begin{aligned} \epsilon_r(m_S) &= \frac{1}{2}\frac{10^{-2}}{1.9} \ , \quad \epsilon_r(m_T) = \frac{1}{2}\frac{10^{-2}}{5.8} \ , \\ \epsilon_r(g_T) &= \frac{1}{2}\frac{10^{-2}}{9.81} \ , \quad \epsilon_r(\frac{r_S}{r_T}) = \frac{1}{2}\frac{1}{112} \ . \end{aligned}$$

Así, directamente,

$$\begin{aligned} \epsilon_r(g_S) &\simeq \epsilon_r(m_S) + \epsilon_r(g_T) + \epsilon_r(m_T) + 2\epsilon_r(\frac{r_S}{r_T}) \\ &= \frac{1}{2}10^{-2}(\frac{1}{1.9} + \frac{1}{9.81} + \frac{1}{5.98}) + \frac{1}{112} \simeq 1.3 \cdot 10^{-2} \ . \end{aligned}$$

El error absoluto está, así, acotado por

$$\epsilon_a(g_S) = g_S \epsilon_r(g_S) \simeq 3.2\,\mathrm{m/s^2} \ ,$$

como antes.

Finalmente,

$$\boxed{g_S = (248.47 \pm 3.2)\,\mathrm{m/s^2}} \ .$$

Problema 1.13 *Consideremos una barra de longitud ℓ y sección rectangular, de anchura a y altura b, empotrada por uno de sus extremos. Si en el extremo libre se aplica una fuerza F perpendicular a la barra, la flexión s experimentada viene dada por la expresión*

$$s = \frac{4}{E}\,\frac{\ell^3}{ab^3}\,F \ ,$$

donde E es una constante que depende sólo del material, llamada módulo de Young.

Sabiendo que una fuerza de 140 Kp, *aplicada sobre una barra de hierro de* 125 cm *de longitud y sección cuadrada de* 2.5 cm *de lado, le produce una flexión de* 1.71 mm, *calcular el módulo de Young del hierro y acotar el error propagado de los errores de los datos (suponiendo que éstos tienen sólo el error de aproximación por corte a las cifras dadas).*

SOLUCIÓN:

El valor aproximado E del módulo de Young se encuentra despejando primero E de la expresión dada para la flexión y sustituyendo después los valores aproximados de los datos

$$E = \frac{4}{s}\frac{\ell^3}{ab^3}F = 16374.27 \ .$$

El error en E se encuentra cómodamente trabajando con los errores relativos; así,

$$\epsilon_r(E) = \epsilon_r(s) + 3\epsilon_r(\ell) + \epsilon_r(a) + 3\epsilon_r(b) + \epsilon_r(F) \ ;$$

con $\epsilon_r(s) = \frac{1}{171}$, $\epsilon_r(\ell) = \frac{1}{125}$, $\epsilon_r(a) = \epsilon_r(b) = \frac{1}{25}$, $\epsilon_r(F) = \frac{1}{140}$.

De donde,

$$\epsilon_r(E) = \frac{1}{171} + \frac{3}{125} + \frac{1}{25} + \frac{3}{25} + \frac{1}{140} = 1.97 \cdot 10^{-1};$$

por tanto,

$$\epsilon_a(E) = E\epsilon_r(E) = 3226 \ .$$

El módulo de Young del hierro, con la cota de error encontrada, viene dado por

$$\boxed{E = (1.6374 \pm 0.3226)10^4} \ .$$

Nótese que únicamente la primera cifra significativa se puede considerar correcta.

Problema 1.14 *Si usamos un ordenador que comete errores relativos acotados por ϵ, en la representación de los números y en las operaciones aritméticas, y por 5ϵ en el cálculo de logaritmos, acotar el error cometido en los cálculos de*

$$\begin{aligned} &a) \quad \sum_{i=1}^{n} a_i \ , \\ &b) \quad \prod_{i=1}^{n} a_i \ , \\ &c) \quad \sum_{i=1}^{n} x_i y_i \ , \\ &d) \quad \sum_{i=1}^{n} x_i \ln x_i \ , \\ &e) \quad \sum_{i=0}^{n} a_i x^i \ , \end{aligned}$$

usando diferentes algoritmos y tratando de decidir (si se puede) cuál sería la mejor manera de realizarlos.

SOLUCIÓN:

El enunciado dice que un número x se almacena como

$$\mathrm{fl}(x) = x(1+\delta) \ ,$$

con $|\delta| \leq \epsilon$; análogamente, $x+y$, xy, $\ln x$ se almacenan respectivamente como

$$\mathrm{fl}(x+y) = (x+y)(1+\delta_+) \ , \mathrm{fl}(xy) = xy(1+\delta_\cdot) \ , \mathrm{fl}(\ln x) = (\ln x)(1+\delta_{\ln}) \ ,$$

con $|\delta_+| \leq \epsilon$, $|\delta_\cdot| \leq \epsilon$ y $|\delta_{\ln}| \leq 5\epsilon$.

a) Para calcular $a_1 + a_2 + a_3$, denotamos $A_i = \mathrm{fl}(a_i) = a_i(1+\delta_i)$ $(i = 1,2,3)$ y encontramos

$$S_2 = \mathrm{fl}(A_1 + A_2) = (A_1 + A_2)(1+\delta_{+2}) \ , \quad S_3 = (S_2 + A_3)(1+\delta_{+3}) \ .$$

El valor calculado de la suma será

$$\mathrm{fl}(a_1 + a_2 + a_3) = \{[a_1(1+\delta_1) + a_2(1+\delta_2)](1+\delta_{+2}) + a_3(1+\delta_3)\}(1+\delta_{+3}) \ ,$$

donde los valores absolutos de δ con cualquier subíndice son menores que ϵ.

Despreciando los términos que darían una contribución no lineal en los diferentes valores de δ, el error en el cálculo se puede acotar por

$$\begin{aligned} &|S_3 - (a_1 + a_2 + a_3)| \\ &\quad \leq \ |a_1\delta_1 + a_2\delta_2 + a_3\delta_3 + (a_1 + a_2)\delta_{+2} + (a_1 + a_2 + a_3)\delta_{+3}| \\ &\quad \leq \ (|a_1| + |a_2| + |a_3| + |a_1 + a_2| + |a_1 + a_2 + a_3|)\epsilon \ , \end{aligned}$$

que también se puede acotar por $(3\,|a_1|\,+3\,|a_2|\,+2\,|a_3\,|)\epsilon$.

En el caso general, acotando los valores absolutos de δ_{+i} por ϵ, queda

$$\begin{aligned}|\,S_n - \sum_{i=1}^{n} a_i\,| \leq & |\,a_1\delta_1 + a_2\delta_2 + \cdots + a_n\delta_n \\ & +(a_1+a_2)\delta_{+2} + (a_1+a_2+a_3)\delta_{+3} + \cdots + (a_1+a_2+\cdots+a_n)\delta_{+n}\,| \\ \leq & \sum_{i=1}^{n} |\,a_i\,||\,\delta_i\,| \\ & +[(n-1)\,|a_1|\,+(n-1)\,|a_2|\,+(n-2)\,|a_3|\,+\cdots+\,|a_n\,|]\epsilon\ ;\ (*)\end{aligned}$$

acotando como antes $|\,\delta_i\,| \leq \epsilon$, tenemos la cota del error absoluto

$$\epsilon_a\left(\sum_{i=1}^{n} a_i\right) = [n\,|a_1|\,+n\,|a_2|\,+(n-1)\,|a_3|\,+\cdots+2\,|a_n\,|]\epsilon\ .$$

En consecuencia, es aconsejable realizar las sumas de los números empezando por los términos de menor valor absoluto y acabando con los de mayor valor absoluto con el fin de minimizar la expresión deducida para la cota del error absoluto.

b) Para el producto $\prod_{i=1}^{n} a_i$ se actúa de forma similar y se encuentra para $n = 3$ el valor calculado P_3

$$\{[(a_1(1+\delta_1)a_2(1+\delta_2)](1+\delta_{.2})a_3(1+\delta_3)\}(1+\delta_{.3})\ .$$

Despreciando los términos no lineales en δ, se escribe como

$$a_1a_2a_3(1+\delta_1+\delta_2+\delta_3+\delta_{.2}+\delta_{.3})$$

y el error relativo está acotado por 5ϵ.

En general,

$$P_n = a_1a_2\cdots a_n(1\pm(2n-1)\epsilon)$$

y el error relativo está acotado por

$$\epsilon_r\left(\prod_{i=1}^{n} a_i\right) = (2n-1)\epsilon\ ,$$

independientemente de la ordenación de los a_i $(i = 1 \div n)$.

c) Repitiendo los procedimientos anteriores, encontraremos fácilmente una cota del error absoluto en el cálculo de

$$\sum_{i=1}^{n} x_iy_i\ ;$$

los productos x_iy_i tendrán errores absolutos acotados por $3|x_iy_i|epsilon$ y errores relativos acotados por $\epsilon_i = 3\epsilon$, según la fórmula anterior para el error relativo en el producto; la suma de todos ellos tendrá un error que, escribiendo $a_i = x_iy_i$, y acotando $|\,\delta_i\,|$ por ϵ_i en $(*)$, está acotado por

$$\begin{aligned}\epsilon_a\left(\sum_{i=1}^{n} x_iy_i\right) = & [(n+2)\,|x_1y_1|\,+(n+2)\,|x_2y_2\,| \\ & +(n+1)\,|x_3y_3|\,+\cdots+4\,|x_ny_n\,|]\epsilon\ .\end{aligned}$$

Así pues, es recomendable, como en el primer caso, comenzar a realizar los cálculos con los menores sumandos, en valor absoluto, y acabar con los mayores.

d) La cota del error absoluto en

$$\sum_{i=1}^{n} x_i \ln x_i$$

se encuentra de forma similar al caso anterior una vez que se conoce una cota del error absoluto de los $\ln x_i$. Obsérvese que los números x_i deben ser positivos con el fin de poder calcular sus logaritmos.

El valor calculado de $\ln x_i$ será

$$\begin{aligned} \mathrm{fl}(\ln x_i) &= \ln[x_i(1+\delta_i)](1+\delta_{\ln}) \\ &\simeq \ln x_i + \delta_i + \ln x_i \delta_{\ln} \ , \end{aligned}$$

donde se ha aplicado la fórmula de propagación del error a la función ln y se han despreciado los términos no lineales en los valores de δ. Así, el error absoluto en los productos $a_i = x_i \ln x_i$ está acotado por $(1 + 5\,|\ln x_i\,|)\epsilon$, y el error relativo por

$$\epsilon_i = \left(5 + \frac{1}{|\ln x_i|}\right)\epsilon \ .$$

Acotando de nuevo $|\delta_i|$ por ϵ_i en $(*)$, tenemos la cota del error absoluto

$$\begin{aligned} \epsilon_a \left(\sum_{i=1}^{n} x_i \ln(x_i)\right) = [&x_1 + x_2 + x_3 + \cdots + x_n \\ &+(n+4)\,|x_1 \ln x_i\,| + (n+4)\,|x_2 \ln x_2\,| \\ &+(n+3)\,|x_3 \ln x_3\,| + \cdots + 6\,|x_n \ln x_n\,|]\epsilon \ . \end{aligned}$$

e) La evaluación del polinomio

$$p(x) = \sum_{i=0}^{n} a_i x^i$$

se analiza aquí con respecto a los errores; se escoge la regla de evaluación de Horner porque es la que realiza menos operaciones. No obstante, nótese que el algoritmo de evaluación consistente en calcular primero todos los términos $a_i x^i$ y sumarlos después puede analizarse usando la técnica que se ha utilizado hasta el momento en este ejercicio: acotar el error en cada sumando y usar $(*)$ para acotar el error absoluto en la suma.

Usando la regla de Horner, $p(x)$ se evalúa mediante la recurrencia

$$b_{n-1} = a_n, \quad b_{j-1} = b_j x + a_j \ (j = n-1 \div 0) \ ; \quad p(x) = b_{-1}$$

(véase el apartado 3.1.2).

Estableceremos ahora una recurrencia entre cotas ϵ_j de los errores absolutos de las b_j $(j = n-1 \div 0)$

$$\begin{aligned} \mathrm{fl}(b_{j-1}) &= \{[(b_j \pm \epsilon_j)x(1\pm\epsilon)](1\pm\epsilon) + a_j(1\pm\epsilon)\}(1\pm\epsilon) \\ &\simeq b_{j-1} \pm [\,|x|\,\epsilon_j + 2\,|b_j x|\,\epsilon + |a_j|\,\epsilon + |b_{j-1}|\,\epsilon] \ , \\ \mathrm{fl}(b_{j-1}) - b_{j-1} &\leq |x|\,\epsilon_j + [2\,|b_j x| + |a_j| + |b_{j-1}|]\epsilon \\ &\leq |x|\,\epsilon_j + [|a_j| + 2\,|b_j x| + |b_{j-1}|]\epsilon \ . \end{aligned}$$

Consideraremos así la siguiente recurrencia para las cotas:

$$\begin{aligned} \epsilon_n &= |a_n|\ \epsilon\ , \\ \epsilon_{j-1} &= |x|\,\epsilon_j + [|a_j| + 2\,|b_j x| + |b_{j-1}|]\epsilon\ (j = n \div 0)\ . \end{aligned}$$

Introduciendo

$$d_j = \frac{\epsilon_j}{\epsilon} + 2\,|b_j|\ \ (j = 0 \div n)\ ,$$

queda

$$\begin{aligned} d_{n-1} &= |a_n| + 2\,|b_{n-1}| = 3\,|a_n|\ , \\ d_{j-1} &= d_j\,|x| + |a_j| + 3\,|b_{j-1}|\ \ (j = n-1 \div 0)\ ; \end{aligned}$$

por lo tanto, como $\epsilon_{-1} = \epsilon(d_{-1} - 2\,|b_{-1}|)$, una cota del error absoluto en $p(x)$ se puede encontrar evaluando en $|x|$ el polinomio de coeficientes positivos

$$c_n = 3\,|a_n|\ ,\ \ c_j = |a_j| + 3\,|b_{j-1}|\ \ (j = n-1 \div 1)\ ,\ \ c_0 = |a_0|\ + |b_{-1}|\ .$$

Multiplicando finalmente por ϵ, encontramos la cota pedida

$$\epsilon_a(p(x)) = \left(\sum_{j=0}^{n} c_j\,|x|^j\right)\epsilon\ .$$

Problema 1.15 *Un ordenador comete errores relativos acotados por ϵ, ϵ, 2ϵ, 4ϵ y 5ϵ en el almacenamiento de datos, operaciones aritméticas, raíz cuadrada, cálculo del logaritmo y cálculo de la exponencial, respectivamente. ¿Para qué valores de a y b el error relativo en el cálculo de a^b como $\exp[b\ln(a)]$ es menor que 12ϵ?*

Aplicación: Estudiar el caso $\sqrt[7]{x}$.

SOLUCIÓN:

Queremos escribir la aproximación $\mathrm{fl}(a^b)$ en la forma $a^b(1+\delta)$ y después acotar $|\delta|$. Para esto, realizaremos un análisis en cadena de las acotaciones de los errores producidos en el almacenamiento y en cada operación, teniendo en cuenta que

$$\mathrm{fl}(x) = x(1+\delta_x);\ \ \mathrm{fl}(x * y) = \mathrm{fl}(x) * \mathrm{fl}(y)(1+\delta_*);$$

con $|\delta_x| \leq \epsilon$ y $|\delta_*| \leq \epsilon$, para las operaciones aritméticas; y

$$\mathrm{fl}(f(x)) = f(\mathrm{fl}(x))(1+\delta_f)\ ,$$

con $|\delta_f| \leq 2\epsilon$, para la raíz cuadrada, $|\delta_f| \leq 4\epsilon$, para el cálculo del logaritmo, y $|\delta_f| \leq 5\epsilon$, para el cálculo de la exponencial.

Así,
$\mathrm{fl}(a) = a(1+\delta_a)$ (representación inicial de a),
$\mathrm{fl}(\ln(a)) = \ln[a(1+\delta_a)](1+\delta_{\ln})$ (logaritmo),
$\mathrm{fl}(b) = b(1+\delta_b)$ (representación inicial de b),
$\mathrm{fl}(b\ln(a)) = b(1+\delta_b)\ln[a(1+\delta_a)](1+\delta_{\ln})(1+\delta.)$ (producto) ,
$\mathrm{fl}(a^b) = \exp\{b\ln[a(1+\delta_a)](1+\delta_b)(1+\delta_{\ln})(1+\delta.)\}(1+\delta_{\exp})$ (exponencial).
Usando la fórmula de propagación del error en una variable, tenemos

$$\ln[a(1+\delta_a)] \simeq \ln(a) + \delta_a \ , \quad \exp[b\ln(a)+\gamma] \simeq a^b(1+\gamma) \ .$$

Aplicándolo en la expresión anterior, con

$$\gamma = b\delta_a + b\ln(a)(\delta_{\ln} + \delta_b + \delta.) \ ,$$

se llega a

$$\mathrm{fl}(a^b) \simeq a^b[1 + \delta_{\exp} + b\delta_a + b\ln(a)(\delta_{\ln} + \delta_b + \delta.)] \ ,$$

que ya da una expresión aproximada para δ en la cual se han tenido en cuenta sólo los errores lineales en ϵ.

Acotando la aproximación de δ en esta expresión y usando la información dada sobre las cotas de los errores cometidos en el almacenamiento y en los cálculos, se obtiene

$$\mathrm{fl}(a^b) = a^b[1 \pm (5 + |b|(1 + 6|\ln(a)|))\epsilon] \ ;$$

de donde, la condición sobre a y b pedida en el enunciado será

$$|b|(1 + 6|\ln(a)|) \leq 7 \ .$$

En consecuencia $|b| \leq 7$ y, en este caso,

$$e^{-c} \leq a \leq e^c \ , \quad c = \frac{7}{6|b|} - \frac{1}{6} \ .$$

La aplicación corresponde a elegir $b = \frac{1}{7}$; así, el error en la raíz séptima, calculada como $\exp[\frac{1}{7}\ln(a)]$, será menor que 12ϵ para los valores de a en el intervalo $[e^{-8}, e^8]$.

Problema 1.16 *Encontrar una expresión de la cota aproximada del error absoluto al evaluar la función*

$$f(x) = \sqrt{3 + \ln^2(x)} \ .$$

Se supone que los errores relativos en la representación de números, operaciones aritméticas, raíces cuadradas y cálculo de logaritmos son, en valor absoluto, menores que ϵ, 2ϵ, 3ϵ y 5ϵ, respectivamente; además, se comete un error relativo menor, en valor absoluto, que 4ϵ al medir x.

SOLUCIÓN:

Para calcular $f(x) = \sqrt{3 + \ln^2(x)}$ realizamos los siguientes pasos:

1. Almacenamiento de x : $\quad x \quad \to x_1 = \mathrm{fl}(x) = x(1 \pm 4\epsilon)(1 \pm \epsilon)$.
2. Cálculo de $\ln x$: $\quad x_1 \to x_2 = \mathrm{fl}(\ln x_1) = \ln x_1(1 \pm 5\epsilon)$.
3. Cálculo de $\ln^2 x$: $\quad x_2 \to x_3 = \mathrm{fl}(x_2^2) = x_2^2(1 \pm 2\epsilon)$.
4. Cálculo de $3 + \ln^2 x$: $\quad x_3 \to x_4 = \mathrm{fl}(3 + x_3) = (3 + x_3)(1 \pm 2\epsilon)$.
5. Cálculo de $\sqrt{3 + \ln^2 x})$: $\quad x_4 \to x_5 = \mathrm{fl}(\sqrt{x_4}) = \sqrt{x_4}(1 + 3\epsilon)$.

Nótese que

$$x_1 \simeq x \ , \quad x_2 \simeq \ln x \ , \quad x_3 \simeq \ln^2 x \ , \quad x_4 \simeq 3 + \ln^2 x \ , \quad x_5 \simeq \sqrt{3 + \ln^2 x} \ .$$

Queremos encontrar una cota para el error absoluto. Usaremos repetidamente la fórmula para el error en un producto y la fórmula aproximada de propagación de los errores:

$$\begin{aligned}
1.\ \epsilon_a(x_1) &= 5x\epsilon \ . \\
2.\ \epsilon_a(x_2) &= \epsilon_a(\ln x_1) + 5\,|x_2|\,\epsilon \simeq \frac{1}{x_1}\epsilon_a(x_1) + 5\,|\ln x_1|\,\epsilon \\
&\simeq 5(1 + |\ln x_1|)\epsilon \ . \\
3.\ \epsilon_a(x_3) &= \epsilon_a(x_2^2) + 2x_2^2\epsilon \simeq 2x_2\epsilon_a(x_2) + 2x_2^2\epsilon \\
&\simeq [10\,|\ln x|\,(1 + |\ln x|) + 2\ln^2 x]\epsilon = (12\ln^2 x + 10\,|\ln x|)\epsilon \ . \\
4.\ \epsilon_a(x_4) &= \epsilon_a(x_3) + 2(3 + x_3)\epsilon \simeq (14\ln^2 x + 10\,|\ln x| + 6)\epsilon \ . \\
5.\ \epsilon_a(x_5) &= \frac{1}{2\sqrt{x_4}}\epsilon_a(x_4) + 3\sqrt{x_4}\epsilon \\
&\simeq \left[\frac{1}{2\sqrt{3 + \ln^2 x}}(14\ln^2 x + 10\,|\ln x| + 6) + 3\sqrt{3 + \ln^2 x}\right]\epsilon \ .
\end{aligned}$$

La expresión final para la cota es

$$\epsilon_a(\sqrt{3 + \ln^2 x}) = \frac{1}{\sqrt{3 + \ln^2 x}}(10\ln^2 x + 5\,|\ln x| + 12)\epsilon \ .$$

Problema 1.17 *Queremos calcular* $\cos(\cos\sqrt{x})$ *y tenemos representado* x *con un error relativo menor que* ϵ *y el cálculo de la raíz cuadrada y del coseno se hace con errores relativos acotados por* 3ϵ *y* 5ϵ*, respectivamente. Acotar aproximadamente el error relativo total en función de* x*, suponiendo* ϵ *suficientemente pequeño, y demostrar que, para cualquier* $x \in (0, 1)$*, la cota es menor que* $(5 + \sqrt{37.25}\tan 1)\epsilon$*.*

SOLUCIÓN:

Para calcular $\cos(\cos\sqrt{x})$ realizamos los siguientes pasos:

1. Almacenamiento de x : $x \to x_1 = \mathrm{fl}(x) = x(1 \pm \epsilon)$.
2. Cálculo de $\sqrt{x}$: $x_1 \to x_2 = \mathrm{fl}(\sqrt{x_1}) = \sqrt{x_1}(1 \pm 3\epsilon)$.
3. Cálculo de $\cos\sqrt{x}$: $x_2 \to x_3 = \mathrm{fl}(\cos x_2) = \cos x_2(1 + 5\epsilon)$.
4. Cálculo de $\cos(\cos\sqrt{x})$: $x_3 \to x_4 = \mathrm{fl}(\cos x_3) = \cos x_3(1 + 5\epsilon)$.

Nótese que

$$x_1 \simeq x \ , \quad x_2 \simeq \sqrt{x} \ , \quad x_3 \simeq \cos\sqrt{x} \ , \quad x_4 \simeq \cos(\cos\sqrt{x}) \ .$$

Queremos encontrar una cota para el error relativo

$$\epsilon_r(x_4) \simeq \frac{\epsilon_a(x_4)}{\cos(\cos\sqrt{x})} \ ;$$

con este fin, usaremos repetidamente la fórmula para el error en un producto y la fórmula aproximada de propagación del error:

$$\begin{aligned}
1.\ \epsilon_a(x_1) &= x\epsilon \ . \\
2.\ \epsilon_a(x_2) &= \epsilon_a(\sqrt{x_1}) + 3\sqrt{x_1}\epsilon \simeq \frac{1}{2\sqrt{x_1}}\epsilon_a(x_1) + 3\sqrt{x_1}\epsilon \\
&\simeq \frac{7}{2}\sqrt{x}\epsilon \ . \\
3.\ \epsilon_a(x_3) &= \epsilon_a(\cos x_2) + 5\cos x_2\epsilon \simeq \operatorname{sen} x_2\epsilon_a(x_2) + 5\cos x_2\epsilon \\
&\simeq (\frac{7}{2}\sqrt{x}\operatorname{sen}\sqrt{x} + 5\cos\sqrt{x})\epsilon \ . \\
4.\ \epsilon_a(x_4) &= \epsilon_a(\cos x_3) + 5\cos x_3\epsilon \simeq \operatorname{sen} x_3\epsilon_a(x_3) + 5\cos x_3\epsilon \\
&\simeq \left[\left(\frac{7}{2}\sqrt{x}\operatorname{sen}\sqrt{x} + 5\cos\sqrt{x}\right)\operatorname{sen}(\cos\sqrt{x})\right) + 5\cos(\cos\sqrt{x})\right]\epsilon \ .
\end{aligned}$$

Así pues tenemos ya la cota del error relativo para cada $x \in (0,1)$

$$\epsilon_r(x_4) = \left[\frac{7}{2}\sqrt{x}\operatorname{sen}\sqrt{x} + 5\cos\sqrt{x})\tan(\cos\sqrt{x}) + 5\right]\epsilon \ ;$$

para demostrar que está acotado por $(5 + \sqrt{37.25}\tan 1)\epsilon$, sólo es necesario ahora ver que la función

$$f(t) = \frac{7}{2}t\operatorname{sen} t + 5\cos t$$

está acotada por $\sqrt{37.25}$ para todo $t \in (0,1)$, dado que la función $t \mapsto \tan(\cos t)$ es decreciente en el intervalo $(0,1)$ y puede acotarse por $\tan 1$. Acotamos $f(t)$ por la función $g(t) = \frac{7}{2}\cos t + 5\operatorname{sen} t$; esta función alcanza su valor máximo para $t = t^*$ tal que $g'(t^*) = 0$ (esto es, $\tan t^* = \frac{7}{10}$); calculando su coseno encontramos

$$\cos t^* = \frac{1}{\sqrt{1 + \tan^2 t^*}} = \frac{10}{\sqrt{149}}$$

y entonces, como se quería demostrar:

$$g(t) \leq g(t^*) = (\frac{7}{2}\tan t* + 5)\cos t^* = (\frac{7}{2}\frac{7}{10} + 5)\frac{10}{\sqrt{149}} = \sqrt{\frac{149}{4}} = \sqrt{37.25}\ .$$

Problema 1.18 *Los errores relativos en la representación de números, operaciones aritméticas* $(+, -, \cdot, /)$*, raíz cuadrada y cálculo del* arcsen *son, en valor absoluto, menores que* ϵ*,* ϵ*,* 2ϵ *y* 5ϵ*, respectivamente. ¿Para qué valores de* x *se puede garantizar que el valor absoluto del error relativo de* $\arctan(x)$ *es menor que* 20ϵ*, si lo calculamos mediante la fórmula*

$$\arctan x = \text{arcsen}\ \frac{x}{\sqrt{1+x^2}}\ ?$$

SOLUCIÓN:

Para el cálculo de $f(x) = \arctan x = \text{arcsen}\ \frac{x}{\sqrt{1+x^2}}$ hay que realizar los siguientes pasos intermedios:

1. Almacenamiento de x: $\quad x \to x_1 = \text{fl}(x) = x(1 \pm \epsilon)$.
2. Cálculo de x^2: $\quad x_1 \to x_2 = \text{fl}(x_1^2) = x_1^2(1 \pm \epsilon)$.
3. Cálculo de $1 + x^2$: $\quad x_2 \to x_3 = \text{fl}(1 + x_2) = (1 + x_2)(1 \pm \epsilon)$.
4. Cálculo de $\sqrt{1 + x^2}$: $\quad x_3 \to x_4 = \text{fl}(\sqrt{x_3}) = \sqrt{x_3}(1 \pm 2\epsilon)$.
5. Cálculo de $\frac{x}{\sqrt{1+x^2}}$): $\quad x_4 \to x_5 = \text{fl}(\frac{x_1}{x_4}) = \frac{x_1}{x_4}(1 \pm \epsilon)$.
6. Cálculo de $\arctan x$: $\quad x_5 \to x_6 = \text{fl}(\text{arcsen}\, x_5) = \text{arcsen}\, x_5(1 \pm 5\epsilon)$.

Nótese que $x_1 \simeq x$, $x_2 \simeq x^2$, $x_3 \simeq 1 + x^2$, $x_4 \simeq \sqrt{1 + x^2}$, $x_5 \simeq \frac{x}{\sqrt{1+x^2}}$, $x_6 \simeq \arctan x$.

Queremos encontrar una cota para el error relativo de x_6. Para ello, usaremos repetidamente la fórmula para el error en un producto y la fórmula aproximada de propagación del error:

1. $\epsilon_a(x_1) = |x|\,\epsilon$.
2. $\epsilon_a(x_2) = \epsilon_a(x_1^2) + x_2\epsilon \simeq 2\,|x_1\,|\epsilon_a(x_1) + x_1^2\epsilon \simeq 3x^2\epsilon$.
3. $\epsilon_a(x_3) = \epsilon_a(1 + x_2) + (1 + x_2)\epsilon \simeq \epsilon_a(x_2) + (1 + x_2)\epsilon \simeq (1 + 4x^2)\epsilon$.
4. $\epsilon_a(x_4) = \epsilon_a(\sqrt{x_3}) + 2\sqrt{x_3}\epsilon \simeq \frac{1}{2\sqrt{x_3}}\epsilon_a(x_3) + 2\sqrt{x_3}\epsilon$
   $\simeq (\frac{1 + 4x^2}{2\sqrt{1 + x^2}} + 2\sqrt{1 + x^2})\epsilon = \frac{5 + 8x^2}{2\sqrt{1 + x^2}}\epsilon$.
5. $\epsilon_a(x_5) = \epsilon_a(\frac{x_1}{x_4}) + \frac{|x_1|}{x_4}\epsilon \simeq \frac{|x_1\epsilon_a(x_4) + x_4\epsilon_a(x_1)|}{x_4^2} + \frac{|x_1|}{x_4}\epsilon \simeq$

$$\simeq \frac{3\,|x|\,(3+4x^2)}{2\sqrt{1+x^2}(1+x^2)}\epsilon\ .$$

$$6.\ \epsilon_a(x_6) = \epsilon_a(\operatorname{arcsen} x_5) + 5\,|\operatorname{arcsen} x_5|\,\epsilon \simeq \frac{\epsilon_a(x_5)}{\sqrt{1-x_5^2}} + 5\operatorname{arcsen} x_5\epsilon$$

$$\simeq \left[\frac{3\,|x|\,(3+4x^2)}{2(1+x^2)} + 5\,|\arctan x|\right]\epsilon\ .$$

La expresión obtenida para la cota del error relativo es

$$\epsilon_r(\arctan x) \simeq \left[\frac{3\,|x|\,(3+4x^2)}{2\,|\arctan x|\,(1+x^2)} + 5\right]\epsilon\ ;$$

la condición de que el error relativo sea menor que 20ϵ puede asegurarse para los números x que cumplen

$$\frac{3\,|x|\,(3+4x^2)}{2\,|\arctan x|\,(1+x^2)} < 15\ .$$

Esta relación se satisface para los valores de x tales que

$$|x| < 3.25\ .$$

Problema 1.19 *Usando un método recurrente, calcular el valor de las integrales*

$$I_j = \int_0^1 x^j \operatorname{sen}(\pi x)\ dx \quad (j = 2, 4, \ldots, 20)\ .$$

Estudiar la estabilidad del método encontrado.

SOLUCIÓN:

La recurrencia se encuentra partiendo de la expresión de I_j e integrándola por partes dos veces:

$$I_j = \int_0^1 x^j \operatorname{sen}(\pi x)dx = -\frac{1}{\pi}x^j\cos(\pi x)\Big|_0^1 - \frac{j}{\pi}\int_0^1 x^{j-1}\cos(\pi x)dx$$

$$= \frac{1}{\pi} - \frac{j}{\pi}\left[x^{j-1}\operatorname{sen}(\pi x)\Big|_0^1 - \frac{j-1}{\pi}\int_0^1 x^{j-2}\operatorname{sen}(\pi x)dx\right]$$

$$= \frac{1}{\pi} - \frac{j(j-1)}{\pi^2}I_{j-2}\ .$$

La recurrencia que cumplen las integrales I_j $(j = 2, 4, 6, \ldots)$ es, por lo tanto,

$$I_0 = \frac{2}{\pi} = 0.636619772... \ , \quad I_j = \frac{1}{\pi} - \frac{j(j-1)}{\pi^2} I_{j-2} \ (j \geq 2) \ .$$

j	I_j
0	0.636620
2	0.189304
4	0.088144
6	0.050384
8	0.032433
10	0.022560
12	0.016588
14	0.012423
16	0.016211
18	−0.018429
20	7.413890

Tabla 1.1: Recurrencia hacia adelante.

j	I_j
0	1.000
2	0.202
4	0.246
6	0.748
8	4.249
10	38.749
12	518.249
14	9556.798
16	232392.245
18	7205154.731
20	277413226.170

Tabla 1.2: Factores de amplificación de errores.

En la tabla 1.1 se muestran los resultados obtenidos al utilizar la recurrencia con precisión simple (que es equivalente a trabajar con 6 cifras decimales).

La recurrencia se muestra decididamente *inestable*: nótese que las integrales tendrían que decrecer al aumentar el valor de j, y esto no se observa; además, ninguna integral I_j puede ser negativa, y I_{18} lo es. Esto se explica por el hecho de que los errores se amplifican a cada paso de la recurrencia por un factor $\frac{j(j-1)}{\pi^2}$. Este factor aumenta a medida que aumenta j, y el factor de amplificación total para el cálculo de I_j es $\frac{j!}{\pi^j}$ que toma los valores aproximados dados en la tabla 1.2.

Los fenómenos anteriores explican la inestabilidad y sugieren una forma recurrente más estable: si se despeja I_{j-2} en la recurrencia anterior se encuentra

$$I_{j-2} = \frac{\pi^2}{j(j-1)} (\frac{1}{\pi} - I_j) \ .$$

j	I_j
26	0.000000
24	0.004833
22	0.005605
20	0.006680
18	0.008094
16	0.010006
14	0.012679
12	0.016574
10	0.022561
8	0.032433
6	0.050384
4	0.088144
2	0.189304
0	0.636620

Tabla 1.3: Recurrencia hacia atrás.

Para utilizar esta nueva recurrencia, es necesario disponer de algún valor de I_j; ahora bien, usando el hecho de que la sucesión de integrales tiende a 0, se puede escoger la aproximación $I_n = 0$, para un valor suficientemente grande de n. Los errores cometidos en esta aproximación se irán reduciendo a medida que se aplique la recurrencia y podremos llevar a cabo el cálculo pedido. Si se escoge $n = 26$ se obtiene la tabla 1.3 que da el valor de las integrales pedidas con 6 cifras decimales correctas, aunque los tres primeros valores no tengan todas estas cifras correctas.

Problema 1.20 *Hacer lo mismo que en el ejercicio anterior con las integrales*

$$I_j = \int_0^1 \frac{x^j}{x+10}\, dx \ (j = 1 \div 20) \ .$$

SOLUCIÓN:

La recurrencia se encuentra simplemente escribiendo el numerador del integrando como $x^j + 10x^{j-1} - 10x^{j-1}$, entonces

$$I_j = \int_0^1 x^{j-1} dx - 10 \int_0^1 \frac{x^{j-1}}{x+10} dx = \frac{1}{j} - 10 I_{j-1} \ .$$

La recurrencia puede iniciarse con $I_0 = \ln(1.1) = 0.0953101798...$ y cabe esperar, como en el ejercicio anterior, una desestabilización numérica del proceso.

En la tabla 1.4 se muestran los resultados obtenidos al usar la recurrencia con precisión doble (que es equivalente a trabajar con 16 cifras decimales), dándose las 6 primeras cifras decimales.

j	I_j	j	I_j
0	0.095310	11	0.007622
1	0.046898	12	0.007114
2	0.031018	13	0.005785
3	0.023154	14	0.013577
4	0.018465	15	0.069105
5	0.015353	16	0.753549
6	0.013138	17	−7.476665
7	0.011481	18	74.822208
8	0.010194	19	−748.169449
9	0.009167	20	7481.744486
10	0.008329		

Tabla 1.4: Recurrencia hacia adelante.

j	I_j	j	I_j
24	0.000000	11	0.007629
23	0.004167	10	0.008328
22	0.003931	9	0.009167
21	0.004152	8	0.010194
20	0.004347	7	0.011481
19	0.004565	6	0.013138
18	0.004807	5	0.015353
17	0.005075	4	0.018465
16	0.005375	3	0.023154
15	0.005713	2	0.031018
14	0.006095	1	0.046898
13	0.006533	0	0.095310
12	0.007039		

Tabla 1.5: Recurrencia hacia atrás.

Las integrales deberían decrecer al aumentar el valor de j, y esto no se observa; además, ninguna integral I_j puede ser negativa, y I_{17} ya lo es. La explicación de este hecho se basa en que los errores se amplifican a cada paso de la recurrencia por un factor 10. A medida que aumenta j, el factor de amplificación total para el cálculo de I_j es 10^j.

Usando la recurrencia al revés

$$I_{j-1} = \frac{1}{10}(\frac{1}{j} - I_j) \ ,$$

el método de cálculo de integrales sí es estable; si partimos de $I_n = 0$ para n suficientemente grande, en cada paso del proceso recurrente el error se divide por 10, se gana aproximadamente una cifra decimal correcta. Si se escoge $n = 24$, ya se obtienen los valores pedidos con 6 cifras decimales correctas; las últimas cifras de I_j ($j = 24, 23, 22, 21$) están aún afectadas de error. Los resultados de los cálculos se muestran en la tabla 1.5.

Problema 1.21 *Queremos evaluar en $x = 1$ las funciones de Besel de orden entero*

$$J_n(x) = \sum_{j=0}^{\infty} (-1)^j \frac{(\frac{x}{2})^{2j+n}}{j!\ (j+n)!} \ ,$$

que cumplen la recurrencia

$$J_{n+1}(x) = \frac{2n}{x} J_n(x) - J_{n-1}(x) \quad (n \geq 1) \ .$$

a) Sabiendo que $J_0(1) = 0.765198$ y $J_1(1) = 0.440051$, calcular $J_7(1)$ usando la recurrencia anterior.

b) Comparar el valor obtenido con el resultado exacto $J_7(1) = 0.1502326 \cdot 10^{-6}$ y explicar el fenómeno que se observa.

c) Usando la relación

$$J_0(x) + 2 \sum_{r=1}^{\infty} J_{2r}(x) = 1 \ ,$$

calcular $J_n(1)$ $(n = 0 \div 20)$, con 6 cifras significativas correctas.

SOLUCIÓN:

n	$J_n(1)$
1	0.440051
2	0.114904
3	0.019565
4	0.002486
5	0.000321
6	0.000725
7	0.008377

Tabla 1.6: Recurrencia hacia adelante.

a) En la tabla 1.6 se muestran los resultados obtenidos al usar la recurrencia

$$J_0(1) = 0.765198 \ , \quad J_1(1) = 0.440051 \ , \quad J_{n+1}(1) = 2nJ_n(1) - J_{n-1}(1) \ ,$$

con precisión doble (que es equivalente a trabajar con 16 cifras decimales), dándose las 6 primeras cifras decimales.

b) No hay ninguna cifra significativa de coincidencia con el valor exacto, los errores relativos cometidos se hacen cada vez mayores. Por un lado, los errores absolutos pueden amplificarse a cada paso por un factor $2n$ (si no tenemos en cuenta los errores de $J_{n-1}(1)$). Su contribución en el factor de amplificación total del error absoluto para el cálculo de $J_n(1)$ es del orden de $(2n-1)!$, muy grande cuando n es grande; por otro lado, los valores de $J_n(1)$ son cada vez más pequeños. El resultado combinado de estos dos efectos es el de un error relativo enorme para $J_7(1)$.

c) Usando la recurrencia al revés

$$J_{n-1}(1) = 2nJ_n(1) - J_{n+1}(1) \ ,$$

n	$\bar{J}_n(1)$	n	$J_n(1)$
24	1	24	$9.507130 \cdot 10^{-32}$
23	48	23	$4.563423 \cdot 10^{-30}$
22	2207	22	$2.098224 \cdot 10^{-28}$
21	97060	21	$9.227621 \cdot 10^{-27}$
20	4074313	20	$3.873502 \cdot 10^{-25}$
19	162875456	19	$1.548478 \cdot 10^{-23}$
18	6185192960	18	$5.880344 \cdot 10^{-22}$
17	222504075264	17	$2.115375 \cdot 10^{-20}$
16	7558953172992	16	$7.186395 \cdot 10^{-19}$
15	241664002097152	15	$2.297531 \cdot 10^{-17}$
14	7242361222463488	14	$6.885407 \cdot 10^{-16}$
13	202544455746584576	13	$1.925617 \cdot 10^{-14}$
12	$5.25891353811773030 \cdot 10^{18}$	12	$4.999718 \cdot 10^{-13}$
11	$1.26011377280803144 \cdot 10^{20}$	11	$1.198007 \cdot 10^{-11}$
10	$2.76699134485810958 \cdot 10^{21}$	10	$2.630615 \cdot 10^{-10}$
9	$5.52138161795883650 \cdot 10^{22}$	9	$5.249249 \cdot 10^{-9}$
8	$9.91081733664729666 \cdot 10^{23}$	8	$9.422343 \cdot 10^{-8}$
7	$1.58020943277800528 \cdot 10^{25}$	7	$1.502326 \cdot 10^{-6}$
6	$2.20238241593444582 \cdot 10^{26}$	6	$2.093834 \cdot 10^{-5}$
5	$2.62705690163896132 \cdot 10^{27}$	5	$2.497577 \cdot 10^{-4}$
4	$2.60503302398405905 \cdot 10^{28}$	4	$2.476639 \cdot 10^{-3}$
3	$2.05775589444304340 \cdot 10^{29}$	3	$1.956335 \cdot 10^{-2}$
2	$1.20860318753651952 \cdot 10^{30}$	2	$1.149035 \cdot 10^{-1}$
1	$4.62863725514910340 \cdot 10^{30}$	1	$4.400506 \cdot 10^{-1}$
0	$8.04867147387741473 \cdot 10^{30}$	0	$7.651977 \cdot 10^{-1}$

Tabla 1.7: Recurrencia hacia atrás.

los errores absolutos van también en aumento de la misma forma que en a); parece en principio que, si procedemos a usar esta fórmula no obtendremos ningún resultado preciso; ahora bien, a medida que vamos calculando valores, observamos que éstos son cada vez mayores y se vislumbra la esperanza de un buen funcionamiento. Además, la relación

$$J_0(1) + 2\sum_{r=1}^{\infty} J_{2r}(1) = 1$$

nos procura una ayuda muy valiosa.

El procedimiento elegido se basa en la recurrencia en sentido contrario, para un valor de a cualquiera

$$\begin{aligned} \bar{J}_{N+1}(1) &= 0\ , \quad \bar{J}_N(1) = a\ , \\ \bar{J}_{n-1}(1) &= 2n\bar{J}_n(1) - \bar{J}_{n+1}(1)\ (n = N \div 1)\ , \end{aligned}$$

donde $\bar{J}_{N+1}(1) = 0$ es una buena aproximación si N es suficientemente grande. El valor de a permite controlar la magnitud de los valores de la sucesión, dado que son proporcionales

a esta a. La sucesión buscada se da para un cierto valor de a que se puede determinar aproximadamente usando la relación anterior. Así, escogiendo N par ($N = 2s$) y algún valor de a, usamos la recurrencia anterior para hallar $\bar{J}_n(1)$ ($n = N - 1 \div 0$). Calculamos

$$S_{2s} = \bar{J}_0(1) + 2\sum_{r=1}^{s} \bar{J}_{2r}(1)$$

ya que es una aproximación de $\frac{a}{J_N(1)}$. La sucesión buscada se encontrará entonces dividiendo por S_{2s} los valores calculados para los $\bar{J}_n(1)$.

En la tabla 1.7 se muestran los valores obtenidos para ambas sucesiones, escogiendo $a = 1$ y $N = 24$. Las 6 primeras cifras significativas de los valores pedidos han estado verificadas como correctas.

El valor obtenido para la suma es $S_{24} = 1.05184210896169095 \cdot 10^{31}$; los valores de $J_n(1)$ se han calculado dividiendo los de $\bar{J}_n(1)$ por este valor.

Problema 1.22 *Sean*

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

a) Determinar una ley de recurrencia para los I_n e indicar cómo usarla para calcularlos de forma numéricamente estable.

b) Si se conocen I_{n-1} y I_n con errores absolutos menores que ϵ y 2ϵ, respectivamente; ¿con cuántas cifras decimales correctas se podrá calcular I_j mediante el método obtenido en a) (suponiendo aritmética exacta)?

c) Explicar cómo se pueden obtener todos los I_j con t cifras decimales correctas sin calcular ninguna integral.

SOLUCIÓN:

a) La recurrencia se deduce escribiendo el numerador del integrando como $x^j + x^{j-1} + 6x^{j-2} - x^{j-1} - 6x^{j-2}$, entonces

$$\begin{aligned} I_j &= \int_0^1 x^{j-2}dx - \int_0^1 \frac{x^{j-1}}{x^2 + x + 6}dx = -6\int_0^1 \frac{x^{j-2}}{x^2 + x + 6}dx \\ &= \frac{1}{j-1} - I_{j-1} - 6I_{j-2} \ . \end{aligned}$$

La recurrencia se ha de plantearse en la forma

$$I_{j-2} = \frac{1}{6}(\frac{1}{j-1} - I_j - I_{j+1})$$

para que sea numéricamente estable; así, la suma de los errores de I_j y I_{j-1} se divide por 6 en cada paso del proceso recurrente, con lo que los errores decrecen.

b) Consideremos I_n y I_{n-1} con cotas de errores absolutos ϵ y 2ϵ. El error absoluto en I_{n-2} puede acotarse por

$$\epsilon_a(I_{n-2}) = \frac{\epsilon_a(I_{n-1}) + \epsilon_a(I_n)}{6} = \frac{\epsilon}{2} \ ,$$

que es la mitad del de I_{n-1}; repitiendo este proceso de nuevo $n-j$ veces, hallamos

$$\epsilon_a(I_j) = \frac{\epsilon}{2^{n-j-1}} \ .$$

El error obtenido no afectará a la cifra decimal t-ésima si

$$\frac{\epsilon}{2^{n-j}} < \frac{1}{2}10^{-t} \ ; \quad (*)$$

así, podemos asegurar que el número t de cifras decimales correctas para I_j es la parte entera de

$$\log \frac{2^{n-j-1}}{\epsilon} \ .$$

c) Si tomamos el valor $j = n$ y aproximamos I_n por cero se comete un error acotado por

$$\epsilon_a(I_n) = \int_0^1 \frac{x^n}{6}\, dx = \frac{1}{6(n+1)} \equiv \epsilon \ ;$$

si tomamos también cero como aproximación de I_{n-1}, tenemos que el error está acotado por

$$\int_0^1 \frac{x^{n-1}}{6} dx = \frac{1}{6n)} \leq 2\epsilon \equiv \epsilon_a(I_{n-1}) \ .$$

Así, aplicando el resultado encontrado en b), tenemos aseguradas t cifras decimales correctas si se cumple la relación $(*)$; es necesario escoger, para cada valor de j, un valor de n tal que

$$3(n+1)2^{n-j-1} > 10^t \ .$$

Nótese que, a partir de $j = 166666$, todas las cifras decimales de I_j son nulas.

PROBLEMAS PROPUESTOS

1. Una calculadora trabaja en base 2 y mantisa de 20 bits y aproxima por redondeo, y un ordenador trabaja en base 16 y mantisa de 24 bits y aproxima por corte. ¿Cuál de los dos se puede considerar más preciso?

2. ¿Con qué exactitud se ha de medir el radio de una esfera y con cuántos decimales se ha de dar el número π para que su volumen se conozca con un error relativo menor que el 0.01%? Considérense ambos efectos por separado.

3. Conocemos el seno y la tangente de un ángulo con la misma precisión. Indicar si es mejor usar la función arcsen o la función arctan para hallar el valor del ángulo.

4. Determinar las raíces de la ecuación de segundo grado $x^2 - 200x + 1 = 0$ con 9 cifras decimales, utilizando una calculadora de bolsillo. ¿Por qué no es correcto usar la expresión $100 \pm \sqrt{100^2 - 1}$?

5. Conociendo las longitudes a y b de dos lados de un triángulo y el valor C del ángulo que forman, la longitud c del lado restante se puede obtener mediante la *fórmula del coseno*

$$c = \sqrt{a^2 + b^2 - 2ab\cos(C)}\ .$$

   a) Si se conocen exactamente a y b, pero de C solamente sabemos que está en el intervalo I, determinar el intervalo donde se encuentra c,

   - exactamente, operando con intervalos,
   - aproximadamente, usando la fórmula de propagación del error maximal,

   en los casos siguientes:

   i) $a = 10$, $b = 20$, $I = [\frac{\pi}{6}, \frac{\pi}{3}]$;

   ii) $a = b = 20$, $I = [-\epsilon, \epsilon]$.

   b) Si $a = b = 20$ y $C = 0.00001$, calcular c con el mismo número de cifras correctas que su calculadora admita.

6. Calcular el error permitido en la inclinación de un cañón para asegurar que acierte un objetivo rectangular de 40 metros de longitud y 20 metros de anchura, situado horizontalmente a la misma altura del cañón, y cuyo centro está a una distancia de 3.000 metros de éste. La velocidad de salida del proyectil es de 600 metros por segundo. Supóngase que todas las magnitudes dadas son exactas y que las dimensiones del objetivo son pequeñas en relación a la distancia a la cual se encuentra el cañón.

Considérese también que se apunta al centro del objetivo y que está permitido tocar en cualquier lugar de éste.

7. ¿Cuál es la mejor manera de sumar 9999, 999, 99 y 9 con una calculadora que utilice representación decimal en punto flotante de 4 dígitos de mantisa?

8. Estudiar, en función de a, b y c, cuál es la mejor manera de calcular en punto flotante

$$a+b+c\ ,\quad abc\ ,\quad (a+b)^2,\quad a^2b^2,\quad (a+b)(a+c)\ .$$

9. Tenemos que calcular x^n para n natural. Suponemos que x está almacenado con un error relativo menor que ϵ, que las multiplicaciones se realizan con error relativo menor que ϵ y que las funciones ln y exp dan errores relativos acotados por 4ϵ y 6ϵ, respectivamente. Comparar la acumulación de los errores que se producen en los dos algoritmos de cálculo siguientes:

   i) $x^n = \exp[n\ln(x)]$;

   ii) se representa n en base 2, $n = a_k \cdots a_1 a_0$ $(a_k \neq 0)$, se calculan las potencias x^{2^j} $(j = 1 \div k)$ y se multiplican aquellas que se corresponden con los unos de la representación en base 2 (ejemplo: x^{21} se calcula mediante
   $x \to x^2 \to x^4 \to x^8 \to x^{16} \to x^{20} = x^{16}\cdot x^4 \to x^{21} = x^{20}\cdot x$).

10. Queremos calcular $\cos(20x)$ para $x \in [0, \pi/2]$. Disponemos de un ordenador que trabaja en base 2 y, por lo tanto, realiza exactamente las multiplicaciones por enteros y las divisiones por potencias de 2. Las operaciones aritméticas y las representaciones de los números se realizan con un error relativo acotado por ϵ. La función que calcula $\cos y$ para $y \in [0, \pi/2]$ comete un error relativo acotado por 3ϵ.

    Acotar aproximadamente el error relativo en el cálculo de $\cos(20x)$ para $x \in [0, \pi/2]$, usando los algoritmos siguientes:

    i) Reducir primeramente el ángulo $20x$ al primer cuadrante y aplicar después la función de cálculo del coseno.

    ii) Aplicar la función directamente a x para calcular $\cos x$ y utilizar la recurrencia: $c_0 = 1$, $c_1 = \cos x$, $c_{k+1} = 2c_1c_k - c_{k-1}$ $(k = 1 \div 19)$, para obtener $\cos(20x) = c_{20}$.

11. Para la evaluación del polinomio $p(x) = a_n x^n + \cdots + a_1 x + a_0$ se dispone de los

siguientes algoritmos:

i)	ii)
$p \leftarrow a_0$	$p \leftarrow a_n$
$z \leftarrow 1$	
$(k = 1 \div n)$	$(k = n-1 \div 1)$
$\quad z \leftarrow zx$	
$\quad p \leftarrow p + a_k z$	$\quad p \leftarrow px + a_k$
	(Algoritmo de Horner) .

Llamaremos ϵ una cota del error relativo al llevar a cabo las operaciones aritméticas y no consideraremos error de representación.

a) Explicitar el error que se comete en el cálculo de p en cada paso, en función del error en el paso anterior en cada uno de los algoritmos.

b) Dar una cota del error absoluto cometido al calcular un valor de $p(x)$, usando cada uno de los algoritmos. (Indicación: hacerlo por recurrencia).

c) Suponiendo que $|a_i| \simeq 1$ $(i = 0 \div n)$, decidir qué algoritmo conviene usar: 1) cuando x toma valores grandes, 2) cuando x toma valores pequeños.

12. Resolver la ecuación

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

donde $a = 1.000 \pm 0.001$ y estudiar los efectos de la incertidumbre de a sobre las soluciones de la ecuación.

13. Estudiar el comportamiento de las soluciones del sistema

$$\left.\begin{array}{rcrcrcl} 2x & + & 3y & + & z & = & 6 \\ 2\epsilon x & + & 2y & + & 2\epsilon z & = & 2+4\epsilon \\ 2\epsilon x & + & 2y & - & \epsilon z & = & 1+\epsilon \end{array}\right\} \ ,$$

para ϵ próximo a 0.

14. Las soluciones de la ecuación de tercer grado $z^3 + a_2 z^2 + a_1 z + a_0 = 0$ pueden expresarse en función de los coeficientes a_2, a_1 y a_0 como se indica a continuación.

Haciendo primeramente

$$q = \frac{a_1}{3} - \frac{a_2^2}{3} \ , \ r = \frac{a_1 a_2 - 3a_0}{6} - \frac{a_2^3}{27} \ , \quad s_{1,2} = \sqrt[3]{r \pm \sqrt{q^3 + r^2}} \ ,$$

las soluciones z_0, z_1 y z_2 se expresan así:

$$z_0 = (s_1 + s_2) - \frac{a_2}{3} \ , \ z_{1,2} = -\frac{s_1 + s_2}{2} - \frac{a_2}{3} \pm \frac{\sqrt{3}(s_1 - s_2)}{2} \ y \ .$$

Para simplificar, supondremos $a_2 = 0$.

a) Estudiar para qué valores de a_0 y a_1 aparecen cancelaciones en el problema cuando se buscan las soluciones usando las fórmulas anteriores.

b) Dar métodos numéricamente mejores para el caso $|a_0| \gg |a_1|$.

c) Aplicación: Para $a_0 = 100$ y $a_1 = 0.01$ y suponiendo que ambos tienen error absoluto acotado por ϵ, estimar el valor de ϵ que asegure aproximadamente una cota de error absoluto para z_0 de 10^{-10}.

CAPÍTULO 2

SISTEMAS LINEALES

Los problemas lineales son los que aparecen, con mayor frecuencia, en Matemática Aplicada de manera directa o por linealización de otros problemas. En este capítulo se presentan diversas técnicas de resolución de sistemas lineales y de problemas de valores propios. La elección concreta de cada método depende de diversos factores, como la estructura del sistema, su eficiencia, su estabilidad numérica, etc.

2.1 CONCEPTOS BÁSICOS

2.1.1 Tipos de matrices

Denotaremos por $\mathcal{M}_{m,n}$ el espacio vectorial de las matrices complejas mxn de m filas y n columnas. Los *vectores de dimensión* m son elementos de $\mathcal{M}_{m,1}$.

Los *elementos de una matriz* $M \in \mathcal{M}_{m,n}$ serán denotados por

$$m_{kj} \quad (k = 1 \div m)\ (j = 1 \div n)$$

y los *elementos de un vector* y, por

$$y_k \quad (k = 1 \div m)\ .$$

En símbolos, $M = (m_{kj})$ y $y = (y_k)$.

Para una matriz $M \in \mathcal{M}_{m,n}$ compleja, denotaremos su *transpuesta* por $M^\top$, su *conjugada* por $\overline{M}$ y su *adjunta* por M^*: $M^\top \in \mathcal{M}_{n,m}$ se forma cambiando filas por columnas, $\overline{M} \in \mathcal{M}_{m,n}$ se forma conjugando los elementos de M y $M^* \in \mathcal{M}_{n,m}$ es la transpuesta de la matriz conjugada de M.

Las matrices tratadas en este capítulo serán consideradas reales, a menos que se especifique lo contrario.

Las matrices A con el mismo número de filas que de columnas ($m = n$) reciben el nombre de *matrices cuadradas*. Denotaremos la *matriz identidad* por I y la *inversa* de la matriz A por A^{-1}, si existe. Los elementos de la matriz identidad se pueden denotar por δ_{ij}, usando la *delta de Kronecker*:

$$\delta_{ii} = 1\ , \quad \delta_{ij} = 0 \quad (i \neq j)\ .$$

Las matrices kxk formadas por las k primeras filas de las k primeras columnas de A serán llamadas *submatrices principales de orden k de A* y denotadas por $(A)_k$.

Denotaremos el *determinante de A* por $\det A$. Llamaremos *determinantes principales de A* a los determinantes de sus submatrices principales y los denotaremos por $\Delta_k \equiv \det (A)_k$.

Veamos ahora diferentes tipos de matrices cuadradas:

- *A singular* : $\det A = 0$.
- *A regular* : $\det A \neq 0$.
- *A hermítica (A simétrica)* : $A^* = A$ $(A^\top = A)$.
- *A unitaria (A ortogonal)* : $A^{-1} = A^*$ $(A^{-1} = A^\top)$.
- *A definida positiva*: A hermítica y $x^*Ax > 0 \ \ \forall x \neq 0$.
- *A estrictamente diagonal dominante* :

$$|\, a_{ii} \,| > \sum_{j \neq y, j=1}^{n} |\, a_{ij} \,| \ \ (i = 1 \div n) \ .$$

- *A banda (p,q)* : $a_{ij} = 0$ $(i \geq j + p$ ó $j \geq i + q)$.
  - *A Hessenberg superior* : $p = 2, \ q = n$.
  - *A Hessenberg inferior* : $p = n, \ q = 2$.
  - *A triangular superior* : $p = 1, \ q = n$.
  - *A triangular inferior* : $p = n, \ q = 1$.
  - *A diagonal, tridiagonal, pentadiagonal, ...* : $p = q = 1, 2, 3, \ldots$

Las definiciones de matrices unitarias y hermíticas corresponden a matrices complejas y las de matrices ortogonales y simétricas, a matrices reales. Nótese que las matrices unitarias y hermíticas reales son, respectivamente, matrices ortogonales y simétricas.

Dos matrices A y B se llaman *semejantes* si existe una matriz inversible C, llamada *transformación de semejanza de A a B*, tal que se cumple $B = C^{-1}AC$.

Se dice que una matriz A es *diagonalizable* cuando es semejante a una matriz diagonal.

2.1.2 Definiciones

Sistemas de ecuaciones lineales

Todo sistema lineal de n ecuaciones con n incógnitas puede ser escrito, en forma matricial, como

$$Ax = b \quad \text{o brevemente} \quad (A \,|\, b) \ ,$$

donde A es una matriz nxn, llamada *matriz del sistema*, y b es un vector de dimensión n, llamado *vector de términos independientes*. Se denotarán los elementos de la matriz A por a_{ij} $(i, j = 1 \div n)$ y, los del vector b, por b_i $(i = 1 \div n)$.

Se supone en este capítulo que el sistema es *compatible* y *determinado*; es decir, que el determinante de la matriz A es no nulo. El objetivo será pues buscar el vector x de componentes x_j $(j = 1 \div n)$ tal que $Ax = b$, llamado *solución del sistema*. En el capítulo siguiente se tratarán sistemas de la forma

$$Ma = y \ ,$$

donde M es una matriz mxn con $m > n$ (con más ecuaciones que incógnitas).

Valores y vectores propios

Dada una matriz nxn $A = (a_{ij})$, diremos que $v \neq 0$ es un *vector propio* de A de *valor propio* λ cuando

$$Av = \lambda v \ .$$

La condición anterior puede ser escrita en la forma $(A - \lambda I)v = 0$. Para que este sistema de ecuaciones tenga soluciones no nulas es necesario y suficiente que λ cumpla la *ecuación característica*

$$p_A(\lambda) \equiv \det(A - \lambda I) = 0 \ ;$$

así , los valores propios de A son los ceros del polinomio p_A, llamado *polinomio característico* de A.

Los vectores propios de A y $A^\top$ reciben también el nombre de *vectores propios por la derecha* y *vectores propios por la izquierda* de A, respectivamente.

Se llama *radio espectral* de A al módulo máximo de todos los valores propios de A y se denota por $\rho(A)$.

A continuación se presentan dos consideraciones que ilustran las definiciones hechas de valores y vectores propios:

INTERPRETACIÓN GEOMÉTRICA

La matriz A puede considerarse como representación, en una cierta base, de una transformación lineal del espacio vectorial $\mathbb{R}^n$ en él mismo. Un vector propio es aquel vector v que, sometido a la transformación, conserva su dirección (si λ es complejo, entendemos que λv representa un vector en la misma dirección que v, respecto al cuerpo de los números complejos). Si $\lambda > 0$ conserva también el sentido y, si $\lambda < 0$, lo invierte. Si $|\lambda| = 1$, conserva el módulo; si $|\lambda| > 1$, el módulo aumenta y, si $|\lambda| < 1$, disminuye.

Si λ es un valor propio de A, la solución del sistema $(A - \lambda I)v = 0$ no es única; existe todo un subespacio vectorial de vectores solución que se llama *subespacio propio* del valor propio λ. Un subespacio generado por un único vector propio se llama también *dirección propia*.

EJEMPLO

La matriz

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$$

tiene por ecuación característica

$$p_A(\lambda) = (3 - \lambda)(3 - \lambda) - 1 = \lambda^2 - 6\lambda + 8 = 0 \ ,$$

de soluciones $\lambda_1 = 4$ y $\lambda_2 = 2$. Estas soluciones son así los valores propios de A.

Los subespacios propios $\mathcal{V}_1$ y $\mathcal{V}_2$ de los valores propios $\lambda_1 = 4$ y $\lambda_2 = 2$ son las direcciones propias generadas, por ejemplo, por los vectores propios

$$v^{(1)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} , \quad v^{(2)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} ,$$

respectivamente.

En la figura 2.1 se representa el comportamiento de la transformación hecha por A sobre las direcciones propias.

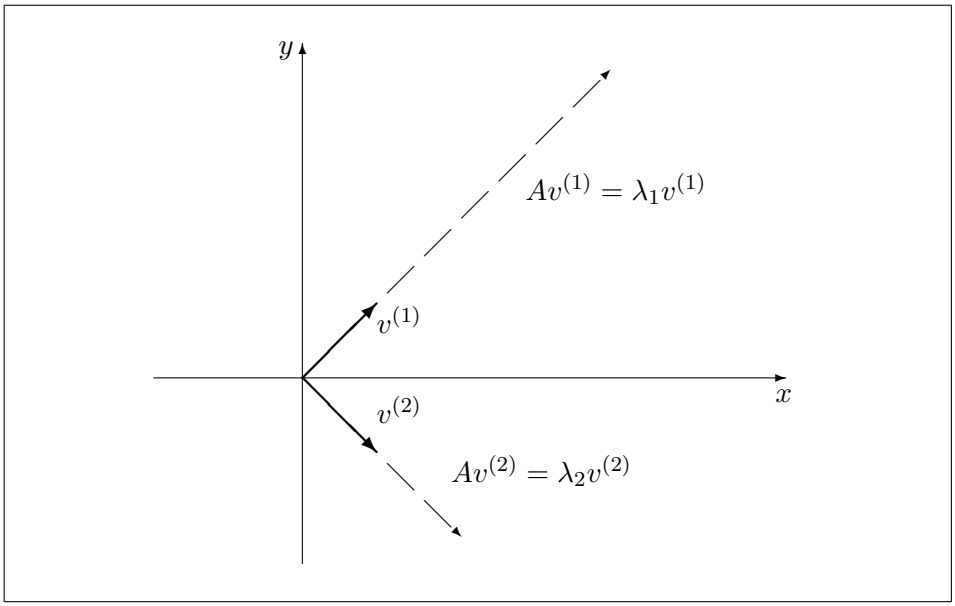


Figura 2.1: Efecto de A sobre sus vectores propios.

Normas vectoriales y normas matriciales

Sea E un espacio vectorial real o complejo. Una *norma* en E es una aplicación

$$\begin{array}{rccc} \| \ \| : & E & \longrightarrow & \mathbb{R}^+ \\ & x & \to & \|x\| \end{array}$$

que cumple

- $\|x\| = 0 \ \Leftrightarrow \ x = 0$;
- $\|cx\| = |c| \ \|x\| \quad \forall$ escalar $c \ \ \forall x \in E$;
- $\|x + y\| \leq \|x\| + \|y\| \ \ \forall x, y \in E$.

Las *normas vectoriales* son aquellas que están definidas en espacios vectoriales de la forma $E = \mathbb{R}^n$ ó $E = \mathbb{C}^n$. Las normas vectoriales más usadas son las *normas de Hölder*

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \ \ (p \geq 1) \ .$$

Los casos particulares más importantes por su uso son:

- *La norma suma de módulos*

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \quad (p = 1) \ ,$$

- *la norma euclídea*

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} \quad (p = 2) \ .$$

Otra norma muy usada es el caso límite (cuando p tiende a infinito) de las normas de Hölder, la *norma del máximo*

$$\|x\|_\infty = \max_{i=1 \div n} |x_i| \ .$$

Una *norma matricial* es una norma en el espacio vectorial $\mathcal{M}_{n,n}$ que sea *multiplicativa*; esto es, que cumpla

$$\|AB\| \ \leq \ \|A\| \, \|B\| \ \ \forall A, B \in \mathcal{M}_{n,n} \ .$$

Una norma matricial $\| \ \|$ es *consistente con una norma vectorial* (que denotaremos igual) si y sólo si

$$\|Ax\| \ \leq \ \|A\| \, \|x\| \ \ \forall A \in \mathcal{M}_{n,n} \ , \ \ \forall x \in \mathbb{R}^n \ .$$

Dada una norma vectorial $\| \ \|$, siempre puede definirse una norma matricial consistente con ella mediante

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \ ,$$

y recibe el nombre de *norma matricial subordinada* a la norma vectorial dada.

2.1.3 Propiedades importantes

Se enuncian a continuación algunas propiedades importantes que pueden ser útiles en la resolución de sistemas lineales, en el cálculo de determinantes e inversas de matrices y en la resolución de problemas de valores y de vectores propios de matrices.

Matrices

1. $(A + B)^\top = A^\top + B^\top$.
2. $\overline{A + B} = \overline{A} + \overline{B}$.
3. $(AB)^\top = B^\top A^\top$.
4. $\overline{AB} = \overline{A} \ \overline{B}$.
5. $\det A^\top = \det A$.
6. $\det AB = \det A \det B$.

7. $(AB)^{-1} = B^{-1}A^{-1}$, si A y B son regulares.

8. Una matriz cuadrada tiene inversa si y sólo si es regular.

9. *Lema de Schur.* Toda matriz A es semejante a una matriz triangular superior, mediante una transformación de semejanza con una matriz unitaria.

10. *Criterio de Sylvester.* Una matriz simétrica es definida positiva si y sólo si tiene todos los determinantes principales positivos.

Valores y vectores propios

1. Las matrices A y $A^\top$ tienen los mismos valores propios. Los vectores propios por la derecha son ortogonales a los vectores propios por la izquierda de valores propios diferentes.

2. Una matriz A es regular si y sólo si tiene todos los valores propios diferentes de cero; en tal caso, los valores propios de A^{-1} son los recíprocos de los valores propios de A y v es un vector propio de valor propio λ de A si y sólo si v es un vector propio de valor propio $\frac{1}{\lambda}$ de A^{-1}.

3. *Teorema de Gerschgorin.* Los valores propios de A están localizados, dentro del plano complejo, en la unión $\mathcal{F}$ de los discos

$$F_i = \{\lambda \mid |\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\} \ (i = 1 \div n) \ ,$$

y también, en la unión $\mathcal{C}$ de los discos

$$C_j = \{\lambda \mid |\lambda - a_{jj}| \leq \sum_{i \neq j} |a_{ij}|\} \ (j = 1 \div n) \ .$$

Además, si una componente conexa de $\mathcal{F}$ (o de $\mathcal{C}$) está formada por d discos F_i (o C_j), ésta contiene exactamente d valores propios de A contados según su multiplicidad. Consecuentemente, toda matriz estrictamente diagonal dominante es inversible.

4. Dos matrices semejantes A y B tienen los mismos valores propios, y v es un vector propio de valor propio λ de A si y sólo si $C^{-1}v$ es un vector propio de valor propio λ de B, siendo C la transformación de semejanza de A a B.

5. Si una matriz A es de la forma

$$A = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & A_{22} \end{array}\right) \ ,$$

con las matrices A_{11} y A_{22} cuadradas, el conjunto de valores propios de A es la unión de los conjuntos de valores propios de A_{11} y de A_{22}. Así, los valores propios de una matriz triangular son los elementos de la diagonal.

6. Si p es un polinomio no nulo, λ es un valor propio de A si y sólo si $p(\lambda)$ lo es de $p(A)$; v es un vector propio de valor propio λ de A si y sólo si es vector propio de valor propio $p(\lambda)$ de $p(A)$.

7. Sean $\lambda_1, \ldots, \lambda_n$ los valores propios de A, repetidos según su multiplicidad; entonces,

$$\sum_{j=1}^{n} \lambda_j = \text{tr } A \equiv \sum_{j=1}^{n} a_{jj} \ , \quad \prod_{j=1}^{n} \lambda_j = \det A \ .$$

8. Sea A diagonalizable por una transformación de semejanza V. Entonces $D = V^{-1}AV$ es diagonal, los elementos de la diagonal de D son los valores propios de A, las columnas de V forman una base de vectores propios (por la derecha) de A, y las filas de V^{-1}, una base de vectores propios por la izquierda de A.

9. Vectores propios correspondientes a valores propios diferentes son linealmente independientes. Por lo tanto, si una matriz tiene todos los valores propios diferentes es diagonalizable.

10. Los valores propios de una matriz hermítica son reales. Toda matriz hermítica A es diagonalizable mediante una transformación de semejanza con una matriz unitaria; si U^*AU es diagonal con U unitaria, las columnas de U forman una base de vectores propios ortonormales de A. En el caso real, si A es simétrica, es diagonalizable mediante una transformación con una matriz ortogonal, cuyas columnas forman también una base de vectores propios ortonormales de A.

Normas vectoriales y matriciales

1. Las normas matriciales subordinadas a las normas vectoriales $\| \ \|_1$, $\| \ \|_2$ y $\| \ \|_\infty$ resultan ser

$$\|A\|_1 = \max_{j=1\div n} \sum_{i=1}^{n} |a_{ij}| \ , \tag{2.1}$$

$$\|A\|_2 = \sqrt{\rho(A^*A)} \ , \tag{2.2}$$

$$\|A\|_\infty = \max_{i=1\div n} \sum_{j=1}^{n} |a_{ij}| \tag{2.3}$$

   (véase el problema 2.4 para las normas $\| \ \|_1$ y $\| \ \|_\infty$).

2. El radio espectral de una matriz es menor o igual que cualquier norma matricial de dicha matriz:

$$\rho(A) \leq \|A\| \ .$$

3. Dada una matriz A, para cualquier $\epsilon > 0$ se puede encontrar una norma matricial, subordinada a alguna norma vectorial, tal que

$$\|A\| \leq \rho(A) + \epsilon \ .$$

2.2 SISTEMAS DE ECUACIONES

2.2.1 Introducción

Esta sección está dedicada a la resolución de sistemas de ecuaciones lineales y a problemas relacionados como son el cálculo de determinantes e inversas de matrices.

Los métodos de resolución de sistemas lineales se clasifican en métodos directos y métodos iterativos.

Los *métodos directos* permiten obtener la solución del sistema después de un número finito de pasos. Hay tipos de matrices con la propiedad de que es extraordinariamente fácil la resolución directa de los sistemas correspondientes. Entre estos tipos, cabe destacar el de las matrices triangulares y el de las matrices unitarias (ortogonales, si son reales). Para otros tipos de matrices más generales presentaremos los métodos gaussianos y los métodos de ortogonalización. Las técnicas usadas serán extendidas de manera natural al cálculo de determinantes e inversas de matrices cuadradas.

El análisis del error en el proceso de resolución de sistemas lineales sirve finalmente para obtener acotaciones de los errores en la solución, producidos por los errores de los datos y por los de las operaciones.

Los *métodos iterativos* permiten obtener la solución del sistema a través de aproximaciones sucesivas. En la práctica, estos métodos son buenos competidores de los métodos directos cuando las matrices son *escasas*; es decir, que tienen relativamente pocos elementos no nulos. Se analizan los métodos iterativos más usuales: de Jacobi, de Gauss-Seidel y de sobrerrelajación.

2.2.2 Resolución de sistemas triangulares

Consideremos el sistema $Ax = b$, donde A es triangular superior:

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots & \\ a_{nn}x_n &= b_n \end{aligned}\right\} .$$

El método de resolución consiste en sustituir $x_n, \ldots, x_{i+1}$ (ya conocidos) en la ecuación i-ésima y despejar después x_i, comenzando por la última ecuación:

$$x_n = \frac{b_n}{a_{nn}}, \quad x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right) \ (i = n-1 \div 1) \ .$$

Nótese que $a_{ii} \neq 0$ $(i = 1 \div n)$, ya que hemos supuesto que $\det A \neq 0$.

El número de operaciones aritméticas a realizar es de $\frac{n(n-1)}{2}$ multiplicaciones, n divisiones y $\frac{n(n-1)}{2}$ sumas/restas.

De ahora en adelante consideraremos que una *operación* será igual a una multiplicación/división junto con una suma/resta. Así, para valores grandes de n, el número de operaciones del método que consideramos es $\frac{n^2}{2} + \mathcal{O}(n)$, donde $\mathcal{O}(n^k)$ indicará un sumando del orden de n^k, cuando $n \to \infty$, según se define en el apartado 4.3.1. En el caso que nos ocupa ahora, no es más que un polinomio de grado k en n.

Este método para sistemas triangulares superiores recibe el nombre de *sustitución hacia atrás*. Análogamente, puede hacerse *sustitución hacia adelante* para sistemas triangulares inferiores.

2.2.3 Métodos gaussianos

Los *métodos gaussianos* son los métodos clásicos de reducción de sistemas a forma triangular y están basados en la anulación de los elementos subdiagonales de la matriz mediante combinaciones lineales simples de las ecuaciones (*eliminación gaussiana*).

En todos ellos se observan dos fases:

- En la primera se produce la eliminación gaussiana de todos los elementos subdiagonales: el método de Gauss trabaja con la matriz y el vector de términos independientes; los métodos LU y de Cholesky, de Doolittle y de Crout producen una factorización de la matriz del sistema como producto de una matriz triangular superior y de una matriz triangular inferior.
- En la segunda fase hay que resolver un sistema triangular superior en el método de Gauss, y dos sistemas triangulares, en los otros métodos citados.

La primera fase marca las diferencias entre los diversos métodos.

Método de Gauss

El *método de Gauss* consta de dos fases. En la fase llamada de eliminación gaussiana, se efectúa la reducción del sistema inicial a un sistema triangular superior.

Este objetivo se consigue mediante un proceso de eliminación que, partiendo de la matriz $A = A_1 = (a_{ij}^{(1)})$, va obteniendo matrices $A_k = (a_{ij}^{(k)})$ $(k = 2 \div n)$, con $a_{ij}^{(k)} = 0$ $(i > j,\ j = 1 \div k-1)$,

$$A_k = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & a_{1k}^{(k)} & \cdots & a_{1n}^{(k)} \\ & a_{22}^{(k)} & \cdots & a_{2k}^{(k)} & \cdots & a_{2n}^{(k)} \\ & & \ddots & \vdots & \cdots & \vdots \\ & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & \vdots & \cdots & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix} \ (k = 2 \div n) \ ,$$

de forma que A_n resulte ser triangular superior. Paralelamente a estas modificaciones sobre la matriz del sistema, se realizan modificaciones sobre el vector b de términos independientes, con el fin de que las transformaciones del sistema no alteren las soluciones; así obtenemos los vectores $b^{(k)}$ de componentes $b_i^{(k)}$ $(i = 1 \div n)$ $(k = 2 \div n)$.

El paso k-ésimo de esta primera fase se realiza cambiando sólo los valores de los elementos de las filas que van de la $(k+1)$-ésima hasta la última y los de los elementos correspondientes del vector de términos independientes, según:

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \ , \quad b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)} \ ,$$

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)} \ (j = k \div n)$$
$$(i = k+1 \div n) \ \ (k = 1 \div n-1) \ . \tag{2.4}$$

En la segunda fase se resuelve el sistema triangular superior $A_n x = b^{(n)}$ por el procedimiento de sustitución hacia atrás detallado en el apartado 2.2.2.

Para n grande, el número de operaciones que se realizan, a lo largo de todo el proceso, es de $\frac{n^3}{3} + \mathcal{O}(n^2)$. Para matrices con características especiales el método puede optimizarse mucho en el sentido de evitar operaciones innecesarias (véase el problema 2.2).

Nótese que, en la etapa k-ésima del proceso de reducción, tenemos que dividir por el elemento diagonal $a_{kk}^{(k)}$, llamado *pivote*. Conviene hacer tres observaciones al respecto:

1. Los determinantes principales coinciden con los productos de pivotes

   $$\Delta_k = a_{11}^{(1)} \cdots a_{kk}^{(k)} \ .$$

   El método de Gauss puede, así, aplicarse directamente si y sólo si todos estos determinantes son no nulos (véase el problema 2.1).

2. Si $a_{kk}^{(k)} = 0$, entonces podemos buscar $a_{ik}^{(k)} \neq 0$ $(i > k)$ (que existe siempre, ya que $\det A \neq 0$) y permutar la ecuación k-ésima con la i-ésima; así se puede continuar el proceso.

3. Si $a_{kk}^{(k)} \neq 0$ pero es pequeño, el método de Gauss, aunque pueda ser aplicado, es muy inestable numéricamente en el sentido de que los errores de la solución, propagados a partir de los errores de los datos, pueden aumentar notablemente. Así, convendrá modificar el método de Gauss también en este caso.

   Podemos escoger entre las opciones siguientes:

   - *Pivotaje maximal por columnas.* Se toma como nuevo pivote $a_{kk}^{(k)}$ el elemento de máximo valor absoluto entre los elementos $a_{ik}^{(k)}$ $(i = k \div n)$.
   - *Pivotaje completo.* El nuevo pivote pasa a ser el elemento de máximo valor absoluto entre los elementos $a_{ij}^{(k)}$ $(i, j = k \div n)$.

   En el primer caso, tenemos que cambiar filas y términos independientes; en el segundo, filas y términos independientes, así como columnas y las incógnitas correspondientes.

Métodos LU y de Cholesky

El *método LU* produce primero la *factorización LU de la matriz*

$$A = LU \ ,$$

donde $L = (l_{ik})$ es triangular inferior con unos en la diagonal y $U = (u_{kj})$, triangular superior.

Esta factorización se puede realizar de manera única siempre que todos los determinantes principales de A sean no nulos: condición equivalente a la de la posibilidad de

aplicación del método de Gauss sin pivotaje. Los elementos de L y de U se pueden calcular entonces usando sólo los elementos de la matriz A mediante la recurrencia:

$$\begin{aligned} l_{kk} = 1, \quad l_{ik} &= m_{ik} \ (i = k+1 \div n) \ , \\ u_{kj} &= a_{kj}^{(k)} \ (j = k \div n) \\ & \quad (k = 1 \div n) \ . \end{aligned} \tag{2.5}$$

En caso de que no se cumpla la condición exigida y siempre que A sea regular, será posible permutar las ecuaciones de manera que la nueva matriz admita una tal factorización.

Una vez realizada la factorización LU, la solución del sistema $Ax = b$ es también la solución del sistema $Ux = y$, siendo y la solución de $Ly = b$. Así, se tendrán que resolver sucesivamente los sistemas triangulares $Ly = b$ y $Ux = y$.

Esta segunda fase es común a todos los métodos gaussianos que siguen y ya no será explicitada.

Para matrices simétricas, es mejor realizar la factorización

$$A = LDL^\top,$$

donde $L = (l_{ik})$ es una matriz triangular inferior con unos en la diagonal y D es una matriz diagonal con elementos diagonales d_{kk} $(k = 1 \div n)$:

$$\begin{aligned} d_{kk} &= a_{kk} - \sum_{r=1}^{k-1} l_{kr}^2 d_{rr} \ , \\ l_{ik} &= \frac{1}{d_{kk}} \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir} d_{rr} l_{kr} \right) \ (i = k+1 \div n) \\ & \quad (k = 1 \div n) \ . \end{aligned} \tag{2.6}$$

Nótese que el número de operaciones se reduce esencialmente a la mitad.

Para matrices definidas positivas, los elementos diagonales d_{kk} son positivos, según el criterio de Sylvester (véase la propiedad 10 del apartado 2.1.3). Si $\mathcal{L}$ es la matriz $LD^{1/2}$, con $D^{1/2} = \mathrm{diag}(d_{11}^{1/2}, \ldots, d_{nn}^{1/2})$, entonces A puede expresarse en la forma

$$A = \mathcal{L}\mathcal{L}^\top,$$

llamada *factorización de Cholesky*. Dicha factorización, con elementos diagonales de $\mathcal{L}$ positivos, existe y es única si y sólo si A es definida positiva.

En el problema 2.1 se estudia el efecto de la eliminación gaussiana sobre matrices definidas positivas y se expone un ejemplo de factorización de Cholesky.

Si la matriz A es compleja, existen factorizaciones análogas, pero sustituyendo la transposición por la conjugación: $A = LDL^*$ y $A = \mathcal{L}\mathcal{L}^*$.

Método de Doolittle

El *método de Doolittle* persigue el mismo objetivo inicial que el método LU, pero ofrece, bajo las mismas condiciones, un algoritmo de cálculo de los elementos de las matrices que

no utiliza la eliminación gaussiana:

$$\begin{aligned} u_{kj} &= a_{kj} - \sum_{r=1}^{k-1} l_{kr}u_{rj} \ (j = k \div n) \ , \\ l_{kk} = 1 \ , \ \ l_{ik} &= \frac{1}{u_{kk}} \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir}u_{rk} \right) \ (i = k+1 \div n) \\ & \quad (k = 1 \div n) \ . \end{aligned} \tag{2.7}$$

Los elementos u_{kk} $(k = 1 \div n)$ serán no nulos, bajo las hipótesis de existencia de la factorización LU: $\det(A)_k \neq 0$ $(k = 1 \div n)$.

El orden de cálculo es importante: u_{1j} $(j = 1 \div n)$, l_{i1} $(i = 2 \div n)$, u_{2j} $(j = 2 \div n)$, l_{i2} $(i = 3 \div n)$, ..., $u_{n-1,n-1}$, $u_{n-1,n}$, $l_{n,n-1}$ y u_{nn}.

Método de Crout

El *método de Crout* es similar al de Doolittle pero con la diferencia de que ahora es la matriz U la que posee los elementos diagonales iguales a 1:

$$\begin{aligned} l_{ik} &= a_{ik} - \sum_{r=1}^{k-1} l_{ir}u_{rk} \ (i = k \div n) \ , \\ u_{kk} = 1 \ , \ \ u_{kj} &= \frac{1}{l_{kk}} \left(a_{kj} - \sum_{r=1}^{k-1} l_{kr}u_{rj} \right) \ (j = k+1 \div n) \\ & \quad (k = 1 \div n) \ . \end{aligned} \tag{2.8}$$

Los elementos l_{kk} $(k = 1 \div n)$ serán no nulos, bajo las mismas hipótesis que antes.

El orden de cálculo es: l_{i1} $(i = 1 \div n)$, u_{1j} $(j = 2 \div n)$, l_{i2} $(i = 2 \div n)$, u_{2j} $(j = 3 \div n)$, ..., $l_{n-1,n-1}$, $l_{n,n-1}$, $u_{n-1,n}$ y l_{nn}.

2.2.4 Métodos de ortogonalización

Los *métodos de ortogonalización* consiguen la reducción a forma triangular mediante el uso de matrices ortogonales. Estos métodos están basados en factorizaciones QR de matrices, que se definen a continuación, y reciben también el nombre de *métodos QR*.

Como en los métodos gaussianos, se observan dos fases:

- En la primera, se factoriza A en la forma:

  $$A = QR,$$

  donde Q es ortogonal y R triangular superior. Esta factorización se llama *factorización QR* de la matriz A.

- En la segunda, se acaba la resolución del sistema $Ax = QRx = b$, resolviendo el sistema triangular superior $Rx = Q^\top b$ por el método de sustitución hacia atrás.

Dicha factorización se realiza usando alguno de los métodos QR siguientes que pueden ser generalizados, de forma natural, a matrices no cuadradas (véase el apartado 3.2.3).

Método de ortogonalización modificado de Gram-Schmidt

El *método de ortogonalización modificado de Gram-Schmidt* es una mejora, desde un punto de vista numérico, del método de ortogonalización clásico de Gram-Schmidt (véanse todos los detalles en el apartado 3.2.3).

Comenzando con $A_1 = A$, una vez conocida

$$A_k = (q_1 \ \dots \ q_{k-1} \ a_k^{(k)} \ \dots \ a_n^{(k)})$$

con columnas q_j $(j = 1 \div k-1)$ y $a_s^{(k)}$ $(s = k \div n)$ cumpliendo las relaciones de ortogonalidad

$$q_j^\top q_l = \delta_{jl} \ , \quad q_j^\top a_s = 0 \ \ (j, l = 1 \div k-1, \ s = k \div n) \ ,$$

se normaliza la k-ésima columna y se ortogonalizan, respecto a ella, todas las que la siguen:

$$r_{kk} = \| \ a_k^{(k)} \ \|_2 \ , \quad q_k = \frac{a_k^{(k)}}{r_{kk}} \ ; \tag{2.9}$$

$$r_{ks} = q_k^\top a_s^{(k)} \ , \quad a_s^{(k+1)} = a_s^{(k)} - r_{ks} q_k \qquad (s = k+1 \div n) \ . \tag{2.10}$$

Se obtiene

$$A_{k+1} = (q_1 \ \dots \ q_k \ a_{k+1}^{(k+1)} \ \dots \ a_n^{(k+1)}) \ ,$$

verificando las mismas condiciones de ortogonalidad anteriores, sustituyendo k por $k+1$ $(k = 1 \div n)$.

Después de n pasos,

$$A_{n+1} = (q_1 \ q_2 \ \dots \ q_n)$$

es una matriz ortogonal, porque $q_j^\top q_l = \delta_{jl}$ $(j, l = 1 \div n)$.

Las matrices $Q = A_{n+1}$ y $R = (r_{ks})$ forman una factorización QR de A. Como A es regular, ésta es única si imponemos que los elementos diagonales de R sean positivos.

Matrices de Householder

Antes de iniciar la descripción del método de ortogonalización de Householder, presentaremos las matrices de Householder, que serán útiles también en otros apartados.

Para cada vector no nulo u de $\mathbb{R}^n$, definimos la *matriz de Householder* $P(u)$ asociada a u mediante

$$P(u) \equiv I - \frac{2}{u^\top u} u u^\top \ .$$

Estas matrices satisfacen las propiedades siguientes:

- $P(cu) = P(u)$, si $c \neq 0$. En otras palabras, las matrices de Householder están realmente asociadas a *direcciones* o subespacios unidimensionales más que a vectores no nulos.
- $P(u)$ es simétrica: $P(u)^\top = P(u)$.

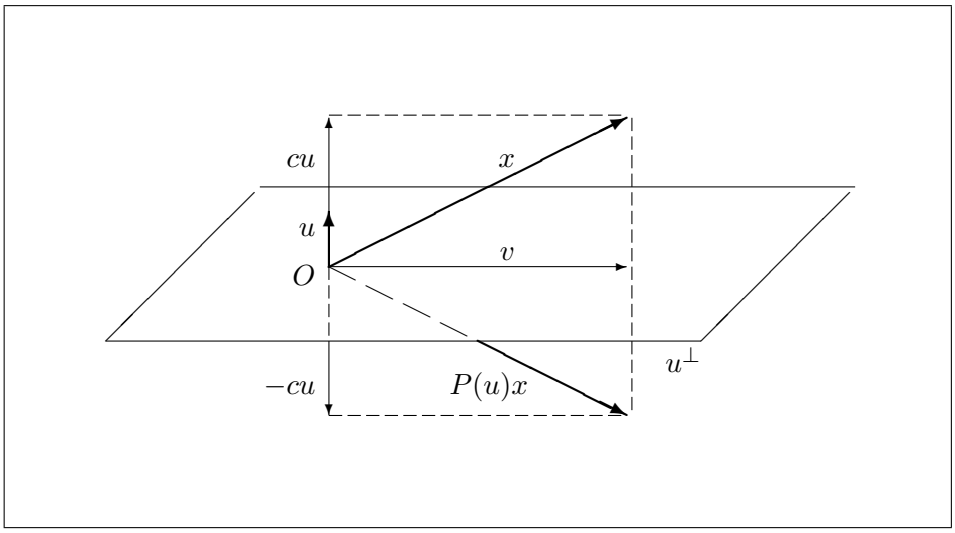


Figura 2.2: $P(u)x$ es la reflexión de x respecto al hiperplano $u^\perp$.

- Si denotamos por $u^\perp$ el subespacio ortogonal a u:

$$u^\perp = \{v \in \mathbb{R}^n : u^\top v = 0\} \ ,$$

entonces $P(u)v = v \ \forall v \in u^\perp$. Como $P(u)u = -u$, resulta que $P(u)$ representa una simetría (o reflexión) respecto al hiperplano $u^\perp$. Por lo tanto todo $x \in \mathbb{R}^n$ se puede descomponer como

$$x = cu + v$$

con

$$c = \frac{u^\top x}{u^\top u} \ , \quad v = x - cu \in u^\perp \ ,$$

y

$$P(u)x = -cu + v$$

(véase la figura 2.2).

- $P(u)P(u) = I$ o $P(u)^{-1} = P(u) = P(u)^\top$. En particular, $P(u)$ es ortogonal y, por lo tanto, conserva la norma euclídea:

$$\|P(u)x\|_2^2 = x^\top P(u)^\top P(u)x = x^\top P(u)P(u)x = x^\top x = \|x\|_2^2 \ .$$

- $\det P(u) = -1$. Esto es consecuencia del hecho de que $P(u)$ es una reflexión.

- Dado un vector $x \in \mathbb{R}^n$, el cálculo de $P(u)x$ se lleva a cabo muy fácilmente de la forma siguiente: si previamente se calcula

$$\alpha = \frac{2}{u^\top u} \ ,$$

entonces

$$P(u)x = x - (\alpha(u^\top x))u$$

sólo requiere $2n + 1$ operaciones.

- Si $\|a\|_2 = \|b\|_2$, $a \neq b$, entonces $P(a-b)a = b$. Esto responde a la pregunta siguiente: dados $a, b \in \mathbb{R}^n$, ¿existe $u \neq 0$ tal que $P(u)a = b$? Como $P(u)$ es ortogonal, hay que suponer $\|a\|_2 = \|b\|_2$. Escribiendo

$$a = cu + v$$

con

$$c = \frac{u^\top a}{u^\top u} \ , \quad v = a - cu \in u^\perp \ ,$$

$P(u)a = b$ equivale a $b = -cu + v$ y, por lo tanto, $a - b = 2cu$; es decir, u es paralelo al vector $a - b$, que suponemos no nulo. Así, podemos escoger u mediante $u = a - b$. Nótese que u está determinado excepto una constante multiplicativa no nula.

- En particular, si $a \in \mathbb{R}^n$ no tiene las $n-1$ últimas componentes nulas: $a_2^2 + a_3^2 + \cdots + a_n^2 > 0$ o, en otras palabras, a no es de la forma $ce^{(1)}$, entonces

$$P(a + \|a\|_2 e^{(1)})a = -\|a\|_2 e^{(1)} \ , \quad P(a - \|a\|_2 e^{(1)})a = \|a\|_2 e^{(1)} \ ;$$

o sea, podemos encontrar dos matrices de Householder que transforman a en un vector muy simple, con todas las componentes nulas, excepto la primera: un múltiplo del primer elemento de la base canónica denotado aquí por $e^{(1)} = (1 \ \ 0 \ \ \ldots \ \ 0)^\top$.

Más brevemente: si $s = \|a\|_2$ o $s = -\|a\|_2$, $P(a - se^{(1)})a = se^{(1)}$.

Con el fin de evitar cancelaciones, se suele escoger s con el signo opuesto al de a_1.

El cálculo con matrices de este tipo se realiza tal como se detalla a continuación.

Cálculo de $P(u) = P(a - se^{(1)})$:
Si $a = (a_1, a_2, \cdots, a_n)^\top$ con $a_2^2 + \cdots + a_n^2 \neq 0$, se calcula:

- s tal que

$$s^2 = \|a\|_2^2 = \sum_{i=1}^{n} a_i^2 \ .$$

Así, $s = \pm\|a\|_2$.

- $u_1 = a_1 - s$, $u_i = a_i$ $(i = 2 \div n)$. Con el fin de evitar cancelaciones en u_1, se suele escoger $s = \|a\|_2$ si $a_1 \leq 0$ y $s = -\|a\|_2$ si $a_1 > 0$.
-

$$\alpha = \frac{2}{u^\top u} = \frac{1}{\|a\|_2(\|a\|_2 + |a_1|)} = \frac{-1}{su_1} \ .$$

Entonces $P(u) = P(a - se^{(1)}) = I - \alpha uu^\top$ satisface $P(u)a = se^{(1)}$, donde el cálculo de u y α requiere en total $n+2$ operaciones y una raíz cuadrada. Una vez tenemos u y α podemos afrontar diversos cálculos.

Cálculo de $P(u)x$:
$P(u)x = x - (\alpha(u^\top x))u$ requiere $2n+1$ operaciones.

Cálculo de $B = P(u)A$:
Se calcula

- $w = \alpha u$.
- $p \in \mathbb{R}^n$ tal que $p^\top = w^\top A$ (es decir, $p = A^\top w$).
- $B = P(u)A = (I - \alpha u u^\top)A = A - up^\top$; en componentes,

$$b_{ij} = a_{ij} - u_i p_j \quad (i, j = 1 \div n) \ .$$

El cálculo requiere, en total, $2n^2 + n$ operaciones.

CÁLCULO DE $A' = P(u)AP(u)$:
$A' = P(u)AP(u) = BP(u) = B - (\alpha Bu)u^\top$. Se calcula:

- $w = \alpha u$, $p = A^\top w$ y $B = A - up^\top$ según el procedimiento anterior.
- $q = Bw$.
- $A' = B - qu^\top$; en componentes,

$$a'_{ij} = b_{ij} - q_i u_j \quad (i, j = 1 \div n) \ .$$

El cálculo requiere, en total, $4n^2 + n$ operaciones.

CÁLCULO DE $A' = P(u)AP(u)$, A SIMÉTRICA:
Solamente se tendrán que conocer los elementos a'_{ij} para $i \leq j$; por ejemplo, calculando $w = \alpha u$ y $p = A^\top w$ en primer lugar. Ahora no hay que calcular B:

$$q = Bw = P(u)Aw = P(u)A^\top w = P(u)p = p - (w^\top p)u$$

se calcula con $2n$ operaciones, y

$$A' = A - up^\top - qu^\top \ ;$$

en componentes,

$$a'_{ij} = a_{ij} - u_i p_j - q_i u_j \quad (1 \leq y \leq j \leq n) \ .$$

Así, el cálculo total requiere $2n^2 + 4n$ operaciones.
De hecho, la expresión para A' se puede escribir de manera simétrica introduciendo el vector

$$q^{(s)} = \frac{q + p}{2} = p - \frac{1}{2}(w^\top p)u \ ,$$

y procediendo a los cálculos:

- $w = \alpha u$.
- $p = Aw$.
- $q^{(s)} = p - \frac{1}{2}(w^\top p)u$.
- $A' = A - uq^{(s)\top} - q^{(s)}u^\top$; en componentes,

$$a'_{ij} = a_{ij} - u_i q_j^{(s)} - q_i^{(s)} u_j \quad (1 \leq y \leq j \leq n) \ .$$

Los procedimientos de cálculo que se acaban de detallar serán muy útiles en el apartado que sigue, en el apartado 2.3.2 al tratar la deflación de matrices y en el apartado 2.3.5 al tratar los métodos de reducción de matrices para el cálculo de sus valores y vectores propios.

Método de ortogonalización de Householder

El *método de ortogonalización de Householder* para establecer una factorización QR de una matriz A con n filas y n columnas consta de $n-1$ pasos. El proceso se inicia con $A_1 = A$ y, en el paso k-ésimo ($k = 1 \div n-1$), se parte de la matriz encontrada A_k de la forma

$$A_k = \left(\begin{array}{ccc|cc} \cdots & \cdots & \cdots & \cdots & \cdots \\ & \ddots & \cdots & \cdots & \cdots \\ & & \ddots & \cdots & \cdots \\ \hline & & & \multicolumn{2}{c}{\tilde{A}_k} \end{array}\right) = \left(\begin{array}{c|c} R_k & M_k \\ \hline 0 & \tilde{A}_k \end{array}\right) ,$$

donde R_k es una matriz $(k-1)$x$(k-1)$ triangular superior, y M_k, $\tilde{A}_k$ son matrices $(k-1)$x$(n-k+1)$ y $(n-k+1)$x$(n-k+1)$, respectivamente, siendo

$$\tilde{a}^{(k)} = \begin{pmatrix} a_{kk}^{(k)} \\ a_{k+1,k}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{pmatrix} \in \mathbb{R}^{n-k+1}$$

la primera columna de $\tilde{A}_k$.

Si las últimas $n-k$ componentes de $\tilde{a}^{(k)}$ son nulas

$$\sum_{i=k+1}^{n} (a_{ik}^{(k)})^2 = 0 ,$$

la matriz formada por las k primeras filas de las k primeras columnas de A_k es ya triangular superior, y escogemos $P_k = I$, $A_{k+1} = P_k A_k = A_k$.

En caso contrario, escogemos s_k tal que

$$s_k^2 = \|\tilde{a}^{(k)}\|_2^2 = \sum_{i=k}^{n} (a_{ik}^{(k)})^2 ;$$

por ejemplo, $s_k = \|\tilde{a}^{(k)}\|_2$, si $a_{kk}^{(k)} \leq 0$, y $s_k = -\|\tilde{a}^{(k)}\|$, si $a_{kk}^{(k)} > 0$. Tomamos entonces la matriz $(n-k+1)$x$(n-k+1)$ de Householder $\tilde{P}_k$ tal que

$$\tilde{P}_k \tilde{a}^{(k)} = \begin{pmatrix} s_k \\ 0 \\ \vdots \\ 0 \end{pmatrix} = s_k e^{(1)} \in \mathbb{R}^{n-k+1} ;$$

es decir,

$$\tilde{P}_k = \tilde{P}(\tilde{u}^{(k)}) = I_{n-k+1} - \alpha_k \tilde{u}^{(k)} \tilde{u}^{(k)^\top} ,$$

con

$$\begin{aligned} \tilde{u}^{(k)} &= \begin{pmatrix} a_{kk}^{(k)} - s_k \\ a_{k+1,k}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{pmatrix} = \tilde{a}^{(k)} - s_k e^{(1)} \in \mathbb{R}^{n-k+1} \ , \\ \alpha_k &= \frac{-1}{s_k(a_{kk}^{(k)} - s_k)} = \frac{1}{s_k(s_k - a_{kk}^{(k)})} \ . \end{aligned}$$

Definiendo entonces

$$P_k = \left(\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & \tilde{P}_k \end{array}\right) \ ,$$

o equivalentemente $P_k = P(u^{(k)})$, con

$$u^{(k)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{u}^{(k)} \end{pmatrix} \ ,$$

resulta que

$$\begin{aligned} A_{k+1} \equiv P_k A_k &= \left(\begin{array}{c|c} R_k & M_k \\ \hline 0 & \tilde{P}_k \tilde{A}_k \end{array}\right) = \left(\begin{array}{ccc|c|c} & R_k & & & \tilde{M}_k \\ \hline 0 & \cdots & 0 & s_k & \cdots \\ \hline & & & 0 & \\ & 0 & & \vdots & \tilde{A}_{k+1} \\ & & & 0 & \end{array}\right) \\ &= \left(\begin{array}{c|c} R_{k+1} & M_{k+1} \\ \hline 0 & \tilde{A}_{k+1} \end{array}\right) \ ; \end{aligned}$$

es decir, la matriz R_{k+1}, formada por las k primeras filas de las k primeras columnas de A_{k+1}, es triangular superior y tiene la matriz $(n-k)$xk nula bajo ella.

La matriz $R \equiv R_n = A_n$ es pues triangular superior y se tiene

$$R = P_{n-1} P_{n-2} \cdots P_1 A$$

con P_k $(k = 1 \div n-1)$ matrices de Householder o matrices identidad y, por lo tanto, cumpliendo $P_k = P_k^\top = P_k^{-1}$; de aquí se deduce la factorización

$$A = QR \ , \ Q = P_1 P_2 \cdots P_{n-1} \ .$$

El cálculo de $\tilde{B}_k \equiv \tilde{P}_k \tilde{A}_k$ requiere $2(n-k+1)^2 + \mathcal{O}(n-k)$ operaciones y una raíz cuadrada. Sumando para $k = 1 \div n-1$, se deduce que deben efectuarse un total de $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ operaciones y $n-1$ raíces cuadradas para obtener

$$R = P_{n-1} P_{n-2} \cdots P_1 A \ .$$

Como el sistema $Ax = b$ equivale a

$$Rx = z \equiv P_{n-1} P_1 P_2 b,$$

el cálculo de z requiere $2n + 2(n-1) + \cdots + 2 \cdot 2 = n^2 + \mathcal{O}(n)$ operaciones y la resolución del sistema $Rx = z$, $\frac{1}{2}n^2 + \mathcal{O}(n)$. Obsérvese que, usando un método QR, deben realizarse el doble de operaciones que con el método de Gauss pero, en contrapartida, se obtiene en general un sistema triangular superior mejor condicionado, como se observará en el apartado 2.2.6.

2.2.5 Cálculo de determinantes e inversas de matrices

Determinantes

Las transformaciones que experimenta una matriz A a lo largo del proceso de eliminación gaussiana con pivotaje consisten en realizar combinaciones lineales de filas de manera que, excepto el signo que puede variar al permutar filas y permutar columnas, se conserva el determinante. Tenemos así que

$$\det A = \sigma \ \det A_n = \sigma \prod_{k=1}^{n} a_{kk}^{(k)} \ ,$$

donde σ vale +1 ó -1 y A_n es la matriz triangular final que tiene los elementos $a_{kk}^{(k)}$ en la diagonal.

Hay que hacer dos observaciones:

- El valor de σ será +1 ó -1 si se han realizado un número par o impar, respectivamente, de intercambios entre filas y entre columnas.
- Podría darse el caso de que, en el k-ésimo paso de la eliminación gaussiana, $a_{ik}^{(k)} = 0$ $(k \leq i \leq n)$. Esta circunstancia terminaría el cálculo del determinante: $\det A = 0$.

El método de ortogonalización de Householder ofrece otra técnica de cálculo de determinantes

$$\det A = \det P_1 P_2 \cdots P_{n-1} R = \sigma \ \det R \ ,$$

donde $\det R$ es el producto de los elementos diagonales de R y σ es el producto de los determinantes de las matrices P_k $(k = 1 \div n-1)$. Como la matriz identidad tiene determinante +1 y las matrices de Householder, -1; σ valdrá +1 o -1 si el número de matrices de Householder usadas es par o impar, respectivamente.

Inversas

Dada una matriz nxn regular A, la matriz $X = A^{-1}$ cumple el sistema de ecuaciones matricial $AX = I$. Separando la columna k-ésima de esta ecuación, tendremos que

$$Ax^{(k)} = e^{(k)} \quad (k = 1 \div n) \ ,$$

donde $e^{(k)}$ indica el k-ésimo vector de la base canónica y $x^{(k)}$, la k-ésima columna de la matriz X.

Así, el cálculo de A^{-1} quedará reducido a la resolución de n sistemas con una matriz común y con términos independientes muy sencillos.

Estos hechos comportan una ventaja de cálculo: si se utiliza el método de Gauss, se podrá realizar la fase de eliminación simultáneamente en los n sistemas, situando la matriz identidad I en el lugar de la columna de términos independientes. Una generalización del procedimiento que se acaba de describir da lugar al *método de Gauss-Jordan*, que parte de los sistemas de ecuaciones $Ax^{(k)} = e^{(k)}$ $(k = 1 \div n)$ expresados en forma matricial

$$(A \,|\, I) = \left(\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1n} & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} & 0 & \cdots & 1 \end{array} \right),$$

y va modificando las filas de $(A \,|\, I)$, usando los elementos de la diagonal como pivotes (como en el método de Gauss) y consiguiendo que se vayan anulando los elementos no diagonales (también los de encima de la diagonal!), columna a columna. Se obtiene, finalmente, $(D \,|\, B)$ con D diagonal, y entonces $A^{-1} = D^{-1}B$.

Alternativamente, el uso de las factorizaciones LU y QR ofrece también la posibilidad de calcular la inversa según:

$$A^{-1} = U^{-1}L^{-1} \ , \quad A^{-1} = R^{-1}Q^{\top} \ .$$

Hay otros procedimientos para el cálculo de la inversa de una matriz A. En el problema 2.9 se describe uno de ellos basado en la construcción de una sucesión de matrices que, bajo ciertas condiciones, converge hacia la matriz inversa A^{-1}. En los problemas propuestos se presentan otros métodos.

2.2.6 Análisis del error

Al resolver un sistema $Ax = b$, A y b aparecen como datos y x como solución. Los errores de x pueden provenir de:

- la propagación de los errores de los datos A y b,
- la acumulación de los errores de redondeo en los cálculos generados en el transcurso del proceso de resolución.

Propagación de los errores de los datos

Supongamos que no disponemos exactamente de la matriz A y del vector b, sino de una matriz $A + \delta A$ y de un vector $b + \delta b$. En estas condiciones, no obtendremos la solución buscada x del sistema $Ax = b$, sino una solución $x + \delta x$ del sistema perturbado

$$(A + \delta A)(x + \delta x) = b + \delta b \ .$$

Disponemos de la fórmula siguiente para acotar, de forma relativa, una norma cualquiera del vector de error δx

$$\frac{\|\delta x\|}{\|x\|} \leq \mu(A) \left[\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \frac{\|x + \delta x\|}{\|x\|} \right] \ ,$$

sobre la cual se tienen que hacer las siguientes observaciones:

- En general y a efectos prácticos, $\|x + \delta x\| \simeq \|x\|$.
- $\mu(A) = \|A\| \, \|A^{-1}\|$ se llama *número de condición* de la matriz A y representa el factor máximo de amplificación de los errores relativos de A y b. Hablaremos de matrices *mal condicionadas* cuando tengan números de condición excesivamente grandes.

  El problema 2.5 presenta una acotación del error relativo de x, bajo unas condiciones particulares, y pone de manifiesto, con un ejemplo concreto, las consecuencias del mal condicionamiento de la matriz sobre la amplificación de los errores de los datos.
- Las normas matriciales usadas en la fórmula han de ser consistentes con las normas vectoriales.
- Si U es una matriz ortogonal (y, por lo tanto, unitaria) tenemos que, usando el número de condición euclídeo $\mu_2(A) = \|A\|_2 \|A^{-1}\|_2$,

  $$\mu_2(UA) = \mu_2(A) \ ;$$

  así, en el método QR, la matriz R del sistema reducido $Rx = Q^\top b$ tiene el mismo número de condición que la matriz A de partida. Se tiene pues la garantía de que el condicionamiento del sistema triangular no empeora, lo cual no puede asegurarse en general para los métodos gaussianos.

En el problema 2.3 se detalla un estudio de propagación de errores en un proceso de factorización de Cholesky.

Acumulación de los errores de redondeo

Al resolver el sistema $Ax = b$, debido a los errores de redondeo durante el proceso, no obtenemos la solución exacta x, sino una solución aproximada $x + \delta x$. Para acotar el error δx (en alguna norma), se puede utilizar el análisis del error hacia atrás, tratando de buscar primeramente δA tal que

$$(A + \delta A)(x + \delta x) = b \ .$$

Se reduce, así, el problema a uno equivalente en el cual se culpa a los elementos de la matriz A de los errores de redondeo; después se deberá aplicar la fórmula de propagación

de los errores de los datos acabada de encontrar para acotar el error en la solución x, ya que no es necesario ahora considerar errores en las operaciones.

La matriz δA depende, evidentemente, del método de resolución que se siga. En concreto, si el método empleado es el de factorización LU, se puede llegar a la acotación siguiente para el error cargado a la matriz A después de llevar a cabo el análisis del error hacia atrás

$$\|\delta A\|_\infty \leq (n^3 + 3n^2)\bar{g}_n \epsilon \|A\|_\infty \ ,$$

donde n es la dimensión del sistema, ϵ es la cota del error relativo en las operaciones aritméticas (véase el apartado 1.1.2) y

$$\bar{g}_n \equiv \frac{\max_{i,j,k} \left|\bar{a}_{ij}^{(k)}\right|}{\max_{i,j} |a_{ij}|}$$

indica el factor máximo de crecimiento del módulo de los elementos calculados $\bar{a}_{ij}^{(k)}$ por el método de eliminación gaussiana respecto al máximo de los módulos de los elementos de la matriz A de partida.

Si definimos g_n de forma análoga, pero considerando los coeficientes $a_{ij}^{(k)}$ sin los errores acumulados, se pueden obtener las acotaciones siguientes:

- $g_n \leq 2^{n-1}$, usando pivotaje maximal por columnas;
- $g_n \leq 3n^{(2+\ln n)/4}$, usando pivotaje completo.

Estas acotaciones son muy pesimistas en la mayor parte de las aplicaciones prácticas; esto es, el error real es mucho menor que la cota obtenida.

No obstante, hay que observar que, si no se usa pivotaje en la resolución de los sistemas lineales no se tiene generalmente ningún control sobre la magnitud del factor $\bar{g}_n$; éste puede llegar a ser muy grande si, durante el proceso, nos encontramos con pivotes cercanos a cero. Estos harán crecer mucho los elementos calculados $\bar{a}_{ij}^{(k)}$ y, consiguientemente, el error en la solución.

2.2.7 Métodos iterativos

Generalidades

Los métodos iterativos están basados en la siguiente idea: dado un sistema lineal $Ax = b$, le asociamos de alguna manera determinada un sistema lineal equivalente $x = Bx + c$ que resolvemos de forma iterativa. Esto es, comenzamos con $x^{(0)}$ arbitrario y obtenemos después $x^{(1)}$, $x^{(2)}$, ... mediante la recurrencia

$$x^{(k+1)} = Bx^{(k)} + c \ \ (k \geq 0) \ ;$$

si la sucesión de vectores $x^{(k)}$ $(k \geq 0)$ es convergente a un vector x, éste será la solución del sistema $x = Bx + c$ y, por lo tanto, del sistema de partida $Ax = b$. La matriz B se acostumbra a llamar *matriz de iteración.*

Analizando el comportamiento de los errores $e^{(k)} \equiv x^{(k)} - x$ cuando k tiende a infinito, y teniendo en cuenta las propiedades de las normas vectoriales y matriciales dadas en el apartado 2.1.3, se llega a los resultados siguientes sobre la convergencia de los métodos iterativos:

- El método iterativo $x^{(k+1)} = Bx^{(k)} + c$ $(k \geq 0)$ es convergente, independientemente de la elección de $x^{(0)}$, si y sólo si el radio espectral de la matriz de iteración $\rho(B)$ es menor que 1. El teorema de Gershgorin, presentado en el apartado 2.1.3, es una buena herramienta para la acotación de $\rho(B)$ (véase el problema 2.11).

- El método iterativo anterior es pues convergente si y sólo si $\|B\| < 1$ para alguna norma matricial consistente con alguna norma vectorial (por ejemplo, las normas matriciales subordinadas $\| \ \|_1$ y $\| \ \|_\infty$).

Describiremos a continuación los métodos iterativos más conocidos.

Método iterativo de Jacobi

Suponiendo que los elementos diagonales de A sean no nulos, si dividimos cada fila por el elemento de la diagonal correspondiente, se obtiene una matriz con unos en la diagonal. Podemos imaginar esta matriz formada por la suma de una matriz triangular inferior L con ceros en la diagonal, de la matriz identidad I y de una matriz triangular superior U con ceros en la diagonal; así queda descompuesta la matriz A como producto de su diagonal D por la suma de L, I y U,

$$A = D(L + I + U) \ .$$

El sistema $Ax = b$ es entonces equivalente al sistema

$$x = -(L + U)x + D^{-1}b \ .$$

Nótese que este sistema equivalente se obtiene simplemente despejando en la ecuación i-ésima la variable x_i $(i = 1 \div n)$.

Este sistema permite definir el *método iterativo de Jacobi*

$$x^{(k+1)} = -(L + U)x^{(k)} + D^{-1}b \ ,$$

que, en componentes, toma la forma

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right) \ (i = 1 \div n) \ .$$

Método iterativo de Gauss-Seidel

Al igual que el método iterativo de Jacobi, sólo será aplicable a sistemas $Ax = b$ tales que los elementos diagonales de A sean no nulos. Está basado en la equivalencia entre el sistema $Ax = b$ y el sistema

$$x = -(I + L)^{-1}Ux + (I + L)^{-1}D^{-1}b \ .$$

Así, se define el *método iterativo de Gauss-Seidel* mediante

$$x^{(k+1)} = -(I + L)^{-1}Ux^{(k)} + (I + L)^{-1}D^{-1}b \ ,$$

que equivale a

$$x^{(k+1)} = -Lx^{(k+1)} - Ux^{(k)} + D^{-1}b \ ,$$

y que, en componentes, toma la forma

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) \ (i = 1 \div n) \ .$$

Los dos métodos que se acaban de describir son muy semejantes, dado que la fórmula de evaluación de la componente i-ésima del nuevo vector es prácticamente la misma: la diferencia radica en el hecho de que, en el método iterativo de Jacobi, dicha componente se calcula a partir de las componentes del vector anterior $x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_{i+1}^{(k)}, \ldots, x_n^{(k)}$, y en el método iterativo de Gauss-Seidel se utilizan las componentes ya calculadas del nuevo vector: $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_{i+1}^{(k)}, \ldots, x_n^{(k)}$.

2.2.8 Métodos iterativos de sobrerrelajación

Se trata de una familia de métodos que generalizan el método iterativo de Gauss-Seidel. Están basados en la equivalencia de $Ax = b$ con

$$x = x - \omega(Lx + (I + U)x - D^{-1}b) \ ,$$

donde ω es un parámetro real, llamado *factor de relajación*. Este sistema también es equivalente al sistema

$$x = B_\omega x + c_\omega \ ,$$

con

$$B_\omega = (I + \omega L)^{-1}[(1 - \omega)I - \omega U] \ , \quad c_\omega = \omega(I + \omega L)^{-1} D^{-1} b \ .$$

Se define el *método iterativo de sobrerrelajación con factor* ω mediante la recurrencia

$$x^{(k+1)} = B_\omega x^{(k)} + c_\omega \ ,$$

que también puede escribirse como

$$x^{(k+1)} = x^{(k)} - \omega(Lx^{(k+1)} + (I + U)x^{(k)} - D^{-1}b) \ .$$

En componentes, toma la forma

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right) \ (i = 1 \div n) \ .$$

Hay que observar que, para $\omega = 1$, se tiene el método iterativo de Gauss-Seidel.

Los métodos de sobrerrelajación que se acaban de describir se basan en una transformación del sistema $Ax = b$ en otro equivalente del tipo $x = B(\omega)x + c(\omega)$, sobre el cual se define la recurrencia. La construcción del sistema $x = B(\omega)x + c(\omega)$ se puede efectuar de varias maneras que dan lugar a nuevos métodos iterativos. En el problema 2.8 se define uno de ellos y se estudia su convergencia.

Convergencia de los métodos iterativos de Jacobi, de Gauss-Seidel y de sobrerrelajación

De acuerdo con lo que se ha expuesto, los métodos iterativos de Jacobi, Gauss-Seidel y sobrerrelajación convergerán si y sólo si los radios espectrales de sus matrices de iteración ($B_J = -(L+U)$, $B_{GS} = -(I+L)^{-1}U$ y B_ω) son menores que 1. En el problema 2.6 se estudia la convergencia de los métodos de Jacobi y de Gauss-Seidel al aplicarlos a un sistema lineal de dos ecuaciones y dos incógnitas.

A continuación se dan algunos resultados relacionados con la convergencia de estos métodos en casos particulares y que no requieren el cálculo del radio espectral:

- Si A es estrictamente diagonal dominante, los métodos iterativos de Jacobi y de Gauss-Seidel convergen.

  La rapidez con que se produce esta convergencia depende de las características del sistema. En el problema 2.7 se presenta una estimación del número de iteraciones del método iterativo de Jacobi que son necesarias para poder garantizar una acotación de error dada al resolver un sistema donde, entre otras condiciones, la matriz se supone estrictamente diagonal dominante.

- Los métodos iterativos de sobrerrelajación divergen para factores ω que no pertenecen al intervalo $(0,2)$.

- Si A es hermítica con elementos positivos en la diagonal, los métodos iterativos de sobrerrelajación con factores $\omega \in (0,2)$ convergen si y sólo si la matriz A es definida positiva.

El método iterativo de sobrerrelajación tiene la ventaja de que permite modificar el factor ω de forma que pueda conseguirse una convergencia más rápida para ciertos tipos de matrices; por ejemplo, se dispone del resultado siguiente en este sentido:

- Para una matriz A, definida positiva, tridiagonal por bloques y tal que los bloques de la diagonal sean matrices diagonales, tenemos que

  $$\rho(B_{GS}) = \rho(B_J)^2 \ .$$

  Si, además, los valores propios de B_J están en $(-1,1)$; entonces, para

  $$\omega = \tilde{\omega} = \frac{2}{1+\sqrt{1-\rho(B_J)^2}} \ ,$$

  tenemos que

  $$\rho(B_{\tilde{\omega}}) = \tilde{\omega} - 1 = \left(\frac{\rho(B_J)}{1+\sqrt{1-\rho(B_J)^2}} \right)^2 \leq \rho(B_\omega) \quad \forall \omega \ .$$

  Así pues, $\tilde{\omega}$ es el valor de ω que minimiza el radio espectral de las matrices B_ω; por lo tanto, el método iterativo de sobrerrelajación correspondiente será el más rápidamente convergente. Por esta razón, $\tilde{\omega}$ se llama *factor óptimo de relajación*.

2.3 VALORES Y VECTORES PROPIOS

2.3.1 Introducción

El problema de búsqueda de valores y vectores propios de una matriz es, de hecho, un problema no lineal, pero tiene una fuerte componente lineal. Así, el cálculo de valores propios se reduce a buscar los ceros de un polinomio y puede ser llevado a cabo mediante los métodos generales de búsqueda de ceros de polinomios (que serán presentados en el capítulo 5), o bien por los métodos iterativos lineales explicados en este capítulo. Una vez conocidos los valores propios, el cálculo de los vectores propios sí que es un problema lineal.

No obstante, en general, es excesivamente laborioso el cálculo directo del polinomio característico de una matriz. Por esta razón el cálculo efectivo de valores y vectores propios no pasa por la obtención del citado polinomio, excepto cuando la especial naturaleza de la matriz lo permita: matrices de dimensiones pequeñas, tridiagonales, de Frobenius (véanse los problemas 2.10 y 2.11), etc.

El concepto de deflación de matrices es similar al de deflación de polinomios (véase el apartado 5.2.2) y permite simplificar el problema de encontrar valores y vectores propios, una vez conocidos algunos de ellos. Se describen los métodos de Hotteling, Wieland y de Householder.

Los métodos de cálculo de valores y vectores propios son de diversa naturaleza: unos, como los métodos iterativos de la potencia, de Jacobi, LR y QR, calculan los valores propios mediante técnicas iterativas; otros, como los de Givens y Householder, reducen la matriz inicial a una matriz semejante Hessenberg superior (tridiagonal simétrica, si la matriz es simétrica), en un número finito de pasos. Para estos tipos de matrices reducidas, se dan métodos específicos de cálculo de los valores y vectores propios que sí pasan por la obtención del polinomio característico. En el problema 2.10 se describe un método de este tipo (el de Danilevski), que reduce la matriz inicial a la forma normal de Frobenius, cuyo polinomio característico se obtiene de inmediato.

2.3.2 Deflación de matrices

Partiendo del conocimiento de un valor propio λ y de un vector propio asociado v de una matriz nxn A, el proceso de *deflación* consiste en obtener una nueva matriz $\tilde{A}$ (de características más simples, a menudo de dimensión menor) de manera que, encontrando los valores y vectores propios de $\tilde{A}$, se puedan obtener los de A, con facilidad.

Conviene emplear la deflación, alternándola con métodos que permitan calcular los valores propios de uno en uno, y así poder obtener todos los valores propios trabajando cada vez con matrices más sencillas. Además, análogamente a lo que ocurre con la deflación polinomial, que se estudiará en el capítulo 5, es mejor que dichos valores propios se vayan calculando de menor a mayor, con el fin de ganar estabilidad numérica; conviene también que los valores y vectores propios encontrados al final del proceso sean corregidos teniendo en cuenta la matriz de partida, para eliminar los errores acumulados en todo el proceso de deflación.

Existen diversos tipos de deflación de matrices; mencionaremos aquí los de Hotteling, Wielandt y Householder. Hay que destacar las ventajas de este último, que usa transformaciones de semejanza con matrices ortogonales.

Deflación de Hotteling

Consideremos conocidos un valor propio λ de A y vectores propios asociados v y u, por la derecha y por la izquierda respectivamente, satisfaciendo $u^\top v \neq 0$; se construye la matriz

$$A_{(H)} = A - \frac{\lambda}{u^\top v} v u^\top \ ,$$

que tiene los mismos valores propios que A, salvo el valor propio λ que ha estado sustituido por el valor propio cero; los vectores propios son los mismos.

Si A es simétrica, puede escogerse $u = v$.

Deflación de Wielandt

Conocido un vector propio v de valor propio λ de A, con $v_1 \neq 0$ (por ejemplo) y denotando por $a^{(1)\top}$ la primera fila de A, la matriz

$$A_{(V)} = A - \frac{1}{v_1} v a^{(1)\top} \ ,$$

tiene la primera fila nula. Solamente hay que buscar ahora valores y vectores propios de la matriz $(n-1)$x$(n-1)$, $\tilde{A}_{(V)}$, formada por las últimas $n-1$ filas de las últimas $n-1$ columnas de A. Esto se debe al hecho de que, si $\lambda' \neq \lambda$ es un valor propio de $\tilde{A}_{(V)}$ con vector propio (de $n-1$ componentes) asociado $\tilde{v}'$ e introducimos

$$w = \begin{pmatrix} 0 \\ \tilde{v}' \end{pmatrix} \ , \quad c = \frac{(\lambda' - \lambda) v_1}{a^{(1)\top} w} \ ,$$

entonces $v' = v + cw$ es un vector propio de A de valor propio λ'.

Deflación de Householder

Es el tipo de deflación al cual prestamos más atención, ya que se lleva a cabo mediante matrices ortogonales; así se consigue un control de la estabilidad numérica y la conservación de la simetría de las matrices, dado el caso.

Conocido un vector propio v de valor propio λ de A, si v es de la forma

$$v = \begin{pmatrix} v_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \ ,$$

podemos tomar $v = e^{(1)}$ y entonces A es de hecho de la forma

$$A = \left(\begin{array}{c|c} \lambda & c^\top \\ \hline 0 & \tilde{A} \\ \vdots & \end{array} \right) \ .$$

Si $\tilde{v}'$ es un vector propio de $\tilde{A}$ de valor propio $\lambda' \neq \lambda$, entonces

$$v' = \begin{pmatrix} b \\ \\ \tilde{v}' \\ \\ \end{pmatrix} \quad \text{con} \quad b = \frac{c^\top \tilde{v}'}{\lambda' - \lambda}$$

será un vector propio de valor propio λ' de A.

Este resultado nos permite calcular los vectores y valores propios de A, a partir de los de $\tilde{A}$, de dimensión $n-1$.

Si A es simétrica, $c = 0$ y $\tilde{A}$ es también simétrica y, si $\tilde{v}'$ es un vector propio de $\tilde{A}$ de valor propio λ', entonces

$$v' = \begin{pmatrix} 0 \\ \\ \tilde{v}' \\ \\ \end{pmatrix}$$

será un vector propio de valor propio λ' de A.

Volviendo al caso general, si v es un vector propio de valor propio λ de A, con $v_2^2 + \cdots + v_n^2 \neq 0$, podemos encontrar matrices de Householder P tales que

$$Pv = se^{(1)} \ , \quad s = \pm \|v\|_2$$

(véase el apartado 2.2.4 para su cálculo efectivo).

Como $P = P^\top = P^{-1}$,

$$PAPe^{(1)} = \frac{1}{s} PAv = \frac{\lambda}{s} Pv = \lambda e^{(1)} \ ;$$

por lo tanto, la matriz PAP es de la forma deseada:

$$A' = PAP = \left(\begin{array}{c|c} \lambda & c'^\top \\ \hline \\ 0 & \tilde{A}' \\ \\ \end{array}\right) .$$

El vector v' es un vector propio de valor propio λ' de A' si y sólo si Pv' es un vector propio de valor propio λ' de A. Si A es simétrica, también A' lo será ($c' = 0$) (véase el problema 2.12).

2.3.3 Métodos de la potencia

Método de la potencia

El método de la potencia y sus variantes son muy útiles cuando sólo se requieren algunos valores propios y sus vectores propios asociados. También sirven para calcular todos los valores propios de la matriz si se combinan con alguno de los tipos de deflación descritos anteriormente.

El método de la potencia propiamente dicho permite calcular aproximaciones sucesivas del valor propio de módulo máximo (si existe), así como vectores propios asociados a éste.

Dada A, matriz nxn, sean λ_j $(j = 1 \div n)$ sus valores propios, repetidos según su multiplicidad y ordenados según su módulo

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n| \ .$$

Supondremos, por simplicidad, que A es diagonalizable y llamaremos $v^{(j)}$ $(j = 1 \div n)$ a los vectores propios linealmente independientes asociados a los valores propios λ_j $(j = 1 \div n)$; supondremos también que existe un único valor propio de módulo máximo λ_1: $|\lambda_1| > |\lambda_2|$. Otros casos son tratados en los problemas propuestos.

El *método de la potencia* considera las iteraciones

$$x^{(k+1)} = Ax^{(k)} \ (k \geq 0)$$

para algún vector

$$x^{(0)} = \sum_{j=1}^{n} \alpha_j v^{(j)} \ \text{ con } \ \alpha_1 \neq 0 \ ;$$

así tenemos

$$x^{(k)} = \lambda_1^k \left[\alpha_1 v^{(1)} + \alpha_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k v^{(2)} + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1} \right)^k v^{(n)} \right] \ .$$

Si tienen sentido las sucesiones de cocientes

$$q_i^{(k)} = \frac{x_i^{(k+1)}}{x_i^{(k)}} \ (i = 1 \div n) \ ;$$

es decir, si $x_i^{(k)} \neq 0$, éstas tienen por límite λ_1; además, cuando $k \to \infty$:

$$q_i^{(k)} = \lambda_1 \left[1 + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right] \ (i = 1 \div n) \ , \tag{2.11}$$

$$\frac{x^{(k)}}{\lambda_1^k} = \alpha_1 v^{(1)} \left[1 + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right] \ . \tag{2.12}$$

Estas son expresiones asintóticas para λ_1 y para algún vector propio asociado a él, tratadas en el apartado 4.3.1; la velocidad de convergencia depende esencialmente de $\left| \frac{\lambda_2}{\lambda_1} \right|$ y puede ser muy pequeña, cuando $|\lambda_2|$ sea próximo a $|\lambda_1|$.

Como la sucesión $(\lambda_1^k)_{k \geq 0}$ generalmente tiende a cero o no está acotada, conviene normalizar de alguna manera la sucesión de los vectores $x^{(k)}$. Un posible procedimiento es el siguiente: una vez tenemos $x^{(k)}$, lo normalizamos según

$$y^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|}$$

(usando una norma cualquiera) y calculamos $x^{(k+1)} = Ay^{(k)}$ $(k \geq 0)$; entonces,

$$\lim_{k \to \infty} \frac{x_i^{(k+1)}}{y_i^{(k)}} = \lambda_1 \ (i = 1 \div n) \ , \quad \lim_{k \to \infty} y^{(k)} = \pm \frac{v^{(1)}}{\|v^{(1)}\|} \ .$$

Con el fin de asegurar la convergencia de las sucesiones a λ_1, se ha tenido que suponer que la componente α_1 de $x^{(0)}$, según $v^{(1)}$, no era nula. En caso contrario y suponiendo, por ejemplo, que $\alpha_2 \neq 0$ y $|\lambda_2| > |\lambda_3|$, se obtendría analíticamente que

$$\lim_{k\to\infty} \frac{x_i^{(k+1)}}{x_i^{(k)}} = \lambda_2 \ .$$

Ahora bien, en la práctica, los errores de redondeo que se cometen durante el cálculo introducen en los $x^{(k)}$ una componente no nula según $v^{(1)}$, que implica la tendencia de la sucesión a λ_1, y no a λ_2.

Para matrices A simétricas, es más eficaz calcular la sucesión de *cocientes de Rayleigh* de $x^{(k)}$

$$\sigma_k = \frac{x^{(k)\top} A x^{(k)}}{x^{(k)\top} x^{(k)}} \ (k \geq 0)$$

usando la recurrencia

$$y^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|_2} \ , \quad x^{(k+1)} = A y^{(k)} \ , \quad \sigma_k = y^{(k)\top} x^{(k+1)} \ (k \geq 0) \ .$$

Entonces obtenemos la convergencia de las sucesiones anteriores:

$$\lim_{k\to\infty} \sigma_k = \lambda_1 \ , \quad \lim_{k\to\infty} y^{(k)} = \pm \frac{v^{(1)}}{\|v^{(1)}\|_2} \ ,$$

según las expresiones asintóticas, cuando $k \to \infty$:

$$\begin{aligned} \sigma_k &= \lambda_1 \left[1 + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^{2k} \right) \right] \ , \\ y^{(k)} &= \pm \frac{v^{(1)}}{\|v^{(1)}\|_2} \left[1 + \mathcal{O}\left(\left| \frac{\lambda_2}{\lambda_1} \right|^{k} \right) \right] \ . \end{aligned} \tag{2.13}$$

La utilización de los cocientes de Rayleigh comporta un aumento de la velocidad de convergencia en el cálculo del valor propio, al ser

$$\left| \frac{\lambda_2}{\lambda_1} \right|^2 < \left| \frac{\lambda_2}{\lambda_1} \right| \ .$$

A continuación se describen algunas posibles variantes del método de la potencia.

Método de la potencia desplazada

Aplicando el método de la potencia a la matriz $A - dI$, que tiene como valores propios $\mu_j = \lambda_j - d$ con vectores propios asociados $v^{(j)}$ $(j = 1 \div n)$, se obtiene el valor propio μ_j de módulo máximo de $A - dI$ (siempre que sea único) y un vector propio asociado $v^{(j)}$: permite, así, calcular el valor propio $\lambda_j = d + \mu_j$ de A más alejado de d.

Método de la potencia inversa

Suponiendo ahora que A es inversible, A^{-1} tiene los mismos vectores propios que A con valores propios μ_j que son los recíprocos de los de A: $\mu_j = \frac{1}{\lambda_j}$ $(j = 1 \div n)$. Si el valor propio de módulo mínimo es único ($|\lambda_n| < |\lambda_{n-1}|$), la aplicación del método de la potencia a A^{-1} proporciona el valor propio $\frac{1}{\lambda_n}$ de A^{-1} y un vector propio asociado: podemos conocer así el valor propio de módulo mínimo λ_n de A.

Método de la potencia inversa desplazada

Es una combinación de las dos variantes anteriores. Consiste en aplicar el método de la potencia a $(A-dI)^{-1}$, que tiene por valores propios los $\mu_j = \frac{1}{\lambda_j - d}$ $(j = 1 \div n)$. La aplicación del método da el valor propio μ_j de módulo máximo de esta matriz, siempre que sea único, y un vector propio asociado $v^{(j)}$: así podemos encontrar el valor propio $\lambda_j = d + \frac{1}{\mu_j}$ de A más próximo a d.

Esta variante es muy útil para refinar aproximaciones d de valores propios, y para obtener vectores propios asociados.

En el problema 2.13 se presenta un ejemplo de uso combinado del método de la potencia con el método de la potencia inversa para el cálculo de todos los valores propios de una matriz 4x4.

2.3.4 Métodos de Jacobi

Método de Jacobi

Sea A una matriz nxn simétrica. Sabemos, por la propiedad 10 de valores y vectores propios del apartado 2.1.3, que existe una matriz ortogonal O ($O^{-1} = O^\top$) tal que $\Lambda = O^\top A O$ es diagonal: $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$; la columna j-ésima de O es un vector propio de A asociado al valor propio λ_j y el conjunto de dichas columnas forma una base de vectores propios ortonormales de A.

El *método de Jacobi* proporciona precisamente un procedimiento de búsqueda de sucesivas aproximaciones A_k y O_k de Λ y O, respectivamente, donde O_k es un producto de matrices ortogonales sencillas ($k \geq 1$). Se trata, por lo tanto, de un método iterativo, cuyo mecanismo de construcción se detalla a continuación.

Iniciamos el proceso con $A_1 = A$ y $O_1 = I$ y, conocidos $A_k \equiv B = (b_{ij})$ y $O_k \equiv \Omega = (\omega_{ij})$, escogemos en B un elemento no diagonal b_{pq} (con $p < q$) tal que $b_{pq} \neq 0$.

Los elementos b_{pp}, b_{pq}, $b_{qp} (= b_{pq})$ y b_{qq} forman una matriz 2x2 simétrica. Buscamos una matriz de rotación

$$\begin{pmatrix} c & -s \\ s & c \end{pmatrix}, \quad c = \cos\varphi\ , \quad s = \mathrm{sen}\,\varphi\ ,$$

de forma que la matriz simétrica

$$\begin{pmatrix} d_{pp} & d_{pq} \\ d_{qp} & d_{qq} \end{pmatrix} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{pmatrix} \begin{pmatrix} c & -s \\ s & c \end{pmatrix}$$

sea diagonal.

Para ello, el ángulo φ ha de cumplir

$$\cot 2\varphi = \frac{b_{pp} - b_{qq}}{2b_{pq}}$$

y puede escogerse cumpliendo además $|\varphi| < \frac{\pi}{4}$.

Las fórmulas que se obtienen así para c y s son:

$$\begin{aligned}
&\text{Si } b_{pp} = b_{qq}: \quad \epsilon = \operatorname{sgn}(b_{pq})\ ,\quad c = \frac{\sqrt{2}}{2}\ ,\quad s = \epsilon\frac{\sqrt{2}}{2}\ . \\
&\text{Si } b_{pp} \neq b_{qq}: \quad \epsilon = \operatorname{sgn}[b_{pq}(b_{pp} - b_{qq})]\ , \qquad (2.14)\\
&\qquad\qquad c = \sqrt{\frac{1}{2}(1 + \frac{|b_{pp} - b_{qq}|}{d})}\ , \\
&\qquad\qquad s = \epsilon\sqrt{\frac{1}{2}(1 - \frac{|b_{pp} - b_{qq}|}{d})}\ , \\
&\qquad\qquad d = \sqrt{(b_{pp} - b_{qq})^2 + 4b_{pq}^2}\ . \qquad (2.15)
\end{aligned}$$

Tomamos ahora la matriz

$$R_k \equiv R = \begin{pmatrix}
1 & & & & & & & & & & \\
& \ddots & & & & & & & & & \\
& & 1 & & & & & & & & \\
& & & c & & & & -s & & & \\
& & & & 1 & & & & & & \\
& & & & & \ddots & & & & & \\
& & & & & & 1 & & & & \\
& & & s & & & & c & & & \\
& & & & & & & & 1 & & \\
& & & & & & & & & \ddots & \\
& & & & & & & & & & 1
\end{pmatrix}\ ,$$

donde los elementos no escritos son nulos; esta matriz representa una rotación de ángulo φ en el plano coordenado $p - q$.

Calculamos ahora $\overline{B} = R^\top BR = (\overline{b}_{ij})$ (cambiando sólo las filas y columnas p-ésima y q-ésima de B) y $\overline{\Omega} = \Omega R = (\overline{\omega}_{ij})$ (cambiando sólo las columnas p-ésima y q-ésima de Ω) según las fórmulas:

$$\begin{aligned}
\overline{b}_{pj} &= \overline{b}_{jp} = b_{pj}c + b_{qj}s\ (j = 1 \div n,\ j \neq p, q)\ , \\
\overline{b}_{iq} &= \overline{b}_{qi} = -b_{ip}s + b_{iq}c\ (i = 1 \div n,\ i \neq p, q)\ , \\
\overline{b}_{pp} &= b_{pp}c^2 + 2b_{pq}cs + b_{qq}s^2 = b_{pp} + b_{pq}t\ , \qquad (2.16)\\
\overline{b}_{qq} &= b_{pp}s^2 - 2b_{pq}cs + b_{qq}c^2 = -b_{pq}t + b_{qq}\ , \\
\overline{b}_{pq} &= \overline{b}_{qp} = 0\ ,
\end{aligned}$$

$$\begin{aligned}
\overline{\omega}_{ip} &= \omega_{ip}c + \omega_{iq}s\ , \qquad (2.17)\\
\overline{\omega}_{iq} &= -\omega_{ip}s + \omega_{iq}c\ (i = 1 \div n)\ ;
\end{aligned}$$

donde se ha usado $t = \frac{s}{c}$; las otras componentes no cambian.

De manera análoga, en la k-ésima iteración del proceso iterativo se obtiene:

$$A_{k+1} = R_k^\top A_k R_k = \cdots = R_k^\top \cdots R_1^\top A R_1 \cdots R_k \ ,$$

$$O_{k+1} = O_k R_k = \cdots = R_1 \cdots R_k \ \ (k \geq 0) \ .$$

Las matrices A_k son, así, semejantes a la matriz A, mediante las transformaciones ortogonales O_k:

$$A_k = O_k^\top A O_k \ (k \geq 0) \ .$$

Variantes del método de Jacobi

La elección de los elementos no diagonales b_{pq} que queremos anular puede ser realizada de diversas maneras que dan lugar a las diferentes variantes del método de Jacobi.

Así, tomando

$$|b_{pq}| = \max_{i<j} |b_{ij}|$$

obtenemos el *método clásico de Jacobi* (véase el problema 2.14).

Los elementos no diagonales $b_{pq} \neq 0$ pueden ser escogidos también de otras maneras: por ejemplo, tomando sus índices según el orden: $(1,2)$, $(1,3)$, ..., $(1,n)$, $(2,3)$, ..., $(2,n)$, ..., $(n-1,n)$. Repitiendo este proceso las veces que haga falta, se obtiene el llamado *método cíclico de Jacobi*. Una variante de éste consiste en efectuar únicamente las iteraciones anuladoras que corresponden a elementos mayores, en valor absoluto, que ciertas cantidades prefijadas, llamadas *umbrales*, y que pueden cambiar en cada ciclo; el método se llama entonces *método cíclico de Jacobi con umbrales*.

Nótese que los elementos no diagonales anulados en una iteración pueden volver a ser diferentes de cero después de alguna de las iteraciones posteriores; a pesar de esto, se puede demostrar que, si definimos

$$\epsilon^2(B) \equiv \sum_{i \neq j} b_{ij}^2 \ ,$$

resulta que, para el método clásico de Jacobi,

$$\epsilon^2(\overline{B}) = \epsilon^2(B) + \sum_{j=1}^{n} b_{jj}^2 \leq - \sum_{j=1}^{n} \bar{b}_{jj}^2 \leq \epsilon^2(B)(1 - \frac{2}{n(n-1)}) \ ;$$

ello implica que los elementos no diagonales de A_k tienden a 0 cuando k tiende a infinito y que, bajo condiciones muy generales, las sucesiones $(A_k)_{k\geq 0}$ y $(O_k)_{k\geq 0}$ tienen límites respectivos Λ y O tales que $\Lambda = O^\top A O$, con Λ diagonal y O ortogonal.

2.3.5 Métodos de reducción

El objetivo de estos métodos consiste en reducir una matriz A real a una matriz Hessenberg superior, mediante transformaciones de semejanza ortogonales. En el caso de que la matriz de partida sea simétrica, quedará reducida a forma tridiagonal simétrica.

La matriz reducida tendrá los mismos valores propios que la matriz inicial, y los vectores propios de la matriz inicial se encontrarán a partir de los de la matriz reducida

por aplicación de la transformación ortogonal hallada. En el apartado siguiente se darán métodos de cálculo de valores y vectores propios de las matrices reducidas.

Trataremos primero el caso simétrico, entendiendo el caso no simétrico como una extensión de éste, donde los procedimientos de transformación conducirán a matrices Hessenberg superior, ya que en este caso no se produce la anulación de los elementos simétricos durante el proceso.

Una vez obtenidos los valores y vectores propios de la matriz tridiagonal simétrica (o Hessenberg superior), los valores propios de la matriz inicial coincidirán con los de aquella; los vectores propios de la matriz reducida deberán ser transformados más adelante a vectores propios de la matriz de partida a través de las transformaciones de semejanza empleadas.

Método de Givens

Como el método de Jacobi, el *método de Givens* usa transformaciones de semejanza sucesivas que son rotaciones en planos coordenados; ahora, en cambio, se busca que, en cada transformación, se anule un elemento no tridiagonal, y también su simétrico respecto a la diagonal, de forma que se mantengan nulos todos los elementos anulados en transformaciones anteriores. Se trata, por lo tanto, de un método directo que consigue anular los $(n-1)(n-2)$ elementos no tridiagonales de una matriz simétrica en un máximo de $m := \frac{1}{2}(n-1)(n-2)$ transformaciones.

En cada paso se busca una rotación R_k en el plano coordenado $p-q$ que anule el elemento $(p-1,q)$ y su simétrico. Este proceso se realiza en el siguiente orden de los planos coordenados $p-q$: $2-3$, ..., $2-n$, $3-4$, ..., $3-n$, ..., $(n-1)-n$ de manera que se vayan anulando los elementos $(1,3)$, ..., $(1,n)$, $(2,4)$, ..., $(2,n)$, ..., $(n-2,n)$, y sus simétricos.

Al cabo de estas m transformaciones de semejanza tendremos la matriz

$$A_{m+1} = R_m^\top \cdots R_1^\top A R_1 \cdots R_m$$

que será ya tridiagonal simétrica.

Usando una notación análoga a la utilizada en el método de Jacobi, el ángulo φ de la rotación en el plano $p-q$ que cumple $\overline{b}_{p-1,q} = 0$ satisface

$$\tan\varphi = \frac{b_{p-1,q}}{b_{p-1,p}} \ ;$$

así, hay que escoger los siguientes valores de c y s:

$$c = \frac{b_{p-1,p}}{\sqrt{b_{p-1,p}^2 + b_{p-1,q}^2}} \ , \quad s = \frac{b_{p-1,q}}{\sqrt{b_{p-1,p}^2 + b_{p-1,q}^2}} \ .$$

Se puede comprobar que los ceros conseguidos son conservados en las sucesivas transformaciones. En este sentido, el orden indicado es importante.

Método de Householder

Las transformaciones de semejanza ortogonales se realizan ahora mediante matrices de Householder (véase el apartado 2.2.4). Se quiere conseguir, en cada transformación, que

los elementos no tridiagonales de la columna correspondiente (y los de la fila simétrica) sean nulos.

Partiendo de $A_2 = A$, hay que considerar en el paso k-ésimo una transformación ortogonal $A_{k+1} = P_k A_k P_k$ de la matriz simétrica

$$A_k = \left(\begin{array}{ccc|c} & & & 0 \\ & T_k & & \\ & & & \tilde{a}^{(k)\top} \\ \hline 0 & & \tilde{a}^{(k)} & \tilde{A}_k \end{array}\right) ,$$

donde

$$T_k = \begin{pmatrix} r_1 & s_2 & & & \\ s_2 & r_2 & s_3 & & \\ & \ddots & \ddots & \ddots & \\ & & s_{k-2} & r_{k-2} & s_{k-1} \\ & & & s_{k-1} & r_{k-1} \end{pmatrix}$$

ya es tridiagonal simétrica y

$$\tilde{a}^{(k)} = \begin{pmatrix} a^{(k)}_{k,k-1} \\ a^{(k)}_{k+1,k-1} \\ \vdots \\ a^{(k)}_{n,k-1} \end{pmatrix} \in \mathbb{R}^{n-k+1} .$$

Si $(a^{(k)}_{k+1,k-1})^2 + \cdots + (a^{(k)}_{n,k-1})^2 = 0$, la matriz $(A_k)_k$ (formada por las k primeras filas de las k primeras columnas de A_k) es ya de hecho una matriz tridiagonal simétrica, con $s_k = a^{(k)}_{k,k-1}$, $r_k = a^{(k)}_{k,k}$ y sólo hay que escoger $P_k = I$, con lo que $A_{k+1} = P_k A_k P_k = A_k$.

En caso contrario, eligiendo s_k tal que

$$s_k^2 = \|\tilde{a}^{(k)}\|_2^2 = \sum_{i=k}^{n} (a^{(k)}_{i,k-1})^2 ,$$

podemos encontrar una matriz $(n-k+1)$x$(n-k+1)$ de Householder $\tilde{P}_k$ tal que

$$\tilde{P}_k \tilde{a}^{(k)} = \begin{pmatrix} s_k \\ 0 \\ \vdots \\ 0 \end{pmatrix} = s_k e^{(1)} \in \mathbb{R}^{n-k+1} .$$

En concreto

$$\tilde{P}_k = I_{n-k+1} - \alpha_k \tilde{u}^{(k)} \tilde{u}^{(k)\top} ,$$

con

$$\tilde{u}^{(k)} = \tilde{a}^{(k)} - s_k e^{(1)} \in \mathbb{R}^{n-k+1} , \quad \alpha_k = \frac{1}{s_k (s_k - a^{(k)}_{k,k-1})}$$

(véase el apartado 2.2.4).

Definiendo entonces

$$P_k = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \tilde{P}_k \end{array}\right) ,$$

o equivalentemente $P_k = I_k - \alpha_k u^{(k)} u^{(k)\top}$, con

$$u^{(k)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{u}^{(k)} \end{pmatrix} ,$$

resulta que

$$A_{k+1} = P_k A_k P_k = \left(\begin{array}{ccc|ccc} & T_k & & & 0 & \\ & & & s_k & 0 \cdots & 0 \\ \hline & & s_k & & & \\ & & 0 & & & \\ 0 & & \vdots & & \tilde{P}_k \tilde{A}_k \tilde{P}_k & \\ & & 0 & & & \end{array}\right)$$

$$= \left(\begin{array}{ccc|c} & T_{k+1} & & 0 \\ & & & \tilde{a}^{(k+1)\top} \\ \hline 0 & & \tilde{a}^{k+1} & \tilde{A}_{k+1} \end{array}\right) .$$

El cálculo efectivo de

$$\tilde{P}_k \tilde{A}_k \tilde{P}_k = \left(\begin{array}{c|c} r_k & \tilde{a}^{(k+1)\top} \\ \hline \tilde{a}^{(k+1)} & \tilde{A}_{k+1} \end{array}\right)$$

se realiza según el procedimiento descrito en el apartado 2.2.4 y requiere por lo tanto $2(n-k+1)^2 + \mathcal{O}(n-k+1)$ operaciones y una raíz cuadrada.

Después de $n-2$ transformaciones de semejanza se obtiene la matriz tridiagonal simétrica

$$A_{n-1} = P_{n-1} \cdots P_2 A P_2 \cdots P_{n-1} ,$$

habiéndose realizado un total de

$$\sum_{k=2}^{n-1}[2(n-k+1)^2+\mathcal{O}(n-k+1)]=\frac{2}{3}n^3+\mathcal{O}(n^2)$$

operaciones y $n-2$ raíces cuadradas.

En el problema 2.15 se presenta un ejemplo de utilización de este método.

2.3.6 Valores y vectores propios de matrices reducidas

Una vez realizada la reducción a forma tridiagonal simétrica o Hessenberg superior estamos en disposición de encontrar los valores y vectores propios. Para estas matrices sí que calcularemos los valores propios como ceros del polinomio característico. Para ello, será necesario hacer una incursión en el mundo de los métodos de búsqueda de ceros de polinomios descritos en el capítulo 5; por ejemplo, utilizaremos el teorema de Sturm para separar los valores propios de matrices tridiagonales simétricas, y los métodos de la secante y de Newton con el fin de calcularlos efectivamente.

Métodos para matrices tridiagonales simétricas

Consideremos una matriz nxn tridiagonal simétrica

$$T=\begin{pmatrix} a_1 & b_2 & & & & \\ b_2 & a_2 & b_3 & & & \\ & b_3 & a_3 & b_4 & & \\ & & \ddots & \ddots & \ddots & \\ & & & b_{n-1} & a_{n-1} & b_n \\ & & & & b_n & a_n \end{pmatrix} .$$

Supondremos que $b_j \neq 0$ $(j=2\div n)$, ya que, en caso contrario, descompondríamos T en dos matrices del mismo tipo T_1 y T_2

$$T=\left(\begin{array}{c|c} T_1 & 0 \\ \hline 0 & T_2 \end{array}\right)$$

y, en virtud de la propiedad 6 de valores y vectores propios del apartado 2.1.3, para encontrar los valores propios de T sólo tendríamos que determinar por separado los de T_1 y T_2.

Como $\det(\lambda I-T)=(-1)^n\det(T-\lambda I)$, buscar los valores propios de T equivale a resolver la ecuación polinomial $p_n(\lambda)=\det(\lambda I-T)=0$.

Los valores $p_n(\lambda)$ se pueden calcular por recurrencia: comenzando con

$$p_0(\lambda)=1\ ,\quad p_1(\lambda)=\lambda-a_1\ ,$$

definimos el determinante $p_{j+1}(\lambda)$ de la matriz formada por las j primeras filas de las j primeras columnas de la matriz $\lambda I - T$; este determinante se encuentra a partir de los dos determinantes análogos anteriores mediante

$$p_j(\lambda) = (\lambda - a_j)p_{j-1}(\lambda) - \gamma_j p_{j-2}(\lambda) \ , \quad \gamma_j = b_j^2 \geq 0 \ (j \geq 2) \ .$$

Obsérvese que $p_j(\lambda)$ es un polinomio de grado j. La recurrencia que se acaba de establecer es muy parecida a la que cumplen los polinomios de Chebichev. El problema 2.16 usa esta similitud.

Nótese que no se precisa conocer explícitamente los coeficientes de $p_n(\lambda)$ con el fin de calcular los valores propios de T, y que, dado un valor cualquiera de λ, $p_n(\lambda)$ se calcula con $2n$ operaciones, aproximadamente.

La sucesión de polinomios $\{p_n(\lambda), p_{n-1}(\lambda), \ldots, p_1(\lambda), p_0(\lambda)\}$ constituye una sucesión de Sturm para $p_n(\lambda)$ sobre $\mathbb{R}$ (véase el apartado 5.2.4 para más detalles); aplicando el teorema de Sturm, tenemos que

- Los valores propios de T son todos reales y diferentes (suponiendo $b_j \neq 0$ $(j = 2 \div n)$). Además, si $p_n(a)p_n(b) \neq 0$, el número de valores propios de T entre a y b es igual a $V(a) - V(b)$, donde $V(\lambda)$ es el número de cambios de signo en la sucesión $\{p_0(\lambda), p_1(\lambda), \ldots, p_n(\lambda)\}$.

Con la ayuda de este resultado, podemos localizar y separar los valores propios de T; después podemos aplicar cualquier método de los descritos en el apartado 5.1.2 para encontrar ceros de una función, como por ejemplo los métodos de la secante o de Newton, con el fin de refinar su valor. Obsérvese que, si se quiere utilizar el método de Newton, el cálculo de $p_n'(\lambda)$ puede llevarse a cabo también por recurrencia

$$p_0'(\lambda) = 0 \ , \quad p_1'(\lambda) = 1 \ ,$$
$$p_j'(\lambda) = p_{j-1}(\lambda) + (\lambda - a_j)p_{j-1}'(\lambda) - \gamma_j p_{j-2}'(\lambda) \ (j = 2 \div n) \ .$$

Esta recurrencia se ha de usar conjuntamente con la anterior para calcular simultáneamente $p_n(\lambda)$ y $p_n'(\lambda)$.

Una vez obtenidos los valores propios λ_j $(j = 1 \div n)$ de T, hay que calcular los vectores propios asociados. Para determinar un vector propio v de T, asociado a un valor propio λ, hay que resolver el sistema

$$\begin{aligned} (\lambda - a_1)v_1 - b_2 v_2 &= 0 \\ -b_i v_{i-1} + (\lambda - a_i)v_i - b_{i+1}v_{i+1} &= 0 \ (i = 2 \div n-1) \ , \\ -b_n v_{n-1} + (\lambda - a_n)v_n &= 0 \end{aligned}$$

con la condición de que $v_1 \neq 0$ para que v no sea nulo.

Tomando $v_1 = 1$, y resolviendo recurrentemente este sistema, se obtiene

$$v_i = \frac{p_{i-1}(\lambda)}{b_2 \cdots b_i} \ (i = 2 \div n)$$

(véase el problema 2.15).

Este procedimiento no siempre está bien condicionado, ya que sólo conocemos aproximaciones de λ. Pueden obtenerse resultados mas precisos aplicando algunas iteraciones del método de la potencia inversa a $T - \lambda I$, aunque se tengan que efectuar más operaciones.

Método para matrices Hessenberg superior

En el caso de matrices H Hessenberg superior puede obtenerse también una relación de recurrencia para los determinantes de las matrices formadas por las j primeras filas de las j primeras columnas de $\lambda I - H$, $p_j(\lambda)$ $(j = 1 \div n)$, que conduce a la ecuación característica $p_n(\lambda) = 0$ que cumplen los valores propios de H.

Supongamos

$$H = \begin{pmatrix} h_{11} & h_{12} & \cdots & \cdots & h_{1,n-1} & h_{1n} \\ h_{21} & h_{22} & h_{23} & \ddots & \ddots & h_{2n} \\ & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & h_{n-1,n-2} & h_{n-1,n-1} & h_{n-1,n} \\ & & & & h_{n,n-1} & h_{n\,n} \end{pmatrix},$$

con $h_{j+1,j} \neq 0$ $(j = 1 \div n-1)$, como en el caso tridiagonal simétrico; en caso contrario pasaríamos a encontrar por separado los valores propios de las matrices Hessenberg superior H_1 y H_2 tales que

$$H = \left(\begin{array}{c|c} H_1 & N_1 \\ \hline 0 & H_2 \end{array}\right).$$

Con el fin de escribir mejor las relaciones de recurrencia de la sucesión de polinomios $p_j(\lambda)$ $(j = 0 \div n)$, tomamos

$$q_0(\lambda) = p_0(\lambda) = 1\ , \quad q_j(\lambda) = \frac{p_j(\lambda)}{h_{21}h_{32}\cdots h_{j+1,j}}\ (j = 2 \div n)\ ;$$

la recurrencia queda entonces en la forma

$$q_0(\lambda) = 1\ , \quad q_1(\lambda) = \frac{\lambda - h_{11}}{h_{21}}\ ,$$

$$q_j(\lambda) = \frac{(\lambda - h_{j\,j})q_{j-1}(\lambda) - [h_{1\,j}q_0(\lambda) + \cdots + h_{j-1,j}q_{j-2}(\lambda)]}{h_{j+1,j}}$$

$$(j = 2 \div n)\ .$$

A diferencia de lo que sucedía con matrices simétricas, los valores propios de matrices arbitrarias no son necesariamente reales. Esto nos indica que los ceros de $q_n(\lambda)$ no siempre serán reales. Se tendrán que buscar usando métodos de cálculo de ceros de polinomios que permitan encontrar ceros complejos (véase el capítulo 5); los vectores propios asociados se podrán buscar entonces con la ayuda del método de la potencia inversa.

2.3.7 Métodos de factorización

Los métodos que se mencionarán están basados en las factorizaciones LU (llamada aquí LR) y QR, presentadas en los apartados 2.2.3 y 2.2.4, respectivamente.

Método iterativo LR

Dada una matriz A nxn real, el *método iterativo LR* parte de la matriz $A_1 = A$ y, cuando conoce la matriz A_k, calcula su factorización LU (si existe)

$$A_k = L_k R_k \ ,$$

con L_k triangular inferior con unos en la diagonal y R_k, triangular superior; finalmente construye $A_{k+1} = R_k L_k$ $(k \geq 1)$.

Como

$$A_{k+1} = L_k^{-1} A_k L_k = \cdots = (L_1 \cdots L_k)^{-1} A (L_1 \cdots L_k) \ ,$$

las matrices A_k $(k \geq 1)$ son semejantes a la matriz A.

Método iterativo QR

El *método iterativo QR* usa la factorización QR en lugar de la LU: así,

$$A_k = Q_k R_k \ , \quad A_{k+1} = R_k Q_k \ ,$$

y entonces

$$A_{k+1} = Q_k^\top A_k Q_k = \cdots = (Q_1 \cdots Q_k)^\top A (Q_1 \cdots Q_k) \ .$$

Por lo tanto, las matrices A_k $(k \geq 1)$ son también semejantes a la matriz A.

Consideraciones generales

Las sucesiones (A_k) $(k \geq 0)$ convergen a una matriz triangular superior, bajo condiciones bastante generales; es suficiente, por ejemplo, que se verifiquen las tres condiciones siguientes:

- los valores propios de A están separados en módulo

$$|\lambda_1| > \cdots > |\lambda_n| \ ,$$

- A es diagonalizable: existe una matriz regular X tal que $D = X^{-1} A X$ es diagonal,
- existe la factorización LU de las matrices X y X^{-1}.

El método iterativo QR conserva la simetría de la matriz; esto es, si la matriz A es simétrica, todas las matrices A_k $(k \geq 1)$ son simétricas, y la convergencia a una matriz triangular superior se traduce en convergencia a una matriz diagonal.

Los métodos iterativos LR y QR conservan además las formas Hessenberg superior de las matrices; el método iterativo QR conserva, así, el carácter de matriz tridiagonal simétrica.

El número de operaciones que requieren estos métodos es elevado y son bastante costosos cuando se aplican a matrices llenas arbitrarias; no obstante, resulta muy eficiente transformar previamente la matriz A a una forma reducida (Hessenberg superior o tridiagonal simétrica) por alguno de los métodos dados en el apartado anterior (de Givens o de Householder) y aplicar después los métodos descritos aquí a las matrices reducidas.

COMENTARIOS BIBLIOGRÁFICOS

Para este tema y el próximo, se han supuesto conocidos los resultados teóricos elementales sobre resolución de sistemas lineales y problemas de valores y vectores propios, que pueden ser consultados, por ejemplo, en [Hir74], [Que71], ... Varios libros dedican a este tema algún capítulo introductorio desde un punto de vista más algorítmico, como [Cia82], [DM73], [Fro69], [Hou64], [IK66], donde se pueden encontrar las propiedades importantes aquí desarrolladas.

[Cia82] es una de las mejores referencias para la resolución de sistemas lineales, junto con [Wil64] para métodos directos, y [You71] para métodos iterativos. Estos tres libros cubren esencialmente todos los resultados de este capítulo. Otras referencias generales básicas son [DB74], [Gas66], [Hil74], [Jac77], [Ral65]. En [CdB72] y, principalmente, [WR71] se presentan programas con la implementación efectiva de los diferentes métodos, tanto de sistemas lineales como de valores y vectores propios. Las matrices escasas no han sido tratadas, aunque aparecen frecuentemente en la resolución de muchos problemas reales; una referencia muy útil puede ser [Ric81]. El estudio de propagación de los errores se encuentra en [IK66] y primordialmente en [Wil64]. Para factorizaciones QR, la referencia estándar es [LH74], que también incorpora rutinas en lenguaje `FORTRAN`. Los cálculos sobre el número de operaciones de los diferentes métodos de este tema aparecen desarrollados en muchas referencias como [Cia82], [IK66] y [LH74].

Como referencias específicas para el cálculo de valores y vectores propios, cabe destacar a [Cia82], [Ham70], [RR78], [Ste73] y [Wil65]. Para diversos resultados sobre localización de valores propios, se puede consultar [SB80]. Los diferentes métodos presentados de deflación se encuentran en [DM73], [Fro69] y [RR78]. Una implementación más efectiva del método de la potencia y variantes utiliza otras técnicas, como la aceleración de Aitken y una elección adecuada de los desplazamientos d para la matriz $A - dI$, y se pueden encontrar en [IK66], [RR78] y [Wil64]. El método de Jacobi y sus diferentes variantes se presentan de una manera detallada en [Cia82] y [IK66]. Este último texto también incorpora los cálculos sobre el número de operaciones necesarias para llevar a cabo los métodos de Givens y de Householder. Un método alternativo para reducir matrices no simétricas a matrices Hessenberg superior consiste en usar eliminación gaussiana con pivotaje por columnas, tal como puede verse en [RR78]. En este libro y en [Cia82] hay una buena discusión sobre el teorema de Sturm. La convergencia de los métodos iterativos LR y QR, así como su implementación efectiva, usando matrices desplazadas, como en el método de la potencia, se puede encontrar en [Cia82], [LH74], [RR78] y, sobre todo, en el artículo de B.N. Parlett en [RW67] y en [Par80].

PROBLEMAS RESUELTOS

Problema 2.1 *a) Si $A = (a_{ij})$ es una matriz definida positiva, demostrar que:*

i) $a_{ii} > 0$ $(i = 1 \div n)$;

ii) $\max_{i,j} |a_{ij}| = \max_i a_{ii}$;

iii) las submatrices $A^{(k)} = (a_{ij}^{(k)})_{i,j=k \div n}$ $(k = 2 \div n)$, *obtenidas al aplicar el método de Gauss, son definidas positivas y, por lo tanto, se puede realizar el proceso sin tener que recurrir a los pivotajes;*

iv) $a_{ii}^{(k+1)} \leq a_{ii}^{(k)}$ $(i = k+1 \div n,\ k = 1 \div n-1)$.

b) Aplicación: Demostrar que la matriz

$$\begin{pmatrix} 13 & 11 & 11 \\ 11 & 13 & 11 \\ 11 & 11 & 13 \end{pmatrix}$$

es definida positiva, hallando su factorización de Cholesky y su determinante.

SOLUCIÓN:

a) i) Se deduce de la definición de matriz definida positiva, tomando $x = e^{(i)}$ (vector i-ésimo de la base canónica) :

$$a_{ii} = e^{(i)\top} A e^{(i)} > 0 \ (i = 1 \div n) \ .$$

ii) El criterio de Sylvester asegura que una matriz es definida positiva si y sólo si los determinantes principales Δ_k $(k \geq 0)$ son estrictamente positivos. Una matriz definida positiva sigue siéndolo después de permutar dos de sus filas al mismo tiempo que se permutan las columnas correspondientes, dado que esto representa sólo un cambio de variable consistente en permutar las componentes correspondientes de los vectores. Así, cambiando las filas y columnas i y j por las filas y columnas 1 y 2 de la matriz A, se obtiene una matriz definida positiva con determinante principal de orden 2

$$\begin{vmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{vmatrix} = a_{ii}a_{jj} - a_{ij}^2 > 0 \ ;$$

por lo tanto,

$$|a_{ij}| < a_{ii} \quad \text{ó} \quad |a_{ij}| < a_{jj} \quad (i, j = 1 \div n,\ i \neq j) \ ,$$

y obtenemos la relación pedida

$$\max_{i,j} |a_{ij}| = \max_i a_{ii} \ .$$

iii) Lo demostraremos por inducción. El enunciado del problema nos dice que $A^{(1)} = A$ es definida positiva. Supondremos que $A^{(k)}$ lo es y deduciremos que $A^{(k+1)}$ también es definida positiva para $k = 1 \div n-1$.

Si $A^{(k)}$ es definida positiva, resulta que:

- $a_{kk}^{(k)} > 0$.

  Se deduce aplicando el apartado i) a la matriz $A^{(k)}$. Por lo tanto, el paso k-ésimo del método de Gauss podrá realizarse sin pivotaje.

- $A^{(k+1)}$ es simétrica.

  Esta afirmación se deduce de la fórmula de obtención de los elementos de $A^{(k+1)}$ a partir de los de $A^{(k)}$ y del hecho de que $A^{(k)}$ es simétrica:

$$\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} \\ &= a_{ji}^{(k)} - \frac{a_{jk}^{(k)}}{a_{kk}^{(k)}} a_{ki}^{(k)} = a_{ji}^{(k+1)} \ (i,j = k+1 \div n) \ . \end{aligned}$$

- $A^{(k+1)}$ es definida positiva.

  Como es simétrica, aplicando el criterio de Sylvester, será suficiente demostrar que todos sus determinantes principales son positivos. Dado que el proceso de eliminación gaussiana sin pivotaje preserva el determinante, tenemos que

$$\det\,(A^{(k)})_j = a_{kk}^{(k)} \det\,(A^{(k+1)})_{j-1} \ \ (j = 2 \div n-k+1) \ .$$

Aplicando el criterio de Sylvester a la matriz $A^{(k)}$, $\det\,(A^{(k)})_j > 0$. Como $a_{kk}^{(k)} > 0$, resulta que $\det\,(A^{(k+1)})_{j-1} > 0$ $(j = 2 \div n-k+1)$, como queríamos demostrar.

Obsérvese que, además de demostrar que las submatrices $A^{(k)}$ $(k = 2 \div n)$ son definidas positivas, hemos deducido que se puede aplicar el método de eliminación gaussiana sin necesidad de recurrir al pivotaje.

iv) Usando la simetría de $A^{(k)}$ i el hecho que $a_{kk}^{(k)} > 0$, resulta que

$$a_{ii}^{(k+1)} = a_{ii}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{ki}^{(k)} = a_{ii}^{(k)} - \frac{a_{ik}^{(k)2}}{a_{kk}^{(k)}} \leq a_{ii}^{(k)} \ (i = k+1 \div n) \ .$$

b) Con el fin de obtener la factorización de Cholesky, buscaremos primero una factorización de la forma $LDL^\top$ válida para una matriz simétrica cualquiera; las fórmulas recurrentes de cálculo de los elementos de las matrices L y D se deducen, por ejemplo, de las fórmulas del método de Crout, entendiendo que la matriz U es ahora $DL^\top$.

Tenemos así,

$$d_{11} = a_{11} = 13 \ , \qquad l_{21} = \tfrac{a_{21}}{d_{11}} = \tfrac{11}{13} \ ,$$
$$d_{22} = a_{22} - l_{21}^2 d_{11} = \frac{48}{13} \ , \qquad l_{31} = \tfrac{a_{31}}{d_{11}} = \tfrac{11}{13} \ ,$$
$$l_{32} = \frac{a_{32} - l_{31} l_{21} d_{11}}{d_{22}} = \frac{11}{24} \ , \quad d_{33} = a_{33} - l_{31}^2 d_{11} - l_{32}^2 d_{22} = \tfrac{35}{12} \ .$$

En forma matricial,

$$A = LDL^\top = \begin{pmatrix} 1 & 0 & 0 \\ \frac{11}{13} & 1 & 0 \\ \frac{11}{13} & \frac{11}{24} & 1 \end{pmatrix} \begin{pmatrix} 13 & 0 & 0 \\ 0 & \frac{48}{13} & 0 \\ 0 & 0 & \frac{35}{12} \end{pmatrix} \begin{pmatrix} 1 & \frac{11}{13} & \frac{11}{13} \\ 0 & 1 & \frac{11}{24} \\ 0 & 0 & 1 \end{pmatrix} \ .$$

La matriz A es definida positiva dado que todos los elementos diagonales de D son positivos y su factorización de Cholesky $A = \mathcal{L}\mathcal{L}^\top$ se obtiene como

$$\mathcal{L} = \begin{pmatrix} \sqrt{13} & 0 & 0 \\ \frac{11}{\sqrt{13}} & \sqrt{\frac{48}{13}} & 0 \\ \frac{11}{\sqrt{13}} & \frac{11}{24}\sqrt{\frac{48}{13}} & \sqrt{\frac{35}{12}} \end{pmatrix} = \begin{pmatrix} 3.6056 & 0 & 0 \\ 3.0509 & 1.9215 & 0 \\ 3.0509 & 0.8807 & 1.7078 \end{pmatrix} \ .$$

Finalmente,

$$\det A = d_{11} d_{22} d_{33} = 13 \frac{48}{13} \frac{35}{12} = 140 \ .$$

Problema 2.2 *Calcular el número de operaciones necesarias para resolver un sistema $Ax = b$, donde A es una matriz pentadiagonal tal que $a_{ij} = 0$, si $|i - j| = 1$.*

SOLUCIÓN:

La matriz A es de la forma

$$A = \begin{pmatrix} a_1 & 0 & c_1 & & & & & \\ 0 & a_2 & 0 & c_2 & & & & \\ d_3 & 0 & a_3 & 0 & c_3 & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & d_{n-2} & 0 & a_{n-2} & 0 & c_{n-2} \\ & & & & d_{n-1} & 0 & a_{n-1} & 0 \\ & & & & & d_n & 0 & a_n \end{pmatrix} \ .$$

Realizamos las dos fases del método de Gauss:

TRIANGULARIZACIÓN

En el paso k-ésimo tenemos que anular únicamente el elemento d_{k+2}; exceptuando esta anulación, los únicos elementos modificados son el a_{k+2} de la matriz A y el b_{k+2} del vector b, que se sustituyen por

$$\bar{a}_{k+2} = a_{k+2} - \frac{d_{k+2}}{\bar{a}_k} c_k \ , \quad \bar{b}_{k+2} = b_{k+2} - \frac{d_{k+2}}{\bar{a}_k} b_k \quad (k = 1 \div n-2) \ ,$$

donde se ha tomado inicialmente $\bar{a}_1 = a_1$, $\bar{a}_2 = a_2$, $\bar{b}_1 = b_1$ y $\bar{b}_2 = b_2$, dado que estos elementos no se modifican en el proceso. En cada paso se han de realizar, así, 2 restas, 2 multiplicaciones y 1 división; en total: $2(n-2)$ restas y multiplicaciones y $n-2$ divisiones.

SUSTITUCIÓN HACIA ATRÁS

Hay que obtener las componentes del vector solución x recurrentemente, según

$$x_n = \frac{\bar{b}_n}{\bar{a}_n} \ , \quad x_{n-1} = \frac{\bar{b}_{n-1}}{\bar{a}_{n-1}} \ ;$$

$$x_k = \frac{\bar{b}_k - c_k x_{k+2}}{\bar{a}_k} \ (k = n-2 \div 1) \ .$$

En esta etapa son necesarias, así, $n-2$ restas y multiplicaciones y n divisiones.

En la tabla siguiente se resumen los números de operaciones encontrados:

	$-$	$*$	$/$
Triangularización	$2(n-2)$	$2(n-2)$	$n-2$
Sustitución	$n-2$	$n-2$	n
Total	$3(n-2)$	$3(n-2)$	$2(n-1)$

Contando las operaciones como multiplicaciones/divisiones+sumas/restas, en total es preciso realizar $5n-8$ operaciones.

Problema 2.3 *Consideremos la matriz nxn*

$$A = \begin{pmatrix} a & 1 & & & & & \\ 1 & a & 1 & & & & \\ & 1 & a & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & 1 & a & 1 \\ & & & & & 1 & a \end{pmatrix} .$$

a) Demostrar que la matriz A es definida positiva para $a \geq 2$.

b) Si $a \geq 2$, encontrar un método recurrente para llevar a cabo la factorización de Cholesky de la matriz A, $A = \mathcal{L}\mathcal{L}^\top$ con

$$\mathcal{L} = \begin{pmatrix} \alpha_1 & & & & & & \\ \beta_2 & \alpha_2 & & & & & \\ & \beta_3 & \alpha_3 & & & & \\ & & \ddots & \ddots & & & \\ & & & \ddots & \ddots & & \\ & & & & \beta_{n-1} & \alpha_{n-1} & \\ & & & & & \beta_n & \alpha_n \end{pmatrix} .$$

c) Si el elemento a_{11} de A tiene un error absoluto δ y los otros elementos son exactos, dar una estimación del error relativo de α_n al efectuar la factorización, suponiendo δ suficientemente pequeño.

SOLUCIÓN:

a) Por el criterio de Sylvester, hay que demostrar que los determinantes principales Δ_j ($j = 1 \div n$) son positivos.

Es evidente que $\Delta_1 = a$ y $\Delta_2 = a\Delta_1 - 1$ y, desarrollando el determinante de $(A)_{j+1}$ por la última fila o columna,

$$\Delta_{j+1} = a\Delta_j - \Delta_{j-1} \quad (j = 2 \div n-1) .$$

Veremos por inducción que $\Delta_{j+1} > \Delta_j > 0$ ($j = 1 \div n-1$).

La relación se cumple para $j = 1$: $\Delta_2 = a^2 - 1 > a = \Delta_1 > 0$, ya que $a \geq 2$; supuesta la validez de la relación hasta $j-1$, tenemos

$$\Delta_{j+1} = a\Delta_j - \Delta_{j-1} \geq 2\Delta_j - \Delta_{j-1} = \Delta_j + (\Delta_j - \Delta_{j-1}) > \Delta_j > 0 ,$$

ya que se ha supuesto que $\Delta_j > \Delta_{j-1} > 0$.

b) Suponemos realizada la factorización $\mathcal{L}\mathcal{L}^\top$ de la matriz A con

$$\mathcal{L} = \begin{pmatrix} \alpha_1 & & & & \\ \beta_2 & \alpha_2 & & & \\ & \ddots & \ddots & & \\ & & \beta_{n-1} & \alpha_{n-1} & \\ & & & \beta_n & \alpha_n \end{pmatrix} .$$

Hay que determinar los valores de α_j ($j = 1 \div n$) y de β_j ($j = 2 \div n$) de manera que se cumpla la igualdad matricial $A = \mathcal{L}\mathcal{L}^\top$ que, componente a componente, se escribe:

$$a = \alpha_1^2 ;$$

$$1 = \alpha_{j-1}\beta_j , \quad a = \alpha_j^2 + \beta_j^2 \ (j = 2 \div n) .$$

despejando las incógnitas, tenemos la siguiente recurrencia:

$$\alpha_1 = \sqrt{a} ; \quad \beta_j = \frac{1}{\alpha_{j-1}}, \quad \alpha_j = \sqrt{a - \beta_j^2} \ (j = 2 \div n) .$$

c) Aplicando la fórmula aproximada de propagación del error a las recurrencias determinadas en el apartado b):

$$e_a(\alpha_1) \simeq \frac{1}{2\alpha_1}\delta \ ;$$

$$e_a(\beta_j) \simeq -\frac{1}{\alpha_{j-1}^2} e_a(\alpha_{j-1}) \ , \quad e_a(\alpha_j) \simeq -\frac{\beta_j}{\alpha_j} e_a(\beta_j) \ (j = 2 \div n) \ .$$

Así,

$$e_a(\alpha_j) \simeq \frac{\beta_j}{\alpha_j \alpha_{j-1}^2} e_a(\alpha_{j-1}) = \frac{1}{\alpha_j \alpha_{j-1}^3} e_a(\alpha_{j-1})$$

y

$$e_a(\alpha_n) \simeq \frac{1}{\alpha_n \alpha_{n-1}^3} \cdots \frac{1}{\alpha_2 \alpha_1^3} e_a(\alpha_1) = \frac{\alpha_n^3 \alpha_1}{\prod_{j=1}^n \alpha_j^4} e_a(\alpha_1) \simeq \frac{\alpha_n^3}{2\prod_{j=1}^n \alpha_j^4}\delta \ .$$

Teniendo en cuenta que

$$\det A = \prod_{j=1}^n \alpha_j^2 \ ,$$

resulta

$$e_a(\alpha_n) \simeq \frac{\alpha_n^3}{2(\det A)^2}\delta \ .$$

Así pues, el error relativo en α_n puede estimarse por

$$e_r(\alpha_n) \simeq \frac{\alpha_n^2}{2(\det A)^2}\delta \ .$$

Problema 2.4 *Demostrar que las normas matriciales subordinadas a las normas vectoriales*

$$\|x\|_1 = \sum_i |x_i| \ , \quad \|x\|_\infty = \max_i |x_i| \ ,$$

son respectivamente:

$$\|A\|_1 = \max_j \sum_i |a_{ij}| \ , \quad \|A\|_\infty = \max_i \sum_j |a_{ij}| \ .$$

SOLUCIÓN:

La definición de norma matricial subordinada a la norma vectorial $\| \ \|_1$ es

$$\|A\|_1 = \max_{x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} \ .$$

Como

$$\|Ax\|_1 = \sum_i \left| \sum_j a_{ij} x_j \right| \leq \sum_j \left(\sum_i |a_{ij}| \right) |x_j| \leq \max_j \sum_i |a_{ij}| \, \|x\|_1 \ ,$$

tenemos que

$$\|A\|_1 \leq \max_j \sum_i |a_{ij}| \ .$$

Ahora bien, si

$$\max_j \sum_i |a_{ij}| = \sum_i |a_{il}| \ ,$$

escogiendo $\bar{x}$ de componentes $\bar{x}_j = \delta_{jl}$ tenemos que $\|\bar{x}\|_1 = 1$ y

$$\|A\bar{x}\|_1 = \sum_i \left| \sum_j a_{ij} \bar{x}_j \right| = \sum_i |a_{il}| = \max_j \sum_i |a_{ij}| \ .$$

De donde,

$$\|A\|_1 = \max_j \sum_i |a_{ij}| \ .$$

De manera análoga, tenemos para $\| \ \|_\infty$:

$$\|Ax\|_\infty = \max_i \left| \sum_j a_{ij} x_j \right| \leq \max_i \sum_j |a_{ij}| \, |x_j| \leq \max_i \sum_j |a_{ij}| \, \|x\|_\infty \ ,$$

lo cual implica

$$\|A\|_\infty \leq \max_i \sum_j |a_{ij}| \ .$$

Si ahora

$$\max_i \sum_j |a_{ij}| = \sum_j |a_{kj}| \ ,$$

escogiendo $\bar{x}$ de componentes $\bar{x}_j = \operatorname{sgn}(a_{kj})$, tenemos que $\|\bar{x}\|_\infty = 1$ y

$$\begin{aligned} \|A\bar{x}\|_\infty &= \max_i \left| \sum_j a_{ij} \bar{x}_j \right| = \max_i \left| \sum_j a_{ij} \ \operatorname{sgn}(a_{kj}) \right| \\ &= \sum_j |a_{kj}| = \max_i \sum_j |a_{ij}| \ . \end{aligned}$$

De donde,

$$\|A\|_\infty = \max_i \sum_j |a_{ij}| \ .$$

Problema 2.5 *a) Partiendo de* $(A+\delta A)(x+\delta x)=b+\delta b$, *demostrar que, si*

$$\|A^{-1}\|\ \|\delta A\|<1\ ,$$

entonces

$$\frac{\|\delta x\|}{\|x\|}\leq\frac{\mu(A)}{1-\mu(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|}\right)\ ,$$

siempre que la norma matricial usada sea consistente con la norma vectorial.

b) Encontrar una matriz 3x3 A *tal que cumpla* $\mu_\infty(A)>10^6$, *a pesar de que* $\|A\|_\infty<10^{-6}$.

c) Resolver los sistemas correspondientes $Ax=b$ *con*

$$i)\ b=b^{(1)}=\begin{pmatrix}1\\1\\1\end{pmatrix}\ ,\quad ii)\ \ b=b^{(2)}=\begin{pmatrix}0.9\\1.1\\1\end{pmatrix}\ .$$

d) Comentar los resultados obtenidos en el apartado c).

SOLUCIÓN:

a) De $(A+\delta A)(x+\delta x)=b+\delta b$ y $Ax=b$, se deduce

$$\delta x=A^{-1}\delta b-A^{-1}\delta A(x+\delta x)\ .$$

Tomando normas consistentes y usando sus propiedades,

$$\begin{aligned}\|\delta x\| &= \|A^{-1}\delta b-A^{-1}\delta A(x+\delta x)\|\\ &\leq \|A^{-1}\delta b\|+\|A^{-1}\delta A(x+\delta x)\|\\ &\leq \|A^{-1}\|\ \|\delta b\|+\|A^{-1}\|\ \|\delta A\|\ \|x+\delta x\|\ ;\end{aligned}$$

como $b=Ax$, resulta que

$$\frac{1}{\|x\|}\leq\frac{\|A\|}{\|b\|}\ ,\quad \text{si } x,b\neq 0\ .$$

Entonces,

$$\begin{aligned}\frac{\|\delta x\|}{\|x\|} &\leq \mu(A)\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|x+\delta x\|}{\|x\|}\frac{\|\delta A\|}{\|A\|}\right)\\ &\leq \mu(A)\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|}\right)+\mu(A)\frac{\|\delta A\|}{\|A\|}\frac{\|\delta x\|}{\|x\|}\ ;\end{aligned}$$

y, usando que

$$\mu(A)\frac{\|\delta A\|}{\|A\|} = \|A^{-1}\|\ \|\delta A\| < 1\ ,$$

tenemos finalmente

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mu(A)}{1-\mu(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta b\|}{\|b\|}+\frac{\|\delta A\|}{\|A\|}\right)\ .$$

b) Consideremos δ suficientemente pequeño y ϵ mucho menor que δ; entonces

$$A = \delta\begin{pmatrix} \epsilon & 1 & 0 \\ -\epsilon & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

cumple que $\|A\|_\infty = \delta(1+\epsilon) \simeq \delta$ y su inversa

$$A^{-1} = \frac{1}{2\epsilon\delta}\begin{pmatrix} 1 & -1 & 0 \\ \epsilon & \epsilon & 0 \\ 0 & 0 & 2\epsilon \end{pmatrix}$$

cumple que $\|A^{-1}\|_\infty = \frac{1}{\epsilon\delta}$; con lo que su número de condición es muy grande

$$\mu_\infty(A) = \frac{1+\epsilon}{\epsilon} \simeq \frac{1}{\epsilon}\ .$$

Eligiendo, por ejemplo, $\epsilon = 10^{-7}$ y $\delta = 10^{-7}$ tenemos:

$$\|A\|_\infty = 10^{-7}(1+10^{-7}) < 10^{-6}\ ,\quad \mu_\infty(A) = \frac{1+10^{-7}}{10^{-7}} > 10^6\ .$$

c) Resolvemos primeramente el sistema con $b = b^{(1)} = (1\ \ 1\ \ 1)^\top$, la solución exacta será

$$x^{(1)} = A^{-1}b^{(1)} = (0\ \ \frac{1}{\delta}\ \ \frac{1}{\delta})^\top\ .$$

Si hacemos lo mismo para $b^{(2)} = (0.9\ \ 1.1\ \ \ 1)^\top$, la solución será $x^{(2)} = x^{(1)} + \delta x$ con $\delta x = A^{-1}\delta b$, siendo $\delta b = (-0.1\ \ 0.1\ \ 0)^\top$; resulta así

$$\delta x = A^{-1}\delta b = -\frac{1}{10\epsilon\delta}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}\ .$$

Para los valores dados de ϵ y δ se tiene:

$$\|x^{(1)}\|_\infty = 10^7\ ,\quad \|\delta x\|_\infty = 10^{13}\ .$$

d) Los resultados del apartado c) nos indican que una perturbación pequeña, en términos relativos, del vector $b^{(1)}$

$$\frac{\|\delta b\|_\infty}{\|b^{(1)}\|_\infty} = 10^{-1}$$

produce una perturbación importante en la solución del sistema

$$\frac{\|\delta x\|_\infty}{\|x^{(1)}\|_\infty} = 10^6 \ .$$

Esta amplificación es debida a que la matriz A tiene un número de condición muy alto, que eleva notablemente el valor de la cota del error relativo de x

$$\frac{\|\delta x\|_\infty}{\|x^{(1)}\|_\infty} \leq \mu_\infty(A)\frac{\|\delta b\|_\infty}{\|b^{(1)}\|_\infty} = \frac{1+10^{-7}}{10^{-7}}\frac{1}{10} = (1+10^{-7})10^6 \ .$$

Nótese que, en este caso, la cota anterior es prácticamente alcanzada.

Problema 2.6 *Demostrar que, para el sistema*

$$\left.\begin{array}{rcrcl} ax & + & by & = & p \\ cx & + & dy & = & q \end{array}\right\} ,$$

una condición necesaria y suficiente de convergencia de los métodos iterativos de Jacobi y Gauss-Seidel es $|bc| < |ad|$.

SOLUCIÓN:

Los métodos iterativos de resolución de sistemas lineales convergen si y sólo si la matriz de iteración tiene todos los valores propios menores que 1 en módulo. Para los métodos iterativos de Jacobi y de Gauss-Seidel, una vez hecha la descomposición $A = D(L+I+U)$ con D diagonal, L triangular inferior con ceros en la diagonal y U triangular superior con ceros en la diagonal, las matrices de iteración son $B_J = -(L+U)$ y $B_{GS} = -(I+L)^{-1}U$, respectivamente.

En el caso planteado:

$$\begin{aligned} A &= \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}\left[\begin{pmatrix} 0 & 0 \\ \frac{c}{d} & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & \frac{b}{a} \\ 0 & 0 \end{pmatrix}\right] \\ &= D(L+I+U) \ , \\ B_J &= -\left[\begin{pmatrix} 0 & 0 \\ \frac{c}{d} & 0 \end{pmatrix} + \begin{pmatrix} 0 & \frac{b}{a} \\ 0 & 0 \end{pmatrix}\right] = \begin{pmatrix} 0 & -\frac{b}{a} \\ -\frac{c}{d} & 0 \end{pmatrix} \ , \\ B_{GS} &= -\left[\begin{pmatrix} 0 & 0 \\ \frac{c}{d} & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]^{-1}\begin{pmatrix} 0 & \frac{b}{a} \\ 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -\frac{b}{a} \\ 0 & \frac{cb}{da} \end{pmatrix} \ . \end{aligned}$$

Los módulos de los valores propios de las matrices de iteración son

$$\sqrt{\left|\frac{bc}{ad}\right|} \ , \quad \text{para } B_J \ ;$$

$$0 \ , \quad \left|\frac{bc}{ad}\right| \ , \quad \text{para } B_{GS} \ .$$

Por lo tanto, los métodos convergen si y sólo si

$$|bc| < |ad| \ .$$

Nótese finalmente que, en este caso especial, la velocidad de convergencia del método iterativo de Gauss-Seidel es el doble de la del de Jacobi.

Problema 2.7 *Queremos resolver por el método iterativo de Jacobi el sistema $Ax = b$ partiendo de $x^{(0)} = 0$. Tenemos la siguiente información sobre los datos:*

$$|a_{ii}| > \alpha \sum_{j \neq i} |a_{ij}| \ \ (i = 1 \div n) \ , \quad \alpha > 1 \ ,$$

$$0 < \gamma < |a_{ii}| \ \ (i = 1 \div n) \ , \quad \|b\|_\infty = \beta \ ,$$

donde α, β, γ son constantes conocidas.

a) ¿Cuántas iteraciones k del método son necesarias para garantizar un error en la norma del máximo menor que ϵ: $\|x^{(k)} - x\|_\infty < \epsilon$?

b) Aplicación: ¿Cuánto vale k si $\alpha = 2$, $\beta = 10$, $\gamma = 1$ y $\epsilon = 10^{-6}$?

SOLUCIÓN:

a) El método iterativo de Jacobi se puede aplicar siempre a matrices de este tipo porque los elementos diagonales son no nulos. Si descomponemos la matriz en la forma $A = D(I + N)$, donde D es una matriz diagonal y N tiene los elementos diagonales nulos, el método iterativo de Jacobi se basa en la equivalencia entre el sistema $Ax = b$ de partida y el sistema $x = D^{-1}b - Nx$; el método iterativo en cuestión será, así,

$$x^{(0)} = 0 \ ; \quad x^{(k+1)} = D^{-1}b - Nx^{(k)} \ \ (k \geq 0) \ .$$

La convergencia del método está asegurada porque A es estrictamente diagonal dominante. El problema consiste ahora en acotar $\|x^{(k)} - x\|_\infty$.

Nótese que el comportamiento de los errores vectoriales $x^{(k)} - x$ es

$$x^{(k)} - x = -N(x^{(k-1)} - x)$$

y, por lo tanto,

$$\|x^{(k)} - x\|_\infty \leq \|N\|_\infty \|x^{(k-1)} - x\|_\infty \ .$$

Calculemos ahora una acotación de $\|N\|_\infty$. Los elementos de $N = (n_{ij})$ son

$$n_{ij} = \begin{cases} \frac{a_{ij}}{a_{ii}} & (\text{ si } i \neq j) \\ 0 & (\text{ si } i = j) \end{cases} \ ;$$

entonces,

$$\|N\|_\infty = \max_i \sum_j |n_{ij}| = \max_i \sum_{j \neq i} \left| \frac{a_{ij}}{a_{ii}} \right| = \max_i \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| < \frac{1}{\alpha} \ .$$

De donde,

$$\|x^{(k)} - x\|_\infty < \frac{1}{\alpha} \|x^{(k-1)} - x\|_\infty < \cdots < \frac{1}{\alpha^k} \|x^{(0)} - x\|_\infty = \frac{1}{\alpha^k} \|x\|_\infty \ .$$

Como $x = D^{-1}b - Nx$, tomando normas, tenemos

$$\|x\|_\infty \leq \|D^{-1}\|_\infty \|b\|_\infty + \|N\|_\infty \|x\|_\infty < \frac{1}{\gamma}\beta + \frac{1}{\alpha}\|x\|_\infty \ ;$$

por lo tanto,

$$\|x\|_\infty < \frac{\alpha\beta}{(\alpha - 1)\gamma} \ , \quad \|x^{(k)} - x\|_\infty < \frac{\beta}{\alpha^{k-1}(\alpha - 1)\gamma} \ .$$

En consecuencia, si

$$\alpha^{k-1} > \frac{\beta}{\epsilon(\alpha - 1)\gamma} \ ,$$

entonces $\|x^{(k)} - x\|_\infty < \epsilon$.

Tomando logaritmos, obtenemos la siguiente acotación inferior del número de iteraciones necesarias para asegurar que el error, en la norma del máximo, sea menor que ϵ:

$$k > 1 + \frac{\log \frac{\beta}{\epsilon(\alpha-1)\gamma}}{\log \alpha} \ .$$

b) En la aplicación pedida, para $\alpha = 2$, $\beta = 10$, $\gamma = 1$ y $\epsilon = 10^{-6}$, se obtiene $k > 24.25$; así pues, 25 iteraciones serán suficientes para garantizar aquella acotación del error.

Problema 2.8 *Sea A una matriz nxn con valores propios* λ_j $(j = 1 \div n)$, *reales entre* $m > 0$ *y* M*:*

$$0 < m = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = M.$$

Para resolver el sistema $Ax = b$, se considera la familia de métodos iterativos

$$x^{(k+1)} = B_\omega x^{(k)} + c_\omega \ ,$$

donde $B_\omega = I - \omega A$ y $c_\omega = \omega b$.

a) Demostrar que son convergentes para todos los valores $\omega \in (0, \frac{2}{M})$.

b) Si $\rho(\omega)$ indica el radio espectral de B_ω, demostrar que

$$\rho(\omega^*) = \min_\omega \rho(\omega) = \frac{M-m}{M+m} \ ,$$

para el valor óptimo de ω, $\omega^* = \frac{2}{M+m}$.

SOLUCIÓN:

a) El método iterativo será convergente cuando todos los valores propios de B_ω tengan módulo menor que 1.

Los valores propios de B_ω son de la forma $\mu_j = 1 - \omega\lambda_j$ $(j = 1 \div n)$, cumpliéndose la relación

$$1 > 1 - \omega m = 1 - \omega\lambda_1 \geq 1 - \omega\lambda_2 \geq \cdots \geq 1 - \omega\lambda_n = 1 - \omega M \ .$$

Se tiene, así, que

$$\rho(\omega) = \max(|1 - \omega M| \, , |1 - \omega m|) \ .$$

Representando las funciones, tal como muestra la figura 2.3, se observa que:

1. Estas funciones coinciden para $\omega = 0$ y para otro valor $\omega = \omega^*$ y el máximo $\rho(\omega)$ de ambas funciones coincide con la segunda de ellas sólo en el intervalo $[0, \omega^*]$, y con la primera, para los otros valores.

2. La función ρ toma el valor 1 para $\omega = 0$ y $\omega = \omega_c$ donde la función $\omega \mapsto |1 - \omega M|$ vale 1; esto es, cuando $\omega_c M - 1 = 1$: $\omega_c = \frac{2}{M}$. Esta función se mantiene menor que 1 en el intervalo $(0, \omega_c)$ y es mayor o igual que 1 para el resto de valores reales.

3. La función ρ es decreciente, hasta el valor de $\omega = \omega^*$, y creciente, a partir de este valor.

Del punto 2 del estudio de la función ρ se deduce la convergencia del método sólo para valores del intervalo $(0, \frac{2}{M})$.

b) El valor óptimo de ω se encuentra cuando el radio espectral de la matriz de iteración es mínimo; es decir, cuando la función ρ es mínima. Según el punto 3, este mínimo se alcanza en $\omega = \omega^*$, descrito en el punto 1 como el valor de ω, diferente de 0, en el cual coinciden las dos funciones. Es decir,

$$1 - \omega^* m = \omega^* M - 1 \ , \quad \omega^* = \frac{2}{M+m} \ .$$

El valor mínimo de la función ρ será, así,

$$\rho(\omega^*) = \omega^* M - 1 = \frac{M-m}{M+m} \ .$$

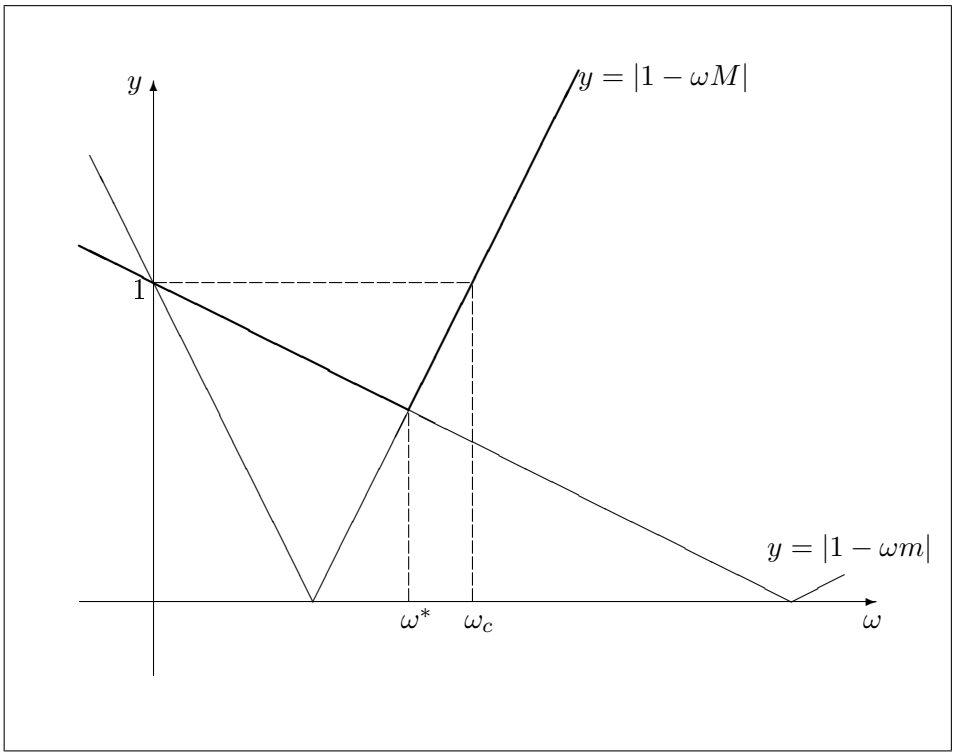


Figura 2.3: Gráficas de las funciones $y = |1 - \omega M|$ e $y = |1 - \omega m|$.

Problema 2.9 *Sean A una matriz nxn regular y X_0, una aproximación de A^{-1}. Consideremos la sucesión $(X_k)_{k\geq 0}$ de matrices dada por la recurrencia:*

$$E_k = I - AX_k \ , \quad X_{k+1} = X_k(I + E_k + E_k^2) \ (k \geq 0) \ .$$

a) Probar que $E_{k+1} = E_k^3$ y que $X_k = A^{-1}(I - E_0^{3k})$ $(k \geq 0)$.
b) Dar condiciones necesarias y suficientes de convergencia de $(X_k)_{k\geq 0}$ hacia A^{-1}.
c) Si partimos de una matriz X_0 tal que

$$E_0 = \begin{pmatrix} 0.1 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.1 \\ 0.3 & 0.1 & 0.2 \end{pmatrix} \ ,$$

¿cuántas iteraciones han de realizarse para que $\|E_k\|_\infty < 10^{-12}$?

SOLUCIÓN:

a) Tenemos

$$\begin{aligned} E_{k+1} &= I - AX_{k+1} = I - AX_k(I + E_k + E_k^2) \\ &= I - (I - E_k)(I + E_k + E_k^2) \\ &= I - (I - E_k + E_k - E_k^2 + E_k^2 - E_k^3) = E_k^3 \end{aligned}$$

i, por inducción,

$$E_k = E_0^{3^k} \ (k \geq 0) \ .$$

De la relación $E_k = I - AX_k$, tenemos

$$X_k = A^{-1}(I - E_k) = A^{-1}(I - E_0^{3^k}) \ (k \geq 0) \ .$$

b) La convergencia de $(X_k)_{k\geq 0}$ hacia A^{-1} equivale a la de $(E_0^{3^k})_{k\geq 0}$ hacia la matriz 0, que se da si y sólo si

$$\rho(E_0) = \rho(I - AX_0) < 1 \ .$$

c) Trabajando con la norma del máximo, tenemos que $\|E_0\|_\infty = 0.6$ y que $\|E_k\|_\infty \leq \|E_0\|_\infty^{3^k} = 0.6^{3^k}$. La condición $0.6^{3^k} < 10^{-12}$ equivale a

$$k > \frac{\log \frac{12}{-\log 0.6}}{\log 3} \simeq 3.63 \ .$$

Por lo tanto, son suficientes cuatro iteraciones.

Problema 2.10 *Sea B una* matriz de Frobenius*; esto es, una matriz de la forma*

$$B = \begin{pmatrix} b_1 & b_2 & b_3 & \cdots & \cdots & b_{n-1} & b_n \\ 1 & 0 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \cdots & 1 & 0 \end{pmatrix} .$$

a) Obtener el polinomio característico de B.

b) Si λ es un valor propio de B, encontrar un vector propio asociado a λ.

c) Sea $A = A_1 = (a_{ij}^{(1)})$ una matriz nxn cualquiera. Consideremos las sucesivas transformaciones de semejanza

$$A_{l+1} = M_{n-l}^{-1} A_l M_{n-l} \ \ (l = 1 \div n-1) \ ,$$

siendo

$$M_k = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ m_{k1} & \cdots & \cdots & m_{kk} & \cdots & \cdots & m_{kn} \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} ,$$

con

$$m_{kj} = -\frac{a_{k+1,j}^{(n-k)}}{a_{k+1,k}^{(n-k)}} \ (j \neq k) \ , \quad m_{kk} = \frac{1}{a_{k+1,k}^{(n-k)}} \ ,$$

suponiendo que $a_{k+1,k}^{(n-k)} \neq 0 \ \ (k = n-1 \div 1)$.

Demostrar que $A^{(n)}$ *es una matriz de Frobenius. Se la llama* forma normal de Frobenius de A. *La combinación de los apartados a), b) y c) ofrece un método de cálculo de valores y vectores propios llamado* método de Danilevski.

d) Aplicar este método de reducción a la forma normal de Frobenius para determinar los valores y vectores propios de la matriz

$$A = \begin{pmatrix} 39 & -16 & -56 \\ 87 & -36 & -126 \\ -2 & 1 & 3 \end{pmatrix} .$$

SOLUCIÓN:

a) El polinomio característico se obtiene desarrollando por la primera fila el determinante

$$\det(B - \lambda I) = \begin{vmatrix} b_1 - \lambda & b_2 & b_3 & \cdots & \cdots & b_{n-1} & b_n \\ 1 & -\lambda & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 1 & -\lambda & \cdots & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \cdots & 1 & -\lambda \end{vmatrix}$$

$$= (b_1 - \lambda) \begin{vmatrix} -\lambda & 0 & \cdots & \cdots & 0 & 0 \\ 1 & -\lambda & \cdots & \cdots & 0 & 0 \\ 0 & 1 & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & -\lambda \end{vmatrix}$$

$$-b_2 \begin{vmatrix} 1 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & -\lambda & \cdots & \cdots & 0 & 0 \\ 0 & 1 & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & -\lambda \end{vmatrix} + \cdots$$

$$
\begin{aligned}
& +(-1)^{n+1}b_n \begin{vmatrix} 1 & -\lambda & \cdots & \cdots & 0 & 0 \\ 0 & 1 & -\lambda & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & 0 & 1 \end{vmatrix} \\
= \; & (b_1 - \lambda)(-\lambda)^{n-1} - b_2(-\lambda)^{n-2} + \cdots + (-1)^{n+1}b_n \\
= \; & (-1)^n(\lambda^n - b_1\lambda^{n-1} - b_2\lambda^{n-2} - \cdots - b_n) \; .
\end{aligned}
$$

Así , el polinomio característico de B será

$$p_B(\lambda) = (-1)^n(\lambda^n - b_1\lambda^{n-1} - b_2\lambda^{n-2} - \cdots - b_n) \; .$$

La expresión anterior del polinomio característico puede obtenerse también por inducción, de forma análoga a la que se usa en el problema 2.11.

b) Si λ es un valor propio de B, ha de cumplir la ecuación característica

$$\lambda^n - b_1\lambda^{n-1} - b_2\lambda^{n-2} - \cdots - b_n = 0 \; .$$

Los vectores propios $y = (y_1 \; y_2 \; \cdots \; y_n)^\top$ de B, asociados a λ, cumplen el sistema lineal $By = \lambda y$ que escrito en componentes queda

$$
\begin{aligned}
b_1y_1 + b_2y_2 + b_3y_3 + \cdots + b_{n-1}y_{n-1} + b_ny_n &= \lambda y_1 \\
y_{i-1} &= \lambda y_i \quad (i = 2 \div n) \; .
\end{aligned}
$$

Las $n-1$ últimas ecuaciones permiten expresar todas las componentes en función de la última y_n:

$$y_{n-i} = \lambda^i y_n \; (i = 1 \div n-1) \; ;$$

la primera ecuación se escribe entonces como

$$(\lambda^n - b_1\lambda^{n-1} - b_2\lambda^{n-2} - \cdots - b_n)y_n = 0 \; ,$$

y se satisface para cualquier y_n, dado que λ es precisamente una raíz de la ecuación característica.

Así, los valores propios y de B asociados a λ se escriben en la forma

$$y = y_n \begin{pmatrix} \lambda^{n-1} \\ \lambda^{n-2} \\ \vdots \\ \lambda \\ 1 \end{pmatrix} \; ;$$

escogiendo $y_n = 1$, obtenemos como vector propio asociado a λ

$$y = \begin{pmatrix} \lambda^{n-1} \\ \lambda^{n-2} \\ \vdots \\ \lambda \\ 1 \end{pmatrix} \; .$$

c) Para llevar a cabo la primera iteración $A_2 = M_{n-1}^{-1} A_1 M_{n-1}$, notemos que la inversa de

$$M_{n-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ -\frac{a_{n1}^{(1)}}{a_{n,n-1}^{(1)}} & \cdots & \cdots & \frac{1}{a_{n,n-1}^{(1)}} & -\frac{a_{nn}^{(1)}}{a_{n,n-1}^{(1)}} \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

vine dada por

$$M_{n-1}^{-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ a_{n1}^{(1)} & \cdots & \cdots & a_{n,n-1}^{(1)} & a_{nn}^{(1)} \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix};$$

entonces,

$$\begin{aligned} A_2 &= M_{n-1}^{-1} A_1 M_{n-1} \\ &= \begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \cdots & \cdots & a_{1,n-1}^{(2)} & a_{1n}^{(2)} \\ \vdots & \vdots & & & \vdots & \vdots \\ a_{n-1,1}^{(2)} & a_{n-1,2}^{(2)} & \cdots & \cdots & a_{n-1,n-1}^{(2)} & a_{n-1,n}^{(2)} \\ 0 & 0 & \cdots & \cdots & 1 & 0 \end{pmatrix}. \end{aligned}$$

Usando el hecho de que la matriz M_k tiene por inversa

$$M_k^{-1} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ a_{k+1,1}^{(n-k)} & \cdots & \cdots & a_{k+1,k}^{(n-k)} & \cdots & \cdots & a_{k+1,n}^{(n-k)} \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix},$$

y sabiendo que A_l tiene las $l-1$ últimas filas de las $l-1$ últimas columnas en forma de matriz de Frobenius, se deduce que

$$A_{l+1} = M_{n-l}^{-1} A_l M_{n-l}$$

tiene las l últimas filas de las l últimas columnas en forma de matriz de Frobenius ($l = 1 \div n-1$).

Así, con $n-1$ pasos se llega a la forma normal de Frobenius de la matriz A

$$B \equiv A_n = \begin{pmatrix} a_{11}^{(n)} & a_{12}^{(n)} & a_{13}^{(n)} & \cdots & a_{1,n-1}^{(n)} & a_{1n}^{(n)} \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

d) Usando el método anterior sobre

$$A_1 \equiv A = \begin{pmatrix} 39 & -16 & -56 \\ 87 & -36 & -126 \\ -2 & 1 & 3 \end{pmatrix} ,$$

encontramos recursivamente:

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & -3 \\ 0 & 0 & 1 \end{pmatrix} , \quad M_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix} ,$$

$$A_2 = M_2^{-1} A_1 M_2 = \begin{pmatrix} 7 & -16 & -8 \\ 1 & -1 & -2 \\ 0 & 1 & 0 \end{pmatrix} ;$$

$$M_1 = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} , \quad M_1^{-1} = \begin{pmatrix} 1 & -1 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} ,$$

$$B = A_3 = M_1^{-1} A_2 M_1 = \begin{pmatrix} 6 & -11 & 6 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} .$$

El polinomio característico de B es pues

$$p_B(\lambda) = -\lambda^3 + 6\lambda^2 - 11\lambda + 6 .$$

Los valores propios de B, y también de A, son:

$$\lambda_1 = 1 , \quad \lambda_2 = 2 , \quad \lambda_3 = 3 .$$

Usando el apartado b), podemos calcular una base de vectores propios de B,

$$y^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} , \quad y^{(2)} = \begin{pmatrix} 4 \\ 2 \\ 1 \end{pmatrix} , \quad y^{(3)} = \begin{pmatrix} 9 \\ 3 \\ 1 \end{pmatrix} .$$

Como B se halla a partir de A aplicando dos transformaciones de semejanza con las matrices M_2 y M_1, podremos determinar una base de vectores propios de A mediante

$$v^{(j)} = M_2 M_1 y^{(j)} \ (j = 1 \div 3) .$$

Resultan

$$v^{(1)} = \begin{pmatrix} 4 \\ 6 \\ 1 \end{pmatrix} , \quad v^{(2)} = \begin{pmatrix} 8 \\ 15 \\ 1 \end{pmatrix} , \quad v^{(3)} = \begin{pmatrix} 14 \\ 28 \\ 1 \end{pmatrix} .$$

Problema 2.11 *Sea B_n una matriz de la forma*

$$\begin{pmatrix} b_1 & b_2 & b_3 & \cdots & \cdots & b_{n-1} & b_n \\ a & 0 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & a & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 0 & a & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \cdots & a & 0 \end{pmatrix} ,$$

con $b_j = \frac{K}{2^j}$ $(j = 1 \div n)$.

a) Consideremos el caso en que a y K son reales no negativos. Responder a las siguientes preguntas:

i) si n es par, ¿pueden ser todos los valores propios de B_n reales positivos?

ii) ¿pueden ser negativas todas las partes reales de los valores propios?

b) Encontrar la ecuación característica y los valores propios de B_n para $a = 2$ y $K = -2$.

c) Consideremos un método iterativo general del tipo

$$x^{(k+1)} = B_n x^{(k)} + c$$

para c cualquiera. Estudiar su convergencia suponiendo $|a| < 1$ y $|K| \leq 1$.

SOLUCIÓN:

a) i) Si n es par, no pueden ser todos los valores propios de B_n reales positivos; si fuese así, el producto de todos ellos seria positivo, pero

$$\prod_{j=1}^{n} \lambda_j = \det\, B_n = (-1)^{n-1} b_n a^{n-1} = (-1)^{n-1} \frac{K}{2^n} a^{n-1} \leq 0 \ ,$$

si $a \geq 0$, $K \geq 0$ y n par.

Las partes reales de los valores propios no pueden ser todas negativas; en tal caso, su suma sería también negativa, pero, por otro lado,

$$\sum_{j=1}^{n} \Re e(\lambda_j) = \sum_{j=1}^{n} \lambda_j = \text{tr}\ B_n = b_1 = \frac{K}{2} \geq 0 \ ,$$

si $K \geq 0$.

La primera igualdad se basa en que los valores propios son los ceros de un polinomio con coeficientes reales; los ceros no reales de este polinomio han de ser conjugados dos a dos de manera que su suma será siempre real.

b) Tenemos que calcular $\Delta_n = \det\,(B_n - \lambda I)$ para

$$b_j = -\frac{2}{2^j} = -\frac{1}{2^{j-1}} \quad \text{y} \quad a = 2 \ .$$

Hagámoslo primeramente para $n = 2$ y para $n = 3$:

$$\Delta_2 = \begin{vmatrix} b_1 - \lambda & b_2 \\ a & -\lambda \end{vmatrix} = \lambda^2 + \lambda + 1 \ ,$$

$$\Delta_3 = \begin{vmatrix} b_1 - \lambda & b_2 & b_3 \\ a & -\lambda & 0 \\ 0 & a & -\lambda \end{vmatrix} = -(\lambda^3 + \lambda^2 + \lambda + 1) \ .$$

Esto nos induce a pensar que

$$\Delta_n = (-1)^n(\lambda^n + \lambda^{n-1} + \cdots + \lambda + 1) \ .$$

Lo demostraremos por inducción. Supongamos que

$$\Delta_{n-1} = (-1)^{n-1}(\lambda^{n-1} + \lambda^{n-2} + \cdots + \lambda + 1) \ ;$$

desarrollando B_n por la última columna, tenemos

$$\Delta_n = -\lambda\Delta_{n-1} + (-1)^{n-1}b_n a^{n-1} \ .$$

Así,

$$\begin{aligned} \Delta_n &= -\lambda(-1)^{n-1}(\lambda^{n-1} + \lambda^{n-2} + \cdots + \lambda + 1) \\ &\quad +(-1)^{n-1}(-\frac{1}{2^{n-1}})2^{n-1} \\ &= (-1)^n(\lambda^n + \lambda^{n-1} + \cdots + \lambda^2 + \lambda) + (-1)^n \\ &= (-1)^n(\lambda^n + \lambda^{n-1} + \cdots + \lambda + 1) \ , \end{aligned}$$

que es lo que se quería demostrar.

Los valores propios de B_n son aquellos valores de λ que anulan

$$\Delta_n(\lambda) = (-1)^n(\lambda^n + \lambda^{n-1} + \cdots + \lambda + 1) = (-1)^n\frac{\lambda^{n+1} - 1}{\lambda - 1} \quad (\lambda \neq 1) \ ;$$

esto es,

$$\lambda^{n+1} = 1 \ , \quad \lambda \neq 1 \ .$$

Los valores propios son pues las raíces $(n+1)$-ésimas de la unidad, exceptuando $\lambda = 1$.

c) Una condición necesaria y suficiente de convergencia es que el radio espectral de B_n sea menor que 1.

Aplicando el teorema de Gerschgorin, sabemos que los valores propios de B_n están en la unión de los discos:

$$\{\lambda : |\lambda - b_1| \leq \sum_{j=2}^{n} |b_j|\} \ , \quad \{\lambda : |\lambda - 0| \leq |a|\} \ .$$

Para los valores de λ en el primer disco, tenemos

$$\left|\lambda - \frac{K}{2}\right| \leq \sum_{j=2}^{n} \frac{|K|}{2^j} = |K|\,(\frac{1}{2} - \frac{1}{2^n}) < \frac{|K|}{2} \ ;$$

así,

$$|\lambda| \leq \left|\lambda - \frac{K}{2}\right| + \left|\frac{K}{2}\right| < |K| \leq 1 \ ,$$

ya que se supone $|K| \leq 1$.

Para los valores de λ en el segundo disco:

$$|\lambda| \leq |a| < 1 \ ,$$

suponiendo $|a| < 1$.

Así pues, ambos discos están contenidos en el disco unidad, centrado en el origen y, por lo tanto, el radio espectral de B_n es menor que 1; consiguientemente, el método iterativo propuesto es siempre convergente.

Problema 2.12 *Se sabe que el vector* $(1 \;\; 1 \;\; 1 \;\; 1)^\top$ *es un vector propio de valor propio 32 de la matriz*

$$A = \begin{pmatrix} 19 & 11 & 1 & 1 \\ 3 & 3 & 13 & 13 \\ 5 & 5 & 15 & 7 \\ 9 & 9 & 7 & 7 \end{pmatrix} .$$

Utilizar la deflación de Householder para determinar los otros valores propios y una base de vectores propios de A.

SOLUCIÓN:

Denotando $v = (1 \;\; 1 \;\; 1 \;\; 1)^\top$, tenemos que escoger s tal que $s^2 = \|v\|_2^2 = 4$ con el fin de formar el vector $u = v - se_1$ tal que $P(u)v = se_1$. Tomamos, por ejemplo, $s = 2$ (normalmente, se escoge s de signo contrario al de v_1, pero en este caso no tenemos problemas graves de cancelación).

Así,

$$u = v - se_1 = (-1 \;\; 1 \;\; 1 \;\; 1)^\top \;, \quad \alpha = \frac{2}{v^\top v} = \frac{-1}{s_1 u_1} = \frac{1}{2} \;.$$

La matriz de Householder $P = P(u) = I - \alpha uu^\top$ satisface $Pv = 2e_1$, y con ella procedemos al cálculo de $A' = PAP$.

Primeramente calculamos

$$B = PA = (I - \alpha uu^\top)A = A - u(\alpha u)^\top A = A - uw^\top A = A - up^\top \;,$$

donde

$$w = \alpha u = \frac{1}{2}(-1 \;\; 1 \;\; 1 \;\; 1)^\top \;, \quad p^\top = w^\top A = (-1 \;\; 3 \;\; 17 \;\; 13) \;;$$

entonces,

$$B = A - \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} -1 & 3 & 17 & 13 \end{pmatrix} = \begin{pmatrix} 18 & 14 & 18 & 14 \\ 4 & 0 & -4 & 0 \\ 6 & 2 & -2 & -6 \\ 10 & 6 & -10 & -6 \end{pmatrix} .$$

Finalmente, procedemos al cálculo de la matriz transformada

$$A' = PAP = BP = B(I - \alpha uu^\top) = B - (Bw)u^\top = B - qu^\top \;,$$

donde $q = Bw = (14 \quad -4 \quad -6 \quad -10)^\top$, resultando

$$A' = B - \begin{pmatrix} 14 \\ -4 \\ -6 \\ -10 \end{pmatrix} \begin{pmatrix} -1 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 32 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \\ 0 & 8 & 4 & 0 \\ 0 & 16 & 0 & 4 \end{pmatrix} .$$

La matriz A' tiene los mismos valores propios que A; uno de los valores propios es $\lambda_1 = 32$ y los otros valores propios son los de la matriz reducida

$$\tilde{A} = \begin{pmatrix} 4 & 0 & 4 \\ 8 & 4 & 0 \\ 16 & 0 & 4 \end{pmatrix} ,$$

que encontramos calculando su polinomio característico

$$\det(\tilde{A} - \lambda I) = \begin{vmatrix} 4-\lambda & 0 & 4 \\ 8 & 4-\lambda & 0 \\ 16 & 0 & 4-\lambda \end{vmatrix} = (4-\lambda)(12-\lambda)(-4-\lambda) .$$

Sus ceros

$$\lambda_2 = 12 , \quad \lambda_3 = 4 , \quad \lambda_4 = -4 ,$$

son entonces los valores propios de $\tilde{A}$.

Nótese que, si $\tilde{v}$ es un vector propio de $\tilde{A}$ de valor propio λ, entonces

$$v' = \begin{pmatrix} v'_1 \\ \\ \tilde{v} \\ \\ \end{pmatrix} , \quad \text{con} \quad v_1 = \frac{c^\top \tilde{v}}{\lambda - \lambda_1} ,$$

es valor propio de

$$A' = \left(\begin{array}{c|c} \lambda_1 & c^\top \\ \hline 0 & \\ & \tilde{A} \\ & \end{array}\right) .$$

Una base de vectores propios de $\tilde{A}$, asociados a λ_2, λ_3 y λ_4, se encuentra directamente:

$$\tilde{v}^{(2)} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} , \quad \tilde{v}^{(3)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} , \quad \tilde{v}^{(4)} = \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix} .$$

Calculando primero

$$v_1'^{(2)} = \frac{4 \cdot 1}{12 - 32} = -\frac{1}{5} , \; v_1'^{(3)} = \frac{4 \cdot 1}{4 - 32} = -\frac{1}{7} , \; v_1'^{(4)} = \frac{4 \cdot 1}{-4 - 32} = -\frac{1}{9} ;$$

tenemos que

$$v'^{(2)} = \begin{pmatrix} -\frac{1}{5} \\ 1 \\ 1 \\ 2 \end{pmatrix} , \quad v'^{(3)} = \begin{pmatrix} -\frac{1}{7} \\ 0 \\ 1 \\ 0 \end{pmatrix} , \quad v'^{(4)} = \begin{pmatrix} -\frac{1}{9} \\ -1 \\ 1 \\ 2 \end{pmatrix} ,$$

son vectores propios de $A' = PAP$ de valores propios asociados λ_2 , λ_3 y λ_4, respectivamente.

Aplicándoles la matriz P encontraremos los vectores propios de A:

$$v^{(2)} = \frac{1}{10}\begin{pmatrix} 19 \\ -11 \\ -11 \\ -1 \end{pmatrix} \ , \quad v^{(3)} = \frac{1}{14}\begin{pmatrix} 6 \\ -8 \\ 6 \\ -8 \end{pmatrix} \ , \quad v^{(4)} = \frac{1}{18}\begin{pmatrix} 17 \\ -37 \\ -1 \\ 17 \end{pmatrix} \ .$$

Una base de vectores propios de A será pues la formada por los vectores:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \ , \quad \begin{pmatrix} 19 \\ -11 \\ -11 \\ -1 \end{pmatrix} \ , \quad \begin{pmatrix} 3 \\ -4 \\ 3 \\ -4 \end{pmatrix} \ , \quad \begin{pmatrix} 17 \\ -37 \\ -1 \\ 17 \end{pmatrix} \ .$$

Problema 2.13 *Determinar todos los valores propios de la matriz*

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 4 \end{pmatrix} \ ,$$

usando los métodos de la potencia y de la potencia inversa.

SOLUCIÓN:

Construimos primeramente la sucesión de vectores definidos por

$$x^{(0)} = (1 \ \ 1 \ \ 1 \ \ 1)^\top \ , \quad x^{(k+1)} = Ax^{(k)} \ (k \geq 0) \ ,$$

y la de vectores de cocientes $q^{(k)}$ de componentes

$$q_i^{(k)} = \frac{x_i^{(k)}}{x_i^{(k-1)}} \ (i = 1 \div n) \ (k \geq 1) \ .$$

Los resultados se exponen en la tabla siguiente:

k	1	2	3	4	5	6	7	8	9
	4	22	126	728	4218	24464	141946	823740	4780650
	5	27	153	881	5099	29563	171509	995249	5775899
$x^{(k)}$	6	34	194	1116	6450	37364	216674	1257088	7294826
	7	43	255	1493	8697	50555	293611	1704573	9894369
	4	5.5	5.73	5.778	5.7999	5.8022	5.80319	5.80359	5.80376
	5	5.4	5.67	5.758	5.7978	5.8014	5.80289	5.80347	5.80371
$q^{(k)}$	6	5.7	5.70	5.753	5.7929	5.7999	5.80175	5.80296	5.80348
	7	6.2	5.93	5.855	5.8129	5.8077	5.80555	5.80460	5.80420

Observamos la convergencia lineal de las componentes de los vectores de cocientes al valor propio de módulo máximo $\lambda_1 \simeq 5.804$.

El uso de cocientes de Rayleigh conduce a una sucesión más rápidamente convergente

$$\begin{aligned} \sigma_k &= \frac{x^{(k+1)\top} x^{(k)}}{x^{(k)\top} x^{(k)}} \\ &= 5.5,\ 5.77,\ 5.7999,\ 5.80319,\ 5.80376,\ 5.803863,\ 5.8038820, \\ &\quad 5.80388554,\ 5.803886331,\ \ldots \quad (k \geq 0) \end{aligned}$$

al valor propio $\lambda_1 \simeq 5.8038863$, que podemos dar con 7 cifras decimales correctas.

La matriz inversa de A es

$$A^{-1} = \frac{1}{6} \begin{pmatrix} 17 & -6 & -3 & -2 \\ -6 & 6 & 0 & 0 \\ -3 & 0 & 3 & 0 \\ -2 & 0 & 0 & 2 \end{pmatrix} ;$$

repitiendo el proceso con $B = 6A^{-1}$, para llevar a cabo las iteraciones con enteros, y partiendo ahora de $x^{(0)} = (1\ 0\ 0\ 0)^\top$, obtenemos

k	1	2	3	4	5	6	7
$x^{(k)}$	17	338	6830	138332	2802884	56796776	1150932152
	−6	−138	−2856	−58116	−1178688	−23889432	−484117248
	3	−60	−1194	−24072	−487212	−9870288	−200001192
	2	−38	−752	−15164	−306992	−6219752	−126033056
$q^{(k)}$		19.	20.21	20.254	20.262	20.26404	20.264114
		23.	20.70	20.349	20.282	20.26491	20.264298
		19.	20.21	20.254	20.262	20.26404	20.263879
		19.	20.21	20.254	20.262	20.26404	20.263972

La convergencia al valor propio dominante μ_1 de B es más rápida, si usamos los cocientes de Rayleigh; notemos que

$$\sigma_5 = 20.26413315\ , \quad \sigma_6 = 20.26413316\ .$$

El valor propio menor en módulo de A será, entonces,

$$\lambda_4 = \frac{6}{\mu_1} \simeq \frac{6}{\sigma_6} = 0.296090\ .$$

Los otros valores propios λ_2 y λ_3 se calculan ahora fácilmente, usando que tr A y det A son respectivamente la suma y el producto de los valores propios:

$$\begin{aligned} \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = \text{tr } A &= 10\ , \\ \lambda_1\lambda_2\lambda_3\lambda_4 = \det A &= 6\ . \end{aligned}$$

O sea,

$$\begin{aligned} \lambda_2 + \lambda_3 &= 10 - \lambda_1 - \lambda_4 = 3.990024\ , \\ \lambda_2\lambda_3 &= \frac{6}{\lambda_1\lambda_4} = 3.491473\ ; \end{aligned}$$

de donde se deduce que los dos valores propios restantes de A son

$$\lambda_2 = 2.507752 \ , \quad \lambda_3 = 1.392272 \ .$$

Problema 2.14 *Aplicar el método de Jacobi para determinar los valores y vectores propios de la matriz simétrica*

$$A = \begin{pmatrix} 4 & 2 & 1 \\ 2 & 5 & 3 \\ 1 & 3 & 6 \end{pmatrix} \ ,$$

hasta que los elementos no diagonales sean todos menores que 10^{-6}.

SOLUCIÓN:

Utilizaremos el método clásico de Jacobi. El elemento maximal no diagonal de $B \equiv A_1 \equiv A$ es $b_{23} = 3$; hay que aplicar primero una rotación R_1, en el plano coordenado 2-3, que anule al elemento $\bar{b}_{23}$ de $\bar{B} = R_1^\top B R_1$. El coseno c_1 y el seno s_1 del ángulo de la rotación se calculan mediante:

$$d_1 = \sqrt{(b_{22} - b_{33})^2 + 4b_{23}^2} = \sqrt{37} \ ,$$

$$c_1 = \sqrt{\frac{1}{2}(1 + \frac{1}{d_1})} = 0.7630199825 \ ,$$

$$s_1 = -\sqrt{\frac{1}{2}(1 - \frac{1}{d_1})} = -0.6463748961 \ ;$$

y los elementos de B se transforman según

$$\begin{aligned} \bar{b}_{12} &= b_{12}c_1 + b_{13}s_1 = 0.8796650689 \ , \\ \bar{b}_{13} &= -b_{12}s_1 + b_{13}c_1 = 2.055769775 \ , \\ \bar{b}_{22} &= b_{22}c_1^2 + 2b_{23}s_1c_1 + b_{33}s_1^2 = b_{22} + b_{23}t_1 = 2.458618735 \ , \\ \bar{b}_{33} &= b_{22}s_1^2 - 2b_{23}s_1c_1 + b_{33}c_1^2 = -b_{23}t_1 + b_{33} = 8.541381266 \ , \end{aligned}$$

donde se ha usado $t_1 = \frac{s_1}{c_1}$. Recordemos que $\bar{b}_{11} = b_{11}$ y que $\bar{b}_{23} = 0$, por construcción.

Tenemos así,

$$A_2 \equiv \bar{B} = \begin{pmatrix} 4 & 0.879665069 & 2.055769775 \\ 0.879665069 & 2.458618735 & 0 \\ 2.055769775 & 0 & 8.541381266 \end{pmatrix} \ .$$

Ahora, el elemento no diagonal maximal de $\bar{B} \equiv A_2$ es $\bar{b}_{13}$; la rotación ha de efectuarse, pues, en el plano 1-3. Los senos y cosenos se calculan de forma análoga, siendo

$$c_2 = 0.9330913253 \ , \quad s_2 = -0.3596395121 \ ;$$

la matriz resultante es

$$A_3 = R_2^\top A_2 R_2 = \begin{pmatrix} 3.207648792 & 0.820807845 & 0 \\ 0.820807845 & 2.458618735 & 0.316362316 \\ 0 & 0.316362316 & 9.333732474 \end{pmatrix} .$$

Repitiendo el proceso, se obtienen los siguientes senos y cosenos de los ángulos de las rotaciones y matrices transformadas:

$$c_3 = 0.8411621222 \ , \quad s_3 = 0.540783029 \ ;$$

$$A_4 = R_3^\top A_3 R_3 = \begin{pmatrix} 3.735346059 & 0 & 0.171083372 \\ 0 & 1.930921469 & 0.266111997 \\ 0.171083372 & 0.266111997 & 9.333732474 \end{pmatrix} .$$

$$c_4 = 0.999356178 \ , \quad s_4 = 0.035877979 \ ;$$

$$A_5 = R_4^\top A_4 R_4 = \begin{pmatrix} 3.735346059 & -0.00613812 & 0.170973224 \\ -0.00613813 & 1.921367757 & 0 \\ 0.170973224 & 0 & 9.343286185 \end{pmatrix} .$$

$$c_5 = 0.999536434 \ , \quad s_5 = -0.0304452993 \ ;$$

$$A_6 = R_5^\top A_5 R_5 = \begin{pmatrix} 3.730138314 & -0.00613528 & 0 \\ -0.00613528 & 1.921367757 & -0.000186877 \\ 0 & -0.00018688 & 9.343286185 \end{pmatrix} .$$

$$c_6 = 0.999994248 \ , \quad s_6 = -0.003391903 \ ;$$

$$A_7 = R_6^\top A_6 R_6 = \begin{pmatrix} 3.730159124 & 0 & 0.000000634 \\ 0 & 1.921346947 & -0.000186877 \\ 0.000000634 & -0.00018688 & 9.348493935 \end{pmatrix} .$$

$$c_7 = 0.999999997 \ , \quad s_7 = 0.0000251595 \ ;$$

$$A_8 = R_7^\top A_7 R_7 = \begin{pmatrix} 3.730159124 & 0.000000000 & 0.000000634 \\ 0.000000000 & 1.921346942 & 0 \\ 0.000000634 & 0 & 9.348493935 \end{pmatrix} .$$

Obsérvese que las 9 primeras cifras significativas de los elementos de la diagonal no han variado en la última iteración del método; damos así las siguientes aproximaciones de los valores propios:

$$\lambda_1 \simeq 3.73015912 \ , \quad \lambda_2 \simeq 1.92134694 \ , \quad \lambda_3 \simeq 9.34849394 \ .$$

Una aproximación de los vectores propios ortonormales asociados a λ_1, λ_2 y λ_3 viene dada por las columnas de la matriz ortogonal O_8 que es producto de todas las rotaciones hechas

$$O_8 = R_1 R_2 \cdots R_7 \ ,$$

dado que

$$O_8^\top A O_8 = A_8 \simeq \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3) \ .$$

Para calcular O_3 a partir de $\Omega = O_2 = R_1$, se tiene que efectuar el producto de matrices

$$\bar{\Omega} = O_3 = R_2 O_2 = R_2 \Omega \ ,$$

en el cual sólo se modifican las filas 1 y 3 de la matriz Ω según

$$\bar{\omega}_{1j} = \omega_{1j} c_2 + \omega_{3j} s_2 \ , \quad \bar{\omega}_{3j} = -\omega_{1j} s_2 + \omega_{3j} c_2 \quad (j = 1, 2, 3) \ .$$

Resulta así,

$$O_3 = \begin{pmatrix} 0.933091325 & 0 & 0.359639512 \\ -0.232461952 & 0.763019983 & 0.603126808 \\ -0.274412134 & 0.646374896 & 0.711967327 \end{pmatrix} \ .$$

Repitiendo el proceso, se obtiene finalmente la matriz formada por los vectores propios de A

$$O_8 = \begin{pmatrix} 0.774686009 & -0.488736998 & 0.401245229 \\ 0.196800996 & 0.789366503 & 0.581523767 \\ -0.600600794 & -0.517791163 & 0.713020526 \end{pmatrix} \ .$$

Problema 2.15 *Reducir la matriz*

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 4 \end{pmatrix}$$

a forma tridiagonal simétrica por el método de Householder y calcular después sus valores y vectores propios.

SOLUCIÓN:

Empecemos con el proceso de reducción a forma tridiagonal simétrica. Esto se conseguirá a través de dos transformaciones de semejanza por matrices de Householder sobre la matriz A.

Para ello, partimos de $A_2 = A$ y escogemos una matriz de Householder de la forma

$$P_2 = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & \tilde{P}_2 \end{array}\right) \ ,$$

con $\tilde{P}_2 = P(\tilde{u}^{(2)}) = I_3 - \alpha_2 \tilde{u}^{(2)} \tilde{u}^{(2)\top}$, donde I_3 indica la matriz identidad de dimensión 3 y el vector $\tilde{u}^{(2)}$, que tiene dimensión 3, es tal que la matriz $\tilde{P}_2$ transforma el vector $\tilde{a}^{(2)} = (1\ 1\ 1)^\top$ en un vector con la segunda y tercera componentes nulas,

$$\tilde{P}_2 \tilde{a}^{(2)} = s_2 \tilde{e}^{(1)} \equiv \begin{pmatrix} s_2 \\ 0 \\ 0 \end{pmatrix} ,$$

donde $s_2^2 = \|\tilde{a}^{(2)}\|_2^2$.

Los cálculos de s_2, α_2 y $\tilde{u}^{(2)}$ se realizan a continuación:

$$\begin{aligned} s_2 &= - \quad \|\tilde{a}^{(2)}\|_2 = -\sqrt{3} , \\ \tilde{u}^{(2)} &= \begin{pmatrix} 1+\sqrt{3} \\ 1 \\ 1 \end{pmatrix} , \\ \alpha_2 &= \frac{1}{\sqrt{3}(\sqrt{3}+1)} = \frac{3-\sqrt{3}}{6} , \end{aligned}$$

La matriz transformada $A_3 = P_2 A_2 P_2$ es entonces de la forma

$$A_3 = \left(\begin{array}{c|ccc} 1 & -\sqrt{3} & 0 & 0 \\ \hline -\sqrt{3} & & & \\ 0 & & \tilde{P}_2 \tilde{A}_2 \tilde{P}_2 & \\ 0 & & & \end{array} \right) .$$

Para el cálculo de $\tilde{P}_2 \tilde{A}_2 \tilde{P}_2$, se realizan los cálculos previos siguientes:

$$\begin{aligned} \tilde{w}^{(2)} &= \alpha_2 \tilde{u}^{(2)} = \frac{1}{3+\sqrt{3}} \begin{pmatrix} 1+\sqrt{3} \\ 1 \\ 1 \end{pmatrix} , \\ \tilde{p}^{(2)} &= \alpha_2 \tilde{A}_2 \tilde{w}^{(2)} = \frac{1}{3+\sqrt{3}} \begin{pmatrix} 4+2\sqrt{3} \\ 5+\sqrt{3} \\ 6+\sqrt{3} \end{pmatrix} , \\ \beta_2 &= \tilde{w}^{(2)\top} \tilde{p}^{(2)} = \frac{21+8\sqrt{3}}{(3+\sqrt{3})^2} , \\ \tilde{q}^{(2)} &= \tilde{p}^{(2)} - \frac{1}{2}\beta_2 \tilde{u}^{(2)} = \frac{1}{12} \begin{pmatrix} 9-9\sqrt{3} \\ 6+\sqrt{3} \\ 12-\sqrt{3} \end{pmatrix} , \\ \tilde{N}_2 &= \tilde{u}^{(2)} \tilde{q}^{(2)\top} = \frac{1}{12} \begin{pmatrix} -18 & 9+7\sqrt{3} & 9+11\sqrt{3} \\ 9-9\sqrt{3} & 6+\sqrt{3} & 12-\sqrt{3} \\ 9-9\sqrt{3} & 6+\sqrt{3} & 12-\sqrt{3} \end{pmatrix} , \\ \tilde{N}_2 + \tilde{N}_2^\top &= \begin{pmatrix} -3 & \frac{3}{2}-\frac{\sqrt{3}}{6} & \frac{3}{2}+\frac{\sqrt{3}}{6} \\ \frac{3}{2}-\frac{\sqrt{3}}{6} & 1+\frac{\sqrt{3}}{6} & \frac{3}{2} \\ \frac{3}{2}+\frac{\sqrt{3}}{6} & \frac{3}{2} & 2-\frac{\sqrt{3}}{6} \end{pmatrix} . \end{aligned}$$

La matriz $\tilde{P}_2\tilde{A}_2\tilde{P}_2$ es entonces igual a $\tilde{A}_2 - (\tilde{N}_2 + \tilde{N}_2^\top)$ y la matriz transformada es

$$A_3 = \begin{pmatrix} 1 & -\sqrt{3} & 0 & 0 \\ -\sqrt{3} & 5 & -\frac{1}{2}+\frac{\sqrt{3}}{6} & -\frac{1}{2}-\frac{\sqrt{3}}{6} \\ 0 & -\frac{1}{2}+\frac{\sqrt{3}}{6} & 2-\frac{\sqrt{3}}{6} & -\frac{1}{2} \\ 0 & -\frac{1}{2}-\frac{\sqrt{3}}{6} & -\frac{1}{2} & 2+\frac{\sqrt{3}}{6} \end{pmatrix},$$

que denotamos en la forma

$$\left(\begin{array}{cc|cc} & & 0 & 0 \\ & T_3 & & \\ & & \multicolumn{2}{c}{\tilde{a}^{(3)\top}} \\ \hline 0 & & & \\ & \tilde{a}^{(3)} & \multicolumn{2}{c}{\tilde{A}_3} \\ 0 & & & \end{array}\right).$$

Escogemos ahora, de manera análoga, una matriz de Householder del tipo

$$P_3 = \left(\begin{array}{c|c} I_2 & 0 \\ \hline & \\ 0 & \tilde{P}_3 \\ & \end{array}\right),$$

con $\tilde{P}_3 = I_2 - \alpha_3\tilde{u}^{(3)}\tilde{u}^{(3)\top}$, donde I_2 indica la matriz identidad de dimensión 2 y el vector $\tilde{u}^{(3)}$, que tiene dimensión 2, es tal que $\tilde{P}_3$ transforma el vector

$$\tilde{a}^{(3)} = \begin{pmatrix} -\frac{1}{2}+\frac{\sqrt{3}}{6} \\ -\frac{1}{2}-\frac{\sqrt{3}}{6} \end{pmatrix}$$

en un vector con la segunda componente nula,

$$\tilde{P}_3\tilde{a}^{(3)} = s_3\tilde{e}^{(2)} \equiv \begin{pmatrix} s_3 \\ 0 \end{pmatrix},$$

donde $s_3^2 = \|\tilde{a}^{(3)}\|_2^2$.

Los cálculos de s_3, α_3 y $\tilde{u}^{(3)}$ se realizan a continuación:

$$\begin{aligned} s_3 &= \|\tilde{a}^{(3)}\|_2 = \frac{\sqrt{6}}{3}, \\ \tilde{u}^{(3)} &= \begin{pmatrix} -\frac{1}{2}+\frac{\sqrt{3}}{6}-\frac{\sqrt{6}}{3} \\ -\frac{1}{2}-\frac{\sqrt{3}}{6} \end{pmatrix}, \\ \alpha_3 &= \frac{6}{4+\sqrt{6}-\sqrt{2}}. \end{aligned}$$

La matriz transformada $A_4 = P_3 A_3 P_3$ es entonces de la forma

$$A_4 = \left(\begin{array}{cc|cc} 1 & -\sqrt{3} & 0 & 0 \\ -\sqrt{3} & 5 & \frac{\sqrt{6}}{3} & 0 \\ \hline 0 & \frac{\sqrt{6}}{3} & & \\ 0 & 0 & \multicolumn{2}{c}{\tilde{P}_3 \tilde{A}_3 \tilde{P}_3} \end{array}\right) .$$

El cálculo de $\tilde{P}_3 \tilde{A}_3 \tilde{P}_3$, se realiza de manera similar y se obtiene finalmente la matriz tridiagonal simétrica

$$T \equiv A_4 = \left(\begin{array}{cccc} 1 & -\sqrt{3} & 0 & 0 \\ -\sqrt{3} & 5 & \frac{\sqrt{6}}{3} & 0 \\ 0 & \frac{\sqrt{6}}{3} & 2 & -\frac{\sqrt{3}}{3} \\ 0 & 0 & -\frac{\sqrt{3}}{3} & 2 \end{array}\right) .$$

El polinomio característico $p_4(\lambda)$ se calcula por recurrencia entre los determinantes $p_j(\lambda)$ $(j = 1 \div 4)$ de las matrices principales de $\lambda I - T$:

$$\begin{aligned} p_0(\lambda) &\equiv 1 , \\ p_1(\lambda) &= \lambda - 1 , \\ p_2(\lambda) &= (\lambda - 5) p_1(\lambda) - 3 p_0(\lambda) , \\ p_3(\lambda) &= (\lambda - 2) p_2(\lambda) - \frac{2}{3} p_1(\lambda) , \\ p_4(\lambda) &= (\lambda - 2) p_3(\lambda) - \frac{1}{3} p_1(\lambda) . \end{aligned}$$

Estos polinomios forman la sucesión de Sturm

$$\{p_4(\lambda) , \ p_3(\lambda) , \ p_2(\lambda) , \ p_1(\lambda) , \ p_0(\lambda)\}$$

a la cual aplicaremos el teorema de Sturm para separar los ceros de $p_4(\lambda)$. Estos son los valores propios de T y también los de A.

Los ceros buscados son todos reales, al ser la matriz simétrica, y por el teorema de Gerschgorin aplicado a la matriz original A, sabemos que se encuentran en el intervalo $[-2, 7]$.

Utilizaremos a continuación el teorema de Sturm para separar los valores propios. Calcularemos, para diversos valores de $\lambda \in [-2, 7]$, el número de variaciones de signo de la sucesión al que llamaremos $V(\lambda)$ y que indicará el número de valores propios de la matriz mayores que λ.

En la tabla siguiente se representan los signos de los valores de los polinomios $p_j(\lambda)$ $(j = 0 \div 4)$ y el número correspondiente de variaciones de signo de la sucesión, para diversos valores de λ,

λ	-2	-1	0	1	2	3	4	5	6
$p_0(\lambda)$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
$p_1(\lambda)$	$-$	$-$	$-$	0	$+$	$+$	$+$	$+$	$+$
$p_2(\lambda)$	$+$	$+$	$+$	$-$	$-$	$-$	$-$	$-$	$+$
$p_3(\lambda)$	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$-$	$+$
$p_4(\lambda)$	$+$	$+$	$+$	$-$	$+$	$-$	$-$	$-$	$+$
$V(\lambda)$	4	4	4	3	2	1	1	1	0

Así tenemos una raíz en $(0,1)$, una en $(1,2)$, una en $(2,3)$ y finalmente otra en $(5,6)$.

Efectuando ahora unas iteraciones de los métodos de bisección y de la secante, descritos en el capítulo 5, se obtienen los siguientes valores propios:

$$\lambda_1 = 5.803886359... \ , \quad \lambda_2 = 2.507748705... \ ,$$

$$\lambda_3 = 1.392275291... \ , \quad \lambda_4 = 0.296089645...$$

Con el fin de encontrar los vectores propios $v''^{(j)}$ $(j = 1 \div 4)$ de

$$T = \begin{pmatrix} a_1 & b_2 & & \\ b_2 & a_2 & b_3 & \\ & b_3 & a_3 & b_4 \\ & & b_4 & a_4 \end{pmatrix} \ ,$$

se resuelven los sistemas tridiagonales simétricos $(\lambda I - T)v'' = 0$:

$$\left.\begin{array}{rrrrr} (\lambda - a_1)v_1'' & -b_2 v_2'' & & & = 0 \\ -b_2 v_1'' & +(\lambda - a_2)v_2'' & -b_3 v_3'' & & = 0 \\ & -b_3 v_2'' & +(\lambda - a_3)v_3'' & -b_4 v_4'' & = 0 \\ & & -b_4 v_3'' & +(\lambda - a_4)v_4'' & = 0 \end{array}\right\} \ ,$$

para $\lambda = \lambda_j$ y $v'' = v''^{(j)}$ $(j = 1 \div 4)$.

Tomando $v_1'' = 1$, resulta

$$v_2'' = \frac{p_1(\lambda)}{b_2} \ , \quad v_3'' = \frac{p_2(\lambda)}{b_2 b_3} \ , \quad v_4'' = \frac{p_3(\lambda)}{b_2 b_3 b_4} \ .$$

Estas fórmulas permiten calcular los vectores propios de T asociados a los valores propios encontrados:

$$v''^{(1)} = \begin{pmatrix} 1.0000000000 \\ -2.7735250825 \\ -0.6093695727 \\ 0.0924895353 \end{pmatrix} \ , \quad v''^{(2)} = \begin{pmatrix} 1.0000000000 \\ -0.8704991207 \\ 4.7784074785 \\ -5.4334256556 \end{pmatrix} \ ,$$

$$v''^{(3)} = \begin{pmatrix} 1.0000000000 \\ -0.2264802449 \\ 3.1220328934 \\ 2.9659918380 \end{pmatrix} \ , \quad v''^{(4)} = \begin{pmatrix} 1.0000000000 \\ 0.4064028329 \\ -0.2200029869 \\ -0.0745454645 \end{pmatrix} \ .$$

Para obtener los vectores propios $v = v^{(j)}$ $(j = 1 \div 4)$ de A, hay que aplicar a los vectores propios de T las transformaciones de semejanza utilizadas:

$$v'^{(j)} = P_3 v''^{(j)} \ , \quad v^{(j)} = P_2 v'^{(j)} \quad (j = 1 \div 4) \ .$$

Nótese que la transformación con P_3 afecta sólo a las dos últimas componentes y la transformación con P_2, a las tres últimas. La aplicación de las matrices de Householder $P_i = I - 2\alpha_i u^{(i)} u^{(i)\top}$ $(i = 2, 3)$ a un vector v'' se calcula simplemente haciendo

$$v' = P_3 v'' = v'' - \alpha_3 (u^{(3)\top} v'') u^{(3)},$$

$$v = P_2 v' = v' - \alpha_2 (u^{(2)\top} v') u^{(2)} \ .$$

Los vectores propios de $T = A_4$ son así transformados por P_3 en los siguientes vectores propios de A_3:

$$v'^{(1)} = \begin{pmatrix} 1.0000000000 \\ -2.7735250825 \\ 0.0683784201 \\ 0.6125438612 \end{pmatrix} \ , \quad v'^{(2)} = \begin{pmatrix} 1.0000000000 \\ -0.8704991207 \\ 4.0115433053 \\ -6.0218612318 \end{pmatrix} \ ,$$

$$v'^{(3)} = \begin{pmatrix} 1.0000000000 \\ -0.2264802449 \\ -3.6729696891 \\ -2.2479970270 \end{pmatrix} \ , \quad v'^{(4)} = \begin{pmatrix} 1.0000000000 \\ 0.4064028329 \\ 0.1289463524 \\ 0.1932127810 \end{pmatrix} \ .$$

Finalmente, éstos son transformados por P_2 en los siguientes vectores propios de $A_2 = A$:

$$v^{(1)} = \begin{pmatrix} 1.0000000000 \\ 1.2081647906 \\ 1.5257780636 \\ 2.0699435048 \end{pmatrix} \ , \quad v^{(2)} = \begin{pmatrix} 1.0000000000 \\ 1.6632404977 \\ 4.9389563722 \\ -5.0944481649 \end{pmatrix} \ ,$$

$$v^{(3)} = \begin{pmatrix} 1.0000000000 \\ 3.5492301578 \\ -2.2909637644 \\ -0.8659911023 \end{pmatrix} \ , \quad v^{(4)} = \begin{pmatrix} 1.0000000000 \\ -0.4206354474 \\ -0.1737706681 \\ -0.1095042395 \end{pmatrix} \ .$$

Problema 2.16 *Obtener una matriz tridiagonal simétrica (no diagonal) nxn que tenga por valores propios los ceros del polinomio de Chebichev T_n en $[-1, 1]$.*

SOLUCIÓN:

El polinomio característico de una matriz tridiagonal

$$\begin{pmatrix} a_1 & b_2 & & & & & \\ b_2 & a_2 & b_3 & & & & \\ & b_3 & a_3 & b_4 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & b_{n-1} & a_{n-1} & b_n \\ & & & & & b_n & a_n \end{pmatrix}$$

se puede calcular recurrentemente. Tomando $p_j(x) = \det(xI - A_j)$, donde A_j es la matriz formada por las j primeras filas de las j primeras columnas de A, se tiene la recurrencia

$$p_0(x) = 1 \ , \quad p_1(x) = x - a_1 \ ;$$

$$p_j(x) = (x - a_j)p_{j-1}(x) - b_j^2 p_{j-2}(x) \ (j \geq 1) \ .$$

El polinomio característico de A se escribe entonces como

$$p_A(\lambda) = \det(A - \lambda I) = (-1)^n \det(\lambda I - A) = (-1)^n p_n(\lambda) \ .$$

No haremos ninguna distinción entre los polinomios $p_A(\lambda)$ y $p_n(\lambda)$ por lo que se refiere a buscar valores propios, y entenderemos que $p_n(\lambda)$ es el polinomio característico de A.

Fijémonos primero que los polinomios obtenidos por la recurrencia anterior són mónicos, ya que lo son p_0 y p_1. Si queremos que p_n tenga los mismos ceros que T_n, p_n ha de ser un múltiplo de T_n; en realidad, debe ser el polinomio mónico de Chebichev (véase el capítulo 3).

Los polinomios de Chebichev cumplen la recurrencia

$$T_0(x) = 1 \ , \quad T_1(x) = x \ ; \quad T_j(x) = 2xT_{j-1}(x) - T_{j-2}(x) \ (j \geq 1) \ .$$

Como el coeficiente principal del polinomio de grado j es 2^{j-1}, para $j \geq 1$, los polinomios mónicos de Chebichev serán

$$\tilde{T}_0(x) = T_0(x) \ , \quad \tilde{T}_j(x) = \frac{T_j(x)}{2^{j-1}} \ (j \geq 1) \ ;$$

despejando T_j y sustituyéndolo en la relación de recurrencia anterior, obtenemos la siguiente relación de recurrencia para estos polinomios mónicos:

$$\tilde{T}_0(x) = 1 \ , \quad \tilde{T}_1(x) = x \ , \quad \tilde{T}_2(x) = x\tilde{T}_1(x) - \frac{1}{2}\tilde{T}_0(x) \ ;$$

$$\tilde{T}_j(x) = x\tilde{T}_{j-1}(x) - \frac{1}{4}\tilde{T}_{j-2}(x) \ (j \geq 2) \ .$$

Imponiendo la igualdad entre los p_j $(j = 0 \div n)$ y los $\tilde{T}_j$ $(j = 0 \div n)$, sobre sus relaciones de recurrencia, tenemos que los coeficientes de la matriz tridiagonal han de cumplir

$$a_j = 0 \ (j = 1 \div n) \ ; \quad b_2^2 = \frac{1}{2} \ , \quad b_j^2 = \frac{1}{4} \ \ (j > 2) \ .$$

Escogiendo, por ejemplo, los coeficientes positivos, encontramos la matriz buscada

$$\begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & & & & & \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{2} & & & & \\ & \frac{1}{2} & 0 & \frac{1}{2} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \frac{1}{2} & 0 & \frac{1}{2} \\ & & & & & \frac{1}{2} & 0 \end{pmatrix} .$$

Problema 2.17 *a) Resolver el siguiente* problema de valores propios en ecuaciones diferenciales lineales*: calcular los valores de λ para los cuales existen funciones $y(x)$ no idénticamente nulas tales que cumplen la ecuación diferencial*

$$y"(x) + \lambda y(x) = 0 \quad \forall x \in (0,1) \ ,$$

y las condiciones de frontera $y(0) = y(1) = 0$.

Discretizamos el problema imponiendo la validez de la ecuación sólo en las $m+1$ abscisas equidistantes $x_k = kh$ $(k = 0 \div m)$, $h = \frac{1}{m}$, y aproximando después la derivada segunda usando diferencias centradas; resulta así un problema de valores propios en dimensión finita en las aproximaciones y_k de y en x_k:

$$y_0 = 0 \ , \quad \frac{1}{h^2}(y_{k+1} - 2y_k + y_{k-1}) + \lambda y_k = 0 \ (k = 1 \div m-1) \ , \quad y_m = 0 \ .$$

b) Resolver exactamente el problema de valores propios resultante.

c) Comparar las soluciones del problema discretizado con las soluciones exactas.

SOLUCIÓN:

a) Las soluciones analíticas (exactas) de la ecuación diferencial son de la forma

$$y(x) = A\cos(\sqrt{\lambda}x) + B\,\mathrm{sen}(\sqrt{\lambda}x) \ .$$

Imponiendo la primera condición de frontera $y(0) = 0$, resulta $A = 0$, e imponiendo la otra condición de frontera $y(1) = 0$, resulta $B = 0$ a menos que $\lambda = \lambda_j = j^2\pi^2$ $(j \geq 0)$.

Estos valores de λ son los valores propios del problema; las funciones propias del problema asociadas a un valor propio $\lambda_j = j^2\pi^2$ son múltiplos de la función $\mathrm{sen}\, j\pi x$.

b) El *problema discretizado* consiste en determinar los valores y vectores propios de la matriz tridiagonal simétrica

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{pmatrix} .$$

Se comprueba fácilmente que los valores propios de la matriz son

$$\tilde{\lambda_j} = \frac{2}{h^2}(1 - \cos\frac{j\pi}{m}) \quad (j = 1 \div m - 1) \ ,$$

y que las componentes de cada uno de los vectores propios $y^{(j)}$ asociados a $\tilde{\lambda_j}$ son

$$y_k^{(j)} = \operatorname{sen}\frac{jk\pi}{m} = \operatorname{sen} jx_k \quad (k = 1 \div m - 1) \ ;$$

los otros vectores propios asociados a $\tilde{\lambda}_j$ son múltiplos de este vector.

c) Nótese que los vectores propios del problema discretizado son las tablas de las funciones propias del problema de partida en las abscisas utilizadas para la discretización.

En el problema discretizado se tienen $m-1$ valores propios que tienden asintóticamente, cuando h tiende a cero o m tiende a infinito, a los infinitos valores propios del problema de partida:

$$\begin{aligned} \tilde{\lambda}_j &= \frac{2}{h^2}(1 - \cos j\pi h) = \frac{2}{h^2}(\frac{j^2\pi^2h^2}{2} - \frac{j^4\pi^4h^4}{24} + \cdots) \\ &= \lambda_j - \frac{j^2\pi^4}{12}h^2 + \cdots \quad (h = \frac{1}{m} \to 0) \end{aligned}$$

(consúltese la tabla 3.5 para el desarrollo de Taylor de la función $\cos x$ cerca de cero).

Tenemos, además, la acotación de error

$$\frac{\left|\tilde{\lambda}_j - \lambda_j\right|}{|\lambda_j|} \leq \frac{\pi^4}{12}\frac{j^2}{m^2} \quad (j = 1 \div m - 1) \ ,$$

que nos pone de manifiesto la poca precisión de los valores aproximados $\tilde{\lambda}_j$ comparados con los exactos λ_j, para los últimos valores del índice j.

PROBLEMAS PROPUESTOS

1. Consideremos el sistema triangular superior $U_n(a)x = e^{(n)}$, donde $e^{(n)}$ es el vector n-ésimo de la base canónica y

$$U_n(a) = (u_{ij}) = \begin{pmatrix} 1 & a & a & \cdots & a \\ & 1 & a & \cdots & a \\ & & 1 & \cdots & a \\ & & & \ddots & a \\ & & & & 1 \end{pmatrix}.$$

   a) Determinar la solución x.

   b) Acotar las componentes x_i de la solución del sistema, suponiendo que $|a| \leq 1$.

2. Resolver el siguiente sistema lineal por eliminación gaussiana sin pivotaje:

$$\left.\begin{array}{rcrcrcrcr} 20x_1 & + & 1.0x_2 & - & 0.1x_3 & + & 1.0x_4 & = & 2.7 \\ 0.4x_1 & + & 0.5x_2 & + & 4.0x_3 & - & 8.5x_4 & = & 21.9 \\ 0.3x_1 & - & 1.0x_2 & + & 1.0x_3 & + & 5.2x_4 & = & -3.9 \\ 1.0x_1 & + & 0.2x_2 & + & 2.5x_3 & - & 1.0x_4 & = & 9.9 \end{array}\right\}.$$

3. Resolver el sistema

$$(A|b) = \left(\begin{array}{ccc|c} 1 & 1 & 0 & 4 \\ 2 & 0 & 2 & 4 \\ 0 & 3 & 3 & 4 \end{array}\right),$$

   por el método de Gauss

   a) sin pivotaje,

   b) con pivotaje maximal por columnas, y

   c) con pivotaje completo.

4. Resolver el sistema

$$(A|b) = \left(\begin{array}{ccc|c} 0.15 & 2.11 & 30.75 & -26.38 \\ 0.64 & 1.21 & 2.05 & 1.01 \\ 3.21 & 1.53 & 1.04 & 5.23 \end{array}\right)$$

   por el método de Gauss

   a) sin pivotaje,

   b) con pivotaje maximal por columnas, y

   c) con pivotaje completo.

5. Resolver los sistemas

$$\left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 1 & 4 & 9 & 16 & 1 \\ 1 & 8 & 27 & 64 & 1 \\ 1 & 16 & 81 & 216 & 1 \end{array}\right), \quad \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 1 \\ 1 & 4 & 9 & 16 & 0 \\ 1 & 8 & 27 & 64 & 0 \\ 1 & 16 & 81 & 216 & 1 \end{array}\right);$$

usando

a) el método LU,

b) el método de Doolittle, y

c) el método de Crout.

6. Efectuar la factorización LU de la matriz

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 3 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 5 \end{pmatrix}.$$

7. Consideremos las matrices A y B dadas por

$$\begin{pmatrix} 1 & 2 & 1 & -1 \\ 2 & -3 & -2 & 0 \\ 0 & 5 & -3 & 2 \\ -3 & 1 & 4 & 0 \end{pmatrix}, \quad \begin{pmatrix} 2 & -5 & 6 & 4 & -1 \\ 3 & 0 & -2 & -1 & 3 \\ 0 & 4 & -5 & 0 & 1 \\ 1 & -1 & 1 & -3 & 0 \\ 5 & 1 & 6 & 1 & -1 \end{pmatrix},$$

respectivamente, y los vectores

$$b^{(1)} = \begin{pmatrix} 7 \\ 7 \\ -17 \\ -5 \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} 4 \\ -6 \\ 3 \\ 9 \end{pmatrix}, \quad b^{(3)} = b^{(1)} - b^{(2)}.$$

Resolver los sistemas

$$Ax^{(1)} = b^{(1)}, \quad Ax^{(2)} = b^{(2)}, \quad Ax^{(3)} = b^{(3)}, \quad Bx^{(4)} = b^{(4)}.$$

8. a) Resolver el sistema

$$(A|b) = \left(\begin{array}{cccc|c} 10 & 7 & 8 & 7 & 4 \\ 7 & 5 & 6 & 5 & 3 \\ 8 & 6 & 10 & 9 & 3 \\ 7 & 5 & 9 & 10 & 1 \end{array}\right),$$

obteniendo primeramente la factorización $A = LDL^\top$ con L triangular inferior con unos en la diagonal y D diagonal.

b) Explicitar la factorización de Cholesky de A.

9. Resolver el siguiente sistema por el método de Cholesky:

$$\left.\begin{array}{rcrcrcrcl} 1.00x_1 & + & 0.42x_2 & + & 0.54x_3 & + & 0.66x_4 & = & 0.3 \\ 0.42x_1 & + & 1.00x_2 & + & 0.32x_3 & + & 0.44x_4 & = & 0.5 \\ 0.54x_1 & + & 0.32x_2 & + & 1.00x_3 & + & 0.22x_4 & = & 0.7 \\ 0.66x_1 & + & 0.44x_2 & + & 0.22x_3 & + & 1.00x_4 & = & 9.9 \end{array}\right\} .$$

10. a) Explicitar el método LU para resolver sistemas tridiagonales, pentadiagonales y heptadiagonales.

b) Contar el número de operaciones requeridas en cada caso.

c) Repetir los apartados anteriores, suponiendo que la matriz del sistema es simétrica.

11. a) Demostrar que el método LU conserva el carácter banda de las matrices; esto es, si A es una matriz banda (p, q), las matrices L y U correspondientes son banda $(p, 1)$ y banda $(1, q)$, respectivamente.

b) Contar el número de operaciones necesarias para resolver un sistema banda (p, q).

c) Explicitar la resolución de un sistema banda $(4, 3)$ de dimensión $n \geq 4$ por el método LU.

12. a) Explicitar un método para resolver directamente el sistema con *matriz tridiagonal por bloques*

$$\left(\begin{array}{ccccccc|c} A_1 & C_1 & & & & & & b^{(1)} \\ B_2 & A_2 & C_2 & & & & & b^{(2)} \\ & B_3 & A_3 & C_3 & & & & b^{(3)} \\ & & \ddots & \ddots & \ddots & & & \vdots \\ & & & \ddots & \ddots & \ddots & & \vdots \\ & & & & B_{n-1} & A_{n-1} & C_{n-1} & b^{(n-1)} \\ & & & & & B_n & A_n & b^{(n)} \end{array}\right) ,$$

donde los bloques A_i $(i = 1 \div n)$ son tridiagonales y los B_i $(i = 2 \div n)$ y C_i $(i = 1 \div n-1)$ son diagonales, todos nxn.

b) Aplicación: Resolver el sistema anterior en el caso $n = 4$ para

$$b^{(i)} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad A_i = \begin{pmatrix} -4 & 1 & 0 & 0 \\ 1 & -4 & 1 & 0 \\ 0 & 1 & -4 & 1 \\ 0 & 0 & 1 & -4 \end{pmatrix} \quad (i = 1 \div 4) ,$$

$$B_{i+1} = C_i = I \quad (i = 1 \div 3) \ .$$

13. a) Indicar un método directo de resolución de sistemas de la forma siguiente, llamados *sistemas casi tridiagonales*:

$$\left(\begin{array}{ccccccc|c} a_1 & c_1 & & & & & e & z_1 \\ b_2 & a_2 & c_2 & & & & & z_2 \\ & b_3 & a_3 & c_3 & & & & z_3 \\ & & \ddots & \ddots & \ddots & & & \vdots \\ & & & \ddots & \ddots & \ddots & & \vdots \\ & & & & b_{n-1} & a_{n-1} & c_{n-1} & z_{n-1} \\ d & & & & & b_n & a_n & z_n \end{array}\right) .$$

b) Si la matriz fuese además simétrica, ¿cómo se tendría que variar el método con el fin de minimizar la memoria usada y el número de operaciones?

c) Aplicación: Resolver el sistema

$$\left(\begin{array}{rrrrr|r} 4 & -1 & 0 & 0 & 1 & 1 \\ -1 & 4 & -1 & 0 & 0 & 2 \\ 0 & -1 & 4 & -1 & 0 & 3 \\ 0 & 0 & -1 & 4 & -1 & 4 \\ 1 & 0 & 0 & -1 & 4 & 5 \end{array}\right) .$$

14. Sea A una matriz regular dada. Para resolver el sistema lineal $A^2x = b$, ¿es mejor calcular A^2 y resolver luego el sistema, o bien, resolver sucesivamente los dos sistemas $Ay = b$, $Ax = y$, siempre usando la factorización LU?

15. Sea A una matriz definida positiva. Efectuamos una factorización de la forma $A = LDL^\top$, con L triangular inferior con unos en la diagonal y D diagonal con elementos diagonales positivos.

a) ¿Cuál es el número total de operaciones aritméticas ($+$, $-$, $*$, $/$) que se deben realizar?

Un *ordenador vectorial* es un ordenador que permite realizar muchas operaciones con una sola instrucción (llamada *instrucción vectorial*); así, si x e y son dos vectores y # es una operación aritmética cualquiera, $x\#y$ calcula el vector de componentes $x_i\#y_i$. Obviamente, en el cálculo de la factorización planteada hay muchas operaciones que se pueden hacer vectorialmente.

b) Calcular el número total de operaciones escalares y operaciones vectoriales que se requieren para llevar a cabo la factorización planteada en un ordenador vectorial.

16. Se quiere factorizar la matriz A como producto de una matriz ortogonal Q y una matriz triangular superior R usando

    i) el método de ortogonalización de Gram-Schmidt,

    ii) el método de ortogonalización de Householder.

    a) Realizar las factorizaciones QR de la matriz

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 216 \end{pmatrix},$$

    por los métodos citados.

    b) Resolver el sistema $Ax = b$ con $b^\top = (1\ 1\ 1\ 1)$.

17. Calcular la inversa de la matriz

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 216 \end{pmatrix},$$

    a) realizando la factorización LU, invirtiendo L y U, y efectuando finalmente el producto $U^{-1}L^{-1}$;

    b) realizando una factorización QR, invirtiendo la matriz R y después efectuando el producto $R^{-1}Q^\top$;

    c) comparar los métodos a) y b) con respecto al número de operaciones.

18. El *método de Gauss-Jordan* de inversión de matrices parte de los sistemas de ecuaciones $Ax^{(j)} = e^{(j)}$ $(j = 1 \div n)$ escritos en forma matricial

$$(A|I) = \left(\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1n} & 1 & 0 & 0 \\ \vdots & & \vdots & 0 & \ddots & 0 \\ a_{n1} & \cdots & a_{nn} & 0 & 0 & 1 \end{array}\right),$$

    i va modificando las filas de $(A|I)$, usando los elementos de la diagonal como pivotes (como en el método de Gauss), de manera que se anulen los elementos no diagonales de A (también los de encima de la diagonal!), columna a columna. Se obtiene finalmente $(D|B)$ con D diagonal; entonces, se tiene $A^{-1} = D^{-1}B$.

    Sea A una matriz regular nxn.

    a) Explicitar el método de Gauss-Jordan para el cálculo de su inversa.

    b) Contar el número de operaciones a realizar, en el caso más general.

c) Aplicación: Invertir la matriz

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 216 \end{pmatrix}$$

por el método citado.

19. Sea A una matriz regular nxn que consideramos partida en cuatro bloques

$$A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix} ,$$

con P y S matrices pxp y sxs, respectivamente $(p + s = n)$.

Suponiendo que P y $S - RP^{-1}Q$ tienen inversa:

a) Dar un método para calcular la inversa de A.

b) En general, ¿cuántas operaciones son necesarias para obtener A^{-1} utilizando este método?

c) ¿Cuándo será útil el método citado?

d) Explicitar un procedimiento para invertir una matriz A, usando $n - 1$ veces el procedimiento anterior.

e) ¿Bajo qué condiciones funcionará el procedimiento del apartado d)?

f) Demostrar que, si A es definida positiva, puede asgurarse que funciona.

g) Aplicación: Calcular

$$\left(\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ \hline 1 & 0 & 1 & 0 \\ 0 & 2 & -1 & 0 \end{array}\right)^{-1} \text{, usando el apartado a),}$$

$$y \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}^{-1} \text{, usando el apartado d) .}$$

20. Calcular el determinante y la inversa de las matrices siguientes:

$$\begin{pmatrix} 2 & 1 & 2 \\ 1 & 2 & 3 \\ 4 & 1 & 2 \end{pmatrix} , \quad \begin{pmatrix} 4 & 2 & 1 & 1 \\ 1 & -3 & -2 & 0 \\ 7 & 10 & 1 & 3 \\ 2 & 4 & 0 & 1 \end{pmatrix} ,$$

$$\begin{pmatrix} 1 & 3 & 6 & 9 & 5 \\ -2 & -7 & -5 & -17 & -12 \\ -1 & -4 & -5 & 1 & -4 \\ 3 & 11 & 2 & 32 & 23 \\ 1 & 4 & 5 & 3 & 4 \end{pmatrix}.$$

21. Queremos calcular la inversa de una *matriz bidiagonal (inferior)*

$$A = \begin{pmatrix} a_1 & & & & & & \\ b_2 & a_2 & & & & & \\ & b_3 & a_3 & & & & \\ & & \ddots & \ddots & & & \\ & & & \ddots & \ddots & & \\ & & & & b_{n-1} & a_{n-1} & \\ & & & & & b_n & a_n \end{pmatrix}.$$

a) Dar un método para el cálculo de cada una de sus columnas.

b) Calcular el número de operaciones necesarias para invertirla, en el caso más general.

c) Encontrar la inversa cuando $a_i = a$ $(i = 1 \div n)$, $b_i = b$ $(i = 2 \div n)$.

d) Si las multiplicaciones y las divisiones se realizan en precisión doble, acotar el error relativo en el elemento $(A^{-1})_{n1}$, debido sólo a las citadas operaciones.

22. a) Calcular el determinante

$$\begin{vmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 3 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{vmatrix},$$

i) usando el método de Gauss,

ii) usando el método de ortogonalización de Householder.

b) Comparar ambos métodos con respecto al número de operaciones, en el caso más general.

23. a) Deducir una recurrencia para el cálculo de determinantes de matrices tridiagonales.

b) Aplicación: Determinar la dimensión máxima de las matrices

$$A_n = \begin{pmatrix} 2 & 1.02 & & & & & & \\ 0.99 & 2 & 1.02 & & & & & \\ & 0.99 & 2 & 1.02 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & 0.99 & 2 & 1.02 & \\ & & & & & 0.99 & 2 & \end{pmatrix} ,$$

para que sean definidas positivas.

c) Establecer también una recurrencia para el cálculo de determinantes de matrices pentadiagonales.

24. Sean $e^{(i)}$ $(i = 1 \div n)$ los vectores (columna) de la base canónica de $\mathbb{R}^n$; dados los números reales a_i $(i = 2 \div n)$, consideramos la matriz

$$S = \sum_{i=1}^{n-1} a_{i+1} e^{(i+1)} (e^{(i)})^\top \ .$$

a) Calcular S^k para todo k entero positivo.

(Indicación: Obtener S^2 y generalizarlo, no hace falta usar inducción).

b) Calcular $(I - S)^{-1}$.

c) Calcular $(I - S)(I - S)^\top$.

d) Aplicación: ¿Cuál es la factorización de Cholesky de la matriz B que tiene como únicos elementos no nulos los siguientes:

$$b_{11} = 1 \ , \quad b_{ii} = i^2 + 1 \quad (i = 2 \div n) \ ,$$

$$b_{i,i-1} = i \quad (i = 2 \div n) \ , \quad b_{i,i+1} = i \quad (i = 1 \div n - 1) \ ?$$

25. a) Dada una matriz nxn regular A y 2 vectores u y v de dimensión n, demostrar que $A - uv^\top$ es regular si y sólo si $v^\top A^{-1} u \neq 1$, y entonces

$$(A - uv^\top)^{-1} = A^{-1} + \alpha A^{-1} uv^\top A^{-1} \ ,$$

donde $\alpha = 1/(1 - v^\top A^{-1} u)$ (*fórmula de Sherman-Morrison*).

b) Sea B una matriz que difiere de A sólo en una fila y en una columna; suponiendo conocida A^{-1}, dar un procedimiento de cálculo de B^{-1} basado en el apartado a).

c) Aplicación: Sabiendo que

$$\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}^{-1} = \frac{1}{4} \begin{pmatrix} 3 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 3 \end{pmatrix} \ ;$$

calcular, usando el apartado b),

$$\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}^{-1} .$$

d) Generalizar la fórmula de Sherman-Morrison a otra que permita efectuar el cálculo de $(A - UV^\top)^{-1}$ a partir de A^{-1}, siendo ahora U y V matrices nxr.

26. Consideremos los vectores u y v de n componentes reales y generemos la matriz $A = I + uv^\top$.

a) Probar que existe un polinomio de segundo grado $P(x) = x^2 + ax + b$ con a, b reales tal que $P(A) = A^2 + aA + bI = 0$.

b) Demostrar que A es inversible si y sólo si $v^\top u \neq -1$ y hallar, en este caso, la inversa de A. Estudiar el caso particular en que u y v son ortogonales.

Consideremos el sistema lineal $Cx = c$, siendo $C = (c_{ij})$ la matriz de término general

$$c_{ij} = \begin{cases} \alpha & (\text{ si } i = j) \\ \beta & (\text{ si } i \neq j) \end{cases} ,$$

donde α y β son no nulos y diferentes.

c) Demostrar que el sistema anterior es compatible y determinado si y sólo si $\alpha + (n-1)\beta \neq 0$. (Indicación: Hallar λ tal que $C - \lambda I$ tenga rango 1).

d) Explicar cómo se pueden aplicar los resultados anteriores a la resolución del citado sistema.

27. a) Demostrar que la norma de matrices (no multiplicativa)

$$\|A\|_E = (\sum_{i,j} |a_{ij}|^2)^{\frac{1}{2}}$$

no está subordinada a ninguna norma vectorial.

(Indicación: Considérese la norma de la matriz identidad).

b) Demostrar que la norma subordinada a la norma euclídea

$$\|x\|_2 = (\sum_{i} |x_i|^2)^{\frac{1}{2}}$$

es

$$(\rho(A^* A))^{\frac{1}{2}} .$$

(Indicación: Usar el hecho de que, si N es hermitiana, existe una matriz U unitaria tal que $U^* N U$ es diagonal).

c) Demostrar las relaciones siguientes:

i) $\|A^*A\|_2 = \|A\|_2^2 = \|A^*\|_2^2$;

ii) $\|U^*AU\|_2 = \|AU\|_2 = \|UA\|_2$, si U es unitaria;

iii) $\|A\|_2 \leq \|A\|_E \leq \sqrt{n}\|A\|_2$ y $\frac{1}{\sqrt{n}}\|A\|_E \leq \|A\|_2 \leq \|A\|_E$;

iv)

$$\|A\|_2 = \max_{x,y\neq 0} \frac{|y^*Ax|}{\|x\|_2\|y\|_2} \ .$$

28. a) Consideremos una matriz A tal que $\max_{i,j}|a_{ij}| \leq a$. Acotar en función de a: $\|A\|_\infty$, $\|A\|_1$, $\|A\|_2$ y $\|A\|_E$.

b) Comprobar el resultado obtenido en el apartado a) con la matriz

$$A = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 2 & 2 \\ 0 & 1 & 1 \end{pmatrix} \ .$$

29. Consideremos una norma matricial $\|\ \|$ subordinada a una norma vectorial cualquiera.

a) Demostrar que, si $\|R\| \leq 1$, $I + R$ es regular, y entonces

$$\frac{1}{1+\|R\|} \leq \|(I+R)^{-1}\| \leq \frac{1}{1-\|R\|} \ .$$

b) Probar que, si A y $A + E$ son regulares, entonces

$$\|(A+E)^{-1} - A^{-1}\| \leq \|A^{-1}\| \frac{\alpha}{1-\alpha} \ ,$$

con $\alpha = \|A^{-1}E\|$.

30. Sea $\|\ \|_2$ la norma matricial subordinada a la norma vectorial euclídea.

a) Demostrar que, para cualquier matriz $A = (a_{ij})$, se verifica

$$\|A\|_2 \geq |a_{ii}| \ \ (i = 1 \div n) \ .$$

b) Demostrar que, si A es triangular con elementos diagonales diferentes de cero, entonces

$$\|A^{-1}\|_2 \geq \frac{1}{|a_{ii}|} \ (i = 1 \div n) \ .$$

Obtener, en este caso, una cota inferior para el número de condición de A en esta norma.

31. a) Demostrar que las series geométricas matriciales

$$A + AB + \cdots + AB^j + \cdots \ , \quad A + BA + \cdots + B^jA + \cdots$$

convergen si $\|B\| < 1$ para alguna norma matricial subordinada a una norma vectorial cualquiera y, en tal caso,

$$\sum_{j=0}^{\infty} AB^j = A(I-B)^{-1} \ , \ \sum_{j=0}^{\infty} B^jA = (I-B)^{-1}A \ .$$

b) Acotar la norma del resto n-ésimo en función de $\|B\|$ y $\|A\|$.

c) Aplicación: Calcular

$$\begin{pmatrix} 0.99 & 1.01 \\ 0.98 & 1.02 \end{pmatrix}^{-1}$$

con un error menor que 10^{-6} en la norma $\| \ \|_\infty$.

32. Para una matriz nxn A, definimos

$$\begin{aligned} e^A &= \sum_{j=0}^{\infty} \frac{A^j}{j!} \ , \\ \operatorname{sen} A &= \sum_{r=0}^{\infty} (-1)^r \frac{A^{2r+1}}{(2r+1)!} \ , \\ \cos A &= \sum_{r=0}^{\infty} (-1)^r \frac{A^{2r}}{(2r)!} \ . \end{aligned}$$

a) Encontrar expresiones para las acotaciones de los restos de estas series matriciales en una norma dada.

b) Aplicación: Dada

$$A = \begin{pmatrix} 0.1 & 0.01 \\ 0.01 & 0.1 \end{pmatrix} \ ,$$

calcular e^A, $\operatorname{sen} A$ y $\cos A$ con un error menor que 10^{-6} en la norma $\| \ \|_\infty$.

33. Sea R una matriz regular. Dada una norma vectorial cualquiera $\| \ \|$ definimos $\|x\|_R \equiv \|Rx\|$.

a) Demostrar que $\| \ \|_R$ es una norma vectorial y que

$$\frac{\|x\|}{\|R^{-1}\|} \leq \|x\|_R \leq \|R\| \ \|x\| \ ,$$

donde la norma matricial $\| \ \|$ indica la norma subordinada a la norma vectorial dada.

b) Expresar la norma matricial $\| \ \|_R$ subordinada a la norma vectorial definida (y denotada de la misma manera) en términos de la norma matricial subordinada a $\| \ \|$.

c) Comprobar que los números de condición

$$\mu_R(A) \equiv \|A\|_R \|A^{-1}\|_R$$

pueden ser arbitrariamente grandes para una matriz A fijada, escogiendo R adecuadamente.

34. Consideremos el sistema

$$(A|b) = \left(\begin{array}{cccc|c} 10 & 7 & 8 & 7 & 32 \\ 7 & 5 & 6 & 5 & 23 \\ 8 & 6 & 10 & 9 & 33 \\ 7 & 5 & 9 & 10 & 31 \end{array}\right) .$$

a) Resolverlo exactamente.

b) Comprobar que los vectores

$$x = x^{(1)} = \begin{pmatrix} 6 \\ -7.2 \\ -2.9 \\ -0.1 \end{pmatrix} , \quad x = x^{(2)} = \begin{pmatrix} 1.50 \\ 0.18 \\ 1.19 \\ 0.89 \end{pmatrix}$$

tienen *restos vectoriales* $r_x = b - Ax$ muy pequeños en $\| \ \|_\infty$.

c) Calcular los números de condición $\mu_\infty(A)$ y $\mu_1(A)$.

d) Explicar el fenómeno descubierto en los apartados a) y b).

35. a) Demostrar que el cociente

$$g_n = \frac{\max_{i,j,k} \left|a_{ij}^{(k)}\right|}{\max_{i,j} |a_{ij}|} ,$$

que aparece en el estudio de los errores en los cálculos al resolver sistemas lineales por el método de Gauss con pivotaje maximal por columnas, es menor que 2^{n-1}.

b) Probar que g_n alcanza este valor máximo 2^{n-1} para las matrices

$$A_n = \begin{pmatrix} 1 & 0 & 0 & \ddots & \ddots & 0 & 1 \\ -1 & 1 & 0 & \ddots & \ddots & 0 & 1 \\ -1 & -1 & 1 & \ddots & \ddots & 0 & 1 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \ddots & \ddots & 0 & 1 \\ -1 & -1 & -1 & \ddots & \ddots & 1 & 1 \\ -1 & -1 & -1 & \ddots & \ddots & -1 & 1 \end{pmatrix} .$$

c) Acotar, en función de n, el error al realizar la factorización LU de estas matrices, debido a los errores en los cálculos, si usamos un ordenador que trabaja con doble precisión.

36. Acotar el valor de g_n, cuando usamos pivotaje maximal por columnas para resolver el sistema $Ax = b$, en los siguientes casos particulares:

i) A es una matriz Hessenberg superior;

ii) A es una matriz tridiagonal.

37. Tenemos que resolver el sistema $Cx = b$, donde $C = A^m$, A es una matriz nxn y $\|A\|_\infty = a$; sabemos también que $\|A^{-1}\|_\infty = \|A\|_\infty$. Disponemos de la siguiente fórmula para el error relativo en x, usando eliminación gaussiana con pivotaje completo

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq \epsilon n^{2+0.25 \ln n} \|C\|_\infty \|C^{-1}\|_\infty \ ,$$

donde ϵ es una cota del error relativo de redondeo en las operaciones aritméticas.

a) Obtener una acotación de este error relativo en función de ϵ, n, a y m.

Un sistema equivalente, en cierto sentido, al anterior es

$$Ay^{(1)} = b \ , \quad Ay^{(j+1)} = y^{(j)} \ (j = 1 \div m - 1) \ ,$$

con $y^{(j)} \in \mathbb{R}^n \ \ (j = 1 \div m)$.

b) Comprobar que $x = y^{(m)}$ y escribir dicho sistema en forma matricial

$$By = d \ , \quad y, d \in \mathbb{R}^{nm} \ . \quad (*)$$

c) Calcular la inversa de B.

d) Usando la acotación dada, establecer una acotación del error relativo para el problema escrito en la forma $(*)$ en función de ϵ, n, a y m.

e) Aplicación: Evaluar las cotas encontradas para $\epsilon = 10^{-13}$, $n = 4$, $a = 2000$ y $m = 5$. ¿Es permisible la acotación obtenida en el primer caso? ¿Y la obtenida en el segundo?

38. Dados los sistemas

$$\left(\begin{array}{ccc|c} 2 & 0 & 3 & 1 \\ 0 & 4 & 2 & 2 \\ 0 & 1 & 6 & 3 \end{array}\right) \ , \quad \left(\begin{array}{ccc|c} 1 & 2 & -2 & 1 \\ 1 & 1 & 1 & 1 \\ 2 & 2 & 1 & 1 \end{array}\right) \ ,$$

$$\left(\begin{array}{ccc|c} 2 & -1 & 1 & 1 \\ 2 & 2 & 2 & 1 \\ -1 & -1 & 2 & 1 \end{array}\right) \ .$$

a) Discutir la convergencia de los métodos iterativos de Jacobi y de Gauss-Seidel.

b) Calcular las soluciones aproximadas después de 10 iteraciones, partiendo del vector cero.

39. Sea la matriz nxn

$$A = \begin{pmatrix} 1 & 2 & 3 & \cdots & n-1 & n \\ n & 1 & 2 & \cdots & n-2 & n-1 \\ n-1 & n & 1 & \cdots & n-3 & n-2 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 3 & 4 & 5 & \cdots & 1 & 2 \\ 2 & 3 & 4 & \cdots & n & 1 \end{pmatrix} ;$$

esto es,

$$a_{ij} = \begin{cases} n+j-i+1 & (j<i) \\ 1+j-i & (j \geq i) \end{cases} .$$

a) Demostrar que los vectores $v^{(j)}$, de componentes

$$v_k^{(j)} = q_j^k \text{ con } q_j = e^{\frac{i2\pi(j-1)}{n}} \quad (k, j = 1 \div n) ,$$

son vectores propios de A de valores propios respectivos

$$\lambda_j = \sum_{k=1}^{n} k q_j^{k-1} \ (j = 1 \div n) .$$

Queremos resolver un sistema $(A + aI)x = b$ por el método iterativo de Jacobi.

b) Discutir la convergencia del citado método en función de a.

c) Para $n = 4$, $a = 19$, $b = (1 \ \ 0 \ \ 0 \ \ 0)^\top$, hallar $x^{(k)}$ tal que

$$\|b - (A + aI)x^{(k)}\|_\infty \leq 0.03 .$$

d) Comprobar la divergencia del método cuando $a = 0$.

e) ¿Para qué valor de a es más rápida la convergencia?

40. Consideremos la matriz por bloques

$$A = \begin{pmatrix} I & S \\ S^* & I \end{pmatrix} ,$$

donde S es una matriz compleja nxn.

a) Si queremos resolver un sistema $Ax = b$, dar una condición necesaria y suficiente de convergencia del método iterativo de Jacobi.

b) Hacer lo mismo para el método iterativo de Gauss-Seidel.

41. Discutir, en función de a, la convergencia de los métodos iterativos de Jacobi y de Gauss-Seidel en sistemas lineales con la matriz

$$A = \begin{pmatrix} 1 & & & & & a \\ -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & -1 & 1 \end{pmatrix} .$$

42. Queremos resolver el sistema lineal

$$\begin{pmatrix} 1 & -a \\ -a & 1 \end{pmatrix} x = b \ ,$$

por el método iterativo de sobrerrelajación con un parámetro ω.

a) Escribir las iteraciones correspondientes.

b) Demostrar que el método converge si y sólo si las dos raíces de

$$\lambda^2 - [2(1-\omega) + (\omega a)^2]\lambda + (1-\omega)^2 = 0$$

son menores que 1, en módulo.

c) Deducir del apartado b) que el método converge si y sólo si $\omega \in (0, 2)$.

d) Demostrar que el valor óptimo de ω es

$$\omega^* = \frac{2}{a^2}(1 - \sqrt{1-a^2}) \ .$$

e) Aplicación: Si $a = \frac{1}{2}$ y usamos el factor óptimo de ω, ¿cuántas iteraciones son necesarias para asegurar que el error inicial se reduce al menos en un factor 10^{-6} en la norma euclídea?

43. Sea B la matriz nxn de elementos

$$b_{ij} = \begin{cases} 1 & (\text{si } i-j=1 \text{ ó } j-i=1) \\ 0 & (\text{en caso contrario}) \end{cases} .$$

a) Demostrar que tiene los valores propios

$$\lambda_j = 2\cos\frac{j\pi}{n+1} \quad (j = 1 \div n)$$

con vectores propios asociados $v^{(j)}$ de componentes

$$v_i^{(j)} = \operatorname{sen}\frac{ij\pi}{n+1} \quad (i, j = 1 \div n) \ .$$

Consideremos los siguientes sistemas de ecuaciones lineales:

$$x_0 = p \ , \quad x_i = a(x_{i-1} - 2x_i + x_{i+1}) \ \ (i = 1 \div n) \ , \quad x_{n+1} = q \ ,$$

para valores positivos de a.

b) Estudiar la convergencia de los métodos iterativos siguientes, en función de a:

$$\begin{aligned} x_0^{(k+1)} &= p \ , \\ x_i^{(k+1)} &= a(x_{i-1}^{(k)} - 2x_i^{(k)} + x_{i+1}^{(k)}) \ \ (i = 1 \div n) \ , \\ x_{n+1}^{(k+1)} &= q \ . \end{aligned}$$

c) Demostrar que el método iterativo de Jacobi aplicado a aquellos sistemas converge para todo valor positivo de a.

44. Consideremos el siguiente problema de valores en la frontera de una ecuación diferencial de segundo orden: determinar la función $u(x)$ en el intervalo $[0,1]$ que cumple

$$\frac{d^2u}{dx^2}(x) - cu(x) = 0 \ (c > 0) \ \ \forall x \in (0,1) \ ; \ \ u(0) = \alpha, \ u(1) = \beta \ .$$

a) Encontrar la solución analítica del problema.

Una aproximación de la solución se puede obtener evaluando la ecuación en las abscisas $x_i = \frac{i}{n}$ $(i = 0 \div n)$, discretizando la derivada segunda en la forma usual

$$\frac{d^2u}{dx^2}(x_i) \simeq n^2[u(x_{i-1}) - 2u(x_i) + u(x_{i+1})]$$

y resolviendo finalmente el sistema lineal resultante (discretizado).

b) Escribir este problema discretizado en las aproximaciones u_i de $u(x_i)$.

c) Estudiar la convergencia de los métodos iterativos de Jacobi y de Gauss-Seidel hacia la solución del problema discretizado.

45. Sea A una matriz regular nxn y X_0 una matriz arbitraria nxn. Definimos una sucesión de matrices, mediante la recurrencia

$$X_{k+1} = X_k + X_k(I - AX_k) \ , \quad (k \geq 0) \ .$$

a) Demostrar que esta sucesión converge a A^{-1} si y sólo si $\rho(I - AX_0) < 1$.

b) Dados

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \ , \quad X_0 = \begin{pmatrix} 1.9 & -0.9 \\ -0.9 & 0.9 \end{pmatrix} \ ,$$

calcular una aproximación de la inversa de A utilizando el método iterativo anterior hasta que $\rho(I - AX_k) < 10^{-6}$.

46. Localizar una zona del plano complejo que contenga los valores propios de la matriz

$$A = \begin{pmatrix} -0.5 & 0.01 & 0 & -0.01 \\ 0.02 & 0 & 0.25 & 0.25 \\ -0.01 & 0.3 & 0 & 0.3 \\ 0 & -0.01 & 0.01 & 0.5 \end{pmatrix} .$$

47. a) Calcular los valores propios de módulo máximo y mínimo de la matriz

$$A = \begin{pmatrix} 9 & 10 & 8 \\ 10 & 5 & -1 \\ 8 & -1 & 3 \end{pmatrix} .$$

b) Calcular el valor propio restante.

48. 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. Calcularlo con 6 cifras decimales correctas.

49. Se supone que la edad máxima a la cual pueden llegar las hembras de cierta población animal es de 15 años. Dividimos la población total de hembras según la edad en 3 clases de 5 años de duración cada una y denotamos por x_i el número de hembras de la clase i-ésima. Sean $a_i \geq 0$ las medias del número de hijas que tienen las hembras que están en la clase i-ésima ($i = 1, 2, 3$), y b_i ($0 < b_i \leq 1$) la fracción de hembras que pertenecen a la clase i-ésima que se espera que sobrevivan y pasen a la clase siguiente de edad ($i = 1, 2$).

Podemos escribir, por lo tanto, la relación entre una distribución de hembras $x^{(k)}$ y su distribución al cabo de 5 años $x^{(k+1)}$:

$$x^{(k+1)} = Lx^{(k)} \ , \quad \text{con} \quad L = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & 0 & 0 \\ 0 & b_2 & 0 \end{pmatrix} \ ;$$

la matriz L se denomina *matriz de Leslie*.

Suponemos que inicialmente hay sólo 1000 hembras en la primera de las 3 clases y que $a_1 = 0$, $a_2 = 1.68$, $a_3 = 1.28$, $b_1 = 0.5$ y $b_2 = 0.25$.

a) ¿Cuál será la distribución después de $5k$ años?

b) Determinar la distribución relativa en el límite (cuando k tiende a infinito); esto es, los límites de los cocientes entre el número de hembras de cada clase y el número total de hembras.

c) Relacionar el comportamiento en el límite con los valores y vectores propios de A.

50. Modificar el método de la potencia para que pueda ser usado en los casos siguientes:

i) el valor propio de módulo máximo es de multiplicidad $m > 1$;

ii) los valores propios de módulo máximo son λ_1 y $\lambda_2 = -\lambda_1$.

51. Supongamos que A es una matriz 3x3 con valores propios λ_1 y λ_2, tales que $|\lambda_1| > |\lambda_2|$ y λ_1 es de multiplicidad 2. Supongamos que A tiene sólo dos vectores propios normalizados linealmente independientes $v^{(1)}$ y $v^{(2)}$ asociados a λ_1 y λ_2, respectivamente. Se puede demostrar que existe un vector no nulo v tal que $(A-\lambda_1 I)v = v^{(1)}$ (v recibe el nombre de *vector propio generalizado*).

a) Demostrar que v, $v^{(1)}$ y $v^{(2)}$ son linealmente independientes.

b) Expresar $A^k v$ como combinación lineal de los vectores linealmente independientes de a) ($k \geq 0$).

c) Sea w un vector cualquiera tal que $w^\top v \neq 0$ y $w^\top v^{(1)} \neq 0$. Calcular

$$\lim_{k\to\infty} \frac{w^\top x^{(k+1)}}{w^\top x^{(k)}} \ ,$$

donde $(x^{(k)})_{k\geq 0}$ es la sucesión obtenida al aplicar el método de la potencia a la matriz A, partiendo de un vector inicial

$$x^{(0)} = av + a_1 v^{(1)} + a_2 v^{(2)} \ ,$$

con $a \neq 0$.

d) Sea e_k el error cometido al aproximar el límite anterior por el término k-ésimo de la sucesión $w^\top x^{(k)}$ ($k \geq 0$). Determinar

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k} \ .$$

52. a) Reducir las matrices siguientes a forma tridiagonal simétrica mediante matrices de Householder:

$$A = \begin{pmatrix} 6 & -1 & 1 & 1 \\ -1 & 5 & -1 & 1 \\ 1 & -1 & 4 & -1 \\ 1 & 1 & -1 & 3 \end{pmatrix} \ , \quad B = \begin{pmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 2 & 0 \\ 0 & 2 & 1 & 2 \\ 2 & 0 & 2 & 1 \end{pmatrix} \ .$$

b) Explicitar las ecuaciones de los hiperplanos respecto a los cuales se han realizado las reflexiones.

53. Sea A una matriz tridiagonal y B la matriz tridiagonal que se obtiene a partir de A, cambiando únicamente el signo de los elementos de la diagonal. Demostrar que λ es un valor propio de B si y sólo si $-\lambda$ lo es de A.

54. Sea una matriz tridiagonal nxn $A = (a_{ij})$ con $a_{ii} = a$ $(i = 1 \div n)$ y $a_{i,i-1} = b$, $a_{i-1,i} = c$ $(i = 2 \div n)$.

    a) Dar una fórmula recurrente para calcular el polinomio característico.

    b) Demostrar que si n es impar, $\lambda = a$ es un valor propio.

    c) Calcular los valores propios para $n = 5$.

    d) Determinar los valores propios de A en el caso general.

55. ¿Cuántos valores propios positivos tienen las matrices $(2n)$x$(2n)$ del tipo

$$\begin{pmatrix} 3 & 2 & & & & & & & \\ 2 & -3 & 1 & & & & & & \\ & 1 & 3 & 2 & & & & & \\ & & 2 & -3 & 1 & & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & \ddots & \ddots & \ddots & & \\ & & & & & 1 & 3 & 2 & \\ & & & & & & 2 & -3 & \end{pmatrix} ?$$

56. Calcular los valores y vectores propios de las matrices

$$\begin{pmatrix} 3 & 2 & 0 \\ 2 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 3 & 2+i & 0 \\ 2-i & 2 & 1+i \\ 0 & 1-i & 2 \end{pmatrix},$$

    obteniendo el polinomio característico por recurrencia y usando el teorema de Sturm para separar los valores propios.

57. Hallar los valores y vectores propios de la matriz

$$A = \begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 2 & 1 & 4 \\ 2 & 1 & 4 & 3 \\ 1 & 4 & 3 & 2 \end{pmatrix},$$

    usando previamente el método de Householder para reducirla a forma tridiagonal simétrica.

58. Determinar los valores y vectores propios de la matriz

$$A = \begin{pmatrix} 3 & 4 & 0 & 0 \\ 2 & 3 & 4 & 0 \\ 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 \end{pmatrix},$$

mediante el cálculo recurrente del polinomio característico.

59. Aplicar los métodos iterativos LR y QR al cálculo de los valores y vectores propios de la matriz

$$\begin{pmatrix} 7 & 6 \\ 3 & 4 \end{pmatrix}.$$

CAPÍTULO 3

INTERPOLACIÓN Y APROXIMACIÓN DE FUNCIONES

La interpolación y la aproximación, tratadas en este capítulo, proporcionan técnicas de obtención de modelos funcionales simples que describen situaciones más complejas, expresadas por datos experimentales o por funciones de evaluación complicada. Ambas técnicas se presentan desde un enfoque unificador, en el cual la interpolación aparece como un caso particular de aproximación. Los distintos métodos de aproximación, que persiguen la minimización del error, se clasifican según la forma con que éste sea medido: mínimos cuadrados, minimax, ... La aproximación constituye un útil indispensable para la generación de los métodos numéricos de derivación, integración y cálculo de ceros de funciones.

3.1 INTERPOLACIÓN

3.1.1 Concepto de interpolación

Sean (x_k, y_k) $(k = 0 \div m)$ pares de valores reales (puntos del plano) dados de manera que $x_k \neq x_i$, si $k \neq i$. Puede preguntarse si existe alguna función p, de un tipo predeterminado, tal que $p(x_k) = y_k$ $(k = 0 \div m)$.

La cuestión anterior recibe el nombre de *problema de interpolación* y el proceso de búsqueda de la función p se llama *interpolación*.

La resolución de este problema es de utilidad en muchas situaciones, en particular cuando la procedencia de los puntos (x_k, y_k) $(k = 0 \div m)$ es experimental. Consideremos un proceso en el que, para determinados valores de una variable x, se obtiene un resultado expresado por la variable y. Supongamos que, conocida experimentalmente la respuesta y_k obtenida bajo condiciones x_k, nos interesa encontrar el resultado y que obtendríamos al tomar condiciones x no experimentadas.

Podemos pensar que los puntos dados forman parte de la gráfica de una función f que queremos conocer al menos aproximadamente y de la que únicamente sabemos que $f(x_k) = y_k$ $(k = 0 \div m)$. Esta situación, frecuente en la ciencia y la técnica, lleva al típico problema de interpolación en el que no aproximamos números, sino funciones.

Ejemplo

El valor de la aceleración de la gravedad en un punto de la superficie terrestre, en el nivel del mar, depende de la latitud. Experimentalmente, se ha comprobado la correspondencia

dada en la tabla 3.1.

Latitud (en grados)	Aceleración gravedad (en m/s^2)
0	9.780350
30	9.793238
45	9.806154
60	9.819099
90	9.832072

Tabla 3.1: Medidas de la gravedad a diversas latitudes.

El problema consiste en encontrar el valor de la gravedad en Barcelona, que está situada a una latitud de $41°25'$.

Teniendo en cuenta este ejemplo no resulta extraño que, a veces, se utilice la expresión "leer entre líneas" para referirse a la interpolación.

Al plantearse un problema de interpolación han de hacerse tres preguntas:

1. ¿De qué tipo (polinomial, trigonométrica, racional, etc.) ha de ser la función p buscada?

   La respuesta a esta cuestión vendrá ligada a las "sospechas" que tengamos respecto al comportamiento de f. Si f expresa algún fenómeno periódico, convendrá buscar p entre las funciones trigonométricas. Si tenemos razones para pensar que f tiene asíntotas, convendrá quizás que p sea de tipo racional. Si f responde a un comportamiento polinomial, deberemos buscar p entre las funciones polinomiales. Sólo trataremos este último tipo de interpolación: el de la *interpolación polinomial*.

2. Una vez elegido el conjunto de funciones del que debemos extraer p, ¿existe la función buscada? y, si existe, ¿es única?

3. ¿Es la función p una buena aproximación de la función f en las abscisas donde no coinciden?

3.1.2 Interpolación polinomial

Planteamiento del problema

Dada una tabla de $m+1$ *puntos de interpolación* (x_k, f_k) $(k = 0 \div m)$ con $x_k \neq x_i$, si $k \neq i$, llamamos *interpolación polinomial* a la determinación de un polinomio $p(x)$ de grado menor o igual que N, tal que

$$p(x_k) = f_k \ (k = 0 \div m) \ .$$

Cuando f_k sea el valor de una función f en x_k $(k = 0 \div m)$, hablaremos de *interpolación polinomial de la función* f en las *abscisas de interpolación* x_k $(k = 0 \div m)$.

A continuación, se da respuesta, en este caso, a las tres preguntas con que acabábamos el apartado anterior.

Tipo de función interpoladora

La función p buscada formará parte del conjunto de polinomios de grado menor o igual que N, para un cierto valor de N; es decir, $p(x)$ será de la forma

$$p(x) = a_N x^N + a_{N-1} x^{N-1} + \cdots + a_1 x + a_0 \ ,$$

y, para determinarla, habrá que hallar los $N+1$ coeficientes a_0, a_1, ..., a_N (reales o complejos, según el caso). En el caso de que a_N sea no nulo, diremos también que $p(x)$ tiene exactamente grado N.

Dado un número α (en general, complejo) y suponiendo $N \geq 1$, la *regla de Horner* nos permite efectuar el proceso de *división sintética*, consistente en encontrar un polinomio $q(x)$, de grado menor o igual que $N-1$, y un número r, tales que

$$p(x) = (x - \alpha) q(x) + r \ , \tag{3.1}$$

a través de la recurrencia

$$b_{N-1} = a_N \ , \quad b_j = b_{j+1}\alpha + a_{j+1} \ (j = N-2 \div -1) \ ,$$

que da

$$q(x) = b_{N-1} x^{N-1} + b_{N-2} x^{N-2} + \cdots + b_1 x + b_0 \ , \quad r = b_{-1} \ .$$

Se deducen dos consecuencias importantes:

- a) $p(\alpha) = r = b_{-1}$ y, por lo tanto, la regla de Horner es un proceso de evaluación de polinomios con sólo N multiplicaciones y sumas para un polinomio de grado N.

  Además, el proceso de división sintética aplicado a $q(x)$ para el número α

  $$c_{N-2} = b_{N-1} \ , \quad c_j = c_{j+1}\alpha + b_{j+1} \ (j = N-3 \div -1) \ ,$$

  da el valor de $p'(\alpha) = c_{-1}$. Repitiendo este proceso,

  $$d_{N-3} = c_{N-2} \ , \quad d_j = d_{j+1}\alpha + c_{j+1} \ (j = N-4 \div -1) \ ,$$

  obtenemos el valor de $p^{(2)}(\alpha) = 2d_{-1}$, como resulta al derivar la expresión (3.1) y evaluarla en $x = \alpha$. Las derivadas de orden superior se obtienen análogamente.

- b) $p(\alpha) = 0$ si y sólo si $p(x) = (x - \alpha)q(x)$ o, en otras palabras, un número α es un *cero* del polinomio $p(x)$ si y sólo si $(x - \alpha)$ es un *factor* de $p(x)$. Dado que el polinomio $q(x)$ es de grado menor o igual que $N-1$, repitiendo este proceso resulta que si un polinomio $p(x)$ de grado menor o igual que N tiene N ceros $\alpha_1, \cdots, \alpha_n$ (es decir, $p(\alpha_k) = 0 \ (k = 1 \div N)$), entonces se escribe como

  $$p(x) = a_N (x - \alpha_1) \cdots (x - \alpha_N) \ ,$$

  con a_N coeficiente del monomio de grado N. Como consecuencia, un polinomio $p(x)$ de grado menor o igual que N no puede tener más de N ceros, a menos que sea idénticamente nulo. Esta propiedad se llama también *propiedad de Haar* para los polinomios y será usada frecuentemente.

Existencia y unicidad del polinomio interpolador

Existe un único polinomio $p_m(x)$, de grado menor o igual que $N = m$, tal que $p_m(x_k) = f_k$ $(k = 0 \div m)$. El polinomio $p_m(x)$ recibe el nombre de *polinomio interpolador* de f en las abscisas x_k $(k = 0 \div m)$.

Para ver esto, consideramos un polinomio $p_m(x) = a_0 + a_1 x + \cdots + a_m x^m$. Cuando imponemos las condiciones de interpolación, obtenemos un sistema lineal de $m+1$ ecuaciones en las $m+1$ incógnitas a_0, a_1, $\cdots$, a_m:

$$\left.\begin{array}{rcl} a_0 + a_1 x_0 + \cdots + a_m x_0^m &=& f_0 \\ a_0 + a_1 x_1 + \cdots + a_m x_1^m &=& f_1 \\ \cdots\cdots\cdots & & \cdots \\ a_0 + a_1 x_m + \cdots + a_m x_m^m &=& f_m \end{array}\right\} .$$

El determinante de la matriz del sistema se llama *determinante de Vandermonde* y tiene la forma

$$\left|\begin{array}{cccc} 1 & x_0 & \cdots & x_0^m \\ 1 & x_1 & \cdots & x_1^m \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 1 & x_m & \cdots & x_m^m \end{array}\right| = \prod_{k>i} (x_k - x_i) \ ;$$

este determinante es diferente de 0 (ya que $x_i \neq x_k$, si $i \neq k$). El sistema planteado es, pues, compatible y determinado y, por lo tanto, existe una única solución: a_0, a_1, ..., a_m, coeficientes del único polinomio interpolador.

Aunque el proceso que acaba de describirse es el primer método en el que se piensa cuando se intenta hallar $p_m(x)$, resulta excesivamente laborioso cuando m no es pequeño. Conviene entonces recurrir a otros métodos, algunos de los cuales se exponen en el apartado 3.1.3. Es importante notar que todos los métodos conducirán al mismo polinomio interpolador (porque es único, como se acaba de ver).

Error de interpolación

Nos interesa tener un criterio para "medir la proximidad" del polinomio $p_m(x)$ a la función f.

Dado un intervalo abierto (a, b), el conjunto de las funciones diferenciables $n+1$ veces con continuidad en (a, b) se denotará por $\mathcal{C}^{n+1}(a, b)$. Asimismo, notaremos por $\mathcal{C}^{n+1}([a, b])$ el conjunto de funciones que son restricciones en el intervalo $[a, b]$ de funciones diferenciables $n+1$ veces con continuidad sobre un intervalo abierto que contiene $[a, b]$.

A continuación, se dan expresiones del error de interpolación para funciones de $\mathcal{C}^{n+1}(a, b)$.

Si $f \in \mathcal{C}^{n+1}(a, b)$ y $x_k \in (a, b)$ $(k = 0 \div m)$; entonces, para todo x de (a, b), tenemos la expresión siguiente para el *error de interpolación*:

$$f(x) - p_m(x) = \frac{f^{(m+1)}(\xi(x))}{(m+1)!}(x - x_0)(x - x_1)\cdots(x - x_m) \ , \tag{3.2}$$

donde $\xi(x)$ depende de x y pertenece al mínimo intervalo que contiene las abscisas x_0, x_1, ..., x_m y x, que indicaremos por $< x_0, \ldots, x_m, x >$.

En efecto: si $x = x_k$ para algún k, la fórmula (3.2) es cierta y, si $x \neq x_k$ $(k = 0 \div m)$, consideramos la función

$$\Phi(z) = f(z) - p_m(z) - a(x)(z - x_0)(z - x_1) \cdots (z - x_m) \ ,$$

con $a(x)$ tal que $\Phi(x) = 0$. La función Φ se anula en las $m + 2$ abscisas: x_k $(k = 0 \div m)$ y x; aplicando $m + 1$ veces el teorema de Rolle, se deduce la existencia de un valor $\xi(x)$ en que se anula la derivada $m + 1$ de Φ:

$$\Phi^{(m+1)}(\xi(x)) = f^{(m+1)}(\xi(x)) - a(x)(m + 1)! = 0 \ ;$$

así tenemos una expresión para $a(x)$ que, substituida en $\Phi(x) = 0$, permite encontrar la expresión del error (3.2).

3.1.3 Métodos de cálculo del polinomio interpolador

Destacaremos aquí los métodos siguientes:

Método de Lagrange

Se toma como expresión del polinomio interpolador la *fórmula de interpolación de Lagrange*:

$$p_m(x) = \sum_{i=0}^{m} f_i l_i(x) \ , \quad l_i(x) = \frac{\prod_{k \neq i}(x - x_k)}{\prod_{k \neq i}(x_i - x_k)} \ (i = 0 \div m) \ . \tag{3.3}$$

Los polinomios $l_i(x)$ reciben el nombre de *polinomios de Lagrange*. Obsérvese que el grado de $l_i(x)$ es igual a m y que $l_i(x_k) = \delta_{ik}$ $(i, k = 0 \div m)$; por lo tanto, $p(x_k) = f_k$ $(k = 0 \div m)$, como se deseaba.

Aunque éste es el método de interpolación más explícito (en los problemas 3.1, 3.2 y 3.3 se presentan diversos ejemplos de aplicación), los métodos que siguen son más eficaces en lo que se refiere al número de operaciones requeridas para obtener el polinomio interpolador.

Métodos de Aitken y Neville

Ambos métodos obtienen el polinomio interpolador construyendo una sucesión de polinomios de grados crecientes que van interpolando cada vez en más abscisas.

Se basan en el principio siguiente:

Sea $C = \{x_0, x_1, \ldots, x_m\}$ y sean $S \subset C$ y $T \subset C$ tales que S y T tienen el mismo número de elementos y que éstos coinciden excepto dos (es decir, existen $x_i \in S$ y $x_j \in T$ tales que $x_i \notin T$ y $x_j \notin S$). Suponemos conocidos los polinomios p_S y p_T tales que $p_S(x_k) = f_k$ (si $x_k \in S$) y $p_T(x_k) = f_k$ (si $x_k \in T$). Entonces puede construirse un nuevo polinomio $p_{S \cup T}$ tal que $p_{S \cup T}(x_k) = f_k$, para todo $x_k \in S \cup T$, de la manera siguiente:

$$p_{S \cup T}(x) = \frac{(x_j - x)p_S(x) - (x_i - x)p_T(x)}{x_j - x_i} \ .$$

Los dos métodos que estudiamos construyen el polinomio interpolador gracias a sucesivas aplicaciones de la fórmula anterior, pero siguiendo un esquema diferente.

Método de Aitken. Se construyen polinomios $p_{i,j}(x)$ $(i = j \div m)$, de grado menor o igual que j, que interpolan en las abscisas $\{x_0, x_1, \ldots, x_{j-1}, x_i\}$ $(j = 0 \div m)$; se llega al polinomio $p_m(x) \equiv p_{m,m}(x)$, de grado menor o igual que m, que interpolará en $\{x_0, x_1, \ldots, x_m\}$:

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

Método de Neville. Se construyen polinomios $p_{i,j}(x)$ $(i = 0 \div m - j)$, de grado menor o igual que j, que interpolan en las abscisas $\{x_i, x_{i+1}, \ldots, x_{i+j}\}$ $(j = 0 \div m)$, se llega al polinomio $p_m(x) \equiv p_{0,m}(x)$, de grado menor o igual que m, que interpolará en $\{x_0, x_1, \ldots, x_m\}$:

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

Los esquemas de construcción para $m = 3$ utilizados por los dos métodos se encuentran en las tablas 3.2 i 3.3.

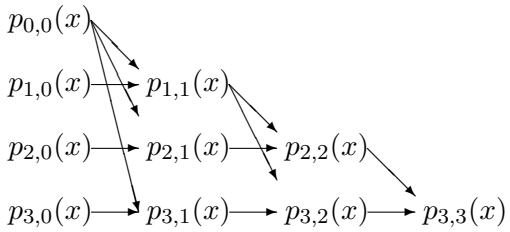


Tabla 3.2: Esquema del método de Aitken.

$$\begin{array}{llll} p_{0,0}(x) & \to p_{0,1}(x) & \to p_{0,2}(x) & \to p_{0,3}(x) \\ & \nearrow & \nearrow & \nearrow \\ p_{1,0}(x) & \to p_{1,1}(x) & \to p_{1,2}(x) & \\ & \nearrow & \nearrow & \\ p_{2,0}(x) & \to p_{2,1}(x) & & \\ & \nearrow & & \\ p_{3,0}(x) & & & \end{array}$$

Tabla 3.3: Esquema del método de Neville.

Aunque estos dos métodos sirven para encontrar el polinomio interpolador, en la práctica, se usan para evaluar directamente $p_m(x) \equiv p_{m,m}(x)$ para un valor concreto de x ya dado. En este caso, las fórmulas que describen los algoritmos respectivos representan expresiones aritméticas, no polinomiales, y los esquemas triangulares de construcción son numéricos, no polinomiales.

Método de las diferencias divididas de Newton

Expresamos ahora el polinomio interpolador de la manera siguiente:

$$\begin{aligned} p_m(x) &= c_0 + c_1(x-x_0) + c_2(x-x_0)(x-x_1) + \cdots \\ &\quad + c_m(x-x_0)(x-x_1)\cdots(x-x_{m-1}) \ . \end{aligned}$$

El método de las diferencias divididas nos permite calcular los coeficientes c_j $(j = 0 \div m)$ mediante la construcción de las llamadas *diferencias divididas*:

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

El esquema de construcción de las diferencias divididas de Newton se da en la tabla 3.4, para $m = 3$, y se puede encontrar, aplicado a diversos casos, en los problemas 3.1–3.4.

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

Tabla 3.4: Esquema de diferencias divididas.

Usando la definición de las diferencias divididas sobre $x \in [a, b]$, para $x \neq x_k$ $(k = 0 \div m)$, se llega, por inducción sobre m, a la fórmula siguiente:

$$\begin{aligned} f(x) &= f[x_0] + f[x_0, x_1](x-x_0) + f[x_0, x_1, x_2](x-x_0)(x-x_1) + \cdots \\ &\quad + f[x_0, x_1, \ldots, x_m](x-x_0)(x-x_1)\cdots(x-x_{m-1}) \\ &\quad + f[x_0, x_1, \ldots, x_m, x](x-x_0)(x-x_1)\cdots(x-x_{m-1})(x-x_m) \ . \end{aligned} \tag{3.4}$$

También, por inducción sobre m, se comprueba que el polinomio $p_m(x)$ de grado menor o igual que m viene también dado por la *fórmula de interpolación de Newton*

$$\begin{aligned} p_m(x) &= f[x_0] + f[x_0, x_1](x-x_0) \\ &\quad + f[x_0, x_1, x_2](x-x_0)(x-x_1) + \cdots \\ &\quad + f[x_0, x_1, \ldots, x_m](x-x_0)(x-x_1)\cdots(x-x_{m-1}) \ , \end{aligned} \tag{3.5}$$

que es, de hecho, el polinomio $p_{0,m}(x)$ construido por el método de Neville y, como consecuencia, es el polinomio interpolador de f en las abscisas $\{x_0, \ldots, x_m\}$; es decir que

$$c_j = f[x_0, x_1, \ldots, x_j] \quad (j = 0 \div m) \ .$$

El método de las diferencias divididas de Newton, además de ser un interesante procedimiento para el cálculo explícito del polinomio interpolador, tiene la ventaja siguiente, que comparte con los métodos de Aitken y de Neville: si se añaden más puntos de interpolación, puede aprovecharse todo el trabajo hecho. Para construir el nuevo polinomio interpolador sólo hace falta continuar el esquema de construcción de diferencias divididas y calcular los nuevos coeficientes c_{m+1}, c_{m+2}, ..., aprovechando así todos los cálculos previos.

Notamos también que c_j es el coeficiente que acompaña al término x^j del polinomio interpolador p_j de f en las abscisas x_0, x_1, ..., x_j y, dado que el polinomio interpolador no varía al permutar éstas, tampoco lo hace c_j; como consecuencia, la diferencia dividida $f[x_0, x_1, \ldots, x_j]$ es una función *simétrica* de x_0, x_1, ..., x_j $(j = 0 \div m)$. Por ejemplo: $f[x_0, x_1] = f[x_1, x_0]$, $f[x_0, x_1, x_2] = f[x_1, x_0, x_2]$, etc.

La fórmula (3.4) da una nueva expresión para el error de interpolación

$$\begin{aligned} & f(x) - p_m(x) \\ & \quad = \quad f[x_0, x_1, \ldots, x_m, x](x - x_0)(x - x_1) \cdots (x - x_{m-1})(x - x_m) \ . \end{aligned}$$

Si las abscisas de interpolación x_k $(k = 0 \div m)$ y la propia x pertenecen a (a, b) y $f \in \mathcal{C}^{m+1}(a, b)$, comparando la expresión anterior con (3.2), tenemos

$$f[x_0, x_1, \ldots, x_m, x] = \frac{f^{(m+1)}(\xi(x))}{(m+1)!} \ ,$$

donde $\xi(x)$ depende de x y pertenece a $< x_0, x_1, \ldots, x_m, x >$.

Cuando las abscisas en que se interpola son equidistantes (o sea, tales que las diferencias $x_{k+1} - x_k$ $(k = 0 \div m - 1)$ son todas iguales), la fórmula de interpolación de Newton admite una nueva expresión, usando operadores en diferencias finitas, que se presentará en el capítulo siguiente.

3.1.4 Interpolación de Taylor

Planteamiento del problema

Se trata de un nuevo tipo de interpolación, conceptualmente diferente del anterior y que se considerará más adelante como el caso límite de aquél cuando todas las abscisas de interpolación tienden a ser la misma.

Queremos ahora construir un polinomio que sea una buena aproximación de una función f, suficientemente derivable, en el entorno de una única abscisa x_0. A tal efecto, lo buscaremos de manera que no sólo interpole a f en x_0, sino también que sus derivadas interpolen a las derivadas de f en x_0, hasta un cierto orden n dado.

Más concretamente, dada una función f, suficientemente derivable en x_0, buscamos un polinomio $p(x)$, de grado menor o igual que N, tal que

$$p^{(j)}(x_0) = f^{(j)}(x_0) \ (j = 0 \div n) \ ,$$

donde $f^{(j)}$ y $p^{(j)}$ denotan las derivadas j-ésimas de las funciones f y p, respectivamente, si $j \geq 1$, y donde tomamos $f^{(0)} = f$ y $p^{(0)} = p$. Éste es el *problema de interpolación de Taylor*.

Existencia, unicidad y error

Para estudiar la existencia, la unicidad y el error del polinomio interpolador, usaremos el *teorema fundamental del cálculo*:

- Si $f \in \mathcal{C}^1(a,b)$ y x, $x_0 \in (a,b)$, entonces

$$f(x) - f(x_0) = \int_{x_0}^{x} f'(s)\, ds \ . \tag{3.6}$$

La fórmula (3.6) nos resuelve el problema de interpolación de Taylor de grado 0 de f en x_0, siendo $p_0(x) = f(x_0)$ la única solución.

Para $f \in \mathcal{C}^{n+1}(a,b)$ y $N = n$, haciendo sucesivas integraciones por partes en (3.6), se obtiene la solución del problema de interpolación de Taylor, dada por la *fórmula de interpolación de Taylor*

$$\begin{aligned} p_n(x) &= f(x_0) + f'(x_0)(x-x_0) + \frac{f^{(2)}(x_0)}{2}(x-x_0)^2 + \cdots \\ &\quad + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \ , \end{aligned} \tag{3.7}$$

juntamente con la *expresión integral del error de la interpolación de Taylor*

$$\begin{aligned} R_n(x) &= f(x) - p_n(x) = \frac{1}{n!}\int_{x_0}^{x} (x-s)^n f^{(n+1)}(s) ds \\ &= \frac{(x-x_0)^{n+1}}{n!}\int_0^1 (1-t)^n f^{(n+1)}(x_0 + t(x-x_0)) dt \ , \end{aligned} \tag{3.8}$$

para x_0, $x \in (a,b)$. Esta última fórmula asegura la unicidad de $p_n(x)$.

El polinomio $p_n(x)$ solución recibe el nombre de *polinomio (interpolador) de Taylor de grado menor igual que n de f en* x_0.

Dado que $(x-s)^n$, considerada como función de s, no cambia de signo en $< x_0, x >$, puede usarse el *teorema del valor medio para integrales*:

- Si g_1 y g_2 son funciones reales definidas en $[0,1]$, tales que g_1 es continua y g_2 no cambia de signo, existe algún $\tau \in [0,1]$ de manera que se cumple

$$\int_0^1 g_1(t) g_2(t) dt = g_1(\tau) \int_0^1 g_2(t) dt \ .$$

Así, podemos obtener la *expresión de Lagrange del error de interpolación de Taylor*, equivalente a (3.8),

$$R_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)^{n+1} \ , \tag{3.9}$$

donde $\xi(x) \in < x_0, x >$ es generalmente desconocido.

OBSERVACIÓN IMPORTANTE

El polinomio interpolador de Taylor puede ser considerado como un límite de polinomios interpoladores en abscisas diferentes. Así, si consideramos el polinomio interpolador $p_m(x) = p_m(x; x_0, \ldots, x_m)$ en $m+1$ abscisas diferentes x_0, x_1, ..., x_m, escrito según la fórmula de interpolación de Newton (3.6)

$$\begin{aligned} p_m(x; x_0, x_1, \ldots, x_m) &= f[x_0] + f[x_0, x_1](x - x_0) \\ &+ f[x_0, x_1, x_2](x - x_0)(x - x_1) + \cdots \\ &+ f[x_0, x_1, \ldots, x_m](x - x_0)(x - x_1) \cdots (x - x_{m-1}) \ , \end{aligned}$$

resulta que el error de interpolación en abscisas diferentes de la fórmula (3.2) tiende a la expresión (3.7), cuando x_k tiende a x_0 ($k = 1 \div m$). Por lo tanto, si definimos

$$p_m(x; x_0, x_0, \ldots, x_0) = \lim_{x_k \to x_0} p_m(x; x_0, x_1, \ldots, x_n) \ ,$$

resulta que $p_m(x; x_0, x_0, \ldots, x_0) = p_n(x)$, donde $p_n(x)$ denota el polinomio interpolador de Taylor de grado menor o igual que $n = m$. Por eso, definiremos también

$$f[x_0, x_0, \ldots, x_0] \equiv \lim_{x_k \to x_0} f[x_0, x_1, \ldots, x_j] = \frac{f^{(j)}(x_0)}{j!} \ .$$

EJEMPLOS

En la tabla 3.5 se dan ejemplos de polinomios de Taylor en $x_0 = 0$, con indicación de los restos R_n respectivos, para diversas funciones definidas sobre los intervalos (a, b) que se reseñan. Es necesario hacer dos observaciones sobre la notación usada en la tabla:

- ξ representa una abscisa en el intervalo $< 0, x >$,
- se ha utilizado
$$\left(\begin{array}{c} \alpha \\ j \end{array} \right) = \frac{\alpha(\alpha - 1) \cdots (\alpha - j + 1)}{j!} \ .$$

En el problema 3.7 puede encontrarse la deducción para algún caso de la tabla 3.5. Otros ejemplos se encuentran en los problemas 3.8 y 3.9.

Desarrollo de Taylor

Mirando la expresión (3.9), nos damos cuenta de que $(x - x_0)^{n+1}$ es un factor de $R_n(x)$ que va multiplicado por $\frac{f^{(n+1)}(\xi(x))}{(n+1)!}$ y que éste está acotado en cualquier entorno cerrado de x_0 contenido en (a, b), si $f \in \mathcal{C}^{n+1}(a, b)$.

Con el fin de expresar de otra manera este hecho, introduciremos las notaciones siguientes:

Dadas f y g, funciones definidas cerca de x_0, diremos que f es *del orden de g en x_0* si existen un entorno I de x_0 y una constante $C \geq 0$ tales que

$$| f(x) | \leq C \, | g(x) | \ ,$$

(a,b)	$f(x) = p_n(x) + R_n(x)$
$(-1,1)$	$(1+x)^{-1} = 1 - x + \cdots + (-1)^n x^n + (-1)^{n+1}\frac{x^{n+1}}{1-x}$
$(-1,1)$	$(1+x)^\alpha = 1 + \alpha x + \cdots + \begin{pmatrix} \alpha \\ n \end{pmatrix} x^n + \frac{\begin{pmatrix} \alpha \\ n+1 \end{pmatrix} x^{n+1}}{(1+\xi)^{n+1-\alpha}}$
$(-\infty,\infty)$	$e^x = 1 + x + \cdots + \frac{x^n}{n!} + e^\xi \frac{x^{n+1}}{(n+1)!}$
$(-\infty,\infty)$	$\operatorname{sen} x = x - \frac{x^3}{3!} + \cdots \pm \frac{x^n}{n!} \mp \cos\xi \frac{x^{n+2}}{(n+2)!}$ (n senar)
$(-\infty,\infty)$	$\cos x = 1 - \frac{x^2}{2!} + \cdots \pm \frac{x^n}{n!} \mp \cos\xi \frac{x^{n+2}}{(n+2)!}$ (n parell)
$(-1,1)$	$\ln(1+x) = x - \frac{x^2}{2} + \cdots + (-1)^{n+1}\frac{x^n}{n} + (1+\xi)^{-1}\frac{x^{n+1}}{n+1}$

Tabla 3.5: Polinomios interpoladores de Taylor y sus restos.

para todo $x \in I$. En símbolos, escribiremos

$$f(x) = \mathcal{O}(g(x)) \ (x \to x_0) \ .$$

Diremos también que f es *de orden menor que g en x_0* cuando

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = 0 \ .$$

En símbolos, escribiremos

$$f(x) = o(g(x)) \ \ (x \to x_0) \ .$$

En nuestro caso, si $f \in \mathcal{C}^{n+1}(a,b)$, podemos escribir

$$R_n(x) = \mathcal{O}((x-x_0)^{n+1}) \ \ (x \to x_0) \ , \tag{3.10}$$

$$R_n(x) = o((x-x_0)^n) \ \ (x \to x_0) \ ; \tag{3.11}$$

estas expresiones indican que, dado $n \geq 0$ y un número ϵ (que será la cota del error permitido en el cálculo de $f(x)$), podemos encontrar un entorno $I = (x_0 - \delta, x_0 + \delta)$ de x_0 donde el error cometido al considerar $p_n(x)$ en lugar de $f(x)$ es menor que ϵ. Observamos que, sobre este intervalo I, puede ser mucho más económico evaluar $p_n(x)$ (sumas y productos) que evaluar $f(x)$.

Por el hecho de cumplirse la relación (3.11), se dice que los polinomios interpoladores de Taylor forman un *desarrollo asintótico, cuando $x \to x_0$*, concepto que será tratado en el apartado 4.3.1.

En los ejemplos de polinomios interpoladores de Taylor dados en la tabla 3.5 se observa que

$$\lim_{n\to\infty} R_n(x) = 0 \quad \forall x \in (a,b) \ .$$

Esto quiere decir que, fijado $x \in (a,b)$, podemos obtener una aproximación de $f(x)$ por un polinomio interpolador de Taylor $p_n(x)$ tan buena como queramos, tomando n suficientemente grande. Las funciones con esta propiedad reciben el nombre de *funciones analíticas* en el intervalo (a,b).

Estas funciones satisfacen

$$f(x) = \lim_{n\to\infty} \sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!}(x-x_0)^j \quad \forall x \in (a,b) \ ,$$

que también suele escribirse como

$$f(x) = \sum_{j=0}^{\infty} \frac{f^{(j)}(x_0)}{j!}(x-x_0)^j \quad \forall x \in (a,b) \ .$$

La sucesión de estos polinomios se llama también *desarrollo en serie de Taylor* de f en un entorno (a,b) de x_0. Este concepto de serie será estudiado en el apartado 4.3.1.

Siguiendo esta terminología, el polinomio de Taylor de grado menor o igual que n recibe también el nombre de *desarrollo de Taylor de orden n*.

Es necesario observar que no todas las funciones indefinidamente derivables en un punto son analíticas. Por ejemplo, la función

$$f(x) = \begin{cases} e^{-\frac{1}{x^2}} & (x \neq 0) \\ 0 & (x = 0) \end{cases}$$

es indefinidamente derivable, pero satisface $f^{(j)}(0) = 0$ $(j \geq 0)$, con lo cual $R_n(x) = f(x)$ $\forall x \in \mathbb{R}$ $(n \geq 0)$ y, por lo tanto, no es analítica en $x_0 = 0$.

Finalmente, destacamos un criterio muy elemental para acotar $R_n(x)$, el *criterio de alternancia de los restos*:

- Si $R_n(x)R_{n+1}(x) < 0$, entonces $f(x) \in < p_n(x), p_{n+1}(x) >$ y, por lo tanto,

$$|\, R_n(x)\,| \leq \frac{|\, f^{(n+1)}(x_0)\,|}{(n+1)!}|\, x - x_0\,|^{n+1} \ ;$$

  esto es, el error es menor que el primer término despreciado.

3.1.5 Interpolación de Hermite

Planteamiento del problema

Sean (x_k, f_k) $(k = 0 \div m)$ $m+1$ puntos conocidos de la gráfica de $f(x)$ y supongamos también conocidas sus pendientes $f'_k \equiv f'(x_k)$ $(k = 0 \div m)$. Se plantea la cuestión siguiente:

¿existe algún polinomio p de grado menor o igual que $N = 2m+1$ (ya que con $2m+2$ condiciones podemos aspirar a determinar $2m+2$ coeficientes) tal que

$$p(x_k) = f_k \ , \quad p'(x_k) = f'_k \quad (k = 0 \div m) \ ?$$

Este problema de interpolación, en el que se busca un polinomio cuya gráfica no sólo pase por unos puntos dados, sino que también pase por ellos con unas pendientes determinadas, se llama *problema de interpolación de Hermite.*

La respuesta a la cuestión planteada anteriormente es que existe un único polinomio con las características exigidas; se llama *polinomio interpolador de Hermite* y lo denotaremos por $p_{2m+1}(x)$.

Expresión del polinomio interpolador de Hermite y de su error

Dicho polinomio admite una expresión explícita análoga a la dada por el método de Lagrange en su caso,

$$p_{2m+1}(x) = \sum_{i=0}^{m} f_i \Phi_i(x) + \sum_{i=0}^{m} f'_i \Psi_i(x) \ ,$$

con

$$\begin{aligned} \Phi_i(x) &= (1 - 2l'_i(x_i)(x - x_i))l_i^2(x) \ , \\ \Psi_i(x) &= (x - x_i)l_i^2(x) \quad (i = 0 \div m) \ , \end{aligned}$$

donde $l_i(x)$ $(i = 0 \div m)$ son los polinomios de Lagrange que aparecen en (3.3).

Los polinomios Φ_i, Ψ_i $(i = 0 \div m)$ se llaman *polinomios básicos de la interpolación de Hermite* porque cumplen:

$$\Phi_i(x_k) = \delta_{ik} \ , \quad \Phi'_i(x_k) = 0 \quad (i, k = 0 \div m) \ ;$$

$$\Psi_i(x_k) = 0 \ , \quad \Psi'_i(x_k) = \delta_{ik} \quad (i, k = 0 \div m) \ .$$

Con el fin de valorar el grado de aproximación del polinomio a la función, se dispone de una expresión del error en el caso de que f sea $2m+2$ veces diferenciable con continuidad sobre un intervalo I tal que $x_k \in I$ $(k = 0 \div m)$:

$$f(x) - p_{2m+1}(x) = \frac{f^{(2m+2)}(\xi(x))}{(2m+2)!}(x - x_0)^2(x - x_1)^2 \cdots (x - x_m)^2$$

para $x \in I$ y donde $\xi(x) \in < x_0, x_1, \ldots, x_m, x >$.

Interpolación de Hermite generalizada

La interpolación de Hermite puede generalizarse al caso en que conocemos la función f en las abscisas x_k $(k = 0 \div m)$ con más "profundidad", aunque esta "profundidad" no sea la misma en todas ellas.

Así, dadas estas abscisas, diferentes dos a dos, y derivadas sucesivas de la función f en ellas

$$x_k \quad , \quad f^{(j)}(x_k) \quad (j = 0 \div n_k) \quad (k = 0 \div m) \ ,$$

existe un único polinomio p_N de grado menor o igual que

$$N = \sum_{k=0}^{m} n_k + m$$

tal que

$$p_N^{(j)}(x_k) = f^{(j)}(x_k) \quad (j = 0 \div n_k) \ (k = 0 \div m) \ .$$

Por lo que se refiere al error en este caso, puede decirse que, si $f \in \mathcal{C}^{N+1}(a,b)$ y $x_k \in (a,b)$ $(k = 0 \div m)$, se cumple

$$f(x) - p_N(x) = \frac{f^{(N+1)}(\xi(x))}{(N+1)!}(x - x_0)^{n_0+1}(x - x_1)^{n_1+1} \cdots (x - x_m)^{n_m+1}$$

para $x \in (a,b)$ y $\xi(x) \in < x_0, x_1, \ldots, x_m, x >$ (véase el problema 3.13).

Método de las diferencias divididas generalizadas

A pesar del innegable interés teórico de la expresión que se ha dado para el polinomio interpolador de Hermite, en la práctica y atendiendo a razones de comodidad, se usa para su cálculo una generalización del método de las diferencias divididas de Newton en la que el esquema triangular se construye de la manera siguiente:

1. En la primera columna se colocan los valores $f[x_i] = f_i$ $(i = 0 \div m)$ de manera que cada uno aparezca dos veces consecutivas: $f[x_0]$, $f[x_0]$, $f[x_1]$, $f[x_1]$, ..., $f[x_m]$, $f[x_m]$.

2. Al construir la segunda columna, la aparición de abscisas repetidas no nos permitirá usar la definición habitual de diferencia dividida para el cálculo de $f[x_i, x_i]$ $(i = 0 \div m)$. En este caso, usaremos el hecho de que:

$$f[x_i, x_i] \equiv \lim_{x_k \to x_i} f[x_i, x_k] = f'(x_i) = f_i' \ (i = 0 \div m) \ .$$

3. La construcción de las demás columnas no comporta ya más problemas. Así, obtendremos un esquema triangular como el de la tabla 3.6, donde las cantidades subrayadas son datos.

   El polinomio interpolador de Hermite será entonces

$$\begin{aligned} p(x) = {} & f[x_0] + f[x_0, x_0](x - x_0) + f[x_0, x_0, x_1](x - x_0)^2 \\ & + f[x_0, x_0, x_1, x_1](x - x_0)^2(x - x_1) + \cdots \\ & + f[x_0, x_0, \ldots, x_m, x_m](x - x_0)^2 \cdots (x - x_{m-1})^2(x - x_m) \ . \end{aligned}$$

Finalmente, es necesario decir que las diferencias divididas generalizadas pueden extenderse también para ser aplicadas al cálculo del polinomio de interpolación de Hermite generalizada (véase el problema 3.12).

Para ilustrar el proceso que se acaba de describir damos el ejemplo siguiente.

EJEMPLO

$$
\begin{array}{c|c|c|c|c|c}
x_0 & \underline{f[x_0]=f_0} & & & & \\
 & & \underline{f[x_0,x_0]=f_0'} & & & \\
x_0 & \underline{f[x_0]=f_0} & & f[x_0,x_0,x_1] & & \\
 & & f[x_0,x_1] & & f[x_0,x_0,x_1,x_1] & \cdots \\
x_1 & \underline{f[x_1]=f_1} & & f[x_0,x_1,x_1] & \vdots & \\
 & & \underline{f[x_1,x_1]=f_1'} & \vdots & & \\
x_1 & \underline{f[x_1]=f_1} & \vdots & & & \\
\vdots & \vdots & & & &
\end{array}
$$

Tabla 3.6: Diferencias divididas generalizadas.

Buscamos un polinomio de tercer grado cuya gráfica pase por los puntos $(1,3)$ y $(2,2)$ con pendientes respectivas 1 y 4.

En este caso: $m=1$, $x_0=1$, $x_1=2$, $f_0=3$, $f_1=2$, $f_0'=1$ y $f_1'=4$. Tendremos el esquema de la tabla 3.7 y el polinomio interpolador de Hermite

$$
\begin{aligned}
p(x) &= 3+1(x-1)-2(x-1)^2+7(x-1)^2(x-2) \\
&= 7x^3-30x^2+40x-14\ .
\end{aligned}
$$

$$
\begin{array}{c|c|c|c|c}
x_0=1 & \underline{f[x_0]=3} & & & \\
 & & \underline{f[x_0,x_0]=1} & & \\
x_0=1 & \underline{f[x_0]=3} & & f[x_0,x_0,x_1]=-2 & \\
 & & f[x_0,x_1]=-1 & & f[x_0,x_0,x_1,x_1]=7 \\
x_1=2 & \underline{f[x_1]=2} & & f[x_0,x_1,x_1]=5 & \\
 & & \underline{f[x_1,x_1]=4} & & \\
x_1=2 & \underline{f[x_1]=2} & & &
\end{array}
$$

Tabla 3.7: Ejemplo de diferencias divididas generalizadas.

3.2 APROXIMACIÓN DE FUNCIONES

3.2.1 Introducción al problema general de aproximación

Todo el cálculo aproximado de funciones desarrollado en la sección anterior ha estado basado en la interpolación como herramienta fundamental; destaquemos las ventajas siguientes:

- El polinomio de interpolación es fácil de calcular y se dispone de una fórmula explícita para el error de interpolación que permite estimar los errores cometidos.
- Es muy útil para generar fórmulas de derivación, integración, etc. de funciones, tal como se verá en el capítulo siguiente.

- Es especialmente apropiada para el cálculo con funciones dadas por tablas; es decir, para funciones bien conocidas sobre conjuntos discretos de abscisas, quizás equiespaciadas, donde el error de redondeo de los valores de las funciones es menor que el error propio de interpolación.

Ahora bien, el proceso de interpolación presenta diversos problemas en otros casos de cálculo de funciones, sobre todo cuando nos encontramos en situaciones totalmente diferentes de las del párrafo anterior:

- Si tenemos un conjunto discreto de valores (x_k, y_k) $(k = 0 \div m)$ que tienen errores de redondeo apreciables, no es conveniente encontrar el polinomio de grado menor o igual que m que interpole exactamente todos estos valores, ya que el carácter "oscilante" (con diferentes máximos y mínimos) del error de la interpolación puede provocar que el polinomio interpolador sea muy diferente de la función interpolada fuera de las abscisas de interpolación. Para disminuir el efecto de los errores en las $m+1$ abscisas, se puede intentar hallar un polinomio p_n de grado menor o igual n, con $n \leq m$, tal que los errores

$$e_k = y_k - p_n(x_k) \ (k = 0 \div m)$$

sean nulos. Si $n < m$ estamos ante un problema *sobredeterminado*; es decir, con más ecuaciones que incógnitas y, por lo tanto, sin solución en general. De hecho, lo modificaremos imponiendo que los errores e_k sean tan pequeños como sea posible en algún sentido que determinaremos más adelante. Este proceso recibe el nombre de *aproximación polinomial*. Si, por diversas razones, quisiéramos "aproximar" los datos no necesariamente por un polinomio sino por una función f_n de un tipo dado como

$$f_n(x) = a_0\varphi_0(x) + \cdots + a_n\varphi_n(x) \ ,$$

nos encontraríamos ante un problema totalmente similar, si $n < m$: encontrar los parámetros $a_0, \ldots, a_n$ de manera que los errores

$$e_k \equiv e_n(x_k) = y_k - f_n(x_k) \quad (k = 0 \div m)$$

sean tan pequeños, en algún sentido, como se quiera.

- Un caso opuesto al anterior se da cuando tenemos una función f que se conoce para todo x de un intervalo I y queremos tener una manera eficiente de calcularla. El procedimento estándar de generación de tablas y interpolación no es adecuado en una calculadora o en un ordenador, por razones de memoria y eficiencia de cálculo, y conviene disponer de otra expresión calculable (un polinomio, una función racional, etc.) o expresiones f_n de manera que la *función error de aproximación*

$$e_n(x) = f(x) - f_n(x)$$

sea suficientemente pequeña sobre I. Si $f_n(x) = p_n(x)$ (polinomio de grado menor o igual que n) nos encontramos de nuevo con el problema de aproximación polinomial. En general, la aproximación buscada tiene la forma

$$f_n(x) = F_n(a_0, \ldots, a_n, \varphi_0(x), \ldots, \varphi_n(x)) \ ,$$

pero muy frecuentemente se toma

$$f_n(x) = a_0\varphi_0(x) + \cdots + a_n\varphi_n(x) \ ,$$

donde φ_0, ..., φ_n son funciones dadas, fácilmente calculables. El problema se reduce entonces a encontrar los parámetros a_0, ..., a_n para que la función error de aproximación $e_n(x)$ sea tan pequeña como se pueda en algún sentido determinado a priori.

Podemos ahora enunciar el *problema general de aproximación*: dado un conjunto I de *abscisas de aproximación* y unas *funciones básicas* φ_j $(j = 0 \div n)$ definidas sobre I; para cada función f definida también sobre I, hace falta buscar

$$f_n(x) = a_0\varphi_0(x) + \cdots + a_n\varphi_n(x)$$

de manera que la magnitud del error de aproximación $e_n(x) = f(x) - f_n(x)$ sea lo más pequeña posible.

Así, para tener totalmente determinado un problema de aproximación, es necesario especificar el conjunto I de abscisas de aproximación, las funciones básicas y la forma de medir la magnitud de las funciones de error.

El conjunto I de abscisas de aproximación

Si I es finito ($I = \{x_0, \ldots, x_m\}$), hablaremos de *aproximación discreta*; y, si I es un intervalo de extremos a y b $(a < b)$, hablaremos de *aproximación continua*. Dar una función f sobre un conjunto finito $I = \{x_0, \ldots, x_m\}$ equivale a dar $y_k \equiv f(x_k)$ $(k = 0 \div m)$.

Las funciones básicas

Las funciones $\varphi_0, \ldots, \varphi_n$, definidas sobre I, pueden escogerse de diversas maneras. El tipo de funciones escogidas dependerá, como en el caso de la interpolación, de las "sospechas" que tengamos respecto al comportamiento de f. Si f expresa algún fenómeno periódico, convendrá escoger entre las funciones trigonométricas: por ejemplo,

$$\varphi_0(x) = 1 \ , \quad \varphi_1(x) = \operatorname{sen} x \ , \quad \varphi_2(x) = \cos x \ , \quad \ldots \ , \quad \varphi_{2s}(x) = \cos 2sx \ ,$$

donde $n = 2s$ y hablaremos de *aproximación trigonométrica*. Si f responde a un comportamiento polinomial, escogeremos cada $\varphi_j(x) = p_j(x)$ entre los polinomios de grado j $(j = 0 \div n)$ (tomando, por ejemplo, $\varphi_j(x) = x^j$, aunque esta elección no será siempre la más adecuada) y hablaremos de *aproximación polinomial*. El conjunto $\mathcal{F}_n$ de funciones f_n que pueden ponerse como combinación lineal de las funciones básicas φ_0, ..., φ_n

$$f_n(x) = \sum_{j=0}^{n} a_j\varphi_j(x) \ , \quad x \in I \ ,$$

se llama también *espacio vectorial generado por las funciones* $\varphi_0, \ldots, \varphi_n$. Otros tipos de aproximaciones, que no son lineales en los parámetros a_0, ..., a_n, como son la racional y la exponencial, no serán tratados aquí.

Medida de la magnitud de las funciones error: normas funcionales

La magnitud de la función error de aproximación $e_n(x) = f(x) - f_n(x)$ en I puede también ser considerada de diferentes maneras.

En el caso discreto, e_n consta de $m+1$ valores

$$e_k \equiv e_n(x_k) = f(x_k) - f_n(x_k) \ (k = 0 \div m) \ ,$$

y medir e_n equivale a medir el vector $(m+1)$-dimensional de componentes e_k $(k = 0 \div m)$, por medio de las normas introducidas en el apartado 2.1.2. Las dos normas más usadas para medir la aproximación son:

- la *norma euclídea*
$$\|e\|_2 = \left(\sum_{k=0}^{m} |e_k|^2 \right)^{\frac{1}{2}} \ ,$$
- la *norma del máximo*
$$\|e\|_\infty = \max_{k=0 \div m} |e_k| \ .$$

Cuando quiere darse una importancia diferente a los errores e_k, se introducen unos *pesos* positivos w_k $(k = 0 \div m)$ y pueden definirse la *norma euclídea* y la *norma del máximo ponderadas*, respecto a esta colección de pesos $w = \{w_k\}_{k=0 \div m}$:

$$\|e\|_{2,w} = \left(\sum_{k=0}^{m} w_k \, |e_k|^2 \right)^{\frac{1}{2}} \ , \quad \|e\|_{\infty,w} = \max_{k=0 \div m} w_k \, |e_k| \ .$$

En el caso continuo, suponemos $I = [a, b]$ y, para cada función e definida en el intervalo I, definimos:

- la *norma euclídea*
$$\|e\|_2 = \left(\int_a^b |e(x)|^2 \, dx \right)^{\frac{1}{2}} \ ,$$
- la *norma del máximo*
$$\|e\|_\infty = \max_{x \in I} |e(x)| \ .$$

Se puede comprobar que las definiciones anteriores cumplen las propiedades de norma sobre el conjunto $\mathcal{C}([a, b])$ de funciones continuas en el intervalo $[a, b]$.

Análogamente al caso discreto, dada una *función peso* $w \in \mathcal{C}([a, b])$ tal que $w(x) > 0$ sobre I, pueden definirse:

- la *norma euclídea ponderada*
$$\|e\|_{2,w} = \left(\int_a^b w(x) \, |e(x)|^2 \right)^{\frac{1}{2}} \ ,$$

- la *norma del máximo ponderada*

$$\|e\|_{\infty,w} = \max_{x\in I} |e(x)|\, w(x) \ .$$

También se admitirán algunas *funciones peso* e *intervalos singulares* como pueden ser: $w(x) = \frac{1}{\sqrt{1-x^2}}$ en $[-1,1]$, $w(x) = e^{-x}$ en $[0,\infty)$ y $w(x) = e^{-x^2}$ en $\mathbb{R}$.

Tanto en el caso discreto como en el continuo, hablaremos de *aproximación por mínimos cuadrados* cuando la elección de los parámetros se haga para minimizar una norma euclídea. En el caso de minimización de normas del máximo, hablaremos de *aproximación minimax*.

3.2.2 Aproximación por mínimos cuadrados

Definición del problema

Consideramos un conjunto I de abscisas de aproximación y unas funciones básicas φ_j $(j = 0 \div n)$. Para cada función f definida sobre I, buscamos $f_n^* \in \mathcal{F}_n$ (el espacio vectorial generado por aquellas) de manera que $\|f - f_n^*\|_2$ sea mínima en $\mathcal{F}_n$; es decir,

$$\|f - f_n^*\|_2 = \min_{f_n\in\mathcal{F}_n} \|f - f_n\|_2 \ ,$$

donde $\|\ \ \|_2 \equiv \|\ \ \|_{2,w}$ representa aquí cualquiera de las normas euclídeas:

a) en el caso discreto, $I = \{x_0, \ldots, x_m\}$ y, si $e = (e_0, \ldots, e_m)$,

$$\|e\|_{2,w} = \left(\sum_{k=0}^{m} w_k\, |e_k|^2\right)^{\frac{1}{2}} \ ,$$

donde $w = \{w_0, \ldots w_m\}$ es una colección de pesos positivos;

b) en el caso continuo, I es un intervalo de la recta real de extremos a y b,

$$\|e\|_{2,w} = \left(\int_a^b w(x)\, |e(x)|^2\, dx\right)^{\frac{1}{2}} \ ,$$

donde $w(x) > 0$ es una función peso sobre I.

Productos escalares asociados

La propiedad fundamental de las normas euclídeas es que provienen de sendos *productos escalares*:

a) en el caso discreto:

$$(u, v) \equiv \sum_{k=0}^{m} w_k u_k v_k \ , \tag{3.12}$$

b) en el caso continuo:

$$(u, v) \equiv \int_a^b w(x)u(x)v(x)dx \ ; \tag{3.13}$$

en el sentido que se cumple, en ambos casos,

$$\|e\|_2^2 = (e, e) \ .$$

Dichos productos escalares satisfacen las propiedades de definición:
i) $(u,u) \geq 0$ y $(u,u) = 0$ si y sólo si $u = 0$;
ii) $(u,v) = (v,u)$;
iii) $(a_1u_1 + a_2u_2, v) = a_1(u_1,v) + a_2(u_2,v)$, para funciones u_1, u_2, u, v sobre I y números reales a_1, a_2 cualesquiera.

La introducción de estos productos escalares proporciona un procedimiento geométrico de resolución del problema que vendrá sugerido por el siguiente caso particular.

Caso particular e interpretación geométrica

Si $I = \{x_0, x_1, x_2\}$, una función real u sobre I viene dada por tres valores reales $u = (u_0, u_1, u_2)$, donde $u_k = u(x_k)$ $(k = 0, 1, 2)$; si además $w_0 = w_1 = w_2 = 1$, dadas dos funciones u, v sobre I, su producto escalar coincide con el producto escalar habitual sobre $\mathbb{R}^3$

$$(u,v) = u_0v_0 + u_1v_1 + u_2v_2 \ .$$

Identificaremos, de ahora en adelante, las funciones sobre I con vectores de $\mathbb{R}^3$.

Consideramos como espacio de aproximación el subespacio $\mathcal{V}_1$ generado por dos vectores $v^{(0)}, v^{(1)}$ de $\mathbb{R}^3$, supuestos linealmente independientes,

$$\mathcal{V}_1 = \{a_0v^{(0)} + a_1v^{(1)},\ a_0, a_1 \in \mathbb{R}\} \ ;$$

Dado un vector $y \in \mathbb{R}^3$, el problema de aproximación por mínimos cuadrados se reduce a encontrar $v^* = a_0^*v^{(0)} + a_1^*v^{(1)}$ tal que

$$\|y - v^*\|_2 = \min_{v \in \mathcal{V}_1} \|y - v\|_2 = \min_{a_0, a_1 \in \mathbb{R}} \|y - a_0v^{(0)} - a_1v^{(1)}\|_2 \ .$$

La representación gráfica de la figura 3.1 nos indica que v^* viene caracterizado por la *propiedad de proyección ortogonal* que asegura que $y - v^*$ es un *vector ortogonal* a todo elemento de $\mathcal{V}_1$; esto es,

$$(y - v^*, v) = 0 \quad \forall v \in \mathcal{V}_1 \ .$$

En efecto, por el teorema de Pitágoras, al ser $y - v^*$ ortogonal a $v - v^* \in \mathcal{V}_1$,

$$\|y - v\|_2^2 = \|y - v^*\|_2^2 + \|v - v^*\|_2^2 \geq \|y - v^*\|_2^2 \ ,$$

y, por lo tanto, si $(y - v^*, v) = 0\ \forall v \in \mathcal{V}_1$,

$$\|y - v^*\|_2 = \min_{v \in \mathcal{V}_1} \|y - v\|_2 \ .$$

Ecuaciones normales

Volviendo a nuestro problema general, sea f_n^* una función sobre I tal que

$$(f - f_n^*, f_n) = 0 \quad \forall f_n \in \mathcal{F}_n, \tag{3.14}$$

tenemos

$$\begin{aligned} \|f - f_n\|_2^2 &= (f - f_n, f - f_n) = (f - f_n^* + f_n^* - f_n, f - f_n^* + f_n^* - f_n) \\ &= (f - f_n^*, f - f_n^*) + 2(f - f_n^*, f_n^* - f_n) + (f_n^* - f_n, f_n^* - f_n) \\ &= \|f - f_n^*\|_2^2 + \|f_n^* - f_n\|_2^2 \end{aligned}$$

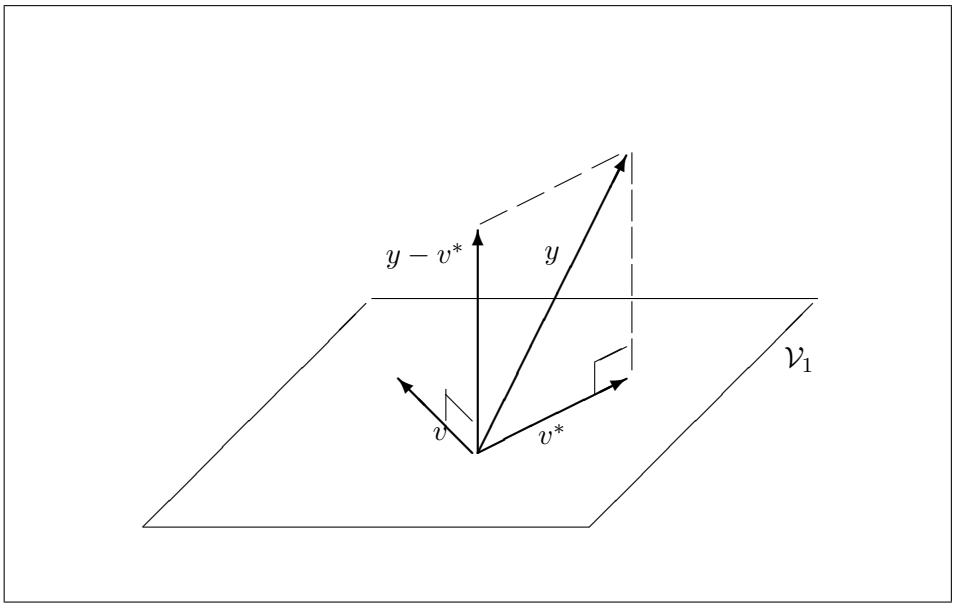


Figura 3.1: Proyección ortogonal de y sobre $\mathcal{V}_1$.

y, por lo tanto, $\|f - f_n\|_2 > \|f - f_n^*\|_2$, si $f_n \neq f_n^*$; es decir, f_n^* es la única función de $\mathcal{F}_n$ que satisface la condición de aproximación por mínimos cuadrados

$$\|f - f_n^*\|_2 = \min_{f_n \in \mathcal{F}_n} \|f - f_n\|_2 \ .$$

Dado que $\mathcal{F}_n$ está generado por φ_i $(i = 0 \div n)$, la condición (3.14) equivale al hecho de que los coeficientes a_j^* $(j = 0 \div n)$ de f_n^* en

$$f_n^*(x) = \sum_{j=0}^{n} a_j^* \varphi_j(x)$$

satisfacen las llamadas *ecuaciones normales*:

$$\sum_{j=0}^{n} (\varphi_i, \varphi_j) a_j^* = (\varphi_i, f) \quad (i = 0 \div n) \ . \tag{3.15}$$

Este sistema puede escribirse, en forma matricial, como

$$Aa^* = b \ ,$$

donde $A = ((\varphi_i, \varphi_j))_{i,j=0\div n}$, $a^* = (a_j^*)_{j=0\div n}$ y $b = ((\varphi_i, f))_{i=0\div n}$.

La matriz A es *semidefinida positiva*; esto es, simétrica y para cualquier vector $a = (a_0 \ \ a_1 \ \ \ldots \ \ a_n)^\top$,

$$a^\top A a = \left(\sum_{j=0}^{n} a_j \varphi_j, \sum_{l=0}^{n} a_l \varphi_l \right) = \| \sum_{j=0}^{n} a_j \varphi_j \|_2^2 = \|f_n\|_2^2 \geq 0 \ ,$$

donde f_n viene dada por

$$f_n(x) = \sum_{j=0}^{n} a_j \varphi_j(x) \ .$$

Esta misma relación nos demuestra que, si las funciones φ_j $(j = 0 \div n)$ son linealmente independientes, $\det A \neq 0$ y viceversa: las ecuaciones normales (3.15) tienen solución única para cualquier f si y sólo si las funciones φ_j $(j = 0 \div n)$ son linealmente independientes.

Ejemplo: recta de regresión

Tenemos un conjunto de datos (x_k, y_k) $(k = 0 \div m)$, con $m > 2$, correspondientes a dos variables x, y y sospechamos que satisfacen una relación lineal del tipo $y = f_1(x) = a_0 + a_1 x$, bien porque la representación gráfica de la nube de puntos en el plano así nos lo sugiere, bien por otras consideraciones aportadas por el contexto del problema real tratado.

A causa de los errores en los datos o porque el modelo lineal $y = a_0 + a_1 x$ no se ajusta suficientemente a la nube de puntos, no encontraremos una recta que una todos los $m+1$ puntos, y nos tendremos que conformar con la recta que pase "más cerca de todos los puntos".

En la aproximación por mínimos cuadrados, se impone que sea mínima la suma de los cuadrados de las *desviaciones* $d_k = y_k - a_0 - a_1 x_k$ $(k = 0 \div m)$:

$$\sum_{k=0}^{m} d_k^2 \ .$$

Éste es un problema de aproximación discreta, con $I = \{x_0, \ldots, x_m\}$ y las funciones básicas $\varphi_0(x) = 1$, $\varphi_1(x) = x$, en el que hemos supuesto que todas las abscisas eran igualmente importantes y hemos tomado por ello todos los pesos iguales a 1 en el producto escalar; esto es

$$(u, v) = \sum_{k=0}^{m} u_k v_k \ .$$

Las ecuaciones normales

$$\begin{pmatrix} (\varphi_0, \varphi_0) & (\varphi_0, \varphi_1) \\ (\varphi_1, \varphi_0) & (\varphi_1, \varphi_1) \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix} = \begin{pmatrix} (\varphi_0, f) \\ (\varphi_1, f) \end{pmatrix}$$

forman el sistema lineal de dos ecuaciones y dos incógnitas:

$$\begin{pmatrix} m+1 & \sum_{k=0}^m x_k \\ \sum_{k=0}^m x_k & \sum_{k=0}^m x_k^2 \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix} = \begin{pmatrix} \sum_{k=0}^m y_k \\ \sum_{k=0}^m x_k y_k \end{pmatrix} \ .$$

La solución de este sistema es:

$$\begin{aligned} a_1^* &= \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - \overline{x}^2} \ , \\ a_0^* &= \overline{y} - a_1^* \overline{x} \ , \end{aligned}$$

donde la barra indica la media:

$$\overline{x} = \frac{1}{m+1} \sum_{k=0}^{m} x_k \ , \quad \overline{y} = \frac{1}{m+1} \sum_{k=0}^{m} y_k \ ,$$

$$\overline{x^2} = \frac{1}{m+1}\sum_{k=0}^{m} x_k^2 \ , \quad \overline{xy} = \frac{1}{m+1}\sum_{k=0}^{m} x_k y_k \ .$$

En particular, observamos que el *punto medio* $(\overline{x}, \overline{y})$ pertenece a la recta buscada $\overline{y} = a_0^* + a_1^* \overline{x}$. Teniendo en cuenta este hecho, podemos buscar la recta $y = a_0 + a_1 x$ en la forma alternativa $y = c_0 + c_1(x - \overline{x}) = c_0\psi_0(x) + c_1\psi_1(x)$, donde claramente $a_1 = c_1$ y $a_0 = c_0 - c_1\overline{x}$; esto es, tomando $\psi_0(x) = 1$ y $\psi_1(x) = x - \overline{x}$ como funciones básicas del espacio vectorial $\mathcal{P}_1$ de los polinomios de grado menor o igual que 1. Las ecuaciones normales son entonces

$$\begin{pmatrix} (\psi_0,\psi_0) & (\psi_0,\psi_1) \\ (\psi_1,\psi_0) & (\psi_1,\psi_1) \end{pmatrix} \begin{pmatrix} c_0^* \\ c_1^* \end{pmatrix} = \begin{pmatrix} (\psi_0, f) \\ (\psi_1, f) \end{pmatrix} \ .$$

Dado que

$$(\psi_1,\psi_0) = \sum_{k=0}^{m}(x_k - \overline{x})\cdot 1 = \sum_{k=0}^{m} x_k - (m+1)\overline{x} = 0 \ ,$$

las ecuaciones normales quedan reducidas al sistema diagonal

$$\begin{pmatrix} m+1 & 0 \\ 0 & \sum_{k=0}^{m}(x_k - \overline{x})^2 \end{pmatrix} \begin{pmatrix} c_0^* \\ c_1^* \end{pmatrix} = \begin{pmatrix} \sum_{k=0}^{m} y_k \\ \sum_{k=0}^{m}(x_k - \overline{x})y_k \end{pmatrix} \ .$$

Usando

$$\sum_{k=0}^{m}(x_k - \overline{x})\overline{y} = \overline{y}\sum_{k=0}^{m}(x_k - \overline{x}) = 0 \ ,$$

obtenemos

$$a_1^* = c_1^* = \frac{\mathrm{cov}(x,y)}{\mathrm{cov}(x,x)} \ , \quad a_0^* = c_0^* - c_1^*\overline{x} = \overline{y} - a_1^*\overline{x} \ ,$$

donde

$$\mathrm{cov}(x,y) \equiv \frac{1}{m+1}\sum_{k=0}^{m}(x_k - \overline{x})(y_k - \overline{y})$$

recibe el nombre de *covariancia* de las variables x e y, y

$$\mathrm{cov}(x,x) \equiv \sigma^2(x) = \frac{1}{m+1}\sum_{k=0}^{m}(x_k - \overline{x})^2$$

recibe el nombre de *variancia* de la variable x; $\sigma(x)$ se llama *desviación típica* o *estándar* de la variable x.

Es necesario observar que se dispone de dos expresiones matemáticamente equivalentes para la pendiente a_1^* de la recta buscada, pero no numéricamente equivalentes, ya que en general habrá cancelaciones más importantes en la primera.

Tomando ahora $y^* = (y_0^*, \ldots, y_m^*)$ donde $y_k^* = a_0^* + a_1^* x_k$ $(k = 0 \div m)$, si continuamos pensando firmemente que los puntos (x_k, y_k) $(k = 0 \div m)$ deberían estar sobre una recta, de la que se han "escapado" por errores en los datos, podemos hacerlos "regresar" a la recta $y = a_0^* + a_1^* x$ admitiendo ahora como puntos válidos los puntos (x_k, y_k^*) $(k = 0 \div m)$. Es

por esto que la recta $y = a_0^* + a_1^* x$ encontrada recibe el nombre de *recta de regresión*. Es claro que esta regresión será más válida cuanto menor sea

$$\sum_{k=0}^{m} d_k^{*2} = \sum_{k=0}^{m} (y_k - y_k^*)^2 = \|y - y^*\|_2^2 \ .$$

En términos de la variancia de y, tenemos

$$\sigma^2(y) = \frac{1}{m+1} \sum_{k=0}^{m} (y_k - \overline{y})^2 = \frac{\|y - \overline{y}\|_2^2}{m+1} \ .$$

Usando la propiedad de ortogonalidad de $y - y^*$ respecto a $y^* - \overline{y}$

$$\|y - \overline{y}\|_2^2 = \|y - y^*\|_2^2 + \|y^* - \overline{y}\|_2^2 \ ,$$

obtenemos

$$\sigma^2(y) = \frac{\|y - y^*\|_2^2}{m+1} + \sigma^2(y^*) \ .$$

La regresión será, pues, más fiable cuanto más cerca de 1 sea el cociente

$$\begin{aligned} \rho_{xy} &= \frac{\sigma(y^*)}{\sigma(y)} = \frac{\|y^* - \overline{y}\|_2}{\|y - \overline{y}\|_2} \\ &= a_1^* \frac{\sigma(x)}{\sigma(y)} = \frac{\mathrm{cov}(x,y)}{\sigma(x)\sigma(y)} \ , \end{aligned}$$

llamado *coeficiente de regresión*. Si ρ_{xy} no está cerca de 1, la nube de puntos no se distribuye como una recta y será necesario recurrir a otros tipos de aproximación no lineal para ajustarla.

3.2.3 Resolución de las ecuaciones normales

Una vez reducido el problema de aproximación por mínimos cuadrados, es necesario resolver las ecuaciones normales asociadas

$$Aa^* = b \ ,$$

donde $A = (a_{ij})$ con $a_{ij} = (\varphi_i, \varphi_j)$ $(i, j = 0 \div n)$, $a^* = (a_0^* \quad \dots \quad a_n^*)^\top$ y $b = (b_i)$ con $b_i = (\varphi_i, f)$ $(i = 0 \div n)$.

Dado que A es definida positiva, si las funciones φ_j $(j = 0 \div n)$ son linealmente independientes, un método especialmente adecuado es el de Cholesky (véase el apartado 2.2.3), que consiste en factorizar A en la forma $LDL^\top$ o $\mathcal{L}\mathcal{L}^\top$ con $L, \mathcal{L}$ matrices triangulares inferiores, y que requiere $\frac{n^3}{6} + \mathcal{O}(n^2)$ operaciones.

El trabajo preliminar de construcción de las ecuaciones normales requiere generalmente $\frac{1}{2}p(n+1)(n+4)$ operaciones, donde p es el número de operaciones necesarias para cada producto escalar que normalmente será mayor que $n+1$: en el caso de la aproximación discreta, $p = m+1 \geq n+1$, si I consta de $m+1$ elementos; en el caso de la aproximación continua, es necesario calcular las integrales

$$\int_a^b \varphi_i(x)\varphi_j(x)w(x)dx \ .$$

Destacamos que la mayor parte del cálculo corresponderá a la formación de las ecuaciones normales (3.15). Ahora bien, esta parte del cálculo está fuertemente condicionada por la elección de las funciones básicas de $\mathcal{F}_n$. En general, deberemos calcular todos los productos escalares posibles. Conviene también tener en cuenta que la matriz A puede estar mal condicionada como consecuencia del hecho de que las funciones φ_j $(j = 0 \div n)$ sean "poco independientes" desde un punto de vista numérico.

En caso de usar una base de funciones ψ_j $(j = 0 \div n)$ *ortogonales* (esto es, $(\psi_i, \psi_j) = 0$, si $i \neq j$, y $(\psi_i, \psi_i) > 0$), el sistema (3.15) es diagonal (véase, por ejemplo, el problema 3.15). Los *coeficientes ortogonales* c_j $(j = 0 \div n)$ de la aproximación por mínimos cuadrados

$$f_n^*(x) = \sum_{j=0}^{n} c_j^* \psi_j(x)$$

se obtienen, de forma inmediata, haciendo

$$c_j^* = \frac{b_j}{a_{jj}} = \frac{(\psi_j, f)}{(\psi_j, \psi_j)} \ (j = 0 \div n) \ .$$

Dada la simplicidad de estas expresiones, los métodos de resolución estándar de las ecuaciones lineales (3.15) están basados en la ortogonalización de las funciones básicas φ_j $(j = 0 \div n)$; es decir, en la expresión de las ecuaciones normales en una base de funciones ortogonales.

A continuación veremos como se realizan estos procesos en cinco tipos de aproximación por mínimos cuadrados.

- Aproximación discreta: métodos de ortogonalización de Householder y de Gram--Schmidt.
- Sistemas lineales sobredeterminados: se reduce al caso anterior.
- Aproximación general: ortogonalización de Gram-Schmidt.
- Aproximación polinomial: polinomios ortogonales.
- Aproximación trigonométrica: polinomios trigonométricos ortogonales, interpolación trigonométrica.

Caso de aproximación discreta: ortogonalización de Householder y de Gram-Schmidt

Si consideramos una aproximación por mínimos cuadrados ponderada con los pesos $w_k > 0$ $(k = 0 \div m)$, tenemos:

$$a_{ij} = (\varphi_i, \varphi_j) = \sum_{k=0}^{m} w_k \varphi_i(x_k) \varphi_j(x_k) \ (i, j = 0 \div n) \ ,$$

$$b_i = (\varphi_i, f) = \sum_{k=0}^{m} w_k \varphi_i(x_k) f(x_k) \ (i = 0 \div n) \ ;$$

entonces podemos escribir la matriz A de las ecuaciones normales en la forma

$$A = M^\top W M$$

y el término independiente b en la forma

$$b = M^\top W y \ ,$$

donde $M = (m_{kj})$ es una matriz $(m+1)$x$(n+1)$ con $m_{kj} = \varphi_j(x_k)$ $(k = 0 \div m)$ $(j = 0 \div n)$, $W = (w_k)$ es la matriz diagonal $(m+1)$x$(m+1)$ formada por los pesos w_k $(k = 0 \div m)$ e $y = (y_k)$ es el vector de $m+1$ componentes con $y_k = f(x_k)$ $(k = 0 \div m)$.

Las ecuaciones normales se escriben, en forma matricial, como

$$M^\top W M a^* = M^\top W y \ , \tag{3.16}$$

sistema que tiene solución única si los vectores

$$(\varphi_j(x_0) \quad \dots \quad \varphi_j(x_m))^\top \ (j = 0 \div n)$$

son linealmente independientes; esto es, si el rango rg$M$ de la matriz M es $n+1$ y necesariamente $m \geq n$.

Suponemos primeramente que $W = I_{m+1}$; es decir que todos los pesos son la unidad. Entonces, el sistema (3.16) queda de la forma

$$M^\top M a^* = M^\top y \ . \tag{3.17}$$

Se exponen a continuación dos métodos que, siguiendo las indicaciones del apartado 2.2.4, permiten encontrar la factorización QR (generalizada) de la matriz M y reducir así las ecuaciones normales a un sistema triangular superior (véase el problema 3.18).

a) *Método de ortogonalización de Householder*

Se halla una matriz ortogonal $(m+1)$x$(m+1)$ P, formada por composición de n matrices de Householder: $P = P_n P_{n-1} \cdots P_1$, de manera que

$$PM = R = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix} \begin{matrix} \}n+1 \\ \}m-n \end{matrix} \ ,$$

con $\tilde{R}$ matriz $(n+1)$x$(n+1)$ triangular superior, regular si rg $M = n+1$. En este caso, $\tilde{R}^\top$ es regular y, si llamamos $\widetilde{Py}$ al vector formado por las $n+1$ primeras componentes de Py, tenemos que el sistema (3.17) es equivalente al sistema triangular $(n+1)$x$(n+1)$

$$\tilde{R} a^* = \widetilde{Py} \ . \tag{3.18}$$

b) *Método de ortogonalización modificado de Gram-Schmidt*

Si rg $M = n+1$, se hallan una matriz $(m+1)$x$(n+1)$ Q, tal que $Q^\top Q = D$ con D matriz diagonal $(n+1)$x$(n+1)$, y una matriz $(n+1)$x$(n+1)$ R, triangular superior con elementos diagonales $r_{jj} = 1$ $(j = 0 \div n)$, de manera que $M = QR$. El sistema (3.17) es entonces equivalente al sistema triangular $(n+1)$x$(n+1)$

$$R a^* = D^{-1} Q^\top y \ . \tag{3.19}$$

Nótese que no es necesario construir las ecuaciones normales (3.16), sino que puede obtenerse directamente uno de los sistemas triangulares superiores (3.18) y (3.19) (véase el problema 3.18).

Remarcamos la conveniencia de usar pesos no iguales para dar más importancia a algunas ecuaciones respecto a las demás. El caso de pesos no iguales a 1 se reduce al caso que se acaba de ver tomando el sistema $M_1 a = y_1$ con $M_1 = \sqrt{W}M$ e $y_1 = \sqrt{W}y$, que equivale a multiplicar la ecuación k-ésima por el número $\sqrt{w_k}$ $(k = 0 \div m)$.

Caso de sistemas lineales sobredeterminados

En el apartado anterior, hemos encontrado $a^* = (a_0^* \quad \dots \quad a_n^*)^\top$ de manera que, para cualquier otro $a = (a_0 \quad \dots \quad a_n)^\top$, tenemos

$$\sum_{k=0}^{m} \left[f(x_k) - \sum_{j=0}^{n} a_j^* \varphi_j(x_k) \right]^2 \leq \sum_{k=0}^{m} \left[f(x_k) - \sum_{j=0}^{n} a_j \varphi_j(x_k) \right]^2 .$$

Recordando que $y_k = f(x_k)$ $(k = 0 \div m)$ y $m_{kj} = \varphi_j(x_k)$ $(k = 0 \div m)$ $(j = 0 \div n)$, vemos que a^* es la solución del problema de minimización

$$\|y - Ma^*\|_2 = \min_{a \in \mathbb{R}^{n+1}} \|y - Ma\|_2 .$$

Consideramos ahora el problema siguiente: dada una matriz M con m_0 filas y n_0 columnas y un vector y de m_0 componentes, encontrar a de n_0 componentes solución del sistema

$$Ma = y . \tag{3.20}$$

Si $m_0 > n_0$, este sistema, en general, no tiene solución: hay más ecuaciones (m_0) que incógnitas (n_0) y recibe el nombre de *sistema lineal sobredeterminado*. Siguiendo la línea de este capítulo, podemos intentar encontrar un vector $a^* \in \mathbb{R}^{n_0}$ que minimice la norma euclídea de $y - Ma$:

$$\|y - Ma^*\|_2 = \min_{a \in \mathbb{R}^{n_0}} \|y - Ma\|_2$$

que corresponde al problema (3.19) y equivale, por lo tanto, a resolver las *ecuaciones normales*

$$M^\top M a^* = M^\top y .$$

Estas ecuaciones tienen solución única si y sólo si rg $M = n_0 \leq m_0$ y son, en este caso, equivalentes a (3.18) y (3.19).

Para dar más importancia a algunas ecuaciones de (3.20) respecto a las demás, se puede introducir una *matriz de pesos* positivos W diagonal en la norma euclídea, tal como ha sido tratado al final del apartado anterior.

Caso de aproximación general: ortogonalización de Gram-Schmidt

Si tenemos una base de funciones φ_j $(j = 0 \div n)$ de $\mathcal{F}_n$, el método (clásico) de ortogonalización de Gram-Schmidt proporciona un algoritmo para el cálculo de una base de

funciones ortogonales ψ_j $(j = 0 \div n)$ de $\mathcal{F}_n$, planteando sencillamente una factorización QR:

$$\begin{aligned} \varphi_0(x) &= r_{00}\psi_0(x) \ , \\ \varphi_1(x) &= r_{01}\psi_0(x) + r_{11}\psi_1(x) \ , \\ \vdots & \qquad\qquad \vdots \\ \varphi_n(x) &= r_{0n}\psi_0(x) + r_{1n}\psi_1(x) + \ldots + r_{nn}\psi_n(x) \ ; \end{aligned} \tag{3.21}$$

o, equivalentemente,

$$(\varphi_0(x) \quad \ldots \quad \varphi_n(x)) = (\psi_0(x) \quad \ldots \quad \psi_n(x)) \begin{pmatrix} r_{00} & r_{01} & \cdots & r_{0n} \\ & r_{11} & \cdots & r_{1n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix} , \tag{3.22}$$

que se resuelve por substitución hacia adelante:

$$\begin{gathered} \psi_0(x) = \varphi_0(x) \ , \quad d_0 = (\psi_0, \psi_0) \ , \quad r_{00} = 1 \ ; \\ \psi_j(x) = \varphi_j(x) - \sum_{i=0}^{j-1} r_{ij}\psi_i(x) \ , \quad d_j = (\psi_j, \psi_j) \ , \\ r_{ij} = \frac{(\varphi_j, \psi_i)}{d_i} \quad (i = 0 \div j-1) \ , \quad r_{jj} = 1 \quad (j = 0 \div n) \ . \end{gathered} \tag{3.23}$$

Notamos que la matriz $R = (r_{ij})$ así obtenida es triangular superior con la diagonal llena de unos, ya que hemos supuesto φ_j $(j = 0 \div n)$ linealmente independientes.

Introduciendo ahora

$$\tilde{\psi}_j = \frac{\psi_j}{\sqrt{(\psi_j, \psi_j)}} = \frac{\psi_j}{\sqrt{d_j}} \ (j = 0 \div n) \ ,$$

$$\tilde{r}_{ij} = \sqrt{d_i} r_{ij} \ (0 \le i \le j \le n) \ ,$$

obtenemos otro sistema triangular. Las funciones $\tilde{\psi}_j$ $(j = 0 \div n)$ forman una *base de funciones ortonormales*:

$$(\tilde{\psi}_i, \tilde{\psi}_j) = \delta_{ij} \quad (i, j = 0 \div n) \ ,$$

y la matriz triangular superior $\tilde{R} = (\tilde{r}_{ij})$ tiene elementos positivos en la diagonal, no necesariamente unitarios,

$$\tilde{r}_{jj} = \sqrt{d_j} = \sqrt{(\psi_j, \psi_j)} \ (j = 0 \div n) \ .$$

En cualquiera de estas dos bases de funciones ortogonales, las ecuaciones normales (3.16) son diagonales y, por lo tanto, resolubles inmediatamente.

Este método de ortogonalización de Gram-Schmidt se aplica tanto en el caso de aproximación continua como discreta. En este último caso, es matemáticamente equivalente al de ortogonalización modificado de Gram-Schmidt, aunque no numéricamente.

En efecto, si tenemos un problema de aproximación discreta sobre un conjunto $I = \{x_0, \ldots, x_m\}$, escribiendo el sistema triangular (3.22) para las abscisas del conjunto I, obtenemos el sistema

$$M = QR \ ,$$

donde las matrices $M = (m_{kj}) = (\varphi_j(x_k))$ y $Q = (q_{kj}) = (\psi_j(x_k))$ son de dimensión $(m+1)$x$(n+1)$ y R es una matriz $(n+1)$x$(n+1)$ triangular superior. Entonces, la condición de ortogonalidad

$$(\psi_i, \psi_j) \equiv \sum_{k=0}^{m} w_k \psi_i(x_k)\psi_j(x_k) = d_j \delta_{ij} \quad (i, j = 0 \div n)$$

equivale a

$$Q^\top W Q = D \ ,$$

donde $W = \mathrm{diag}(w_0, \ldots, w_n)$ y $D = \mathrm{diag}(d_0, \ldots, d_m)$.

Si los pesos son todos iguales a 1, el método clásico de Gram-Schmidt (3.23) proporciona una factorización QR (generalizada) de la matriz M de dimensión $(m+1)$x$(n+1)$ con Q de dimensión $(m+1)$x$(n+1)$ tal que $Q^\top Q = D$ y R triangular superior con unos en la diagonal. En cambio, la factorización con $\tilde{\psi}_j$, $\tilde{r}_{ij}$ da lugar a una factorización $A = \tilde{Q}^\top \tilde{R}$ con $\tilde{Q}^\top \tilde{Q} = I_{n+1}$ y $\tilde{R}$ triangular superior con elementos positivos en la diagonal: $\tilde{r}_{jj} > 0$.

Este método clásico presenta problemas si las funciones iniciales φ_j $(j = 0 \div n)$ son "poco independientes". En tal caso, los coeficientes de la matriz R pueden ser grandes y la fórmula

$$\psi_j(x) = \varphi_j(x) - \sum_{i=0}^{j-1} r_{ij}\psi_i(x)$$

puede presentar problemas de cancelación. Ello puede provocar que la función ψ_j calculada numéricamente no sea suficientemente ortogonal a las funciones anteriores ψ_i $(i = 0 \div j-1)$ $(j = 0 \div n)$.

Observando que el valor (φ_j, ψ_i) no cambia si φ_j es modificado por

$$\varphi_j^{(i)} = \varphi_j - \sum_{l=0}^{i-1} b_l \psi_l \quad (i = 0 \div j-1)$$

e imponiendo que $\|\varphi_j^{(i)}\|$ sea mínima, se obtiene $b_l = r_{lj}$ $(l = 0 \div j-1)$. Para evitar las cancelaciones se recomienda tomar

$$\varphi_j^{(i)} = \varphi_j - \sum_{l=0}^{i-1} b_l \psi_l$$

o, de manera recursiva,

$$\varphi_j^{(0)} = \varphi_j \ (j = 0 \div n) \ ,$$

$$\varphi_j^{(i+1)} = \varphi_j^{(i)} - r_{ij}\psi_i \ (j = i+1 \div n) \ (i = 0 \div n) \ .$$

Se obtiene así el *método de ortogonalización modificado de Gram-Schmidt*

$$\varphi_j^{(0)} = \varphi_j \ (j = 0 \div n) \ ,$$

$$\psi_i(x) = \varphi_i^{(i)}(x) \ , \quad d_i = (\psi_i, \psi_i) \ ,$$

$$r_{ij} = \frac{(\varphi_j^{(i)}, \psi_i)}{d_i} \ , \quad \varphi_j^{(i+1)} = \varphi_j^{(i)} - r_{ij}\psi_i \ (j = i+1 \div n) \ (i = 0 \div n)$$

que fue ya introducido en el apartado 2.2.4.

Caso de aproximación polinomial: polinomios ortogonales

Consideramos ahora el caso en que φ_j es un polinomio de grado j (por ejemplo, $\varphi_j(x) = x^j$). El espacio vectorial $\mathcal{F}_n$ generado por las funciones φ_j $(j = 0 \div n)$ es ahora el espacio vectorial $\mathcal{P}_n$ de polinomios de grado menor o igual que n y el problema de aproximación por mínimos cuadrados polinomial de grado menor o igual que n toma la forma que se enuncia a continuación.

Dada $f : I \longrightarrow \mathbb{R}$, buscamos $p_n^* \in \mathcal{P}_n$ tal que

$$\|f - p_n^*\| = \min_{p_n \in \mathcal{P}_n} \|f - p_n\| \ , \tag{3.24}$$

donde $\| \ \| = \| \ \|_{2,w}$ es una norma euclídea; recordamos que I puede ser discreto ($I = \{x_0, \ldots, x_m\}, m \leq n$) o continuo (un intervalo).

Escribiendo $p_n^*(x)$ en la forma

$$p_n^*(x) = \sum_{j=0}^{n} a_j^* \varphi_j(x) \ ,$$

las ecuaciones normales que resuelven el problema (3.24) de aproximación por mínimos cuadrados

$$\sum_{j=0}^{n} (\varphi_i, \varphi_j) a_j^* = (\varphi_i, f) \ (i = 0 \div n) \tag{3.25}$$

tienen solución única, a causa de la independencia lineal de $\varphi_0(x)$, ..., $\varphi_n(x)$ sobre I; notamos que, en el caso discreto, es necesaria la condición $n \leq m$, debido a que los polinomios satisfacen la propiedad de Haar introducida en el apartado 3.1.2.

La resolución de las ecuaciones normales puede presentar en general muchos problemas numéricos. Por esto es necesario usar métodos de ortogonalización. En el caso de polinomios, es muy eficiente calcular los polinomios ortogonales a través de relaciones de recurrencia obtenidas por ortogonalización de Gram-Schmidt, tal como se explica a continuación.

El primer polinomio (de grado 0) debe ser una constante: $\psi_0(x) = A_0$, con $A_0 \neq 0$. Sea ahora $j \geq 0$ y supongamos que hemos encontrado ya los $j+1$ primeros polinomios ortogonales $\psi_l(x)$ $(l = 0 \div j)$. Cada polinomio $\psi_j(x) = A_j x^j + \ldots$ tiene grado j; esto es, su *coeficiente principal* A_j es no nulo.

Consideramos la ortogonalidad tanto en el caso discreto como en el continuo; es decir, para los dos productos escalares (3.12) y (3.13). Para los dos se cumple la relación $(xu, v) = (u, xv) \ \forall u, v$. Además, como que los polinomios ortogonales $\psi_l(x)$ $(l = 0 \div j)$ son linealmente independientes (en el caso discreto, es necesario que $j \leq m$), todo polinomio $p_i(x)$ de grado $i \leq j$ se expresa de manera única como combinación lineal de $\psi_0(x), \ldots, \psi_i(x)$:

$$p_i(x) = c_0\psi_0(x) + \ldots + c_i\psi_i(x) \ .$$

Usando las relaciones de ortogonalidad, tenemos que

$$(\psi_j, p_i) = 0, \text{ para todo } p_i(x) \text{ de grado } i < j \ . \tag{3.26}$$

Aplicamos ahora el método clásico de Gram-Schmidt (3.23) al conjunto formado por las funciones $\psi_0, \ldots, \psi_j, \varphi_{j+1}$, donde escogemos $\varphi_{j+1}(x) = \alpha_j x\psi_j(x)$ (polinomio de grado $j+1$, si $\alpha_j \neq 0$). Dado que los primeros $j+1$ polinomios ya son ortogonales, sólo es necesario encontrar $\psi_{j+1}(x) = A_{j+1}x^{j+1} + \ldots$, donde prefijamos un valor no nulo para el coeficiente principal A_{j+1} de $\psi_{j+1}(x)$.

Encontramos

$$\psi_{j+1}(x) = \varphi_{j+1}(x) - \sum_{i=0}^{j} r_{i,j+1}\psi_i(x) \ ,$$

donde

$$r_{i,j+1} = \frac{(\varphi_{j+1}, \psi_i)}{d_i} = \alpha_j \frac{(x\psi_j, \psi_i)}{d_i} = \alpha_j \frac{(\psi_j, x\psi_i)}{d_i} \ , \quad d_i = (\psi_i, \psi_i) \ .$$

De la relación (3.26), tenemos

$$\begin{aligned} r_{i,j+1} &= 0 \quad (i = 0 \div j-2) \\ r_{j,j+1} &= \alpha_j \frac{(\psi_j, x\psi_j)}{d_j} \equiv \alpha_j \beta_j \ , \\ r_{j-1,j+1} &= \alpha_j \frac{(\psi_j, x\psi_{j-1})}{d_{j-1}} \equiv \gamma_j \ ; \end{aligned}$$

se obtiene, pues, la siguiente *recurrencia de los polinomios ortogonales*:

$$\psi_0(x) = A_0 \ , \quad \psi_{j+1}(x) = \alpha_j(x - \beta_j)\psi_j(x) - \gamma_j\psi_{j-1}(x) \ (j \geq 0) \ , \tag{3.27}$$

con

$$\begin{aligned} \alpha_j &= \frac{A_{j+1}}{A_j} \quad (j \geq 0) \ , \\ \beta_j &= \frac{(\psi_j, x\psi_j)}{(\psi_j, \psi_j)} \quad (j \geq 0) \ , \\ \gamma_j &= \alpha_j \frac{(\psi_j, x\psi_{j-1})}{(\psi_{j-1}, \psi_{j-1})} = \frac{\alpha_j}{\alpha_{j-1}} \frac{(\psi_j, \psi_j)}{(\psi_{j-1}, \psi_{j-1})} \quad (j \geq 1) \ , \end{aligned}$$

donde hemos tomado, por convenio, $\psi_{-1}(x) = 0$.

La última relación para los γ_j se obtiene a partir de la relación recurrente (3.27), substituyendo $j+1$ por j, haciendo el producto escalar por $\psi_j(x)$ y aplicando (3.26). Obtenemos así, de manera única, $\psi_{j+1}(x)$ a partir de $\psi_j(x)$ y $\psi_{j-1}(x)$ y el coeficiente principal escogido $A_{j+1} = \alpha_j A_j \neq 0$.

Este proceso se puede ir repitiendo indefinidamente en el caso continuo, pero no en el caso discreto, ya que era necesario que $j \leq m$. De hecho, si $I = \{x_0, \ldots, x_m\}$, de la unicidad de los polinomios ortogonales (una vez fijados los coeficientes principales) tenemos

$$\psi_{m+1}(x) = A_{m+1}(x - x_0) \cdots (x - x_m) \ ,$$

que es nulo sobre I y

$$\|\psi_{m+1}\|^2 = \sum_{k=0}^{m} w_k \psi_{m+1}^2(x_k) = 0 \ .$$

COMENTARIOS

1. Si buscamos polinomios ortogonales *mónicos* (esto es, con coeficientes principales $A_j = 1$ $(j \geq 0)$), entonces tomaremos $\alpha_j = 1$ $(j \geq 0)$.

2. Si los pesos son *simétricos* respecto a la abscisa media $x = c$, entonces $\beta_j = c$ $(j \geq 0)$ y $\psi_j(c+x) = (-1)^j \psi_j(c-x)$. En el caso particular $c = 0$, tenemos $\beta_j = 0$ $(j \geq 0)$; entonces, si j es par, $\psi_j(x)$ es par $(\psi_j(-x) = \psi_j(x))$ y, si j es impar, $\psi_j(x)$ es impar $(\psi_j(-x) = -\psi_j(x))$.

Usando ahora la relación recurrente de los polinomios ortogonales (3.27), la solución del problema de aproximación polinomial por mínimos cuadrados (3.24) viene dada por

$$p_n^*(x) = \sum_{j=0}^{n} c_j^* \psi_j(x) \ , \tag{3.28}$$

donde los coeficientes ortogonales c_j^* son las componentes de la solución de las ecuaciones normales (3.25), que son diagonales. Por lo tanto,

$$c_j^* = \frac{(\psi_j, f)}{(\psi_j, \psi_j)} \quad (j = 0 \div n) \ .$$

Además, la expresión (3.28) juntamente con la recurrencia (3.27) proporcionan un método eficaz para la evaluación de $p_n^*(x)$, llamado *regla de Clenshaw* (véase el problema 3.18):

$$\begin{aligned} q_{n+2} &= q_{n+1} = 0 \ , \\ q_j &= \alpha_j(x - \beta_j) q_{j+1} - \gamma_{j+1} q_{j+2} + c_j^* \quad (j = n \div 0) \ , \\ p_n^*(x) &= A_0 q_0 \ . \end{aligned}$$

En efecto,

$$\begin{aligned} p_n^*(x) &= \sum_{j=0}^{n} c_j^* \psi_j(x) = \sum_{j=0}^{n} [q_j - \alpha_j(x - \beta_j) q_{j+1} + \gamma_{j+1} q_{j+2}] \psi_j(x) \\ &= \sum_{j=0}^{n} [\psi_{j+1}(x) - \alpha_j(x - \beta_j)\psi_j(x) + \gamma_j \psi_{j-1}(x)] q_{j+1} + q_0 \psi_j(x) \\ &= A_0 q_0 \ . \end{aligned}$$

NOTA

La interpolación de una función f en unas abscisas x_k $(k = 0 \div m)$ por un polinomio de grado menor o igual que m,

$$p_m(x_k) = f(x_k) \ (k = 0 \div m) \ ,$$

no es más que un caso particular del problema de aproximación por mínimos cuadrados en el caso discreto para pesos arbitrarios, en el que buscamos un polinomio $p_m^*(x)$ de manera que minimice

$$\|f - q_m\|_{2,w}^2 = \sum_{k=0}^{m} [f(x_k) - q_m(x_k)]^2 w_k$$

entre los polinomios $q_m(x)$ de grado menor o igual que m. Ahora bien, este mínimo se alcanza para el polinomio de interpolación $p_m(x)$ debido a que se anula para él.

Por lo tanto, otra manera de resolver el problema de interpolación, diferente de las expuestas en el apartado 3.1.3, consiste en buscar una base de polinomios ortogonales $\psi_j(x)$ $(j = 0 \div m)$ respecto al producto escalar discreto

$$(u, v) = \sum_{k=0}^{m} w_k u(x_k) v(x_k) \ ,$$

y entonces $p_m^*(x)$ viene dado explícitamente por

$$p_m^*(x) = \sum_{j=0}^{m} c_j^* \psi_j(x) \ , \quad c_j^* = \frac{(\psi_j, f)}{(\psi_j, \psi_j)} \ (j = 0 \div m) \ .$$

Ejemplos de familias de polinomios ortogonales

1. Los polinomios de Legendre $P_j(t)$ $(j \geq 0)$

Si $I = [-1, 1]$ y $w(t) = 1$, entonces los llamados *polinomios de Legendre*, definidos por

$$P_0(t) = 1 \ , \quad P_j(t) = \frac{1}{2^j j!} \frac{d^j}{dt^j} [(t^2 - 1)^j] \ (j \geq 1) \ ,$$

son polinomios ortogonales asociados.

En efecto, si $i \neq j$, integrando por partes

$$(P_i, P_j) = \int_{-1}^{1} P_i(t) P_j(t) dt \ ,$$

se ve que esta integral es nula; además, por inducción, resulta que

$$d_j = (P_j, P_j) = \frac{(2j)!}{(2^j j!)^2} \int_{-1}^{1} (1 - t^2)^j dt = \frac{2}{2j + 1} \ (j \geq 0) \ .$$

Dado que el coeficiente principal es

$$A_j = \frac{2j(2j - 1) \cdots (j + 1)}{2^j j!} \ (j \geq 0) \ ,$$

entonces

$$\alpha_j = \frac{A_{j+1}}{A_j} = \frac{2j + 1}{j + 1} \ , \quad \gamma_j = \frac{\alpha_j d_j}{\alpha_{j-1} d_{j-1}} = \frac{j}{j + 1} \ (j \geq 0) \ .$$

Usando ahora el hecho de que la función peso $w(t) = 1$ es *simétrica* respecto a la abscisa media $c = 0$ del intervalo I, tenemos que $\beta_j = 0$ $(j \geq 0)$ y que los polinomios de Legendre satisfacen

$$P_j(-t) = (-1)^j P_j(t)$$

y la recurrencia:

$$P_0(t) = 1 \ , \quad P_1(t) = t \ ,$$
$$P_{j+1}(t) = \frac{2j+1}{j+1} t P_j(t) - \frac{j}{j+1} P_{j-1}(t) \ (j \geq 1) \ .$$

2. Los polinomios de Gram $P_{j,m}(t)$ $(j = 0 \div m)$

Si $t_k = k$ y $w_k = 1$ $(k = 0 \div m)$, los *polinomios de Gram*, definidos por

$$P_{j,m}(t) = \sum_{i=0}^{j} (-1)^i \left(\begin{array}{c} j \\ i \end{array} \right) \left(\begin{array}{c} j+i \\ i \end{array} \right) \frac{t(t-1)\cdots(t-i+1)}{m(m-1)\cdots(m-i+1)}$$

son polinomios ortogonales asociados: si $j \neq l$ $(j, l = 0 \div m)$,

$$(P_{j,m}, P_{l,m}) = \sum_{k=0}^{m} P_{j,m}(x_k) P_{l,m}(x_k) = 0 \ ;$$

además,

$$d_j = (P_{j,m}, P_{j,m}) = \frac{(m+j+1)!(m-j)!}{(2j+1)(m!)^2} \ (j = 0 \div m) \ .$$

Dado que el coeficiente principal de $P_{j,m}(x)$ es

$$A_j = (-1)^j \left(\begin{array}{c} 2j \\ j \end{array} \right) \frac{1}{m(m-1)\cdots(m-j+1)} \ ,$$

entonces:

$$\alpha_j = \frac{A_{j+1}}{A_j} = -\frac{2(2j+1)}{(j+1)(m-j)} \ ,$$
$$\gamma_j = \frac{\alpha_j d_j}{\alpha_{j-1} d_{j-1}} = \frac{j(m+j+1)}{(j+1)(m-j)} \ .$$

Usando ahora el hecho de que el conjunto de abscisas de aproximación y la función peso son simétricos respecto a $\beta = \frac{m}{2}$, tenemos que $\beta_j = \frac{m}{2}$ $(j = 0 \div m)$ y los polinomios de Gram satisfacen la recurrencia:

$$\begin{aligned} P_{0,m}(t) &= 1 \ , \quad P_{1,m}(t) = 1 - \frac{2t}{m} \ , \\ (j+1)(m-j) P_{j+1,m}(t) &= (2j+1)(m-2t) P_{j,m}(t) \\ &\quad - j(m+j+1) P_{j-1,m}(t) \quad (j = 1 \div m-1) \ . \end{aligned}$$

3. Los polinomios de Legendre en el intervalo $[a, b]$

Si $I = [a, b]$ y $w(x) = 1$, los polinomios ortogonales asociados $\psi_j(x)$ $(j \geq 0)$ se obtienen de los polinomios de Legendre en $[-1, 1]$, a través del cambio de variable afín:

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

es decir,

$$\psi_j(x) \equiv P_j(\frac{2}{b-a}(x - \frac{a+b}{2})) \quad (j \geq 0)$$

son polinomios ortogonales y satisfacen la recurrencia:

$$\begin{aligned} \psi_0(x) &= 1 \ , \quad \psi_1(x) = \frac{2}{b-a}\left(x - \frac{a+b}{2}\right) \ , \\ \psi_{j+1}(x) &= \frac{2j+1}{j+1}\frac{2}{b-a}\left(x - \frac{a+b}{2}\right)\psi_j(x) \\ &\quad -\frac{j}{j+1}\psi_{j-1}(x) \ (j \geq 1) \ . \end{aligned}$$

Análogamente, si $x_k = x_0 + kh$ y $w_k = 1$ $(k = 0 \div m)$, entonces

$$\psi_{j,m}(x) = P_{j,m}\left(\frac{x - x_0}{h}\right) \ (j = 0 \div m)$$

son polinomios ortogonales. Por ejemplo, si $x_k = -1 + \frac{2k}{m}$ $(k = 0 \div m)$, entonces los polinomios $\psi_{j,m}(x) = (-1)^j P_{j,m}(\frac{m}{2}(x+1))$ satisfacen

$$\psi_{j,m}(-x) = (-1)^j \psi_{j,m}(x) \quad (j = 0 \div m)$$

y la recurrencia:

$$\begin{aligned} \psi_{0,m}(x) &= 1 \ , \quad \psi_{1,m}(x) = x \ , \\ (j+1)(m-j)\psi_{j+1,m}(x) &= (2j+1)mx\psi_{j,m}(x) \\ &\quad -j(m+j+1)\psi_{j-1,m}(x) \quad (j = 1 \div m-1) \ . \end{aligned}$$

Caso de aproximación trigonométrica: ortogonalidad de los polinomios trigonométricos, interpolación trigonométrica

Consideramos el caso en que $\psi_0(\theta) = \frac{1}{2}$, $\psi_1(\theta) = \cos\theta$, $\psi_2(\theta) = \operatorname{sen}\theta$, ..., $\psi_{2n-1}(\theta) = \cos n\theta$ y $\psi_{2n}(\theta) = \operatorname{sen} n\theta$.

El espacio vectorial $\mathcal{T}_n$ generado por las funciones $\psi_j(\theta)$ $(j = 0 \div 2n)$ se llama espacio de las *sumas trigonométricas* t_n de grado menor o igual que n

$$t_n(\theta) = \frac{a_0}{2} + \sum_{j=1}^{n} a_j \cos j\theta + \sum_{j=1}^{n} b_j \operatorname{sen} j\theta \ , \tag{3.29}$$

donde suponemos que los coeficientes ortogonales a_0, a_j y b_j $(j = 0 \div n)$ son reales. Es importante tener en cuenta que, para cualquier n, toda suma trigonométrica $t_n(\theta)$ de grado menor que n es 2π-periódica; esto es,

$$t_n(\theta) = t_n(\theta + 2\pi) \quad \forall\theta \in \mathbb{R} \ .$$

Usando ahora la función exponencial

$$e^{i\theta} = \cos\theta + i\,\mathrm{sen}\,\theta \ , \quad \cos\theta = \frac{e^{i\theta}+e^{-i\theta}}{2} \ , \quad \mathrm{sen}\,\theta = \frac{e^{i\theta}-e^{-i\theta}}{2i} \ ,$$

donde $i^2 = -1$, juntamente con la *fórmula de Moivre*

$$\cos j\theta + i\,\mathrm{sen}\,j\theta = e^{ij\theta} = (e^{i\theta})^j = (\cos\theta + i\,\mathrm{sen}\,\theta)^j \ ,$$

se obtiene otra representación para las sumas trigonométricas (3.29), llamada *desarrollo de Fourier de orden n*

$$t_n(\theta) = \sum_{j=-n}^{n} c_j e^{ij\theta} \ , \tag{3.30}$$

donde

$$c_0 = \frac{a_0}{2} \ , \quad c_j = \frac{1}{2}(a_j + ib_j) \ , \quad c_{-j} = \bar{c}_j = \frac{1}{2}(a_j - ib_j) \ (j = 1 \div n) \ ;$$

inversamente,

$$\begin{aligned} a_0 = 2c_0 \ , \quad a_j &= c_j + c_{-j} = 2\mathrm{Re}(c_j) \ , \\ b_j &= i(c_{-j} - c_j) = 2\mathrm{Im}(c_j) \\ &\quad (j = 1 \div n) \ . \end{aligned} \tag{3.31}$$

Estos coeficientes ortogonales de los desarrollos de Fourier reciben el nombre de *coeficientes de Fourier*.

Notamos que los coeficientes de Fourier c_j $(j = -n \div n)$ en (3.30) son complejos, pero satisfacen que $c_{-j} = \bar{c}_j$, si la suma es real; esto es, si proviene de una suma (3.29) con coeficientes reales.

Las sumas trigonométricas de grado menor o igual que n, en la forma (3.29) y en la forma (3.30), reciben también el nombre de *polinomios trigonométricos* de grado menor o igual que n, ya que pueden escribirse en la forma

$$t_n(\theta) = p_n(\cos\theta, \mathrm{sen}\,\theta) \equiv \sum_{j+l\le n} d_{j,l}(\cos\theta)^j(\mathrm{sen}\,\theta)^l \ ,$$

y recíprocamente (basta usar la fórmula de Moivre).

El *problema de aproximación trigonométrica* de grado menor o igual que n se enuncia, a continuación, tanto para el caso continuo como para el discreto.

Caso continuo

Dada $F : [0, 2\pi] \longrightarrow \mathbb{R}$ continua y tal que $F(0) = F(2\pi)$, buscamos $t_n^* \in \mathcal{T}_n$ tal que

$$\|F - t_n^*\| = \min_{t_n \in \mathcal{T}_n} \|F - t_n\| \ , \tag{3.32}$$

donde $\|\ \|$ es la norma euclídea sobre $[0, 2\pi]$

$$\|F\|^2 \equiv \int_0^{2\pi} F^2(\theta)d\theta \ .$$

Caso discreto

Dado $m \geq 2n$ y dada $F : I_m \longrightarrow \mathbb{R}$ sobre $m+1$ abscisas equiespaciadas en cualquier intervalo de la forma $[\phi, \phi + 2\pi]$ con paso $h_m = \frac{2\pi}{m+1}$:

$$I_m = \{\theta_k = \theta_k^{(m)} = \phi + kh_m = \phi + \frac{2\pi k}{m+1} \ (k = 0 \div m) \ \} \ ,$$

buscamos $t_n^* \in \mathcal{T}_n$ tal que

$$\|F - t_n^*\|_m = \min_{t_n \in \mathcal{T}_n} \|F - t_n\|_m \ , \tag{3.33}$$

donde $\| \ \|_m$ es la norma euclídea sobre I_m

$$\|F\|_m^2 \equiv \sum_{k=0}^{m} F^2(\theta_k) = \sum_{k=0}^{m} F^2\left(\phi + \frac{2\pi k}{m+1}\right) \ .$$

La gran ventaja de esta aproximación es que las funciones

$$\{\psi_j(\theta)\}_{j\geq 0} = \{1, \cos\theta, \operatorname{sen}\theta, \cos 2\theta, \operatorname{sen} 2\theta, \ldots\}$$

forman una base de funciones ortogonales en el caso continuo, y que las funciones

$$\{\psi_j(\theta)\}_{j=0\div 2n} = \{1, \cos\theta, \operatorname{sen}\theta, \cos 2\theta, \operatorname{sen} 2\theta, \ldots, \cos 2n\theta, \operatorname{sen} 2n\theta\}$$

forman una base de funciones ortogonales en el caso discreto si $2n \leq m$:

$$(\psi_j, \psi_l) = \int_0^{2\pi} \psi_j(\theta)\psi_l(\theta)d\theta = \begin{cases} 0 & \text{si } j \neq l, \ j, l \geq 0 \\ \frac{\pi}{2} & \text{si } j = l = 0 \\ \pi & \text{si } j = l > 0 \end{cases} \ , \tag{3.34}$$

$$(\psi_j, \psi_l)_m = \sum_{k=0}^{m} \psi_j(\theta_k)\psi_l(\theta_k) = \begin{cases} 0 & \text{si } j \neq l, \ j, l = 0 \div 2n \\ \frac{m+1}{4} & \text{si } j = l = 0 \\ \frac{m+1}{2} & \text{si } j = l = 1 \div 2n \end{cases} \ . \tag{3.35}$$

Para demostrar (3.34), basta usar las fórmulas trigonométricas:

$$\begin{aligned} \cos A \cos B &= \frac{1}{2}[\cos(A-B) + \cos(A+B)] \ , \\ \operatorname{sen} A \operatorname{sen} B &= \frac{1}{2}[\cos(A-B) - \cos(A+B)] \ , \\ \operatorname{sen} A \cos B &= \frac{1}{2}[\operatorname{sen}(A-B) + \operatorname{sen}(A+B)] \ . \end{aligned} \tag{3.36}$$

La relación (3.35) se obtiene también de la (3.36) y de la suma de la progresión geométrica

$$\sum_{k=0}^{m} \cos kA + i \sum_{k=0}^{m} \operatorname{sen} kA = \sum_{k=0}^{m} e^{ikA} = \frac{e^{i(m+1)A} - 1}{e^{iA} - 1} \ ,$$

cuando $e^{iA} \neq 1$.

Ahora las ecuaciones normales que resuelven los problemas de aproximación por mínimos cuadrados (3.32) y (3.33) son diagonales y, por lo tanto, la solución t_n^* viene dada por

$$t_n^*(\theta) = \frac{a_0^*}{2} + \sum_{j=1}^{n} a_j^* \cos j\theta + \sum_{j=1}^{n} b_j^* \operatorname{sen} j\theta \ ,$$

donde a_0^*, a_j^* y b_j^* $(j = 1 \div n)$ vienen dadas por las expresiones siguientes:

$$\begin{array}{rcll} & & \text{Caso continuo} & \text{Caso discreto } (2n \le m) \\ a_0^* = \dfrac{(\psi_0, F)}{(\psi_0, \psi_0)} = & & \frac{1}{\pi}\int_0^{2\pi} F(\theta)d\theta \ , & \dfrac{2}{m+1}\displaystyle\sum_{k=0}^{m} F(\theta_k) \\ a_j^* = \dfrac{(\psi_{2j-1}, F)}{(\psi_{2j-1}, \psi_{2j-1})} = & & \frac{1}{\pi}\int_0^{2\pi} F(\theta)\cos j\theta d\theta \ , & \dfrac{2}{m+1}\displaystyle\sum_{k=0}^{m} F(\theta_k)\cos j\theta_k \\ b_j^* = \dfrac{(\psi_{2j}, F)}{(\psi_{2j}, \psi_{2j})} = & & \frac{1}{\pi}\int_0^{2\pi} F(\theta)\operatorname{sen} j\theta d\theta \ , & \dfrac{2}{m+1}\displaystyle\sum_{k=0}^{m} F(\theta_k)\operatorname{sen} j\theta_k \end{array}$$

Debido a que la aproximación $t_n^*(\theta)$ así obtenida es 2π-periódica, la aproximación trigonométrica es especialmente útil para aproximar funciones $F : \mathbb{R} \longrightarrow \mathbb{R}$ 2π-periódicas (véase el problema 3.15). En este caso, el intervalo $[0, 2\pi]$ puede substituirse por cualquier intervalo de la forma $[a, a + 2\pi]$ (por ejemplo, $[-\pi, \pi]$ es muy usual) sin que esto altere la aproximación $t_n^*(\theta)$.

Otra base importante de funciones trigonométricas ortogonales es la formada por las funciones $\psi_j(\theta) = \cos j\theta$ $(j \ge 0)$. Es ortogonal respecto al producto escalar continuo en el intervalo $I = [0, \pi]$

$$(F, G) = \int_0^{\pi} F(\theta)G(\theta)d\theta \ ,$$

con $\|\psi_j\|^2 = \frac{\pi}{2}$, si $j > 0$ y $\|\psi_0\|^2 = \pi$.

Las funciones $\psi_j(\theta) = \cos j\theta$ $(j = 0 \div m)$ son también ortogonales respecto al producto escalar discreto

$$(F, G)_m = \sum_{k=0}^{m} F(\theta_k)G(\theta_k) \ , \quad \theta_k = \frac{(2k+1)\pi}{2(m+1)} \quad (k = 0 \div m) \ ,$$

con $\|\psi_j\|_m^2 = \frac{m+1}{2}$ $(j = 1 \div m)$ y $\|\psi_0\|_m^2 = m + 1$.

3.2.4 Aproximación minimax

Definición del problema

Dado un intervalo acotado $I = [a, b]$ y unas funciones φ_j $(j = 0 \div n)$, para cada función continua f definida en I, buscamos $\hat{f}_n \in \mathcal{F}_n$ (el espacio generado por aquéllas) de manera que

$$\|f - f_n\|_\infty \equiv \sup_{x \in I} |f(x) - f_n(x)|$$

sea mínima en $\mathcal{F}_n$ para $f_n = \hat{f}_n$; esto es,

$$\|f - \hat{f}_n\|_\infty = \min_{f_n \in \mathcal{F}_n} \|f - f_n\|_\infty \ .$$

Es necesario decir aquí que sólo consideraremos el caso de la norma del máximo sobre un intervalo, con función peso $w(x) = 1$, entre todas las normas del máximo introducidas en el apartado 3.2.1, y nos concentraremos en los casos de aproximación polinomial y trigonométrica, porque tenemos una condición especial sobre sus ceros, la propiedad de Haar:

a) Si $p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$ es un polinomio de grado menor o igual que n, no puede tener más de n ceros a menos que sea idénticamente nulo.

Esta propiedad fue deducida en el apartado 3.1.2, usando sólo la regla de Horner y, por esto, también es válida para el caso de coeficientes complejos a_j $(j = 0 \div n)$ y ceros complejos.

b) Si

$$t_n(\theta) = \frac{a_0}{2} + \sum_{j=1}^{n} a_j \cos j\theta + \sum_{j=1}^{n} b_j \operatorname{sen} j\theta = \sum_{j=-n}^{n} c_j e^{ij\theta}$$

es un *polinomio trigonométrico* de grado menor o igual que n, no puede tener más de $2n$ ceros en ningún intervalo $[a, a + 2\pi)$, a menos que sea el polinomio nulo.

Esta propiedad es consecuencia de a) aplicada sobre el polinomio de grado menor o igual que $2n$

$$p_{2n}(x) = \sum_{j=0}^{n} c_{n+j} x^j \ ,$$

ya que $t_n(\theta) = e^{-in\theta} p_{2n}(e^{i\theta}) = 0$ implica $p_{2n}(e^{i\theta}) = 0$.

Caracterización

Usando la propiedad de Haar, puede deducirse el *teorema de Chebichev* que caracteriza las aproximaciones minimax polinomial y trigonométrica:

- Dada $f : I \longrightarrow \mathbb{R}$ continua, sea $\hat{p}_n(x)$ un polinomio de grado menor o igual que n tal que la función $e_n(x) = f(x) - \hat{p}_n(x)$ tenga al menos $n + 2$ *abscisas de extremo* en (a, b) $a \leq \xi_0 < \xi_1 < \cdots < \xi_{n+1} \leq b$ cumpliendo la *propiedad de equioscilación*

  $$|e_n(\xi_j)| = \|e_n\|_\infty = \max_{x \in I} |e_n(x)| \ \ (j = 0 \div n + 1) \ ,$$

  con alternancia de signos

  $$e_n(\xi_0) = -e_n(\xi_1) = \cdots = (-1)^{n+1} e_n(\xi_{n+1}) \ ;$$

  entonces, la norma del supremo sobre el intervalo I de e_n se minimiza dentro del espacio $\mathcal{P}_n$ de polinomios $p_n(x)$ de grado menor o igual que n para $p_n = \hat{p}_n$; esto es,

  $$\|f - \hat{p}_n\|_\infty \leq \|f - p_n\|_\infty \ .$$

- Dada $F : I = [0, 2\pi] \longrightarrow \mathbb{R}$ continua con $F(0) = F(2\pi)$, sea $\hat{t}_n(\theta)$ un polinomio trigonométrico de grado menor o igual que n tal que la función $e_n(x) = F(x) - \hat{t}_n(x)$ tenga al menos $2n + 2$ *abscisas de extremo* en $(0, 2\pi)$ $0 \leq \eta_0 < \eta_1 < \cdots < \eta_{2n+1} \leq 2\pi$ cumpliendo la *propiedad de equioscilación*

  $$|e_n(\eta_j)| = \|e_n\|_\infty = \max_{x \in I} |e_n(x)| \ \ (j = 0 \div 2n + 1) \ ,$$

con alternancia de signos

$$e_n(\eta_0) = -e_n(\eta_1) = \cdots = -e_n(\eta_{2n+1}) \ ;$$

entonces, la norma del máximo sobre el intervalo I de e_n se minimiza dentro del espacio $\mathcal{T}_n$ de polinomios trigonométricos t_n de grado menor o igual que n para $t_n = \hat{t}_n$; esto es,

$$\|F - \hat{t}_n\|_\infty \leq \|F - t_n\|_\infty \ .$$

Ambos resultados se obtienen por reducción al absurdo. Por ejemplo, en el primero, supongamos que existe un polinomio $p_n(x)$ tal que

$$\|f - p_n\|_\infty < \|f - \hat{p}_n\|_\infty \ ;$$

entonces,

$$|f(\xi_j) - p_n(\xi_j)| < |f(\xi_j) - \hat{p}_n(\xi_j)| = \|e_n\|_\infty \ (j = 0 \div n+1) \ .$$

Por lo tanto, el polinomio $p_n - \hat{p}_n = (f - \hat{p}_n) - (f - p_n) \in \mathcal{P}_n$ tiene el mismo signo en ξ_j $(j = 0 \div n+1)$ que la función error $e_n = f - \hat{p}_n$. Así, $p_n - \hat{p}_n$ tiene al menos $n+1$ ceros en (a, b), ello contradice la propiedad de Haar y nos lleva a un absurdo.

Se puede demostrar también que los problemas de aproximación minimax presentados tienen solución única, ésta se caracteriza por la propiedad de oscilación uniforme para la función error, que hemos dado aquí como condición suficiente. Ahora bien, el cálculo concreto de los polinomios $\hat{p}_n(x)$ y de los polinomios trigonométricos $\hat{t}_n(\theta)$ requiere, en general, el uso de un cierto tipo de métodos iterativos debidos a Remes. No obstante, veremos a continuación algunos casos en los que pueden calcularse explícitamente las aproximaciones minimax.

EJEMPLOS

1. Sea $f \in C^2(a, b)$ tal que $f^{(2)}(x) \neq 0 \ \forall x \in [a, b]$. Si llamamos c a la solución única de

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

en (a, b), entonces la aproximación lineal minimax a f sobre el intervalo $[a, b]$ viene dada por $\hat{p}_1(x) = \hat{a}_0 + \hat{a}_1 x$, donde

$$\hat{a}_1 = \frac{f(b) - f(a)}{b - a} = f'(c) \ , \quad 2\hat{a}_0 = f(a) + f(c) + \hat{a}_1(a + c)$$

(véase el problema 3.19).

2. Sea

$$F(\theta) = a_0 + \sum_{j=1}^{n+1} a_j \cos j\theta + \sum_{j=1}^{n+1} b_j \operatorname{sen} j\theta$$

un polinomio trigonométrico de grado menor o igual que $n+1$; entonces, la aproximación minimax trigonométrica, de grado menor o igual que n, a F sobre $[0, 2\pi]$ viene dada por

$$\begin{aligned} \hat{t}_n(\theta) &= F(\theta) - a_{n+1} \cos(n+1)\theta - b_{n+1} \operatorname{sen}(n+1)\theta \\ &= a_0 + \sum_{j=1}^{n} a_j \cos j\theta + \sum_{j=1}^{n} b_j \operatorname{sen} j\theta \end{aligned}$$

(es decir, se halla simplemente suprimiendo los términos de orden $n+1$).

En efecto, notemos que, con esta elección de $\hat{t}_n(\theta)$, la función error es

$$e_n(\theta) = a_{n+1}\cos(n+1)\theta + b_{n+1}\,\mathrm{sen}(n+1)\theta = A\cos((n+1)\theta - \theta_0)\ ;$$

donde A y θ_0 son las coordenadas "polares" del punto (a_{n+1}, b_{n+1}):

$$a_{n+1} = A\cos\theta_0\ ,\quad b_{n+1} = A\,\mathrm{sen}\,\theta_0\ .$$

Dado que $\cos((n+1)\theta - \theta_0)$ tiene exactamente $2n+2$ abscisas de extremo en $[0, 2\pi]$, en las que toma alternativamente los valores ± 1, el teorema de Chebichev asegura que

$$\|f - \hat{t}_n\|_\infty \leq \|f - t_n\|_\infty \quad \forall t_n \in \mathcal{T}_n\ .$$

3. Sea ahora

$$F(\theta) = a_0 + \sum_{j=1}^{n+1} a_j \cos j\theta$$

un polinomio trigonométrico de grado menor o igual que $n+1$, que es combinación lineal de sólo las funciones $\psi_j(\theta) = \cos j\theta$ $(j = 0 \div n+1)$; entonces, el polinomio trigonométrico

$$\hat{t}_n(\theta) = a_0 + \sum_{j=1}^{n} a_j \cos j\theta$$

da la aproximación minimax trigonométrica de grado menor o igual que n sobre el intervalo $[0, 2\pi]$. Al ser F y $\hat{t}_n(\theta)$ *polinomios trigonométricos en cosenos*, ambas funciones quedan totalmente determinadas por sus valores sobre el intervalo $[0, \pi]$, ya que $F(\pi+\theta) = F(\pi-\theta)$ y análogamente para $\hat{t}_n$; en otras palabras, se trata de funciones simétricas respecto a $\theta = \pi$.

Polinomios de Chebichev

Este último ejemplo permite encontrar una estrecha relación de los polinomios de Chebichev con la aproximación minimax.

Sobre el intervalo $[0, \pi]$ tiene sentido pensar en el cambio de variable

$$x = \cos\theta \in [-1, 1] \Leftrightarrow \theta = \arccos x \in [0, \pi]\ ,$$

ya que la función coseno es biyectiva entre $[0, \pi]$ y $[-1, 1]$; los polinomios de Chebichev se encuentran haciendo este cambio sobre las funciones $\psi_j(\theta) = \cos j\theta$ $(j \geq 0)$:

$$T_j(t) = \psi_j(\arccos t) = \cos(j \arccos t) \quad \forall t \in [-1, 1]\ .$$

Observamos que $T_0(t) = 1$ y que $T_1(t) = t$.

Recordamos que las funciones ψ_j $(j \geq 0)$ forman una base de funciones trigonométricas ortogonales respecto al producto escalar continuo

$$(F, G) = \int_0^\pi F(\theta)G(\theta)d\theta$$

y que las $m+1$ primeras son también ortogonales respecto al producto escalar discreto

$$(F,G)_m = \sum_{k=0}^{m} F(\theta_k)G(\theta_k) \ , \quad \theta_k = \frac{(2k+1)\pi}{2(m+1)} \quad (k = 0 \div m) \ .$$

Además:

$$\|\psi_0\|_2^2 = \pi \ , \quad \|\psi_j\|_2^2 = \frac{\pi}{2} \ (j > 0) \ ;$$

$$\|\psi_0\|_{2,m}^2 = m+1 \ , \quad \|\psi_j\|_{2,m}^2 = \frac{m+1}{2} \ (j = 1 \div m) \ .$$

Usando las relaciones trigonométricas (3.36), las funciones ψ_j $(j \geq 0)$ satisfacen la relación de recurrencia

$$\psi_{j+1} = 2\psi_1\psi_j - \psi_{j-1} \quad (j \geq 1) \ ;$$

así resulta que cada $T_j(t)$ es un polinomio de grado menor o igual que j (que puede definirse para todo $t \in \mathbb{R}$) y que se llama *polinomio de Chebichev* de grado j.

Los polinomios de Chebichev satisfacen, pues, las propiedades siguientes:

- Si $t \in [-1,1]$, $T_j(t) = \cos(j \arccos t)$ $(j \geq 1)$.
- $T_0(t) = 1$, $T_1(t) = t$, $T_{j+1}(t) = 2tT_j(t) - T_{j-1}(t)$ $(j \geq 1)$. Esta relación asegura que los polinomios de Chebichev tienen *paridad definida*:

$$T_j(-t) = (-1)^j T_j(t) \quad (j \geq 0) \ .$$

- El coeficiente principal de $T_j(t)$ es 2^{j-1}: $T_j(t) = 2^{j-1}t^j + \cdots$ $(j \geq 0)$.
- Los polinomios $T_j(t)$ $(j \geq 0)$ son ortogonales respecto al producto escalar continuo

$$(f,g) = \int_{-1}^{1} \frac{f(t)g(t)}{\sqrt{1-t^2}} dt \ .$$

Además:

$$\|T_0\|_2^2 = \pi \ , \quad \|T_j\|_2^2 = \frac{\pi}{2} \ (j > 0) \ .$$

- Los polinomios $T_j(t)$ $(j = 0 \div m)$ son ortogonales respecto al producto escalar discreto

$$(f,g)_m = \sum_{k=0}^{m} f(t_k)g(t_k) \ ,$$

donde $t_k = \cos \frac{2k+1}{m+1}\frac{\pi}{2}$ $(k = 0 \div m)$ son los $m+1$ ceros de $T_{m+1}(t)$.

Además:

$$\|T_0\|_{2,m}^2 = m+1 \ , \quad \|T_j\|_{2,m}^2 = \frac{m+1}{2} \ (j = 1 \div m) \ .$$

Esta propiedad y la anterior se deducen de la de ortogonalidad de las funciones ψ_j, mencionadas antes, via el cambio $t = \cos\theta$.

- Sea $p_{n+1}(x) = a_0 + a_1 x + \cdots + a_{n+1}x^{n+1}$ un polinomio de grado menor o igual que $n+1$; entonces,
$$\hat{p}_n(x) = p_{n+1}(x) - \frac{a_{n+1}}{2^{n+1}} T_{n+1}(x)$$
da la aproximación minimax de grado menor o igual que n sobre el intervalo $[-1,1]$, debido a que $\hat{p}_n \in \mathcal{P}_n$ y $T_{n+1}(t) = \cos((n+1)\arccos t)$ tiene exactamente $n+2$ abscisas de extremo $\xi_k = \cos\frac{k\pi}{n+1}$ $(k = 0 \div n+1)$ sobre el intervalo $[-1,1]$, donde toma alternativamente los valores ± 1.

  Otra manera de expresar esta propiedad es la siguiente:

- El polinomio mónico de Chebichev de grado $n+1$
$$\frac{1}{2^n} T_{n+1}(t)$$
minimiza la norma del máximo en el intervalo $[-1,1]$ sobre todos los polinomios mónicos de grado $n+1$.

  En efecto, si $p_{n+1}(t) = t^{n+1} - p_n(t)$ es un polinomio mónico de grado $n+1$, entonces
$$\begin{aligned} \|\frac{T_{n+1}}{2^n}\|_\infty &= \|t^{n+1} - (t^{n+1} - \frac{T_{n+1}}{2^n})\|_\infty \\ &\leq \|t^{n+1} - p_n\|_\infty = \|p_{n+1}\|_\infty \ . \end{aligned}$$

Economización de Lanczos

Si podemos aproximar una función f por un polinomio en un intervalo $[a,b]$ (por ejemplo, hallando su polinomio de Taylor de grado n cerca de un punto a) de manera que
$$f(x) = b_0 + b_1(x-a) + b_2(x-a)^2 + \cdots + b_n(x-a)^n + r_n(x) \ , \quad x \in [a,b] \ ;$$
entonces, haciendo el cambio afín de variables
$$\frac{x-a}{b-a} = \frac{t+1}{2} \quad (x \in [a,b] \Leftrightarrow t \in [-1,1]) \ ,$$
podemos trasladarlo todo al intervalo $[-1,1]$
$$g(t) = a_0 + a_1 t + \cdots + a_n t^n + s_n(t) \ , \quad t \in [-1,1] \ ,$$
donde
$$g(t) = f(\frac{a+b}{2} + \frac{b-a}{2}t) \ , \quad s_n(t) = r_n(\frac{a+b}{2} + \frac{b-a}{2}t) \ ,$$
y aproximar
$$g_n(t) = a_0 + a_1 t + \cdots + a_n t^n$$
por
$$g_{n-1}(t) = g_n(t) - \frac{a_n}{2^{n-1}} T_n(t) \ .$$

Tenemos así
$$g(t) = g_{n-1}(t) + e_n(t) + s_n(t) \ , \quad t \in [-1,1] \ ,$$

con $|e_n(t)| \leq \frac{a_n}{2^{n-1}}$, $t \in [-1,1]$.

Este proceso puede repetirse sobre g_{n-1}

$$g(t) = g_{n-2}(t) + e_{n-1}(t) + e_n(t) + s_n(t)$$

y así sucesivamente, siempre que los errores $e_j(t)$ $(j \leq n)$ sean suficientemente pequeños respecto a la precisión exigida en los cálculos. La ventaja de estas nuevas expresiones es que no requieren tantos cálculos como $g_n(t)$.

El proceso anterior puede expresarse, de forma equivalente, escribiendo $g_n(t)$ en la base ortogonal de polinomios de Chebichev

$$g(t) = a_0 + a_1 T_1(t) + a_2 T_2(t) + \cdots + a_n T_n(t) + s_n(t) \ ;$$

entonces, $e_j(t) = a_j T_j(t)$ $(j \leq n)$. y, si $|s_n(t)| \leq C \ \forall t \in [-1,1]$,

$$\begin{aligned} g(t) &= a_0 + a_1 T_1(t) + \cdots + a_{n-2} T_{n-2}(t) \\ &\quad \pm \left(\frac{|a_{n-1}|}{2^{n-2}} + \frac{|a_n|}{2^{n-1}} + C \right) \quad \forall t \in [-1,1] \ . \end{aligned}$$

Interpolación de Chebichev

La *interpolación de Chebichev* de grado menor o igual que m en el intervalo $[-1,1]$ consiste en interpolar una función $g : [-1,1] \longrightarrow \mathbb{R}$ por un polinomio de grado menor o igual que m en $m+1$ *abscisas de Chebichev*; esto es, en los ceros de $T_{m+1}(t)$:

$$t_k = \cos \frac{(2k+1)\pi}{2(m+1)} \quad (k = 0 \div m) \ .$$

El error cometido es así

$$\begin{aligned} g(t) - p_m(t) &= \frac{g^{(m+1)}(\xi(t))}{(m+1)!}(t-t_0)(t-t_1)\cdots(t-t_m) \\ &= \frac{g^{(m+1)}(\xi(x))}{(m+1)!}\frac{T_{m+1}(t)}{2^m} \end{aligned}$$

donde $\xi(x) \in (-1,1)$.

Este error puede acotarse por

$$\|g - p_m\|_\infty \leq \frac{\|g^{(m+1)}\|_\infty}{2^m (m+1)!} \ .$$

Observamos que, en la interpolación de Chebichev se hace mínima la norma del máximo de la función $(t-t_0)\cdots(t-t_m)$ entre todas las posibles elecciones de $m+1$ abscisas de interpolación en el intervalo $[-1,1]$.

Además, dado que los polinomios $T_j(t)$ $(j = 0 \div m)$ son ortogonales respecto al producto escalar discreto

$$(f,g)_m = \sum_{k=0}^{m} f(t_k) g(t_k) \ ,$$

los polinomios de aproximación discreta por mínimos cuadrados vienen dados directamente por

$$p_n^*(t) = \sum_{j=0}^{n} c_j^* T_j(t) \ ,$$

con

$$c_j^* = \frac{(f, T_j)_m}{(T_j, T_j)_m} \quad (j = 0 \div n) \ .$$

El caso límite $n = m$ da lugar a la interpolación de Chebichev de grado menor o igual que m, que se expresa ahora como

$$p_m^*(t) = \sum_{j=0}^{m} c_j^* T_j(t) \ ,$$

con

$$c_j^* = \frac{(f, T_j)_m}{(T_j, T_j)_m} \quad (j = 0 \div m) \ .$$

Si la función que queremos interpolar está definida en un intervalo $[a, b]$, el cambio afín de variables

$$\frac{x - a}{b - a} = \frac{t + 1}{2}$$

da las $m + 1$ abscisas de Chebichev sobre el intervalo $[a, b]$:

$$x_k = \frac{a + b}{2} + \frac{b - a}{2} t_k = \frac{a + b}{2} + \frac{b - a}{2} \cos \frac{(2k + 1)\pi}{2(m + 1)} \quad (k = 0 \div m) \ ;$$

ahora, los polinomios ortogonales respecto al producto escalar

$$(f, g)_m = \sum_{k=0}^{m} f(x_k) g(x_k)$$

vienen dados por

$$T_j(-1 + \frac{2}{b - a}(x - a)) \quad (j = 0 \div m)$$

.

La interpolación de Chebichev en el intervalo $[a, b]$ puede obtenerse también usando la misma técnica de aproximación discreta por mínimos cuadrados (véase el problema 3.22). Conviene destacar la eficiencia de este método de mínimos cuadrados, cuando quieren hallarse polinomios de interpolación de Chebichev, respecto a otros métodos de interpolación.

COMENTARIOS BIBLIOGRÁFICOS

Una de las referencias más completas para aproximación e interpolación es [Dav75] y también [Che66], [Ham73], [Hil74], [IK66], [RR78]. Una buena introducción a la interpolación se encuentra en [Hen64]. En muchos libros clásicos, el enfoque dado del cálculo del polinomio interpolador está basado en el uso de las diferencias finitas y también de los operadores lineales, conceptos que son introducidos en el capítulo siguiente. Una referencia básica sobre polinomios ortogonales es [Sze59]. Varias tablas y métodos numéricos para su cálculo se muestran en [AS65], uno de los manuales más completos sobre funciones matemáticas. Los algoritmos de Remes pueden encontrarse detallados en [Ral65].

PROBLEMAS RESUELTOS

Problema 3.1 *a) Calcular $f(3)$ por interpolación cuadrática de la tabla*

x_k	1	2	4	5
f_k	0	2	12	21

i) utilizando los valores en $x = 1,\ 2,\ 4$;

ii) utilizando los valores en $x = 2,\ 4,\ 5$.

b) Calcular $f(3)$ por interpolación cúbica.

SOLUCIÓN:

Usaremos el método de Lagrange.
a) i) En este caso, $m = 2$:

$$x_0 = 1,\ f_0 = 0;\quad x_1 = 2,\ f_1 = 2;\quad x_2 = 4,\ f_2 = 12.$$

Los polinomios de Lagrange asociados a las abscisas de la tabla son:

$$\begin{aligned} l_0(x) &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{1}{3}(8-6x+x^2)\ , \\ l_1(x) &= \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{1}{-2}(4-5x+x^2)\ , \\ l_2(x) &= \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{1}{6}(2-3x+x^2)\ , \end{aligned}$$

y el polinomio interpolador es

$$p_2(x) = f_0 l_0(x) + f_1 l_1(x) + f_2 l_2(x) = -x + x^2\ .$$

El polinomio interpolador en $x = 3$ vale 6.
ii) En este caso, también $m = 2$:

$$x_0 = 2,\ f_0 = 2;\quad x_1 = 4,\ f_1 = 12;\quad x_2 = 5,\ f_2 = 21.$$

Los polinomios de Lagrange asociados a las abscisas de la tabla son:

$$\begin{aligned} l_0(x) &= \frac{1}{6}(20-9x+x^2)\ , \\ l_1(x) &= \frac{1}{-2}(10-7x+x^2)\ , \\ l_2(x) &= \frac{1}{3}(8-6x+x^2)\ , \end{aligned}$$

y el polinomio interpolador es

$$p_2(x) = \frac{1}{3}(8 - 9x + 4x^2) \ .$$

El polinomio interpolador en $x = 3$ vale $\frac{17}{3} = 5.666...$

b) En este caso, $m = 3$:

$$x_0 = 1, \ f_0 = 0; \quad x_1 = 2, \ f_1 = 2; \quad x_2 = 4, \ f_2 = 12; \quad x_3 = 5, \ f_3 = 21.$$

Los polinomios de Lagrange asociados a las abscisas de la tabla son:

$$\begin{aligned} l_0(x) &= \frac{1}{-12}(-40 + 38x - 11x^2 + x^3) \ , \\ l_1(x) &= \frac{1}{6}(-20 + 29x - 10x^2 + x^3) \ , \\ l_2(x) &= \frac{1}{-6}(-10 + 17x - 8x^2 + x^3) \ , \\ l_3(x) &= \frac{1}{12}(-8 + 14x - 7x^2 + x^3) \ , \end{aligned}$$

y el polinomio interpolador es

$$p_3(x) = \frac{1}{12}(-8 + 2x + 5x^2 + x^3) \ .$$

El polinomio interpolador en $x = 3$ vale $\frac{35}{6} = 5.8333...$

Problema 3.2 *Dada la siguiente tabla de la función $f(x) = e^x$:*

x	0.0	0.2	0.4	0.6
$f(x)$	1.0000	1.2214	1.4918	1.8221

.

a) Hallar valores aproximados de $\sqrt[3]{e}$ por interpolación lineal y cúbica, usando los métodos de Lagrange y de Newton.

b) Dar cotas respectivas de los errores debidos a la interpolación. Comparar dichas cotas con el error exacto, sabiendo que $\sqrt[3]{e} = 1.395612425\ldots$

SOLUCIÓN:

a) Tomaremos todos los puntos para hacer la interpolación cúbica y sólo los puntos centrales de la tabla para la lineal.

INTERPOLACIÓN LINEAL POR EL MÉTODO DE LAGRANGE

$$x_0 = 0.2,\ f_0 = 1.2214;\quad x_1 = 0.4,\ f_1 = 1.4918.$$

Los polinomios de Lagrange asociados a las abscisas de la tabla son:

$$\begin{aligned} l_0(x) &= \frac{1}{-0.2}(-0.4 + x) = 2 - 5x\ , \\ l_1(x) &= \frac{1}{0.2}(-0.2 + x) = 5x - 1\ , \end{aligned}$$

y el polinomio interpolador es

$$p_1(x) = 0.9510 + 1.3520x\ .$$

El polinomio interpolador en $x = \frac{1}{3}$ vale $p_1(\frac{1}{3}) = \frac{4.205}{3} = 1.401666...$

Interpolación cúbica por el método de Lagrange

$$x_0 = 0,\ f_0 = 1;\quad x_1 = 0.2,\ f_1 = 1.2214;$$

$$x_2 = 0.4,\ f_2 = 1.4918;\quad x_3 = 0.6,\ f_3 = 1.8221.$$

Los polinomios de Lagrange asociados a las abscisas de la tabla son:

$$\begin{aligned} l_0(x) &= \frac{1}{-0.048}(-0.048 + 0.44x - 1.2x^2 + x^3)\ , \\ l_1(x) &= \frac{1}{0.016}(0.24x - x^2 + x^3)\ , \\ l_2(x) &= \frac{1}{-0.016}(0.12x - 0.8x^2 + x^3)\ , \\ l_3(x) &= \frac{1}{0.048}(0.08x - 0.6x^2 + x^3)\ , \end{aligned}$$

y el polinomio interpolador es

$$p_3(x) = \frac{1}{48}(48 + 48.128x + 22.86x^2 + 10.9x^3)\ .$$

El polinomio interpolador en $x = \frac{1}{3}$ vale $p_3(\frac{1}{3}) = \frac{113.0395}{81} \simeq 1.395549$.

Interpolaciones lineal y cúbica por el método de las diferencias divididas de Newton

La tabla de diferencias divididas de Newton es:

0.0	1.0000			
		1.107		
0.2	1.2214		0.6125	
		1.352		$\frac{0.68125}{3}$
0.4	1.4918		0.74875	
		1.6515		
0.6	1.8221			

Los polinomios interpoladores serán entonces:

$$\begin{aligned} p_1(x) &= 1.2214 + 1.352(x - 0.2) \ , \\ p_3(x) &= 1 + 1.107x + 0.6125x(x - 0.2) \\ &\quad + \frac{0.68125}{3} x(x - 0.2)(x - 0.4) \ , \end{aligned}$$

que, en $x = \frac{1}{3}$, valen $\frac{4.205}{3}$ y $\frac{113.0395}{81}$, respectivamente, igual que antes.

b) La expresión para el error de interpolación puede acotarse por

$$\frac{e}{2} \left| (\frac{1}{3} - 0.2)(\frac{1}{3} - 0.4) \right| = \frac{e}{225} \simeq 0.012 \ ,$$

en la interpolación lineal, y por

$$\frac{e}{24} \left| (\frac{1}{3} - 0.0)(\frac{1}{3} - 0.2)(\frac{1}{3} - 0.4)(\frac{1}{3} - 0.6) \right| = \frac{e}{30375} \simeq 0.000089 \ ,$$

en la cúbica.

Comparando con el resultado exacto dado $1.395612425\ldots$, los errores en las interpolaciones lineal y cúbica son, respectivamente:

$$\begin{aligned} p_1(\frac{1}{3}) - \sqrt[3]{e} &= 0.0061 \ , \\ p_3(\frac{1}{3}) - \sqrt[3]{e} &= 0.000063 \ . \end{aligned}$$

Las cotas halladas son, pues, aproximadamente el doble y una vez y media de los valores de los errores exactos, respectivamente.

Problema 3.3 *Hallar los polinomios de interpolación en abscisas equidistantes de grados 1, 2, 3 y 4 a las funciones:*

a) e^x, *en* $[0, 1]$;

b) $\operatorname{sen} x$, *en* $[0, \pi/2]$;

c) e^{x^2}, *en* $[0, 1]$;

d) $\cos(\cos x)$, *en* $[0, \pi/2]$;

usando los métodos de interpolación de Lagrange y de las diferencias divididas de Newton.

Acotar el error cometido en cualquier punto de los intervalos respectivos.

SOLUCIÓN:

MÉTODO DE LAGRANGE

a) Para la función $f(x) = e^x$, que suponemos conocida en cada punto del intervalo $[0, 1]$ con 6 decimales, obtendremos los polinomios de interpolación.

En el caso $m = 1$, los valores de la función en las abscisas $x_i = \frac{i}{1}$ $(i = 0 \div 1)$ son:

$$f_0 = 1,\ f_1 = 2.718282,$$

y los polinomios de Lagrange asociados a dichas abscisas vienen dados por:

$$\begin{aligned} l_0(x) &= 1 - x\ , \\ l_1(x) &= x\ . \end{aligned}$$

El polinomio interpolador es

$$p_1(x) = 1 + 1.718282x\ .$$

En el caso $m = 2$, los valores de la función en las abscisas $x_i = \frac{i}{2}$ $(i = 0 \div 2)$ son:

$$f_0 = 1,\ f_1 = 1.648721,\ f_2 = 2.718282,$$

y los polinomios de Lagrange asociados a dichas abscisas vienen dados por:

$$\begin{aligned} l_0(x) &= 1 - 3x + 2x^2\ , \\ l_1(x) &= 4x - 4x^2\ , \\ l_2(x) &= -x + 2x^2\ . \end{aligned}$$

El polinomio interpolador es

$$p_2(x) = 1 + 0.8766033x + 0.8416786x^2\ .$$

En el caso $m = 3$, los valores de la función en las abscisas $x_i = \frac{i}{3}$ $(i = 0 \div 3)$ son:

$$f_0 = 1,\ f_1 = 1.395612,\ f_2 = 1.947734,\ f_3 = 2.718282,$$

y los polinomios de Lagrange asociados a dichas abscisas vienen dados por:

$$\begin{aligned} l_0(x) &= 1 - 5.5x + 9x^2 - 4.5x^3\ , \\ l_1(x) &= 9x - 22.5x^2 + 13.5x^3\ , \\ l_2(x) &= -4.5x + 18x^2 - 13.5x^3\ , \\ l_3(x) &= x - 4.5x^2 + 4.5x^3\ . \end{aligned}$$

El polinomio interpolador es

$$p_3(x) = 1 + 1.01399x + 0.4256649x^2 + 0.2786264x^3\ .$$

En el caso $m = 4$, los valores de la función en las abscisas $x_i = \frac{i}{4}$ $(i = 0 \div 4)$ son:

$$f_0 = 1,\ f_1 = 1.284025,\ f_2 = 1.648721,\ f_3 = 2.117,\ f_4 = 2.718282,$$

y los polinomios de Lagrange asociados a dichas abscisas vienen dados por:

$$\begin{aligned} l_0(x) &= 1 - 8.333333x + 23.33333x^2 - 26.66667x^3 + 10.66667x^4 \ , \\ l_1(x) &= 16x - 69.33333x^2 + 96x^3 - 42.66667x^4 \ , \\ l_2(x) &= -12x + 76x^2 - 128x^3 + 64x^4 \ , \\ l_3(x) &= 5.333333x - 37.33333x^2 + 74.66667x^3 - 42.66667x^4 \ , \\ l_4(x) &= -x + 7.333333x^2 - 16x^3 + 10.66667x^4 \ . \end{aligned}$$

El polinomio interpolador es

$$p_4(x) = 1 + 0.998803x + 0.5097871x^2 + 0.140276x^3 + 0.06941567x^4 \ .$$

b) Para la función $f(x) = \operatorname{sen} x$, calculada con 7 decimales en el intervalo $[0, \frac{\pi}{2}]$, obtendremos los polinomios interpoladores.

En el caso $m = 1$, los valores de la función en las abscisas $x_i = \frac{i\pi}{2}$ $(i = 0 \div 1)$ son:

$$f_0 = 0, \ f_1 = 1,$$

y los polinomios de Lagrange asociados a dichas abscisas vienen dados por:

$$\begin{aligned} l_0(x) &= 1 - 0.6366198x \ , \\ l_1(x) &= 0.6366198x \ . \end{aligned}$$

El polinomio interpolador es

$$p_1(x) = 0.6366198x \ .$$

En el caso $m = 2$, los valores de la función en las abscisas $x_i = \frac{i\pi}{4}$ $(i = 0 \div 2)$ son:

$$f_0 = 0, \ f_1 = 0.7071068, \ f_2 = 1,$$

y los polinomios de Lagrange asociados a dichas abscisas vienen dados por:

$$\begin{aligned} l_0(x) &= 1 - 1.909859x + 0.8105695x^2 \ , \\ l_1(x) &= 2.546479x - 1.621139x^2 \ , \\ l_2(x) &= -0.6366198x + 0.8105695x^2 \ . \end{aligned}$$

El polinomio interpolador es

$$p_2(x) = 1.164013x - 0.3357489x^2 \ .$$

En el caso $m = 3$, los valores de la función en las abscisas $x_i = \frac{i\pi}{6}$ $(i = 0 \div 3)$ son:

$$f_0 = 0, \ f_1 = 0.5, \ f_2 = 0.8660254, \ f_3 = 1,$$

y los polinomios de Lagrange asociados a dichas abscisas vienen dados por:

$$\begin{aligned} l_0(x) &= 1 - 3.501409x + 3.647563x^2 - 1.161055x^3 \ , \\ l_1(x) &= 5.729578x - 9.118907x^2 + 3.483166x^3 \ , \\ l_2(x) &= -2.864789x + 7.295125x^2 - 3.483166x^3 \ , \\ l_3(x) &= 0.6366198x - 1.823781x^2 + 1.161055x^3 \ . \end{aligned}$$

El polinomio interpolador es

$$p_3(x) = 1.020429x - 0.0654708x^2 - 0.1138719x^3 \ .$$

En el caso $m = 4$, los valores de la función en las abscisas $x_i = \frac{i\pi}{8}$ $(i = 0 \div 4)$ son:

$$f_0 = 0,\ f_1 = 0.3826834,\ f_2 = 0.7071068,\ f_3 = 0.9238795,\ f_4 = 1,$$

y los polinomios de Lagrange asociados a dichas abscisas vienen dados por:

$$\begin{aligned}
l_0(x) &= 1 - 5.305165x + 9.456644x^2 - 6.880327x^3 + 1.752061x^4 \ , \\
l_1(x) &= 10.18592x - 28.09974x^2 + 24.76918x^3 - 7.008244x^4 \ , \\
l_2(x) &= -7.639437x + 30.80164x^2 - 33.02557x^3 + 10.51237x^4 \ , \\
l_3(x) &= 3.395305x - 15.13063x^2 + 19.26492x^3 - 7.008244x^4 \ , \\
l_4(x) &= -0.6366198x + 2.972088x^2 - 4.128196x^3 + 1.752061x^4 \ .
\end{aligned}$$

El polinomio interpolador es

$$p_4(x) = 0.996317x + 0.01995143x^2 - 0.2035855x^3 + 0.02871423x^4 \ .$$

Nótese que todos los polinomios de Lagrange de este apartado se obtienen substituyendo x por $\frac{2}{\pi}x \simeq 0.636619772x$ en las expresiones de los respectivos polinomios de Lagrange del apartado a).

c) Para la función $f(x) = \exp(x^2)$ en el intervalo $[0, 1]$, calculada con 6 cifras decimales exactas, ya no escribimos los polinomios de Lagrange debido a que son los mismos que en el apartado a).

En el caso $m = 1$, los valores de la función en las abscisas $x_i = \frac{i}{1}$ $(i = 0 \div 1)$ son:

$$f_0 = 1,\ f_1 = 2.718282.$$

El polinomio interpolador es

$$p_1(x) = 1 + 1.718282x \ .$$

En el caso $m = 2$, los valores de la función en las abscisas $x_i = \frac{i}{2}$ $(i = 0 \div 2)$ son:

$$f_0 = 1,\ f_1 = 1.284025,\ f_2 = 2.718282.$$

El polinomio interpolador es

$$p_2(x) = 1 - 0.5821802x + 2.300462x^2 \ .$$

En el caso $m = 3$, los valores de la función en las abscisas $x_i = \frac{i}{3}$ $(i = 0 \div 3)$ son:

$$f_0 = 1,\ f_1 = 1.117519,\ f_2 = 1.559623,\ f_3 = 2.718282.$$

El polinomio interpolador es

$$p_3(x) = 1 + 0.2576477x - 0.3032243x^2 + 1.763858x^3 \ .$$

En el caso $m = 4$, los valores de la función en las abscisas $x_i = \frac{i}{4}$ $(i = 0 \div 4)$ son:

$$f_0 = 1,\ f_1 = 1.064494,\ f_2 = 1.284025,$$

$$f_3 = 1.755055,\ f_4 = 2.718282.$$

El polinomio interpolador es

$$p_4(x) = 1 - 0.06771732x + 1.526342x^2 - 1.27888x^3 + 1.538537x^4 \ .$$

d) Para la función $f(x) = \cos(\cos(x))$, calculada con 6 cifras decimales exactas en el intervalo $[0, \frac{\pi}{2}]$, tampoco escribiremos los polinomios de Lagrange, debido a que aparecen en el apartado b).

En el caso $m = 1$, los valores de la función en las abscisas $x_i = \frac{i\pi}{2}$ $(i = 0 \div 1)$ son:

$$f_0 = 0.5403023,\ f_1 = 1.$$

El polinomio interpolador es

$$p_1(x) = 0.5403023 + 0.2926526x \ .$$

En el caso $m = 2$, los valores de la función en las abscisas $x_i = \frac{i\pi}{4}$ $(i = 0 \div 2)$ son:

$$f_0 = 0.5403023,\ f_1 = 0.7602446,\ f_2 = 1.$$

El polinomio interpolador es

$$p_2(x) = 0.5403023 + 0.2674258x + 0.0160599x^2 \ .$$

En el caso $m = 3$ los valores de la función en las abscisas $x_i = \frac{i\pi}{6}$ $(i = 0 \div 3)$ son:

$$f_0 = 0.5403023,\ f_1 = 0.6478593,\ f_2 = 0.8775826,\ f_3 = 1.$$

El polinomio interpolador es

$$p_3(x) = 0.5403023 - 0.05732768x + 0.6413111x^2 - 0.2664296x^3 \ .$$

En el caso $m = 4$, los valores de la función en las abscisas $x_i = \frac{i\pi}{8}$ $(i = 0 \div 4)$ son:

$$f_0 = 0.5403023,\ f_1 = 0.602729,\ f_2 = 0.7602446,$$

$$f_3 = 0.927666,\ f_4 = 1.$$

El polinomio interpolador es

$$\begin{aligned} p_4(x) \ = \ & 0.5403023 - 0.02179655x + 0.525614x^2 \\ & -0.1526548x^3 - 0.0347085x^4 \ . \end{aligned}$$

Método de las diferencias divididas de Newton

a) Las tablas de diferencias divididas para la función $f(x) = e^x$ en el intervalo $[0, 1]$ se dan a continuación para los valores de $m = 1, 2, 3, 4$, así como los polinomios interpoladores resultantes:

$$\begin{array}{c|cc} 0 & 1.000000 & \\ & & 1.718282 \\ 1 & 2.718282 & \end{array}$$

$$p_1(x) = 1 + 1.718282x \ .$$

$$\begin{array}{c|ccc} 0 & 1.000000 & & \\ & & 1.297443 & \\ 0.5 & 1.648721 & & 0.841679 \\ & & 2.139121 & \\ 1 & 2.718282 & & \end{array}$$

$$p_2(x) = 1 + 1.297443x + 0.841679x(x - \frac{1}{2}) \ .$$

$$\begin{array}{c|ccccc} 0 & 1.000000 & & & & \\ & & 1.186837 & & & \\ 0.333333 & 1.395612 & & 0.704291 & & \\ & & 1.656365 & & 0.278626 & \\ 0.666666 & 1.947734 & & 0.982918 & & \\ & & 2.311643 & & & \\ 1 & 2.718282 & & & & \end{array}$$

$$p_3(x) = 1 + 1.186837x + 0.704291x(x - \frac{1}{3}) + 0.278626x(x - \frac{1}{3})(x - \frac{2}{3}) \ .$$

$$\begin{array}{c|ccccc} 0 & 1.000000 & & & & \\ & & 1.136102 & & & \\ 0.25 & 1.284025 & & 0.645363 & & \\ & & 1.458783 & & 0.244400 & \\ 0.5 & 1.648721 & & 0.828663 & & 0.069416 \\ & & 1.873115 & & 0.313815 & \\ 0.75 & 2.117000 & & 1.064025 & & \\ & & 2.405127 & & & \\ 1 & 2.718282 & & & & \end{array}$$

$$\begin{aligned} p_4(x) \quad = \quad & 1 + 1.136102x + 0.645363x(x - \frac{1}{4}) \\ & +0.244400x(x - \frac{1}{4})(x - \frac{1}{2}) \\ & +0.069416x(x - \frac{1}{4})(x - \frac{1}{2})(x - \frac{3}{4}) \ . \end{aligned}$$

b) Las tablas de diferencias divididas para la función $f(x) = \operatorname{sen} x$ en el intervalo $[0, \frac{\pi}{2}]$ se dan a continuación para los valores de $m = 1, 2, 3, 4$, así como los polinomios interpoladores resultantes:

$$\begin{array}{c|cc} 0 & 0.000000 & \\ & & 0.636620 \\ \frac{\pi}{2} & 1.000000 & \end{array}$$

$$p_1(x) = 0.636620x \ .$$

$$\begin{array}{c|ccc} 0 & 0.000000 & & \\ & & 0.900316 & \\ \frac{\pi}{4} & 0.707107 & & -0.335749 \\ & & 0.372923 & \\ \frac{\pi}{2} & 1.000000 & & \end{array}$$

$$p_2(x) = 0.900316x - 0.335749x(x - \frac{\pi}{4}) \ .$$

$$\begin{array}{c|cccc} 0 & 0.000000 & & & \\ & & 0.954930 & & \\ \frac{\pi}{6} & 0.500000 & & -0.244340 & \\ & & 0.699057 & & -0.113872 \\ \frac{\pi}{3} & 0.866025 & & -0.423210 & \\ & & 0.255873 & & \\ \frac{\pi}{2} & 1.000000 & & & \end{array}$$

$$p_3(x) = 0.954930x - 0.244340x(x - \frac{\pi}{6}) - 0.113872x(x - \frac{\pi}{6})(x - \frac{\pi}{3}) \ .$$

$$\begin{array}{c|ccccc} 0 & 0.000000 & & & & \\ & & 0.974495 & & & \\ \frac{\pi}{8} & 0.382683 & & -0.188895 & & \\ & & 0.826137 & & -0.135929 & \\ \frac{\pi}{4} & 0.707107 & & -0.349033 & & 0.028714 \\ & & 0.552007 & & -0.090825 & \\ \frac{3\pi}{8} & 0.923880 & & -0.456034 & & \\ & & 0.193839 & & & \\ \frac{\pi}{2} & 1.000000 & & & & \end{array}$$

$$\begin{aligned} p_4(x) \quad = \quad & 0.974495x - 0.188895x(x - \frac{\pi}{8}) \\ & -0.135929x(x - \frac{\pi}{8})(x - \frac{\pi}{4}) \\ & +0.028714x(x - \frac{\pi}{8})(x - \frac{\pi}{4})(x - \frac{3\pi}{8}) \ . \end{aligned}$$

c) Las tablas de diferencias divididas para la función $f(x) = \exp(x^2)$ en el intervalo $[0, 1]$ se dan a continuación para los valores de $m = 1, 2, 3, 4$, así como los polinomios interpoladores resultantes:

$$\begin{array}{c|cc} 0 & 1.000000 & \\ & & 1.718282 \\ 1 & 2.718282 & \end{array}$$

$$p_1(x) = 1 + 1.718282x \ .$$

$$\begin{array}{c|cc} 0 & 1.000000 & \\ & & 0.568051 \\ 0.5 & 1.284025 & \\ & & 2.868513 \\ 1 & 2.718282 & \end{array}$$

$$p_2(x) = 1 + 0.568051x + 2.300462(x - \frac{1}{2}) \ .$$

$$\begin{array}{c|ccc} 0 & 1.000000 & & \\ & & 0.352557 & \\ 0.333333 & 1.117519 & & 1.460634 \\ & & 1.326313 & \\ 0.666666 & 1.559623 & & 3.224493 \\ & & 3.475975 & \\ 1 & 2.718282 & & \end{array}$$

$$p_3(x) = 1 + 0.352557x + 1.460634x(x - \frac{1}{3}) + 1.763858x(x - \frac{1}{3})(x - \frac{2}{3}) \ .$$

$$\begin{array}{c|ccccc} 0 & 1.000000 & & & & \\ & & 0.257978 & & & \\ 0.25 & 1.064494 & & 1.240292 & & \\ & & 0.878124 & & 1.028926 & \\ 0.5 & 1.284025 & & 2.011986 & & 1.538537 \\ & & 1.884117 & & 2.567463 & \\ 0.75 & 1.755055 & & 3.937583 & & \\ & & 3.852909 & & & \\ 1 & 2.718282 & & & & \end{array}$$

$$\begin{aligned} p_4(x) \quad = \quad & 1 + 0.257978x + 1.240292x(x - \frac{1}{4}) \\ & +1.028926x(x - \frac{1}{4})(x - \frac{1}{2}) \\ & +1.538537x(x - \frac{1}{4})(x - \frac{1}{2})(x - \frac{3}{4}) \ . \end{aligned}$$

d) Las tablas de diferencias divididas para la función $f(x) = \cos(\cos x)$ en el intervalo $[0, \frac{\pi}{2}]$ se dan a continuación para los valores de $m = 1, 2, 3, 4$, así como los polinomios interpoladores resultantes:

$$\begin{array}{c|cc} 0 & 0.540302 & \\ & & 0.292653 \\ \frac{\pi}{2} & 1.000000 & \end{array}$$

$$p_1(x) = 0.540302 + 0.292653x \ .$$

$$\begin{array}{c|ccc} 0 & 0.540302 & & \\ & & 0.280039 & \\ \frac{\pi}{4} & 0.760245 & & 0.016060 \\ & & 0.305266 & \\ \frac{\pi}{2} & 1.000000 & & \end{array}$$

$$p_2(x) = 0.545302 + 0.280039x + 0.016060x(x - \frac{\pi}{4}) \ .$$

$$\begin{array}{c|cccc}
0 & 0.540302 & & & \\
 & & 0.205419 & & \\
\frac{\pi}{6} & 0.647859 & & 0.222804 & \\
 & & 0.438739 & & -0.266430 \\
\frac{\pi}{3} & 0.877583 & & -0.195702 & \\
 & & 0.233800 & & \\
\frac{\pi}{2} & 1.000000 & & &
\end{array}$$

$$\begin{aligned}
p_3(x) &= 0.540302 + 0.205419x + 0.222804x(x - \frac{\pi}{6}) \\
&\quad -0.266430x(x - \frac{\pi}{6})(x - \frac{\pi}{3}) \ .
\end{aligned}$$

$$\begin{array}{c|ccccc}
0 & 0.540302 & & & & \\
 & & 0.158968 & & & \\
\frac{\pi}{8} & 0.602729 & & 0.308304 & & \\
 & & 0.401110 & & -0.234435 & \\
\frac{\pi}{4} & 0.760245 & & 0.032117 & & -0.034708 \\
 & & 0.426335 & & -0.288955 & \\
\frac{3\pi}{8} & 0.927666 & & -0.308299 & & \\
 & & 0.184197 & & & \\
\frac{\pi}{2} & 1.000000 & & & &
\end{array}$$

$$\begin{aligned}
p_4(x) &= 0.540302 + 0.158968x + 0.308304x(x - \frac{\pi}{8}) \\
&\quad -0.234435x(x - \frac{\pi}{8})(x - \frac{\pi}{4}) \\
&\quad -0.034708x(x - \frac{\pi}{8})(x - \frac{\pi}{4})(x - \frac{3\pi}{8}) \ .
\end{aligned}$$

Errores cometidos

Seguidamente pasamos a acotar los errores en todos los casos usando la fórmula del error de interpolación.

Acotamos en un intervalo cualquiera $[a, b]$ el factor del error para abscisas equidistantes

$$\omega_m(x) = (x - x_0)(x - x_1) \cdots (x - x_m) \ .$$

Lo podemos considerar escrito, si hacemos el cambio de variable $x = c + h_m u$, como

$$h_m^{m+1}(u + \frac{m}{2})(u + \frac{m}{2} - 1) \cdots (u - \frac{m}{2})$$

en el intervalo $[-\frac{m}{2}, \frac{m}{2}]$, donde hemos hecho $h_m = \frac{b-a}{m}$ $(m = 1, 2, 3, 4)$ y $c = \frac{a+b}{2}$.

Definiendo

$$\bar{\omega}_m(u) \equiv (u + \frac{m}{2})(u + \frac{m}{2} - 1) \cdots (u - \frac{m}{2}) \ ,$$

su valor absoluto, en el intervalo $[-\frac{m}{2}, \frac{m}{2}]$, presenta máximos en las abscisas:

$$u^* = 0,\ \pm\frac{\sqrt{3}}{3}, \pm\frac{\sqrt{5}}{2}, \pm\sqrt{\frac{15+\sqrt{145}}{10}}\quad (m = 1, 2, 3, 4)\ ;$$

los valores respectivos de estos máximos resultan ser:

$$\Omega_m = |\bar{\omega}_m(u^*)| = \frac{1}{4},\ \frac{2\sqrt{3}}{9},\ 1,\ \frac{\sqrt{950+58\sqrt{145}}}{5\sqrt{5}}\quad (m = 1, 2, 3, 4)\ .$$

Así pues, el factor $\omega_m(x)$ queda acotado por $h_m^{m+1}\Omega_m$; de manera que, si M_{m+1} es una cota de la función $f^{(m+1)}$ en todo el intervalo, el error de interpolación en cualquier punto del intervalo está acotado por

$$\frac{M_{m+1}}{(m+1)!}h_m^{m+1}\Omega_m\ .$$

a) Para la función $f(x) = e^x$ en el intervalo $[0, 1]$ puede tomarse $M_{m+1} = e$, para todos los valores de m, y $h_m = \frac{1}{m}$.

Las cotas de los errores para los diferentes valores de m serán, pues:

$m = 1$	$\frac{e}{2}h_1^2\Omega_1 \simeq 0.341$
$m = 2$	$\frac{e}{6}h_2^3\Omega_2 \simeq 0.022$
$m = 3$	$\frac{e}{24}h_3^4\Omega_3 \simeq 0.0014$
$m = 4$	$\frac{e}{120}h_4^5\Omega_4 \simeq 0.000080$

b) Para $f(x) = \operatorname{sen} x$ en $[0, \frac{\pi}{2}]$, $M_{m+1} = 1$ y $h_m = \frac{\pi}{2m}$:

$m = 1$	$\frac{1}{2}h_1^2\Omega_1 \simeq 0.31$
$m = 2$	$\frac{1}{6}h_2^3\Omega_2 \simeq 0.031$
$m = 3$	$\frac{1}{24}h_3^4\Omega_3 \simeq 0.0031$
$m = 4$	$\frac{1}{120}h_4^5\Omega_4 \simeq 0.00028$

c) Para $f(x) = \exp(x^2)$ en $[0,1]$, $h_m = \frac{1}{m}$ y tenemos:

$$\begin{aligned} f'(x) &= 2xe^{x^2}, \\ f^{(2)}(x) &= (2+4x^2)e^{x^2} \leq 6e \equiv M_2, \\ f^{(3)}(x) &= (12x+8x^3)e^{x^2} \leq 20e \equiv M_3, \\ f^{(4)}(x) &= (12+48x^2+16x^4)e^{x^2} \leq 76e \equiv M_4, \\ f^{(5)}(x) &= (120x+160x^3+32x^5)e^{x^2} \leq 312e \equiv M_5; \end{aligned}$$

así,

$m=1$	$M_2 \frac{1}{2} h_1^2 \Omega_1 \simeq 2.04$
$m=2$	$M_3 \frac{1}{6} h_2^3 \Omega_2 \simeq 0.44$
$m=3$	$M_4 \frac{1}{24} h_3^4 \Omega_3 \simeq 0.11$
$m=4$	$M_5 \frac{1}{120} h_4^5 \Omega_4 \simeq 0.025$

d) Para $f(x) = \cos(\cos x)$ en $[0, \frac{\pi}{2}]$, $h_m = \frac{\pi}{2m}$ y

$$\begin{aligned} f'(x) &= s(c)s \ , \\ f^{(2)}(x) &= -c(c)s^2 + s(c)c \ , \\ f^{(3)}(x) &= -3c(c)sc - s(c)(s^3+s) = -3c(c)sc - s(c)s(s^2+1) \ , \\ f^{(4)}(x) &= c(c)(s^4+4s^2-3c^2) - s(c)(6s^2c+c) \\ &= c(c)(s^4+7s^2-3) - 6s(c)(sc)s - s(c)c \ , \\ f^{(5)}(x) &= c(c)(10s^3c+15sc) + s(c)(s^5+10s^3-15sc^2+s) \\ &= c(c)sc(10s^2+15) + s(c)(s^5+25s^3-14s) \ ; \end{aligned}$$

donde se ha usado $s(c) = \operatorname{sen}(\cos x)$, $c(c) = \cos(\cos x)$, $s = \operatorname{sen} x$ y $c = \cos x$.

Acotando $s(c)$ por $\operatorname{sen} 1$, $c(c)$ por 1 y sc por $\frac{1}{2}$, y acotando correctamente los polinomios en s, podemos tomar:

$$M_2 = 1 + \operatorname{sen} 1 \ , \quad M_3 = \frac{3}{2} + 2\operatorname{sen} 1 \ , \quad M_4 = 5 + 4\operatorname{sen} 1 \ , \quad M_5 = \frac{25}{2} + 12\operatorname{sen} 1 \ .$$

Así,

$m=1$	$M_2\frac{1}{2}h_1^2\Omega_1 \simeq 0.57$
$m=2$	$M_3\frac{1}{6}h_2^3\Omega_2 \simeq 0.099$
$m=3$	$M_4\frac{1}{24}h_3^4\Omega_3 \simeq 0.026$
$m=4$	$M_5\frac{1}{120}h_4^5\Omega_4 \simeq 0.0064$

Problema 3.4 *Disponemos de la tabla siguiente:*

x	2.0	2.1	2.2	2.3	2.4	2.5
$J_0(x)$	0.2239	0.1666	0.1104	0.0555	0.0025	-0.0484

de la función de Bessel de orden 0

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

usar el método de las diferencias divididas de Newton para hallar los valores de $J_0(2.15)$, $J_0(2.25)$ *y* $J_0(2.35)$ *con errores menores que* $\frac{1}{2}10^{-3}$.

SOLUCIÓN:

Se propone hacer la siguiente elección de abscisas para los cálculos pedidos: si se quiere calcular $J_0(x)$ para $x = 2.15, 2.25, 2.35$, tomamos las abscisas consecutivas de la tabla x_0, x_1, ..., x_m de manera que $x_0 < x < x_1$.

La expresión del error en la interpolación

$$J_0(x) - p_m(x) = \frac{J_0^{(m+1)}(\xi(x))}{(m+1)!}\omega_m(x)$$

nos permitirá hallar el valor de m para el que puede asegurarse que este error sea menor que $\frac{1}{2}10^{-3}$.

Una cota de este error, para $x = 2.15, 2.25, 2.35$, se halla haciendo

$$\begin{aligned} |\, J_0(x) - p_m(x)\,| &\leq \frac{|\, J_0^{(m+1)}(\xi(x))\,|}{(m+1)!}\,|\,\omega_m(x)\,| \\ &\leq \frac{M_{m+1}}{(m+1)!}h^{m+1}\frac{1}{2}\frac{1}{2}\frac{3}{2}\cdots\frac{2m-1}{2} \ . \end{aligned}$$

Hallemos seguidamente cotas para las derivadas de la función J_0:

$$\begin{aligned}
J_0'(x) &= -\frac{1}{\pi}\int_0^\pi \operatorname{sen}(x\operatorname{sen}t)\operatorname{sen}t\ dt\ , \\
|J_0'(x)| &\le \frac{1}{\pi}\int_0^\pi \operatorname{sen}t\ dt = \frac{2}{\pi}\ ; \\
J_0^{(2)}(x) &= \frac{1}{\pi}\int_0^\pi \cos(x\operatorname{sen}t)\operatorname{sen}^2 t\ dt\ , \\
|J_0^{(2)}(x)| &\le \frac{1}{\pi}\int_0^\pi \operatorname{sen}^2 t\ dt = \frac{1}{2}\ ; \\
J_0^{(3)}(x) &= -\frac{1}{\pi}\int_0^\pi \operatorname{sen}(x\operatorname{sen}t)\operatorname{sen}^3 t\ dt\ , \\
|J_0^{(3)}(x)| &\le \frac{1}{\pi}\int_0^\pi \operatorname{sen}^3 t\ dt = \frac{4}{3\pi}\ .
\end{aligned}$$

Así, tomando $M_2 = \frac{1}{2}$ y $M_3 = \frac{4}{3\pi}$, tenemos:

$$\begin{aligned}
|J_0(x) - p_1(x)| &\le \frac{1}{2\cdot 2}(0.1)^2\frac{1}{4} = \frac{1}{1600}\ , \\
|J_0(x) - p_2(x)| &\le \frac{4}{3\pi\cdot 6}(0.1)^3\frac{3}{8} = \frac{1}{12\pi}10^{-3} < \frac{1}{2}10^{-3}\ ;
\end{aligned}$$

por lo tanto, no parece suficiente hacer una interpolación lineal. En cambio, una interpolación cuadrática produce un error sensiblemente inferior a la cota pedida.

Construimos la tabla de diferencias divididas hasta diferencias de segundo orden usando todas las abscisas excepto la primera, que no se necesita,

2.1	0.1666		
		−0.562	
2.2	0.1104		0.065
		−0.549	
2.3	0.0555		0.095
		−0.530	
2.4	0.0025		0.105
		−0.509	
2.5	−0.0484		

Las aproximaciones pedidas, halladas por interpolación de segundo grado según se ha indicado, son:

$$\begin{aligned}
J_0(2.15) &\simeq 0.1666 - 0.562(2.15-2.1) + 0.065(2.15-2.1)(2.15-2.2) \\
&= 0.1383\ , \\
J_0(2.25) &\simeq 0.1104 - 0.549(2.25-2.2) + 0.095(2.25-2.2)(2.25-2.3) \\
&= 0.0827\ , \\
J_0(2.35) &\simeq 0.0555 - 0.530(2.35-2.3) + 0.105(2.35-2.3)(2.35-2.4) \\
&= 0.0287\ .
\end{aligned}$$

Problema 3.5 *Disponemos de una gráfica de $y = f(x)$. Una tabla digitalizadora da las coordenadas de los puntos de la curva con un error absoluto (en x, y) menor que 0.02 dm. Cinco puntos de aquella gráfica dados por la tabla son:*

x_k (en dm)	0.20	0.50	1.00	1.20	1.80
y_k (en dm)	0.80	0.80	0.60	0.40	0.60

Hallar una cota del error absoluto del valor de y, correspondiente a la abscisa $x = 1.50$, dado por el polinomio interpolador en los puntos de la gráfica, debido únicamente a los errores de las medidas.

SOLUCIÓN:

Para encontrar la cota pedida haremos un estudio aproximado de la propagación de los errores de los datos en el valor del polinomio interpolador en la abscisa $x = 1.50$, supuesto exacto, y sin tener en cuenta los errores en las operaciones.

Se elige la fórmula de interpolación de Lagrange para su cálculo,

$$p(x) = \sum_{i=0}^{4} y_i l_i(x) = \sum_{i=0}^{4} y_i \prod_{k \neq i} \frac{x - x_k}{x_i - x_k} \ .$$

Usando las fórmulas aproximadas para el error en sumas, productos y divisiones, hallaremos una cota para el error absoluto en $p(x)$ en función de la cota del error absoluto en la medida de las coordenadas de los puntos $\epsilon = 0.02$dm:

$$\begin{aligned}
\epsilon_a(p(x)) &= \sum_{i=0}^{4} |\, l_i(x) \,|\, \epsilon_a(y_i) + \sum_{i=0}^{4} |\, y_i \,|\, \epsilon_a(\prod_{k \neq i} \frac{x - x_k}{x_i - x_k}) \\
&= \sum_{i=0}^{4} |\, l_i(x) \,|\, \epsilon_a(y_i) \\
&\quad + \sum_{i=0}^{4} |\, l_i(x) \,|\, |\, y_i \,| \sum_{k \neq i} \left(\frac{\epsilon_a(x_k)}{|\, x - x_k \,|} + \frac{\epsilon_a(x_i)}{|\, x_i - x_k \,|} + \frac{\epsilon_a(x_k)}{|\, x_i - x_k \,|} \right) \\
&= \sum_{i=0}^{4} |\, l_i(x) \,| \left[1 + |\, y_i \,| \sum_{k \neq i} \left(\frac{1}{|\, x - x_k \,|} + \frac{2}{|\, x_i - x_k \,|} \right) \right] \epsilon
\end{aligned}$$

Substituyendo las coordenadas por los valores de la tabla, tenemos

$$\begin{aligned}
\epsilon_a(p(x)) = 0.02 \cdot & \\
& \left[\tfrac{1.0 \cdot 0.5 \cdot 0.3 \cdot 0.3}{0.3 \cdot 0.8 \cdot 1.0 \cdot 1.6} \left[1 + 0.8 \left(\tfrac{1}{1.0} + \tfrac{1}{0.5} + \tfrac{1}{0.3} + \tfrac{1}{0.3} + \tfrac{2}{0.3} + \tfrac{2}{0.8} + \tfrac{2}{1.0} + \tfrac{2}{1.6} \right) \right] \right. \\
& + \tfrac{1.3 \cdot 0.5 \cdot 0.3 \cdot 0.3}{0.3 \cdot 0.5 \cdot 0.7 \cdot 1.3} \left[1 + 0.8 \left(\tfrac{1}{1.3} + \tfrac{1}{0.5} + \tfrac{1}{0.3} + \tfrac{1}{0.3} + \tfrac{2}{0.3} + \tfrac{2}{0.5} + \tfrac{2}{0.7} + \tfrac{2}{1.3} \right) \right] \\
& + \tfrac{1.3 \cdot 1.0 \cdot 0.3 \cdot 0.3}{0.8 \cdot 0.5 \cdot 0.2 \cdot 0.8} \left[1 + 0.6 \left(\tfrac{1}{1.3} + \tfrac{1}{1.0} + \tfrac{1}{0.3} + \tfrac{1}{0.3} + \tfrac{2}{0.8} + \tfrac{2}{0.5} + \tfrac{2}{0.2} + \tfrac{2}{0.8} \right) \right] \\
& \left. + \tfrac{1.3 \cdot 1.0 \cdot 0.5 \cdot 0.3}{1.0 \cdot 0.7 \cdot 0.2 \cdot 0.6} \left[1 + 0.4 \left(\tfrac{1}{1.3} + \tfrac{1}{1.0} + \tfrac{1}{0.5} + \tfrac{1}{0.3} + \tfrac{2}{1.0} + \tfrac{2}{0.7} + \tfrac{2}{0.2} + \tfrac{2}{0.6} \right) \right] \right] \ .
\end{aligned}$$

Una vez realizados los cálculos resulta

$$\epsilon_a(p(x)) \simeq 71\epsilon \simeq 0.35 \ ,$$

una cota de error extremadamente grande si observamos que el valor aproximado encontrado para $p(1.50)$ es 0.198.

Este error tan grande es debido principalmente al error en las abscisas. Si éstas fuesen exactas, puede comprobarse fácilmente que

$$\epsilon_a(p(1.5)) \leq \epsilon \sum_{k=0}^{4} |\, l_k(1.5)\,| = 5\epsilon = 2.5 \cdot 10^{-2} \ ,$$

apreciablemente menor que el resultado obtenido anteriormente.

Problema 3.6 *Queremos tabular la* función error

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

en abscisas equidistantes en el intervalo $[0,1]$ *con paso* $h = 0.01$. *Hallar el número máximo* t *de cifras decimales redondeadas con las que han de darse los valores de la tabla para que el error debido a la interpolación por polinomios de grados 1, 3 y 5 (usando los puntos más cercanos) no exceda del error propagado en el cálculo del polinomio interpolador debido al error de redondeo de los datos.*

SOLUCIÓN:

Acotaremos primero las derivadas sucesivas de la función $f(x) = \mathrm{erf}(x)$ para poder acotar después las expresiones de los errores de interpolación para $x \in [0,1]$:

$$\begin{aligned} |\, f(x) - p_1(x)\,| &\leq \frac{M_2}{2}(0.01)^2\Omega_1 \ , \\ |\, f(x) - p_3(x)\,| &\leq \frac{M_4}{24}(0.01)^4\Omega_3 \ , \\ |\, f(x) - p_5(x)\,| &\leq \frac{M_6}{720}(0.01)^6\Omega_5 \ ; \end{aligned}$$

donde M_{m+1} es una cota del valor absoluto de $f^{(m+1)}$ en el intervalo $[0,1]$ y los Ω_m $(m = 1,3,5)$ están definidos de forma análoga a los del problema 3.3. Notemos que aquí el máximo de ω_m se busca en el subintervalo central $[-\frac{1}{2}, \frac{1}{2}]$ en la variable u, definida en aquel problema.

En el caso de interpolación de grado $m = 2s+1$ a f en x, escogeremos las abscisas x_0, $\ldots$, x_{2s+1} de forma que x esté en el subintervalo central x_s, x_{s+1} de centro $c = \frac{x_s + x_{s+1}}{2}$. La variable u cumple, pues, $x - c = hu$, con $h = 0.01$.

Debido a la simetría, el máximo se alcanza ahora en el punto central $u = 0$ correspondiente a $x = c$. Resultan así: $\Omega_1 = \frac{1}{4}$, $\Omega_3 = \frac{9}{16}$, $\Omega_5 = \frac{225}{64}$.

Las primeras derivadas de $f(x)$ son:

$$\begin{aligned} f'(x) &= \frac{2}{\sqrt{\pi}} e^{-x^2}, \\ f^{(2)}(x) &= -\frac{4}{\sqrt{\pi}} x e^{-x^2}, \\ f^{(3)}(x) &= \frac{2}{\sqrt{\pi}} (-2 + 4x^2) e^{-x^2}, \\ f^{(4)}(x) &= -\frac{2}{\sqrt{\pi}} (-12x + 8x^3) e^{-x^2}, \\ f^{(5)}(x) &= \frac{2}{\sqrt{\pi}} (12 - 48x^2 + 16x^4) e^{-x^2}, \\ f^{(6)}(x) &= -\frac{2}{\sqrt{\pi}} (120x - 160x^3 + 32x^5) e^{-x^2} \ . \end{aligned}$$

Los polinomios $H_j(x)$ $(j \geq 0)$ que cumplen

$$f^{(j+1)}(x) = (-1)^j \frac{2}{\sqrt{\pi}} H_j(x) e^{-x^2} \ ,$$

o equivalentemente

$$H_j(x) = (-1)^j e^{x^2} \frac{d^j}{dx^j} e^{-x^2} \ ,$$

se llaman polinomios de Hermite y se tratan en los problemas propuestos.

Así las cotas M_2, M_4 y M_6 de las derivadas segunda, cuarta y sexta, requeridas en el cálculo, pueden hallarse buscando los máximos y mínimos locales de estas funciones derivadas en el intervalo $[0, 1]$; éstos se dan en los ceros de las funciones derivadas de éstas y, por lo tanto, en los ceros de $H_2(x)$, $H_4(x)$ y $H_6(x)$, que son respectivament $\frac{\sqrt{2}}{2} = 0.7071...$, $\sqrt{\frac{3-\sqrt{6}}{2}} = 0.5246...$ y $0.4361...$

Las cotas resultantes para los errores e_m^I debidos a la interpolación serán:

$$\begin{aligned} \epsilon_1^I &\simeq \frac{|f^{(2)}(0.7071)|}{2} 10^{-4} \frac{1}{4} \simeq 0.12 \cdot 10^{-4} \ , \\ \epsilon_3^I &\simeq \frac{|f^{(4)}(0.5246)|}{24} 10^{-8} \frac{9}{16} \simeq 0.62 \cdot 10^{-9} \ , \\ \epsilon_5^I &\simeq \frac{|f^{(6)}(0.4361)|}{720} 10^{-12} \frac{225}{64} \simeq 0.19 \cdot 10^{-12} \ . \end{aligned}$$

En el caso de interpolación lineal, el error e_1^R propagado en

$$p_1(x) = f(x_0) l_0(x) + f(x_1) l_1(x)$$

estará acotado por

$$|\, l_0(x) \,|\, \epsilon + |\, l_1(x) \,|\, \epsilon \leq 2\epsilon = \epsilon_1^R \ ,$$

si los datos tienen un error absoluto acotado por ϵ, de acuerdo con el problema 3.5.

En el caso de interpolación de grado $m = 3, 5$, el error propagado e_m^R en

$$p_m(x) = \sum_{i=0}^{m} f(x_i) l_i(x) \ ,$$

debido al error de redondeo en $f(x_i)$ $(i = 0 \div m)$, estará acotado por

$$\epsilon_m^R = \epsilon \sum_{i=0}^{m} |\, l_i(x) \,| \ ,$$

que puede demostrarse que es menor que 2.5ϵ, si x pertenece al subintervalo central.

Parece, pues, adecuado dar los valores de la función con $t = 4, 8, 12$ cifras decimales redondeadas, respectivamente; entonces, una cota del error de redondeo de los datos es $\epsilon = \frac{1}{2}10^{-t}$ y las cotas ϵ_m^R no son sobrepasadas por las cotas ϵ_m^I encontradas para $m = 1, 3, 5$.

Problema 3.7 *a) Dar, de forma explícita, los polinomios de Taylor en el punto $x = 0$ de las funciones siguientes:*

i) $\cosh x = \frac{e^x + e^{-x}}{2}$, $\operatorname{senh} x = \frac{e^x - e^{-x}}{2}$;

ii) $(1 + x)^\alpha$ $(\alpha = 2, \frac{1}{2}, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{3})$.

b) Hallar expresiones de los errores y cotas de éstos.

c) Estudiar la convergencia de los desarrollos de Taylor de las funciones del apartado a).

d) Evaluar en $x = 0.001$ las funciones dadas en el apartado a), con un error menor que 10^{-20}.

SOLUCIÓN:

a) i) Tomamos primero $f(x) = \cosh x$. Las derivadas sucesivas de esta función son $f'(x) = \operatorname{senh} x$, $f^{(2)}(x) = \cosh x$, ...: las derivadas pares son $f^{(2r)}(x) = \cosh x$ y las impares, $f^{(2r+1)}(x) = \operatorname{senh} x$ $(r \geq 0)$. En el punto $x = x_0 = 0$, tendremos, pues,

$$f^{(2r)}(0) = 1 \ , \quad f^{(2r+1)}(0) = 0 \quad (r \geq 0) \ .$$

Ello permite ya escribir los polinomios de Taylor pedidos

$$p_{2s}(x) = p_{2s+1}(x) = 1 + \frac{1}{2}x^2 + \frac{1}{24}x^4 + \cdots + \frac{x^{2s}}{(2s)!} \ .$$

Para $f(x) = \operatorname{senh} x$, es fácil observar que los polinomios de Taylor son

$$p_{2s+1}(x) = p_{2s+2}(x) = x + \frac{1}{6}x^3 + \cdots + \frac{x^{2s+1}}{(2s+1)!} \quad (s \geq 0) \ .$$

ii) Ahora $f(x) = (1+x)^\alpha$, la derivada j-ésima es

$$f^j(x) = \alpha(\alpha-1)\dots(\alpha-j+1)(1+x)^{\alpha-j}$$

y los coeficientes del polinomio de Taylor,

$$\frac{f^j(0)}{j!} = \frac{\alpha(\alpha-1)\dots(\alpha-j+1)}{j!} \equiv \left(\begin{array}{c}\alpha\\ j\end{array}\right) .$$

Los polinomios de Taylor resultan ser

$$p_n(x) = \sum_{j=0}^{n} \left(\begin{array}{c}\alpha\\ j\end{array}\right) x^j .$$

Para el caso $\alpha = 2$ tenemos:

$$p_0(x) = 1 \ , \quad p_1(x) = 1+2x \ , \quad p_n(x) = 1+2x+x^2 \ (n \geq 2) .$$

Para el caso $\alpha = \frac{1}{2}$, tenemos que los coeficientes del polinomio de Taylor son

$$\begin{aligned}\left(\begin{array}{c}\frac{1}{2}\\ j\end{array}\right) &= \frac{\frac{1}{2}(-\frac{1}{2})\cdots(-\frac{2j-3}{2})}{j!} = (-1)^{j-1}\frac{1\cdot 3\cdot 5\cdots(2j-3)}{2\cdot 4\cdot 6\cdots(2j)} \\ &= (-1)^{j-1}\frac{(2j-3)!!}{(2j)!!} \quad (j \geq 0) ,\end{aligned}$$

y los polinomios de Taylor,

$$p_n(x) = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3 - \frac{5}{128}x^4 + \cdots + (-1)^{n-1}\frac{(2n-3)!!}{(2n)!!}x^n .$$

Para el caso $\alpha = -\frac{1}{2}$, tenemos que los coeficientes del polinomio de Taylor son

$$\begin{aligned}\left(\begin{array}{c}-\frac{1}{2}\\ j\end{array}\right) &= \frac{(-\frac{1}{2})(-\frac{3}{2})\cdots(-\frac{2j-1}{2})}{j!} = (-1)^{j}\frac{1\cdot 3\cdot 5\cdots(2j-1)}{2\cdot 4\cdot 6\cdots(2j)} \\ &= (-1)^{j}\frac{(2j-1)!!}{(2j)!!} \quad (j \geq 0) ,\end{aligned}$$

y los polinomios de Taylor,

$$p_n(x) = 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \frac{5}{16}x^3 - \frac{35}{128}x^4 + \cdots + (-1)^{n}\frac{(2n-1)!!}{(2n)!!}x^n .$$

Para el caso $\alpha = \frac{1}{3}$, tenemos que los coeficientes del polinomio de Taylor son

$$\begin{aligned}\left(\begin{array}{c}\frac{1}{3}\\ j\end{array}\right) &= \frac{\frac{1}{3}(-\frac{2}{3})\cdots(-\frac{3j-4}{3})}{j!} = (-1)^{j-1}\frac{2\cdot 5\cdot 8\cdots(3j-4)}{3\cdot 6\cdot 9\cdots(3j)} \\ &= (-1)^{j-1}\frac{(3j-4)!!!}{(3j)!!!} \quad (j \geq 0) ,\end{aligned}$$

y los polinomios de Taylor,

$$p_n(x) = 1 + \frac{1}{3}x - \frac{1}{9}x^2 + \frac{5}{81}x^3 - \frac{10}{243}x^4 + \cdots + (-1)^{n-1}\frac{(3n-4)!!!}{(3n)!!!}x^n \ .$$

Para el caso $\alpha = -\frac{1}{3}$, tenemos que los coeficientes del polinomio de Taylor son

$$\begin{aligned}\binom{-\frac{1}{3}}{j} &= \frac{(-\frac{1}{3})(-\frac{4}{3})\cdots(-\frac{3j-2}{3})}{j!} = (-1)^j\frac{1\cdot 4\cdot 7\cdots(3j-2)}{3\cdot 6\cdot 9\cdots(3j)} \\ &= (-1)^j\frac{(3j-2)!!!}{(3j)!!!} \quad (j \geq 0) \ ,\end{aligned}$$

y los polinomios de Taylor,

$$p_n(x) = 1 - \frac{1}{3}x + \frac{2}{9}x^2 - \frac{14}{81}x^3 + \frac{35}{243}x^4 + \cdots + (-1)^n\frac{(3n-2)!!!}{(3n)!!!}x^n \ .$$

b) i) La expresión de Lagrange para los errores de los polinomios de Taylor

$$R_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}x^{n+1}$$

permite encontrar las cotas

$$|R_n(x)| \leq \frac{M_{n+1}}{(n+1)!}x^{n+1} \ ,$$

siendo M_{n+1} una cota del valor absoluto de la derivada $(n+1)$-ésima de f en el intervalo $< 0, x >$.

En el caso de que $f(x) = \cosh x$, tomando $n = 2s+1$, a los polinomios de Taylor $p_{2s}(x) = p_{2s+1}(x)$ les corresponde la expresión siguiente del error:

$$R_{2s}(x) = R_{2s+1}(x) = \frac{\cosh\xi(x)}{(2s+2)!}x^{2s+2}$$

y la cota

$$\frac{\cosh x}{(2s+2)!}x^{2s+2} \ ;$$

debido a que sus derivadas pares coinciden con la función, que es par y creciente para valores positivos de x y, por lo tanto, puede acotarse por $\cosh x$.

En el otro caso, $f(x) = \operatorname{senh} x$, tenemos análogamente

$$R_{2s+1}(x) = R_{2s+2}(x) = \frac{\cosh\xi(x)}{(2s+3)!}x^{2s+3}$$

que puede acotarse, en valor absoluto, por

$$\frac{\cosh x}{(2s+3)!}x^{2s+3} \ .$$

ii) Para $f(x) = (1+x)^\alpha$, la expresión de Lagrange para los errores

$$R_n(x) = \frac{1}{(1+\xi(x))^{n+1-\alpha}}\binom{\alpha}{n+1}x^{n+1}$$

puede acotarse por

$$\left(\begin{array}{c} \alpha \\ n+1 \end{array} \right) x^{n+1}$$

para valores de $x \geq 0$ y por

$$\frac{1}{(1-|x|)^{n+1-\alpha}} \left(\begin{array}{c} \alpha \\ n+1 \end{array} \right) x^{n+1}$$

para valores de x en $(-1,0)$. Esta última expresión toma valores arbitrariamente grandes a medida que x tiende a -1, para cada n fijo.

c) i) La convergencia de los polinomios de Taylor a la función está asegurada aquí en ambos casos, para cualquier valor de x: es fácil comprobar que las cotas encontradas para los errores tienden a cero cuando s tiende a infinito.

ii) La convergencia en este caso es obvia cuando α es natural, debido a que se trata de una suma finita. Si $x \in (0,1)$, es también fácil demostrar que la cota del error encontrada tiende a cero, cuando n tiende a infinito. No se puede llegar directamente al mismo resultado para todos los valores de $(-1,0)$ (aunque pueda hacerse de manera análoga en el intervalo $(-\frac{1}{2},0)$), así pues, es necesario seguir otro camino para decidir sobre la convergencia.

La fórmula integral es mejor en este caso; tomando $y=-x$, tenemos

$$\begin{aligned} R_n(x) &= (n+1)\left(\begin{array}{c} \alpha \\ n+1 \end{array} \right) \int_0^{-y} \left(\frac{-y-t}{1+t} \right)^n (1+t)^{\alpha-1} dt \\ &= (-1)^{n+1}(n+1)\left(\begin{array}{c} \alpha \\ n+1 \end{array} \right) \int_0^{y} u^n \left(\frac{1-y}{1-u} \right)^{\alpha} \frac{du}{1-u} \\ &= (-1)^{n+1}(n+1)\left(\begin{array}{c} \alpha \\ n+1 \end{array} \right) \frac{(1-y)^{\alpha}}{(1-\eta)^{\alpha+1}} \int_0^{y} u^n du \\ &= \left(\begin{array}{c} \alpha \\ n+1 \end{array} \right) \left(\frac{1+x}{1+\xi} \right)^{\alpha} \frac{x^{n+1}}{1+\xi} , \end{aligned}$$

donde se ha usado la variable $u = \frac{y+t}{1+t}$, $0 < \eta < y$, $x < \xi = -\eta < 0$.

Claramente, la misma expresión es válida también para $x \geq 0$, y tenemos así

$$R_n(x) = \left(\begin{array}{c} \alpha \\ n+1 \end{array} \right) \left(\frac{1+x}{1+\xi} \right)^{\alpha} \frac{x^{n+1}}{1+\xi} , \quad \xi \in <0, x> \ .$$

La expresión hallada puede acotarse ahora para $x \in (-1,1)$,

$$\begin{aligned} |R_n(x)| &= \left| \left(\begin{array}{c} \alpha \\ n+1 \end{array} \right) \right| \left| \frac{1+x}{1+\xi} \right|^{\alpha} \left| \frac{x^{n+1}}{1+\xi} \right| \\ &\leq \left(\begin{array}{c} \alpha \\ n+1 \end{array} \right) \left(\frac{1+|x|}{1-|x|} \right)^{\alpha} \frac{|x|^{n+1}}{1-|x|} . \end{aligned}$$

De esta cota se infiere la convergencia de los polinomios de Taylor a la función para todo valor de $x \in (-1,1)$.

d) i) Los cálculos de $\cosh 0.001$ y $\operatorname{senh} 0.001$, aproximando amb polinomios de Taylor adecuados, dan:

$$\begin{aligned}
\cosh 0.001 &= 1 + \frac{10^{-6}}{2} + \frac{10^{-12}}{24} \pm \cosh 0.001 \frac{10^{-18}}{720} \\
&= 1.000000500000004166667 \pm 0.5 \cdot 10^{-20} , \\
\operatorname{senh} 0.001 &= 10^{-3} + \frac{10^{-9}}{6} + \frac{10^{-15}}{120} \pm \cosh 0.001 \frac{10^{-21}}{5040} \\
&= 0.001000000166666675000000 \pm 0.2 \cdot 10^{-24} .
\end{aligned}$$

ii) El cálculo de $(1+x)^\alpha$ se realiza de la misma manera para los diferentes valores de α pedidos. Mostramos aquí sólo los cálculos para $\alpha = \frac{1}{2}$:

$$\begin{aligned}
(1+0.001)^{\frac{1}{2}} &= 1 + \frac{10^{-3}}{2} - \frac{10^{-6}}{8} + \frac{10^{-9}}{16} - \frac{5 \cdot 10^{-12}}{128} \\
&\quad + \frac{7 \cdot 10^{-15}}{256} - \frac{21 \cdot 10^{-18}}{1024} \pm \frac{33 \cdot 10^{-21}}{2048} \\
&= 1.000500125062539089864 2 \pm 10^{-22} .
\end{aligned}$$

Problema 3.8 *Calcular* $e^{0.1}$, $\ln 1.1$, $1.1^{-3/2}$ *y* $\operatorname{sen} 0.1$ *con error absoluto acotado por* 10^{-10}, *usando desarrollos de Taylor en* $x_0 = 0$.

SOLUCIÓN:

Partiremos del conocimiento de los desarrollos de Taylor de las funciones e^x, $\ln(1+x)$, $(1+x)^{-\frac{3}{2}}$ y $\operatorname{sen} x$ dados en la tabla 3.5. Los evaluaremos en $x = 0.1$ para hallar los valores pedidos, cuidando efectuar las operaciones con suficientes decimales (por ejemplo, 12), y acotaremos el error cometido:

$$\begin{aligned}
e^{0.1} &= 1 + 10^{-1} + \frac{10^{-2}}{2} + \frac{10^{-3}}{6} + \frac{10^{-4}}{24} + \frac{10^{-5}}{120} + \frac{10^{-6}}{720} \\
&\quad \pm e^{0.1} \frac{10^{-7}}{5040} \\
&= 1.10517091805 \pm 2 \cdot 10^{-11} .
\end{aligned}$$

$$\begin{aligned}
\ln 1.1 &= 10^{-1} - \frac{10^{-2}}{2} + \frac{10^{-3}}{3} - \frac{10^{-4}}{4} + \frac{10^{-5}}{5} - \frac{10^{-6}}{6} \\
&\quad + \frac{10^{-7}}{7} - \frac{10^{-8}}{8} + \frac{10^{-9}}{9} \pm \frac{10^{-10}}{10} \\
&= 0.09531017989 \pm 10^{-11} .
\end{aligned}$$

$$
\begin{aligned}
1.1^{-\frac{3}{2}} &= 1-\frac{3}{2}10^{-1}+\frac{3\cdot 5}{2\cdot 4}10^{-2}-\frac{3\cdot 5\cdot 7}{2\cdot 4\cdot 6}10^{-3} \\
&\quad +\frac{3\cdot 5\cdot 7\cdot 9}{2\cdot 4\cdot 6\cdot 8}10^{-4}+\cdots \\
&\quad +\frac{3\cdot 5\cdots 21}{2\cdot 4\cdots 20}10^{-10}\pm\frac{3\cdot 5\cdots 23}{2\cdot 4\cdots 22}10^{-11} \\
&= 1-\frac{3\cdot 10^{-1}}{2}+\frac{15\cdot 10^{-2}}{8}-\frac{35\cdot 10^{-3}}{16}+\frac{315\cdot 10^{-4}}{128} \\
&\quad -\frac{693\cdot 10^{-5}}{256}+\frac{3003\cdot 10^{-6}}{1024}-\frac{6435\cdot 10^{-7}}{2048} \\
&\quad +\frac{109395\cdot 10^{-8}}{32768}-\frac{230945\cdot 10^{-9}}{65536} \\
&\quad +\frac{969969\cdot 10^{-10}}{262144}\pm\frac{2028117\cdot 10^{-11}}{524288} \\
&= 0.86678417276\pm 4\cdot 10^{-11}\ .
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{sen} 0.1 &= 10^{-1}-\frac{10^{-3}}{6}+\frac{10^{-5}}{120}\pm\frac{10^{-7}}{5040} \\
&= 0.099833416667\pm 2\cdot 10^{-11}\ .
\end{aligned}
$$

Problema 3.9 *Hallar, como mínimo, los cinco primeros términos no nulos de los desarrollos de Taylor en $x_0=0$ de las funciones siguientes:*

a) e^{3x}, $e^x \operatorname{sen} x^2$, $x^4(1+x^2)^{\frac{1}{2}}$, $\tan(3x)$, $\operatorname{sen}^3 x$;

b) $\dfrac{\operatorname{sen} x}{x}$, $\dfrac{\cos x-1}{x^2}$, $\dfrac{\operatorname{sen} x-x-x^3/6}{x^5}$, $\ln\left(\dfrac{1+x}{1-x}\right)^{\frac{1}{x}}$;

c) $\dfrac{x^5}{\operatorname{sen}^3 x}$, $\dfrac{x(\cos x-1)}{\operatorname{sen} x-\operatorname{sen}^3 x}$, $\dfrac{(1+x)^{\frac{1}{3}}-(1-x)^{\frac{1}{3}}}{(1+x)^{\frac{1}{3}}+(1-x)^{\frac{1}{3}}}$;

d) $e^{\operatorname{sen} x}$, $e^{\sqrt{\cos x}}$, $\ln(\cos x)$.

SOLUCIÓN:

a) Usando el desarrollo de la exponencial, tenemos

$$
\begin{aligned}
e^{3x} &= 1+3x+\frac{3^2x^2}{2!}+\frac{3^3x^3}{3!}+\frac{3^4x^4}{4!}+\frac{3^5x^5}{5!}+\cdots \\
&= 1+3x+\frac{9}{2}x^2+\frac{9}{2}x^3+\frac{27}{8}x^4+\frac{81}{40}x^5+\cdots
\end{aligned}
$$

Haciendo el producto de los desarrollos de e^x y de $\operatorname{sen} x^2$, hallamos

$$\begin{aligned} e^x \operatorname{sen} x^2 &= (1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots)(x^2 - \frac{x^6}{3!} + \cdots) \\ &= x^2 + x^3 + \frac{x^4}{2} + \frac{x^5}{6} - \frac{x^6}{8} + \cdots \end{aligned}$$

De manera similar se tiene

$$\begin{aligned} x^4(1+x^2)^{\frac{1}{2}} &= x^4(1 + \frac{1}{2}x^2 - \frac{1}{2\cdot 4}x^4 + \frac{1\cdot 3}{2\cdot 4\cdot 6}x^6 \\ &\quad - \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6\cdot 8}x^8 + \frac{1\cdot 3\cdot 5\cdot 7}{2\cdot 4\cdot 6\cdot 8\cdot 10}x^{10} - \cdots) \\ &= x^4 + \frac{1}{2}x^6 - \frac{1}{8}x^8 + \frac{1}{16}x^{10} \\ &\quad - \frac{5}{128}x^{12} + \frac{7}{256}x^{14} - \cdots \end{aligned}$$

Para obtener el desarrollo de $\tan 3x$, hallamos primero el de $\tan x$, partiendo del conocimiento de los de $\operatorname{sen} x$ y de $\cos x$. Nótese que la función tangente es impar y, por lo tanto, su desarrollo está formado sólo por monomios con exponente impar

$$\tan x = a_1 x + a_3 x^3 + a_5 x^5 + a_7 x^7 + a_9 x^9 + a_{11} x^{11} + \cdots$$

El desarrollo anterior cumple la relación

$$\begin{aligned} &(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \frac{x^{10}}{10!} + \cdots)\tan x \\ &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \cdots \end{aligned}$$

y, por lo tanto, sus coeficientes satisfacen el sistema de ecuaciones lineales siguiente:

$$\begin{aligned} 1 &= a_1 \\ -\frac{1}{3!} &= a_3 - \frac{a_1}{2!} \\ \frac{1}{5!} &= a_5 - \frac{a_3}{2!} + \frac{a_1}{4!} \\ -\frac{1}{7!} &= a_7 - \frac{a_5}{2!} + \frac{a_3}{4!} - \frac{a_1}{6!} \\ \frac{1}{9!} &= a_9 - \frac{a_7}{2!} + \frac{a_5}{4!} - \frac{a_3}{6!} + \frac{a_1}{8!} \\ -\frac{1}{11!} &= a_{11} - \frac{a_9}{2!} + \frac{a_7}{4!} - \frac{a_5}{6!} + \frac{a_3}{8!} - \frac{a_1}{10!} \\ &\cdots \end{aligned}$$

que puede escribirse en la forma:

$$\begin{aligned} a_1 &= 1 \\ 6a_3 &= 3a_1 - 1 \end{aligned}$$

$$\begin{aligned}
120a_5 &= 60a_3 - 5a_1 + 1\\
5040a_7 &= 2520a_5 - 210a_3 + 7a_1 - 1\\
362880a_9 &= 181440a_7 - 15120a_5 + 504a_3 - 9a_1 + 1\\
39916800a_{11} &= 19958400a_9 - 1663200a_7 + 55440a_5\\
&\quad -990a_3 + 11a_1 - 1\\
&\quad \cdots
\end{aligned}$$

Así obtenemos el desarrollo

$$\tan x = x + \frac{1}{3}x^3 + \frac{2}{15}x^5 + \frac{17}{315}x^7 + \frac{62}{2835}x^9 + \frac{1382}{155925}x^{11} + \cdots$$

El desarrollo pedido es, pues,

$$\begin{aligned}
\tan 3x &= 3x + \frac{1}{3}3^3x^3 + \frac{2}{15}3^5x^5 + \frac{17}{315}3^7x^7\\
&\quad + \frac{62}{2835}3^9x^9 + \frac{1382}{155925}3^{11}x^{11} + \cdots\\
&= 3x + 9x^3 + \frac{162}{5}x^5 + \frac{4131}{35}x^7\\
&\quad + \frac{15066}{35}x^9 + \frac{3022434}{1925}x^{11} + \cdots
\end{aligned}$$

Para el cálculo del desarrollo de $f(x) = \operatorname{sen}^3 x$, usaremos una fórmula trigonométrica que se deduce de la fórmula de Moivre

$$\begin{aligned}
\operatorname{sen}^3 x &= \frac{3}{4}\operatorname{sen} x - \frac{1}{4}\operatorname{sen} 3x\\
&= \frac{3}{4}(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + \cdots)\\
&\quad -\frac{1}{4}(3x - \frac{3^3x^3}{3!} + \frac{3^5x^5}{5!} - \frac{3^7x^7}{7!} + \frac{3^9x^9}{9!} - \frac{3^{11}x^{11}}{11!} + \cdots)\\
&= \frac{3}{4}\frac{8}{3!}x^3 - \frac{3}{4}\frac{3^4-1}{5!}x^5 + \frac{3}{4}\frac{3^6-1}{7!}x^7\\
&\quad -\frac{3}{4}\frac{3^8-1}{9!}x^9 + \frac{3}{4}\frac{3^{10}-1}{9!}x^9 - \cdots\\
&= x^3 - \frac{1}{2}x^5 + \frac{13}{120}x^7 - \frac{41}{3024}x^9 + \frac{671}{604800}x^{11} - \cdots
\end{aligned}$$

b) Los tres primeros desarrollos pedidos se hallan fácilmente si se parte de los desarrollos conocidos de $\operatorname{sen} x$ y $\cos x$.

Así, directamente:

$$\begin{aligned}
\frac{\operatorname{sen} x}{x} &= 1 - \frac{x^2}{6} + \frac{x^4}{120} - \frac{x^6}{5040}\\
&\quad + \frac{x^8}{362880} - \frac{x^{10}}{39916800} + \cdots\\
\frac{\cos x - 1}{x^2} &= -\frac{1}{2} + \frac{x^2}{24} - \frac{x^4}{720} + \frac{x^6}{40320}
\end{aligned}$$

$$\begin{aligned}\frac{\operatorname{sen} x - x + \frac{x^3}{6}}{x^5} &= -\frac{x^8}{3628800} + \frac{x^{10}}{479001600} - \cdots \\ &= \frac{1}{120} - \frac{x^2}{5040} + \frac{x^4}{362880} - \frac{x^6}{39916800} \\ &\quad + \frac{x^8}{6227020800} - \frac{x^{10}}{1307674368000} - \cdots)\end{aligned}$$

La última función de este apartado puede escribirse como

$$\frac{1}{x}[\ln(1+x) - \ln(1-x)] \ ;$$

dicha función tiene el desarrollo

$$\ln\left(\frac{1+x}{1-x}\right)^{\frac{1}{x}} = 2\left(1 + \frac{x^2}{3} + \frac{x^4}{5} + \frac{x^6}{7} + \frac{x^8}{9} + \frac{x^{10}}{11} + \cdots\right) \ ,$$

que se deduce directamente del de $\ln(1+x)$.

c) Notamos primero que el desarrollo buscado para la función $f(x) = \frac{x^5}{\operatorname{sen}^3 x}$ ha de ser de la forma

$$f(x) = x^2 + a_4x^4 + a_6x^6 + a_8x^8 + a_{10}x^{10} + a_{12}x^{12} + \cdots \ ;$$

partiendo del conocimiento del desarrollo de $\operatorname{sen}^3 x$ hallado en a), estableceremos un sistema de ecuaciones lineales para sus coeficientes imponiendo

$$(x^3 - \frac{x^5}{2} + \frac{13x^7}{120} - \frac{41x^9}{3024} + \frac{671x^{11}}{604800} - \cdots)f(x) = x^5 \ .$$

Resulta:

$$\begin{aligned} a_4 - \frac{1}{2} &= 0 \\ a_6 - \frac{a_4}{2} + \frac{13}{120} &= 0 \\ a_8 - \frac{a_6}{2} + \frac{13a_4}{120} - \frac{41}{3024} &= 0 \\ a_{10} - \frac{a_8}{2} + \frac{13a_6}{120} - \frac{41a_4}{3024} + \frac{671}{604800} &= 0 \\ &\cdots \end{aligned}$$

El desarrollo hallado es entonces

$$\frac{x^5}{\operatorname{sen}^3 x} = x^2 + \frac{x^4}{2} + \frac{17x^6}{120} + \frac{457x^8}{15120} + \frac{3287x^{10}}{604800} + \cdots$$

Para $f(x) = \frac{x(\cos x - 1)}{\operatorname{sen} x - \operatorname{sen}^3 x}$ el desarrollo tiene la misma forma que en el caso anterior

$$f(x) = x^2 + a_4x^4 + a_6x^6 + a_8x^8 + a_{10}x^{10} + a_{12}x^{12} + \cdots$$

Los coeficientes ahora se hallan imponiendo

$$\frac{\operatorname{sen} x - \operatorname{sen}^3 x}{x} f(x) = \cos x - 1 \ ;$$

esto es,

$$(1-\frac{7x^2}{6}+\frac{61x^4}{120}-\frac{547x^6}{5040}+\frac{703x^8}{51840}-\frac{44287x^{10}}{39916800}+\cdots)f(x)$$
$$=\frac{x^2}{2}-\frac{x^4}{4!}+\frac{x^6}{6!}-\frac{x^8}{8!}+\frac{x^{10}}{10!}+\cdots$$

La resolución del sistema lineal correspondiente da el desarrollo buscado. Los detalles del proceso se dejan al lector.

El desarrollo de

$$f(x)=\frac{(1+x)^{\frac{1}{3}}-(1-x)^{\frac{1}{3}}}{(1+x)^{\frac{1}{3}}+(1-x)^{\frac{1}{3}}}\ ,$$

que es de la forma

$$f(x)=a_1x+a_3x^3+a_5x^5+a_7x^7+a_9x^9+a_{11}x^{11}+\cdots\ ,$$

se encontrará partiendo del desarrollo

$$\begin{aligned}(1+x)^{\frac{1}{3}} &= 1+\frac{1}{3}x-\frac{2}{3\cdot 6}x^2+\frac{2\cdot 5}{3\cdot 6\cdot 9}x^3-\frac{2\cdot 5\cdot 8}{3\cdot 6\cdot 9\cdot 12}x^4+\cdots\\ &= 1+\frac{x}{3}-\frac{x^2}{9}+\frac{5x^3}{81}-\frac{10x^4}{243}+\frac{22x^5}{729}-\frac{154x^6}{6561}+\frac{374x^7}{19683}\\ &\quad -\frac{935x^8}{59049}+\frac{21505x^9}{1594323}-\frac{223652x^{10}}{4782969}+\frac{147407x^{11}}{14398907}-\cdots\end{aligned}$$

y entonces

$$(2-\frac{2x^2}{9}-\frac{20x^4}{243}-\frac{308x^6}{6561}-\frac{1870x^8}{59049}-\frac{111826x^{10}}{4782969}-\cdots)f(x)$$
$$=\frac{2x}{3}+\frac{10x^3}{81}+\frac{44x^5}{729}+\frac{758x^7}{19683}+\frac{43010x^9}{1594323}+\frac{294814x^{11}}{14398907}+\cdots$$

La resolución del sistema lineal correspondiente también se deja al lector.

d) El desarrollo de $f(x)=e^{\operatorname{sen} x}$ podría hallarse por composición de los desarrollos del seno y de la exponencial; otra manera de operar consiste en derivar y hallar después el desarrollo

$$f(x)=a_1x+a_2x^2+a_3x^3+a_4x^4+a_5x^5+a_6x^6+\cdots$$

imponiendo la relación resultante

$$f'(x)=f(x)\cos x$$

y usando el desarrollo de $\cos x$.

El sistema lineal correspondiente es:

$$\begin{aligned}a_1 &= 1\\ 2a_2 &= a_1\\ 3a_3 &= a_2-\frac{1}{2}\\ 4a_4 &= a_3-\frac{a_1}{2}\end{aligned}$$

$$\begin{aligned} 5a_5 &= a_4 - \frac{a_2}{2} + \frac{1}{24} \\ 6a_6 &= a_5 - \frac{a_3}{2} + \frac{a_1}{24} \\ &\cdots \end{aligned}$$

y el desarrollo buscado

$$e^{\operatorname{sen} x} = 1 + x + \frac{x^2}{2} - \frac{x^4}{8} - \frac{x^5}{15} - \frac{x^6}{240} + \cdots$$

De manera similar, derivando ahora $g(x) = \sqrt{\cos x}$, hallamos

$$g'(x) = -\frac{1}{2}\frac{\operatorname{sen} x}{\sqrt{\cos x}} \ ;$$

esto es,

$$2g'(x) = -g(x)\tan x \ .$$

Esta relación permite hallar el desarrollo de la función exponente $g(x) = \sqrt{\cos x}$ de $f(x) = e^{\sqrt{\cos x}}$, partiendo del desarrollo conocido de $\tan x$; una vez conocido éste, el de $f(x)$ se encuentra imponiendo la relación entre las derivadas $f'(x) = g'(x)f(x)$.

Derivando $f(x) = \ln(\cos x)$, tenemos

$$f'(x) = -\tan x \ ;$$

el desarrollo buscado se halla simplemente integrando, término a término, el desarrollo conocido de $\tan x$ e imponiendo $f(0) = 0$:

$$f(x) = -\frac{1}{2}x^2 - \frac{1}{12}x^4 - \frac{1}{45}x^6 - \frac{17}{2520}x^8 - \frac{62}{28350}x^{10} - \frac{691}{935550}x^{12} + \cdots$$

Problema 3.10 *Consideramos la funciî*

$$f_0(n) = [(\sqrt[n]{2} - 1)n - \ln 2]n.$$

Calcularla directamente para $n = 10^k$ $(k = 1 \div 10)$, *explicar los resultados y hallar* $f_0(10^{10})$ *con 15 cifras significativas.*

Hacer lo mismo con las funciones

$$f_1(n) = \left(1 + \frac{1}{n}\right)^n - e \ , \quad f_j(n) = \sqrt[j]{n + \frac{1}{n}} - \sqrt[j]{n - \frac{1}{n}} \quad (j = 2, 3, 4) \ .$$

Usar $e = 2.718281828459045...$, $\ln 2 = 0.6931471805599453...$

k	$f_0(10^k)$
1	0.2458744
2	0.2407825
3	0.2402820
4	0.2402321
5	0.2402271
6	0.2402265
7	0.2402252
8	0.2406785
9	0.2600646
10	0.3449082

Tabla 3.8: Clculo directo de $f_0(10^k)$ $(k = 1 \div 10)$.

SOLUCIÓN:

En la tabla 3.8 se muestran los resultados obtenidos trabajando con doble precisiî (con aproximadamente 16 cifras significativas). Se observa que, a medida que aumenta k, los resultados obtenidos van distorsionndose por el efecto de la cancelaciî que se observa en la expresiî de f_0 usada; notamos, por ejemplo, que $\sqrt[n]{2} \simeq 1$ cuando $n = 10^k$ tiende a infinito. Una forma de evitar la cancelaciî consiste en desarrollar $\sqrt[n]{2}$ para valores de n grandes.

En efecto,

$$\sqrt[n]{2} = e^{\frac{\ln 2}{n}} = \sum_{j=0}^{\infty} \frac{\ln^j 2}{j! n^j} = 1 + \frac{\ln 2}{n} + \frac{\ln^2 2}{2n^2} + \frac{\ln^3 2}{6n^3} + \cdots$$

y entonces

$$f_0(n) \simeq \frac{\ln^2 2}{2} + \frac{\ln^3 2}{6n}$$

se puede usar para el clculo de $f_0(10^{10})$ con 20 cifras decimales exactas, haciendo

$$\begin{aligned} f_0(10^{10}) &= \frac{\ln^2 2}{2} + \frac{\ln^3 2}{6} 10^{-10} \pm 10^{-22} \\ &= 0.240226506964651(1 \pm \frac{1}{2} 10^{-15}) \ . \end{aligned}$$

La cota dada para el error se deduce del hecho que

$$R_3(n) \equiv \sqrt[n]{2} - \sum_{j=0}^{3} \frac{\ln^j 2}{j! n^j} = e^{\xi} \frac{\ln^4 2}{4! n^4} \ ,$$

con $0 < \xi < \frac{\ln 2}{n}$ y, por lo tanto,

$$\bar{R}_2(n) \equiv f_0(n) - \sum_{j=2}^{3} \frac{\ln^j 2}{j! n^{j-2}} = e^{\xi} \frac{\ln^4 2}{4! n^2}$$

se puede acotar por

$$|\bar{R}_2(n)| \leq e^{\frac{\ln 2}{n}} \frac{\ln^4 2}{4! n^2} < 10^{-22} \ ,$$

k	$f_1(10^k)$	$f_2(10^k)$	$f_3(10^k)$	$f_4(10^k)$
1	$-1.245394 \cdot 10^{-1}$	$3.162317 \cdot 10^{-2}$	$1.436316 \cdot 10^{-2}$	$8.891592 \cdot 10^{-3}$
2	$-1.346800 \cdot 10^{-2}$	$1.000000 \cdot 10^{-3}$	$3.094393 \cdot 10^{-4}$	$1.581139 \cdot 10^{-4}$
3	$-1.357896 \cdot 10^{-3}$	$3.162278 \cdot 10^{-5}$	$6.666667 \cdot 10^{-6}$	$2.811707 \cdot 10^{-6}$
4	$-1.359016 \cdot 10^{-4}$	$1.000000 \cdot 10^{-6}$	$1.436290 \cdot 10^{-7}$	$5.000000 \cdot 10^{-8}$
5	$-1.359127 \cdot 10^{-5}$	$3.162278 \cdot 10^{-8}$	$3.094393 \cdot 10^{-9}$	$8.891398 \cdot 10^{-10}$
6	$-1.359363 \cdot 10^{-6}$	$1.000089 \cdot 10^{-9}$	$6.670222 \cdot 10^{-11}$	$1.581279 \cdot 10^{-11}$
7	$-1.343270 \cdot 10^{-7}$	$3.370437 \cdot 10^{-11}$	$1.530873 \cdot 10^{-12}$	$2.996804 \cdot 10^{-13}$
8	$-3.011169 \cdot 10^{-8}$	$1.776534 \cdot 10^{-11}$	0	$8.882478 \cdot 10^{-14}$
9	$2.406785 \cdot 10^{-7}$	0	0	0
10	$-1.103034 \cdot 10^{-7}$	0	0	0

Tabla 3.9: Clculo directo de $f_j(10^k)$ $(j = 1 \div 4)$ $(k = 1 \div 10)$.

si $n = 10^{10}$.

Para los dems casos pedidos, se obtienen los valores de la tabla 3.9.

El uso de desarrollos permitir tambiî hacer los clculos evitando el efecto causado por las cancelaciones producidas que se observa claramente en la tabla para los valores mayores de k.

Asl

$$\begin{aligned}
f_1(n) &= (1+\frac{1}{n})^n - e = \exp\left(n\ln(1+\frac{1}{n})\right) - e \\
&= \exp(1 - \frac{1}{2n} + \frac{1}{3n^2} - \cdots) - e = e\exp(-\frac{1}{2n} + \frac{1}{3n^2} - \cdots) - e \\
&= -\frac{e}{2n} + \frac{e}{24n^2}\cdots , \\
f_1(10^{10}) &= -\frac{e}{2}10^{-10} + \frac{e}{24}10^{-20} \pm 10^{-30} \\
&= = -1.359140914218197 \cdot 10^{-10}(1 \pm \frac{1}{2}10^{-15}) . \\
f_j(n) &= \sqrt[j]{n}\left[\left(1+\frac{1}{n^2}\right)^{\frac{1}{j}} - \left(1-\frac{1}{n^2}\right)^{\frac{1}{j}}\right] \\
&= \sqrt[j]{n}\left[\frac{2}{jn^2} + \frac{(j-1)(2j-1)}{3j^3n^6} + \cdots\right] ; \\
f_2(10^{10}) &= 10^5(10^{-20} \pm \frac{1}{2}10^{-60}) \\
&= 1.000000000000000 \cdot 10^{-15} \pm \frac{1}{2}10^{-30} . \\
f_3(10^{10}) &= 10^{\frac{10}{3}}(\frac{2}{3}10^{-20} \pm 10^{-60}) \\
&= 1.43628979335459 \cdot 10^{-17}(1 \pm \frac{1}{2}10^{-15}) . \\
f_4(10^{10}) &= 10^{\frac{5}{2}}(\frac{1}{2}10^{-20} \pm 10^{-60}) \\
&= 1.58113883008419 \cdot 10^{-18}(1 \pm \frac{1}{2}10^{-15}) .
\end{aligned}$$

Problema 3.11 *Dadas las funciones* $f_t(x) = \frac{xe^{tx}}{e^x-1}$, *se definen los* polinomios de Bernoulli $B_j(t)$ $(j \geq 0)$ *por* $B_j(t) = f_t^{(j)}(0)$; *así,*

$$f_t(x) = B_0(t) + \frac{B_1(t)}{1!}x + \frac{B_2(t)}{2!}x^2 + \frac{B_3(t)}{3!}x^3 + \cdots$$

Los valores $B_j \equiv B_j(0)$ $(j \geq 0)$ *reciben el nombre de* números de Bernoulli *y cumplen:*

$$\frac{x}{e^x - 1} = B_0 + \frac{B_1}{1!}x + \frac{B_2}{2!}x^2 + \frac{B_3}{3!}x^3 + \cdots$$

a) Probar la recurrencia

$$B_0 = 1\ , \quad B_j = \sum_{l=0}^{j} \left(\begin{array}{c} j \\ l \end{array} \right) B_l \quad (j \geq 2)\ .$$

b) Probar que

$$B_j(t+h) = \sum_{l=0}^{j} \left(\begin{array}{c} j \\ l \end{array} \right) B_l(t) h^{j-l} \ (j \geq 0)\ ; \quad B_j(1) = B_j(0)\ (j \geq 2)\ .$$

c) Establecer la recurrencia

$$B_0(t) = 1\ , \quad B_1(t) = t - \frac{1}{2}\ ,$$

$$B_j'(t) = jB_{j-1}(t)$$

que, juntamente con $B_j(1) = B_j(0)$ $(j \geq 2)$, *determina completamente los* $B_j(t)$ $(j \geq 0)$; *probar también que* $B_j(t)$ *es un polinomio mónico de grado* j *que cumple*

$$(-1)^j B_j(1-t) = B_j(t) \quad (j \geq 0)\ .$$

d) Probar, para $r \geq 1$, *que*

$$B_{2r+1} = 0; \quad (-1)^r B_{2r-1}(t) > 0,\ t \in (0, \frac{1}{2})\ ;$$

$$(-1)^{r+1} B_{2r} > 0\ , \quad (-1)^r (B_{2r}(t) - B_{2r}) > 0\ , \quad t \in (0,1)\ .$$

SOLUCIÓN:

a) De la definición de los polinomios de Bernoulli, tenemos

$$e^{tx} = \frac{e^x - 1}{x}(B_0(t) + B_1(t)x + \frac{B_2(t)}{2!}x^2 + \cdots)\ ;$$

es decir,

$$(1 + tx + \frac{t^2x^2}{2!} + \cdots) = (1 + \frac{x}{2!} + \frac{x^2}{3!} + \cdots)(B_0(t) + B_1(t)x + \frac{B_2(t)}{2!}x^2 + \cdots)\ .$$

Igualando los coeficientes de x^{j-1} $(j \geq 1)$, obtenemos

$$\frac{t^{j-1}}{(j-1)!} = \sum_{l=0}^{j-1} \frac{1}{(j-l)!} \frac{B_l(t)}{l!} \ (j \geq 1) \ ;$$

esto es,

$$jt^{j-1} = \sum_{l=0}^{j-1} \left(\begin{array}{c} j \\ l \end{array} \right) B_l(t) \ (j \geq 1) \ . \quad (*)$$

Tomando $j = 1$, obtenemos $B_0(t) = 1$ y, tomando $j \geq 2$ y $t = 0$, obtenemos

$$\sum_{l=0}^{j-1} \left(\begin{array}{c} j \\ l \end{array} \right) B_l = 0 \ ,$$

donde $B_l = B_l(0)$ $(l \geq 1)$ y $B_0 = B_0(t) = 1$. Sumando B_j a los dos miembros de la igualdad, obtenemos la relación pedida. Dicha relación suele escribir se para recordarla como $(B+1)_j = B_j$, ya que, si B es un número cualquiera, la fórmula del binomio de Newton da

$$(B+1)^j = \sum_{l=0}^{j} \left(\begin{array}{c} j \\ l \end{array} \right) B^l$$

y la expresión hallada para los B_j se obtiene "bajando los índices" de ésta.

b) Usando ahora que $B_j(t+h) = f_{t+h}^{(j)}(0)$, tenemos

$$\frac{xe^{(t+h)x}}{e^x - 1} = (B_0(t+h) + B_1(t+h)x + \frac{B_2(t+h)}{2!}x^2 + \cdots \ ;$$

pero,

$$\begin{aligned} \frac{xe^{(t+h)x}}{e^x - 1} &= \frac{xe^{tx}}{e^x - 1} e^{hx} \\ &= (B_0(t) + B_1(t)x + \frac{B_2(t)}{2!}x^2 + \cdots)(1 + hx + \frac{h^2}{2!}x^2 + \cdots) \ . \end{aligned}$$

Por la unicidad de los polinomios de interpolación de Taylor de un grado dado, podemos igualar los coeficientes de x^j en ambas expresiones y se obtiene la recurrencia. De $(*)$, tenemos $B_j(t+1) = B_j(t) + jt^{j-1}$ y, si $j \geq 2$ y $t = 0$, resulta $B_j(1) = B_j(0)$.

c) De b), se deduce

$$\frac{B_j(t+h) - B_j(t)}{h} = jB_{j-1}(t) + h \sum_{l=0}^{j-2} B_l(t) h^{j-l-2} \ ;$$

por lo tanto, pasando al límite cuando h tiende a 0, se obtiene

$$B_j'(t) = jB_{j-1}(t) \ \ (j \geq 2) \ .$$

Es claro que $B_0(t) = 1$. De la relación $B_0(t) + 2B_1(t) = 2t$ (obtenida de $(*)$ para $j = 2$), se obtiene ahora $B_1(t) = t - \frac{1}{2}$. Además, $B_2(t) = t^2 - t + C_2$ satisface $B_2(1) = B_2(0)$ ya

que $\int_0^1 B_1(s)ds = 0$ y la constante C_2 se determina de manera que $\int_0^1 B_2(s)ds = 0$ y, por lo tanto, $B_3(1) = B_3(0)$. Es decir, la recurrencia:

$$B_0(t) = 1 \ , \quad B_1(t) = t - \frac{1}{2} \ ;$$

$$B_j'(t) = jB_{j-1}(t) \ , \quad B_j(1) = B_j(0) \ \ (j \geq 2) \ ,$$

también determina unívocamente los polinomios de Bernoulli.

Notamos que $B_j(1) = B_j(0)$ $(j \geq 2)$ es equivalente a $\int_0^1 B_{j-1}(t) = 0$, debido a la relación $B_j'(t) = jB_{j-1}(t)$. Por inducción, se deduce inmediatamente de esta relación que los polinomios $B_j(t)$ són mónicos.

Finalmente,

$$B_0(t) + B_1(t)x + \frac{B_2(t)}{2!}x^2 + \cdots = \frac{xe^{tx}}{e^x - 1} = \frac{(-x)e^{(1-t)(-x)}}{e^{-x} - 1}$$
$$= \quad B_0(1-t) + B_1(1-t)(-x) + \frac{B_2(1-t)}{2!}(-x)^2 + \cdots \ ;$$

de donde, igualando términos, $B_j(t) = (-1)^j B_j(1-t)$ $(j \geq 0)$. En particular, para conocer $B_j(t)$ en $[0,1]$, basta conocer $B_j(t)$ para $t \in [0, \frac{1}{2})$: si j es impar, $B_j(t) = -B_j(1-t)$ y, si j es par, $B_j(t) = B_j(1-t)$.

d) Si $j = 2r+1$, $B_{2r+1}(t) = -B_{2r+1}(1-t)$ implica la relación

$$B_{2r+1}(0) = -B_{2r+1}(1) \ ;$$

ya sabemos que se cumple $B_{2r+1}(0) = B_{2r+1}(1)$ si $r \geq 1$, entonces

$$B_{2r+1} = B_{2r+1}(0) = B_{2r+1}(1) = 0 \ (r \geq 1) \ .$$

Probaremos ahora, por inducción sobre $r \geq 1$:
$(-1)^r B_{2r-1}(t) > 0, \ t \in (0, \frac{1}{2})$;
$(-1)^r (B_{2r}(t) - B_{2r}) > 0, \ t \in (0,1)$;
$(-1)^{r+1} B_{2r} > 0$.

Si $r = 1$ y $t \in (0, \frac{1}{2})$, $(-1)^r B_{2r-1}(t) = \frac{1}{2} - t > 0$. Suponemos ahora que $(-1)^r B_{2r-1}(t) > 0$ para $t \in (0, \frac{1}{2})$, entonces

$$(-1)^r (B_{2r}(t) - B_{2r}) = (2r-1) \int_0^t (-1)^r B_{2r-1}(s)ds > 0 \ , \text{ si } t \in (0, \frac{1}{2}]$$

y, usando c), $(-1)^r (B_{2r}(t) - B_{2r}) = (-1)^r (B_{2r}(1-t) - B_{2r}) > 0$, si $t \in (\frac{1}{2}, 1)$.

Debido a que $\int_0^1 B_{2r}(s)ds = 0$,

$$0 < (-1)^r \int_0^1 (B_{2r}(t) - B_{2r})dt = (-1)^{r+1} B_{2r} \ .$$

Basta probar ahora $(-1)^{r+1} B_{2r+1}(t) > 0$, para $t \in (0, \frac{1}{2})$.

Si $B_{2r+1}(t_1) = 0$ para $t_1 \in (0, \frac{1}{2})$, querría decir que $B_{2r+1}(t)$ se anularía en $0, t_1, \frac{1}{2}$; por el teorema de Rolle, $B_{2r+1}'(t)$ se anularía en t_2, t_3, cumpliéndose la ordenación $0 < t_2 < t_1 < t_3 < \frac{1}{2}$ y, de nuevo, $B_{2r+1}^{(2)}(t)$ se anularía en una abscisa $t_3 \in (0, \frac{1}{2})$; pero, si $r \geq 1$,

$$B_{2r+1}^{(2)}(t) = (2r+1)B_{2r}'(t) = (2r+1)(2r)B_{2r-1}(t)$$

y, por lo tanto, $B_{2r-1}(t_3) = 0$, $t_3 \in (0, \frac{1}{2})$, que entra en contradicción con la suposición hecha. Así pues, el polinomio $B_{2r+1}(t)$ no se anula en $(0, \frac{1}{2})$ y, como que es continuo, no cambia de signo en el intervalo; su signo será así igual al signo de $B'_{2r+1}(0) = (2r+1)B_{2r}(0)$, que es el de $(-1)^{r+1}$. En definitiva, tenemos $(-1)^{r+1}B_{2r+1}(t) > 0$ para $t \in (0, \frac{1}{2})$.

Como ilustración y para facilitar su uso en el capítulo siguiente, damos a continuación una tabla de los primeros números de Bernoulli B_0, B_1, B_2, B_4, B_6, ..., B_{30} (recuérdese que, por d), $B_{2r+1} = 0$ $(r \geq 1)$):

B_0	$=$	1	B_{16}	$=$	$-\frac{3617}{510}$
B_1	$=$	$-\frac{1}{2}$	B_{18}	$=$	$\frac{43867}{798}$
B_2	$=$	$\frac{1}{6}$	B_{20}	$=$	$-\frac{174611}{330}$
B_4	$=$	$-\frac{1}{30}$	B_{22}	$=$	$\frac{854513}{138}$
B_6	$=$	$\frac{1}{42}$	B_{24}	$=$	$-\frac{236364091}{2730}$
B_8	$=$	$-\frac{1}{30}$	B_{26}	$=$	$\frac{8553103}{6}$
B_{10}	$=$	$\frac{5}{66}$	B_{28}	$=$	$-\frac{23749461029}{870}$
B_{12}	$=$	$-\frac{691}{2730}$	B_{30}	$=$	$\frac{8615841276005}{14322}$
B_{14}	$=$	$\frac{7}{6}$		$\ldots$	

Problema 3.12 *Calcular* $\tan 22.5°$ *usando interpolación de Hermite en* $0°$ *y* $45°$. *Acotar el error cometido y comparar la cota encontrada con el error exacto.*

SOLUCIÓN:

Si definimos $f(x) = \tan \pi x$ y trabajamos con radianes, se pide el cálculo de $f(\frac{1}{8}) = \tan \frac{\pi}{8}$ por interpolación de Hermite en $x_0 = 0$ y $x_1 = \frac{1}{4}$.

Usaremos el método de las diferencias divididas generalizadas repitiendo las abscisas x_0, x_1 y aplicando la definición de aquéllas haciendo $f[x_0, x_0] \equiv f'(x_0)$ y $f[x_1, x_1] \equiv f'(x_1)$.

Recogemos primero la información requerida:

$$f(x_0) = 0 \ , \quad f(x_1) = 1 \ , \quad f'(x_0) = \pi \ , \quad f'(x_1) = 2\pi \ ,$$

debido a que $f'(x) = (1 + f^2(x))\pi$.

El esquema resultante es el dado en la tabla

$$\begin{array}{c|cccc} 0 & 0 & & & \\ & & \pi & & \\ 0 & 0 & & 4(4-\pi) & \\ & & 4 & & 16(3\pi-8) \\ \frac{1}{4} & 1 & & 8(\pi-2) & \\ & & 2\pi & & \\ \frac{1}{4} & 1 & & & \end{array} \quad .$$

El polinomio interpolador de Hermite es, pues,

$$p_3(x) = \pi x + 4(4-\pi)x^2 + 16(3\pi-8)x^2(x-\frac{1}{4})$$

y permite hallar la aproximación

$$\tan(\frac{\pi}{8}) \simeq p_3(\frac{1}{8}) = 0.4018...$$

La fórmula trigonométrica $\tan\frac{\theta}{2} = \frac{\operatorname{sen}\theta}{1+\cos\theta}$ permite hallar el valor exacto

$$\tan(\frac{\pi}{8}) = \frac{\frac{1}{\sqrt{2}}}{1+\frac{1}{\sqrt{2}}} = \frac{1}{1+\sqrt{2}} = \sqrt{2}-1 = 0.4142213562...$$

El error cometido puede acotarse usando la expresión del error en la interpolación de Hermite

$$f(x) - p_3(x) = \frac{f^{(4)}(\xi)}{4!}x^2(x-\frac{1}{4})^2, \quad 0 < \xi < \frac{1}{4}\ ,$$

que, en $x = \frac{1}{8}$, vale

$$f(\frac{1}{8}) - p_3(\frac{1}{8}) = \frac{f^{(4)}(\xi)}{4!8^4}.$$

Daremos como cota del error $\frac{M_4}{4!8^4}$, donde M_4 es una cota de $f^{(4)}$ en el intervalo $[0, \frac{1}{4}]$ que buscamos a continuación. Tenemos

$$f' = \pi(1+f^2) > 0, \quad f^{(2)} = 2\pi f f' > 0, \quad f^{(3)} = 2\pi(f^{(2)}f + f'^2) > 0,$$

$$f^{(4)} = 2\pi(f^{(3)}f + 3f^{(2)}f') > 0, \quad f^{(5)} > 0\ ;$$

así $f^{(4)}$, al ser creciente, toma el valor máximo sobre $[0, \frac{1}{4}]$ en $x = \frac{1}{4}$. Para calcularlo, hallamos $f(\frac{1}{4}) = 1$, $f'(\frac{1}{4}) = 2\pi$, $f^{(2)}(\frac{1}{4}) = 4\pi^2$, $f^{(3)}(\frac{1}{4}) = 16\pi^3$; entonces,

$$M_4 \equiv f^{(4)}(\frac{1}{4}) = 80\pi^4$$

y la cota del error da 0.079, que es sensiblemente mayor que el error real 0.0124.

Problema 3.13 (Interpolación hermitiana generalizada).

a) Demostrar la existencia y unicidad del polinomio $p_N(x)$ de grado menor o igual que N cumpliendo:

$$p_N^{(l)}(x_k) = f_k^{(l)} \quad (l = 0 \div n_k)\ (k = 0 \div m)\ ,$$

con $N = m + \sum_{k=0}^{m} n_k$.

b) Hallar una expresión para los errores $f(x) - p_N(x)$ en $[a, b]$, si la función $f \in \mathcal{C}^{N+1}(a, b)$.

c) Explicar, con detalle, cómo se generaliza el método de las diferencias divididas a este caso.

d) Hallar el polinomio $p_{11}(x)$ de grado 11 cumpliendo:

$$\begin{aligned} p_{11}^{(l)}(-1) &= f^{(l)}(-1) \quad (l = 0 \div 3)\ , \\ p_{11}^{(l)}(0) &= f^{(l)}(0) \quad (l = 0 \div 2)\ , \\ p_{11}^{(l)}(1) &= f^{(l)}(1) \quad (l = 0 \div 4) \end{aligned}$$

para $f(x) = x^{12}$. Hallar el error de interpolación en el intervalo $[-1, 1]$.

SOLUCIÓN:

a) Las condiciones de interpolación de Hermite generalizada imponen un sistema de ecuaciones a los coeficientes del polinomio $p_N(x)$. La existencia y unicidad de la solución equivale a la existencia y unicidad de la solución del sistema homogéneo; esto es, que el único polinomio que cumpla

$$p_N^{(l)}(x_k) = 0 \quad (l = 0 \div n_k)\ (k = 0 \div m)\ ,$$

sea el polinomio 0.

Veamos que esto es cierto por reducción al absurdo: La condición $p_N^{(l)}(x_k) = 0$ $(l = 0 \div n_k)$ equivale a que $p_N(x)$ tenga el cero x_k de multiplicidad $n_k + 1$ y a que $(x - x_k)^{n_k+1}$ sea un factor de $p_N(x)$ $(k = 0 \div m)$. Si el polinomio $p_N(x)$ no fuese nulo, el polinomio $(x - x_0)^{n_0+1}(x - x_1)^{n_1+1} \cdots (x - x_m)^{n_m+1}$ (de grado $N + 1$) sería un factor de $p_N(x)$ (de grado máximo N); de donde se deduce que $p_N(x)$ ha de ser nulo.

b) Para cada $x \in (a, b)$, $x \neq x_k$ $(k = 0 \div m)$, consideramos la función

$$\Phi(z) = f(z) - p_N(z) - a(x)(z - x_0)^{n_0+1}(z - x_1)^{n_1+1} \cdots (z - x_m)^{n_m+1}\ ,$$

donde $a(x)$ se elige de forma que $\Phi(x) = 0$. Así la función Φ tiene los siguientes ceros: x, x_k (de multiplicidad $n_k + 1$) $(k = 0 \div m)$; esto es, $N + 1$ ceros (contados con su multiplicidad). Una generalización simple del teorema de Rolle nos permite afirmar que la derivada $N + 1$ se anula en una abscissa ξ del intervalo (a, b)

$$\Phi^{(N+1)}(\xi) = f^{(N+1)}(\xi) - a(x)(N + 1)! = 0\ ;$$

de donde, despejando $a(x)$ y substituyéndolo en la expresión inicial, tenemos la fórmula del error en la interpolación de Hermite generalizada

$$f(x) - p_N(x) = \frac{f^{(N+1)}(\xi)}{(N + 1)!}(x - x_0)^{n_0+1}(x - x_1)^{n_1+1} \cdots (x - x_m)^{n_m+1}\ .$$

c) El proceso a seguir es casi el mismo que en el caso de interpolación usual (en el que $n_k = 0$ $(k = 0 \div m)$). Recordemos primero que la interpolación de Taylor puede entenderse como una interpolación normal en la que todas las abscisas de interpolación se han hecho coincidir en una misma abscisa x_0 y que pueden generalizarse así las diferencias divididas al caso de que haya abscisas repetidas haciendo

$$f[x_0, \underbrace{x_0, x_0, \ldots, x_0}_{j}] \equiv \frac{f^j(x_0)}{j!} \ .$$

La interpolación de Hermite generalizada es una extensión de la de Taylor a varias abscisas; así, poniendo $n_k + 1$ veces cada abscisa x_k $(k = 0 \div m)$, puede construirse una tabla de diferencias divididas (generalizadas) basándonos en la expresión anterior para todas las abscisas

$$f[x_k, \underbrace{x_k, x_k, \ldots, x_k}_{j}] \equiv \frac{f^j(x_k)}{j!} \quad (k = 0 \div m) \ .$$

La obtención del polinomio interpolador se hace entonces de manera análoga:

$$\begin{aligned} p_N(x) = {} & f[x_0] + \cdots + f[x_0, \ldots, x_0](x - x_0)^{n_0} + \ldots \\ & + f[x_0, \ldots, x_0, x_1, \ldots, x_1](x - x_0)^{n_0+1}(x - x_1)^{n_1} + \ldots \\ & + \ldots \\ & + f[x_0, \ldots, x_0, \ldots\ldots, x_m, \ldots, x_m](x - x_0)^{n_0+1} \cdots (x - x_m)^{n_m} \ . \end{aligned}$$

d) Construimos la tabla de diferencias divididas generalizadas partiendo de la información siguiente:

$$\begin{array}{llll} f(-1) = 1 & f'(-1) = -12 & f^{(2)}(-1) = 132 & f^{(3)}(-1) = -1320 \\ f(0) = 0 & f'(0) = 0 & f^{(2)}(0) = 0 & \\ f(1) = 1 & f'(1) = 12 & f^{(2)}(1) = 132 & f^{(3)}(1) = 1320 \\ f^{(4)}(1) = 11880 \ . & & & \end{array}$$

El resultado es el siguiente:

−1	1											
		−12										
−1	1		66									
		−12		−220								
−1	1		66		165							
		−12		−55		−120						
−1	1		11		45		84					
		−1		−10		−36		−34				
0	0		1		9		16		14			
		0		−1		−4		−6		−4		
0	0		0		1		4		6		4	
		0		1		4		6		4		1
0	0		1		9		16		14		6	
		1		10		36		34		16		
1	1		11		45		84		46			
		12		55		120		126				
1	1		66		165		210					
		12		220		330						
1	1		66		495							
		12		220								
1	1		66									
		12										
1	1											

El polinomio interpolador pedido es

$$\begin{aligned} p_{11}(x) &= 1 - 12(x+1) + 66(x+1)^2 - 220(x+1)^3 \\ &\quad +165(x+1)^4 - 120(x+1)^4 x + 84(x+1)^4 x^2 \\ &\quad -34(x+1)^4 x^3 + 14(x+1)^4 x^3 (x-1) \\ &\quad -4(x+1)^4 x^3 (x-1)^2 + 4(x+1)^4 x^3 (x-1)^3 \\ &\quad +(x+1)^4 x^3 (x-1)^4 \ . \end{aligned}$$

La fórmula del error hallada en b) nos dice que

$$x^{12} - p_{11}(x) = (x-1)^4 x^3 (x+1)^5 \ ,$$

exactamente. Así, el polinomio interpolador también se escribe

$$p_{11}(x) = x^{12} - (x-1)^4 x^3 (x+1)^5 \ .$$

El polinomio de error $(x^2-1)(x+1)[(x-1)x(x+1)]^3$ puede acotarse por el producto de las cotas de cada factor; esto es, usando el problema 3.3 sobre $[-1,1]$, por $1 \cdot 2 \cdot \frac{2\sqrt{3}}{9} = \frac{4\sqrt{3}}{9}$. La mejor cota se hallaría buscando el valor absoluto máximo del error en los máximos y mínimos relativos de éste.

Problema 3.14 *Hallar la constante que aproxima mejor $f(x) = e^x$ en $[0,1]$, según las normas: i) $\| \ \|_1$, ii) $\| \ \|_2$, iii) $\| \ \|_\infty$.*

SOLUCIÓN:

La aproximación por una constante es el caso más sencillo de aproximación; puede tratarse a partir de la misma definición del problema como la minimización de la función

$$\Phi(c) = \min_{c\in\mathbb{R}} \|f - c\| \ .$$

Esto es lo que haremos para las tres normas sugeridas.

i) La función $\Phi_1(c) = \|f - c\|_1$ está definida por

$$\Phi_1(c) = \int_0^1 |e^x - c|\, dx \ .$$

Tenemos, así,

$$\begin{aligned}
\Phi_1(c) &= \int_0^1 (e^x - c)dx = e - 1 - c \ , \text{ si } c \le 1 \ ; \\
\Phi_1(c) &= \int_0^{\ln c} (c - e^x)dx + \int_{\ln c}^1 (e^x - c)dx = 2c\ln c - 3c + e + 1 \ , \\
&\quad \text{si } 1 \le c \le e \ ; \\
\Phi_1(c) &= \int_0^1 (c - e^x)dx = c - e + 1 \ , \text{ si } c \ge e \ .
\end{aligned}$$

Dicha función es decreciente en el tramo $c \le 1$ y creciente en $c \ge 1$; tiene, por lo tanto, una abscisa de mínimo $\tilde{c}$ en $[1, e]$ que puede hallarse por anulación de la derivada:

$$\Phi_1'(\tilde{c}) = 2\ln\tilde{c} - 1 = 0 \ , \quad \tilde{c} = \sqrt{e} \ .$$

Dicha abscisa de mínimo $\tilde{c}$ es la constante óptima en esta norma.

ii) Análogamente, tenemos

$$\Phi_2(c) = \int_0^1 (e^x - c)^2 dx = \frac{1}{2}(e^2 - 1) - 2(e - 1)c + c^2 \ ;$$

se trata de una parábola en la variable c. El valor de $c = c^*$ correspondiente al mínimo de la misma nos da la constante óptima buscada

$$c^* = e - 1 \ ,$$

que, al ser $\| \ \|_2$ una norma euclídea, también se obtiene de la resolución de la ecuación normal asociada

$$c^* = \frac{\int_0^1 1 \cdot e^x dx}{\int_0^1 1^2 dx} = e - 1 \ .$$

iii) En este caso, la función que es necesario minimizar

$$\Phi_\infty(c) = \|e^x - c\|_\infty = \max\{c - 1, e - c\}$$

resulta ser

$$\begin{aligned}
\Phi_\infty(c) &= e - c \ , \text{ si } c \le \frac{e+1}{2} \ ; \\
\Phi_\infty(c) &= c - 1 \ , \text{ si } c \ge \frac{e+1}{2} \ .
\end{aligned}$$

La gráfica de la función está formada por dos rectas: la primera con pendiente -1, la segunda con pendiente 1 que coinciden para

$$c = \hat{c} = \frac{e+1}{2} \ ;$$

éste es, pues, el valor para el que se da el mínimo buscado de la función Φ_∞.

Problema 3.15 *Hallar la mejor aproximación por mínimos cuadrados, en los puntos*

$$(-\frac{\pi}{2}, 1),\ (0,0),\ (\frac{\pi}{2}, \frac{1}{2}),\ (\pi, 1)\ ,$$

que sea del tipo

$$t_1(\theta) = a + r\,\mathrm{sen}(\theta + \alpha)\ ,$$

con a, r reales y $\alpha \in [0, 2\pi]$.

SOLUCIÓN:

Primero, usamos la expresión del seno de la suma

$$t_1(\theta) = a + r\,\mathrm{sen}\,\alpha\cos\theta + r\cos\alpha\,\mathrm{sen}\,\theta$$

y denotamos $a_0 \equiv a$, $a_1 \equiv r\,\mathrm{sen}\,\alpha$, $b_1 \equiv r\cos\alpha$.

Se trata de una aproximación discreta por mínimos cuadrados en las abscisas angulares $\theta_k = -\frac{\pi}{2} + kh$ $(k = 0 \div 3)$ con $h = \frac{\pi}{2}$: es necesario encontrar los coeficientes a_0^*, a_1^* y b_1^* para que

$$\|f - t_1\|_2^2 = \sum_{k=0}^{3} [f(\theta_k) - t_1(\theta_k)]^2$$

sea mínima para $a_0 = a_0^*$, $a_1 = a_1^*$ y $b_1 = b_1^*$.

La solución del problema es la solución de las ecuaciones normales

$$\begin{pmatrix} (1,1) & (1,\cos\,) & (1,\mathrm{sen}\,) \\ (\cos\,,1) & (\cos\,,\cos\,) & (\cos\,,\mathrm{sen}\,) \\ (\mathrm{sen}\,,1) & (\mathrm{sen}\,,\cos\,) & (\mathrm{sen}\,,\mathrm{sen}\,) \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \\ b_1^* \end{pmatrix} = \begin{pmatrix} (1,f) \\ (\cos\,,f) \\ (\mathrm{sen}\,,f) \end{pmatrix}$$

halladas usando el producto escalar asociado a la norma euclídea definida antes.

Las componentes diagonales de la matriz son:

$$\begin{aligned} (1,1) &= \sum_{k=0}^{3} 1 \cdot 1 = 4\ , \\ (\cos\,,\cos\,) &= \sum_{k=0}^{3} \cos^2\theta_k = 0 + 1 + 0 + 1 = 2\ , \\ (\mathrm{sen}\,,\mathrm{sen}\,) &= \sum_{k=0}^{3} \mathrm{sen}^2\,\theta_k = 1 + 0 + 1 + 0 = 2\ ; \end{aligned}$$

las componentes no diagonales resultan ser todas nulas. Se trata de una aproximación trigonométrica equidistante, con $\theta_k = -\frac{\pi}{2} + \frac{k\pi}{2}$ $(k = 0 \div 3)$.

Por lo que se refiere a las componentes del término independiente, tomando

$$f_0 \equiv f(\theta_0) = 1 \ , \quad f_1 \equiv f(\theta_1) = 0 \ , \quad f_2 \equiv f(\theta_2) = \frac{1}{2} \ , \quad f_3 \equiv f(\theta_3) = 1 \ ,$$

tenemos

$$(1, f) = \sum_{k=0}^{3} f_k = 1 + 0 + \frac{1}{2} + 1 = 2 \ ,$$

$$(\cos \ , f) = \sum_{k=0}^{3} \cos\theta_k f_k = 0 + 0 + 0 - 1 = -1 \ ,$$

$$(\text{sen} \ , f) = \sum_{k=0}^{3} \text{sen}\,\theta_k f_k = -1 + 0 + \frac{1}{2} + 0 = -\frac{1}{2} \ .$$

La resolución de las ecuaciones normales es inmediata y da:

$$a_0^* = \frac{1}{2} \ , \quad a_1^* = -\frac{1}{2} \ , \quad b_1^* = -\frac{1}{4} \ .$$

Los valores óptimos, en el sentido de la aproximación por mínimos cuadrados, de los coeficientes a, r y α serán, pues:

$$a^* = a_0^* = \frac{1}{2} \ , \quad r^* = \sqrt{a_1^{*2} + b_1^{*2}} = \frac{\sqrt{5}}{4} = 0.559017... \ ,$$

$$\alpha^* = \arctan\frac{a_1}{b_1} + \pi = \arctan 2 + \pi = 4.24874...$$

Problema 3.16 *Hallar la familia de polinomios ortogonales mónicos* $\psi_j(x)$ $(j = 0 \div 4)$ *correspondiente a la función peso* $w(x) = x^2 + 1$ *en el intervalo* $[-1, 1]$.

SOLUCIÓN:

Los polinomios ortogonales respecto al producto escalar asociado

$$(f, g) = \int_{-1}^{1} (x^2 + 1) f(x) g(x) dx \ ,$$

se pueden hallar usando la fórmula recurrente de los polinomios ortogonales con $A_j = 1$ $(j \geq 0)$:

$$\psi_{-1}(x) = 0 \ , \quad \psi_0(x) = 1 \ ,$$

$$\psi_{j+1}(x) = \alpha_j(x - \beta_j)\psi_j(x) - \gamma_j\psi_{j-1}(x) \ (j \geq 0) \ ,$$

donde

$$\begin{aligned} \alpha_j &= \frac{A_{j+1}}{A_j} = 1 \ , \\ \beta_j &= \frac{(\psi_j, x\psi_j)}{(\psi_j, \psi_j)} = 0 \ , \\ \gamma_j &= \frac{\alpha_j}{\alpha_{j-1}} \frac{(\psi_j, \psi_j)}{(\psi_{j-1}, \psi_{j-1})} = \frac{(\psi_j, \psi_j)}{(\psi_{j-1}, \psi_{j-1})} \ . \end{aligned}$$

La anulación de los coeficientes β_j $(j \geq 0)$ se deduce del hecho de que los polinomios ortogonales tienen paridad $\psi_j(-x) = (-1)^j\psi_j(x)$ $(j \geq 0)$, debido a que la función peso $w(x) = x^2 + 1$ es par en el intervalo simétrico $[-1, 1]$.

La fórmula recurrente queda, así, reducida a la forma

$$\psi_{j+1}(x) = x\psi_j(x) - \gamma_j\psi_{j-1}(x) \ (j \geq 0) \ .$$

Efectuando los cálculos correspondientes y definiendo

$$I_{2r} \equiv \int_{-1}^{1} x^{2r} dx = \frac{2}{2r+1} \ (r \geq 0) \ ,$$

tenemos

$$\begin{aligned} \psi_1(x) &= x\psi_0(x) = x \ ; \\ \gamma_1 &= \frac{(x\psi_1, \psi_0)}{(\psi_0, \psi_0)} = \frac{I_4 + I_2}{I_2 + I_0} = \frac{2}{5} \ , \\ \psi_2(x) &= x\psi_1(x) - \gamma_1\psi_0(x) = x^2 - \frac{2}{5} \ ; \\ \gamma_2 &= \frac{(x\psi_2, \psi_1)}{(\psi_1, \psi_1)} = \frac{I_6 + \frac{3}{5}I_4 - \frac{2}{5}I_2}{I_4 + I_2} = \frac{17}{70} \ , \\ \psi_3(x) &= x\psi_2(x) - \gamma_2\psi_1(x) = x^3 - \frac{9}{14}x \ ; \\ \gamma_3 &= \frac{(x\psi_3, \psi_2)}{(\psi_2, \psi_2)} = \frac{I_8 - \frac{3}{70}I_6 - \frac{55}{70}I_4 + \frac{9}{35}I_2}{I_6 + \frac{1}{5}I_4 - \frac{16}{25}I_2 + \frac{4}{25}} = \frac{185}{714} \ , \\ \psi_4(x) &= x\psi_3(x) - \gamma_3\psi_2(x) = x^4 - \frac{46}{51}x^2 + \frac{37}{357} \ . \end{aligned}$$

Los polinomios ortogonales mónicos buscados son, pues,

$$1 \ , \quad x \ , \quad x^2 - \frac{2}{5} \ , \quad x^3 - \frac{9}{14}x \ , \quad x^4 - \frac{46}{51}x^2 + \frac{37}{357} \ .$$

Problema 3.17 *a) Demostrar que la familia de polinomios*

$$Q_j(y) = T_j(2y-1) \quad (j \geq 0)$$

es ortogonal respecto al peso $w(y) = (y(1-y))^{-\frac{1}{2}}$ *en* $[0,1]$*, donde los polinomios* $T_j(x)$ $(j \geq 0)$ *son los polinomios de Chebichev.*

b) Aproximar y^{10} *en el intervalo* $[0,1]$ *por un polinomio* $p_4^*(y)$ *de grado menor o igual que 4, de manera que el error de aproximación por mínimos cuadrados*

$$\int_0^1 \frac{(y^{10} - p_4(y))^2}{(y(1-y))^{\frac{1}{2}}} dy$$

sea mínimo para $p_4 = p_4^*$.

SOLUCIÓN:

a) Sabemos que los polinomios de Chebichev $T_j(x)$ $(j \geq 0)$, dados por

$$T_j(x) = \cos(j \arccos x) \quad \forall x \in [-1,1]$$

o también por la recurrencia

$$T_0(x) = 1 \ , \quad T_1(x) = x \ , \quad T_{j+1}(x) = 2xT_j(x) - T_{j-1}(x) \quad (j \geq 1) \ ,$$

son ortogonales respecto al peso $w(x) = \frac{1}{\sqrt{1-x^2}}$ en $[-1,1]$.

En efecto, de

$$\begin{aligned} (T_j, T_l) &= \int_{-1}^1 \frac{\cos(j \arccos x)\cos(l \arccos x)}{\sqrt{1-x^2}} dx \\ &= \int_0^\pi \cos(j\theta)\cos(l\theta) d\theta \ , \end{aligned}$$

tenemos

$$(T_j, T_l) = \begin{cases} 0 & \text{si } j \neq l \\ \pi & \text{si } j = l = 0 \\ \frac{\pi}{2} & \text{si } j = l \neq 0 \end{cases} .$$

Para el caso que nos ocupa y haciendo el cambio $x = 2y - 1$,

$$\begin{aligned} ((Q_j, Q_l)) &= \int_0^1 \frac{Q_j(y)Q_l(y)}{\sqrt{y(1-y)}} dy = \int_0^1 \frac{T_j(2y-1)T_l(2y-1)}{\sqrt{y(1-y)}} dy \\ &= \int_{-1}^1 \frac{T_j(x)T_l(x)}{\sqrt{1-x^2}} = (T_j, T_l) \ , \end{aligned}$$

de donde se desprende la relación de ortogonalidad enunciada.

b) La aproximación polinomial por mínimos cuadrados pedida se escribe como una combinación lineal de los polinomios ortogonales $Q_j(y)$ hasta grado 4

$$p_4^*(y) = \sum_{j=0}^4 c_j^* Q_j(y) \ .$$

Los coeficientes ortogonales correspondientes son

$$c_j^* = \frac{((y^{10}, Q_j))}{((Q_j, Q_j))} \ (j = 0 \div 4) \ .$$

A continuación, los calculamos.
Usando los cambios $y = \frac{1+x}{2}$ y $x = \cos\theta$, tenemos primero

$$\begin{aligned} ((y^{10}, Q_j)) &= \int_0^1 \frac{y^{10} Q_j(y)}{\sqrt{y(1-y)}} dy = \frac{1}{2^{10}} \int_{-1}^1 \frac{(1+x)^{10} T_j(x)}{\sqrt{1-x^2}} dx \\ &= \frac{1}{2^{10}} \int_0^\pi (1+\cos\theta)^{10} \cos j\theta \ d\theta \\ &= \int_0^\pi \cos^{20}\frac{\theta}{2} \cos j\theta \ d\theta \ . \end{aligned}$$

La integral resultante puede evaluarse fácilmente, después de desarrollar el término $\cos^{20}\frac{\theta}{2}$, así:

$$\begin{aligned} \cos^{20}\frac{\theta}{2} &= \left(\frac{\exp(i\frac{\theta}{2}) + \exp(-i\frac{\theta}{2})}{2}\right)^{20} \\ &= \frac{1}{2^{20}} \sum_{l=0}^{20} \left(\begin{array}{c} 20 \\ l \end{array} \right) e^{i(l-10)\theta} = \frac{1}{2^{20}} \sum_{l=-10}^{10} \left(\begin{array}{c} 20 \\ 10+l \end{array} \right) e^{il\theta} \\ &= \frac{1}{2^{20}} \left(\begin{array}{c} 20 \\ 10 \end{array} \right) + \frac{1}{2^{19}} \sum_{l=1}^{10} \left(\begin{array}{c} 20 \\ 10+l \end{array} \right) \frac{e^{il\theta} + e^{-il\theta}}{2} \\ &= \frac{1}{2^{20}} \left(\begin{array}{c} 20 \\ 10 \end{array} \right) + \frac{1}{2^{19}} \sum_{l=1}^{10} \left(\begin{array}{c} 20 \\ 10+l \end{array} \right) \cos l\theta \ . \end{aligned}$$

Tenemos, pues,

$$((y^{10}, Q_j)) = \frac{\pi}{2^{20}} \left(\begin{array}{c} 20 \\ 10-j \end{array} \right) \ (j = 0 \div 10) \ ;$$

$$\begin{aligned} c_0^* &= \frac{1}{2^{20}} \left(\begin{array}{c} 20 \\ 10 \end{array} \right) \ , \\ c_j^* &= \frac{1}{2^{19}} \left(\begin{array}{c} 20 \\ 10-j \end{array} \right) \ (j = 1 \div 4) \ . \end{aligned}$$

De

$$\frac{1}{2^{20}} \left(\begin{array}{c} 20 \\ 10 \end{array} \right) = \frac{46189}{8192} \ ,$$

resulta:

$$c_0^* = \frac{1}{2^{20}} \left(\begin{array}{c} 20 \\ 10 \end{array} \right) = \frac{46189}{8192} \ ,$$

$$
\begin{aligned}
c_1^* &= \frac{1}{2^{19}}\left(\begin{array}{c} 20 \\ 9 \end{array}\right) = \frac{10}{11}2c_0^* = \frac{20995}{10240}\ , \\
c_2^* &= \frac{1}{2^{19}}\left(\begin{array}{c} 20 \\ 8 \end{array}\right) = \frac{9}{12}c_1^* = \frac{62985}{40960}\ , \\
c_3^* &= \frac{1}{2^{19}}\left(\begin{array}{c} 20 \\ 7 \end{array}\right) = \frac{8}{13}c_2^* = \frac{4845}{5120}\ , \\
c_4^* &= \frac{1}{2^{19}}\left(\begin{array}{c} 20 \\ 6 \end{array}\right) = \frac{7}{14}c_2^* = \frac{4845}{10240}\ .
\end{aligned}
$$

Problema 3.18 *Ajustar por mínimos cuadrados la tabla de datos:*

x	0.25	0.50	0.75	1.00	1.25	1.50	1.75
y	0.40	0.50	0.90	1.28	1.60	1.66	2.02

a funciones de los tipos siguientes:

a) $y = a_0 + a_1 x$,
b) $y = a_0 + a_1 x + a_2 x^2$,
c) $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$,
d) $y = a x^\alpha$.

SOLUCIÓN:

a) La tabla considerada (x_k, y_k) $(k = 0 \div 6)$ conduce al sistema lineal sobredeterminado

$$a_0 + x_k a_1 = y_k \ (k = 0 \div 6)\ ;$$

esto es,

$$
\left(\begin{array}{cc} 1 & 0.25 \\ 1 & 0.50 \\ 1 & 0.75 \\ 1 & 1.00 \\ 1 & 1.25 \\ 1 & 1.50 \\ 1 & 1.75 \end{array}\right)\left(\begin{array}{c} a_0 \\ a_1 \end{array}\right) = \left(\begin{array}{c} 0.40 \\ 0.50 \\ 0.90 \\ 1.28 \\ 1.60 \\ 1.66 \\ 2.02 \end{array}\right)\ ,
$$

que, en notación matricial, escribimos como $Ma = y$.

El método de ortogonalización de Gram-Schmidt permite factorizar la matriz $M = (m^{(1)}\ m^{(2)})$ del sistema en la forma QR, donde Q es una matriz 7x2 con columnas ortogonales ($D = Q^\top Q$ es diagonal), y R es una matriz 2x2 triangular superior con unos en la

diagonal:

$$\begin{aligned} q^{(1)} &= m^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} , \\ d_1 &= \|q^{(1)}\|_2^2 = 7 \ ; \\ r_{12} &= \frac{q^{(1)\top} m^{(2)}}{d_1} = \frac{7}{7} = 1 \ , \\ q^{(2)} &= m^{(2)} - r_{12} q^{(1)} = \begin{pmatrix} -0.75 \\ -0.50 \\ -0.25 \\ 0.00 \\ 0.25 \\ 0.50 \\ 0.75 \end{pmatrix} , \\ d_2 &= \|q^{(2)}\|_2^2 = 1.75 \ . \end{aligned}$$

Así pues, resulta

$$M = QR = \begin{pmatrix} 1 & -0.75 \\ 1 & -0.50 \\ 1 & -0.25 \\ 1 & 0.00 \\ 1 & 0.25 \\ 1 & 0.50 \\ 1 & 0.75 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} , \quad D = \begin{pmatrix} 7 & 0 \\ 0 & 1.75 \end{pmatrix} .$$

El sistema de ecuaciones normales asociado al sistema $QRa = y$ es entonces $M^\top M a^* = M^\top y$, equivalente a $Ra^* = D^{-1} Q^\top y$:

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix} = \begin{pmatrix} 1.194285714 \\ 1.125714286 \end{pmatrix} ,$$

de solución

$$a_0^* = 0.0685714282 \ , \quad a_1^* = 1.125714286 \ .$$

Como ilustración, usaremos a continuación el método de ortogonalización de Householder para hallar otro tipo de factorización QR de la matriz M.

Si u es un vector no nulo, la matriz de Householder $P \equiv P(u) = I - \alpha u u^\top$ con $\alpha = \frac{2}{u^\top u}$ es una matriz ortogonal simétrica: $P = P^\top = P^{-1}$ y, para cualquier vector a de norma euclídea s, se cumple

$$P(a + se^{(1)})a = -se^{(1)} \ , \quad P(a - se^{(1)})a = se^{(1)} \ ,$$

donde $e^{(1)} = (1 \;\; 0 \;\; \cdots \;\; 0)^\top$.

En nuestro caso, si escribimos $M = (m^{(1)} \; m^{(2)})$ y tomamos $a = m^{(1)}$, se cumple $s = \|m^{(1)}\|_2 = \sqrt{7}$. Si $u = m^{(1)} \pm s e^{(1)}$, tenemos $P(u)m^{(1)} = \mp s e^{(1)}$; es decir, la matriz $P_1 = P(u)$ envía el vector $m^{(1)}$ a un vector con sólo la primera componente no nula. Dado que la primera componente de $m^{(1)}$ es positiva, escogemos el vector $u = m^{(1)} + s e^{(1)}$ para evitar cancelaciones y podemos proceder a calcular la matriz $M_2 \equiv P_1 M = (P_1 m^{(1)} \; P_1 m^{(2)})$ y el vector $P_1 y$:

$$
\begin{aligned}
s &= \|m^{(1)}\|_2 = \sqrt{7}\,, \\
u &= m^{(1)} + \sqrt{7} e^{(1)} = \begin{pmatrix} 1+\sqrt{7} \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \\
\alpha &= \frac{2}{u^\top u} = \frac{1}{s(s+1)} = \frac{1}{\sqrt{7}(\sqrt{7}+1)} = \frac{1}{6} - \frac{\sqrt{7}}{42}\,, \\
P_1 m^{(1)} &= -\sqrt{7} e^{(1)} = \begin{pmatrix} -\sqrt{7} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \\
P_1 m^{(2)} &= m^{(2)} - \alpha (u^\top m^{(2)}) u = m^{(2)} - \frac{9-\sqrt{7}}{8} u \\
&= \frac{1}{8} \begin{pmatrix} -8\sqrt{7} \\ -5+\sqrt{7} \\ -3+\sqrt{7} \\ -1+\sqrt{7} \\ 1+\sqrt{7} \\ 3+\sqrt{7} \\ 5+\sqrt{7} \end{pmatrix}, \\
P_1 y &= y - \alpha (u^\top y) u = y - (\frac{199}{150} - \frac{139}{1050}\sqrt{7}) u \\
&= \frac{1}{1050} \begin{pmatrix} -1254\sqrt{7} \\ -868+139\sqrt{7} \\ -448+139\sqrt{7} \\ -49+139\sqrt{7} \\ 287+139\sqrt{7} \\ 350+139\sqrt{7} \\ 728+139\sqrt{7} \end{pmatrix}.
\end{aligned}
$$

Tenemos, así,

$$M_2 = P_1 M = \begin{pmatrix} -\sqrt{7} & -\sqrt{7} \\ 0 & \frac{-5+\sqrt{7}}{8} \\ 0 & \frac{-3+\sqrt{7}}{8} \\ 0 & \frac{-1+\sqrt{7}}{8} \\ 0 & \frac{1+\sqrt{7}}{8} \\ 0 & \frac{3+\sqrt{7}}{8} \\ 0 & \frac{5+\sqrt{7}}{8} \end{pmatrix} = \left(\begin{array}{c|c} -\sqrt{7} & -\sqrt{7} \\ \hline 0 & \overline{M}_2 \end{array}\right) ,$$

donde $\overline{M}_2$ es ahora una matriz 6x1; es decir, un vector de 6 filas que llamamos $\overline{m}^{(1)}$

$$\overline{M}_2 = \overline{m}^{(1)} = \frac{1}{8} \begin{pmatrix} -5+\sqrt{7} \\ -3+\sqrt{7} \\ -1+\sqrt{7} \\ 1+\sqrt{7} \\ 3+\sqrt{7} \\ 5+\sqrt{7} \end{pmatrix}$$

y que queremos transformar en un vector con sólo la primera componente no nula, mediante otra transformación $\overline{P}_2$ tal que

$$P_2 = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & \overline{P}_2 \end{array}\right) .$$

Hacemos, así,

$$\begin{aligned} s &= \|\overline{m}^{(1)}\|_2 = \frac{\sqrt{7}}{2} , \\ \overline{u} &= \overline{m}^{(1)} - s\overline{e}^{(1)} = \frac{1}{8} \begin{pmatrix} -5-3\sqrt{7} \\ -3+\sqrt{7} \\ -1+\sqrt{7} \\ 1+\sqrt{7} \\ 3+\sqrt{7} \\ 5+\sqrt{7} \end{pmatrix} , \\ \alpha &= \frac{2}{\overline{u}^\top \overline{u}} = \frac{1}{s(s+\frac{5-\sqrt{7}}{8})} = \frac{168-40\sqrt{7}}{133} , \\ \overline{P}_2 \overline{m}^{(1)} &= \frac{\sqrt{7}}{2}\overline{e}^{(1)} = \begin{pmatrix} \frac{\sqrt{7}}{2} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} , \end{aligned}$$

$$\overline{P}_2\ \overline{P_1 y} = \overline{P_1 y} - \alpha(\overline{u}^\top \overline{P_1 y})\overline{u} = \overline{P_1 y} - (\frac{10304+688\sqrt{7}}{9975})\overline{u}$$
$$= \begin{pmatrix} \frac{197}{350}\sqrt{7} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{pmatrix} ,$$

donde las cinco últimas componentes de $\overline{P}_2\ \overline{P_1 y}$ no han sido calculadas porque no se usarán.

Tenemos, pues,

$$M_3 = P_2 P_1 M = \sqrt{7}\begin{pmatrix} -1 & -1 \\ 0 & \frac{1}{2} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \left(\begin{array}{cc} \multicolumn{2}{c}{\tilde{R}} \\ \hline 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{array}\right) ,$$

$$P_2 P_1 y = \begin{pmatrix} -\frac{1254}{1050}\sqrt{7} \\ \frac{197}{350}\sqrt{7} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{pmatrix} = \left(\begin{array}{c} \widetilde{P_2 P_1 y} \\ \hline \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{array}\right)$$

y, por lo tanto, el sistema de ecuaciones normales $M^\top M a^* = M^\top y$ es equivalente a $\tilde{R}a^* = \widetilde{P_2 P_1 y}$:

$$\sqrt{7}\begin{pmatrix} -1 & -1 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix} = \sqrt{7}\begin{pmatrix} -\frac{1254}{1050} \\ \frac{197}{350} \end{pmatrix} ,$$

de solución:

$$a_0^* = \frac{12}{175} = 0.06857142857... , \quad a_1^* = \frac{197}{175} = 1.125714286...$$

b) y c) Estamos en el caso de aproximación polinomial. De hecho, el caso del apartado anterior era también de aproximación polinomial y podríamos aplicar los mismos métodos que allí. Resulta, pero, más conveniente generar polinomios ortogonales asociados al producto escalar discreto

$$(f,g) = \sum_{k=0}^{6} f(x_k)g(x_k) ,$$

de norma euclídea asociada $\|f\|_2 = \sqrt{(f,f)}$, a través de la recurrencia:

$$\psi_{-1}(x) = 0 , \quad \psi_0(x) = 1 ,$$

$$\psi_{j+1}(x) = (x-\beta_j)\psi_j(x) - \gamma_j\psi_{j-1}(x)\ (j \geq 0) ;$$

donde

$$\beta_j = \frac{(\psi_j, x\psi_j)}{(\psi_j, \psi_j)} \ (j \geq 0) \ , \quad \gamma_j = \frac{(\psi_j, \psi_j)}{(\psi_{j-1}, \psi_{j-1})} \ (j \geq 1) \ .$$

Tendremos en cuenta las observaciones siguientes:

1. Por comodidad, se ha escogido que los polinomios ortogonales sean mónicos: $\psi_j(x) = x^j + \cdots$ $(j \geq 0)$.

2. El conjunto de abscisas de aproximación consta de 7 abscisas equidistantes con paso $h = \frac{1}{4}$: $x_k = \frac{1}{4} + kh = (k+1)h$ $(k = 0 \div 6)$; por lo tanto, la familia de polinomios ortogonales sobre esta partición se compone de los 7 polinomios $\psi_0(x)$, $\psi_1(x), \ldots, \psi_6(x)$.

3. Al ser el conjunto de abscisas de aproximación simétrico respecto a la abscisa central $x_3 = 1$, se tiene $\beta_j = 1$ y $\psi_j(1+u) = (-1)^j \psi_j(1-u)$ $(j = 0 \div 6)$. Así, si expresamos los polinomios $\psi_j(x)$ en la variable $u = x - 1$, éstos serán impares (pares) en esta variable $x - 1$, si j es impar (par) $(j = 0 \div 6)$.

4. Usando estos polinomios ortogonales $\psi_j(x)$ $(j = 0 \div 6)$, la solución $p_n^*(x)$ del problema de aproximación polinomial por mínimos cuadrados

$$\|f - p_n^*\|_2 = \min_{p_n \in \mathcal{P}_n} \|f - p_n\|_2 \ ,$$

donde $\mathcal{P}_n$ denota el espacio vectorial de los polinomios de grado menor o igual que n, viene dada por

$$p_n^*(x) = \sum_{j=0}^{n} c_j^* \psi_j(x) \ , \quad c_j^* = \frac{(\psi_j, f)}{(\psi_j, \psi_j)} \ (j = 0 \div n) \ .$$

Por lo tanto, sólo es necesario conocer los polinomios ortogonales $\psi_0(x), \ldots, \psi_n(x)$ (en nuestro caso, para $n = 2, 4$); tampoco es necesario conocerlos explícitamente: basta conocer los valores $\psi_j(x_k)$ $(k = 0 \div 6)$ $(j = 0 \div n)$ que, además, se hallan sucesivamente, usando la fórmula recurrente.

Calculamos en la tabla siguiente los valores $\psi_j(x_k)$ $(k = 0 \div 6)$, así como $d_j = (\psi_j, \psi_j)$, c_j^* y γ_j, empezando con $\psi_0(x) = 1$, $\psi_1(x) = x - 1$, y haciendo $\psi_{j+1}(x) = (x-1)\psi_j(x) - \gamma_j \psi_{j-1}(x)$ $(j = 1 \div 3)$:

k	0	1	2	3	4	5	6	d_j	c_j^*	γ_j
x_k	0.25	0.50	0.75	1.00	1.25	1.50	1.75			
y_k	0.40	0.50	0.90	1.28	1.60	1.66	2.02			
$\psi_0(x_k)$	1	1	1	1	1	1	1	7	$\frac{209}{175}$	
$\psi_1(x_k)$	$-\frac{3}{4}$	$-\frac{2}{4}$	$-\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{4}$	$\frac{7}{4}$	$\frac{197}{175}$	$\frac{1}{4}$
$\psi_2(x_k)$	$\frac{5}{16}$	0	$-\frac{3}{16}$	$-\frac{4}{16}$	$-\frac{3}{16}$	0	$\frac{5}{16}$	$\frac{21}{64}$	$-\frac{52}{525}$	$\frac{3}{16}$
$\psi_3(x_k)$	$-\frac{3}{32}$	$\frac{3}{32}$	$\frac{3}{32}$	0	$-\frac{3}{32}$	$-\frac{3}{32}$	$\frac{3}{32}$	$\frac{27}{512}$	$-\frac{32}{75}$	$\frac{9}{56}$
$\psi_4(x_k)$	$\frac{9}{448}$	$-\frac{21}{448}$	$\frac{3}{448}$	$\frac{18}{448}$	$\frac{3}{448}$	$-\frac{21}{448}$	$\frac{9}{448}$	$\frac{99}{14336}$	$\frac{1856}{825}$	

Podemos decir así que las mejores aproximaciones por mínimos cuadrados de f por polinomios de grado 2 y de grado 4, respectivamente, son

$$\begin{aligned} p_2^*(x) &= c_0^* + c_1^* \psi_1(x) + c_2^* \psi_2(x) \ , \\ p_4^*(x) &= c_0^* + c_1^* \psi_1(x) + c_2^* \psi_2(x) + c_3^* \psi_3(x) + c_4^* \psi_4(x) \ , \end{aligned}$$

donde

$$\psi_0(x) = 1 \ , \quad \psi_1(x) = x - 1 \ , \quad \psi_2(x) = (x-1)\psi_1(x) - \gamma_1 \psi_0(x) \ ,$$

$$\psi_3(x) = (x-1)\psi_2(x) - \gamma_2\psi_1(x) \ , \quad \psi_4(x) = (x-1)\psi_3(x) - \gamma_3\psi_2(x) \ ,$$

con

$$\gamma_1 = \frac{1}{4} \ , \quad \gamma_2 = \frac{3}{16} \ , \quad \gamma_3 = \frac{9}{56} \ .$$

Una manera eficaz de evaluar $p_2^*(x)$ y $p_4^*(x)$ consiste en aplicar la regla de Clenshaw. Por ejemplo, para el cálculo de $p_2^*(x)$ ésta da:

$$q_2 = c_2^* \ , \quad q_1 = (x-1)q_2 + c_1^* \ , \quad p_2^*(x) = q_0 = (x-1)q_1 - \gamma_1 q_2 + c_0^* \ .$$

Notemos ahora que la solución del apartado a) también la tenemos resuelta aquí y viene dada por

$$p_1^*(x) = c_0^* + c_1^*\psi_1(x) = \frac{209}{175} + \frac{197}{175}(x-1) \ .$$

Además, los cálculos efectuados para obtener una aproximación $p_n^*(x)$ son muy útiles si quiere calcularse después una aproximación $p_{n'}^*(x)$, con $n' > n$.

Finalmente, calculemos los errores $e_j \equiv \|f - p_j^*\|_2$ de las aproximaciones $p_j^*(x)$ ($j = 0 \div 4$). Para ello, obtendremos primero

$$e_0^2 \equiv \|f - p_0^*\|_2^2 = \sum_{k=0}^{6} (f(x_k) - c_0^*)^2 = 2.2701714... \ ;$$

usando ahora que $f - p_1^*$ es ortogonal a $p_1^* - p_0^*$, el teorema de Pitágoras da

$$e_0^2 \equiv \|f - p_0^*\|_2^2 = \|f - p_1^*\|_2^2 + \|p_1^* - p_0^*\|_2^2 = e_1^2 + \|p_1^* - p_0^*\|_2^2 \ ;$$

$$\begin{aligned}
e_1^2 &= e_0^2 - \|p_1^* - p_0^*\|_2^2 = e_0^2 - c_1^{*2}\|\psi_1\|_2^2 = 0.052514... \ , \\
e_2^2 &= e_1^2 - \|p_2^* - p_1^*\|_2^2 = e_1^2 - c_2^{*2}\|\psi_2\|_2^2 = 0.049295... \ , \\
e_3^2 &= e_2^2 - \|p_3^* - p_2^*\|_2^2 = e_2^2 - c_3^{*2}\|\psi_3\|_2^2 = 0.039695... \ , \\
e_4^2 &= e_3^2 - \|p_4^* - p_3^*\|_2^2 = e_3^2 - c_4^{*2}\|\psi_4\|_2^2 = 0.0047445...
\end{aligned}$$

Observamos que las aproximaciones de grados 2 y 3 no hacen bajar substancialmente el error de la aproximación por mínimos cuadrados. Esto muestra que la aproximación de grado 1 es ya bastante buena en este caso, sobre todo si tenemos en cuenta los pocos cálculos que requiere.

d) Con el fin de hallar una expresión lineal de aproximación, tomamos las variables $Y = \ln y$, $X = \ln x$ y entonces hacemos $a_0 = \ln A$ y $a_1 = \alpha$; resulta la relación $Y = a_0 + a_1 X$.

Consideramos ahora la tabla en las variables (X, Y):

X	Y
-1.386294	-0.916290
-0.693147	-0.693147
-0.287682	-0.105361
0.000000	0.246860
0.223144	0.470004
0.405465	0.506817
0.559616	0.703098

Haciendo los mismos cálculos del apartado a), utilizando el método de Gram-Schmidt, tenemos

$$A = QR = \begin{pmatrix} 1 & -1.217880 \\ 1 & -0.524733 \\ 1 & -0.119268 \\ 1 & 0.168414 \\ 1 & 0.391558 \\ 1 & 0.573879 \\ 1 & 0.728030 \end{pmatrix} \begin{pmatrix} 1 & -0.168414 \\ 0 & 1 \end{pmatrix} ,$$

$$D = \begin{pmatrix} 7 & 0 \\ 0 & 2.813847552 \end{pmatrix} .$$

El sistema de ecuaciones normales tiene entonces la solución:

$$a_0^* = 0.18114 \ , \quad a_1^* = 0.89577 \ ;$$

así, los valores correspondientes de A y α serán:

$$A^* = \exp(a_0^*) = 1.19859 \ , \quad \alpha^* = a_1^* = 0.89577 \ .$$

Problema 3.19 *Hallar la aproximación minimax de la forma* $x^2 + \hat{a}_1 x + \hat{a}_0$ *a la función* $\operatorname{sen} x$ *en el intervalo* $[0, 1]$.

SOLUCIÓN:

El problema equivale a obtener la aproximación minimax polinomial de grado menor o igual que 1 a la función $f(x) = \operatorname{sen} x - x^2$:

$$\hat{p}_1(x) = \hat{a}_0 + \hat{a}_1 x \ .$$

El teorema de caracterización de Chebichev permite hallar la aproximación buscada, imponiendo la propiedad de equioscilación que asegura que la función error $e_1(x) = f(x) - \hat{a}_0 - \hat{a}_1 x$ presenta 3 abscisas ξ_0, ξ_1 y ξ_2 en las cuales alcanza valores extremos $\pm\|e_1\|_\infty$ con alternancia de signos:

$$e_1(\xi_0) = -e_1(\xi_1) = e_1(\xi_2) = \pm\|e_1\|_\infty \ ;$$

además, dado que $f^{(2)}(x) = -\operatorname{sen} x - 2$ no cambia de signo sobre $[0, 1]$, ξ_0 y ξ_2 coinciden con los extremos del intervalo, $a = 0$, $b = 1$. La otra abscisa ξ_1 se caracteriza por la anulación de la derivada primera del error

$$e_1'(\xi_1) = f'(\xi_1) - \hat{a}_1 = 0 \ .$$

De la propiedad de equioscilación en los extremos $e_1(a) = e_1(b)$, sale la fórmula para la pendiente de la recta de aproximación minimax

$$\hat{a}_1 = \frac{f(b) - f(a)}{b - a} \ ,$$

y, de la propiedad de equioscilación en ξ_1 y a,

$$\hat{a}_0 = \frac{1}{2}[f(a) + f(\xi_1) - (a + \xi_1)\hat{a}_1] \ .$$

Para el caso propuesto, tenemos que

$$\hat{a}_1 = \operatorname{sen} 1 - 1 = -0.158529 \ ;$$

la abscisa ξ_1 cumple la ecuación

$$\cos \xi_1 - 2\xi_1 + 1 - \operatorname{sen} 1 = 0$$

que, usando los métodos del capítulo 5, tiene la solución $\xi_1 = 0.5145272...$

Así,

$$\hat{a}_0 = \frac{1}{2}(\operatorname{sen} \xi_1 - \xi_1^2 - \hat{a}_1 \xi_1) = 0.154476...$$

Problema 3.20 *Hallar el polinomio de aproximación minimax de grado 3 a la función $f(x) = x^4$ en el intervalo $[0, 10]$. Acotar el error cometido en todo el intervalo.*

SOLUCIÓN:

La obtención de este polinomio es muy simple si nos basamos en el hecho de que el polinomio mónico de grado $n+1$ con norma del máximo mínima en $[-1, 1]$ es el polinomio mónico de Chebichev

$$\tilde{T}_{n+1}(t) = \frac{1}{2^n} T_{n+1}(t) \ .$$

En el intervalo $[0, 10]$, el polinomio de Chebichev puede hallarse aplicando la transformación afín

$$t \in [-1, 1] \quad \longrightarrow \quad x = 5(t + 1) \in [0, 10] \ ,$$

de inversa

$$x \in [0, 10] \quad \longrightarrow \quad t = \frac{x - 5}{5} \in [-1, 1] \ ,$$

al anterior polinomio $T_{n+1}(t)$; obtenemos así el polinomio

$$T_{n+1}\left(\frac{x - 5}{5}\right) \ .$$

El polinomio mónico asociado

$$\tilde{T}_{n+1}\left(\frac{x-5}{5}\right) = \frac{5^{n+1}}{2^n} T_{n+1}\left(\frac{x-5}{5}\right)$$

será consecuentemente el polinomio mónico de norma mínima en $[0,10]$; por lo tanto, el polinomio

$$x^{n+1} - \tilde{T}_{n+1}(x)$$

es el polinomio de aproximación minimax de grado menor o igual que n a x^{n+1}.

En el caso en cuestión, basta tomar $n = 3$.

Tenemos:

$$\begin{aligned} T_4(t) &= 8t^4 - 8t^2 + 1 \ , \\ T_4\left(\frac{x-5}{5}\right) &= \frac{8(x^4 - 20x^3 + 150x^2 - 500x + 625)}{625} \\ &\quad -\frac{8(x^2 - 10x + 25)}{25} + 1 \\ &= \frac{8x^4 - 160x^3 + 1000x^2 - 2000x + 625}{625} \ , \\ \tilde{T}_4\left(\frac{x-5}{5}\right) &= x^4 - (20x^3 - 125x^2 + 250x - \frac{625}{8}) \ . \end{aligned}$$

El polinomio de aproximación minimax buscado es, pues,

$$\hat{p}_3(x) = 20x^3 - 125x^2 + 250x - \frac{625}{8} \ .$$

El error cometido estará acotado por $\|\tilde{T}_4\|_\infty$ en $[0,10]$, que resulta ser

$$\frac{625}{8}\|T_4(\frac{x-5}{5})\|_{\infty,[0,10]} = \frac{625}{8}\|T_4(t)\|_{\infty,[-1,1]} = \frac{625}{8} \ .$$

Problema 3.21 *Utilizar la economización de Lanczos para calcular un polinomio de cuarto grado que aproxime, con un error de magnitud menor que* $2 \cdot 10^{-4}$*, a la función*

$$f(x) = \int_0^x \frac{e^t - 1}{t} dt \ , \quad \forall x \in [-1,1] \ .$$

SOLUCIÓN:

Hallamos primero el desarrollo de Taylor de f cerca de $x_0 = 0$.
Tenemos $f(0) = 0$ y

$$f'(x) = \frac{e^x - 1}{x} \equiv \frac{F(x)}{x} = \frac{1}{x}(x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + R_n(x)) \ ,$$

donde

$$R_n(x) = \frac{F^{(n+1)}(\xi(x))}{(n+1)!}x^{n+1} = \frac{e^{\xi(x)}}{(n+1)!}x^{n+1} \ , \quad \xi(x) \in <0, x> \ .$$

Así,

$$f'(x) = 1 + \frac{x}{2!} + \frac{x^2}{3!} + \cdots + \frac{x^{n-1}}{n!} + \frac{e^{\xi(x)}}{(n+1)!}x^n \ , \quad \xi(x) \in <0, x> \ ;$$

por lo tanto,

$$\begin{aligned} f(x) &= x + \frac{x^2}{2 \cdot 2!} + \frac{x^3}{3 \cdot 3!} + \cdots + \frac{x^n}{n \cdot n!} \\ &\quad + \int_0^x \frac{e^{\xi(t)}}{(n+1)!}t^n dt \ , \quad \xi(t) \in <0, t> \ . \end{aligned}$$

Usando el teorema del valor medio para integrales el resto del desarrollo puede escribirse también como

$$\frac{e^\eta}{(n+1)!}\int_0^x t^n dt = \frac{e^\eta}{(n+1)(n+1)!}x^{(n+1)} \ , \quad \eta \in <0, x> \ .$$

Para el cálculo de f podemos escoger una primera posibilidad que consiste en aproximarla por su desarrollo de Taylor hasta un cierto orden n de manera que produzca unos errores acotados por $2 \cdot 10^{-4}$.

El resto del desarrollo hallado hasta orden n puede acotarse por $\frac{e}{(n+1)(n+1)!}$ en el intervalo $[-1, 1]$. $n = 6$ es el valor menor para el que dicha cota sea menor que el error permitido y se tiene

$$f(x) = x + \frac{x^2}{2 \cdot 2!} + \frac{x^3}{3 \cdot 3!} + \cdots + \frac{x^6}{6 \cdot 6!} \pm 7.7 \cdot 10^{-5} \ , \quad x \in [-1, 1] \ .$$

La segunda posibilidad intenta economizar los cálculos, tratando de hallar un polinomio de aproximación a la función de grado más bajo que 6 que permita calcularla con un error menor que el permitido. Se usa a continuación la economización de Lanczos para obtenerlo.

Se reduce primero el desarrollo de grado 6 hallado a un polinomio de grado 5, restándole un polinomio de grado 6 que lo pueda conseguir y que sea de norma mínima: el múltiplo del polinomio de Chebixev $T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1$ que tenga el mismo coeficiente principal.

Dado que el coeficiente principal de $\frac{x^6}{6 \cdot 6!}$ es $\frac{1}{4320}$, será necesario restarle

$$\frac{1}{2^5 \cdot 4320}T_6(x) = \frac{1}{138240}T_6(x) \ ,$$

que produce un error acotado por $0.072 \cdot 10^{-4}$ en todo el intervalo.

Notamos que esta modificación no alterará el coeficiente de x^5 del desarrollo f, debido a la paridad de los polinomios de Chebichev. Así, si de manera análoga le restamos también un múltiplo adecuado del polinomio de grado 5 de Chebichev $T_5(x) = 16x^5 - 20x^3 + 5x$ a $f(x) - \frac{1}{138240}T_6(x)$, conseguiremos un polinomio de grado 4 que quizás todavía permitirá hacer el cálculo de la función f con un error menor que el permitido.

Tenemos, así,

$$\begin{aligned} f(x) &= x + \frac{x^2}{2 \cdot 2!} + \frac{x^3}{3 \cdot 3!} + \frac{x^4}{3 \cdot 4!} \\ &\quad + \frac{x^5 - \frac{1}{16}T_5(x)}{5 \cdot 5!} + \frac{x^6 - \frac{1}{32}T_6(x)}{6 \cdot 6!} \\ &\quad + \frac{1}{16 \cdot 600}T_5(x) + \frac{1}{32 \cdot 4320}T_6(x) \pm 0.77 \cdot 10^{-4} \ , \quad x \in [-1, 1] \ . \end{aligned}$$

Acotamos ahora los errores de economización, teniendo en cuenta que se cumple $\|T_j\|_\infty = 1$,

$$\left| \frac{1}{16 \cdot 600}T_5(x) + \frac{1}{32 \cdot 4320}T_6(x) \right| < 1.12 \cdot 10^{-4} \ ;$$

dado que el error del desarrollo de Taylor inicial era menor que $0.77 \cdot 10^{-4}$, el error total todavía es menor que el error permitido.

Utilizando las expresiones de $T_5(x)$ y $T_6(x)$, resulta el desarrollo economizado buscado

$$\begin{aligned} \int_0^x \frac{e^t - 1}{t} dt &= \frac{1}{138240}(1 + 138168x + 34542x^2 + 1056x^3 + 1488x^4) \\ &\quad \pm 2 \cdot 10^{-4} \ , \quad x \in [-1, 1] \ . \end{aligned}$$

Notamos que con otro paso de economización no sería posible asegurar que los errores estuviesen todavía dentro de los límites permitidos.

Problema 3.22 *a) La función $\omega_m(x) = (x - x_0)(x - x_1) \cdots (x - x_m)$ es un factor del error en la interpolación polinomial en las abscisas x_k $(k = 0 \div m)$. Si hacemos interpolación de Chebichev en un intervalo $[a, b]$ cualquiera, hallar $\|\omega_m\|_\infty$ en función de m, a y b.*

b) Queremos interpolar la función $f(x) = \frac{1}{1+2x}$ en el intervalo $[4, 5]$. Dar una estimación del número de abscisas de interpolación que se necesitan para que el error de interpolación sea menor que 10^{-6} en todo el intervalo.

c) Comparar este error con el que se produciría con interpolación equidistante con el mismo número de abscisas que el hallado en b).

SOLUCIÓN:

a) La interpolación de Chebichev de grado m en el intervalo $[-1,1]$ tiene lugar en las llamadas abscisas de Chebichev: los ceros del polinomio de grado $m+1$ de Chebichev

$$T_{m+1}(t) = \cos((m+1)\arccos t) \ ;$$

es decir, en las abscisas

$$t_k = \cos\frac{(2k+1)\pi}{2(m+1)} \ (k = 0 \div m) \ .$$

En otro intervalo $[a,b]$, consideraremos como interpolación de Chebichev la que tiene lugar en las abscisas de Chebichev que se obtienen por la transformación de las del intervalo $[-1,1]$ mediante el cambio afín

$$x = \frac{b-a}{2}t + \frac{a+b}{2} \ ;$$

esto es, en las abscisas

$$x_k = \frac{b-a}{2}\cos\frac{(2k+1)\pi}{2(m+1)} + \frac{a+b}{2} \ (k = 0 \div m) \ .$$

El factor $(t-t_0)(t-t_1)\cdots(t-t_m)$ del error en la interpolación de Chebichev en $[-1,1]$ tiene, pues, los mismos ceros que $T_{m+1}(t)$; debido a que el coeficiente principal de $T_{m+1}(t)$ es 2^m, tenemos

$$(t-t_0)(t-t_1)\cdots(t-t_m) = \frac{1}{2^m}T_{m+1}(t) \ .$$

Además,

$$\omega_m(t) = \left(\frac{b-a}{2}\right)^{m+1}(t-t_0)(t-t_1)\cdots(t-t_m) \ .$$

Notamos finalmente que la relación $\|T_{m+1}\|_\infty = 1$ nos permite hallar la norma de ω_m:

$$\|\omega_m\|_{\infty,[a,b]} = \left(\frac{b-a}{2}\right)^{m+1}\frac{1}{2^m} = 2\left(\frac{b-a}{4}\right)^{m+1} \ .$$

b) La expresión anterior nos permite dar una fórmula de acotación del error en la interpolación de Chebichev en un intervalo $[a,b]$ cualquiera a una función f, supuesta $m+1$ veces diferenciable con continuidad

$$|f(x) - p_m(x)| \le 2\frac{M_{m+1}}{(m+1)!}\left(\frac{b-a}{4}\right)^{m+1} \ ,$$

donde M_{m+1} es una cota superior de $\left|f^{(m+1)}\right|$ en $[a,b]$.

La derivada $m+1$ de la función $f(x) = \frac{1}{1+2x}$ es

$$f^{(m+1)}(x) = (-1)^{m+1}2^{m+1}(m+1)!\frac{1}{(1+2x)^{m+2}} \ ;$$

su valor absoluto, al ser decreciente en el intervalo $[a,b] = [4,5]$, puede acotarse por su valor en el extremo inferior $x = 4$; así, para $x \in [4,5]$,

$$|f(x) - p_m(x)| \le 2\cdot 2^{m+1}\frac{1}{9^{m+2}}\frac{1}{4^{m+1}} = \frac{4}{18^{m+2}} \ .$$

Para $m = 3$, esta cota vale aproximadamente $2.1 \cdot 10^{-6}$ y, para $m = 4$, $1.17 \cdot 10^{-7}$. Resulta así que, haciendo una interpolación de Chebichev en 5 abscisas ($m = 4$), podemos estar seguros que el error de ésta será menor que 10^{-6} en todo el intervalo $[4, 5]$.

c) Si llevásemos a término una interpolación equidistante en las 5 abscisas $x_k = -4+kh_4$ ($k = 0 \div 4$), con $h_4 = \frac{1}{4}$, el error de interpolación sería menor que

$$\frac{M_5}{5!} h_4^5 \Omega_4 \simeq 2.13 \cdot 10^{-7} \ ,$$

donde

$$\Omega_4 = \frac{\sqrt{950 + 58\sqrt{145}}}{5\sqrt{5}} \simeq 3.63$$

está introducida en el problema 3.3.

Tendríamos así un error de interpolación menor que 10^{-6}, aunque dos veces mayor que el error de interpolación de Chebichev en 5 abscisas. La gran ventaja de la interpolación de Chebichev es que, además, requiere menos cálculo que la equidistante.

PROBLEMAS PROPUESTOS

1. Hallar el polinomio de interpolación a la tabla

$$\begin{array}{c|cccc} x_k & 0 & 1 & 2 & 4 \\ \hline f_k & 1 & 5 & 10 & 24 \end{array},$$

usando los métodos de: a) Lagrange, b) las diferencias divididas de Newton, c) Aitken y d) Neville.

2. Repetir el ejercicio anterior con la tabla

$$\begin{array}{c|ccccc} x_k & 0 & 1 & 2 & 4 & 8 \\ \hline f_k & 1 & 5 & 10 & 24 & 50 \end{array}.$$

¿Se pueden aprovechar los cálculos hechos con los cuatro métodos del problema anterior? ¿Cuáles de los cuatro métodos permiten aprovechar los cálculos hechos en el ejercicio anterior?

3. Queremos tabular la *función de Bessel de orden 0*

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

en abscisas equidistantes del intervalo $[0,1]$. ¿Qué paso de tabulación ha de usarse para que todos los valores de $J_0(x)$ $(x \in [0,1])$ obtenidos por interpolación lineal (usando los dos puntos de la tabla más próximos) tengan un error debido a la interpolación menor que $\frac{1}{2}10^{-6}$? ¿Y si hacemos interpolación cúbica (usando los cuatro puntos más próximos)?

4. Interpolamos la función e^x en $[0,1]$ por un polinomio de grado n en abscisas equidistantes. Dar el valor de n mínimo para que el error de interpolación sea menor que 10^{-6} en cualquier punto de $[0,1]$. Hacer lo mismo tomando como cota de error 10^{-12}.

5. Sea $f \in \mathcal{C}^{\infty}(\mathbb{R})$ con $\mid f^{(l)}(x) \mid \leq 4^l$ $(l \geq 0), \forall x \in [0,1]$. Queremos construir polinomios de interpolación $p_m(x)$ $(m \geq 0)$ en abscisas equidistantes. Sabemos que, debido a los errores de redondeo, f se evalua con un error menor que 10^{-5} y llamamos $\bar{p}_m(x)$ al polinomio obtenido. ¿Qué valor de m hace que la cota del error

$$\max_{x\in[0,1]} \mid f(x) - \bar{p}_m(x) \mid$$

sea mínima?

6. Consideramos la función
$$f(x) = \frac{1}{1+x}$$
y sea $p_m(x)$ el polinomio de interpolación a f en las abscisas 1, a, a^2, ..., a^m con $0 < a < 1$ $(m \geq 0)$; demostrar que
$$\lim_{m\to\infty} p_m(0) = f(0) \ .$$

7. Sea $p_m(x)$ el polinomio interpolador a la función
$$f(x) = \frac{1}{1+x^2}$$
en las abscisas equidistantes $x_k^{(m)} = -5 + \frac{10k}{m}$ $(k = 0 \div m)$ $m \geq 0$. Verificar la divergencia de la sucesión $(p_m(4))_{m\geq 0}$ (*fenómeno de Runge*).

8. a) Hallar el polinomio de Taylor $p_n(x)$ de grado n de la función $f(x) = e^x$ para cualquier n.

   b) Dar expresiones para los errores $f(x) - p_n(x)$ y acotarlas.

   c) ¿Para qué valores de n puede asegurarse que el polinomio de Taylor dará $f(x)$ con 8 cifras decimales correctas, para todo $x \in [-1, 1]$?

   Repetir los tres apartados para $f(x) = \operatorname{sen} x$ en $[0, \pi/2]$.

9. Considerar los desarrollos de Taylor de las funciones siguientes en $x_0 = 0$ en los intervalos que se indican:

   i) $\ln(1+x)$ en $(-1, 1)$;

   ii) $(1+x)^{-1}$ en $(-1, 1)$;

   iii) $\cos x$ en $(-\pi/2, \pi/2)$.

   Responder, para cada función ,a las cuestiones siguientes:

   a) ¿Cuántos términos deberíamos tomar para asegurar una precisión en la aproximación de 8 cifras decimales correctas, para cualquier valor de x en el intervalo dado?

   b) ¿Para qué valores de x tenemos aquella precisión, si tomamos sólo los tres primeros términos no nulos del desarrollo?

10. a) Hallar los desarrollos de Taylor de las funciones siguientes en cualquier x_0:

    i) e^x, $\cosh x$, $\operatorname{senh} x$, a^x $(a > 0)$;

    ii) $\operatorname{sen} x$, $\cos x$;

iii) $\ln x$, $\log x$, x^α.

b) Hacer explícitos estos desarrollos para $x_0 = 2$ y $x_0 = \pi$.

11. Mediante el uso de desarrollos de Taylor en $x_0 = 0$, hallar los límites, cuando x tiende a 0, de las funciones siguientes:

$$\frac{\operatorname{sen} x^2}{\cos x - 1} \ , \quad \frac{x^4}{(e^x - 1)^2} \ , \quad \left(\frac{1+x}{1-x}\right)^{\frac{1}{x}} \ .$$

12. Queremos calcular la función

$$f(x) = \frac{-2x^3 + 3x - 3 \operatorname{sen} x \cos x}{3x^3} \ ,$$

en $\bar{x} = 0.1$.

a) ¿Con qué precisión ha de conocerse x cerca de $\bar{x}$ para que el error propagado a $f(x)$ sea menor que 10^{-3}?

b) Dar otra fórmula para calcular $f(\bar{x})$ que sea mejor desde el punto de vista numérico, si se tienen en cuenta también los errores en los cálculos.

13. Hallar los desarrollos de Taylor de grado n en $x_0 = 0$ de las funciones siguientes, expresándolos en términos de los números de Bernoulli:

a) $\cot x \ \left(= i\frac{e^{2ix}+1}{e^{2ix}+1}\right)$;

b) $\tan x \ \left(= \cot x - 2i\left(1 + \frac{2}{e^{4ix}-1}\right)\right)$;

c) $\frac{x}{\operatorname{sen} x} \ \left(= \frac{2ix}{e^{ix}-1} - \frac{2ix}{e^{2ix}-1}\right)$;

d) $\ln \frac{\operatorname{sen} x}{x}$.

14. Los *números de Euler* $(E_j)_{j\geq 0}$ se definen por

$$\frac{1}{\cos x} = E_0 - \frac{E_2}{2!}x^2 - \frac{E_4}{4!}x^4 + \cdots$$

a) Probar que se cumplen las relaciones

$$E_0 = 1 \ , \quad \sum_{r=0}^{s} \binom{2s}{2r} E_{2r} = 0 \quad (s \geq 1) \ .$$

b) Calcular el desarrollo de Taylor en $x_0 = 0$ de $\ln(\cos x)$.

15. Probar que los coeficientes c_j $(j \geq 0)$ del desarrollo de Taylor

$$(1 - x - x^2)^{-1} = c_0 + c_1 x + c_2 x^2 + \cdots$$

verifican la relación $c_{j+1} = c_j + c_{j-1}$ $(j \geq 1)$; son los llamados *números de Fibonacci*.

16. Los *polinomios de Legendre* $(P_j(t))_{j\geq 0}$ pueden definirse a partir de su *función generatriz* $F(t,x) = (1 - 2tx + x^2)^{-1}$. Dichos polinomios son entonces los coeficientes del desarrollo de Taylor de $F(t,x)$ respecto a x en $x = 0$:

$$F(t,x) = P_0(t) + P_1(t)x + P_2(t)x^2 + \cdots \ , \quad |x| < 1, \quad |\, t \,| \leq 1 \ .$$

a) Comprobar que

$$P_0(t) = 1 \ , \quad P_1(t) = t \ , \quad P_2(t) = \frac{1}{2}(3t^2 - 1) \ , \quad P_3(t) = \frac{1}{2}(5t^3 - 3t) \ .$$

b) Demostrar que $P_j(-t) = (-1)^j P_j(t)$ $(j \geq 0)$.

c) Para $j \geq 1$, hallar las relaciones de recurrencia:

$$(j+1)P_{j+1}(t) = (2j+1)tP_j(t) - jP_{j-1}(t) \ ,$$

$$2P_j(t) = P'_{j+1}(t) - 2tP'_j(t) + P'_{j-1} \ ,$$

$$(2j+1)P_j(t) = P'_{j+1}(t) - P'_{j-1}(t) \ .$$

17. Los polinomios de Chebichev $(T_j(t))_{j\geq 0}$ pueden definirse también por la *función generatriz*

$$\frac{1 - tx}{1 - 2tx + x^2} = T_0(t) + T_1(t)x + T_2(t)x^2 + \cdots \ , \quad |\, x \,| < 1, \quad |\, t \,| \leq 1 \ .$$

a) Hallar $T_j(t)$ $(j = 0, 1, 2, 3, 4)$.

b) Demostrar que $T_j(-t) = (-1)^j T_j(t)$ $(j \geq 0)$.

c) Probar que $T_{j+1}(t) = 2tT_j(t) - T_{j-1}(t)$ $(j \geq 1)$.

d) Demostrar que $T_j(t) = \cos(j \arccos t)$, $|\, t \,| < 1$ $(j \geq 0)$.

18. Los *polinomios de Hermite* $(H_j(t))_{j\geq 0}$ pueden definirse por la función generatriz

$$e^{2tx - x^2} = H_0(t) + \frac{H_1(t)}{1!}x + \frac{H_2(t)}{2!}x^2 + \cdots \ , \quad |x| < 1 \ , \quad t \in \mathbb{R} \ .$$

a) Hallar $H_j(t)$ $(j = 0, 1, 2, 3, 4)$.

b) Demostrar que $H_j(-t) = (-1)^j H_j(t)$ $(j \geq 0)$.

c) Probar las relaciones de recurrencia:

$$H_{j+1}(t) = 2tH_j(t) - 2jH_{j-1}(t) \ , \quad H_j'(t) = 2jH_{j-1}(t) \ \ (j \geq 1) \ .$$

d) Probar que

$$H_j(t) = (-1)^j e^{t^2} \frac{d^j}{dx^j} e^{-t^2} \ .$$

19. Sea

$$f(x) = x^9 \operatorname{sen} \frac{1}{x} \ , \quad \text{si } x \neq 0 \ ; \quad f(0) = 0 \ .$$

a) Probar que el polinomio de Taylor de grado n $(0 \leq n \leq 8)$ es idénticamente nulo.

b) Si

$$R_4(x) = x^9 \operatorname{sen} \frac{1}{x} - \left(x^8 - \frac{x^6}{3!} + \frac{x^4}{5!} \right) \ ,$$

probar que $|R_4(x)| \leq \frac{x^2}{7!}$ y que la fórmula aproximada

$$x^9 \operatorname{sen} \frac{1}{x} \simeq x^8 - \frac{x^6}{3!} + \frac{x^4}{5!}$$

es especialmente apropiada para x cumpliendo $\frac{1}{\sqrt{72}} \leq |\, x \,| \leq \frac{1}{\sqrt{42}}$; es decir, que es mejor que las fórmulas

$$x^9 \operatorname{sen} \frac{1}{x} \simeq x^8 - \frac{x^6}{3!} \ , \quad x^9 \operatorname{sen} \frac{1}{x} \simeq x^8 - \frac{x^6}{3!} + \frac{x^4}{5!} - \frac{x^2}{7!} \ , \quad \text{etc.}$$

20. *(Interpolación de Hermite)* a) Demostrar que el polinomio p_{2m+1}, de grado menor o igual que $2m+1$, cumpliendo:

$$p_{2m+1}(x_k) = f_k \ , \quad p_{2m+1}'(x_k) = f_k' \ \ (k = 0 \div m) \ ,$$

existe siempre y es único.

b) Demostrar que

$$p_{2m+1}(x) = \sum_{i=0}^{m} f_i \Phi_i(x) + \sum_{i=0}^{m} f_i' \Psi_i(x) \ ,$$

con

$$\Phi_i(x) = (1 - 2l_i'(x)(x - x_i)) l_i^2(x) \ , \quad \Psi_i(x) = (x - x_i) l_i^2(x) \ ,$$

siendo $l_i(x)$ el polinomio de Lagrange asociado a la abscisa x_i $(i = 0 \div m)$.

c) Demostrar también:

$$\sum_{i=0}^{m} \Phi_i(x) = 1 \ ,$$

$$\sum_{i=0}^{m} \Phi_i(x) x_i^j + j \sum_{i=0}^{m} \Psi_i(x) x_i^{j-1} = x^j \ (j = 1 \div 2m+1) \ ,$$

$$\sum_{i=0}^{m} \Phi_i(x) x_i^{2m+2} + (2m+2) \sum_{i=0}^{m} \Psi_i(x) x_i^{2m+1} = x^{2m+2} - (x-x_0)^2 \cdots (x-x_m)^2 \ .$$

d) Hallar el polinomio interpolador de Hermite $p_5(x)$ a $f(x) = e^x$ en las abscisas $x_0 = -1$, $x_1 = 0$ y $x_2 = 1$, usando las fórmulas de a) y diferencias divididas generalizadas. Acotar el error $|f(x) - p_5(x)|$ en el intervalo $[-1, 1]$.

21. a) Dada la función $f(x) = \frac{1}{1+x}$ en el intervalo $[0, 1]$, calcular las normas

$$\|f\|_p = \left(\int_0^1 |f(x)|^p \right)^{\frac{1}{p}} dx \quad (p \geq 1)$$

y $\|f\|_\infty$.

b) Comprobar que

$$\lim_{p \to 1} \|f\|_p = \|f\|_1 \ , \quad \lim_{p \to \infty} \|f\|_p = \|f\|_\infty \ .$$

22. Sea w una función peso en $[a, b]$ normalizada por $\int_a^b w(x)dx = 1$. Considerar las normas $\|f\|_\infty$ y $\|f\|_{2,w}$ en $[a, b]$.

a) ¿Existe alguna función f tal que $\|f\|_\infty < \|f\|_{2,w}$?

b) ¿Existe alguna función f tal que $\|f\|_\infty > \|f\|_{2,w}$?

En caso de existencia, dar un ejemplo.

23. Hallar a, r reales y $\alpha \in [0, 2\pi]$ tales que la expresión

$$y(x) = a + r \operatorname{sen}(x + \alpha)$$

aproxime de manera óptima, en el sentido de los mínimos cuadrados, a los puntos:

$$(-\frac{\pi}{2}, 1), \ (0, 0), \ (\frac{\pi}{2}, \frac{1}{2}), \ (\pi, 1) \ .$$

24. Resolver por mínimos cuadrados los siguientes sistemas lineales sobredeterminados:

$$\left(\begin{array}{cc|c} 1 & 1 & 5 \\ 1 & 2 & 9 \\ 1 & 3 & 12 \\ 1 & 4 & 16 \end{array} \right) \ , \quad \left(\begin{array}{ccc|c} 1 & 1 & 1 & 5 \\ 1 & 1 & 2 & 9 \\ 1 & 1 & 3 & 12 \\ 1 & 1 & 4 & 16 \end{array} \right) \ .$$

25. Sabemos que los pesos atómicos del oxígeno y del nitrógeno son aproximadamente $O = 16$ y $N = 14$; utilizar los pesos moleculares de los seis óxidos de nitrógeno dados a continuación para ajustarlos por mínimos cuadrados

Compuesto	NO	N_2O	NO_2	N_2O_3	N_2O_5	N_2O_4
Peso molecular	30.006	44.013	46.006	76.012	108.010	92.011

.

26. En los procesos termodinámicos adiabáticos de los gases, la presión P y el volumen V siguen una ley del tipo $PV^\gamma = C$, donde C es constante a lo largo del proceso. Ajustar por mínimos cuadrados los valores de C y de γ en un proceso adiabático según la tabla de medidas experimentales siguiente:

P (atm)	1.62	1.00	0.75	0.62	0.52	0.46
V (litros)	0.5	1.0	1.5	2.0	2.5	3.0

.

27. Se supone que el cometa Tentax, descubierto en el año 1968, es un objeto del Sistema Solar. En cierto sistema de coordenadas polares (r, φ), centrado en el Sol, se han medido experimentalmente las siguientes posiciones del cometa:

r	2.70	2.00	1.61	1.20	1.02
φ	48°	67°	83°	108°	126°

.

Las leyes de Kepler garantizan que el cometa se moverá en una órbita elíptica, parabólica o hiperbólica (si se desprecian las perturbaciones de los planetas), que, en dichas coordenadas polares, tendrá por ecuación

$$r = \frac{p}{1 + e\cos\varphi} \ ,$$

donde p es un parámetro y e la excentricidad. Ajustar por mínimos cuadrados los valores de p y e, a partir de las medidas hechas.

28. Consideremos un vector $y = (y_1 \ \cdots \ y_m)^T$ y una matriz M de dimensión mxn con $m \geq n \geq 2$, y rango n. Denotamos las columnas de M por $m^{(j)}$ $(j = 1 \div n)$. Para resolver el sistema $Ma = y$, se considera el método iterativo

$$a^{(k+1)} = \frac{1}{m}\sum_{j=1}^{m} b^{(k,j)} \ ,$$

donde $b^{(k,j)}$ es la proyección del vector $a^{(k)}$ sobre el hiperplano de ecuación $m^{(j)T}a = y_j$ $(j = 1 \div n)$.

a) Demostrar que los valores propios de la matriz M^TM pertenecen al intervalo $(0, m)$.

b) Escribir el método iterativo anterior en la forma $a^{(k+1)} = Ba^{(k)} + c$ y demostrar que converge a la solución del sistema de ecuaciones normales correspondiente al sistema $Ma = y$, partiendo de cualquier aproximación $a^{(0)}$ inicial.

c) Si $a^{(0)}$ no es ya solución del sistema de ecuaciones normales, ¿se puede dar la convergencia del método en un número finito de iteraciones?

29. Determinar las rectas que aproximan la curva $y(x) = \text{sen}(\pi x)$ haciendo que:

i) la norma euclídea discreta del error en el conjunto de abscisas

$$\{-0.5, -0.25, 0, 0.25, 0.5\}$$

sea mínima;

ii) la norma euclídea del error en el intervalo $[-0.5, 0.5]$ sea mínima.

30. Dada una tabla de la función f en abscisas equidistantes $x_k = x_0 + kh$ $(k = -2 \div 2)$, queremos aproximar la derivada de la función f en la abscisa x_0 por la derivada en esta abscisa de la función f^* de la forma

$$f^*(x) = \sum_{j=0}^{4} a_j^*(x - x_0)^j$$

tal que

$$\sum_{k=-2}^{2} (f(x_k) - f^*(x_k))^2$$

sea mínima entre todas las funciones de aquel tipo.

Dar la expresión de esta derivada aproximada en función de los valores $f(x_k)$ $(k = -2 \div 2)$ y de h.

31. Sea w una función peso en el intervalo $[a, b]$. Su *momento de orden* j se define por

$$\mu_j = \int_a^b w(x)x^j dx \ .$$

Demostrar que el sistema de ecuaciones

$$-\begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_{n-1} \\ \mu_1 & \mu_2 & \cdots & \mu_n \\ \vdots & \vdots & \cdots & \vdots \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-2} \end{pmatrix} \begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ d_{n-1} \end{pmatrix} = \begin{pmatrix} \mu_n \\ \mu_{n+1} \\ \vdots \\ \mu_{2n-1} \end{pmatrix}$$

tiene como solución los coeficientes del polinomio mónico de grado n

$$\varphi_n(x) = x^n + \sum_{j=0}^{n-1} d_j x^j$$

de la familia de polinomios ortogonales asociada al peso w en $[a,b]$. La matriz del sistema anterior recibe el nombre de *matriz de Hankel*.

32. En el espacio vectorial de los polinomios a coeficientes reales se define el producto escalar

$$< p, q >= \sum_{j\geq 0} a_j b_j \ ,$$

siendo

$$p(x) = \sum_{j\geq 0} a_j x^j \ , \quad q(x) = \sum_{j\geq 0} b_j x^j \ .$$

¿Existe algún intervalo $[a,b]$ y algún peso w en dicho intervalo tales que

$$< p, q >= \int_a^b p(x)q(x)w(x)dx \ ?$$

33. Consideramos la norma euclídea asociada al peso w en $[a,b]$. Demostrar que, entre todos los polinomios mónicos de grado n, el polinomio mónico de grado n de la familia de polinomios ortogonales respecto al producto escalar asociado es el que tiene norma mínima.

34. Los polinomios de Hermite quedan definidos por la recurrencia:

$$H_0(x) = 1 \ , H_1(x) = x \ , \quad H_{j+1}(x) = 2xH_j(x) - 2jH_{j-1}(x) \ .$$

a) Calcular los diez primeros.

b) Escribir los polinomios 1, x, ..., x^9 en función de ellos.

c) Demostrar que los polinomios de Hermite pueden definirse también según la fórmula

$$H_j(x) = (-1)^j \exp(-x^2)\frac{d^j}{dx^j}(\exp(-x^2)) \ , (j \geq 0) \ .$$

d) Ver que son ortogonales respecto al producto escalar correspondiente al peso $w(x) = \exp(-x^2)$ en $\mathbb{R}$.

35. a) Demostrar que los polinomios de Laguerre, definidos por

$$L_j(x) = e^x \frac{d^j}{dx^j}(x^j e^{-x}) \ ,$$

son ortogonales respecto al producto escalar de peso $w(x) = e^{-x}$ en el intervalo $[0,\infty)$.

b) Calcular los 6 primeros.

c) Demostrar la relación de recurrencia:

$$L_0(x) = 1 \ , \quad L_1(x) = 1 - x \ ,$$

$$L_{j+1}(x) = (2j + 1 - x)L_j(x) - j^2 L_{j-1}(x) \ (j \geq 1) \ .$$

36. a) Determinar los polinomios ortogonales $\psi_j(x)$ $(j = 0 \div 4)$ correspondientes a la función peso $x^2 + x^4$ en el intervalo $[-1, 1]$.

b) Dada $f(x) = \text{sen}(\pi x)$, hallar el polinomio $p_4^*(x)$ que hace mínima

$$\int_{-1}^{1} (x^2 + x^4)(f(x) - p_4(x))^2 dx$$

entre los polinomios $p_4(x)$ de grado menor o igual que 4.

37. a) Sea $f(x) = \cos x$, obtener la combinación de polinomios de Legendre $P_j(x)$ $(j \geq 0)$ de la forma

$$p_4(x) = \sum_{j=0}^{4} c_j P_j(x)$$

que hace que sea mínimo el error cuadrático

$$\int_{-1}^{1} (f(x) - p_4(x))^2 dx \ .$$

b) ¿Cuánto vale este mínimo?

38. Hallar el polinomio de primer grado obtenido por interpolación de Chebichev de la función $f(x) = \frac{1}{1+3x}$ en el intervalo $(-1, 1)$. ¿Cuál es la norma del máximo del error en dicho intervalo?

39. Hallar el polinomio de interpolación de Chebichev de grado 4 a la función $f(x) = \ln(1 + x)$ en el intervalo $[2, 4]$,

i) usando el método de las diferencias divididas de Newton,

ii) hallando los coeficientes ortogonales de Chebichev.

¿Cuál de los dos métodos parece más eficiente?

40. Dada la función $f(\theta) = \theta(\pi - \theta)$ en $[0, \pi]$, determinar el polinomio en cosenos

$$f_n^*(\theta) = \sum_{j=0}^{n} c_j^* \cos(j\theta)$$

que minimiza la norma euclídea del error

$$\|f - f_n\|_2 = \left(\int_0^\pi (f(\theta) - f_n(\theta))^2 d\theta \right)^{\frac{1}{2}} .$$

41. a) Sea la función f definida en el intervalo $[-\pi, \pi]$ por

$$f(\theta) = \begin{cases} -1 & (-\pi < \theta < 0) \\ 1 & (0 < \theta < \pi) \\ 0 & \theta = -\pi, 0, \pi \end{cases} .$$

Calcular los coeficientes a_0, a_j, b_j $(j = 1 \div n)$ de la función

$$t^*_{2n-1}(\theta) = a_0 + \sum_{j=1}^{2n-1} a_j \cos(j\theta) + \sum_{j=1}^{2n-1} b_j \operatorname{sen}(j\theta)$$

que minimizan la norma euclídea del error.

b) Demostrar, por inducción, que

$$t^*_{2n-1}(\theta) = \frac{2}{\pi} \int_0^\theta \frac{\operatorname{sen} 2n\varphi}{\operatorname{sen} \varphi} d\varphi$$

y deducir que t^*_{2n-1} tiene el primer máximo local positivo en $\frac{\pi}{2n}$.

c) Ver también que

$$\lim_{n\to\infty} t^*_{2n-1}\left(\frac{\pi}{2n}\right) = \frac{2}{\pi} \int_0^\pi \frac{\operatorname{sen} \varphi}{\varphi} d\varphi .$$

42. Hallar el polinomio de aproximación minimax de grado 4 a $f(x) = x^5 - 3x^3$ en el intervalo $[-3, 3]$ y acotar el error cometido en todo el intervalo.

43. Hallar la aproximación minimax polinomial de grado 1 a $f(x) = e^x$ en el intervalo $[0, 1]$ y acotar el error cometido en todo el intervalo.

44. a) Aproximar la función $f(x) = \frac{1}{x+2}$ en el intervalo $[-1, 1]$ por un polinomio de grado 1 de las maneras siguientes:

i) por interpolación en los extremos,

ii) por interpolación de Chebichev,

iii) minimizando la norma del máximo del error.

b) Dar cotas para los errores cometidos en todos los casos y comparar los resultados hallados, justificándolos teóricamente.

45. Hallar la aproximación minimax de la forma ax^2 a la función $f(x) = x^4$ en el intervalo $[0, 1]$.

46. Utilizar el desarrollo de Taylor de

$$f(x) = \int_0^x \frac{\operatorname{sen} t}{t} dt$$

hasta grado 9 para hallar un polinomio de grado mínimo que aproxime a la función dada sobre el intervalo $[-\frac{\pi}{2}, \frac{\pi}{2}]$ con un error menor que 10^{-4}, utilizando economización de Lanczos.

CAPÍTULO 4

DERIVACIÓN, INTEGRACIÓN Y SUMACIÓN

La derivación, integración y sumación de funciones son las operaciones esenciales del Análisis Matemático. Su aproximación numérica constituye un útil de gran importancia en la realización de dichas operaciones, cuando no puedan llevarse a cabo de manera exacta, analíticamente. Los métodos presentados en este capítulo estan basados en técnicas de aproximación de funciones, fundamentalmente en las de interpolación.

4.1 DERIVACIÓN NUMÉRICA

4.1.1 Introducción

Aunque haya reglas bien conocidas para derivar las funciones más usuales, no siempre pueden ser utilizadas (por ejemplo, en funciones dadas por tablas de valores o numéricamente), o no es conveniente hacerlo (por ejemplo, en funciones con expresiones analíticas demasiado complicadas). En estos casos deberemos recurrir a técnicas numéricas que, partiendo de los valores de la función en diversas abscisas, nos permitirán calcular una aproximación al valor de alguna de sus derivadas en una abscisa próxima.

4.1.2 Derivadas primeras

Fórmulas de derivación interpolatoria y errores

La *derivación numérica* de una función f diferenciable en $a \in \mathbb{R}$ consta de dos etapas:

- Construcción del polinomio interpolador $p_m(x)$ a la función f en una familia de abscisas x_0, x_1, ..., x_m (que convendrá tomar próximas a a).
- Derivación del polinomio $p_m(x)$ y evaluación en a, según la fórmula de derivación numérica:

$$f'(a) \simeq p'_m(a) \ .$$

Las fórmulas así obtenidas reciben el nombre de *fórmulas de derivación interpolatorias* (véase el problema 4.1).

Si denotamos por $e_m(x)$ la función de error de interpolación $f(x)-p_m(x)$ de la función f por el polinomio $p_m(x)$ y suponemos que f es $m+1$ veces diferenciable con continuidad en un intervalo I que contiene las abscisas x_k $(k=0\div m)$, entonces, para $x\in I$, disponemos de la siguiente expresión para el error de interpolación:

$$e_m(x)=F_{m+1}(x)\omega_m(x)\ ,$$

con

$$F_{m+1}(x)=\frac{f^{(m+1)}(\xi(x))}{(m+1)!}=f[x_0,\ldots,x_m,x]\ ,$$

donde $\omega_m(x)=(x-x_0)\cdots(x-x_m)$ y $\xi(x)\in<x_0,\ldots,x_m,x>$ (véanse las fórmulas (3.2) y (3.6)). Resulta así que el error en la derivación numérica viene dado por

$$e'_m(a)=f'(a)-p'_m(a)=F'_{m+1}(a)\omega_m(a)+F_{m+1}(a)\omega'_m(a)\ ;$$

esta fórmula presenta el problema del cálculo de $F'_{m+1}(a)$, si a es arbitrario, debido a que la función F_{m+1} no es conocida en general (véase un ejemplo en el problema 4.2).

Ahora bien, si a pertenece al conjunto de abscisas x_0, ..., x_m (es decir, si $a=x_k$ para algún $k=0\div m$), entonces se tiene

$$e'_m(x_k)=\frac{f^{(m+1)}(\xi_k)}{(m+1)!}\prod_{i\neq k}(x_k-x_i)\ ,$$

donde $\xi_k=\xi(x_k)\in<x_0,\ldots,x_m>$.

Ejemplos

Si $m=1$, $x_0=a$, $x_1=a+h$ y $f\in\mathcal{C}^2([a,a+h])$:

$$f'(a)=\frac{f(a+h)-f(a)}{h}+\frac{f^{(2)}(\xi)}{2}h\ ,\quad \xi\in<a,a+h>\ .$$

Si $m=2$, $x_0=a-h$, $x_1=a$, $x_2=a+h$ y $f\in\mathcal{C}^3([a-h,a+h])$:

$$f'(a)=\frac{f(a+h)-f(a-h)}{2h}-\frac{f^{(3)}(\xi)}{6}h^2\ ,\quad \xi\in<a-h,a+h>\ .$$

4.1.3 Derivadas de orden superior

Fórmulas de derivación interpolatoria

El procedimiento anterior puede repetirse para calcular $f^{(2)}(a)$, ..., $f^{(d)}(a)$, siempre que $d\leq m$. Así, si conocemos la función f en las abscisas x_0, ..., x_m próximas a a, derivando d veces el polinomio interpolador $p_m(x)$ y evaluando en a, tenemos

$$f^{(d)}(a)\simeq d!f[x_0,x_1,\ldots,x_d]\ .$$

Para obtener expresiones del error es conveniente usar los errores de interpolación de Taylor para la fórmula hallada.

Ejemplo

Si $m = 2$, $x_0 = a - h$, $x_1 = a$, $x_2 = a + h$,

$$f^{(2)}(a) \simeq 2f[x_0, x_1, x_2] = \frac{f(a+h) - 2f(a) + f(a-h)}{h^2} \ ;$$

si suponemos $f \in \mathcal{C}^4([a-h, a+h])$, obtenemos una expresión del error usando

$$f(a+h) = f(a) + f'(a)h + \frac{f^{(2)}(a)}{2}h^2 + \frac{f^{(3)}(a)}{3!}h^3 + \frac{f^{(4)}(\xi_1)}{4!}h^4 \ ,$$

$$f(a-h) = f(a) - f'(a)h + \frac{f^{(2)}(a)}{2}h^2 - \frac{f^{(3)}(a)}{3!}h^3 + \frac{f^{(4)}(\xi_2)}{4!}h^4 \ ,$$

donde $\xi_1 \in < a, a+h >$ y $\xi_2 \in < a-h, h >$.

Sumandolas se obtiene

$$f^{(2)}(a) = \frac{f(a+h) - 2f(a) + f(a-h)}{h^2} + R(h) \ ,$$

con

$$R(h) = -\frac{f^{(4)}(\xi_1) + f^{(4)}(\xi_2)}{24}h^2 \ ,$$

y así, $|R(h)| \leq \frac{M}{12}h^2$ si $|f^{(4)}(x)| \leq M$ para $x \in < a-h, a+h >$; es decir,

$$R(h) = \mathcal{O}(h^2) \ (h \to 0) \ . \tag{4.1}$$

Ahora consideraremos el resultado general siguiente que será utilizado a menudo:

- Si F es una función real y continua sobre un intervalo I, dadas las abscisas de I, $\xi_1, \xi_2, \ldots, \xi_m$ y los coeficientes positivos $\alpha_1, \alpha_2, \ldots, \alpha_m$, existe una abscisa $\xi \in < \xi_1, \xi_2, \ldots, \xi_m >$ tal que
$$\sum_{k=1}^{m} \alpha_k F(\xi_k) = F(\xi) \sum_{k=1}^{m} \alpha_k$$
(*teorema del valor medio para sumas).*

Aplicando esta propiedad general a la función $f^{(4)}$, se obtiene la siguiente expresión para el error

$$R(h) = -\frac{f^{(4)}(\xi)}{12}h^2 \ , \quad \xi \in < a-h, a+h > \ . \tag{4.2}$$

No siempre podremos obtener una expresión explícita para $R(h)$ como en (4.2) y, a menudo, tendremos que conformarnos con una expresión asintótica como en (4.1).

Suponiendo $f \in \mathcal{C}^6([a-h, a+h])$, podríamos obtener

$$R(h) = -\frac{f^{(4)}(a)}{12}h^2 + \mathcal{O}(h^4) \ (h \to 0) \ ,$$

$$R(h) = -\frac{f^{(4)}(a)}{12}h^2 - \frac{2f^{(6)}(\xi)}{6!}h^4 \ , \quad \xi \in < a-h, a+h > \ ,$$

y también desarrollos análogos, si f fuese todavía más derivable.

Observaciones

1. El error de las fórmulas de derivación numérica halladas depende de términos en h, h^2, etc., donde $h = \max |a - x_k|$; por lo tanto, decrece al decrecer h. Esto no quiere decir que convenga escoger h muy pequeño, ya que en las fórmulas mencionadas se produce entonces una notable cancelación de términos (véase el problema 4.15). En esta situación convendrá tomar h prudencialmente pequeño y emplear el método de extrapolación repetida de Richardson, tal como veremos en el apartado 4.4.2.

2. La idea fundamental usada aquí es que si $p(x)$ es cercana a $f(x)$ en a, entonces $p^{(d)}(x)$ será próxima a $f^{(d)}(x)$ en a. Desgraciadamente con mucha frecuencia esto no es cierto, ya que dos funciones pueden ser próximas y tener pendientes muy diferentes (véase la figura 4.1).

3. En el apartado 4.5.3, mediante el uso de operadores, se expondrá una manera sistemática de generación de fórmulas generales de derivación numérica.

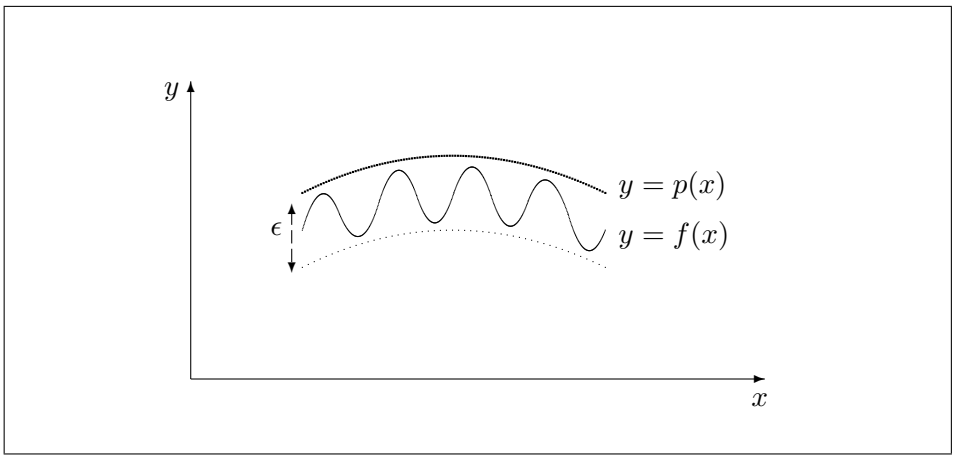


Figura 4.1: Funciones próximas con derivadas alejadas.

4.2 INTEGRACIÓN NUMÉRICA

4.2.1 Introducción

Dada una función f definida sobre un intervalo acotado $[a, b]$, queremos calcular

$$J(f) = \int_a^b f(x)dx \ ,$$

suponiendo que esta integral tenga sentido para la función f (por ejemplo, si f es continua sobre $[a, b]$). La *cuadratura* o *integración numérica* consiste en dar fórmulas aproximadas para el cálculo de la integral $J(f)$ de f; estas fórmulas pueden ser de gran utilidad cuando la integral no pueda calcularse por métodos analíticos (por ejemplo, usando la regla de Barrow), o bien, cuando no conviene usarlos porque resultan complicados, y nos conformamos con conocer $J(f)$ con una precisión finita dada.

Para ello, aproximaremos f por un polinomio interpolador $p(x)$ y calcularemos exactamente $J(p)$, obteniendo así fórmulas de integración numérica.

4.2.2 Integración con abscisas dadas

Fórmulas de integración interpolatoria

Sean $a \leq x_0 < x_1 < \cdots < x_m \leq b$ una partición en $m+1$ abscisas del segmento $[a,b]$ y consideremos el polinomio $p_m(x)$ de grado menor o igual que m verificando

$$p_m(x_k) = f(x_k) \ (k = 0 \div m) \ ;$$

entonces, aproximamos $J(f)$ por $J(p_m)$:

$$\int_a^b f(x)dx \simeq \int_a^b p_m(x)dx \ .$$

Así, integrando la fórmula de interpolación de Lagrange

$$p_m(x) = \sum_{k=0}^{m} f_k l_k(x) \ ,$$

donde

$$l_k(x) = \prod_{i \neq k} \frac{x - x_i}{x_k - x_i} \ ,$$

hallamos la fórmula de integración numérica

$$\int_a^b f(x)dx \simeq \sum_{k=0}^{m} W_k f_k \ , \quad W_k = \int_a^b l_k(x)dx \ (k = 0 \div m) \ . \tag{4.3}$$

Debido a que la fórmula anterior se halla por integración de un polinomio interpolador, recibe el nombre de *fórmula de integración interpolatoria de $m+1$ abscisas*.

Los coeficientes W_k $(k = 0 \div m)$, llamados *pesos* de la fórmula de integración, dependen del intervalo $[a,b]$ y de las abscisas $x_0, \ldots, x_m$, pero no de f. Por la unicidad del polinomio interpolador, la fórmula (4.3) es exacta para cualquier polinomio de grado menor o igual que m, ello nos proporciona una manera de calcular los pesos W_k $(k = 0 \div m)$ sin necesidad de integrar $l_k(x)$ $(k = 0 \div m)$: se impone que la fórmula (4.3) sea exacta para los monomios $1, \ x, \ x^2, \ \ldots, \ x^m$ (por ejemplo), y se resuelve el sistema lineal resultante (*método de los coeficientes indeterminados*).

Ejemplo: fórmula de Simpson

Se quieren calcular los pesos de integración W_{-1}, W_0 y W_1 para que la fórmula de integración numérica

$$\int_{-1}^{1} g(t)dt \simeq W_{-1}g_{-1} + W_0 g_0 + W_1 g_1$$

sea exacta para todos los polinomios de grado menor o igual que 2, donde se ha tomado $g_k = g(k)$ $(k = -1, 0, 1)$; después, se quiere evaluar el error cometido al aplicarla a polinomios de grado menor o igual que 4.

Imponiendo la exactitud de la fórmula para $g(t) = 1, \ t, \ t^2$, se obtiene el sistema

$$\left.\begin{array}{rcrcrcl} W_{-1} & + & W_0 & + & W_1 & = & 2 \\ -W_{-1} & & & + & W_1 & = & 0 \\ W_{-1} & & & + & W_1 & = & \frac{2}{3} \end{array}\right\} \ ,$$

de solución $W_1 = W_{-1} = \frac{1}{3}$, $W_0 = \frac{4}{3}$ y, como consecuencia, la fórmula de integración numérica

$$\int_{-1}^{1} g(t)dt \simeq \frac{1}{3}(g_{-1} + 4g_0 + g_1) \ . \tag{4.4}$$

Esta fórmula puede trasladarse a cualquier intervalo $[a, b]$, a través del cambio de variables

$$t = 2\frac{x-a}{b-a} - 1 \quad \text{o, equivalentemente,} \quad x = \frac{b-a}{2}t + \frac{a+b}{2} \ ,$$

dando lugar a la *fórmula de Simpson*

$$\int_a^b f(x)dx \simeq \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \ , \tag{4.5}$$

que también se escribe como

$$\int_{c-h}^{c+h} f(x)dx \simeq \frac{h}{3}\left[f(c-h) + 4f(c) + f(c+h)\right] \ , \tag{4.6}$$

donde $c = \frac{a+b}{2}$ y $h = \frac{b-a}{2}$.

Sobre los polinomios de grado 3, las fórmulas de Simpson (4.5) y (4.6) resultan ser también exactas: si $g(t) = t^3$, ambos miembros de (4.4) son nulos.

Si tomamos $g(t) = t^4$ en (4.4), obtenemos

$$\int_{-1}^{1} g(t)dt - \frac{1}{3}(g_{-1} + 4g_0 + g_1) = \frac{2}{5} - \frac{2}{3} = -\frac{4}{15} = -\frac{24}{90} = -\frac{g^{(4)}(\xi)}{90} \tag{4.7}$$

y, tomando análogamente $f(x) = x^4$ en (4.5),

$$\int_{c-h}^{c+h} f(x)dx - \frac{h}{3}\left(f(c-h) + 4f(c) + f(c+h)\right) = -\frac{f^{(4)}(\xi)}{90}h^5 \ . \tag{4.8}$$

Las abscisas ξ que aparecen en (4.7) y (4.8) son arbitrarias, ya que $g^{(4)}$ y $f^{(4)}$ son funciones constantes.

Error de las fórmulas de integración interpolatoria

Volviendo a la fórmula (4.3), observamos que el error de la aproximación que expresa es la integral del error de interpolación; así pues, si $f \in \mathcal{C}^{m+1}([a, b])$, tomando la expresión del error de interpolación (3.2), se obtiene

$$\begin{aligned} E_m &= \int_a^b f(x)dx - \sum_{k=0}^{m} W_k f_k \\ &= \int_a^b \frac{f^{(m+1)}(\xi(x))}{(m+1)!}(x-x_0)\cdots(x-x_m)dx \ , \end{aligned}$$

con $\xi(x) \in (a, b)$.

Esta fórmula da la expresión general para el *error de la fórmula de integración interpolatoria de* $m+1$ *abscisas.*

Si $|f^{(m+1)}(x)| \leq M_{m+1}$ para todo $x \in [a,b]$, podremos escribir

$$|E_m| \leq \frac{M_{m+1}}{(m+1)!} \int_a^b |(x-x_0)\cdots(x-x_m)|\, dx\ .$$

Una manera de obtener cotas más precisas consiste en desarrollar por Taylor las fórmulas de cuadratura dadas, tal como se hace en el apartado que sigue.

Error de la fórmula de Simpson

Continuamos con la fórmula de Simpson y definimos

$$E_S(h) = \int_{c-h}^{c+h} f(x)dx - \frac{h}{3}[f(c-h) + 4f(c) + f(c+h)] \ ;$$

entonces se cumple que $E_S(0) = E_S'(0) = E_S^{(2)}(0) = 0$ y

$$E_S^{(3)}(h) = -\frac{h}{3}[f^{(3)}(c+h) - f^{(3)}(c-h)] \ .$$

Si $f \in \mathcal{C}^4([c-h, c+h])$, tomamos

$$F(h) = \begin{cases} \frac{f^{(3)}(c+h) - f^{(3)}(c-h)}{2h} & (\text{si} \quad h \neq 0) \ , \\ f^{(4)}(c) & (\text{si} \quad h = 0) \ . \end{cases}$$

Resulta que F es continua, ya que $F(h) \to F(0)$ (cuando $h \to 0$); además, por el teorema del valor medio, para cualquier $\xi \in (0, h]$, $F(\xi) = f^{(4)}(\eta)$, para algún $\eta \in (c-h, c+h)$.

Usando la expresión integral del error de interpolación de Taylor (3.8), se tiene

$$E_S(h) = \frac{1}{2}\int_0^h (h-t)^2 E_S^{(3)}(t)dt = -\frac{1}{3}\int_0^h (h-t)^2 t^2 F(t)dt \ .$$

Dado que $G(t) = (h-t)^2t^2$ es una función continua que no cambia de signo en $(0, h)$, el teorema del valor medio para integrales nos permite extraer $F(t)$ fuera del signo integral y así,

$$E_S(h) = -\frac{1}{3}F(\xi)\int_0^h (h-t)^2 t^2 dt = -\frac{f^{(4)}(\eta)}{90}h^5 \ , \quad \eta \in (c-h, c+h) \ ,$$

que justamente coincide con la expresión (4.8).

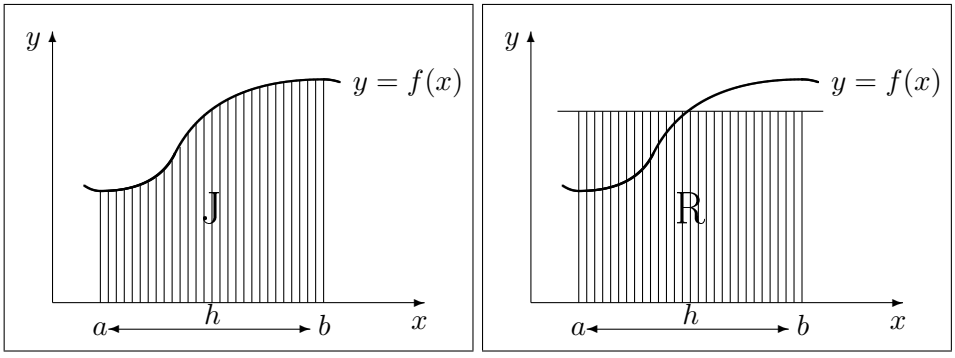


Figura 4.2: Aproximación de la integral J por la fórmula del rectángulo R

Fórmulas del rectángulo y del trapecio

Tomando $h = b-a$ y haciendo un proceso análogo al explicado para la fórmula de Simpson, obtenemos la *fórmula del rectángulo*

$$\int_a^b f(x)dx = hf\left(\frac{a+b}{2}\right) + \frac{f^{(2)}(\xi)}{24}h^3 \ , \quad \xi \in (a,b) \ ,$$

y la *fórmula del trapecio*

$$\int_a^b f(x)dx = \frac{h}{2}[f(a)+f(b)] - \frac{f^{(2)}(\xi)}{12}h^3 \ , \quad \xi \in (a,b) \ , \tag{4.9}$$

usando interpolación únicamente en la abscisa media y en las abscisas extremas, respectivamente.

Para una representación gráfica, véanse las figuras 4.2 y 4.2.2.

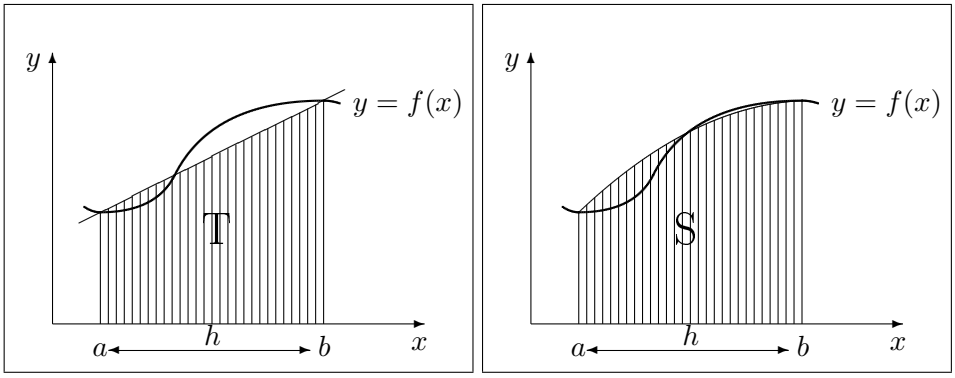


Figura 4.3: Aproximaciones por las fórmulas del trapecio T y de Simpson S

Fórmulas de Newton-Cotes

Las fórmulas del trapecio y de Simpson son un caso particular de dichas *fórmulas (cerradas) de Newton-Cotes* que se comentan seguidamente.

Si, en la fórmula (4.3), tomamos $m+1$ abscisas equidistantes sobre el intervalo $[a,b]$

$$x_k = a + kh \ (k = 0 \div m) \ , \quad h = \frac{b-a}{m}$$

obtenemos la *fórmula de Newton-Cotes de $m+1$ abscisas*

$$\int_a^b f(x)dx \simeq h\sum_{k=0}^{m} \alpha_k f_k \ , \quad \alpha_k = \int_0^m \prod_{i\neq k} \frac{t-i}{k-i}dt \ (k = 0 \div m) \ ,$$

donde se ha usado la notación $f_k = f(a+kh)$ $(k = 0 \div m)$.

Los coeficientes α_k $(k = 0 \div m)$ no dependen ni del intervalo $[a,b]$ ni de la función f; sólo dependen del grado m. La *expresión del error de la fórmula de Newton-Cotes* de $m+1$ abscisas viene dada por

$$E_m = \int_a^b f(x)dx - h\sum_{k=0}^{m} \alpha_k f_k = K_m \frac{f^{(p+1)}(\xi)}{(p+1)!}h^{p+2} \ , \quad \xi \in (a,b) \ , \tag{4.10}$$

donde

$$K_m = \int_0^m \pi_m(t)dt \ , \quad p = m \ \ (\text{ si } m \text{ es impar }) \ ,$$
$$K_m = \int_0^m t\pi_m(t)dt \ , \quad p = m+1 \ \ (\text{ si } m \text{ es par }) \ ;$$

siendo $\pi_m(t) = t(t-1)\cdots(t-m)$.

Así pues, las fórmulas de Newton-Cotes de $m+1$ abscisas son exactas para todos los polinomios de grado $m+1$, cuando m es par (o equivalentemente, cuando el número de abscisas es impar) tal como sucedía con la fórmula de Simpson ($m = 2$) en el ejemplo anterior.

Nótese que K_m puede obtenerse también aplicando (4.10) a un polinomio de grado $m+1$ (resp. $m+2$) si m es impar (resp. par), ya que $f^{(p+1)}$ es constante en tales casos.

Otras fórmulas de integración interpolatorias

Las fórmulas interpolatorias que acaban de presentarse utilizan únicamente el polinomio interpolador a la función que se integra. Otras fórmulas de integración numérica pueden deducirse usando tipos diferentes de interpolación; esto es, interpolación de Taylor, interpolación de Hermite e, incluso, interpolación de Hermite generalizada. Algunos ejemplos de estas fórmulas y sus aplicaciones pueden encontrarse en los problemas 4.4 y 4.5.

4.2.3 Reglas (compuestas) de integración numérica

Las fórmulas de integración numérica no se aplican normalmente sobre todo el intervalo $[a,b]$, sino sobre subintervalos de $[a,b]$, dando lugar así a las *reglas (compuestas) de integración numérica.*

Regla de los trapecios

Si dividimos $[a,b]$ en M partes iguales y, en cada una de ellas, aplicamos la fórmula del trapecio, obtenemos la *regla de los trapecios*

$$T(h) = \frac{h}{2}[f(a) + 2f(a+h) + 2f(a+2h) + \cdots + 2f(b-h) + f(b)] \ , \tag{4.11}$$

que aparece al descomponer la integral inicial como la suma de las integrales en las M partes de longitud $h = \frac{b-a}{M}$ en las que se ha dividido el intervalo $[a,b]$:

$$J(f) = \int_a^b f(x)dx = \sum_{k=0}^{M-1} \int_{x_k}^{x_{k+1}} f(x)dx = \sum_{k=0}^{M-1} J_k(f) \ ,$$

con $x_k = a + kh$ $(k = 0 \div M)$.

Dado que

$$J_k(f) = \int_{x_k}^{x_{k+1}} f(x)dx = \frac{h}{2}[f(x_k) + f(x_{k+1})] - \frac{f^{(2)}(\xi_k)}{12}h^3 \ ,$$

con $\xi_k \in (x_k, x_{k+1})$, entonces

$$J(f) - T(h) = -\frac{1}{12}\sum_{k=0}^{M-1} f^{(2)}(\xi_k)h^3 = -\frac{b-a}{12M}\sum_{k=0}^{M-1} f^{(2)}(\xi_k)h^2 \ ,$$

donde suponemos que $f \in \mathcal{C}^2([a,b])$; finalmente, teniendo en cuenta el teorema del valor medio para sumas, existe $\xi \in (a,b)$ tal que

$$\int_a^b f(x)dx - T(h) = -\frac{b-a}{12} f^{(2)}(\xi)h^2 \ . \tag{4.12}$$

Regla de Simpson

Dividiendo ahora $[a,b]$ en $2M$ partes iguales y aplicando, en cada intervalo de longitud $\frac{b-a}{M} = 2h$, la fórmula de Simpson, obtenemos la *regla de Simpson*,

$$\begin{aligned} S(h) \quad = \quad & \frac{h}{3}[f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) + \cdots \\ & +2f(b-2h) + 4f(b-h) + f(b)] \ . \end{aligned}$$

Suponiendo $f \in \mathcal{C}^4([a,b])$, se dispone de la siguiente expresión para el error asociado:

$$\int_a^b f(x)dx - S(h) = -\frac{b-a}{180} f^{(4)}(\xi)h^4 \ , \tag{4.13}$$

donde $\xi \in (a,b)$.

Fórmula de Euler-Maclaurin

Una manera bastante útil de generar fórmulas de integración numérica es mediante la integración por partes. Así, por ejemplo, escribimos la integral

$$J(f) = \int_a^b f(x)dx$$

como

$$\int_a^b f(x)b_1'(x)dx \ ,$$

donde $b_1(x)$ es un polinomio de grado 1 tal que $b_1'(x) = 1$ $(b_1(x) = x + c_1)$; entonces

$$J(f) = f(x)b_1(x)\Big|_a^b - \int_a^b f'(x)b_1(x)dx \ .$$

Si ahora introducimos $b_2(x)$ tal que $b_2'(x) = b_1(x)$ $(b_2(x) = \frac{(x+c_1)^2}{2} + c_2)$, resulta

$$J(f) = f(x)b_1(x)\Big|_a^b - f'(x)b_2(x)\Big|_a^b + \int_a^b f^{(2)}(x)b_2(x)dx \ . \tag{4.14}$$

Si tomamos $c_1 = -\frac{a+b}{2}$, $c_2 = -\frac{(b-a)^2}{8}$, entonces $b_2(x) = \frac{1}{2}(x-a)(x-b)$ se anula sobre a y b y obtenemos

$$J(f) = \frac{b-a}{2}[f(a)+f(b)] + \frac{1}{2}\int_a^b f^{(2)}(x)(x-a)(x-b)dx \ .$$

Dado que $(x-a)(x-b) < 0$ sobre (a,b), si $f \in \mathcal{C}^2([a,b])$, aparece la ya conocida fórmula del trapecio con su error correspondiente (4.12):

$$\begin{aligned} J(f) - \frac{b-a}{2}[f(a)+f(b)] &= \frac{1}{2}\int_a^b f^{(2)}(x)(x-a)(x-b)dx \\ &= -\frac{f^{(2)}(\xi)}{12}(b-a)^3 \ . \end{aligned}$$

El procedimiento de (4.14) puede iterarse para obtener una expresión asintótica del error de la regla de los trapecios. Para simplificar, podemos considerar que el intervalo $[a,b]$ sea el intervalo $[0,1]$; más concretamente, haciendo el cambio $x = a + th$ e introduciendo $g(t) = f(a+th)$, tenemos

$$\int_a^b f(x)dx = h\int_0^1 g(t)dt \ ,$$

con $h = b-a$.

Tomamos $b_1(t) = t - \frac{1}{2}$ como antes (ahora la abscisa media es $\frac{1}{2}$) y definimos $b_2(t)$, $b_3(t)$, ..., de manera que

$$b'_{j+1}(t) = b_j(t) \ (j \geq 1) \ , \tag{4.15}$$

y con las constantes adecuadas para que

$$\int_0^1 b_j(t)dt = 0 \ (j \geq 1) \tag{4.16}$$

o, equivalentemente, $b_{j+1}(0) = b_{j+1}(1)$ $(j \geq 1)$.

Repetidas integraciones por partes dan lugar a

$$\begin{aligned} \int_0^1 g(t)dt &= \frac{1}{2}(g(0)+g(1)) \\ &\quad -b_2(0)(g'(1)-g'(0)) + b_3(0)(g^{(2)}(1)-g^{(2)}(0)) + \ldots \\ &\quad +(-1)^{(n+1)}b_n(0)(g^{(n-1)}(1)-g^{(n-1)}(0)) \\ &\quad +(-1)^{(n+2)}\int_0^1 g^{(n)}(t)b_n(t)dt \ , \end{aligned} \tag{4.17}$$

para $g \in \mathcal{C}^n([0,1])$.

En el problema 3.11, se demuestra que

$$(-1)^{r+1}b_{2r}(0) > 0 \ , \quad b_{2r+1}(0) = 0 \ \ (r \geq 1) \ , \tag{4.18}$$

$$(-1)^r(b_{2r}(t) - b_{2r}(0)) > 0 \ \ \forall t \in (0,1) \ (r \geq 1) \ . \tag{4.19}$$

Tomando ahora $n = 2s+1$ e integrando por partes el último sumando de (4.17) de signo negativo, se obtiene

$$\begin{aligned}\int_0^1 g^{(2s+1)}(t)b_{2s+1}(t)dt &= [b_{2s+2}(t) - b_{2s+2}(0)]g^{(2s+1)}(t)\Big|_0^1 \\ &\quad - \int_0^1 (b_{2s+2}(t) - b_{2s+2}(0))g^{(2s+2)}(t)dt \\ &= b_{2s+2}(0)g^{(2s+2)}(\xi)\ , \quad \xi \in (0,1)\ ,\end{aligned}$$

donde se ha usado (4.15), (4.16) y (4.19).

A continuación, ponemos la fórmula (4.17) en términos de los polinomios *mónicos* $B_0(t) = 1$, $B_j(t) = j!b_j(t) = t^j + \cdots$ $(j \geq 1)$ (es decir, polinomios con coeficientes principales unitarios). Éstos reciben el nombre de *polinomios de Bernoulli* y los valores $B_j \equiv B_j(0)$ $(j \geq 0)$ se denominan *números de Bernoulli* (véase de nuevo el problema 3.11 para generarlos):

$$\begin{aligned}\int_0^1 g(t)dt &= \frac{1}{2}(g(0)+g(1)) \qquad (4.20)\\ &\quad - \sum_{r=1}^{s} \frac{B_{2r}}{(2r)!}\left[g^{(2r-1)}(1) - g^{(2r-1)}(0)\right] \\ &\quad - \frac{B_{2s+2}}{(2s+2)!}g^{(2s+2)}(\xi)\ , \quad \xi \in (0,1)\ , \qquad (4.21)\end{aligned}$$

válida para toda $g \in \mathcal{C}^{2s+2}([0,1])$.

Deshaciendo el cambio $g(t) = f(a+th)$ y reordenando los términos, obtenemos la *expresión asintótica del error de la fórmula del trapecio*

$$\begin{aligned}\frac{h}{2}[f(a)+f(a+h)] &= \int_a^{a+h} f(x)dx \qquad (4.22)\\ &\quad + \sum_{r=1}^{s} h^{2r}\frac{B_{2r}}{(2r)!}\left[f^{(2r-1)}(a+h) - f^{(2r-1)}(a)\right] \\ &\quad + \frac{B_{2s+2}}{(2s+2)!}f^{(2s+2)}(\xi)h^{2s+3}\ , \quad \xi \in (a, a+h)\ ,\end{aligned}$$

válida para $f \in \mathcal{C}^{2s+2}([a, a+h])$.

Haciendo una subdivisión del intervalo $[a,b]$ en M partes iguales, y aplicando en cada segmento $[x_k, x_{k+1}]$ la fórmula del trapecio con $h = \frac{b-a}{M}$ $(k = 0 \div M-1)$, se obtiene, de manera análoga al procedimiento de (4.12), la *expresión asintótica del error de la regla de los trapecios*

$$\begin{aligned}T(h) &= \int_a^b f(x)dx + \sum_{r=1}^{s} h^{2r}\frac{B_{2r}}{(2r)!}\left[f^{(2r-1)}(b) - f^{(2r-1)}(a)\right] \\ &\quad + \frac{B_{2s+2}}{(2s+2)!}(b-a)f^{(2s+2)}(\xi)h^{2s+2}\ , \quad \xi \in (a,b)\ , \qquad (4.23)\end{aligned}$$

válida para $f \in \mathcal{C}^{2s+2}([a,b])$.

Las fórmulas (4.21), (4.22) y (4.23) reciben el nombre de *fórmulas de Euler-Maclaurin*. Al aplicarlas, conviene saber acotar el resto. Por ejemplo, en la (4.23),

$$R_s = \frac{B_{2s+2}}{(2s+2)!}(b-a)f^{(2s+2)}(\xi)h^{2s+2} \ , \ \xi \in (a,b)$$

y, claramente,

$$|\, R_s \,| \leq \frac{|\, B_{2s+2} \,|}{(2s+2)!}(b-a) \sup_{x\in[a,b]} |\, f^{(2s+2)}(x) \,|\, h^{2s+2} \ .$$

Ahora bien, según (4.18) los números de Bernoulli alternan el signo (esto es, $B_{2s}B_{2s+2} < 0$ $(s \geq 0)$). Si $f^{(2s+2)}$ y $f^{(2s+4)}$ no se anulan en (a,b) y tienen el mismo signo, entonces $R_sR_{s+1} < 0$ y, por lo tanto, $|\, R_s \,|$ es menor que el valor absoluto del primer término despreciado, según el criterio de alternancia de los restos expuesto en el apartado 3.1.4,

$$|\, R_s \,| \leq \frac{|\, B_{2s+2} \,|}{(2s+2)!} \,|\, f^{(2s+1)}(b) - f^{(2s+1)}(a) \,|\, h^{2s+2} \ .$$

Es necesario advertir que los números de Bernoulli crecen rápidamente y, por lo tanto, no ha de esperarse que $R_s \to 0$ cuando $s \to \infty$. En la práctica, conviene tomar el número de términos s en (4.23) de manera que $|\, R_s \,|$ sea menor que la tolerancia ϵ pedida, siempre que ello sea posible.

4.2.4 Integración gaussiana

Las fórmulas de integración interpolatoria de $m+1$ abscisas x_0, x_1, ..., x_m, obtenidas integrando el polinomio interpolador en estas abscisas, son exactas para los polinomios de grado menor o igual que m; esto ocurre para cualquier elección que se haga de las abscisas dentro del intervalo de integración. Veremos ahora que una elección adecuada de estas $m+1$ abscisas nos proporcionará fórmulas de integración numérica de $m+1$ abscisas, exactas para polinomios de grado menor o igual que $2m+1$, que recibirán el nombre de *fórmulas gaussianas*.

Ejemplo motivador

La regla de los trapecios nos proporcionará el primer ejemplo de estas fórmulas, al ser aplicada convenientemente sobre polinomios trigonométricos.

Así, sea

$$t_n(\theta) = \frac{a_0}{2} + \sum_{j=1}^{n} a_j \cos j\theta + \sum_{j=1}^{n} b_j \operatorname{sen} j\theta = \sum_{j=-n}^{n} c_j e^{ij\theta}$$

un polinomio trigonométrico de grado menor o igual que n, y calculamos

$$J(t_n) = \int_0^{2\pi} t_n(\theta)d\theta = \pi a_0 = 2\pi c_0 \ ,$$

usando la regla de los trapecios con paso $h = \frac{2\pi}{M}$ para $M > n$.

Debido a la periodicidad de $t_n(\theta)$,

$$t_n(\theta + 2\pi) = t_n(\theta) \ \ \forall\theta \in \mathbb{R} \ ;$$

para cualquier $\phi \in \mathbb{R}$ se cumple

$$J(t_n) = \int_{\phi}^{2\pi+\phi} t_n(\varphi)d\varphi = \int_0^{2\pi} t_n(\phi+\theta)d\theta = \pi a_0 \ .$$

La aplicación de la regla de los trapecios a esta última integral da el resultado exacto

$$\begin{aligned} T(\frac{2\pi}{M}) &= \frac{2\pi}{M}\sum_{k=0}^{M-1} t_n(\phi+\frac{2\pi k}{M}) \\ &= \frac{2\pi}{M}\sum_{j=-n}^{n} c_j e^{ij\phi}\sum_{k=0}^{M-1} e^{i\frac{2\pi jk}{M}} = 2\pi c_0 = J(t_n) \ , \end{aligned}$$

donde hemos usado que

$$\sum_{k=0}^{M-1} e^{i\frac{2\pi jk}{M}} = \begin{cases} M & (j=0) \ , \\ 0 & (0<|j|\leq n<M) \ , \end{cases}$$

sabiendo que $e^{i\frac{2\pi j}{M}} \neq 1$ $(0<|j|\leq n<M)$.

Si ahora consideramos un polinomio trigonométrico en cosenos de grado menor o igual que n

$$t_n(\theta) = \frac{a_0}{2} + \sum_{j=1}^{n} a_j \cos j\theta \ ,$$

tenemos, además de la periodicidad $t_n(\theta) = t_n(\theta+2\pi)$, la simetría respecto a $\theta=\pi$: $t_n(\theta) = t_n(2\pi-\theta)$; por lo tanto, si $M > n$,

$$\begin{aligned} J(t_n) &= \int_0^{\pi} t_n(\theta)d\theta = \frac{1}{2}\int_0^{2\pi} t_n(\theta)d\theta = \frac{\pi}{2}a_0 \\ &= \frac{1}{2}T(\frac{2\pi}{M}) = \frac{\pi}{M}\sum_{k=0}^{M-1} t_n(\phi+\frac{2\pi k}{M}) \ . \end{aligned}$$

Tomamos ahora $M = 2(m+1) > n$ y $\phi = \frac{\pi}{M}$ para que el conjunto de abscisas sea simétrico respecto a π. Entonces $\phi+\frac{2\pi k}{M} = \frac{2k+1}{m+1}\frac{\pi}{2}$ $(k = 0 \div 2m+1)$ y los valores $t_n(\frac{(2k+1)\pi}{2(m+1)})$ aparecen dos veces: si $l = 2m+1-k$, cuando $k = 0 \div m$, tenemos que $l = 2m+1 \div m+1$, y

$$t_n\left(\frac{(2k+1)\pi}{2(m+1)}\right) = t_n\left(2\pi - \frac{(2k+1)\pi}{2(m+1)}\right) = t_n\left(\frac{(2l+1)\pi}{2(m+1)}\right) \ .$$

Obtenemos así la siguiente fórmula de integración numérica de $m+1$ abscisas:

$$\int_0^{\pi} F(\theta)d\theta \simeq \frac{\pi}{m+1}\sum_{k=0}^{m} F\left(\frac{(2k+1)\pi}{2(m+1)}\right) \ ,$$

exacta para los polinomios trigonométricos en cosenos de grado menor o igual que $2m+1$.

Haciendo ahora el cambio $t = \cos\theta$, usual en la aproximación a través de polinomios de Chebichev (véase el apartado 3.2.4), resulta que $f(t) = F(\arccos t)$ es un polinomio de grado menor o igual que n si F es un polinomio trigonométrico en cosenos de grado

menor o igual que n (recuérdese la fórmula de Moivre) y se tiene la siguiente fórmula de integración numérica de $m+1$ abscisas

$$\int_{-1}^{1} \frac{f(t)}{\sqrt{1-t^2}} dt \simeq \frac{\pi}{m+1} \sum_{k=0}^{m} f\left(\cos \frac{(2k+1)\pi}{2(m+1)}\right) , \tag{4.24}$$

exacta para los polinomios de grado menor o igual que $2m+1$, llamada *fórmula de Gauss-Chebichev* (véase un ejemplo de su aplicación en el problema 4.7).

Las fórmulas anteriores también pueden escribirse como:

$$\begin{aligned} \int_0^{\pi} F(\theta) d\theta &\simeq \frac{\pi}{m+1} \sum_{k=0}^{m} F(\theta_k) , \\ \int_{-1}^{1} \frac{f(t)}{\sqrt{1-t^2}} dt &\simeq \frac{\pi}{m+1} \sum_{k=0}^{m} f(t_k) , \end{aligned}$$

donde θ_k $(k = 0 \div m)$ son los ceros de $\psi_{m+1}(\theta) = \cos((m+1)\theta)$ y t_k $(k = 0 \div m)$, los de la función $T_{m+1}(t) = \cos((m+1)\arccos t) = \psi_{m+1}(\arccos t)$. Recuérdese, del apartado 3.2.4, que $\psi_{m+1}(\theta)$ y $T_{m+1}(t)$ forman parte de las familias $\psi_j(\theta)$ $(j \geq 0)$ y $T_j(t)$ $(j \geq 0)$ ortogonales respecto a los productos escalares

$$(F, G) = \int_0^{\pi} F(\theta) G(\theta) d\theta , \quad (f, g) = \int_{-1}^{1} \frac{f(t) g(t)}{\sqrt{1-t^2}} dt ,$$

respectivamente.

De la exactitud de (4.24) para los polinomios de grado menor o igual que $2m+1$, resulta que $\psi_j(\theta) = \cos j\theta$ y $T_j(t) = \psi_j(\arccos t)$ $(j = 0 \div m)$ son ortogonales respecto a los productos escalares discretos

$$\begin{aligned} (F, G)_m &= \sum_{k=0}^{m} F(\theta_k) G(\theta_k) , \\ (f, g)_m &= \sum_{k=0}^{m} f(t_k) g(t_k) , \end{aligned}$$

tal como ya se ha demostrado en el capítulo 3, con

$$\begin{aligned} (\psi_j, \psi_j) = (T_j, T_j) &= \frac{\pi}{m+1} (\psi_j, \psi_j)_m = \frac{\pi}{m+1} (T_j, T_j)_m \\ &= \begin{cases} \pi & (j = 0) , \\ \frac{\pi}{2} & (j = 1 \div m) . \end{cases} \end{aligned}$$

Fórmulas gaussianas

Veremos a continuación que la elección de las abscisas x_k $(k = 0 \div m)$ como ceros de un polinomio $\psi_{m+1}(x)$, de grado $m+1$, de una familia de polinomios ortogonales llevará siempre a fórmulas de cuadratura exactas para los polinomios de grado menor o igual que $2m+1$.

Así, sea $w : [a,b] \longrightarrow \mathbb{R}$ una función peso positiva y continua sobre el intervalo $[a,b]$ y sea $\psi_{m+1}(x) = A_{m+1}x^{m+1} + \cdots$ el polinomio ortogonal de grado $m+1$, con coeficiente principal A_{m+1}, asociado al producto escalar

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

Este polinomio $\psi_{m+1}(x)$ tiene $m+1$ ceros simples x_k $(k = 0 \div m)$ (por lo tanto, diferentes dos a dos), que se encuentran en el intervalo (a,b).

En efecto, si $\psi_{m+1}(x)$ sólo cambiase de signo en i abscisas α_1, ..., α_i de $[a,b]$, con $1 \leq i \leq m$, entonces el polinomio

$$q_i(x)\psi_{m+1}(x) \equiv (x-\alpha_1)\cdots(x-\alpha_i)\psi_{m+1}(x) \ ,$$

de grado $m+i+1$, no cambiaría de signo sobre (a,b) y, por lo tanto, la integral

$$\int_a^b w(x)q_i(x)\psi_{m+1}(x)dx = (q_i, \psi_{m+1})$$

sería no nula, en clara contradicción con el hecho de que $\psi_{m+1}(x)$ es ortogonal a cualquier polinomio de grado menor o igual que m.

Consideramos ahora la siguiente fórmula de integración numérica de $m+1$ abscisas, evaluada sobre los ceros de $\psi_{m+1}(x)$:

$$\int_a^b w(x)f(x)dx \simeq \sum_{k=0}^{m} W_k f(x_k) \ . \tag{4.25}$$

De entrada, imponiendo exactitud para los polinomios de grado menor o igual que m, obtenemos como en (4.3) que los pesos W_k vienen dados por

$$W_k = \int_a^b l_k(x)w(x)dx \ , \quad l_k(x) = \prod_{k \neq i} \frac{x - x_i}{x_k - x_i} \quad (k = 0 \div m) \ . \tag{4.26}$$

Comprobaremos a continuación que, con esta elección de abscisas y pesos, la fórmula anterior es exacta también para los polinomios de grado menor o igual que $2m+1$. Las fórmulas así obtenidas se llaman *fórmulas gaussianas de $m+1$ abscisas*.

Sea $p_{2m+1}(x)$ un polinomio de grado menor o igual que $2m+1$ y sean $q_m(x)$ y $r_m(x)$ los polinomios de grado menor o igual que m que son respectivamente el cociente y el resto obtenidos al dividir $p_{2m+1}(x)$ por el polinomio $\psi_{m+1}(x)$ de grado $m+1$

$$p_{2m+1}(x) = q_m(x)\psi_{m+1}(x) + r_m(x) \ .$$

El polinomio $q_m(x)$ será, pues, ortogonal a $\psi_{m+1}(x)$: $(q_m, \psi_{m+1}) = 0$; por lo tanto,

$$\begin{aligned} \int_a^b w(x)p_{2m+1}(x)dx &= \int_a^b w(x)q_m(x)\psi_{m+1}(x) \\ &\quad + \int_a^b w(x)r_m(x)dx \\ &= \int_a^b w(x)r_m(x)dx = \sum_{k=0}^{m} W_k r_m(x_k) \ , \end{aligned}$$

ya que (4.25) es exacta para el polinomio $r_m(x)$ de grado menor o igual que m.

Usando ahora que $\psi_{m+1}(x_k) = 0$ $(k = 0 \div m)$, obtenemos la exactitud de (4.25) para $p_{2m+1}(x)$:

$$\begin{aligned}\sum_{k=0}^{m} W_k p_{2m+1}(x_k) &= \sum_{k=0}^{m} W_k q_m(x_k)\psi_{m+1}(x_k) + \sum_{k=0}^{m} W_k r_m(x_k) \\ &= \int_a^b w(x)p_{2m+1}(x)dx \ .\end{aligned}$$

Error de las fórmulas gaussianas

Para funciones $f \in \mathcal{C}^{2m+2}([a,b])$ podemos dar una expresión del error de las fórmulas gaussianas. Para ello, consideramos el polinomio interpolador de Hermite $p_{2m+1}(x)$ a f en las abscisas x_k $(k = 0 \div m)$ (de grado menor o igual que $2m+1$); por una parte, tenemos que la fórmula gaussiana es exacta para este polinomio y, por otra, disponemos de una fórmula del error de interpolación para todo $x \in [a,b]$

$$f(x) - p_{2m+1}(x) = \frac{f^{(2m+2)}(\xi(x))}{(2m+2)!}\omega_m^2(x) \ ,$$

donde $\xi(x) \in < x_0, x_1, \ldots, x_m, x > \subset [a,b]$ y

$$\omega_m(x) = (x-x_0)(x-x_1)\cdots(x-x_m) = \frac{\psi_{m+1}(x)}{A_{m+1}} \ .$$

Multiplicando por $w(x)$ e integrando entre a y b, se obtiene la expresión siguiente del error de la fórmula gaussiana de $m+1$ abscisas:

$$\begin{aligned}&\int_a^b w(x)f(x)dx - \sum_{k=0}^{m} W_k f(x_k) \qquad (4.27)\\ &= \frac{f^{(2m+2)}(\xi)}{(2m+2)!}\frac{1}{A_{m+1}^2}\int_a^b w(x)\psi_{m+1}^2(x)dx \ ,\end{aligned}$$

donde $\xi \in (a,b)$ y A_{m+1} indica el coeficiente principal del polinomio ortogonal $\psi_{m+1}(x)$ elegido.

Es necesario hacer notar que el factor del error

$$\frac{(\psi_{m+1}, \psi_{m+1})}{A_{m+1}^2} = \int_a^b w(x)\omega_m^2(x)dx$$

puede obtenerse por aplicación de la fórmula gaussiana a $f(x) = x^{2m+2}$, ya que entonces el otro factor es la unidad

$$\frac{f^{(2m+2)}(\xi)}{(2m+2)!} = 1 \ .$$

Pesos de las fórmulas gaussianas

Los pesos W_k, que pueden hallarse por (4.26), admiten otras expresiones equivalentes.

Usando la exactitud de (4.25) para los polinomios $l_k^2(x)$ de grado $2m$, obtenemos

$$W_k = \int_a^b w(x) l_k^2(x) dx > 0 \quad (k = 0 \div m) \ ,$$

de donde resulta que todos los pesos de una fórmula gaussiana son positivos.

Imponiendo ahora la exactitud para

$$f(x) = \frac{\psi_{m+1}(x)}{x - x_k} = A_{m+1} \prod_{i \neq k} (x - x_i) \ ,$$

resulta que $f(x_i) = 0$ $(i \neq k)$ y $f(x_k) = \psi'_{m+1}(x_k)$, dando lugar a

$$W_k = \frac{1}{\psi'_{m+1}(x_k)} \int_a^b w(x) \frac{\psi_{m+1}(x)}{x - x_k} dx \quad (k = 0 \div m) \ ;$$

análogamente, para

$$f(x) = \frac{\psi_{m+1}^2(x)}{(x - x_k)^2} = A_{m+1}^2 \prod_{i \neq k} (x - x_i)^2 \ ,$$

se obtiene

$$W_k = \frac{1}{(\psi'_{m+1}(x_k))^2} \int_a^b w(x) \frac{\psi_{m+1}^2(x)}{(x - x_k)^2} dx > 0 \quad (k = 0 \div m) \ .$$

Si substituimos en la fórmula (4.27) el polinomio de grado $2m + 2$

$$f(x) = \frac{\psi_{m+1}(x)}{x - x_k} \psi_{m+2}(x) \ ,$$

y tenemos en cuenta el hecho de que

$$\left(\frac{\psi_{m+1}}{x - x_k}, \psi_{m+2}\right) = 0 \ ,$$

llegamos a la expresión

$$W_k = -\frac{A_{m+2}(\psi_{m+1}, \psi_{m+1})}{A_{m+1} \psi'_{m+1}(x_k) \psi_{m+2}(x_k)} \ .$$

Finalmente, usando la relación de recurrencia de los polinomios ortogonales, obtenemos

$$W_k = \frac{A_{m+1}(\psi_m, \psi_m)}{A_m \psi'_{m+1}(x_k) \psi_m(x_k)} \ .$$

Ejemplos de fórmulas gaussianas

1. Volvemos a la fórmula de Gauss-Chebichev (4.24). El polinomio ortogonal de grado $m+1$ y coeficiente principal $A_{m+1} = 2^m$ correspondiente es el polinomio de Chebichev $T_{m+1}(t) = \cos((m+1)\arccos t)$.

Sus ceros $x_k \equiv t_k = \cos\theta_k$, $\theta_k = \frac{(2k+1)\pi}{2(m+1)}$ $(k = 0 \div m)$ son las abscisas de la fórmula; entonces, los pesos de la fórmula de Gauss-Chebichev son

$$\begin{aligned} W_k &= -\frac{\pi}{T'_{m+1}(t_k)T_{m+2}(t_k)} = -\frac{\pi}{\frac{(-1)^k(m+1)}{\operatorname{sen}\theta_k}(-1)^{k+1}\operatorname{sen}\theta_k} \\ &= \frac{\pi}{m+1} \ (k = 0 \div m) \ , \end{aligned}$$

tal como se había visto; la fórmula con el término de error correspondiente queda finalmente así

$$\begin{aligned} \int_{-1}^{1} \frac{f(t)}{\sqrt{1-t^2}}dt &= \frac{\pi}{m+1}\sum_{k=0}^{m} f\left(\cos\frac{(2k+1)\pi}{2(m+1)}\right) \\ &\quad + \frac{\pi}{2^{2m+1}(2m+2)!}f^{(2m+2)}(\xi) \ , \quad \xi \in (-1,1) \ . \end{aligned}$$

2. Si consideramos ahora la función peso $w(x) = 1$ en el intervalo $[-1,1]$, el polinomio ortogonal de grado $m+1$ y coeficiente principal

$$A_{m+1} = \frac{(2m+2)!}{2^{m+1}[(m+1)!]^2}$$

es el polinomio de Legendre

$$P_{m+1}(t) = \frac{1}{2^{m+1}(m+1)!}\frac{d^{m+1}}{dt^{m+1}}[(t^2-1)^{m+1}] \ ,$$

introducido en el apartado 3.2.3. Sus ceros $t_0, t_1, \ldots, t_m$ están distribuidos simétricamente respecto al origen y la fórmula gaussiana, llamada de *Gauss-Legendre*, con el error término de error correspondiente queda (véase el problema 4.8)

$$\int_{-1}^{1} f(t)dt = \sum_{k=0}^{m} W_k f(t_k) + \frac{2^{2m+3}[(m+1)!]^4}{(2m+3)[(2m+2)!]^3}f^{(2m+2)}(\xi) \ ,$$

con $\xi \in (-1,1)$ y

$$W_k = \frac{2}{(1-t_k^2)[P'_{m+1}(t_k)]^2} \ (k = 0 \div m) \ .$$

Esta fórmula de Gauss-Legendre puede ser extendida a cualquier intervalo $[a,b]$, mediante el cambio

$$x = \frac{b-a}{2}t + \frac{a+b}{2} \ .$$

4.3 SUMACIÓN NUMÉRICA

4.3.1 Introducción

Consideramos a continuación las sumas (finitas)

$$S_n = \sum_{j=0}^{n} a_j \ , \tag{4.28}$$

donde a_j $(j = 0 \div n)$ son los $n+1$ primeros términos de la sucesión $(a_j)_{j\geq 0}$. Las sumas S_j $(j \geq 0)$ se denominan *sumas parciales* de dicha sucesión.

A menudo, los términos de la sucesión podrán escribirse en la forma

$$a_j = f(j) \ \ (j \geq 0) \ ,$$

siendo $f : [0, \infty) \longrightarrow \mathbb{R}$ (también podría tomar valores complejos).

La sucesión $(S_j)_{j\geq 0}$ recibe el nombre de *serie* y, si es convergente a un valor S, escribiremos

$$S = \sum_{j=0}^{\infty} a_j$$

y diremos que S es la *suma de la serie de término general* a_j.

Esto quiere decir que, para cualquier $\epsilon > 0$, existe un índice n, dependiente de ϵ, tal que $|S - S_n| \leq \epsilon$ o, también,

$$S = S_n \pm \epsilon \ . \tag{4.29}$$

El problema de la *sumación numérica* consiste en proporcionar métodos para aproximar sumas del tipo (4.28) o, si existe, su límite cuando n tiende a infinito. Desde el punto de vista numérico, el caso límite se reduce a calcular S_n con $n = n(\epsilon)$ suficientemente grande para que se cumpla (4.29), dado un error permisible ϵ. Ahora bien, si n es muy grande, el cálculo directo de (4.28) es muy costoso y se hace necesario recurrir a métodos alternativos.

También consideraremos sumas cuyos términos son funciones de una variable real (o compleja)

$$S_n(x) = \sum_{j=0}^{n} f_j(x) \ ,$$

que dan lugar, en el caso límite $n \to \infty$, a las llamadas *series de funciones*

$$S(x) = \sum_{j=0}^{\infty} f_j(x) \ .$$

Estas series tienen sentido sólo para aquellos valores de x para los cuales la serie numérica correspondiente es convergente. El conjunto de abscisas en las que la serie de funciones converge recibe el nombre de *región de convergencia*.

Un caso muy importante es el de las *series de potencias* del tipo

$$S(x) = \sum_{j=0}^{\infty} a_j (x - x_0)^j \ ,$$

que aparecen al hacer desarrollos de Taylor de una función cerca de una abscisa x_0. En efecto, si $f : I \longrightarrow \mathbb{R}$ es una función indefinidamente diferenciable sobre un intervalo I, podemos escribir, para cualquier $x_0 \in I$,

$$f(x) = p_n(x) + R_n(x) \quad (n \geq 0) \ ,$$

donde

$$p_n(x) = \sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j$$

es el polinomio de Taylor de grado menor o igual que n de f en x_0 y $R_n(x)$, el error de la interpolación de Taylor que admite las expresiones (3.8) y (3.9).

Recordemos que, para cualquier n fijo,

$$\lim_{x \to x_0} \frac{|R_n(x)|}{|x - x_0|^n} = 0 \ .$$

La serie de potencias formada por la sucesión de polinomios de Taylor de una función f indefinidamente diferenciable cerca de x_0 recibió el nombre de *desarrollo en serie de Taylor de f cerca de x_0* en el apartado 3.1.4. De hecho, se trata de un caso particular de los llamados desarrollos asintóticos que se introducen a continuación.

Diremos que

$$f(x) = \sum_{j=0}^{n} a_j(x - x_0)^j + R_n(x)$$

es un *desarrollo asintótico de una función f cerca de x_0* cuando

$$\lim_{x \to x_0} \frac{|R_n(x)|}{|x - x_0|^n} = 0 \ .$$

En tal caso, escribiremos

$$f(x) \sim \sum_{j=0}^{\infty} a_j(x - x_0)^j \quad (x \to x_0) \ .$$

De manera análoga, diremos que

$$f(x) = \sum_{j=0}^{n} \frac{a_j}{x^j} + R_n(x)$$

es un *desarrollo asintótico de f cerca de infinito* y escribiremos

$$f(x) \sim \sum_{j=0}^{\infty} \frac{a_j}{x^j} \quad (x \to \infty)$$

cuando

$$\lim_{x \to \infty} |x|^n \, |R_n(x)| = 0 \ .$$

Dada una abscisa x, el desarrollo asintótico de una función f

$$f(x) = S_n(x) + R_n(x)$$

no es necesariamente convergente a $f(x)$, cuando n tiende a infinito; esto es: los restos $R_n(x)$ no siempre convergen a 0.

En el caso de los desarrollos de Taylor, si se cumplía la condición de convergencia

$$\lim_{n\to\infty} R_n(x) = 0 \quad \forall x \in (a,b) \ ,$$

decíamos que la función f era analítica en el intervalo (a,b). En la tabla 3.5, se ecuentran diversos ejemplos de funciones analíticas.

Un desarrollo asintótico puede usarse para el cálculo de una función, aunque no se cumpla la condición de convergencia

$$\lim_{n\to\infty} R_n(x) = 0 \ .$$

En este caso, el error cometido no será arbitrariamente pequeño.

Si se dispone de una expresión para cada $R_j(x)$ $(j \geq 0)$, será necesario hallar un valor n para el cual $|R_n(x)|$ sea menor que el error permitido. Esto no siempre será posible; en caso de que no lo sea, interesará hallar el valor de n que haga mínimo $|R_n(x)|$ y deberemos conformarnos con tener $f(x)$ calculado con este error (véase el problema 4.11).

Las series divergentes que aparecen en la aplicación de desarrollos asintóticos de funciones reciben el nombre de *series semiconvergentes* porque permiten también el cálculo de dichas funciones, aunque no con un error arbitrariamente pequeño.

4.3.2 Cotas de los restos de las series

Dada una serie convergente

$$S = \sum_{j=0}^{\infty} a_j \ ,$$

donde a_j puede ser eventualmente del tipo $f(j)$ o $f_j(x)$, es necesario, en primer lugar, saber estimar sus *restos*

$$R_n = S - S_n = \sum_{j=n+1}^{\infty} a_j$$

para poder aproximar correctamente S mediante el cálculo directo de S_n; si n resulta ser demasiado grande, deberá buscarse otro método para aproximar S_n o S.

Por ejemplo, si

$$|a_{j+1}| \leq \rho\,|a_j| \ \ (j \geq n) \ ,$$

con $\rho < 1$; entonces, por inducción,

$$|a_{j+l}| \leq \rho^l\,|a_j| \ \ (l \geq 1)\ (j \geq n)$$

y obtenemos la cota del resto

$$|R_n| \leq \sum_{j=n+1}^{\infty} |a_j| \leq \left(\sum_{j=0}^{\infty} \rho^j\right) |a_{n+1}| = \frac{|a_{n+1}|}{1-\rho} \ ;$$

es decir, la magnitud del resto es menor o igual que la magnitud del primer término despreciado dividida por el factor $1-\rho$.

Si, en cambio, tenemos $a_j = f(j)$, con $f : [0,\infty) \to \mathbb{R}$ decreciente (al menos sobre $[n,\infty)$); entonces, si $x \in [j, j+1]$ $(j \geq n)$, se cumple $f(j) \geq f(x) \geq f(j+1)$ y, por lo tanto,

$$f(j) = \int_j^{j+1} f(j)dx \geq \int_j^{j+1} f(x)dx \geq f(j+1) \ .$$

Así, resulta

$$R_{n-1} = \sum_{j=n}^{\infty} f(j) \geq \int_n^{\infty} f(x)dx \geq R_n \ ;$$

en particular, se asegura que el resto

$$\sum_{j=n+1}^{\infty} f(j)$$

es convergente si y sólo si la integral

$$\int_n^{\infty} f(x)dx$$

es convergente.

Recordemos el criterio de alternancia de los restos del apartado 3.1.4: si los restos R_n y R_{n+1} tienen signos diferentes, entonces

$$S \in < S_n, S_{n+1} > \ , \quad |R_n| \leq |a_{n+1}| \ ;$$

esto es: la magnitud del resto es menor que la del primer término despreciado. En particular, si tenemos una *serie alternada* a partir del término $n+1$,

$$a_j a_{j+1} < 0 \quad (j \geq n+1) \ ,$$

que sea monótona decreciente, en módulo, hacia 0, a partir del término n,

$$|a_{j+1}| \leq |a_j| \ \ (j \geq n+1) \ , \quad \lim_{j\to\infty} a_j = 0 \ ;$$

entonces

$$\operatorname{sgn}(a_j + a_{j+1}) = \operatorname{sgn}(a_{j+2} + a_{j+3}) \ (j \geq n+1)$$

y, por lo tanto,

$$R_{j-1} = (a_j + a_{j+1}) + (a_{j+2} + a_{j+3}) + \cdots$$

tiene el mismo signo que a_j $(j \geq n+1)$. Resulta, pues, que

$$R_n R_{n+1} < 0 \ , \quad |R_n| \leq |a_{n+1}| \ .$$

Recopilaremos ahora estas cotas de R_n en los criterios de acotación siguientes:

1. *Criterio de comparación con una serie geométrica*

   Si existe $\rho < 1$ tal que $|a_{j+1}| \leq \rho\,|a_j|$ $(j \geq n)$, entonces

$$|R_n| \leq \frac{|a_{n+1}|}{1-\rho} \leq \frac{\rho}{1-\rho}\,|a_n| \ .$$

2. *Criterio integral*

   Si $a_j = f(j)$ con $f : [0, \infty) \longrightarrow \mathbb{R}$ decreciente sobre $[n, \infty)$, entonces

   $$R_n \leq \int_n^\infty f(x)dx \leq R_{n-1} \ .$$

   En particular, bajo la condición $|a_j| \leq f(j)$ $(j \geq n)$,

   $$|R_n| \leq \int_n^\infty f(x)dx \ .$$

3. *Criterio de alternancia de restos*

   Si $R_n R_{n+1} < 0$, entonces

   $$S = S_n + R_n = S_{n+1} + R_{n+1} \in < S_n, S_{n+1} > \ , \ \ |R_n| \leq |a_{n+1}| \ .$$

4. *Criterio para series alternadas*

   Si $a_j a_{j+1} < 0$, $|a_{j+1}| \leq |a_j|$ $(j \geq n+1)$ y $a_j \to 0$ $(j \to \infty)$; entonces

   $$S \in < S_n, S_{n+1} > \ , \ \ |R_n| \leq |a_{n+1}| \ .$$

Naturalmente, estos criterios se aplican de forma directa también al cálculo de sumas finitas.

4.3.3 Métodos de sumación numérica

Usando las cotas del resto R_n, el cálculo de la serie $S = S_n + R_n$ se reduce al cálculo de S_n para n tal que la magnitud de R_n sea suficientemente pequeña de acuerdo con la precisión deseada. Ahora bien, el cálculo directo de S_n puede llegar a ser muy costoso, para series llamadas *lentamente convergentes*. Presentaremos primero un ejemplo de este tipo de series y, a continuación, expondremos algunos métodos que permiten aproximar S_n (y S) de una forma alternativa más eficaz que el cálculo directo, y que llamaremos *métodos de aceleración de la sumación*.

EJEMPLO

Si queremos sumar la serie

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

con un error menor que 10^{-9} mediante el cálculo directo de las sumas parciales, el criterio integral aconseja tomar más de 10^9 términos.

La fórmula de Euler-Maclaurin

Recordemos la fórmula de Euler-Maclaurin (4.23) para una función cualquiera f de $\mathcal{C}^{2s+2}([a, b])$: si dividimos $[a, b]$ en n partes iguales de longitud $h = \frac{b-a}{n}$, entonces

$$\begin{aligned} T(h) \ &= \ \int_a^b f(x)dx + \sum_{r=1}^{s} h^{2r} \frac{B_{2r}}{(2r)!} \left[f^{(2r-1)}(b) - f^{(2r-1)}(a) \right] \\ &\quad + \frac{B_{2s+2}}{(2s+2)!}(b-a) f^{(2s+2)}(\xi) h^{2s+2} \ , \ \ \xi \in (a, b) \ , \end{aligned}$$

donde los coeficientes B_{2r} son los números de Bernoulli y

$$\begin{aligned} T(h) &= h\left[\frac{f(a)}{2}+\sum_{j=1}^{n-1} f(a+jh)+\frac{f(b)}{2}\right] \\ &= h\sum_{j=0}^{n} f(a+jh)-\frac{h}{2}\left[f(a)+f(a+nh)\right] \ . \end{aligned}$$

Dicha fórmula puede escribirse de la manera siguiente, recibiendo, en este caso, el nombre de *fórmula de Euler-Maclaurin par sumas*:

$$\begin{aligned} \sum_{j=0}^{n} f(a+jh) = \frac{1}{h}\int_a^{a+nh} f(x)dx + \frac{1}{2}\left[f(a)+f(a+nh)\right] \\ +\sum_{r=1}^{s} h^{2r-1}\frac{B_{2r}}{(2r)!}\left[f^{(2r-1)}(a+nh)-f^{(2r-1)}(a)\right] \\ +R_s \ , \end{aligned}$$

donde

$$R_s = nh^{2s+2}\frac{B_{2s+2}}{(2s+2)!}f^{(2s+2)}(\xi) \ , \quad \xi\in(a,a+nh) \ . \tag{4.30}$$

Por la alternancia de los signos de los números de Bernoulli (véase (4.18)), se deduce primero que, si $f\in\mathcal{C}^{2s+4}([a,a+nh])$ y $f^{(2s+2)}f^{(2s+4)}\geq 0$ en $(a,a+nh)$, entonces $R_sR_{s+1}<0$ y, por el criterio de alternancia de los restos,

$$|R_s|\leq h^{2s+1}\frac{|B_{2s+2}|}{(2s+2)!}\left|f^{(2s+1)}(a+nh)-f^{(2s+1)}(a)\right| \ . \tag{4.31}$$

La *integral impropia*

$$\int_a^{\infty} f(x)dx$$

es convergente cuando existe

$$J(f)=\lim_{n\to\infty}\int_a^{a+nh} f(x)dx \ .$$

Usando integrales impropias convergentes, en caso de que

$$\lim_{x\to\infty} f^{(2r-1)}(x)=0 \quad (r=1\div s) \ ,$$

la fórmula de Euler-Maclaurin para sumas finitas puede generalizarse al cálculo de sumas de series, dando lugar a la *fórmula de Euler-Maclaurin para series*

$$\begin{aligned} \sum_{j=0}^{\infty} f(a+jh) &= \frac{1}{h}\int_a^{\infty} f(x)dx+\frac{f(a)}{2} \\ &\quad -\sum_{r=1}^{s} h^{2r-1}\frac{B_{2r}}{(2r)!}f^{(2r-1)}(a)+R_s \ . \end{aligned}$$

Una expresión para el resto R_s, análoga a (4.30) no es ahora válida, porque $n \to \infty$ y ha de recurrirse a (4.17) para obtener una expresión correcta del mismo (véase el problema 4.10). Si $f^{(2s+2)}f^{(2s+4)} > 0$ en $[a, \infty)$ y

$$\lim_{x\to\infty} f^{(2s+1)}(x) = 0 \ ,$$

la fórmula (4.31) nos asegura que $|R_s|$ es menor que la magnitud del primer término despreciado

$$|R_s| \leq h^{2s+1}\frac{|B_{2s+2}|}{(2s+2)!}\left|f^{(2s+1)}(a)\right| \ .$$

Esta fórmula se usa, tomando $h = 1$, para calcular sumas de series de forma muy eficiente en un gran número de casos.

Finalmente, notemos que el desarrollo de la fórmula de Euler-Maclaurin no es siempre convergente; es decir, no tiene que cumplirse necesariamente que $R_s \to 0$, cuando $s \to \infty$. De hecho, la magnitud de los números de Bernoulli B_{2s+2} crece indefinidamente cuando s tiende a infinito, de manera que su cociente con $(2s+2)!$ decrece potencialmente en s; así que, si $f^{(2s+1)}$ tiene un comportamiento de crecimiento más fuerte que aquel cociente, tendremos que R_s crecerá indefinidamente. A menudo, al variar s de 0 a ∞, se observa que la magnitud de R_s decrece primero, para crecer después; se trata de un ejemplo típico de series semiconvergentes. Para este tipo de series divergentes, interesará tomar s de forma que la magnitud de R_s sea cuanto más pequeña mejor; ahora bien, a diferencia de las series convergentes, el valor más pequeño de R_s es fijo y no arbitrariamente pequeño, como en aquéllas. Es decir, puede darse el caso que, si la precisión pedida en la suma es demasiado grande, no pueda ser alcanzada para ningún valor de s (véase un ejemplo en el problema 4.11).

Método de comparación

Para calcular la suma S de una serie lentamente convergente

$$\sum_{j=0}^{\infty} a_j \ ,$$

buscamos otra serie

$$\sum_{j=0}^{\infty} b_j$$

tal que sea convergente a un valor conocido T y que la serie

$$\sum_{j=0}^{\infty}(a_j - b_j)$$

sea más rápidamente convergente; así, para el cálculo de S, basta calcular la suma de la última serie, más rápidamente convergente que la inicial, y después sumar T al resultado obtenido,

$$S = T + \sum_{j=0}^{\infty}(a_j - b_j) \ .$$

Será necesario, pues, conocer algunos ejemplos de sumas de series para poder aplicar este método de comparación al cálculo de las sumas desconocidas de otras series.

Los ejemplos más obvios de sumas finitas (y también, de series) calculables exactamente corresponden a las llamadas *sumas telescópicas*; esto es, sumas

$$\sum_{j=m}^{n} b_j \ ,$$

donde $b_j = \Delta T_j = T_{j+1} - T_j$, ya que, entonces

$$\sum_{j=m}^{n} b_j = T_{n+1} - T_m \ .$$

Así, si consideramos las *funciones factoriales*, para $\alpha \in \mathbb{Z}$,

$$x^{(\alpha)} = \begin{cases} x(x-1)\cdots(x-\alpha+1) & (\alpha > 0) \\ 1 & (\alpha = 0) \\ \frac{1}{(x+1)(x+2)\cdots(x-\alpha)} & (\alpha < 0) \end{cases}$$

resulta que $\Delta x^{(\alpha)} = \alpha x^{(\alpha-1)}$ y, por lo tanto, si $\alpha \neq -1$,

$$x^{(\alpha)} = \frac{\Delta x^{(\alpha+1)}}{\alpha+1} \ , \quad \sum_{j=m}^{n} j^{(\alpha)} = \frac{(n+1)^{(\alpha+1)} - m^{(\alpha+1)}}{\alpha+1} \ .$$

En particular, si $\alpha < -1$,

$$\sum_{j=m}^{\infty} j^{(\alpha)} = -\frac{m^{(\alpha+1)}}{\alpha+1} \ .$$

En el problema 4.9 se presenta una aplicación práctica de este método.

4.4 EXTRAPOLACIÓN

4.4.1 Introducción

En la resolución numérica de muchos problemas matemáticos, cuyo objetivo consiste en el cálculo de un valor v, pueden distinguirse dos etapas:

- Discretización: Se calculan aproximaciones numéricas $F(h)$ al valor v, dependientes de un parámetro h, llamado *paso de discretización*. Por ejemplo, en el cálculo numérico de derivadas e integrales, h es la separación entre abscisas consecutivas. En el caso de que las cantidades aproximadas dependan de una variable entera n de forma que el valor v se alcance cuando n tiende a ∞, tomaremos $h = \frac{1}{n}$.
- Paso al límite: Se considera el límite de las aproximaciones $F(h)$, cuando h tiende a 0.

En el contexto numérico, el paso al límite consiste en realizar los cálculos con pasos h cada vez más pequeños lo que comporta dos tipos de dificultades:

1. El número de operaciones aumenta enormemente.

2. Aparecen importantes errores de redondeo, por exceso de operaciones o por cancelación de términos en restas de cantidades muy cercanas, que desfiguran el valor de $F(h)$.

4.4.2 Método de Richardson de extrapolación repetida

Para evitar estas dificultades, si conocemos el desarrollo asintótico de $F(h)$ cuando h tiende a 0, el *método de Richardson* proporciona, a partir de una sucesión $F(h_1)$, $F(h_2)$, ... de valores de F calculados directamente, nuevas sucesiones más rápidamente convergentes hacia v (diversas aplicaciones al cálculo de derivadas e integrales pueden consultarse en los problemas 4.12, 4.13 y 4.15).

Así, si sabemos que el desarrollo asintótico de $F(h)$ es de la forma

$$F(h) = v + a_1 h^{p_1} + a_2 h^{p_2} + \cdots \ , \quad 0 < p_1 < p_2 < \cdots \ ,$$

tomamos un número q cualquiera mayor que 1 y definimos la sucesión de extrapolaciones siguientes:

$$F_1(h) = F(h) \ , \quad F_{j+1}(h) = F_j(h) + \frac{F_j(h) - F_j(qh)}{q^{p_j} - 1} \ (j \geq 1) \ .$$

Puede probarse, por inducción, que

$$F_j(h) = v + a_j^{(j)} h^{p_j} + a_{j+1}^{(j)} h^{p_{j+1}} + \cdots \ ,$$

con lo cual los términos que contaminan v son ahora mucho más pequeños, si tomamos el paso h suficientemente pequeño.

Para calcular $F_{j+1}(h)$ necesitamos conocer $F_j(h)$ y $F_j(qh)$; para calcular estos últimos, $F_{j-1}(h)$, $F_{j-1}(qh)$ y $F_{j-1}(q^2h)$, y así sucesivamente. Por ello, la forma más cómoda de proceder al cálculo de los $F_j(h)$ $(j \geq 1)$ es mediante un esquema triangular como el que se muestra en la tabla 4.1.

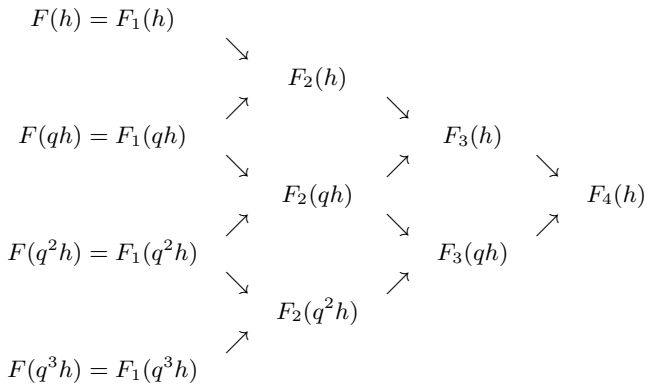


Tabla 4.1: Esquema del método de Richardson.

OBSERVACIONES

1. Para utilizar este algoritmo basta con conocer los exponentes p_1, p_2, ... y no los coeficientes a_1, a_2, ...

2. Conocidos $F(h_1)$, $F(h_2)$, ..., se pretende calcular $F(0)$; es decir, el valor de F en una abscisa fuera de $< h_1, h_2, \ldots >$. Por esta razón, el proceso recibe el nombre de *extrapolación*.

3. El paso de extrapolación de $F_j(h)$ a $F_{j+1}(h)$ se lleva a cabo añadiendo a $F_j(h)$ la diferencia entre $F_j(h)$ y $F_j(qh)$ dividida por $q^{p_j} - 1$. Por esta razón, se le suele llamar *extrapolación del tipo* $\frac{\Delta}{q^{p_j}-1}$.

4.5 CÁLCULO CON OPERADORES

4.5.1 Introducción

Los métodos basados en el cálculo formal con operadores se revelan como herramientas eficaces y elegantes para la construcción de fórmulas (de interpolación, de derivación numérica, de integración numérica, ...) con abscisas equidistantes, exactas para polinomios hasta un cierto grado y, por lo tanto, aproximadas para otro tipo de funciones.

Definiciones

Sobre el conjunto $\mathcal{P}$ de polinomios en una variable (que, de hecho, es un espacio vectorial), definimos los operadores siguientes, mediante su aplicación a una función polinomial f en una abscisa x cualquiera:

$(Ef)(x) = f(x+h)$ (*Operador desplazamiento hacia adelante*).

$(\Delta f)(x) = f(x+h) - f(x)$ (*Operador diferencia hacia adelante*).

$(\nabla f)(x) = f(x) - f(x-h)$ (*Operador diferencia hacia atrás*).

$(\delta f)(x) = f(x+\frac{h}{2}) - f(x-\frac{h}{2})$ (*Operador diferencia centrada*).

$(\mu f)(x) = \frac{1}{2}(f(x+\frac{h}{2}) + f(x-\frac{h}{2}))$ (*Operador media*).

$(Df)(x) = f'(x)$ (*Operador de derivación*).

$(Jf)(x) = \int_x^{x+h} f(t)dt$ (*Operador de integración*).

Todos estos operadores, excepto el operador D, dependen de un paso h. Si es necesaria una notación más explícita respecto al paso h usado (por ejemplo, cuando trabajemos con más de un paso), escribiremos $E_h f$, $\Delta_h f$, etc.

Los cinco primeros operadores E, Δ, ∇, δ y μ pueden definirse también para cualquier función $f : \mathbb{R} \longrightarrow \mathbb{R}$ y aún, para funciones tabuladas sobre abscisas equidistantes, ..., x_{-2}, x_{-1}, x_0, x_1, ..., siempre que h sea un múltiplo entero de $x_{k+1} - x_k$ para E, Δ y ∇, y de $2(x_{k+1} - x_k)$ para δ y μ. Los operadores D y J pueden definirse sobre funciones más

generales que los polinomios (funciones derivables e integrables, respectivamente); pero, entonces requieren un proceso de paso al límite.

Dados dos operadores L_1 y L_2, su *producto (por composición)* L_1L_2 y su *suma* L_1+L_2 vienen dados por

$$(L_1L_2)f = L_1(L_2f)\ , \quad (L_1+L_2)f = L_1f + L_2f\ .$$

En particular, las potencias L^j $(j \geq 0)$ de un operador se definen inductivamente por $L^0 = 1$, donde 1 es el *operador identidad* (es decir, $1f = f\ \forall f \in \mathcal{P}$), $L^j = LL^{j-1}$ $(j \geq 1)$.

4.5.2 Propiedades de los operadores

Se presentan a continuación algunas propiedades muy útiles de los operadores definidos:

1. Si L designa uno cualquiera de los operadores definidos, $L : \mathcal{P} \longrightarrow \mathcal{P}$ es un operador lineal; esto es,

$$L(\alpha f + \beta g) = \alpha Lf + \beta Lg\ ,$$

donde α, β son coeficientes y $f, g \in \mathcal{P}$.

2. Si $f \in \mathcal{P}$ es de grado N, entonces Δf, ∇f, δf y Df son de grado $N-1$ (si $N \geq 1$), Ef y μf son de grado N y Jf es de grado menor o igual que N. En particular, $L^jf = 0$ (si $j \geq N+1$ y $L = \Delta, \nabla, \delta, D$).

3. Toda suma del tipo

$$\sum_{j=0}^{\infty} a_jL^j = a_0 + a_1L + a_2L^2 + a_3L^3 + \dots\ ,$$

con $L = \Delta, \nabla, \delta, D$, y a_j $(j \geq 0)$ coeficientes cualesquiera, se reduce a una suma finita al aplicarla a un polinomio y, por lo tanto, la *serie de operadores*

$$\sum_{j=0}^{\infty} a_jL^j$$

es un operador bien definido sobre $\mathcal{P}$.

A continuación, se dan algunos ejemplos.

Si $t \in \mathbb{R}$ y $L = \Delta, \nabla, \delta, D$:

$$\begin{aligned}
(1+L)^t &\equiv \sum_{j=0}^{\infty} \binom{t}{j} L^j \\
&= 1 + tL + \frac{t(t-1)}{2}L^2 + \frac{t(t-1)(t-2)}{6}L^3 + \cdots\ , \\
e^{tL} &\equiv \sum_{j=0}^{\infty} \frac{t^j}{j!}L^j \\
&= 1 + tL + \frac{t^2}{2}L^2 + \frac{t^3}{6}L^3 + \cdots\ , \\
\ln(1+L) &\equiv \sum_{j=1}^{\infty} \frac{(-1)^{j-1}}{j}L^j
\end{aligned}$$

$$\begin{aligned}
&= L - \frac{1}{2}L^2 + \frac{1}{3}L^3 - \cdots \ , \\
\arg\operatorname{senh} L &\equiv \sum_{r=0}^{\infty} \frac{1}{2r+1} \binom{-\frac{1}{2}}{r} L^{2r+1} \\
&= L - \frac{1}{2}\frac{L^3}{3} + \frac{1\cdot 3}{2\cdot 4}\frac{L^5}{5} - \cdots \ , \\
\operatorname{senh}(tL) &\equiv \frac{1}{2}(e^{tL} - e^{-tL}) = \sum_{r=0}^{\infty} \frac{t^{2r+1}}{(2r+1)!} L^{2r+1} \\
&= tL + \frac{t^3}{3!}L^3 + \frac{t^5}{5!}L^5 + \cdots \ , \\
\cosh(tL) &\equiv \frac{1}{2}(e^{tL} + e^{-tL}) = \sum_{r=0}^{\infty} \frac{t^{2r}}{(2r)!} L^{2r} \\
&= 1 + \frac{t^2}{2!}L^2 + \frac{t^4}{4!}L^4 + \frac{t^6}{6!}L^6 + \cdots
\end{aligned}$$

Los *operadores inversos* (respecto al producto por composición) de $1+L$ y e^{tL} se indican, respectivamente, por $(1+L)^{-1}$ y e^{-tL} . En notación multiplicativa, suele escribirse:

$$\frac{1}{1+L} \equiv (1+L)^{-1} \ , \quad \frac{1}{e^{tL}} \equiv e^{-tL} \ , \quad \sqrt{1+L} = (1+L)^{\frac{1}{2}} \ , \quad \text{etc.}$$

4. Los siete operadores E, Δ, δ, ∇, D, μ y J *conmutan entre ellos.* Esto es que, si L_1 y L_2 son dos de estos operadores, $L_1L_2 = L_2L_1$; por ejemplo, $JD = DJ = \Delta$.

5. Se cumple la relación:

$$(E^t f)(x) = ((1+\Delta)^t f)(x) = f(x+th) \quad \forall t \in \mathbb{R} \quad \forall f \in \mathcal{P} \ .$$

En efecto, esta relación es claramente cierta para $t \in \mathbb{N}$. Veamos ahora que también es válida para $t \in \mathbb{R}$. Esto es consecuencia del hecho de que, si x_0, x_1, x_2, ... forman una partición equidistante con $x_{k+1} - x_k = h$ $(k \geq 0)$, entonces

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

y, por inducción,

$$f[x_0, x_1, \ldots, x_j] = \frac{1}{j!h^j}(\Delta^j f)(x_0) \ .$$

Si $f \in \mathcal{P}$ tiene grado m, usando la variable t tal que $x = x_0 + th$ tenemos $x_0 + th - x_k = (t-k)h$ y, usando la fórmula de interpolación proporcionada por el método de las diferencias divididas, obtenemos la relación buscada:

$$\begin{aligned}
f(x_0 + th) &= \sum_{j=0}^{m} f[x_0, x_1, \ldots, x_j] th(t-1)h \cdots (t-j+1)h \\
&= \sum_{j=0}^{m} \binom{t}{j} (\Delta^j f)(x_0) = ((1+\Delta)^t f)(x_0) \ .
\end{aligned}$$

Nótese que el operador $\ln E = \ln(1+\Delta)$ tiene también sentido y aparece definido en la propiedad 3.

6. Se cumplen las relaciones entre operadores dadas en la tabla 4.2, deducidas a partir de las definiciones y de la *fórmula de Taylor usando operadores*

$$E = e^{hD} \ .$$

	E	Δ	δ	∇	D
E	E	$1+\Delta$	$1+\frac{1}{2}\delta^2+\delta(1+\frac{1}{4}\delta^2)^{\frac{1}{2}}$	$\frac{1}{1-\nabla}$	e^{hD}
Δ	$E-1$	Δ	$\delta(1+\frac{1}{4}\delta^2)^{\frac{1}{2}}+\frac{1}{2}\delta^2$	$\frac{\nabla}{1-\nabla}$	$e^{hD}-1$
δ	$E^{\frac{1}{2}}-E^{-\frac{1}{2}}$	$\frac{\Delta}{(1+\Delta)^{\frac{1}{2}}}$	δ	$\frac{\nabla}{(1-\nabla)^{\frac{1}{2}}}$	$2\,\mathrm{senh}\,\frac{h}{2}D$
∇	$1-E^{-1}$	$\frac{\Delta}{1+\Delta}$	$\delta(1+\frac{1}{4}\delta^2)^{\frac{1}{2}}-\frac{1}{2}\delta^2$	∇	$1-e^{-hD}$
D	$\frac{1}{h}\ln E$	$\frac{1}{h}\ln(1+\Delta)$	$\frac{2\,\mathrm{arg\,senh}(\frac{1}{2}\delta)}{h}$	$-\frac{\ln(1-\nabla)}{h}$	D
μ	$\frac{1}{2}(E^{\frac{1}{2}}+E^{-\frac{1}{2}})$	$\frac{1+\frac{1}{2}\Delta}{(1+\Delta)^{\frac{1}{2}}}$	$(1+\frac{1}{4}\delta^2)^{\frac{1}{2}}$	$\frac{1-\frac{1}{2}\nabla}{(1-\nabla)^{\frac{1}{2}}}$	$\cosh(\frac{h}{2}D)$

Tabla 4.2: Relaciones entre operadores.

Por ejemplo, $\mu = (1+\frac{1}{4}\delta^2)^{\frac{1}{2}}$, $D = \frac{2\,\mathrm{arg\,senh}(\frac{1}{2}\delta)}{h}$, etc.

7. Se tiene la siguiente expresión para el operador J:

$$J = h\Delta(\ln(1+\Delta))^{-1} = h(1+\frac{1}{2}\Delta-\frac{1}{12}\Delta^2+\frac{1}{24}\Delta^3-\frac{19}{720}\Delta^4+\cdots) \ .$$

Notemos que, a partir de $JD=\Delta$ y $D=\frac{1}{h}\ln(1+\Delta)$, resulta la relación $J\ln(1+\Delta) = h\Delta$ de la que se desprende la primera igualdad dada. Para calcular el desarrollo de J en función de Δ, se escribe

$$J = h\sum_{j=0}^{\infty} c_j\Delta^j$$

y se impone la ecuación satisfecha por J, $J\ln(1+\Delta) = h\Delta$, obteniendo

$$h(c_0+c_1\Delta+c_2\Delta^2+c_3\Delta^3+\ldots)(\Delta-\frac{1}{2}\Delta^2+\frac{1}{3}\Delta^3-\ldots) = h\Delta \ ;$$

de donde, se pueden determinar los sucesivos coeficientes

$$c_0 = 1 \ , \quad c_1 = \frac{1}{2} \ , \quad c_2 = -\frac{1}{12} \ , \quad c_3 = \frac{1}{24} \ , \quad \ldots$$

4.5.3 Aplicaciones del cálculo con operadores

Generamos ahora de manera sistemática toda una serie de fórmulas por medio de las relaciones establecidas entre los operadores definidos. Estas fórmulas, que utilizarán siempre abscisas equidistantes, serán exactas sólo para polinomios; pero, para otras funciones, además de ser aproximadas, darán a menudo un desarrollo asintótico del error cometido.

Interpolación

A partir de la relación $E^t = (1+\Delta)^t$ para $t \in \mathbb{R}$, se obtiene

$$f(x_0+th) = \sum_{j=0}^{\infty} \left(\begin{array}{c} t \\ j \end{array} \right) (\Delta^j f)(x_0) \ , \quad t, x_0 \in \mathbb{R} \ ,$$

exacta para cualquier polinomio $f \in \mathcal{P}$, siendo además la suma finita.

Si fijamos $m \geq 0$, obtenemos una fórmula exacta para polinomios f de grado menor o igual que m

$$f(x_0+th) = \sum_{j=0}^{m} \frac{t(t-1)\cdots(t-j+1)}{j!} (\Delta^j f)(x_0) \ , \quad t, x_0 \in \mathbb{R} \ ;$$

de donde, haciendo $x = x_0 + th$ ($t = \frac{x-x_0}{h}$), $\Delta^j f_0 = (\Delta^j f)(x_0)$ ($j = 0 \div m$), $x_k = x_0 + kh$, $f_k = f(x_k)$ ($k = 0 \div m$), se deduce la fórmula de interpolación siguiente que aproxima el valor de $f(x)$, suponiendo conocidos f_0, f_1, ..., f_m:

$$f(x) \simeq p_m(x) = \sum_{j=0}^{m} \frac{\Delta^j f_0}{j! h^j} (x-x_0)\cdots(x-x_{j-1}) \ ,$$

llamada *fórmula de interpolación de Newton hacia adelante*, exacta para polinomios de grado menor o igual que m.

Como ya se ha visto,

$$\frac{\Delta^j f_0}{j! h^j} = f[x_0, x_1, \ldots, x_j] \ ;$$

por lo tanto, $p_m(x)$ es el polinomio de grado menor o igual que m que interpola a f en las abscisas x_0, $x_1 = x_0 + h$, ..., $x_m = x_0 + mh$.

Igualmente, usando que $E^{-t} = (1-\nabla)^t$, obtenemos la relación

$$f(x) \simeq p_m(x) = \sum_{j=0}^{m} \frac{\nabla^j f_0}{j! h^j} (x-x_0)\cdots(x-x_{-(j-1)}) \ ,$$

llamada *fórmula de interpolación de Newton hacia atrás*, exacta para polinomios de grado menor o igual que m; en este caso, el polinomio $p_m(x)$ interpola a f en las abscisas x_0, x_{-1}, ..., x_{-m}, ya que

$$\frac{\nabla^j f_0}{j! h^j} = f[x_{-j}, x_{-(j-1)}, \ldots, x_0] \ .$$

Otras fórmulas de interpolación, para particiones centradas, se encuentran en el problema 4.14.

Derivación numérica

A partir de

$$\begin{aligned} D &= \frac{1}{h}\ln(1+\Delta) = -\frac{1}{h}\ln(1-\nabla) \\ &= \frac{2}{h}\operatorname{arg\,senh}(\frac{\delta}{2}) = \frac{2}{h}\mu\left(1+\frac{\delta^2}{4}\right)^{-\frac{1}{2}} \operatorname{arg\,senh}(\frac{\delta}{2}) \ , \end{aligned}$$

se obtienen las fórmulas de derivación numérica siguientes, exactas para los polinomios de grado menor o igual que m, si cortamos después de los términos en Δ^m, δ^m y ∇^m, respectivamente:

$$\begin{aligned}
f_0' &\simeq \frac{1}{h}(\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \cdots) \ , \\
f_0' &\simeq \frac{1}{h}(\nabla f_0 + \frac{1}{2}\nabla^2 f_0 + \frac{1}{3}\nabla^3 f_0 + \cdots) \ , \\
f_0' &\simeq \frac{1}{h}(\delta f_0 - \frac{1}{2}\frac{1}{2^2\cdot 3}\delta^3 f_0 + \frac{1\cdot 3}{2\cdot 4}\frac{1}{2^4\cdot 5}\delta^5 f_0 - \cdots) \ , \\
f_0' &\simeq \frac{1}{h}(\mu\delta f_0 - \frac{1}{6}\mu\delta^3 f_0 + \frac{1}{30}\mu\delta^5 f_0 - \cdots) \\
&= \frac{1}{2h}[(f_1 - f_{-1}) - \frac{1}{6}(\delta^2 f_1 - \delta^2 f_{-1}) \\
&\quad + \frac{1}{30}(\delta^4 f_1 - \delta^4 f_{-1}) - \cdots] \ ;
\end{aligned}$$

donde $f_0' \equiv f'(x_0)$.

La primera es una fórmula de derivación hacia adelante y la segunda, hacia atrás. Las dos últimas son fórmulas centradas; la última tiene, además, la ventaja de no hacer intervenir valores de la función en abscisas intermedias de la partición equidistante con paso h: por ejemplo, la tercera contiene $\delta f_0 = f_{\frac{1}{2}} - f_{-\frac{1}{2}}$ y requiere valores que no aparecen en una tabla de f hecha con paso h.

Para calcular derivadas de orden superior, es muy frecuente usar las relaciones:

$$\begin{aligned}
D^{2r} &= \left[\frac{2}{h}\operatorname{arg\,senh}(\frac{\delta}{2})\right]^{2r} \ , \\
D^{2r+1} &= \mu\left(1 + \frac{1}{4}\delta^2\right)^{-\frac{1}{2}}\left[\frac{2}{h}\operatorname{arg\,senh}(\frac{\delta}{2})\right]^{2r+1} \ .
\end{aligned}$$

Así, por ejemplo:

$$\begin{aligned}
f_0^{(2)} &\simeq \frac{1}{h^2}(\delta^2 f_0 - \frac{1}{12}\delta^4 f_0 + \frac{1}{90}\delta^6 f_0 - \cdots) \ , \\
f_0^{(3)} &\simeq \frac{1}{h^3}(\mu\delta^3 f_0 - \frac{1}{4}\mu\delta^5 f_0 + \frac{7}{120}\mu\delta^7 f_0 - \cdots) \ , \\
f_0^{(4)} &\simeq \frac{1}{h^4}(\delta^4 f_0 - \frac{1}{6}\delta^6 f_0 + \frac{7}{240}\delta^8 f_0 - \cdots) \ , \\
f_0^{(5)} &\simeq \frac{1}{h^5}(\mu\delta^5 f_0 - \frac{1}{3}\mu\delta^7 f_0 + \cdots) \ .
\end{aligned}$$

Nótese finalmente que todas las expresiones asintóticas obtenidas permiten la aplicación del método de Richardson (véase el problema 4.15).

Integración numérica

A partir de

$$(E^{m-1}Jf)(x) = (JE^{m-1}f)(x) = \int_{x+(m-1)h}^{x+mh} f(t)dt \ ,$$

resulta

$$\begin{aligned} J_m f_0 &\equiv \int_{x_0}^{x_0+mh} f(t)dt = [(1+E+\cdots+E^{m-1})J]f_0 \\ &= [(E^m-1)(E-1)^{-1}J]f_0) \ , \end{aligned}$$

para cualquier polinomio $f \in \mathcal{P}$. Desarrollando en potencias de Δ, se obtiene

$$\begin{aligned} J_m f_0 &= h[((1+\Delta)^m-1)\Delta^{-1}][\Delta(\ln(1+\Delta))^{-1}]f_0 \\ &= h\left[\Delta^{m-1}+m\Delta^{m-2}+\binom{m}{2}\Delta^{m-3}+\cdots\right. \\ &\quad \left.+\binom{m}{m-2}\Delta+\binom{m}{m-1}\right][c_0+c_1\Delta+c_2\Delta^2+\cdots]f_0 \\ &= h\left[d_0^{(m)}+d_1^{(m)}\Delta+d_2^{(m)}\Delta^2+d_3^{(m)}\Delta^3+\cdots\right]f_0 \ , \end{aligned}$$

donde

$$d_j^{(m)} = \sum_{l=0}^{\min(j,m-1)} \binom{m}{m-l-1} c_{j-l} = \sum_{l=0}^{\min(j,m-1)} \binom{m}{l+1} c_{j-l} \ ;$$

pueden calcularse estos coeficientes, usando la propiedad 7 de los operadores

$$\Delta[\ln(1+\Delta)]^{-1} = c_0 + c_1\Delta + c_2\Delta^2 + \cdots \ ,$$

con $c_0 = 1$, $c_1 = \frac{1}{2}$, $c_2 = -\frac{1}{12}$, $c_3 = \frac{1}{24}$, $c_4 = -\frac{19}{720}$, $c_5 = \frac{3}{160}$, ...

En particular, en los casos $m = 1, 2$, obtenemos:

$$\begin{aligned} J_1 f_0 = \int_{x_0}^{x_0+h} f(t)dt &\simeq h(f_0 + \frac{1}{2}\Delta f_0 - \frac{1}{12}\Delta^2 f_0 \\ &\quad + \frac{1}{24}\Delta^3 f_0 - \frac{19}{720}\Delta^4 f_0 + \cdots) \ , \\ J_2 f_0 = \int_{x_0}^{x_0+2h} f(t)dt &\simeq 2h(f_0 + \Delta f_0 + \frac{1}{6}\Delta^2 f_0 \\ &\quad - \frac{1}{180}\Delta^4 f_0 + \frac{1}{180}\Delta^5 f_0 + \cdots) \ . \end{aligned}$$

Las fórmulas de integración numérica que se han deducido para las integrales $J_m f_0$ ($m \geq 1$) son exactas para los polinomios de grado menor o igual que n si cortamos después del término en $\Delta^n f_0$. Si $n = m$, las fórmulas obtenidas dependen linealmente sólo de los valores f_0, f_1, ..., f_m, y son, por lo tanto, las llamadas fórmulas de Newton-Cotes de orden m.

Por ejemplo, si $m = 1, 2$, se tiene:

$$\begin{aligned} \int_{x_0}^{x_0+h} f(t)dt &\simeq \frac{h}{2}(f_0+f_1) - \frac{h}{12}(\Delta^2 f_0 - \frac{1}{2}\Delta^3 f_0 + \cdots) \ , \\ \int_{x_0}^{x_0+2h} f(t)dt &\simeq \frac{h}{3}(f_0+4f_1+f_2) - \frac{h}{90}(\Delta^4 f_0 - \Delta^5 f_0 + \cdots) \ . \end{aligned}$$

Estas fórmulas son las del trapecio y de Simpson con términos correctores, respectivamente. Obsérvese también que, substituyendo $\Delta^2 f_0$ y $\Delta^4 f_0$, por $h^2 f^{(2)}(\xi)$ y $h^4 f^{(4)}(\xi)$ en los primeros términos correctores, volvemos a encontrar la expresión del error de integración numérica de ambas fórmulas.

La fórmula de Euler-Maclaurin se deduce en el problema 4.16, usando también técnicas de cálculo con operadores.

COMENTARIOS BIBLIOGRÁFICOS

La presentación hecha del polinomio interpolador, herramienta fundamental para la derivación e integración numéricas, es similar a la que se encuentra en [Hen64]. Una buena colección de fórmulas de derivación e integración numéricas se encuentran en [AS65]. La referencia [IK66] contiene las expresiones del error en la derivación numérica y en las fórmulas (cerradas) de Newton-Cotes presentadas, así como las correspondientes a otras fórmulas de Newton-Cotes, llamadas abiertas; el teorema del valor medio para sumas y la convergencia de las reglas de integración numérica también se encuentran allí. Más completo es [DR67] que abarca una parte muy importante de métodos de integración numérica, entre los cuales destacamos los de *integración adaptable*, que proporcionan diversas estrategias dependiendo del comportamiento de la función que se integra en diferentes zonas del intervalo de integración (puede consultarse también [RR78]). Para técnicas de *integración múltiple*, véase [Str71] y, para la integración gaussiana, [SS66]. Los criterios de acotación de los restos de las series se encuentran en los libros estándar de análisis, por ejemplo [Apo57]. Otros métodos de sumación, diferentes de los presentados, pueden encontrarse en [DB74], [DM73] y [RR78]. La fórmula de Euler-Maclaurin, con la notación quizás más estándar de los números de Bernoulli, se encuentra en [SB80], que también trata con detalle la extrapolación en el marco de la integración numérica. Otras referencias sobre extrapolación son [DB74], [Hen64] y [RR78]. La generación, con ayuda de operadores lineales, de fórmulas de interpolación, derivación e integración numéricas, es bastante clásica y se encuentra, por ejemplo, en [DB74], [Fro69], [Hil74] y [Sch67]. Este último aporta, además, muchos ejercicios resueltos sobre otros temas tratados en este libro.

PROBLEMAS RESUELTOS

Problema 4.1 *a) Hallar una fórmula de derivación interpolatoria para el cálculo aproximado de $f'(a)$, por derivación del polinomio interpolador en las abscisas $a+h$, $a+\frac{h}{2}$, $a+\frac{h}{4}$ y a.*

b) Suponiendo que $f \in \mathcal{C}^4([a, a+h])$, dar una expresión para el error cometido.

c) Calcular la derivada de la función $f(x) = \cosh x^2$ en $x = a = 1$, usando la fórmula hallada en a) con $h = 0.1$ y acotar el error cometido.

d) Comparar el error exacto con la cota de error hallada.

SOLUCIÓN:

a) Obtenemos el polinomio interpolador a partir de la tabla de diferencias divididas

$$\begin{array}{l|llll} x_0 = a & f_0 & & & \\ & & \frac{4(f_1-f_0)}{h} & & \\ x_1 = a+\frac{h}{4} & f_1 & & \frac{8(f_2-2f_1+f_0)}{h^2} & \\ & & \frac{4(f_2-f_1)}{h} & & \frac{8(f_3-6f_2+8f_1-3f_0)}{3h^3} \\ x_2 = a+\frac{h}{2} & f_2 & & \frac{8(f_3-3f_2+2f_1)}{3h^2} & \\ & & \frac{2(f_3-f_2)}{h} & & \\ x_3 = a+h & f_3 & & & \end{array}$$

El polinomio interpolador que resulta es

$$\begin{aligned} p_3(x) &= f[x_0] + f[x_0,x_1](x-a) + f[x_0,x_1,x_2](x-a)(x-a-\frac{h}{4}) \\ &\quad + f[x_0,x_1,x_2,x_3](x-a)(x-a-\frac{h}{4})(x-a-\frac{h}{2}) \ . \end{aligned}$$

La derivada de este polinomio en $x = a$ es la aproximación pedida de $f'(a)$

$$\begin{aligned} p_3'(a) &= f[x_0,x_1] + f[x_0,x_1,x_2](-\frac{h}{4}) + f[x_0,x_1,x_2,x_3](-\frac{h}{4})(-\frac{h}{2}) \\ &= f[x_0,x_1] - \frac{h}{4} f[x_0,x_1,x_2] + \frac{h^2}{8} f[x_0,x_1,x_2,x_3] \\ &= \frac{4}{h}(f_1 - f_0) - \frac{h}{4}\frac{8}{h^2}(f_2 - 2f_1 + f_0) \\ &\quad + \frac{h^2}{8}\frac{8}{3h^3}(f_3 - 6f_2 + 8f_1 - 3f_0) \\ &= \frac{f_3 - 12f_2 + 32f_1 - 21f_0}{3h} \ . \end{aligned}$$

Así, queda

$$f'(a) \simeq \frac{1}{3h}[f(a+h) - 12f(a+\frac{h}{2}) + 32f(a+\frac{h}{4}) - 21f(a)] \ .$$

b) El error en la interpolación se expresa como

$$f(x) - p_3(x) = \frac{f^{(4)}(\xi(x))}{4!}(x-a)(x-a-\frac{h}{4})(x-a-\frac{h}{2})(x-a-h) \ ,$$

y el error de la fórmula de derivación hallada,

$$e_3'(a) = f'(a) - p_3'(a) = \frac{f^{(4)}(\xi)}{4!}(-\frac{h}{4})(-\frac{h}{2})(-h) = -\frac{f^{(4)}(\xi)}{192}h^3 \ ,$$

donde $\xi \in < a, a+h >$.

c) Tomando $h = 0.1$, resulta

$$\begin{aligned} f'(1) &\simeq \frac{1}{3 \cdot 0.1}[f(1.1) - 12f(1.05) + 32f(1.025) - 21f(1)] \\ &\simeq 2.351006625 \ . \end{aligned}$$

Para hallar una cota del error es necesario acotar primero la derivada cuarta de f en el intervalo $[1, 1.1]$; tenemos

$$\begin{aligned} f'(x) &= 2x \operatorname{senh} x^2 \ , \\ f^{(2)}(x) &= 4x^2 \cosh x^2 + 2 \operatorname{senh} x^2 \ , \\ f^{(3)}(x) &= 8x^3 \operatorname{senh} x^2 + 12x \cosh x^2 \ , \\ f^{(4)}(x) &= (16x^4 + 12) \cosh x^2 + 48x^2 \operatorname{senh} x^2 \ . \end{aligned}$$

Dado que $f^{(4)}$ es una función creciente y positiva, podemos tomar como cota de ésta su valor en $x = 1.1$ que resulta ser $M_4 = 153.408$.

La cota hallada para el error es entonces

$$\frac{M_4}{192}10^{-3} \simeq 0.8 \cdot 10^{-3} \ .$$

d) Usando directamente la expresión de la derivada de f en $x = a = 1$, tenemos

$$f'(a) = 2a \operatorname{senh} a^2 = 2 \operatorname{senh} 1 = 2.3504023873 \ ;$$

el error real es, pues, $0.60424 \cdot 10^{-3}$, no mucho menor que la cota hallada en c).

Problema 4.2 *Disponemos de los siguientes datos sobre una función f:*

x	0.4	0.5
$f(x)$	1.554284	1.561136
$f'(x)$	0.243031	−0.089618

.

a) Hallar la abscisa del máximo de f en $[0.4, 0.5]$, aproximándola por el máximo del polinomio interpolador de Hermite $p_3(x)$ a la tabla dada de f.

b) Suponiendo que $f \in \mathcal{C}^4([x_0, x_1])$, hallar la siguiente expresión para la derivada e_3' del error en la interpolación de Hermite en dos abscisas $x_0 < x_1$:

$$e_3'(x) \equiv f'(x) - p_3'(x) = \frac{f^{(4)}(\eta(x))}{3!}(x - x_0)(x - x_1)(x - \xi) \ ,$$

donde $\xi \in (x_0, x_1)$ y $\eta(x) \in < x_0, x_1, x >$.

c) Acotar el error en la abscisa del máximo debido a la interpolación, sabiendo que que $|f^{(4)}(x)| < 10^3$ i $|f^{(2)}(x)| > 1$ $\forall x \in (0.4, 0.5)$.

SOLUCIÓN:

a) El polinomio interpolador de Hermite se halla a partir de la siguiente tabla de diferencias divididas generalizadas, donde $f[x_i, x_i] \equiv f'(x_i)$:

$$\begin{array}{c|cccc} 0.4 & 1.554284 & & & \\ & & 0.243031 & & \\ 0.4 & 1.554284 & & -1.7451 & \\ & & 0.068520 & & 1.637 \\ 0.5 & 1.561136 & & -1.5814 & \\ & & -0.089618 & & \\ 0.5 & 1.561136 & & & \end{array} \ ,$$

lo escribimos así:

$$\begin{aligned} p_3(x) &= d + c(x - x_0) + b(x - x_0)^2 + a(x - x_0)^2(x - x_1) \\ &= d + c\delta + b\delta^2 - a(x_1 - x_0)\delta^2 + a\delta^3 \ , \end{aligned}$$

con $a = 1.637$, $b = -1.7451$, $c = 0.243031$ y $d = 1.554284$, haciendo $\delta = x - x_0$. El máximo se da en el cero del polinomio derivado en el intervalo $(0.4, 0.5)$. La derivada del polinomio resulta ser

$$p_3'(x) = c - 2[a(x_1 - x_0) - b]\delta + 3a\delta^2 = C - 2B\delta + A\delta^2 \ ,$$

con $C = c = 0.243031$, $B = a(x_1 - x_0) - b = 1.9088$ y $A = 3a = 4.911$. De los dos ceros de este polinomio

$$\delta = \frac{B \pm \sqrt{B^2 - AC}}{A} = \frac{C}{B \mp \sqrt{B^2 - AC}}$$

elegimos el menor, porque $f'(x_0) > 0$ y $f'(x_1) < 0$; entonces el máximo buscado será

$$x_M = x_0 + \frac{C}{B + \sqrt{B^2 - AC}} = 0.4700 \ .$$

b) El error de interpolación e_3 y su derivada e_3' se anulan en las abscisas de interpolación x_0 y x_1. Aplicando el teorema de Rolle, existe $\xi \in (x_0, x_1)$ tal que $e_3'(\xi) = 0$. La fórmula propuesta para $e_3'(x)$ es claramente cierta para $x = x_0, x_1, \xi$ y cualquier valor de η escogido.

Para los demás valores de x, definimos para $z \in [x_0, x_1]$ la función

$$F(z) = e_3'(z) - a(x)(z - x_0)(z - x_1)(z - \xi) \ ,$$

donde

$$a(x) = \frac{e_3'(x)}{(x-x_0)(x-x_1)(x-\xi)} \ .$$

Es claro que $F \in \mathcal{C}^3([x_0, x_1])$ y se anula en las cuatro abscisas diferentes x, x_0, x_1, ξ. Aplicando sucesivamente el teorema de Rolle a F, F' y $F^{(2)}$, existe $\eta(x) \in < x_0, x_1, x >$ tal que $F^{(3)}(\eta(x)) = 0$; es decir,

$$f^{(4)}(\eta(x)) - 3!a(x) = 0$$

y, por lo tanto,

$$0 = F(x) = e_3'(x) - \frac{f^{(4)}(\eta(x))}{3!}(x-x_0)(x-x_1)(x-\xi) \ .$$

Acotando ahora $|e_3'(x)|$, obtenemos, para $x \in [x_0, x_1]$,

$$|e_3'(x)| \leq \frac{M_4}{6} |(x-x_0)(x-x_1)(x_1-x_0)| \leq \frac{M_4}{24}(x_1-x_0)^3 \ ,$$

con $M_4 = \max_{\eta \in [x_0,x_1]} \left| f^{(4)}(\eta) \right|$.

Por ejemplo, en las condiciones del apartado c), $|e_3'(x)| < \frac{1}{24} \equiv \epsilon$. Es decir, en todo el intervalo $(0.4, 0.5)$, la derivada se evalúa con un error acotado por ϵ.

c) Sean $\bar{x}_M$ y x_M los máximos de f y $p_3(x)$, respectivamente, en (x_0, x_1). Entonces, por el teorema del valor medio,

$$e_3'(x_M) = f'(x_M) = f^{(2)}(\zeta)(x_M - \bar{x}_M) \ , \quad \zeta \in < x_M, \bar{x}_M > \ .$$

Por lo tanto, si hacemos

$$m_2 = \min_{\zeta \in [x_0,x_1]} \left| f^{(2)}(\zeta) \right| \ ,$$

obtenemos la cota pedida

$$\begin{aligned} |x_M - \bar{x}_M| &\leq \frac{|e'(x_M)|}{m_2} \leq \frac{M_4}{6m_2} |(x_M-x_0)(x_M-x_1)(x_M-\xi)| \\ &< \frac{10^3}{6}(0.07)(0.03)(0.07) \simeq 2.5 \cdot 10^{-2} \ . \end{aligned}$$

Problema 4.3 *a) Determinar la fórmula de integración numérica*

$$\int_{-1}^{1} g(t)dt \simeq \sum_{i=0}^{1} W_i g(t_i)$$

que es exacta para todos los polinomios de grado menor o igual que 3.

b) Si $g \in \mathcal{C}^4([-1,1])$, hallar una expresión para el error cometido.

SOLUCIÓN:

a) Han de determinarse cuatro parámetros: las abscisas $t_0, t_1 \in [-1,1]$ y los pesos W_0, W_1.

Debido a la linealidad de la fórmula respecto a g, el hecho de imponer exactitud en $\mathcal{P}_3$ es equivalente a hacerlo para la base de polinomios $\{1, t, t^2, t^3\}$:

$$\begin{array}{l|lcllcl} g(t)=1 & 2 & = & W_0 & + & W_1 \\ g(t)=t & 0 & = & W_0 t_0 & + & W_1 t_1 \\ g(t)=t^2 & \frac{2}{3} & = & W_0 t_0^2 & + & W_1 t_1^2 \\ g(t)=t^3 & 0 & = & W_0 t_1^3 & + & W_1 t_1^3 \end{array} \quad .$$

Este sistema no lineal es sencillo de resolver; resulta:

$$t_0 = -\frac{\sqrt{3}}{3} \ , \quad t_1 = \frac{\sqrt{3}}{3} \ ; \quad W_0 = W_1 = 1 \ .$$

La fórmula de integración numérica buscada es, pues,

$$\int_{-1}^{1} g(t)dt \simeq g\left(-\frac{\sqrt{3}}{3}\right) + g\left(\frac{\sqrt{3}}{3}\right) \ .$$

b) Esta fórmula es exacta para todos los polinomios de grado menor o igual que 3, pero no lo es para todos los de grado 4. En efecto, tomando

$$g(t) = (t-t_0)^2(t-t_1)^2 = \left(t^2 - \frac{1}{3}\right)^2 = t^4 - \frac{2}{3}t^2 + \frac{1}{9} \ ,$$

tenemos

$$\int_{-1}^{1} (t-t_0)^2(t-t_1)^2 dt = \frac{8}{45}$$

y la aproximación daría cero, ya que $g(t_i) = 0 \ (i = 0, 1)$.

Suponiendo que $g \in \mathcal{C}^2([-1,1])$, consideremos el polinomio $p_3(x)$ de interpolación de Hermite a f en las abscisas t_0, t_1. De la fórmula del error en la interpolación de Hermite se tiene

$$g(t) - p_3(t) = \frac{g^{(4)}(\xi(t))}{4!}(t-t_0)^2(t-t_1)^2 \ , \quad \xi(t) \in < t_0, t_1, t > \ .$$

Debido a que el polinomio de interpolación de Hermite se integra exactamente (al ser de grado menor o igual que 3), hallamos el error de integración numérica usando el teorema del valor medio para integrales:

$$\begin{aligned} E_3 &\equiv \int_1^1 g(t)dt - g(-\frac{\sqrt{3}}{3}) - g(\frac{\sqrt{3}}{3}) \\ &= \int_{-1}^{1} \frac{g^{(4)}(\xi(t))}{4!}(t-t_0)^2(t-t_1)^2 dt \\ &= \frac{g^{(4)}(\xi)}{4!}\int_{-1}^{1}(t-t_0)^2(t-t_1)^2 dt \\ &= \frac{1}{135}g^{(4)}(\xi) \ , \ \xi \in (-1,1) \ . \end{aligned}$$

Nótese que no hemos hecho más que redescubrir, usando únicamente métodos elementales, la fórmula de Gauss-Legendre de dos abscisas que aparece deducida, en general, en el problema 4.8.

Problema 4.4 *Queremos hallar fórmulas de integración en $m+1$ abscisas equidistantes de la forma*

$$\int_a^b f(x)dx \simeq \sum_{k=0}^{m}[A_k f(x_k) + B_k f'(x_k)] \ ,$$

exactas para todos los polinomios de grado menor o igual que $2m+1$.

a) Demostrar que dichas fórmulas pueden hallarse integrando el polinomio interpolador de Hermite de f en las $m+1$ abscisas de la fórmula.

b) Explicitar estas fórmulas para $m = 1, 2$, dando asimismo una expresión para los errores cometidos.

c) Aplicación: Calcular

$$\int_0^1 e^{e^x} dx \ ,$$

acotando el error cometido, usando todas las fórmulas halladas en b).

SOLUCIÓN:

a) Los coeficientes de estas fórmulas pueden hallarse por integración del polinomio interpolador de Hermite, debido a que la fórmula ha de ser exacta para este polinomio, que tiene grado menor o igual que $2m+1$ y

$$p_{2m+1}(x_k) = f(x_k) \ , \quad p'_{2m+1}(x_k) = f'(x_k) \quad (k = 0 \div m) \ .$$

Así resulta,

$$\begin{aligned} \int_a^b f(x)dx &\simeq \sum_{k=0}^{m}[A_k f(x_k) + B_k f'(x_k)] \\ &= \sum_{k=0}^{m}[A_k p_{2m+1}(x_k) + B_k p'_{2m+1}(x_k)] \\ &= \int_a^b p_{2m+1}(x)dx \ . \end{aligned}$$

b) Hallaremos primero las fórmulas usando la expresión del polinomio interpolador de Hermite en función de los polinomios básicos

$$\begin{aligned} \Phi_k(x) &= (1 - 2l'_k(x_k)(x - x_k))l_k^2(x) \ , \\ \Psi_k(x) &= (x - x_k)l_k^2(x) \quad (k = 0 \div m) \ . \end{aligned}$$

Recuérdese que estos polinomios cumplen

$$\Phi_k(x_i) = \delta_{ki} \ , \quad \Phi'_k(x_i) = 0 \quad (k, i = 0 \div m) \ ;$$

$$\Psi_k(x_i) = 0 \ , \quad \Psi'_k(x_i) = \delta_{ki} \quad (k, i = 0 \div m) \ ;$$

y que el polinomio interpolador de Hermite se expresa como

$$p_{2m+1}(x) = \sum_{k=0}^{m} f(x_k)\Phi_k(x) + \sum_{k=0}^{m} f'(x_k)\Psi_k(x) \ .$$

Substituyendo esta expresión en el integrando, hallamos

$$\begin{aligned} A_k &= \int_a^b \Phi_k(x)dx = \int_a^b (1 - 2l'_k(x_k)(x - x_k))l_k^2(x)dx \\ &= h\int_0^m (1 - 2\frac{d\ell_k}{ds}(k)(s-k))\ell_k^2(s)ds \equiv h\alpha_k^{(m)} \ , \\ B_k &= \int_a^b \Psi_k(x)dx = \int_a^b (x - x_k)l_k^2(x)dx \\ &= h^2\int_0^m (s-k)\ell_k^2(s)ds \equiv h^2\beta_k^{(m)} \\ &\quad (k = 0 \div m) \ , \end{aligned}$$

donde

$$\ell_k(s) = \prod_{i\neq k} \frac{s-i}{k-i} \ , \quad \frac{d\ell_k}{ds}(k) = \sum_{i\neq k} \frac{1}{k-i} \ .$$

Conviene remarcar que los coeficientes $\alpha_k^{(m)}$ y $\beta_k^{(m)}$ $(k = 0 \div m)$ dependen sólo de k y de m, pero no dependen ni de los intervalos, ni de los pasos elegidos.

La expresión del error en estas fórmulas se halla por integración de la expresión del error en la interpolación de Hermite

$$\int_a^b e_{2m+1}(x)dx = \int_a^b \frac{f^{(2m+2)}(\xi(x))}{(2m+2)!}\omega_m^2(x)dx \ ;$$

aplicando el teorema del valor medio para integrales, resulta

$$\begin{aligned} E_{2m+1} &= \int_a^b e_{2m+1}(x)dx = \frac{f^{(2m+2)}(\xi)}{(2m+2)!}\int_a^b \omega_m^2(x)dx \\ &= \frac{f^{(2m+2)}(\xi)}{(2m+2)!}h^{2m+3}\int_0^m \prod_{i=0}^{m}(s-i)^2 ds \\ &= f^{(2m+2)}(\xi)h^{2m+3}\delta_m \ , \quad \xi \in (a,b) \ . \end{aligned}$$

Nótese también que el factor δ_m sólo depende de m.

Calculamos a continuación los coeficientes $\alpha_k^{(m)}$ y $\beta_k^{(m)}$ $(k = 0 \div m)$, así como el factor δ_m del error, para $m = 1, 2$.

En el caso $m = 1$, tenemos:

$$\begin{aligned}
\alpha_0^{(1)} &= \int_0^1 (1 + 2(s-0))\frac{(s-1)^2}{(0-1)^2}ds = \frac{1}{2} \ , \\
\alpha_1^{(1)} &= \int_0^1 (1 - 2(s-1))\frac{(s-0)^2}{(1-0)^2}ds = \frac{1}{2} \ , \\
\beta_0^{(1)} &= \int_0^1 (s-0)\frac{(s-1)^2}{(0-1)^2}ds = \frac{1}{12} \ , \\
\beta_1^{(1)} &= \int_0^1 (s-1)\frac{(s-0)^2}{(1-0)^2}ds = -\frac{1}{12} \ , \\
\delta_1 &= \frac{1}{24}\int_0^1 s^2(s-1)^2 ds = \frac{1}{720} \ .
\end{aligned}$$

En el caso $m = 2$, tenemos:

$$\begin{aligned}
\alpha_0^{(2)} &= \int_0^2 (1 - 2(-\frac{1}{2} - 1)(s-0))\frac{(s-1)^2(s-2)^2}{(0-1)^2(0-2)^2}ds = \frac{7}{15} \ , \\
\alpha_1^{(2)} &= \int_0^2 (1 - 2(1-1)(s-1))\frac{(s-0)^2(s-2)^2}{(1-0)^2(1-2)^2}ds = \frac{16}{15} \ , \\
\alpha_2^{(2)} &= \int_0^2 (1 - 2(\frac{1}{2} + 1)(s-2))\frac{(s-0)^2(s-1)^2}{(2-0)^2(2-1)^2}ds = \frac{7}{15} \ , \\
\beta_0^{(2)} &= \int_0^2 (s-0)\frac{(s-1)^2(s-2)^2}{(0-1)^2(0-2)^2}ds = \frac{1}{15} \ , \\
\beta_1^{(2)} &= \int_0^2 (s-1)\frac{(s-0)^2(s-2)^2}{(1-0)^2(1-2)^2}ds = 0 \ , \\
\beta_2^{(2)} &= \int_0^2 (s-2)\frac{(s-0)^2(s-1)^2}{(2-0)^2(2-1)^2}ds = -\frac{1}{15} \ , \\
\delta_2 &= \frac{1}{720}\int_0^2 s^2(s-1)^2(s-2)^2 ds = \frac{1}{4725} \ .
\end{aligned}$$

Las fórmulas buscadas son, pues:

$$\begin{aligned}
\int_{x_0}^{x_1} f(x)dx &\simeq \frac{h}{2}[f(x_0) + f(x_1)] \\
&\quad + \frac{h^2}{12}[f'(x_0) - f'(x_1)] \\
&\quad + \frac{h^5}{720}f^{(4)}(\xi) \ , \quad \xi \in (x_0, x_1) \ ; \\
\int_{x_0}^{x_2} f(x)dx &\simeq \frac{h}{15}[7f(x_0) + 16f(x_1) + 7f(x_2)] + \frac{h^2}{15}[f'(x_0) - f'(x_2)] \\
&\quad + \frac{h^7}{4725}f^{(6)}(\xi) \ , \ \xi \in (x_0, x_2) \ .
\end{aligned}$$

c) Para la función $f(x) = e^{e^x}$ en $[0, 1]$, estas fórmulas de integración ofrecen, respectivamente, las aproximaciones:

$$\int_0^1 e^{e^x} dx \simeq \frac{1}{2}(e + e^e) + \frac{1}{12}(e - e^e e) = 5.73 \ ,$$

$$\int_0^1 e^{e^x} dx \simeq \frac{1}{30}(7e + 16e^{\sqrt{e}} + 7e^e) + \frac{1}{60}(e - e^e e) = 6.3025 \ .$$

Los errores cometidos pueden acotarse teniendo en cuenta las expresiones siguientes de las derivadas de f:

$$\begin{aligned}
f'(x) &= e^{e^x} e^x \ , \\
f^{(2)}(x) &= e^{e^x}(e^{2x} + e^x) \ , \\
f^{(3)}(x) &= e^{e^x}(e^{3x} + 3e^{2x} + e^x) \ , \\
f^{(4)}(x) &= e^{e^x}(e^{4x} + 6e^{3x} + 7e^{2x} + e^x) \ , \\
f^{(5)}(x) &= e^{e^x}(e^{5x} + 10e^{4x} + 25e^{3x} + 15e^{2x} + e^x) \ , \\
f^{(6)}(x) &= e^{e^x}(e^{6x} + 15e^{5x} + 65e^{4x} + 90e^{3x} + 31e^{2x} + e^x) \ .
\end{aligned}$$

Las derivadas cuarta y sexta están acotadas, pues, por:

$$M_4 = e^e(e^4 + 6e^3 + 7e^2 + e) < 3479 \ ,$$

$$M_6 = e^e(e^6 + 15e^5 + 65e^4 + 90e^3 + 31e^2 + e) < 124538 \ ,$$

respectivamente.

Las cotas halladas de los errores serán:

$$\begin{aligned}
|E_3| &< \frac{1}{720} M_4 < 4.832... < 4.9 \ , \\
|E_5| &< \frac{1}{4725 \cdot 2^7} M_6 < 0.20592... < 0.21 \ .
\end{aligned}$$

Problema 4.5 *El período de un péndulo simple de longitud l, dejado libre desde un ángulo inicial α con la dirección vertical en un lugar de la tierra donde la aceleración de la gravedad vale g, es*

$$T = 4\sqrt{\frac{l}{g}} \int_0^{\frac{\pi}{2}} \frac{d\varphi}{\sqrt{1 - K^2 \operatorname{sen}^2 \varphi}} \ , \quad K = \operatorname{sen} \frac{\alpha}{2} \ .$$

a) Hallar el desarrollo de Taylor de T como función de K.

b) Descubrir en él la fórmula aproximada $T \simeq 2\pi\sqrt{\frac{l}{g}}$, válida para pequeñas oscilaciones.

c) Acotar el error relativo cometido al usar la fórmula de b), cuando $\alpha = 5°$.

SOLUCIÓN:

a) El desarrollo de Taylor de la integral propuesta, en función de K, se halla integrando, término a término, el desarrollo del integrando de la forma que sigue:

$$\int_0^{\frac{\pi}{2}} \frac{d\varphi}{\sqrt{1 - K^2 \operatorname{sen}^2 \varphi}} = \int_0^{\frac{\pi}{2}} (1 - K^2 \operatorname{sen}^2 \varphi)^{-\frac{1}{2}} d\varphi$$

$$\begin{aligned} &= \int_0^{\frac{\pi}{2}} \sum_{j=0}^{\infty} \binom{-\frac{1}{2}}{j} (-K^2 \operatorname{sen}^2 \varphi)^j d\varphi \\ &= \sum_{j=0}^{\infty} (-1)^j \binom{-\frac{1}{2}}{j} \left(\int_0^{\frac{\pi}{2}} \operatorname{sen}^{2j} \varphi d\varphi \right) K^{2j} \\ &= \frac{\pi}{2} + \frac{\pi}{2} \sum_{j=1}^{\infty} \left(\frac{(2j-1)!!}{(2j)!!} \right)^2 K^{2j} \ . \end{aligned}$$

Explicitándolo y substituyéndolo en la expresión del período del péndulo hallamos

$$\begin{aligned} T &= 4\sqrt{\frac{l}{g}} \int_0^{\frac{\pi}{2}} \frac{d\varphi}{\sqrt{1 - K^2 \operatorname{sen}^2 \varphi}} \\ &= 2\pi \sqrt{\frac{l}{g}} \left(1 + \frac{K^2}{4} + \frac{9K^4}{64} + \frac{25K^6}{256} + \cdots \right) \ . \end{aligned}$$

b) Nótese que el primer término del desarrollo da la expresión aproximada cuando se trata de pequeñas oscilaciones del péndulo

$$T \simeq 2\pi \sqrt{\frac{l}{g}} \ .$$

c) El error relativo que se comete usando la fórmula aproximada de b) vendrá dado por el desarrollo

$$\frac{K^2}{4} + \frac{9K^4}{64} + \frac{25K^6}{256} + \cdots \ ,$$

que puede ser mayorado por una serie geométrica de razón $K^2 = \operatorname{sen}^2 2.5^\circ$, resultando la cota aproximada

$$\frac{K^2}{4} \frac{1}{1 - K^2} = \frac{1}{4} \tan^2 2.5^\circ = 0.4766... \cdot 10^{-3} < 0.5 \cdot 10^{-3} \ .$$

Problema 4.6 *Consideramos las integrales del tipo*

$$\int_{-1}^{1} (1 + x^4) f(x) dx \ .$$

a) Hallar una fórmula de tres abscisas que sea exacta para todo polinomio de grado menor o igual que 5 y dar una expresión para el error cometido en el caso de que $f \in \mathcal{C}^6([-1, 1])$.

b) Aplicación: Calcular la integral para $f(x) = \operatorname{sen}^2 x$, *acotando el error.*

SOLUCIÓN:

a) La fórmula buscada es necesariamente gaussiana; es decir, tiene como abscisas los ceros del polinomio $\psi_3(x)$ ortogonal respecto al producto escalar

$$(f,g)=\int_{-1}^{1}(1+x^4)f(x)g(x)dx$$

y como pesos aquéllos que hacen la fórmula exacta para todos los polinomios de grado menor o igual que 2.

Los polinomios ortogonales pueden hallarse por recurrencia; pero, dado que sólo necesitamos calcularlos hasta grado 3, pueden obtenerse más directamente.

La simetría del intervalo de integración y la paridad de la función peso implican la paridad de los polinomios ortogonales: $\psi_j(-x)=(-1)^j\psi_j(x)$ $(j\geq 0)$. Así, si tomamos los polinomios ortogonales mónicos, tenemos

$$\psi_0(x)=1\ ,\quad \psi_1(x)=x\ ,\quad \psi_2(x)=x^2-\alpha\ ,\ \psi_3(x)=x^3-\beta x\ .$$

El polinomio $\psi_3(x)$, debido a que no tiene términos pares, es automáticamente ortogonal a $\psi_0(x)$ y $\psi_2(x)$. El coeficiente indeterminado β de este polinomio se halla imponiendo su ortogonalidad con $\psi_1(x)$, resultando

$$\beta=\frac{(x^3,x)}{(x,x)}=\frac{49}{75}\ .$$

El polinomio ortogonal $\psi_3(x)$ tiene, pues, los ceros $x_0=-\sqrt{\frac{49}{75}}$, $x_1=0$ y $x_2=\sqrt{\frac{49}{75}}$ y, teniendo en cuenta que $\psi_3(x)$ es mónico, la fórmula gaussiana buscada tiene la forma

$$\begin{aligned}\int_{-1}^{1}(1+x^4)f(x)dx &= W_0 f\left(-\sqrt{\frac{49}{75}}\right)+W_1 f(0)+W_2 f\left(\sqrt{\frac{49}{75}}\right)\\ &\quad +\frac{f^{(6)}(\xi)}{6!}(\psi_3,\psi_3)\ ,\quad \xi\in(-1,1)\ .\end{aligned}$$

Hallaremos ahora sus pesos, imponiendo exactitud para los polinomios de grado menor o igual que 2, y explicitaremos la expresión del error.

El sistema lineal que aparece al imponer la exactitud de la fórmula para la base $\{1,x,x^2\}$ es

$$\begin{array}{l|rcrcrcr} f(x)=1 & \frac{12}{5} & = & W_0 & + & W_1 & + & W_2\\ f(x)=x & 0 & = & -\sqrt{\frac{49}{75}}W_0 & & & + & \sqrt{\frac{49}{75}}W_2\\ f(x)=x^2 & \frac{20}{21} & = & \frac{49}{75}W_0 & & & + & \frac{49}{75}W_2 \end{array}\ .$$

La resolución del sistema da $W_0=W_2=\frac{250}{343}$ y $W_1=\frac{1616}{1715}$.

Calculando primeramente

$$(\psi_3,\psi_3)=\int_{-1}^{1}(1+x^4)\psi_3^2(x)dx=\frac{15856}{259875}\ ,$$

hallamos la expresión del error.

Resumiendo, la fórmula gaussiana buscada es

$$\begin{aligned}\int_{-1}^{1}(1+x^4)f(x)dx &= \frac{250}{343}f\left(-\sqrt{\frac{49}{75}}\right)+\frac{1616}{1715}f(0)+\frac{250}{343}f\left(\sqrt{\frac{49}{75}}\right)\\ &\quad+\frac{991}{11694375}f^{(6)}(\xi)\ ,\quad \xi\in(-1,1)\ .\end{aligned}$$

b) La fórmula da la aproximación siguiente de la integral:

$$J=\int_{-1}^{1}(1+x^4)\,\mathrm{sen}^2\,x\;dx\simeq\frac{500}{343}\,\mathrm{sen}^2\sqrt{\frac{49}{75}}=0.76222...$$

Para acotar el error cometido

$$\left|\frac{991}{11694375}f^{(6)}(\xi)\right|\leq\frac{991}{11694375}M_6\ ,$$

es necesario hallar una cota M_6 de la derivada sexta de la función. A partir de la relación $\mathrm{sen}^2\,x=\frac{1}{2}(1-\cos 2x)$, se hallan fácilmente las derivadas sucesivas de la función y tenemos

$$f^{(6)}(x)=2^5\cos 2x\ ,$$

que puede acotarse por $M_6=32$.

Resulta, finalmente,

$$J\simeq 0.76222...\pm\frac{991\cdot 32}{11694375}=0.76222...\pm 2.72\cdot 10^{-3}\ .$$

De hecho, la integral pedida puede calcularse exactamente, descomponiéndola como

$$\begin{aligned}J &= \int_0^1(1-\cos 2x)dx+\int_0^1 x^4dx-\int_0^1 x^4\cos 2x\;dx\\ &= \frac{6}{5}-\frac{1}{4}\,\mathrm{sen}\,2+\frac{1}{2}\cos 2=0.7646022...\end{aligned}$$

Finalmente, el cálculo

$$J-0.76222...\simeq 2.38\cdot 10^{-3}\ ,$$

muestra la gran similitud entre la cota de error estimada y el error real.

Problema 4.7 *El período de un péndulo simple de longitud l, dejado libre desde un ángulo inicial α con la dirección vertical en un lugar de la tierra donde la aceleración de la gravedad vale g, es*

$$T=4\sqrt{\frac{l}{g}}\int_0^{\frac{\pi}{2}}\frac{d\varphi}{\sqrt{1-K^2\,\mathrm{sen}^2\,\varphi}}\ ,\quad K=\mathrm{sen}\,\frac{\alpha}{2}\ .$$

Calcular T con un error relativo menor que 10^{-4} para $\alpha=30°$, usando una fórmula de Gauss-Chebichev adecuada.

SOLUCIÓN:

La integral propuesta puede convertirse, mediante el cambio $t = \operatorname{sen}\varphi$, en una integral en el intervalo $[-1, 1]$ en la que aparece el peso $w(t) = \frac{1}{\sqrt{1-t^2}}$ de las fórmulas de Gauss-Chebichev en este intervalo,

$$\begin{aligned} T &= 4\sqrt{\frac{l}{g}}\int_0^1 \frac{dt}{\sqrt{1-t^2}\sqrt{1-K^2t^2}} \\ &= 2\sqrt{\frac{l}{g}}\int_{-1}^1 \frac{dt}{\sqrt{1-t^2}\sqrt{1-K^2t^2}} \ , \end{aligned}$$

debido a las paridades del peso y de la función

$$f(t) = \frac{1}{\sqrt{1-K^2t^2}} \ .$$

La fórmula de Gauss-Chebichev de $m+1$ abscisas se escribe ahora así

$$\begin{aligned} T &= 2\sqrt{\frac{l}{g}}\left[\frac{\pi}{m+1}\sum_{k=0}^{m}\frac{1}{\sqrt{1-K^2\cos^2(\frac{(2k+1)\pi}{2(m+1)})}}\right. \\ &\qquad \left. +\frac{\pi}{2^{2m+1}(2m+2)!}f^{(2m+2)}(\xi)\right] \ , \end{aligned}$$

con $\xi \in (-1, 1)$ y $K = \operatorname{sen} 15^\circ = 0.258819045...$

Para $m = 0, 1, 2$, tenemos, respectivamente,

$$\begin{aligned} T &= 2\pi\sqrt{\frac{l}{g}}\left[1+\frac{1}{4}f^{(2)}(\xi)\right] \ , \\ T &= 2\pi\sqrt{\frac{l}{g}}\left[\frac{1}{2}\left(\frac{1}{\sqrt{1-\frac{1}{2}K^2}}+\frac{1}{\sqrt{1-\frac{1}{2}K^2}}\right)+\frac{1}{192}f^{(4)}(\xi)\right] \ , \\ T &= 2\pi\sqrt{\frac{l}{g}}\left[\frac{1}{3}\left(\frac{1}{\sqrt{1-\frac{3}{4}K^2}}+1+\frac{1}{\sqrt{1-\frac{3}{4}K^2}}\right)+\frac{1}{23040}f^{(6)}(\xi)\right] \ . \end{aligned}$$

Con el fin de evaluar los errores para hallar un valor adecuado de m, es necesario acotar las correspondientes derivadas pares de la función f:

$$\begin{aligned} f(t) &= (1-K^2t^2)^{-\frac{1}{2}} \ , \\ f'(t) &= K^2tf^3(t) \ , \\ f^{(2)}(t) &= K^2(1+2K^2t^2)f^5(t) \ , \\ f^{(3)}(t) &= 3K^4t(3+2K^2t^2)f^7(t) \ , \\ f^{(4)}(t) &= 3K^4(3+24K^2t^2+8K^4t^4)f^9(t) \ , \\ f^{(5)}(t) &= 15K^6t(15+40K^2t^2+8K^4t^4)f^{11}(t) \ , \\ f^{(6)}(t) &= 45K^6(5+90K^2t^2+120K^4t^4+16K^6t^6)f^{13}(t) \ . \end{aligned}$$

Por lo tanto,

$$M_{2r} \equiv \max_{t\in[-1,1]} \left|f^{(2r)}(t)\right| = f^{(2r)}(1) \quad (r = 1, 2, 3) \ .$$

Así, obtenemos

$$M_2 \simeq 0.0903 \ , \quad M_4 \simeq 0.0854 \ , \quad M_6 \simeq 0.2457 \ .$$

Las cotas de los errores correspondientes son, para $m = 0, 1, 2$, respectivamente,

$$\frac{M_2}{4} \leq 226 \cdot 10^{-4} \ , \quad \frac{M_4}{192} \leq 4.5 \cdot 10^{-4} \ , \quad \frac{M_6}{23040} \leq 0.11 \cdot 10^{-4} \ .$$

La fórmula adecuada es, pues, la de $m = 2$ que da la siguiente aproximación de T:

$$\begin{aligned} T &= 2\pi\sqrt{\frac{l}{g}}\left[\frac{1.0261082 + 1 + 1.0261082}{3} \pm 1.1 \cdot 10^{-5}\right] \\ &= 2\pi\sqrt{\frac{l}{g}}(1.017405 \pm 2 \cdot 10^{-5}) \ . \end{aligned}$$

Problema 4.8 *a) Demostrar que los polinomios de Legendre cumplen las siguientes relaciones de recurrencia para $t \in \mathbb{R}$ y $j \geq 1$:*

i) $P'_j(t) - tP'_{j-1}(t) = jP_{j-1}(t)$,

ii) $tP'_j(t) - P'_{j-1}(t) = jP_j(t)$,

iii) $P'_{j+1}(t) - P'_{j-1}(t) = (2j+1)P_j(t)$,

iv) $(t^2-1)P'_j(t) = jtP_j(t) - jP_{j-1}(t)$.

b) Deducir la fórmula de Gauss-Legendre de $m+1$ abscisas

$$\int_a^b f(x)dx = \frac{b-a}{2}\sum_{k=0}^{m} W_k f(x_k) + E_{m+1}(f) \ ,$$

con

$$x_k = \frac{b-a}{2}t_k + \frac{a+b}{2} \ , \quad W_k = \frac{2}{(1-t_k)^2[P'_{m+1}(t_k)]^2} \ ,$$

donde $t_k \in (-1, 1)$ $(k = 0 \div m)$ son los ceros del polinomio de Legendre $P_{m+1}(t)$ y, si la función $f \in \mathcal{C}^{2m+2}([a, b])$, el error de la fórmula anterior viene dado por la expresión

$$E_{m+1}(f) = \frac{(b-a)^{2m+3}[(m+1)!]^4}{(2m+3)[(2m+2)!]^3} f^{(2m+2)}(\xi) \ , \quad \xi \in (a, b) \ .$$

c) Explicitar las fórmulas de Gauss-Legendre de 1, 2 y 3 abscisas sobre $[-1, 1]$.

SOLUCIÓN:

Recordemos del capítulo anterior que los polinomios de Legendre están definidos por

$$P_0(t) = 1 \ , \quad P_j(t) = \frac{1}{2^j j!}\frac{d^j}{dt^j}[(t^2-1)^j] \ (j \geq 1)$$

y son ortogonales respecto al peso $w(t) = 1$ en el intervalo $[-1, 1]$:

$$(P_j, P_l) = \int_{-1}^{1} P_j(t)P_l(t)dt = \frac{2}{2j+1}\delta_{jl} \ (j, l \geq 0) \ .$$

Además, tienen coeficiente principal

$$A_j = \frac{(2j)!}{2^j (j!)^2} \ ,$$

y satisfacen la relación de recurrencia

$$\begin{aligned} P_0(t) &= 1 \ , \\ P_1(t) &= t \quad , \\ (j+1)P_{j+1}(t) &= (2j+1)tP_j(t) - jP_{j-1}(t) \ (j \geq 1) \ . \quad (*) \end{aligned}$$

a) De ahora en adelante supondremos $j \geq 1$.
Derivando en la definición de $P_j(t)$, obtenemos i):

$$\begin{aligned} P_j'(t) &= \frac{1}{2^j j!}\frac{d^{j+1}}{dt^{j+1}}[(t^2-1)^j] = \frac{1}{2^j j!}\frac{d^j}{dt^j}[2jt(t^2-1)^{j-1}] \\ &= \frac{1}{2^{j-1}(j-1)!}\left(t\frac{d^j}{dt^j}[(t^2-1)^{j-1}] + j\frac{d^{j-1}}{dt^{j-1}}[(t^2-1)^{j-1}]\right) \\ &= tP_{j-1}'(t) + jP_{j-1}(t) \ , \end{aligned}$$

donde hemos usado que

$$\frac{d^j}{dt^j}[tg(t)] = t\frac{d^j}{dt^j}[g(t)] + j\frac{d^{j-1}}{dt^{j-1}}[g(t)] \ .$$

Ahora las demás relaciones se obtienen rápidamente. Así, derivando la relación de recurrencia $(*)$ y substituyendo $P_{j+1}'(t)$ en la relación i) para el índice $j+1$, se obtiene ii). Sumando la relación ii) a la relación i) para el índice $j+1$, sale iii). Finalmente, iv) se halla despejando $P_{j-1}'(t)$ en las relaciones i), ii) e igualando.

b) Empezaremos suponiendo $[a, b] = [-1, 1]$. Por la ortogonalidad de los polinomios de Legendre respecto a la función $w(t) = 1$ en $[-1, 1]$, tenemos la fórmula gaussiana de $m+1$ abscisas $(m \geq 0)$:

$$\int_{-1}^{1} g(t)dt = \sum_{k=0}^{m} W_k g(t_k) + E_{m+1}(g) \ ,$$

donde $t_k \in (-1,1)$ $(k = 0 \div m)$ son los ceros del polinomio de Legendre $P_{m+1}(t)$, y

$$\begin{aligned} E_{m+1}(g) &= \frac{(P_{m+1}, P_{m+1})}{A_{m+1}^2} \frac{g^{(2m+2)}(\eta)}{(2m+2)!} \\ &= \frac{2^{2m+3}[(m+1)!]^4}{(2m+3)[(2m+2)!]^3} g^{(2m+2)}(\eta) \ , \quad \eta \in (-1,1) \ , \end{aligned}$$

ya que $g \in \mathcal{C}^{2m+2}([-1,1])$. Los pesos W_k $(k = 0 \div m)$ pueden obtenerse imponiendo que la fórmula gaussiana sea exacta para los polinomios de grado menor o igual que m; esto es, que $E_{m+1}(t^i) = 0$ $(i = 0 \div m)$.

Ahora bien, se tiene también la expresión siguiente para los pesos de la fórmula gaussiana:

$$W_k = \frac{A_{m+1}(P_m, P_m)}{A_m P'_{m+1}(t_k) P_m(t_k)} = \frac{2}{(m+1) P_m(t_k) P'_{m+1}(t_k)} \ ;$$

aplicando la relación iv), obtenemos

$$(t_k^2 - 1) P'_{m+1}(t_k) = -(m+1) P_m(t_k) \ ,$$

y, por lo tanto,

$$W_k = \frac{2}{(1 - t_k^2)[P'_{m+1}(t_k)]^2} \ (k = 0 \div m) \ .$$

El caso general de una integral

$$\int_a^b f(x) dx$$

se reduce al caso estudiado anteriormente, mediante el siguiente cambio afín de variables:

$$x = \frac{b-a}{2} t + \frac{a+b}{2} \ ;$$

entonces

$$\int_a^b f(x) dx = \frac{b-a}{2} \int_{-1}^1 g(t) dt \ ,$$

donde

$$g(t) = f\left(\frac{b-a}{2} t + \frac{a+b}{2} \right) \ , \quad t \in [-1,1] \ ,$$

cumple

$$g^{(2m+2)}(\eta) = \left(\frac{b-a}{2} \right)^{2m+2} f^{(2m+2)}(\xi) \ ,$$

con

$$\xi = \frac{b-a}{2} \eta + \frac{a+b}{2} \in (a,b) \ , \eta \in (-1,1) \ .$$

c) Aplicando el apartado b), hallamos las fórmulas de Gauss-Legendre sobre $[-1,1]$ para $m=0,1,2$:

$$m=0: \quad P_1(t)=t\ ,$$

$$t_0=0\ ,$$

$$\textstyle\int_{-1}^{1} g(t)dt = 2g(0)+\frac{1}{3}g^{(2)}(\eta)\ .$$

$$m=1: \quad P_2(t)=\tfrac{1}{2}(3t^2-1)\ ,$$

$$t_1=-t_0=\tfrac{\sqrt{3}}{3}\ ,$$

$$\textstyle\int_{-1}^{1} g(t)dt = g(\frac{-\sqrt{3}}{3})+g(\frac{\sqrt{3}}{3})+\frac{g^{(4)}(\eta)}{135}\ .$$

$$m=2: \quad P_3(t)=\tfrac{5}{2}t(t^2-\tfrac{3}{5})\ ,$$

$$t_1=0\ , \quad t_2=-t_0=\tfrac{\sqrt{15}}{5}\ ,$$

$$\textstyle\int_{-1}^{1} g(t)dt = \frac{5}{9}g(-\frac{\sqrt{15}}{5})+\frac{8}{9}f(0)+\frac{5}{9}g(\frac{\sqrt{15}}{5})+\frac{g^{(6)}(\eta)}{15750}\ .$$

Problema 4.9 *Calcular la suma de la serie*

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

con 4 cifras exactas, usando sucesivamente el método de comparación con series telescópicas.

SOLUCIÓN:

Consideramos primero la serie telescópica

$$\sum_{j=1}^{\infty}\frac{1}{j(j+1)}=\sum_{j=0}^{\infty}\frac{1}{(j+1)(j+2)}=\sum_{j=0}^{\infty}j^{(-2)}=0^{(-1)}=1\ ;$$

así,

$$S=1+\sum_{j=1}^{\infty}\left(\frac{1}{j^2}-\frac{1}{j(j+1)}\right)=1+\sum_{j=1}^{\infty}\frac{1}{j^2(j+1)}\equiv 1+S_1\ .$$

Tomando ahora la segunda serie telescópica

$$\begin{aligned}\sum_{j=2}^{\infty}\frac{1}{(j-1)j(j+1)} &= \sum_{j=0}^{\infty}\frac{1}{(j+1)(j+2)(j+3)}\\ &= \sum_{j=0}^{\infty} j^{(-3)} = \frac{0^{(-2)}}{2} = \frac{1}{4}\ ;\end{aligned}$$

tenemos,

$$\begin{aligned}S &= 1+\frac{1}{2}+\sum_{j=2}^{\infty}\frac{1}{j^2(j+1)}\\ &= 1+\frac{1}{2}+\frac{1}{4}+\sum_{j=2}^{\infty}\left(\frac{1}{j^2(j+1)}-\frac{1}{(j-1)j(j+1)}\right)\\ &= \frac{7}{4}-\sum_{j=2}^{\infty}\frac{1}{j^2(j^2-1)}\ .\end{aligned}$$

Considerando

$$\begin{aligned}&\sum_{j=3}^{\infty}\frac{1}{(j-2)(j-1)j(j+1)}\\ &\quad= \sum_{j=0}^{\infty}\frac{1}{(j+1)(j+2)(j+3)(j+4)}\\ &\quad= \sum_{j=0}^{\infty} j^{(-4)} = \frac{0^{(-3)}}{3} = \frac{1}{18}\ ,\end{aligned}$$

tendremos

$$\begin{aligned}S &= \frac{7}{4}-\frac{1}{12}-\sum_{j=3}^{\infty}\frac{1}{j^2(j^2-1)}\\ &= \frac{5}{3}-\frac{1}{18}-\sum_{j=3}^{\infty}\left(\frac{1}{j^2(j^2-1)}-(j-3)^{(-4)}\right)\\ &= \frac{29}{18}+2\sum_{j=3}^{\infty}\frac{1}{j^2(j-1)(j+1)(j-2)}\ .\end{aligned}$$

Para esta última serie, los restos pueden acotarse por el criterio integral:

$$\begin{aligned}R_n &= 2\sum_{j=n+1}^{\infty}\frac{1}{j^2(j-1)(j^2-j-2)} < 2\sum_{j=n+1}^{\infty}\frac{1}{(j-1)^5}\\ &< 2\int_n^{\infty}\frac{dx}{(x-1)^5} = \frac{1}{2(n-1)^4}\ ;\end{aligned}$$

si $n=6$, $0\le R_6\le 0.8\cdot 10^{-3}$ y

$$\begin{aligned}S &= \frac{29}{18}+2\sum_{j=3}^{6}\frac{1}{j^2(j^2-1)(j-2)}+R_6\\ &= 1.64456+R_6 = 1.64496\pm 0.4\cdot 10^{-3}\ .\end{aligned}$$

Notamos que el resultado exacto es $S = \frac{\pi^2}{6} = 1.644934...$, que es un caso particular del resultado general

$$\zeta(2s) = \sum_{j=1}^{\infty} \frac{1}{j^{2s}} = \frac{(2\pi)^{2s}}{2(2s)!} |B_{2s}| \quad (s \geq 1) \ ,$$

donde ζ es la llamada función *zeta de Riemann* y los coeficientes B_{2s} son los números de Bernoulli.

Problema 4.10 *La* constante de Euler *se define como*

$$\gamma = \lim_{n\to\infty} F(n) \ ,$$

siendo

$$F(n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n-1} + \frac{1}{2n} - \ln n \ .$$

Calcular γ con 10 cifras correctas.

SOLUCIÓN:

Para calcular

$$F(N) = \sum_{j=1}^{N-1} \frac{1}{j} + \frac{1}{2N} - \ln N = \sum_{j=1}^{N} \frac{1}{j} - \frac{1}{2N} - \ln N \ ,$$

escogemos $M < N$ y aplicamos la fórmula de Euler-Maclaurin para sumas con $a = M \geq 1$, $n = N - M$, $h = 1$ y $f(x) = \frac{1}{x}$.

Dado que $f^{(j)}(x) = (-1)^j j! x^{-(j+1)}$, la fórmula se convierte en

$$\begin{aligned} F(N) - F(M) = & \sum_{j=M}^{N} \frac{1}{j} - \ln N + \ln M - \frac{1}{2}\left(\frac{1}{M} + \frac{1}{N}\right) \\ + & \sum_{r=1}^{s} \frac{B_{2r}}{2r}\left(\frac{1}{M^{2r}} - \frac{1}{N^{2r}}\right) + R_s \ , \end{aligned}$$

donde

$$R_s = (N-M) B_{2s+2} \frac{1}{\xi^{2s+3}} \ , \quad \xi \in (M, N) \ .$$

Los números B_{2r} $(r \geq 1)$ que aparecen en la fórmula son los números de Bernoulli (consúltese el problema 3.11)

$$B_{2r} = \frac{1}{6} \ , \quad -\frac{1}{30} \ , \quad \frac{1}{42} \ , \quad -\frac{1}{30} \ , \ldots \ (r \geq 1) \ ,$$

que satisfacen la propiedad $B_{2r}B_{2r+2} < 0$ $(r \geq 1)$. Por lo tanto, $R_s R_{s+1} < 0$ y, por el criterio de alternancia de los restos, la magnitud de R_s es menor o igual que la del primer término despreciado

$$|R_s| \leq \frac{|B_{2s+2}|}{2s+2}\left(\frac{1}{M^{2s+2}} - \frac{1}{N^{2s+2}}\right) \leq \frac{|B_{2s+2}|}{2s+2}\frac{1}{M^{2s+2}} \ .$$

Así, haciendo tender N a infinito en la expresión hallada de $F(N) - F(M)$, tenemos

$$\begin{aligned} \gamma &= F(M) + \sum_{r=1}^{s} \frac{B_{2r}}{2r}\frac{1}{M^{2r}} \pm \frac{|B_{2s+2}|}{2s+2}\frac{1}{M^{2s+2}} \\ &= F(M) + \frac{1}{12M^2} - \frac{1}{120M^4} + \frac{1}{252M^6} - \frac{1}{240M^8} + \cdots \\ &\quad \pm\frac{|B_{2s+2}|}{2s+2}\frac{1}{M^{2s+2}} \ . \end{aligned}$$

Por comodidad de cálculo, tomamos $M = 10$; nótese que el término en M^8 ya es menor que el error permitido para el cálculo de γ

$$\frac{1}{240M^8} < \frac{1}{2}10^{-10} \ .$$

Así pues,

$$\begin{aligned} \gamma &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \frac{1}{20} - \ln 10 \\ &\quad + \frac{1}{1200} - \frac{1}{1200000} + \frac{1}{252000000} \pm \frac{1}{2}10^{-10} \\ &= 0.57721566494 \pm \frac{1}{2}10^{-10} \ . \end{aligned}$$

Problema 4.11 *La* función de Airy

$$Ai(x) = \frac{1}{\pi}\int_0^\infty \cos(\frac{t^3}{3} + xt)dt \ ,$$

admite el desarrollo asintótico siguiente:

$$Ai(x) \sim \frac{1}{2\sqrt{\pi}x^{\frac{1}{4}}}e^{-\zeta}\sum_{j=0}^{\infty}(-1)^j\frac{c_j}{\zeta^j} \quad (x \to \infty) \ ,$$

con

$$\zeta = \frac{2}{3}x^{\frac{3}{2}} \ ,$$

y

$$c_0 = 1 \ , \quad c_j = \frac{(2j+1)(2j+3)\cdots(6j-1)}{216^j j!} \quad (j \geq 1) \ .$$

Usando que el error de aproximación de la función de Airy por los primeros términos del desarrollo es menor que la magnitud del primer término despreciado, calcular $Ai(4)$ *con el menor error que permite dicho desarrollo.*

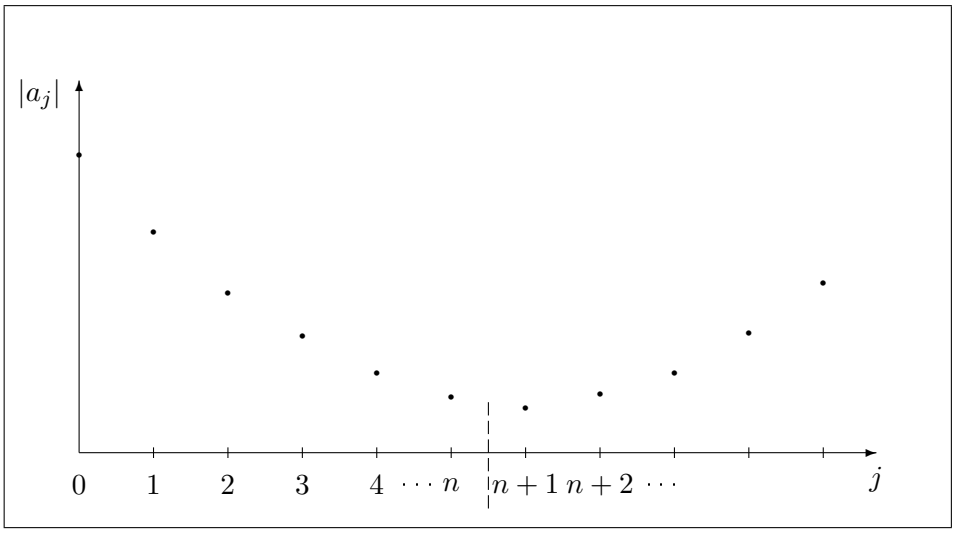


Figura 4.4: Comportamiento de los términos de una serie semiconvergente.

SOLUCIÓN:

Para $x = 4$ resulta $\zeta = \frac{16}{3}$ y

$$\mathrm{Ai}(4) \sim B \sum_{j=0}^{n} a_j \ ,$$

con

$$B = \frac{1}{2\sqrt{2\pi}} e^{-\frac{16}{3}} \ , \quad a_j = (-1)^j \left(\frac{3}{16}\right)^j c_j \quad (j \geq 0) \ .$$

Buscaremos el valor de n que haga mínimo $|a_{n+1}|$ y entonces calcularemos

$$\mathrm{Ai}(4) = B \left(\sum_{j=0}^{n} a_j \pm |a_{n+1}| \right) \ .$$

Para hallar el mínimo de los $|a_{j+1}|$ calculamos

$$\begin{aligned} \left| \frac{a_{j+1}}{a_j} \right| &= \frac{3}{16} \frac{c_{j+1}}{c_j} = \frac{3}{16} \frac{(6j+5)(6j+3)(6j+1)}{216(2j+1)(j+1)} \\ &= \frac{(6j+5)(6j+1)}{384(j+1)} \ ; \end{aligned}$$

nótese que, si $|a_{n+1}|$ es mínimo, cumplirá

$$\left| \frac{a_{n+1}}{a_n} \right| < 1 \ , \quad \left| \frac{a_{n+2}}{a_{n+1}} \right| > 1 \ .$$

La condición

$$\left| \frac{a_{j+1}}{a_j} \right| < 1$$

equivale a

$$36j^2 - 348j - 379 < 0 \ ;$$

dado que los ceros del polinomio $36x^2 - 348x - 379$ son aproximadamente -0.988 y 10.65, resulta que $n = 10$ es el número natural mayor para el que

$$\left|\frac{a_{n+1}}{a_n}\right| < 1 \ ,$$

y a_{11} es el menor de los términos, en valor absoluto.

Así pues, calcularemos

$$\mathrm{Ai}(4) = B\left(\sum_{j=0}^{10} a_j \pm |a_{11}|\right) = 0.0009515652 \pm 3 \cdot 10^{-9} \ .$$

Nótese que, usando esta serie semiconvergente, no puede obtenerse el valor pedido con precisión arbitraria, ya que el error mínimo de aproximación es el hallado. No serviría de nada calcular más términos del desarrollo: el error cometido sería todavía mayor. Véase la figura 4.4.

Problema 4.12 *Consideramos la integral*

$$I = \int_0^{\frac{\pi}{2}} \operatorname{sen} x \ dx = 1 \ .$$

a) Escribir la fórmula de Euler-Maclaurin para dicha integral, hasta términos de orden 8 en el paso h.

b) Aproximar la integral por la regla de los trapecios, dividiendo el intervalo en $M = 2, 4, 8, 16, 32, 64$ subintervalos i trabajando con 9 cifras significativas.

c) Realizar una tabla de extrapolación repetida de Richardson, basada en la fórmula hallada en a), a partir de los resultados hallados en b).

d) Obtener los primeros términos de las fórmulas asintóticas de error para las tres primeras fórmulas extrapoladas, cuando el paso h tiende a cero.

SOLUCIÓN:

a) Para $f(x) = \operatorname{sen} x$ y $[a, b] = [0, \frac{\pi}{2}]$, los factores $f^{(2r-1)}(b) - f^{(2r-1)}(a)$ que aparecen en la fórmula de Euler-Maclaurin resultan ser iguales a $(-1)^r$.

Tenemos, pues, el siguiente desarrollo asintótico de la fórmula de los trapecios en este caso:

$$\begin{aligned} T(h) &= I + \sum_{r=1}^{\infty} (-1)^r \frac{B_{2r}}{(2r)!} h^{2r} \\ &= I - \frac{h^2}{12} - \frac{h^4}{720} - \frac{h^6}{30240} - \frac{h^8}{1209600} - \cdots \end{aligned}$$

b) Para $M = 2, 4, 8, 16, 32, 64$, ha de calcularse

$$T(h) = \frac{h}{2}(f_0 + 2f_1 + \cdots + 2f_{M-1} + f_M) \ ,$$

con $h = \frac{\pi}{2M}$ y $f_k = \text{sen}\, \frac{k\pi}{2M}$ $(k = 0 \div M)$. Nótese que los cálculos realizados para evaluar cualquier $T(h)$ pueden aprovecharse para el siguiente $T(\frac{h}{2})$.

Los resultados obtenidos se muestran en la tabla siguiente:

M	$T(h)$
2	0.948059449
4	0.987115801
8	0.996785172
16	0.999196681
32	0.999799194
64	0.999949800

c) Tenemos una expresión asintótica de la forma

$$T(h) = I + a_1 h^2 + a_2 h^4 + a_3 h^6 + a_4 h^8 + \cdots \ , \quad a_r = (-1)^r \frac{B_{2r}}{(2r)!} \neq 0 \ .$$

Son, pues, adecuadas las extrapolaciones de los tipos

$$\frac{\Delta}{2^{p_j} - 1} \ , \quad p_j = 2j \quad (j \geq 1) \ ;$$

es decir,

$$\frac{\Delta}{3} \ , \quad \frac{\Delta}{15} \ , \quad \frac{\Delta}{63} \ , \quad \frac{\Delta}{255} \ , \quad \frac{\Delta}{1023} \ , \ \cdots$$

La evaluación efectiva de integrales extrapolando la regla de los trapecios tal como se indica anteriormente recibe el nombre de *método de Romberg*.

El método de Romberg realiza pues estas extrapolaciones según el esquema

$$T_1(h) = T(h) \ , \quad T_{j+1}(h) = T_j(h) + \frac{T_j(h) - T_j(2h)}{4^j - 1} \quad (j \geq 1) \ ;$$

los resultados obtenidos se recogen en la tabla siguiente:

T_1	Δ	$T_2 = T_1 + \frac{\Delta}{3}$	Δ	$T_3 = T_2 + \frac{\Delta}{15}$
0.948059449				
0.987115801	0.039056352	1.000134585		
0.996785172	0.009669371	1.000008296	−0.000126289	0.999999876
0.999196681	0.002411509	1.000000517	−0.000007779	0.999999998
0.999799194	0.000602513	1.000000032	−0.000000485	1.000000000
0.999949800	0.000150706	1.000000002	−0.000000030	1.000000000

d) Las expresiones asintóticas de los errores en las diversas fórmulas extrapoladas son de la forma

$$T_j(h) - I = a_j^{(j)} h^{2j} + a_{j+1}^{(j)} h^{2j+2} + \cdots \ ,$$

donde los coeficientes se determinan mediante la recurrencia

$$a_l^{(j+1)} = a_l^{(j)} \left(1 + \frac{1-4^l}{4^j-1}\right) = \frac{4^j-4^l}{4^j-1} a_l^{(j)} \quad (l \geq j)\ (j \geq 1)$$

que se deduce substituyendo las expresiones asintóticas de $T_j(h)$ y de $T_j(2h)$ en la fórmula de extrapolación. Nótese que se consigue la finalidad de la extrapolación repetida, anular $a_j^{(j+1)}$.

Así, los factores de variación de los diversos coeficientes serán:

$$\begin{aligned}
\frac{a_l^{(2)}}{a_l^{(1)}} &= -\frac{4^l-4}{3} = 0, -4, -20, -84, -340, \cdots \quad (l \geq 1)\ ; \\
\frac{a_l^{(3)}}{a_l^{(2)}} &= -\frac{4^l-16}{15} = 0, -\frac{16}{5}, -16, -\frac{336}{5}, \cdots \quad (l \geq 2)\ ; \\
\frac{a_l^{(4)}}{a_l^{(3)}} &= -\frac{4^l-64}{63} = 0, -\frac{64}{21}, -\frac{320}{21}, \cdots \quad (l \geq 3)\ .
\end{aligned}$$

Las fórmulas extrapoladas tendrán, pues, los errores asintóticos siguientes, cuando h tienda a 0:

$$\begin{aligned}
T_2(h) - I &= \frac{h^4}{180} + \frac{h^6}{1512} + \frac{h^8}{14400} + \cdots\ , \\
T_3(h) - I &= -\frac{2h^6}{945} - \frac{h^8}{900} - \cdots\ , \\
T_4(h) - I &= \frac{16h^8}{4725} + \cdots
\end{aligned}$$

Problema 4.13 *Supongamos el desarrollo asintótico para la regla de los trapecios*

$$T(h) = \int_a^b f(x)dx + a_1 h^2 + a_2 h^4 + \ldots + a_r h^{2r} + \ldots\ ,$$

con $a_r \neq 0$ $(r \geq 1)$.

Consideramos las fórmulas de cuadratura $T_j(\frac{h_0}{2^{j-1}})$ *obtenidas por el método de Richardson aplicado a la regla de los trapecios* $T_1(h) \equiv T(h)$ *con los pasos siguientes:* h_0, $\frac{h_0}{2}$, ..., $\frac{h_0}{2^{j-1}}$, *donde* $h_0 = b - a$.

a) Hallar dichas fórmulas $T_j(\frac{h_0}{2^{j-1}})$ *para* $j = 2, 3$ *e identificarlas con fórmulas de Newton-Cotes.*

b) Demostrar que la fórmula extrapolada $T_j(h)$ *tiene un error asintótico de la forma* $a_j^{(j)} h^{2j} + \cdots$, *con* $a_j^{(j)} \neq 0$.

c) Constatar, apoyándose en el resultado de b), que ninguna de estas fórmulas extrapoladas coincidirá con fórmulas de Newton-Cotes para $j \geq 4$.

SOLUCIÓN:

a) En la deducción de estas fórmulas, consideramos para mayor simplicidad sólo las abscisas $x_k = a + kh$ $(k = 0 \div 4)$ con $h = \frac{b-a}{4}$, debido a que son suficientes para poder realizar los tres primeros pasos del método de Richardson con $p_j = 2j$ $(j \geq 1)$ y $q = 2$.

Si tomamos $T_1(h_0)$ y $T_1(\frac{h_0}{2})$ y hacemos la primera extrapolación (del tipo $\frac{\Delta}{3}$), hallamos

$$\begin{aligned} T_2(\frac{h_0}{2}) &= T_1(\frac{h_0}{2}) + \frac{T_1(\frac{h_0}{2}) - T_1(h_0)}{3} \\ &= \frac{h_0}{4}(f_0 + 2f_2 + f_4) + \frac{\frac{h_0}{4}(f_0 + 2f_2 + f_4) - \frac{h_0}{2}(f_0 + f_4)}{3} \\ &= \frac{h_0}{6}(f_0 + 4f_2 + f_4) \ , \end{aligned}$$

que es la fórmula de Simpson con paso $h = \frac{h_0}{2}$. Se deduce, pues, que $T_2(h)$ es la regla de Simpson $S(h)$.

Si consideramos ahora $T_2(\frac{h_0}{2}) = S(\frac{h_0}{2})$ y $T_2(\frac{h_0}{4}) = S(\frac{h_0}{4})$ y hacemos la siguiente extrapolación (del tipo $\frac{\Delta}{15}$), tenemos

$$\begin{aligned} T_3(\frac{h_0}{4}) &= T_2(\frac{h_0}{4}) + \frac{T_2(\frac{h_0}{4}) - T_2(\frac{h_0}{2})}{15} \\ &= \frac{h_0}{12}(f_0 + 4f_1 + 2f_2 + 4f_3 + f_4) \\ &\quad + \frac{\frac{h_0}{12}(f_0 + 4f_1 + 2f_2 + 4f_3 + f_4) - \frac{h_0}{6}(f_0 + 4f_2 + f_4)}{15} \\ &= \frac{h_0}{90}(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4) \ , \end{aligned}$$

que es la fórmula de Newton-Cotes para $n = 4$.

b) El proceso de extrapolación modifica los coeficientes del desarrollo asintótico del error según la recurrencia

$$a_l^{(j+1)} = \frac{4^j - 4^l}{4^j - 1} a_l^{(j)} \quad (l \geq j) \ ,$$

así se consigue el objetivo de anular $a_j^{(j+1)} = 0$ y que los coeficientes $a_l^{(j+1)}$ $(l > j)$ del error asintótico de T_{j+1} sean no nulos, si lo son los coeficientes correspondientes de T_j. Dado que los coeficientes $a_l^{(1)}$ $(l \geq 1)$ se suponen no nulos, deducimos que los errores asintóticos son como se indica en el enunciado.

c) Para $j \geq 1$, la fórmula hallada para T_j utiliza $2^{j-1} + 1$ abscisas y tiene una expresión del error asintótico que empieza con términos de orden $\mu_j = 2j$ en h. Esta fórmula se tendría que comparar con la de Newton-Cotes del mismo número de abscisas $m + 1 = 2^{j-1} + 1$. Si $j = 1$, la expresión del error de la fórmula de Newton-Cotes de 2 abscisas es del orden de $\nu_1 = 2$ en h. Si $j \geq 2$, el número de abscisas $2^{j-1} + 1$ es impar y el orden del error de la fórmula de Newton-Cotes correspondiente es $\nu_j = 2^{j-1} + 2$. La coincidencia de estos valores se da para $j = 1, 2, 3$:

$$\mu_1 = \nu_1 = 2 \ , \quad \mu_2 = \nu_2 = 4 \ , \quad \mu_3 = \nu_3 = 6 \ ;$$

pero no se da para $j > 3$. Esto demuestra que las fórmulas extrapoladas no pueden coincidir con las de Newton-Cotes cuando $j > 3$.

Problema 4.14 *a) Deducir la* fórmula de Gauss hacia adelante *de interpolación equidistante en las abscisas* $\{x_0, x_1, x_{-1}, \ldots, x_s, x_{-s}\}$

$$\begin{aligned} f(x_0+th) &= f_0 + \sum_{r=1}^{s}\left[\binom{t+r-1}{2r-1}\delta^{2r-1}f_{\frac{1}{2}} + \binom{t+r-1}{2r}\delta^{2r}f_0\right] \\ &\quad + h^{2s+1}f^{(2s+1)}(\xi(t))\binom{t+s}{2s} \end{aligned}$$

y la fórmula de Gauss hacia atrás *de interpolación equidistante en las abscisas* $\{x_0, x_{-1}, x_1, \ldots, x_{-s}, x_s\}$

$$\begin{aligned} f(x_0+th) &= f_0 + \sum_{r=1}^{s}\left[\binom{t+r-1}{2r-1}\delta^{2r-1}f_{-\frac{1}{2}} + \binom{t+r}{2r}\delta^{2r}f_0\right] \\ &\quad + h^{2s+1}f^{(2s+1)}(\xi(t))\binom{t+s}{2s} \ . \end{aligned}$$

b) Deducir, haciendo la media de las fórmulas de Gauss hacia adelante y hacia atrás, la fórmula de interpolación de Stirling

$$\begin{aligned} f(x_0+th) &= f_0 + \sum_{r=1}^{s}\left[\binom{t+r-1}{2r-1}\mu\delta^{2r-1}f_0 + \frac{t}{2r}\binom{t+r-1}{2r-1}\delta^{2r}f_0\right] \\ &\quad + h^{2s+1}f^{(2s+1)}(\xi(t))\binom{t+s}{2s} \ . \end{aligned}$$

c) Aplicación: Usar la fórmula de interpolación de Stirling para hallar $\operatorname{sen} 1.05$*, con 9 cifras decimales correctas, a partir de la tabla de* $\operatorname{sen}(1+x)$ *en las abscisas* $x_k = \frac{k}{10}$ $(k = -10 \div 10)$.

SOLUCIÓN:

a) En las abscisas indicadas, la fórmula de las diferencias divididas de Newton toma la siguiente forma:

$$\begin{aligned} f(x_0+th) &= f_0 + \sum_{r=1}^{s}\Big[f[x_0,\ldots,x_{r-1},x_{-(r-1)},x_r]\cdot \\ &\qquad (x-x_0)\cdots(x-x_{r-1})(x-x_{-(r-1)}) \\ &\quad + f[x_0,\ldots,x_{r-1},x_{-(r-1)},x_r,x_{-r}]\cdot \\ &\qquad (x-x_0)\cdots(x-x_{r-1})(x-x_{-(r-1)})(x-x_r)\Big] \\ &\quad + \frac{f^{(2s+1)}(\xi(t))}{(2s+1)!}(x-x_0)\cdots(x-x_s)(x-x_{-s}) \ . \end{aligned}$$

Debido a su simetría, las diferencias divididas pueden escribirse, usando los operadores Δ y δ, en la forma:

$$\begin{aligned} f[x_0,\ldots,x_{r-1},x_{-(r-1)},x_r] &= f[x_{-(r-1)},\ldots,x_{r-1},x_r] \\ &= \frac{\Delta^{2r-1}f_{-(r-1)}}{(2r-1)!h^{2r-1}} = \frac{\delta^{2r-1}f_{\frac{1}{2}}}{(2r-1)!h^{2r-1}} \ , \end{aligned}$$

$$\begin{aligned} f[x_0,\ldots,x_{r-1},x_{-(r-1)},x_r,x_{-r}] &= f[x_{-r},x_{-(r-1)},\ldots,x_{r-1},x_r] \\ &= \frac{\Delta^{2r}f_{-r}}{(2r)!h^{2r}} = \frac{\delta^{2r}f_0}{(2r)!h^{2r}} \ . \end{aligned}$$

Los polinomios en x que aparecen en la fórmula se escriben como polinomios en la variable $t = \frac{x-x_0}{h}$:

$$\begin{aligned} &(x-x_0)\cdots(x-x_{r-1})(x-x_{-(r-1)}) \\ &= h^{2r-1}(t+r-1)\cdots(t-r+1) \\ &= h^{2r-1}(2r-1)!\begin{pmatrix} t+r-1 \\ 2r-1 \end{pmatrix} \ , \end{aligned}$$

$$\begin{aligned} &(x-x_0)\cdots(x-x_{r-1})(x-x_{-(r-1)})(x-x_r) \\ &= h^{2r}(t+r-1)\cdots(t-r+1)(t-r) \\ &= h^{2r}(2r)!\begin{pmatrix} t+r-1 \\ 2r \end{pmatrix} \ ; \end{aligned}$$

$$\begin{aligned} (x-x_0)\cdots(x-x_s)(x-x_{-s}) &= h^{2s+1}(t+s)\cdots(t-s) \\ &= h^{2s+1}(2s+1)!\begin{pmatrix} t+s \\ 2s+1 \end{pmatrix} \ . \end{aligned}$$

Substituyendo estas expresiones en la primera expresión hallada, obtenemos la fórmula de Gauss hacia adelante del enunciado.

De manera análoga, se deduce la fórmula de Gauss hacia atrás.

b) Al hacer la media de ambas fórmulas, aparecen los siguientes cálculos para los dos primeros términos del sumatorio de la fórmula resultante:

$$\frac{1}{2}(\delta^{2r-1}f_{\frac{1}{2}} + \delta^{2r-1}f_{-\frac{1}{2}}) = \mu\delta^{2r-1}f_0 \ ,$$

$$\begin{aligned} &\frac{1}{2}\left[\begin{pmatrix} t+r-1 \\ 2r \end{pmatrix} + \begin{pmatrix} t+r \\ 2r \end{pmatrix}\right] = \frac{1}{2(2r)!}\cdot \\ &\qquad [(t+r-1)\cdots(t-r+1)(t-r) + (t+r)(t+r-1)\cdots(t-r+1)] \\ &= \frac{t}{2r}\begin{pmatrix} t+r-1 \\ 2r-1 \end{pmatrix} \ . \end{aligned}$$

Así, resulta la fórmula de interpolación de Stirling del enunciado, que hacemos explícita a continuación:

$$\begin{aligned} f(x_0+th) &= f_0+\mu\delta f_0 t+\frac{1}{2}\delta^2 f_0 t^2 \\ &\quad +\frac{1}{3!}\mu\delta^3 f_0(t+1)t(t-1)+\frac{1}{4!}\delta^4 f_0(t+1)t^2(t-1) \\ &\quad +\frac{1}{5!}\mu\delta^5 f_0(t+2)(t+1)t(t-1)(t-2) \\ &\quad +\frac{1}{6!}\delta^6 f_0(t+2)(t+1)t^2(t-1)(t-2)+\cdots \\ &\quad +\frac{f^{(2s+1)}(\xi(t))}{(2s+1)!}h^{2s+1}(t+s)\cdots(t-s)\ . \end{aligned}$$

c) Por lo que se refiere a la aplicación pedida, notamos que el error permitido se alcanza para $s=3$. Acotamos, por ello, el término de error en este caso en el que $t=\frac{1}{2}$ y $h=0.1$:

$$\begin{aligned} &\left|\frac{f^{(7)}(\xi)}{7!}10^{-7}(\frac{1}{2}+3)\cdots(\frac{1}{2}-3)\right| \\ &< \frac{1}{5040}10^{-7}\frac{7}{2}\frac{5}{2}\frac{3}{2}\frac{1}{2}\frac{1}{2}\frac{3}{2}\frac{5}{2}M_7=\frac{5}{2048}10^{-7}M_7\ , \end{aligned}$$

donde M_7 es una cota de la derivada séptima de $f(x)=\mathrm{sen}(1+x)$ en $[-0.3,0.3]$ que podemos acotar, a su vez, por 1; la cota del error resultante es menor que $\frac{1}{2}10^{-9}$.

En resumen, para $s=3$ y trabajando con valores de la función dados con más de 10 decimales, deberíamos hallar el valor pedido con 9 decimales correctos.

La fórmula que tenemos que aplicar, para $t=\frac{1}{2}$, es pues

$$\begin{aligned} f(0.05) &= f_0+\frac{1}{2}\mu\delta f_0+\frac{1}{8}\delta^2 f_0 \\ &\quad -\frac{1}{16}\mu\delta^3 f_0-\frac{1}{128}\delta^4 f_0 \quad (*) \\ &\quad +\frac{3}{256}\mu\delta^5 f_0+\frac{1}{1024}\delta^6 f_0 \\ &\quad \pm\frac{1}{2}10^{-9}\ . \end{aligned}$$

Con el fin de calcular cómodamente los términos de la forma $\mu\delta^{(2r-1)}f_0$ y $\delta^{2r}f_0$ ($r=1\div 3$), conviene tener en cuenta que los primeros son las medias de los términos

$$\begin{aligned} \delta^{(2r-1)}f_{\frac{1}{2}} &= \Delta^{(2r-1)}f_{-(r-1)}\ , \\ \delta^{(2r-1)}f_{-\frac{1}{2}} &= \Delta^{(2r-1)}f_{-r} \end{aligned}$$

de la tabla de diferencias finitas, y los segundos son los términos $\Delta^{2r}f_{-r}$ de la misma tabla.

Esto es, la tabla de diferencias finitas que tenemos que aplicar puede escribirse también

como

$$\begin{array}{c|ccccccc}
x_{-3} & f_{-3} & & & & & & \\
 & & \delta f_{-\frac{5}{2}} & & & & & \\
x_{-2} & f_{-2} & & \delta^2 f_{-2} & & & & \\
 & & \delta f_{-\frac{3}{2}} & & \delta^3 f_{-\frac{3}{2}} & & & \\
x_{-1} & f_{-1} & & \delta^2 f_{-1} & & \delta^4 f_{-1} & & \\
 & & \delta f_{-\frac{1}{2}} & & \delta^3 f_{-\frac{1}{2}} & & \delta^5 f_{-\frac{1}{2}} & \\
x_0 & f_0 & & \delta^2 f_0 & & \delta^4 f_0 & & \delta^6 f_0 \\
 & & \delta f_{\frac{1}{2}} & & \delta^3 f_{\frac{1}{2}} & & \delta^5 f_{\frac{1}{2}} & \\
x_1 & f_1 & & \delta^2 f_1 & & \delta^4 f_1 & & \\
 & & \delta f_{\frac{3}{2}} & & \delta^3 f_{\frac{3}{2}} & & & \\
x_2 & f_2 & & \delta^2 f_2 & & & & \\
 & & \delta f_{\frac{5}{2}} & & & & & \\
x_3 & f_3 & & & & & &
\end{array}$$

Concretamos, a continuación, esta tabla de diferencias finitas para la función $f(x) = \operatorname{sen}(1+x)$ en las abscisas $x_k = \frac{k}{10}$ $(k = -3 \div 3)$:

$$\begin{array}{c|ccccccc}
0.7 & 0.64421768724 & & & & & & \\
 & & 0.07313840366 & & & & & \\
0.8 & 0.71735609090 & & -0.00716758493 & & & & \\
 & & 0.06597081873 & & -0.00782674355 & & -0.00065915861 & \\
0.9 & 0.78332690963 & & -0.00782674355 & & -0.00058095638 & & \\
 & & 0.05814407518 & & -0.00840769993 & & -0.00049694942 & \\
1.0 & 0.84147098481 & & -0.00840769993 & & -0.00049694942 & & \\
 & & 0.04973637525 & & -0.00890464935 & & -0.00040797711 & \\
1.1 & 0.89120736006 & & -0.00890464935 & & -0.00040797711 & & \\
 & & 0.04083172591 & & -0.00931262646 & & & \\
1.2 & 0.93203908597 & & -0.00931262646 & & & & \\
 & & 0.03151909945 & & & & & \\
1.3 & 0.96355818542 & & & & & &
\end{array}$$

$$\begin{array}{ccccc}
 & & 0.00007820223 & & \\
 & & & 0.00000580472 & \\
0.00008400696 & & & & -0.00000083937 \\
 & & & 0.00000496535 & \\
 & & 0.00008897231 & &
\end{array}$$

Aplicando la fórmula de interpolación de Stirling $(*)$, resulta

$$\begin{aligned}
\operatorname{sen} 1.05 \;=\; & 0.84147098481 \\
& + \frac{1}{4}(0.04973637525 + 0.05814407518) \\
& + \frac{1}{8}(-0.00840769993) \\
& - \frac{1}{32}(-0.00049694942 - 0.00058095638) \\
& - \frac{1}{128}0.00008400696 \\
& + \frac{3}{512}(0.00000496535 + 0.00000580472)
\end{aligned}$$

$$\begin{aligned} & +\frac{1}{1024}(-0.00000083937) \pm \frac{1}{2}10^{-9} \\ = \; & 0.86742322546 \pm \frac{1}{2}10^{-9} \; , \end{aligned}$$

resultado correcto, con la precisión pedida, si comparamos con el valor exacto.

Problema 4.15 *Consideramos la fórmula de derivación para el cálculo de derivadas segundas de f en x_0*

$$f_0^{(2)} \simeq \frac{1}{h^2}(\delta^2 - \frac{1}{12}\delta^4)f_0 \; .$$

a) Expresar esta fórmula en función de los valores de la tabla de f dada por $\{(x_k = x_0 + kh, f_k = f(x_k))\}_{k=-2\div 2}$.

b) Demostrar que la fórmula considerada se halla por extrapolación de la fórmula

$$f_0^{(2)} \simeq \frac{1}{h^2}\delta^2 f_0$$

y hallar su error de truncamiento, si $f \in \mathcal{C}^6([x_0 - 2h, x_0 + 2h])$.

c) Si f se evalúa con un error acotado por ϵ, ¿cuál será el valor del paso h^* que hemos de usar para que sea mínima la suma de las cotas de error debidas a la discretización y a la evaluación de la fórmula?

d) Aplicación: Hallar la derivada segunda de la función $f(x) = \mathrm{sen}(1+x)$ en $x_0 = 0$ con los pasos $h = 10^{-3}, 10^{-2}, 10^{-1}, 1$ y comparar los errores cometidos con el que se produce al hacer el cálculo con el paso h^*, suponiendo que la función se calcula con 6 cifras decimales correctas.

SOLUCIÓN:

a) Tenemos

$$\delta^2 f_0 = f_1 - 2f_0 + f_{-1} \; , \quad \delta^4 f_0 = f_2 - 4f_1 + 6f_0 - 4f_{-1} + f_{-2} \; ;$$

la fórmula dada se expresa, por lo tanto, así:

$$\begin{aligned} f_0^{(2)} \; & \simeq \; \frac{1}{h^2}\left(f_1 - 2f_0 + f_{-1} - \frac{f_2 - 4f_1 + 6f_0 - 4f_{-1} + f_{-2}}{12}\right) \\ & = \; \frac{1}{12h^2}(-f_2 + 16f_1 - 30f_0 + 16f_{-1} - f_{-2}) \; . \end{aligned}$$

b) Llamando $D_1^2 f_0(h) = \frac{1}{h^2}\delta^2 f_0$, se tiene

$$D_1^2 f_0(h) = \frac{1}{h^2}(f_1 - 2f_0 + f_{-1}) \; .$$

A partir de los desarrollos de Taylor hasta orden 5, puede hallarse una expresión para el error de esta aproximación de $f_0^{(2)}$:

$$\begin{aligned} D_1^2 f_0(h) = \frac{1}{h^2}\cdot & \\ & \left[f_0 + hf_0' + \frac{h^2}{2}f_0^{(2)} + \frac{h^3}{3!}f_0^{(3)} + \frac{h^4}{4!}f_0^{(4)} + \frac{h^5}{5!}f_0^{(5)} + \frac{h^6}{6!}f^{(6)}(\xi_+) \right. \\ & -2f_0 \\ & \left. +f_0 - hf_0' + \frac{h^2}{2}f_0^{(2)} - \frac{h^3}{3!}f_0^{(3)} + \frac{h^4}{4!}f_0^{(4)} - \frac{h^5}{5!}f_0^{(5)} + \frac{h^6}{6!}f^{(6)}(\xi_-) \right] \\ = \quad & f_0^{(2)} + \frac{h^2}{12}f_0^{(4)} + \frac{h^4}{360}f^{(6)}(\xi) \ , \end{aligned}$$

con $\xi_+, \xi_-, \xi \in (x_0 - h, x_0 + h)$, donde se ha aplicado el teorema del valor medio para sumas, atendiendo a la continuidad de $f^{(6)}$.

Suponiendo que $f_0^{(4)} \neq 0$ y considerando la misma fórmula con paso $2h$

$$D_1^2 f_0(2h) = \frac{1}{4h^2}(f_2 - 2f_0 + f_{-2}) \ ,$$

la extrapolación adecuada es la del tipo $\frac{\Delta}{3}$, que da origen, como se quería demostrar, a la fórmula extrapolada

$$\begin{aligned} D_2^2 f_0(h) &= D_1^2 f_0(h) + \frac{1}{3}(D_1^2 f_0(h) - D_1^2 f_0(2h)) \\ &= \frac{1}{h^2}\left[f_1 - 2f_0 + f_{-1} + \frac{1}{3}(f_1 - 2f_0 + f_{-1} - \frac{f_2 - 2f_0 + f_{-2}}{4}) \right] \\ &= \frac{1}{12h^2}(-f_2 + 16f_1 - 30f_0 + 16f_{-1} - f_{-2}) = \frac{1}{h^2}\left(\delta^2 - \frac{1}{12}\delta^4 \right) f_0 \ . \end{aligned}$$

El error de esta fórmula extrapolada se obtiene usando la fórmula de error hallada anteriormente. Se tiene, por lo tanto,

$$\begin{aligned} D_2^2 f_0(h) - f_0^{(2)} &= \frac{h^4}{360}f^{(6)}(\xi) + \frac{1}{3}\left[\frac{h^4}{360}f^{(6)}(\xi) - \frac{16h^4}{360}f^{(6)}(\eta) \right] \\ &= -\frac{h^4}{90}f^{(6)}(\zeta) \ , \end{aligned}$$

con $\xi, \eta, \zeta \in (x_0 - 2h, x_0 + 2h)$, donde se ha aplicado de nuevo el teorema del valor medio para sumas.

c) Denotando por e_R y e_T los errores absolutos de evaluación y discretización, hallamos las cotas siguientes de los mismos:

$$\begin{aligned} |e_R| &\leq \frac{1 + 16 + 30 + 16 + 1}{12h^2}\epsilon = \frac{16\epsilon}{3h^2} \ , \\ |e_T| &\leq \frac{h^4}{90}M_6 \ , \quad M_6 = \max_{x \in [x_0 - 2h, x_0 + 2h]} \left| f^{(6)}(x) \right| \ . \end{aligned}$$

La suma de las cotas de error halladas es la siguiente función de h:

$$\epsilon(h) = \frac{16\epsilon}{3h^2} + \frac{h^4}{90} M_6 \ .$$

La función $\epsilon(h)$ se minimiza para el paso buscado $h = h^*$ que anula su derivada, $\epsilon'(h^*) = 0$; esto es,

$$\frac{32\epsilon}{3h^{*3}} = \frac{4h^{*3}}{90} M_6 \ , \quad h^* = \sqrt[6]{\frac{240\epsilon}{M_6}} \ .$$

d) Calculamos las tablas de diferencias finitas de la función $f(x) = \text{sen}(1+x)$ con los pasos dados, trabajando con 6 cifras decimales. Los términos $\delta^2 f_0$ y $\delta^4 f_0$ con los que se hallan las correspondientes aproximaciones de la derivada segunda aparecen subrayados:

−0.002	0.840389				
		0.000541			
−0.001	0.840930		−0.000000		
		0.000541		0.000000	
0.000	0.841471		<u>−0.000001</u>		<u>0.000000</u>
		0.000540		0.000000	
0.001	0.842011		−0.000001		
		0.000539			
0.002	0.842550				

$$f_0^2 \simeq D_2^2 f_0(10^{-3}) = 10^6(-0.000001) = -1 \ .$$

−0.02	0.830497				
		0.005529			
−0.01	0.836026		−0.000085		
		0.005445		−0.000001	
0.00	0.841471		<u>−0.000084</u>		<u>0.000000</u>
		0.005361		−0.000001	
0.01	0.846832		−0.000085		
		0.005276			
0.02	0.852108				

$$f_0^2 \simeq D_2^2 f_0(10^{-2}) = 10^4(-0.000084) = -0.84 \ .$$

−0.2	0.717356				
		0.065971			
−0.1	0.783327		−0.007827		
		0.058144		−0.000581	
0.0	0.841471		<u>−0.008408</u>		<u>0.000085</u>
		0.049736		−0.000496	
0.1	0.891297		−0.008904		
		0.040832			
0.2	0.932939				

$$f_0^2 \simeq D_2^2 f_0(10^{-1}) = 10^2(-0.008408 - \frac{1}{12}0.000085) = -0.841508 \ .$$

$$\begin{array}{r|llll}
-2 & -0.841471 & & & \\
 & & 0.841471 & & \\
1 & 0.000000 & & 0.000000 & \\
 & & 0.841471 & & -0.773645 \\
0 & 0.841471 & & \underline{-0.773645} & & \underline{0.711285} \\
 & & 0.067826 & & -0.062359 & \\
1 & 0.909297 & & -0.836004 & \\
 & & -0.768177 & & \\
2 & 0.141120 & & &
\end{array}$$

$$f_0^2 \simeq D_2^2 f_0(1) = -0.773645 - \frac{1}{12}0.711285 = -0.832919 \ .$$

Restando ahora el valor exacto de la derivada $-\operatorname{sen} 1 = -0.841471...$, hallamos los errores:

$$D_2^2 f_0(h) - f_0^2 \simeq -0.16,\ 0.001471,\ -0.000037,\ 0.008552,$$

para $h = 10^{-3}, 10^{-2}, 10^{-1}, 1$, respectivamente.

Ahora hacemos el mismo cálculo con el paso

$$h^* = \sqrt[6]{\frac{240\epsilon}{M_6}} \simeq 0.23 \ ,$$

donde hemos tomado $\epsilon = \frac{1}{2}10^{-6}$ y hemos acotado la derivada sexta por $M_6 = 1$; se tiene

$$\begin{array}{r|llll}
-0.46 & 0.514136 & & & \\
 & & 0.181999 & & \\
-0.23 & 0.696135 & & -0.036663 & \\
 & & 0.145336 & & -0.007655 \\
0.0 & 0.841471 & & \underline{-0.044318} & & \underline{0.002334} \\
 & & 0.101018 & & -0.005321 & \\
0.23 & 0.942489 & & -0.049639 & \\
 & & 0.051379 & & \\
0.46 & 0.993868 & & &
\end{array}$$

$$f_0^2 \simeq D_2^2 f_0(0.23) = \frac{1}{0.23^2}(-0.044318 - \frac{1}{12}0.002334) = -0.841446 \ .$$

Esta aproximación da un error de 0.000025, menor que los anteriores, tal como se esperaba.

Problema 4.16 *a) Deducir la fórmula de Euler-Maclaurin mediante el cálculo formal con operadores.*

b) Explicitar dicha fórmula hasta los términos de orden 14 en el paso h.

SOLUCIÓN:

a) Fijado h, definimos el operador T que, aplicado a una función f en x_0, da la aproximación de la regla de los trapecios con paso h

$$Tf_0 = h(\frac{1}{2}f_0 + f_1 + f_2 + \cdots + f_{M-1} + \frac{1}{2}f_M)$$

a la integral definida de f entre x_0 y $x_0 + Mh$.

Queremos hallar una relación entre el operador T y el operador que da la integral definida que, aplicando la regla de Barrow y usando notación multiplicativa para los operadores, es

$$\frac{E^M - 1}{D} \equiv (E^M - 1)D^{-1} \ .$$

El operador T puede escribirse también en función del operador E

$$T = h(\frac{1}{2} + E + E^2 + \cdots + E^{M-1} + \frac{1}{2}E^M) \ .$$

Operando en esta relación, mediante la fórmula de Taylor con operadores $E = e^{hD}$, hallamos su relación con el operador integral definida y potencias del operador D:

$$\begin{aligned} T &= h\left(-\frac{1}{2} + \frac{E^M - 1}{E - 1} + \frac{1}{2}E^M\right) = h(E^M - 1)\left(\frac{1}{E-1} + \frac{1}{2}\right) \\ &= h(E^M - 1)\frac{E+1}{2(E-1)} = h(E^M - 1)\frac{hD}{2}\frac{E+1}{E-1}\frac{1}{hD} \\ &= h(E^M - 1)\left[1 + \sum_{j=1}^{\infty} c_j (hD)^j\right]\frac{1}{hD} \\ &= \frac{E^M - 1}{D} + h\sum_{j=1}^{\infty} c_j (E^M - 1)(hD)^{j-1} \ , \end{aligned}$$

donde c_j $(j \geq 1)$ son los coeficientes del desarrollo de Taylor de la función

$$G(t) = \frac{t}{2}\frac{e^t + 1}{e^t - 1} = \frac{t}{2}\ \frac{e^{\frac{t}{2}} + e^{-\frac{t}{2}}}{e^{\frac{t}{2}} - e^{-\frac{t}{2}}} = \frac{t}{e^t - 1} + \frac{t}{2} = 1 + \sum_{j=1}^{\infty} c_j t^j \ .$$

Así, tenemos la fórmula de Euler-Maclaurin que relaciona la regla de los trapecios con la integral definida y da unos términos correctivos en función de diferencias entre las derivadas impares de la función en los extremos del intervalo de integración:

$$T(h) \equiv Tf_0 = \left[\frac{E^M - 1}{D} + h\sum_{j=1}^{\infty} c_j (E^M - I)(hD)^{j-1}\right] f_0$$

$$
\begin{aligned}
&= \int_{x_0}^{x_0+Mh} f(x)dx \\
&\quad + \sum_{r=1}^{\infty} h^{2r} c_{2r} \left[f^{(2r-1)}(x_0+Mh) - f^{(2r-1)}(x_0) \right] \ ,
\end{aligned}
$$

donde se ha usado la paridad de la función G: $G(-t) = G(t)$ y, por lo tanto, $c_{2r-1} = 0$ $(r \geq 1)$.

b) Los números de Bernoulli pueden definirse por el desarrollo

$$
\frac{t}{e^t - 1} = \sum_{j=0}^{\infty} \frac{B_j}{j!} t^j \ ,
$$

y satisfacen la recurrencia

$$
(j+1)B_j = -\frac{1}{j+1} \sum_{j=0}^{\infty} \left(\begin{array}{c} j+1 \\ i \end{array} \right) B_l \ ;
$$

ésta se escribe también $(B+1)_{j+1} = B_{j+1}$ (véase el problema 3.11).

Los coeficientes c_j $(j \geq 1)$ están relacionados con los números de Bernoulli, según

$$
c_0 = B_0 = 1 \ , \quad c_1 = B_1 + \frac{1}{2} = 0 \ , \quad c_{2r+1} = 0 \ , \quad c_{2r} = \frac{B_{2r}}{(2r)!} \ .
$$

La relación de recurrencia anterior, da los valores siguientes:

$$
\begin{aligned}
B_j &= 1, \ -\frac{1}{2}, \ \frac{1}{6}, \ 0, \ -\frac{1}{30}, \ 0, \ \frac{1}{42}, \ 0, \ -\frac{1}{30}, \ 0, \ \frac{5}{66}, \\
&\quad 0, \ -\frac{691}{2730}, \ 0, \ \frac{7}{6}, \ \ldots \quad (j \geq 0) \ ;
\end{aligned}
$$

de donde

$$
\begin{aligned}
c_{2r} = \frac{B_{2r}}{(2r)!} &= \frac{1}{12}, \ -\frac{1}{720}, \ \frac{1}{30240}, \ -\frac{1}{1209600}, \ \frac{1}{47900160}, \\
&\quad -\frac{691}{2730 \cdot 12!}, \ \frac{7}{6 \cdot 14!}, \ \ldots \quad (r \geq 1) \ .
\end{aligned}
$$

La fórmula de Euler-Maclaurin, expresada hasta el término pedido, es pues

$$
\begin{aligned}
T(h) &= \int_{x_0}^{x_0+Mh} f(x)dx \\
&\quad + \frac{h^2}{12}(f'(x_0+Mh) - f'(x_0)) \\
&\quad - \frac{h^4}{720}(f^{(3)}(x_0+Mh) - f^{(3)}(x_0)) \\
&\quad + \frac{h^6}{30240}(f^{(5)}(x_0+Mh) - f^{(5)}(x_0)) \\
&\quad - \frac{h^8}{1209600}(f^{(7)}(x_0+Mh) - f^{(7)}(x_0))
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{h^{10}}{47900160}(f^{(9)}(x_0+Mh)-f^{(9)}(x_0))\\
&-\frac{691h^{12}}{2730\cdot 12!}(f^{(11)}(x_0+Mh)-f^{(11)}(x_0))\\
&+\frac{7h^{14}}{6\cdot 14!}(f^{(13)}(x_0+Mh)-f^{(13)}(x_0))+\cdots
\end{aligned}
$$

PROBLEMAS PROPUESTOS

1. Sea f una función definida en un intervalo que contenga las abscisas a y b; queremos aproximar el valor de su derivada en la abscisa a mediante una fórmula del tipo

$$f'(a) \simeq \sum_{k=0}^{m} a_k f(x_k)$$

donde las abscisas x_k $(k = 0 \div m)$ son equidistantes en el intervalo $[a, b]$.

Hallar los coeficientes a_k $(k = 0 \div m)$ de manera que la fórmula sea exacta para las funciones $f_j(x) = x^j$ $(j = 0 \div m)$, con $[a, b] = [0, 1]$.

2. a) Determinar la fórmula de derivación numérica

$$g'(0) \simeq W_{-1}g(-1) + W_0 g(0) + W_1 g(1)$$

de manera que sea exacta para todos los polinomios de grado menor o igual que 2.

b) ¿Es exacta también para los polinomios de grado menor o igual que 3?

c) Usando los apartados anteriores, deducir una fórmula de derivación de la forma

$$f'(c) \simeq a_{-1}f(c-h) + a_0 f(c) + a_1 f(c+h)$$

de manera que sea exacta para todos los polinomios de grado menor o igual que 3.

3. a) Escribir de manera explícita una fórmula de derivación para el cálculo de $f'(a)$, deducida por derivación del polinomio interpolador en las abscisas a, $a+h$, $a+2h$, $a+3h$ y $a+4h$.

b) Obtener una expresión exacta para el error, si $f \in \mathcal{C}^5([a, a+4h])$.

c) Hallar una expresión asintótica para el error, cuando f sea suficientemente diferenciable.

4. Repetir el problema anterior, cambiando las abscisas de interpolación por las nuevas abscisas $a-2h$, $a-h$, a, $a+h$ y $a+2h$.

5. a) Calcular numéricamente $f^{(2)}(0.6)$ para $f(x) = \operatorname{sen} x$, usando las fórmulas:

$$\begin{aligned} f^{(2)}(a) &\simeq \frac{1}{4h^2}[f(a+2h) - 2f(a) + f(a-2h)] \ , \\ f^{(2)}(a) &\simeq \frac{1}{h^2}[f(a+h) - 2f(a) + f(a-h)] \ , \end{aligned}$$

con pasos $h = 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005$ y usando los valores de la función seno con 8 cifras decimales correctas.

b) Analizar los errores cometidos atendiendo tanto a la fórmula usada, como a los pasos utilizados.

c) Hallar cotas de los errores de redondeo y de los errores de truncamiento de las fórmulas de derivación empleadas, dando finalmente el valor (óptimo, en teoría) del paso h que minimiza la cota del error total.

6. Determinar las abscisas de la fórmula de integración numérica

$$\int_0^1 f(x)dx \simeq 0.72f(x_0) + 0.28f(x_1)$$

de manera que sea exacta para todos los polinomios de grado menor o igual que 2.

7. Determinar los pesos y las abscisas de la fórmula de integración numérica

$$\int_{-1}^1 g(t)dt \simeq W_0 g(t_0) + W_1 g(t_1)$$

para que sea exacta para todos los polinomios del grado más alto posible.

8. Suponiendo que todos los pesos son iguales ($W_{-1} = W_0 = W_1 = W$), hallar el valor W de éstos y las abscisas de la fórmula de integración numérica

$$\int_{-1}^1 g(t)dt \simeq W_{-1}g(t_{-1}) + W_0 f(t_0) + W_1(t_1) \ ,$$

para que sea exacta para todos los polinomios de grado menor o igual que 3.

9. a) Determinar los pesos de la siguiente fórmula de integración numérica para que sea exacta para todos los polinomios de grado menor o igual que 3:

$$\int_a^b f(x)dx = A_0 f(a) + A_1 f(b) + C_0 f^{(2)}(a) + C_1 f^{(2)}(b) \ .$$

b) Demostrar que también es exacta para los polinomios de grado 4.

10. Determinar las incógnitas W_0, W_1, W_2 y c de la fórmula de integración numérica siguiente (de *tipo Lobatto*):

$$\int_{-1}^1 g(t)dt \simeq W_0 f(0) + W_1[g(-1) + g(1)] + W_2[g(-c) + g(c)]$$

para que tenga el grado de precisión más alto posible.

11. a) Hallar todas las fórmulas de Newton-Cotes (cerradas) de $m+1$ abscisas $m = 1 \div 8$ para el cálculo de

$$\int_{x_0}^{x_m} f(x)dx \ .$$

b) Demostrar que dichas fórmulas son exactas para todos los polinomios de grado menor o igual que m (si m es impar) y que $m+1$ (si m es par).

c) Suponiendo que f es suficientemente diferenciable, explicitar la fórmula del error de integración numérica para $m = 1 \div 8$, sabiendo que es de la forma

$$\frac{f^{(m+1)}(\xi)}{(m+1)!} \int_{x_0}^{x_m} \omega_m(x)dx \quad (\text{si } m \text{ es impar}) \ ,$$

$$\frac{f^{(m+2)}(\xi)}{(m+2)!} \int_{x_0}^{x_m} x\omega_m(x)dx \quad (\text{si } m \text{ es par}) \ ,$$

donde $\omega_m(x) = (x - x_0) \cdots (x - x_m)$ y $\xi \in (x_0, x_m)$.

12. a) Evaluar la integral

$$\int_{-1}^{1} \frac{dx}{1+x^2},$$

usando las fórmulas de Newton-Cotes de $m+1$ abscisas ($m = 1 \div 8$). Observar que los errores cometidos van decreciendo cuando m aumenta.

b) Hacer ahora los mismos cálculos con la integral

$$\int_{-5}^{5} \frac{dx}{1+x^2};$$

¿qué se observa? Explicarlo.

13. a) Hallar las *fórmulas de Newton-Cotes abiertas* de $m-1$ abscisas (exactas para todos los polinomios de grado menor o igual que $m-2$):

$$\int_{x_0}^{x_m} f(x)dx \simeq \sum_{k=1}^{m-1} A_k f(x_k) \ ,$$

donde las abscisas x_k ($k = 0 \div m$) son equidistantes. Hacerlo para $m = 2 \div 8$.

b) Aplicación: Calcular con ellas las integrales

$$\int_0^1 \tanh(sx)\ dx \ (s = 1, 2, 4, 8) \ ,$$

donde

$$\tanh z = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

y comentar los resultados obtenidos.

14. a) Determinar las fórmulas de integración interpolatoria en $m+1$ abscisas equidistantes ($m = 2, 3, 4, 5$) para las integrales de la forma

$$\int_{-1}^{1} w(x) f(x) dx \ ,$$

y las funciones peso siguientes:

i) $w(x) = x$,

ii) $w(x) = |x|$,

iii) $w(x) = |x|^{\frac{1}{2}}$,

iv) $w(x) = \ln(1+x)$,

v) $w(x) = (1-x^2)^{-\frac{1}{2}}$.

b) Hallar expresiones para los errores cometidos cuando f sea suficientemente diferenciable.

c) Aplicación: Calcular con ellas las integrales

$$\int_{-1}^{1} w(x) e^{-x^2} dx$$

y acotar los errores cometidos.

15. Consideramos una función f suficientemente diferenciable.

a) Dar, de forma explícita, las *fórmulas de integración numérica de Taylor*

$$\int_a^b f(x) dx \simeq \sum_{j=0}^{n} C_j f^{(j)}(c) \ ,$$

halladas por integración numérica del polinomio de interpolación de Taylor de grado menor o igual que n a f en una abscisa c del intervalo $[a, b]$.

b) Explicitar dicha fórmula cuando $c = a$ y cuando $c = b$.

c) Obtener una expresión para el error en todos los casos.

16. Hallar la integral

$$\int_0^1 \frac{\cos x - 1}{x^2} dx$$

por desarrollo de Taylor del integrando con un error menor que 10^{-10}.

17. Calcular con 5 cifras decimales correctas

$$\int_{10}^{\infty} (x^3 + x)^{-\frac{1}{2}} dx \ .$$

18. Calcular con 5 cifras decimales correctas

$$\int_0^{0.1} (1 - 0.1 \operatorname{sen} t)^{\frac{1}{2}} dt \ .$$

19. a) Partiendo del conocimiento del desarrollo

$$(1-z)^{-1} = 1 + z + z^2 + z^3 + \cdots + z^n + \frac{z^{n+1}}{1-z} \quad (|z| < 1) \ ,$$

obtener el desarrollo de Taylor en $x = 0$ de $\arctan x$.

b) Hallar de forma análoga el de $\operatorname{arcsen} x$.

20. a) Explicitar la función inversa de

$$\tanh y = \frac{e^y - e^{-y}}{e^y + e^{-y}}$$

y determinar su campo de definición. Esta función inversa se llama *argumento de la tangente hiperbólica* y se escribe $\arg\tanh$.

b) Hallar el desarrollo de Taylor de

$$F(x) = \int_0^x \arg\tanh t \ dt$$

cerca de $x_0 = 0$ y estudiar su convergencia.

c) ¿Cuántos términos del desarrollo son necesarios para calcular $F(0.1)$ y $F(0.9)$ con un error menor que $\frac{1}{2}10^{-15}$?

21. Sea

$$F(x) = \int_0^{\frac{\pi}{2}} \cos(x \operatorname{sen} t) dt \ .$$

a) Hallar su desarrollo en serie de Taylor en $x_0 = 0$.

b) Calcular $F(2)$ con un error menor que 10^{-6}.

c) Si quisiéramos calcular $F(20)$, ¿cuántos términos serían necesarios para obtener una precisión parecida?

d) ¿Con cuántas cifras, como mínimo, tendrían que hacerse los cálculos en el apartado c)?

22. La longitud de una elipse de semieje mayor a y excentricidad e (entre 0 y 1) viene dada por

$$L = 4a \int_0^{\frac{\pi}{2}} \sqrt{1 - e^2 \operatorname{sen}^2 \varphi} d\varphi \ .$$

a) Hallar el desarrollo de Taylor de L como función de e.

b) Si consideramos L aproximada por el primer término del desarrollo, ¿para qué valores de e el error cometido es menor que el uno por ciento?

c) Notamos que $L = 4a$ para $e = 1$; ¿para qué valores de e menores que 1 podemos aproximar L por $4a$ con un error menor que el uno por ciento?

Indicación: Usar

$$\int_0^{\frac{\pi}{2}} \operatorname{sen}^{2r} \varphi d\varphi = \frac{\pi}{2} \frac{(2r-1)!!}{(2r)!!} \ .$$

23. Queremos hallar fórmulas de integración en $m+1$ abscisas equidistantes de la forma

$$\int_a^b \frac{f(x)}{\sqrt{x}} dx \simeq \sum_{k=0}^{m} [A_k f(x_k) + B_k f'(x_k)] \ ,$$

exactas para todos los polinomios de grado menor o igual que $m+1$.

a) Explicitar estas fórmulas para $m = 1, 2, 3$, dando asimismo una expresión para los errores cometidos.

b) Aplicación: Calcular

$$\int_0^1 \frac{e^x}{\sqrt{x}} dx \ ,$$

acotando el error cometido, usando todas las fórmulas halladas en a).

24. Calcular

$$\int_0^{\pi} \cosh(3x) dx \ ,$$

a) directamente, dando el resultado con 6 cifras decimales correctas;

b) usando la regla de los trapecios con un paso h adecuado para que el error sea menor que 10^{-6};

c) usando la regla de Simpson con las mismas condiciones que en b).

25. Calcular

$$\int_0^{\pi} \frac{\cos x - 1}{x^2} dx \ ,$$

con un error menor que 10^{-10}, usando las reglas de los trapecios y de Simpson.

26. a) Hallar las reglas (compuestas) a partir de las fórmulas (simples) de integración numérica de Newton-Cotes cerradas de $m+1$ abscisas ($m = 2, 3, 4, 5$).

b) Obtener expresiones para los errores de estas reglas, suponiendo f suficientemente diferenciable.

c) Repetir el apartado a) para las fórmulas de Newton-Cotes abiertas.

27. a) Consideramos las aproximaciones dadas por la regla de los trapecios $T(h)$ para valores del paso h tendiendo a 0, al cálculo de la integral J de una función continua f en un intervalo acotado $[a, b]$. Demostrar que

$$\lim_{h \to 0} T(h) = J \ .$$

b) Generalizar este resultado a todas las fórmulas de cuadratura

$$Q_m(f) = \sum_{k=0}^{m} A_k f(x_k) \ ,$$

con

$$\sum_{k=0}^{m} A_k = b - a \ .$$

c) Deducir de b) que el resultado es cierto para todas las reglas compuestas de Newton-Cotes cerradas y abiertas.

28. a) Hallar una fórmula de integración de dos abscisas para integrales del tipo

$$\int_{-1}^{1} (1 + x^2) f(x) dx \ ,$$

que sea exacta para todos los polinomios de grado menor o igual que 3.

b) Dar una cota del error cometido en su aplicación al cálculo de

$$\int_{-1}^{1} (1 + x^2) x^4 dx \ .$$

29. Consideramos la fórmula de integración gaussiana

$$\int_a^b w(x) f(x) dx \simeq \sum_{k=0}^{m} W_k f(x_k) \ .$$

Demostrar que, si $\{\psi_j(x)\}_{j \geq 0}$ es una familia de polinomios ortogonales respecto al producto escalar continuo asociado al peso w en $[a, b]$, entonces $\{\psi_j(x)\}_{j=0 \div m}$ es una familia de polinomios ortogonales respecto al producto escalar discreto

$$(f, g)_d = \sum_{k=0}^{m} W_k f(x_k) g(x_k) \ .$$

30. a) Evaluar exactamente

$$\int_0^\pi \cos^3\theta d\theta \ ,$$

usando una fórmula de Gauss-Chebichev adecuada.

b) Calcular

$$\int_0^\pi e^{\cos x} dx \ ,$$

mediante una fórmula de Gauss-Chebichev con un número adecuado de abscisas para que el error sea menor que 10^{-8}.

31. Calcular las sumas finitas siguientes y analizar su comportamiento cuando n tiende a infinito:

$$\sum_{j=1}^{n} j \ , \quad \sum_{j=1}^{n} r^j \ , \quad \sum_{j=1}^{n} \ln\frac{j+1}{j} \ , \quad \sum_{j=1}^{n} \frac{1}{4j^2+4j+1} \ .$$

32. a) Mediante el uso de funciones factoriales, calcular y analizar el comportamiento de las sumas siguientes cuando n tiende a infinito:

$$\sum_{j=1}^{n} j^k \quad (k=2,3,4,5,6,7) \ , \quad \sum_{j=1}^{n} (3j^2+3j^5-7j^7) \ .$$

b) Estudiar el caso general

$$\sum_{j=1}^{n} P_k(j) \ ,$$

siendo $P_k(x)$ un polinomio de grado k.

33. a) Usar el método de comparación para el cálculo de las sumas de las series siguientes con un error menor que 10^{-6}:

$$\sum_{j=1}^{\infty} \frac{1}{j^2+k} \quad (k=1,2,3) \ , \quad \sum_{j=1}^{\infty} \frac{1}{j^2+j+1} \ ,$$

$$\sum_{j=1}^{\infty} \frac{1}{j^4+1} \ , \quad \sum_{j=1}^{\infty} \frac{1}{j^8+1} \ , \quad \sum_{j=1}^{\infty} \frac{j^2+3j+2}{j^4+1} \ .$$

b) Estudiar el caso general

$$\sum_{j=1}^{\infty} \frac{P_r(j)}{Q_s(j)} \quad (s \geq r+2) \ ,$$

siendo $P_r(x)$ y $Q_s(x)$ polinomios de grados r y s, respectivamente.

34. a) Sea

$$S = \sum_{j=1}^{\infty} a_j \quad (a_j > 0) \ .$$

Definimos

$$c_j = a_j + 2a_{2j} + 4a_{4j} + 8a_{8j} + \cdots$$

Probar que

$$S = \sum_{j=1}^{\infty} (-1)^{j+1} c_j \ , \quad a_j = c_j - 2c_{2j} \ .$$

b) Utilizar el resultado anterior para calcular

$$\sum_{j=1}^{\infty} \frac{1}{j^3} \ , \quad \sum_{j=1}^{\infty} \frac{1}{(2j-1)^2} \ ,$$

con 6 cifras decimales correctas.

35. Usando la fórmula de Euler-Maclaurin para sumas, calcular las sumas del problema 32.

36. a) Calcular, con un error acotado por una millonésima,

$$\sum_{j=1}^{\infty} \frac{1}{j^{\frac{4}{3}}} \ ,$$

usando la fórmula de Euler-Maclaurin para series.

b) ¿Cuántos términos se tendrían que sumar directamente para alcanzar la misma precisión?

37. Calcular, con un error acotado por una millonésima,

$$\sum_{j=123}^{369} \frac{1}{j^2+1} \ ,$$

usando la fórmula de Euler-Maclaurin para sumas.

38. La función *zeta de Riemann*

$$\zeta(x) = \sum_{j=1}^{\infty} \frac{1}{j^x} \quad (x > 1)$$

se puede aproximar sumando préviamente los 9 primeros términos y aplicando al resto la fórmula de Euler-Maclaurin para series hasta los términos que contienen derivadas quintas.

a) Demostrar que el error de la aproximación hecha es menor que 10^{-10} para cualquier x.

b) Hacer una tabla de $\zeta(x)$ para x entre 2 y 25 con paso $h = 1$ y otra para x entre 1.01 y 2 con paso $h = 0.01$.

c) Utilizar las tablas halladas para recalcular la suma de las series del problema 33.

39. Calcular con 6 cifras decimales correctas las sumas

$$\sum_{j=1}^{\infty}\left[\exp\left(\frac{1}{j^3}\right)-1\right]\ ,\ \ \sum_{j=1}^{\infty}\tan^2\frac{1}{j}\ ,\ \ \sum_{j=1}^{\infty}\operatorname{sen}^2\frac{1}{j}\ ,$$

usando las tablas de la función zeta de Riemann del problema anterior.

40. Definimos la función *digamma* por

$$\begin{aligned}\Psi(1+x) &= \sum_{j=1}^{\infty}\frac{x}{j(j+x)}-\gamma \\ &= \sum_{j=1}^{\infty}(\frac{1}{j}-\frac{1}{j+x})-\gamma\ \ (x\neq -1,-2,\ldots)\ ,\end{aligned}$$

donde γ es la constante de Euler, introducida en el problema 4.10.

a) Comprobar la fórmula recurrente

$$\Psi(k+x)=\sum_{j=1}^{k-1}\frac{1}{j+x}+\Psi(1+x)\ .$$

b) Tabular $\Psi(1+x)$, $\Psi'(1+x)$, $\Psi^{(2)}(1+x)$ y $\Psi^{(3)}(1+x)$, con 7 cifras decimales correctas y paso $h = 0.01$ en el intervalo $[0, 1]$, usando la fórmula de Euler-Maclaurin para sumas.

c) Tabular $\Psi(1+x)$, $\Psi'(1+x)$, $\Psi^{(2)}(1+x)$ y $\Psi^{(3)}(1+x)$, con 7 cifras decimales correctas y paso $h = 0.1$ en el intervalo $[-10, 10]$.

41. Usar las tablas del problema anterior para sumar las series siguientes con 6 cifras decimales correctas:

$$\sum_{j=1}^{\infty}\frac{1}{(2j+1)(j+1)^2}\ ,\ \ \sum_{j=1}^{\infty}\frac{1}{(4j+2)(4j+1)^2(4j+3)^3}\ ,$$

$$\sum_{j=1}^{\infty} \frac{2j+5}{(2j+7)j^4} \ .$$

42. Calcular con 6 cifras decimales correctas

$$P = \prod_{j=1}^{\infty} (1 + j^{-\frac{3}{2}}) \ ,$$

usando la fórmula de Euler-Maclaurin para series, aplicada a la serie $\ln P$.

43. Dibujamos un triángulo equilátero circunscrito a una circunferencia; dibujamos a continuación una circunferencia que pase por sus vértices; dibujamos después el cuadrado circunscrito a esta última circunferencia, y así sucesivamente, con todos los polígonos regulares. Estudiar el comportamiento asintótico del radio de las sucesivas circunferencias cuando el número de lados tiende a infinito.

44. a) Hallar desarrollos asintóticos asociados a las siguientes funciones, expresadas por integrales:

$$\int_x^{\infty} \frac{e^{-t}}{t} dt \ , \quad \int_x^{\infty} \frac{\operatorname{sen} t}{t} dt \ , \quad \int_x^{\infty} \frac{\cos t}{t} dt \ , \int_x^{\infty} \operatorname{sen} t^2 dt \ .$$

$$\int_x^{\infty} \frac{e^{-t}}{t^{\alpha}} dt \ , \quad \int_x^{\infty} \frac{\operatorname{sen} t}{t^{\alpha}} dt \ , \quad \int_x^{\infty} \frac{\cos t}{t^{\alpha}} dt \ , \int_x^{\infty} \operatorname{sen} t^{\alpha} dt \ , \quad (\alpha \geq 1).$$

$$\int_x^{\infty} \exp\left(-\frac{t^2}{2}\right) dt \ , \quad \int_0^{\infty} \frac{e^t}{1+xt} dt \ .$$

b) ¿Con qué precisión pueden calcularse, en función de x?

c) Aplicación: Calcularlas para $x = 10, 100, 1000$.

45. Una función F, relacionada con las integrales de Fresnel empleadas en óptica, tiene el desarrollo asintótico siguiente:

$$F(x) = 1 + \sum_{j=1}^{\infty} (-1)^j \frac{(4j-1)!!}{(\pi x^2)^{2j}} \ .$$

Sabemos que la magnitud del error al tomar n términos es menor que la del término $n+1$.

a) ¿Cuál es el valor óptimo de n para el cálculo de $F(6)$? ¿Y para el cálculo de $F(3)$?

b) Acotar los errores cometidos en ambos casos.

46. Las funciones

$$E_k(x) = \int_1^\infty \frac{e^{-xt}}{t^k} dt$$

admiten las expresiones siguientes:

$$\text{i)} \qquad \frac{e^{-x}}{x}\left(1 - \frac{k}{x} + \frac{k(k+1)}{x^2} - \frac{k(k+1)(k+2)}{x^3} + \cdots\right) ,$$

$$\text{ii)} \qquad \frac{(-x)^{k-1}}{(k-1)!}[-\ln x + \Psi(k)] - \sum_{j=0}^{\infty} \frac{(-x)^j}{(j-k+1)j!} ,$$

donde Ψ es la función digamma del problema 40 que cumple

$$\Psi(1) = -\gamma = -0.5772156649... , \quad \Psi(k) = -\gamma + \sum_{j=1}^{k-1} \frac{1}{j} \quad (k > 1) .$$

Para $n = 2, 6, 10$, indicar hasta qué valor de $x > 0$ es mejor usar la expresión i) que la ii).

47. Consideramos la *función gamma de Euler*

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

que cumple

$$\Gamma(z+1) = z\Gamma(z)$$

y, para valores enteros positivos de z,

$$\Gamma(z) = (z-1)! .$$

La *fórmula de Stirling* da el siguiente desarrollo asintótico para el logaritmo de la función gamma de Euler,

$$\ln \Gamma(x) = (x - \frac{1}{2}) \ln x - x + \frac{1}{2} \ln 2\pi + \sum_{r=1}^{\infty} \frac{B_{2r}}{2r(2r-1)x^{2r-1}} .$$

a) Para cada x, ¿cuál es el número de términos que es necesario sumar para cometer el menor error posible?

b) Calcular $999999! = \Gamma(1000000)$ con 5 cifras significativas correctas.

c) ¿Qué error mínimo tiene la fórmula cuando se aplica al cálculo de 2!?

48. La función error

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

admite los desarrollos siguientes:

$$\frac{2}{\sqrt{\pi}}\sum_{j=0}^{\infty}\frac{(-1)^j x^{2j+1}}{(2j+1)j!}\ ,\quad \frac{2}{\sqrt{\pi}}\exp(-x^2)\sum_{j=0}^{\infty}\frac{2^j}{(2j+1)!!}x^{2j+1}\ ,$$

$$1-\frac{1}{\sqrt{\pi}x}\exp(-x^2)\left(1+\sum_{j=1}^{\infty}(-1)^j\frac{(2j-1)!!}{(2x^2)^j}\right)\ .$$

Analizar cuál de ellos es mejor usar para los diferentes valores de x.

49. a) Calcular aproximaciones de la derivada de la función $f(x)=\ln x$ en $x=2$ mediante las fórmulas

$$\begin{aligned} f'(a) &\simeq \frac{1}{h}(f(a+h)-f(a))\ ,\\ f'(a) &\simeq \frac{1}{2h}(f(a+h)-f(a-h))\ , \end{aligned}$$

con pasos $h=0.1, 0.01, 0.001, 0.0001, 0.00001$.

b) Usando la expresión asintótica del error de truncamiento de las fórmulas de derivación numérica de a), deducir de ellas expresiones mejores, cuando el paso h tiende a 0.

c) Aplicar las fórmulas halladas por extrapolación en b) al cálculo de la derivada propuesta en a) y analizar los resultados.

50. a) Calcular la derivada aproximada de la función $f(x)=e^{x^2}$, en $x=2$, mediante extrapolación repetida de Richardson, a partir de la tabla de los valores de la función en las abscisas $x_i=\frac{i}{4}$ $(i=0\div 16)$, dados con 8 cifras decimales correctas redondeadas.

b) Estudiar los efectos de los errores de redondeo sobre las diferentes aproximaciones halladas.

51. a) Deducir una fórmula para el cálculo de $f^{(n)}(a)$ basada en la derivación del polinomio interpolador de grado menor o igual que $n+1$ en las abscisas $x_0=a$, x_1, x_2, $\ldots$, x_{n+1}.

b) Explicitarla para $n=1,2,3,4$ y x_i $(i=0\div n+1)$ equidistantes con paso h.

c) Aplicación: Calcular, con las fórmulas halladas en b), las 4 primeras derivadas de la función $f(x)=\cos(\cos(\operatorname{sen} x))$ en $x=a=\frac{\pi}{4}$, usando los pasos $h=0.1, 0.01$.

d) Extrapolar los resultados obtenidos.

52. Consideramos el cálculo de la integral

$$\int_0^1 \frac{\operatorname{sen} x}{x}dx\ ,$$

usando la regla de los trapecios $T(h)$ con pasos h cada vez menores.

a) Calcular $T(h)$ para $h = 2^{-k}$ $(k = 0 \div 20)$, trabajando con simple precisión (o, equivalentemente, con 6 cifras decimales y redondeo); ¿a partir de qué valor de k es inútil continuar haciendo los cálculos porque los errores de redondeo son ya superiores a los de truncamiento?

b) Desde un punto de vista teórico, supongamos que f se evalúa con un error relativo acotado por ϵ. Hallar la cota del error relativo total, que se obtiene sumando las cotas de los errores de truncamiento y de redondeo y el valor del paso (óptimo, desde este punto de vista) que minimiza esta cota del error.

53. a) Calcular la integral

$$J = \int_1^2 \ln x \; dx \; ,$$

usando la regla de los trapecios con pasos $h = 0.1, 0.01, 0.001$.

b) Evaluar los errores cometidos $E_T(h) = T(h) - J$, comparando con el resultado obtenido directamente.

c) Ajustar los resultados de a) a una expresión asintótica del error de la forma

$$E_T(h) = a_1 h^2 + a_2 h^4 + \cdots$$

d) Extrapolar los resultados anteriores y hallar una mejor aproximación del valor de J. ¿Cómo sería la expresión asintótica del error de estos valores extrapolados?

54. a) Calcular

$$\int_0^{\frac{1}{2}} e^x dx$$

de las maneras siguientes:

i) buscando la función primitiva,

ii) mediante la regla de los trapecios con pasos $h = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$,

iii) extrapolando los resultados obtenidos en b).

b) Hallar los cocientes entre los errores observados y analizar el resultado.

55. Supongamos que la función f se evalúa con un error absoluto acotado por ϵ en el intervalo $[a, b]$. Demostrar que el error en la evaluación de las aproximaciones a la integral de f en $[a, b]$, dadas por el método de Romberg, es menor que $2(b - a)\epsilon$, independientemente del paso usado.

56. a) Deducir la *fórmula de Gauss hacia adelante* de interpolación equidistante en un número par de abscisas $\{x_0, x_1, x_{-1}, \ldots, x_s, x_{-s}, x_{s+1}\}$

$$\begin{aligned} f(x_0+th) &= \sum_{r=0}^{s}\left[\binom{t+r}{2r+1}\delta^{2r+1}f_{\frac{1}{2}} + \binom{t+r-1}{2r}\delta^{2r}f_0\right] \\ &\quad + h^{2s+2}f^{(2s+2)}(\xi(t))\binom{t+s}{2s+1}\,. \end{aligned}$$

b) A partir de la fórmula anterior, deducir la *fórmula de Everett*

$$\begin{aligned} f(x_0+th) &= \sum_{r=0}^{s}\left[\binom{t+r}{2r+1}\delta^{2r}f_1 - \binom{t+r-1}{2r+1}\delta^{2r}f_0\right] \\ &\quad + h^{2s+1}f^{(2s+1)}(\xi(t))\binom{t+s}{2s+1}\,, \end{aligned}$$

y la *fórmula de Bessel*

$$\begin{aligned} f(x_0+th) &= \sum_{r=0}^{s}\left[\binom{t+r-1}{2r}\mu\delta^{2r}f_{\frac{1}{2}}\right. \\ &\quad \left. + \frac{1}{2r+1}\left(t-\frac{1}{2}\right)\binom{t+r-1}{2r}\delta^{2r+1}f_{\frac{1}{2}}\right] \\ &\quad + h^{2s+1}f^{(2s+1)}(\xi(t))\binom{t+s}{2s+1}\,. \end{aligned}$$

c) Aplicación: Usar las fórmulas de Everett y de Bessel para calcular $\operatorname{sen} 1.05$, con 9 cifras decimales correctas, a partir de la tabla de $\operatorname{sen}(1+x)$ en las abscisas $x_k = \frac{k}{10}$ $(k = -10 \div 10)$.

57. a) Derivar las fórmulas de interpolación de Newton con diferencias hacia adelante y hacia atrás para encontrar fórmulas de derivación en la abscisa x_0 y el error cometido correspondiente.

b) Derivar la fórmula de Stirling del problema 4.14 para hallar una fórmula de derivación en la abscisa central x_0 y el error cometido.

c) Aplicación: Calcular, mediante dichas fórmulas, la derivada de la función $f(x) = \operatorname{sen}(1+x)$ en $x = 0$, usando la tabla mencionada en el problema anterior.

d) Repetir los apartados anteriores, añadiéndoles el cálculo de fórmulas para derivadas segundas y terceras.

58. a) Deducir la *regla de los trapecios con términos correctivos hasta orden 5*

$$\int_{x_0}^{x_M} f(x)dx \simeq \frac{h}{2}(f_0 + 2f_1 + 2f_2 + \cdots + 2f_{M-1} + f_M)$$

$$-\frac{h}{12}(\mu\delta f_M - \mu\delta f_0) + \frac{11h}{720}(\mu\delta^3 f_M - \mu\delta^3 f_0)$$
$$-\frac{191h}{60480}(\mu\delta^5 f_M - \mu\delta^5 f_0)\ ,$$

integrando sobre el intervalo $[x_i, x_{i+1}]$ la fórmula de interpolación de Bessel de grado 5 en las abscisas x_{i-k} $(k = -2 \div 3)$ y efectuando la suma para $i = 0 \div M-1$.

b) Deducir la *regla de Simpson con términos correctivos hasta orden 6*

$$\begin{aligned}\int_{x_0}^{x_{2M}} f(x)dx \ \simeq\ & \frac{h}{3}(f_0 + 4f_1 + 2f_2 + \cdots \\ & \quad +2f_{2M-2} + 4f_{2M-1} + f_{2M}) \\ & -\frac{h}{90}(\delta^4 f_1 + \delta^4 f_3 + \cdots + \delta^4 f_{2M-1}) \\ & +\frac{h}{756}(\delta^6 f_1 + \delta^6 f_3 + \cdots + \delta^6 f_{2M-1})\ ,\end{aligned}$$

integrando sobre el intervalo $[x_{2r}, x_{2r+1}]$ la fórmula de interpolación de Stirling de grado 6 en las abscisas x_{i-k} $(k = -3 \div 3)$ y efectuando la suma para $r = 0 \div 2M-1$.

c) Deducir la *fórmula de Gregory con términos correctivos hasta orden 6*

$$\begin{aligned}\int_{x_0}^{x_M} f(x)dx \ \simeq\ & \frac{h}{2}(f_0 + 2f_1 + 2f_2 + \cdots + 2f_{M-1} + f_M) \\ & -\frac{h}{12}(\nabla f_M - \Delta f_0) - \frac{h}{24}(\nabla^2 f_M + \Delta^2 f_0) \\ & -\frac{19h}{720}(\nabla^3 f_M - \Delta^3 f_0) - \frac{3h}{160}(\nabla^4 f_M + \Delta^4 f_0) \\ & -\frac{863h}{60480}(\nabla^5 f_M - \Delta^5 f_0)\ ,\end{aligned}$$

a partir de la fórmula de Euler-Maclaurin, substituyendo las derivadas sucesivas en las abscisas x_0 y x_M por las derivadas correspondientes de las fórmulas de Newton con diferencias hacia adelante y hacia atrás, respectivamente.

CAPÍTULO 5

ECUACIONES NO LINEALES

La resolución de ecuaciones no lineales es un problema habitual en cualquier rama de la ciencia aplicada y la técnica. Los métodos de resolución numérica son iterativos, partiendo de aproximaciones iniciales de la solución buscada. La generación de los distintos métodos iterativos se presenta de forma unificadora a partir de la utilización de la interpolación directa e inversa. En particular, se describen métodos específicos de localización, separación y aproximación numérica de raíces de ecuaciones polinomiales.

5.1 ECUACIONES EN UNA VARIABLE

5.1.1 Introducción

Denominamos *ecuación no lineal* a una ecuación del tipo $f(x) = 0$ (a veces la expresaremos por $x = g(x)$) en la cual f es una función real de variable real no lineal. La función f puede ser polinomial (caso que estudiaremos en la segunda parte de este capítulo), trascendente (aparecen en su expresión funciones exponenciales, logarítmicas y trigonométricas) e, incluso, puede ocurrir que no se disponga de una expresión explícita de $f(x)$, pero que se conozcan las reglas para su cálculo. Un ejemplo de este caso, frecuente en matemática aplicada, se da cuando $f(x)$ es el valor de la solución de una ecuación diferencial, después de un tiempo determinado, partiendo de unas condiciones iniciales representadas por x.

Llamamos *raíz* o *solución* de una ecuación no lineal $f(x) = 0$ a un valor α tal que $f(\alpha) = 0$. En este caso se dice también que α es un *cero* de f.

En general, las raíces de una ecuación no lineal no pueden ser calculadas de forma exacta. El objetivo de este capítulo consiste en ofrecer métodos numéricos que permitan obtener aproximaciones numéricas de las mismas.

En cualquier proceso de cálculo de raíces de una ecuación no lineal pueden distinguirse tres fases:

- Localización

  Interesa tener un cierto conocimiento de la zona en la que se encuentran las raíces. En general, esta información se obtendrá bien mediante la confección de tablas de la función, bien a partir de un estudio analítico o, incluso, a partir de una representación gráfica aproximada, cuando la complejidad de la función no lo impida. En muchos

casos, las ecuaciones provienen de un problema técnico o científico, cuyo conocimiento puede ayudar a la localización de las raíces.

- Separación

  A veces dos raíces diferentes de una ecuación están muy próximas. En estos casos, antes de aplicar cualquier método numérico, conviene separar las raíces; esto es, determinar intervalos que contengan una y sólo una raíz de la ecuación.

- Aproximación numérica

  Su objetivo, en general, consistirá en construir una sucesión de valores que converja hacia la raíz buscada. Esta construcción se hará, normalmente, de manera iterativa partiendo de valores iniciales que supondremos suficientemente próximos a la raíz buscada: a partir de $x_0, \ldots, x_m$, obtendremos $x_{m+1} = G(x_m, \ldots, x_0)$ y, de una manera más general, los iterados $x_{k+1} = G(x_k, \ldots, x_{k-m})$.

A continuación se presentan algunos métodos de aproximación de raíces. Más adelante dedicaremos un apartado al estudio de su convergencia.

5.1.2 Métodos iterativos de aproximación de soluciones

Método de bisección

Una vez encontrados unos valores a y b donde la función f cambia de signo, el teorema de Bolzano asegura que, si f es continua en $[a, b]$ y $f(a)f(b) < 0$, entonces existe $\alpha \in (a, b)$ tal que $f(\alpha) = 0$. De hecho, puede existir más de una raíz, pero nosotros supondremos que existe sólo una, entendiendo que previamente se habrán efectuado procesos de localización y separación. El método de bisección construye entonces una sucesión de intervalos encajados,

$$(a_0, b_0) \supset (a_1, b_1) \supset \ldots \supset (a_k, b_k) \supset \ldots,$$

de manera que siempre contengan la raíz buscada y que la amplitud de cada uno sea la mitad de la del anterior. Cuando la amplitud del intervalo sea suficientemente pequeĄa de acuerdo con la precisión deseada para la raíz, podremos considerar como una buena aproximación de ésta uno cualquiera de sus extremos.

Para la construcción de la sucesión de intervalos se parte de a_0 y b_0 tales que $f(a_0)f(b_0) < 0$ y se considera la abscisa media de (a_0, b_0), $c = \frac{1}{2}(a_0 + b_0)$. Si $f(c) = 0$, c es la raíz buscada. De lo contrario, se escogerá como (a_1, b_1) el intervalo (a_0, c) o el (c, b_0) según sea $f(a_0)f(c) < 0$ o $f(c)f(b_0) < 0$, respectivamente. El proceso iterativo se continúa análogamente (véase la figura 5.1).

Este método, aunque converge siempre hacia la raíz buscada, tiene el inconveniente de no aprovechar, salvo el signo, las características concretas de la función f, razón por la cual es bastante lento.

Método de Newton

Supongamos ahora que f es derivable en un entorno de la raíz buscada.

El objetivo de este método consiste en construir una sucesión $(x_k)_{k\geq 0}$, convergente hacia la raíz. Para ello se parte de un valor x_0, se traza la tangente a la curva $y = f(x)$

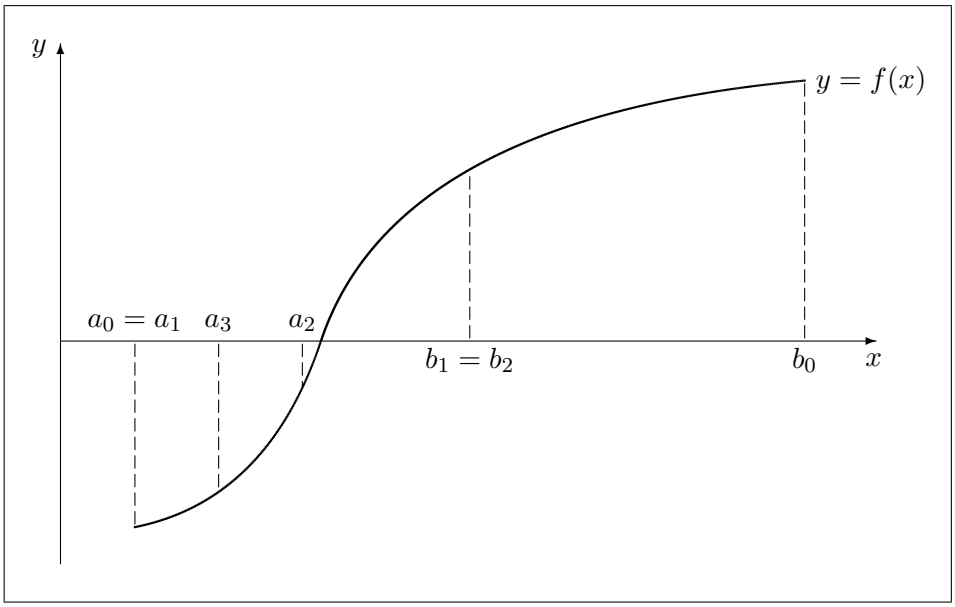


Figura 5.1: Método de bisección.

por el punto $(x_0, f(x_0))$ y se busca su punto de corte con el eje $y = 0$. La abscisa de este punto será x_1 y así sucesivamente (véase la figura 5.2).

En general,

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \ (k \geq 0) \ ,$$

suponiendo que f' no se anula en los diferentes x_k.

La convergencia de este método no está asegurada para cualquier x_0 y es conveniente escoger la aproximación inicial x_0 tan próxima como sea posible a la raíz. Ahora bien, tal como se verá más adelante, su convergencia suele ser bastante rápida; por esto, es uno de los métodos más usados.

El método de Newton presenta problemas cuando el cero α de f que se busca es múltiple; esto es, tal que $f'(\alpha) = 0$. Pero puede ser modificado de manera trivial para adaptarlo a este caso (véase el problema 5.3).

El criterio general usado para acabar la construcción de esta sucesión es que $\mid x_k - x_{k-1} \mid$ sea suficientemente pequeĄo. En este caso, x_k podrá considerarse como una buena aproximación de la raíz buscada.

Hay que hacer notar aquí, aunque lo que se dice vale para cualquier método, que el criterio para terminar las iteraciones no tiene porqué ser siempre el que la corrección $|x_k - x_{k-1}|$ sea menor que un valor δ dado. Así, si α es la solución exacta y suponemos que $f'(\alpha)$ tiene un valor relativamente pequeĄo y que los errores en el cálculo de f están acotados por ϵ, entonces no se pueden pedir errores en la determinación de α menores que $\frac{\epsilon}{|f'(\alpha)|}$. Si $\frac{\epsilon}{|f'(\alpha)|} > \delta$, entonces el criterio adecuado para acabar las iteraciones no será $\mid x_k - x_{k-1} \mid < \delta$, sino $\mid f(x_k) - f(x_{k-1}) \mid < 2\epsilon$.

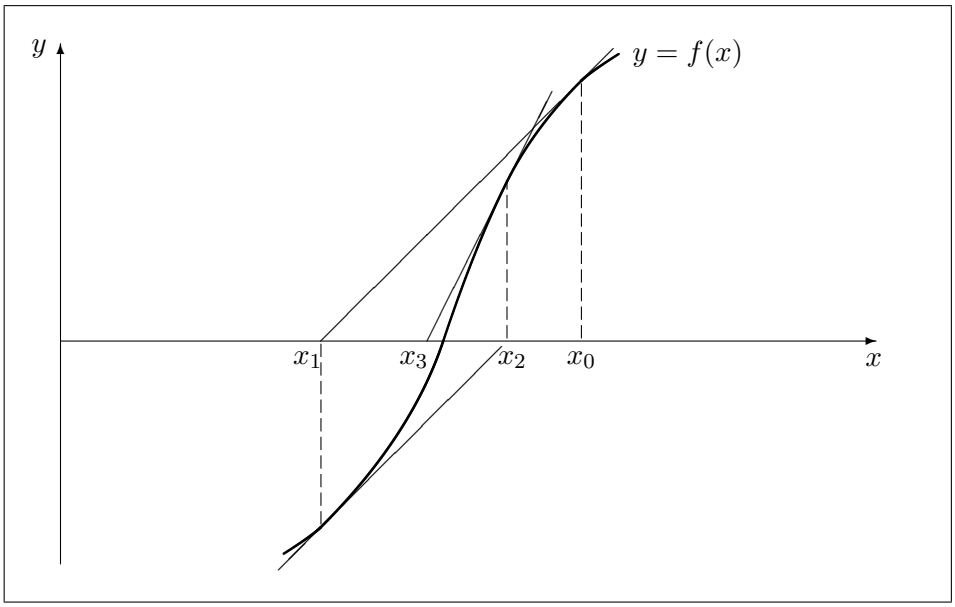


Figura 5.2: Método de Newton.

Método de la secante

El objetivo del método de la secante consiste en la construcción de una sucesión x_0, x_1, x_2, ..., x_k, ... que converja hacia la raíz buscada. Para llevar a cabo esta construcción, se parte de dos valores x_0 y x_1 y se traza la recta secante (interpolación lineal) a la curva $y = f(x)$ por los puntos $(x_0, f(x_0))$ y $(x_1, f(x_1))$, tomándose x_2 como la abscisa del punto de corte de esta recta con el eje $y = 0$. El proceso continúa encontrando x_3 a partir de x_1 y x_2 y así sucesivamente (véase la figura 5.3); en general:

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \ (k \geq 1) \ .$$

De hecho, es como el método de Newton, pero tomando la secante como aproximación de la tangente.

Si bien la convergencia no está asegurada (conviene escoger x_0 y x_1 tan próximos como sea posible a la raíz), cuando se da, es bastante rápida, como se verá más adelante.

La elección entre la utilización del método de la secante y la del método de Newton depende básicamente de la magnitud de la labor necesaria para el cálculo de $f'(x_k)$. Puede decirse que, si el coste para este cálculo es mayor que 0.44 veces el coste de cálculo de $f(x_k)$, resulta más eficiente usar el método de la secante (véase el problema 5.5).

Una variante del método de la secante que recuerda al de bisección es el llamado *método de la regula falsi*. En éste se parte de dos valores x_0 y x_1 tales que $f(x_0)f(x_1) < 0$ y se encuentra x_2 por el método de la secante. En el cálculo del valor siguiente x_3 no se utilizarán necesariamente x_1 y x_2 (como en el método de la secante), sino x_0 y x_2 o x_1 y x_2 según sea $f(x_0)f(x_2) < 0$ o $f(x_1)f(x_2) < 0$, respectivamente, y así sucesivamente (véase la figura 5.4). Aunque este método sea más lento que el de la secante, tiene la ventaja de que siempre es convergente.

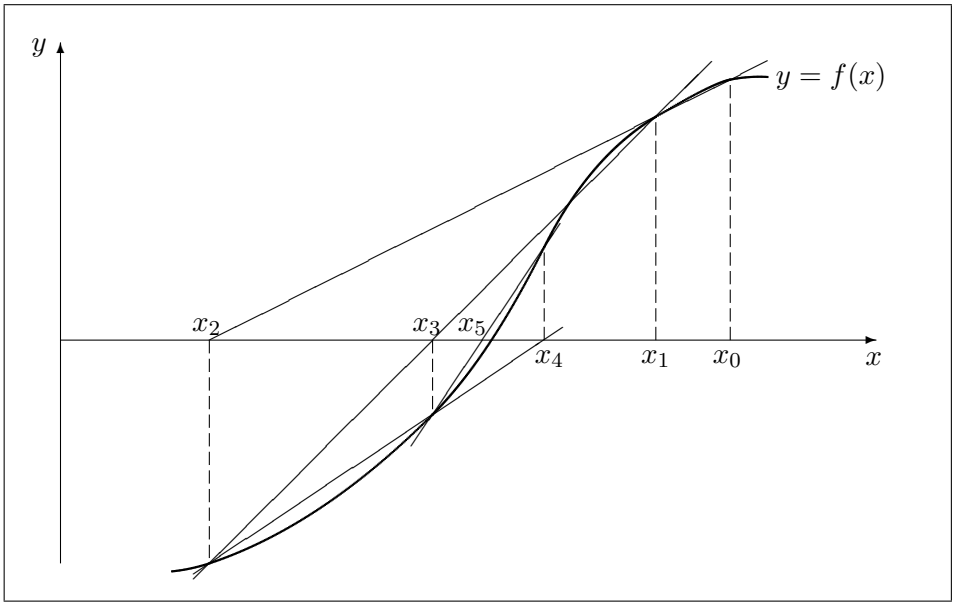


Figura 5.3: Método de la secante.

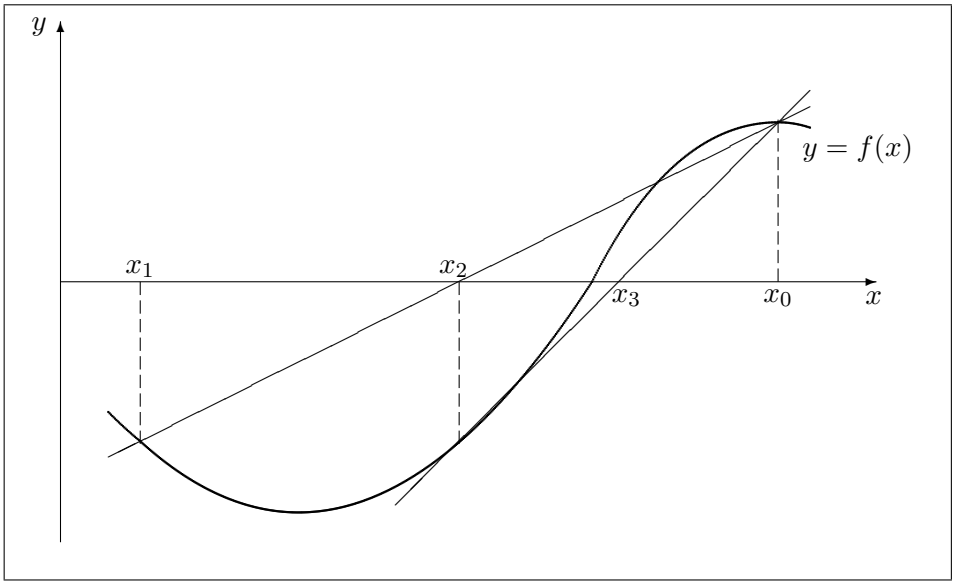


Figura 5.4: Método de la regula falsi.

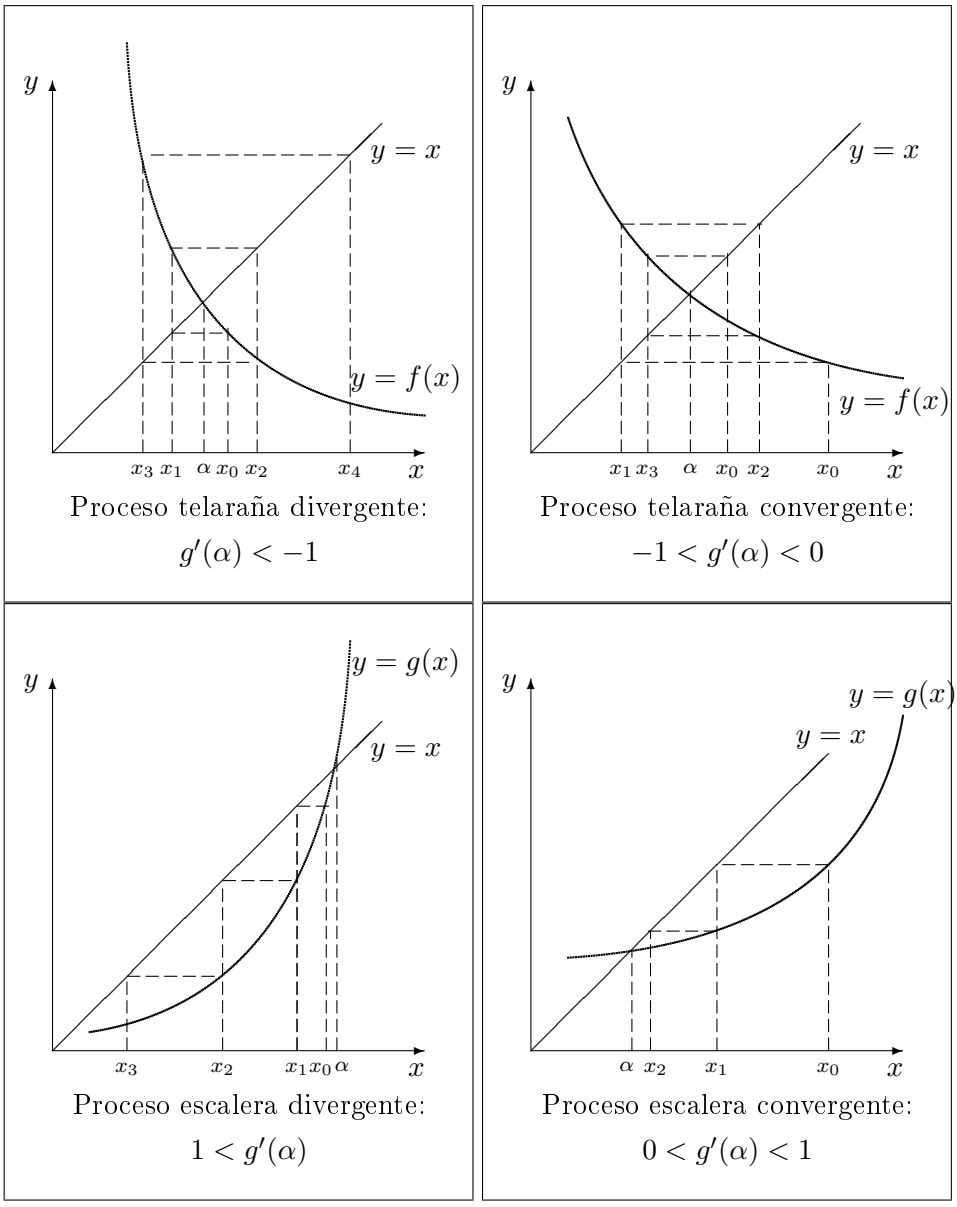


Figura 5.5: Procesos telaraña y escalera.

Método de iteración simple

En este caso escribiremos la ecuación a resolver en la forma $x = g(x)$; a una raíz α de una ecuación de este tipo, $\alpha - g(\alpha) = 0$, la llamaremos *punto fijo de* g. Nuevamente, pretendemos generar una sucesión de valores tendiendo a la raíz: $x_0, x_1, x_2, \ldots, x_k, \ldots$ La construcción de esta sucesión se realiza mediante *iteraciones simples* $x_{k+1} = g(x_k)$ $(k \geq 0)$, a partir de un x_0 escogido próximo a la raíz buscada.

La convergencia, que no está asegurada, depende de la elección de x_0 y de la función g. En caso de que g sea derivable, un criterio para que, escogiendo x_0 suficientemente

próximo a la raíz α, tengamos convergencia es que $|\,g'\,| \leq r < 1$ en un entorno del valor α (en este caso diremos que la función g es *contractiva* cerca del punto fijo α); este entorno existe si g' es continua y $|g'(\alpha)| < 1$.

Las cuatro representaciones de la figura 5.5 ilustran la convergencia del método según el valor de la derivada de la función g en α.

5.1.3 Orden de convergencia y constante asintótica del error

Diremos que una sucesión $(x_k)_{k\geq 0}$, convergente a un límite α, tiene *orden de convergencia al menos* p si existen $N \geq 0$ y $C > 0$ tales que

$$|x_{k+1} - \alpha| \ \leq C \mid x_k - \alpha \mid^p$$

para cualquier $k \geq N$. Cuando $p = 1$, hay que exigir que $C < 1$.

En particular, en el caso en que exista

$$L = \lim_{k\to\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^p} \ ,$$

la sucesión $(x_k)_{k\geq 0}$ tendrá orden de convergencia al menos p (si $p = 1$, tendremos que exigir además que $|L| < 1$). Si $L \neq 0$, diremos que la sucesión $(x_k)_{k\geq 0}$ tiene *orden de convergencia* p y que L es la *constante asintótica del error*. Para $p = 1, 2, 3$, hablaremos de convergencia al menos *lineal, cuadrática, cúbica*, respectivamente.

Diremos que un método iterativo $x_{k+1} = G(x_k, x_{k-1}, \ldots, x_{k-m})$ $(k \geq m)$ es al menos de *orden* p cerca de un punto fijo α de G: $\alpha = G(\alpha, \ldots, \alpha)$, cuando exista un entorno $(\alpha - \epsilon, \alpha + \epsilon)$ de manera que, para cualquier elección de valores iniciales $x_0, \ldots, x_m$ en $(\alpha - \epsilon, \alpha + \epsilon)$, la sucesión $(x_k)_{k\geq 0}$ tenga al menos orden de convergencia p.

Si tenemos un método de orden de convergencia al menos p y no tenemos en cuenta los errores de redondeo (es decir, dados $x_k, \ldots, x_{k-m}$, suponemos que podemos calcular exactamente $x_{k+1} = G(x_k, \ldots, x_{k-m})$), entonces, llamando ϵ_k a $|\,x_k - \alpha\,|$ y suponiendo que $\epsilon_i \leq C\epsilon_{i-1}^p$ $(i = 1 \div k)$, se cumple:

$$\begin{aligned} \epsilon_k &\leq C^k \epsilon_0 \quad (p = 1) \ , \\ \epsilon_k &\leq C^{1+p+\ldots+p^{k-1}} \epsilon_0^{p^k} = \frac{(\overline{C}\epsilon_0)^{p^k}}{\overline{C}} \ , \quad (p > 1) \ , \end{aligned}$$

con

$$\overline{C} = C^{\frac{1}{p-1}} \ .$$

Así, cuanto mayor sea el orden de convergencia, más rápidamente convergerá $(x_k)_{k\geq 0}$ a α.

De todas maneras, en el estudio de la eficiencia de un método, no se ha de tener en cuenta tan sólo su orden de convergencia, sino también la cantidad de cálculo necesario en cada iteración $x_{k+1} = G(x_k, \ldots, x_{k-m})$ $(k \geq m)$.

Seguidamente agrupamos algunas consideraciones sobre orden de convergencia:

- I. Consideremos la ecuación $x = g(x)$ y sea α una solución.

1. Si g es contractiva en un entorno de α, el método de iteración simple $x_{k+1} = g(x_k)$ $(k \geq 0)$ es al menos de orden 1.
2. Si g es m veces derivable y $g'(\alpha) = \ldots = g^{(m-1)}(\alpha) = 0$, entonces el método de iteración simple $x_{k+1} = g(x_k)$ $(k \geq 0)$ es al menos de orden m; si $g^{(m)}(\alpha) \neq 0$, es de orden m.

- II. Consideremos ahora la ecuación $f(x) = 0$ y sea α un cero de f: $f(\alpha) = 0$.

1. Si $f'(\alpha) \neq 0$, el método de Newton tiene convergencia al menos cuadrática.
2. Si $f'(\alpha) = 0$, el método de Newton tiene convergencia lineal.
3. Si $f'(\alpha) \neq 0$, el método de la secante tiene orden de convergencia al menos $p = \frac{1+\sqrt{5}}{2}$.
4. Si $f^{(2)}(\alpha) \neq 0$, el método de la regula falsi tiene convergencia al menos lineal. En comparación con el método de la secante, se asegura la convergencia a cambio de una pérdida de velocidad de aproximación.
5. Independientemente del comportamiento de f, puede considerarse que el método de bisección tiene convergencia lineal ya que, en cada paso, reducimos a la mitad el intervalo de error donde se encuentra el cero buscado.

5.1.4 Aceleración de la convergencia

Observemos que los métodos propuestos para encontrar raíces de ecuaciones no lineales tienen como objetivo común la construcción de una sucesión de valores $(x_k)_{k\geq 0}$ convergente (en el mejor de los casos) hacia la raíz buscada. La rapidez con que se produce esta convergencia depende del método usado y de la propia ecuación. Los llamados *métodos de aceleración de la convergencia* tienen como objetivo la obtención, a partir de la sucesión $(x_k)_{k\geq 0}$, de una nueva sucesión $(x_k^{,})_{k\geq 0}$ más rápidamente convergente hacia la raíz. Veremos dos de estos métodos.

Método de aceleración de Aitken

Supongamos que la sucesión $(x_k)_{k\geq 0}$ de límite α, con $x_k \neq \alpha$, es tal que el error se comporta asintóticamente como una sucesión geométrica de razón menor que 1, en valor absoluto; es decir,

$$x_{k+1} - \alpha = (L + \delta_k)(x_k - \alpha) \ ,$$

con

$$\lim_{k\to 0} \delta_k = 0 \ , \quad |L| < 1 \ ;$$

entonces, la sucesión $(x_k^{,})_{k\geq 0}$ dada por

$$x_k^{,} = x_k - \frac{(\Delta x_k)^2}{\Delta^2 x_k} = x_k - \frac{(x_{k+1} - x_k)^2}{x_{k+2} - 2x_{k+1} + x_k} \ (k \geq 0) \ ,$$

está bien definida para k suficientemente grande y cumple que

$$\lim_{k\to\infty} \frac{x_k^{,} - \alpha}{x_k - \alpha} = 0 \ .$$

Por tanto, $(x'_k)_{k\geq 0}$ converge hacia α más rápidamente que $(x_k)_{k\geq 0}$; este método permite, así, acelerar la convergencia de métodos con convergencia lineal.

Método de aceleración de Steffensen

Si tenemos una iteración simple $x_{k+1} = g(x_k)$ $(k \geq 0)$, el *método de aceleración de Steffensen* construye la iteración simple dada por $y_{k+1} = G(y_k)$, con

$$G(x) = x - \frac{(g(x) - x)^2}{g(g(x)) - 2g(x) + x} \quad (k \geq 0) \ .$$

Se cumple:

1. Si $G(\alpha) = \alpha$, entonces $g(\alpha) = \alpha$.
2. Si $g(\alpha) = \alpha$ y $g'(\alpha) \neq 1$, entonces $G(\alpha) = \alpha$.
3. Si $(x_k)_{k\geq 0}$ tiene convergencia al menos lineal, entonces $(y_k)_{k\geq 0}$ tiene convergencia al menos cuadrática.
4. Si $(x_k)_{k\geq 0}$ tiene orden de convergencia al menos $p > 1$, entonces $(y_k)_{k\geq 0}$ tiene orden de convergencia al menos $2p - 1$.

Al aplicar este método al caso de convergencia lineal, obtenemos una aceleración de convergencia ya que $(y_k)_{k\geq 0}$ converge más rápidamente que $(x_k)_{k\geq 0}$. Cuando el orden de convergencia del método iterativo inicial sea $p > 1$, la aplicación del método no tiene ventajas: el método de aceleración de Steffensen consigue acelerar la convergencia hasta orden $2p - 1$ mientras que, tomando sencillamente $\overline{G}(x) = g(g(x))$ (cálculo por otro lado necesario en la evaluación de $G(x)$ por el método de Steffensen), puede generarse una sucesión de iterados de orden p^2. En el problema 5.2 se presenta otra expresión de la iteración simple de Steffensen utilizable cuando tenemos ecuaciones del tipo $f(x) = 0$.

La diferencia fundamental entre los métodos de aceleración de Aitken y de Steffensen aplicados a un método de iteración simple, radica en el hecho de que, en el método de Aitken, la aceleración se produce sobre la sucesión de iterados ya construida de entrada, mientras que, en el método de Steffensen, se construye directamente la sucesión acelerada. En general, el método de Steffensen ofrece mejores resultados que el de Aitken ya que, para el cálculo de los sucesivos términos de $(y_k)_{k\geq 0}$, utiliza valores que, de hecho, han estado ya acelerados.

5.1.5 Clasificación de métodos iterativos

Se acaban de presentar diversos métodos iterativos para el cálculo aproximado de raíces de ecuaciones. Todos ellos tienen en común la construcción de una sucesión x_0, x_1, x_2, ..., x_k, ... que, bajo determinadas condiciones, tiende a la raíz α buscada. El objetivo de este apartado es dar una visión global de estos procedimientos, abriendo el horizonte a otros métodos que, si bien tienen el mismo fundamento que los presentados, no son tan usados por la dificultad de cálculo que comportan.

Los métodos iterativos de aproximación de ceros de una función f pueden clasificarse atendiendo a dos rasgos que los caracterizan:

1. La herramienta que se usa para el estudio de la función:

   - El desarrollo de Taylor: calculamos x_{k+1} a partir del conocimiento de la función y de sus derivadas en un punto x_k. Se hablará de *métodos de Taylor*.
   - La interpolación: calculamos x_{k+1} a partir del conocimiento de la función en $x_k, x_{k-1}, \ldots$ Se hablará de *métodos de interpolación*.

2. La función que se desarrolla o se interpola: si $g = f^{-1}$ (nótese que la inversa de f existirá siempre cerca de ceros simples de f), será equivalente buscar α tal que $f(\alpha) = 0$ a buscar α tal que $g(0) = \alpha$. Así podremos dar dos enfoques del problema según usemos el desarrollo de Taylor o la interpolación de la función f o de la función g:

   - Si usamos la función f, se hablará de *métodos directos*.
   - Si usamos la función $g = f^{-1}$, se hablará de *métodos inversos*.

Estas características clasifican los métodos iterativos tratados aquí en cuatro familias: métodos de Taylor directos, métodos de Taylor inversos, métodos de interpolación directos y métodos de interpolación inversos.

Métodos de Taylor directos

El cálculo de x_{k+1} a partir de x_k, en estos métodos, se basa en el procedimiento siguiente que, finalmente, se resume en una fórmula iterativa:

- Se toma como aproximación de $f(x)$ su polinomio de Taylor $p_n(x)$ de grado menor o igual que n en x_k.
- Se resuelve la ecuación $p_n(x) = 0$. La solución obtenida (si hay más de una, se escoge la más próxima a x_k) será el valor x_{k+1}.
- Se repite el proceso a partir de x_{k+1} hasta obtener la precisión deseada.

Los diferentes métodos de esta familia aparecen según el valor de n que se tome:

- Para $n = 1$, resulta el método de Newton ya estudiado.
- Para $n = 2$, resulta la iteración simple

$$x_{k+1} = x_k - \frac{2f(x_k)}{f'(x_k) \pm \sqrt{f'^2(x_k) - 2f(x_k)f^{(2)}(x_k)}} .$$

- Para $n > 2$, la ecuación $p_n(x) = 0$ no permite un cálculo sencillo de x_{k+1}.

Métodos de Taylor inversos

El cálculo de x_{k+1} a partir de x_k, en estos métodos, se basa en el procedimiento siguiente que, finalmente, se resume en una fórmula iterativa:

- Se toma como aproximación de $g(y) = f^{-1}(y)$ su polinomio de Taylor $q_n(y)$ de grado menor o igual que n en $y_k = f(x_k)$. Las derivadas sucesivas de g se obtienen derivando repetidamente la relación $g(f(x)) = x$ (véase el problema 5.1).
- Se escoge $x_{k+1} = q_n(0)$. Será una nueva aproximación del valor α buscado.
- Se repite el proceso a partir de x_{k+1} hasta obtener la precisión deseada.

Los diferentes métodos de esta familia aparecen según el valor de n que se tome:

- Para $n = 1$, resulta nuevamente el método de Newton.
- Para $n = 2$, resulta el *método de Chebichev*, contemplado detalladamente en el problema 5.1.
- Para $n = 3$, resulta el método iterativo:

$$\begin{aligned} x_{k+1} &= x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f^{(2)}(x_k)f^2(x_k)}{2f'^3(x_k)} \\ &\quad - \frac{3f^{(2)2}(x_k) - f^{(3)}(x_k)f'(x_k)}{6f'^5(x_k)} f^3(x_k) \ . \end{aligned}$$

  Si bien el orden de este método es alto, el exceso de cálculo de cada iteración impone una seria dificultad en lo referente a su eficiencia.
- Para $n > 3$, las expresiones que aparecen no permiten un cálculo sencillo de x_{k+1}.

Métodos de interpolación directos

El cálculo de x_{k+1} a partir de x_k, x_{k-1}, ..., x_0, en estos métodos, se basa en el procedimiento siguiente, que, finalmente, se resume en una fórmula iterativa:

- Se consideran $m + 1$ aproximaciones del valor α buscado: x_k, x_{k-1}, ..., x_{k-m}.
- Se construye el polinomio interpolador $p_m(x)$ a la función f en las abscisas x_k, x_{k-1}, ..., x_{k-m}.
- Se resuelve la ecuación $p_m(x) = 0$. La solución obtenida más próxima a x_k será el valor x_{k+1}.
- Se repite el proceso a partir de x_{k+1}, x_k, ..., x_{k-m+1} hasta obtener la precisión deseada.

Los diferentes métodos de esta familia aparecen según el valor de m que se tome:

- Para $m = 1$, resulta el método de la secante ya estudiado.
- Para $m = 2$, resulta el *método de Muller-Traub* que se expone en la segunda parte de este capítulo ya que generalmente se usa para el cálculo de ceros de polinomios.
- Para $m > 2$, la ecuación $p_m(x) = 0$ no permite un cálculo sencillo de x_{k+1}.

Métodos de interpolación inversos

El cálculo de x_{k+1} a partir de x_k, x_{k-1}, ..., x_0, en estos métodos, se basa en el procedimiento siguiente que, finalmente, se resume en una fórmula iterativa:

- Se consideran $m+1$ aproximaciones del valor α buscado: x_k, x_{k-1}, ... , x_{k-m}.
- Se considera la función $g = f^{-1}$.
- Se construye el polinomio interpolador $q_m(y)$ a la función g en los valores $f(x_k)$, $f(x_{k-1})$, ..., $f(x_{k-m})$.
- Se escoge $x_{k+1} = q_m(0)$. Será una nueva aproximación del valor $\alpha = g(0)$ buscado.
- Se repite el proceso a partir de x_{k+1}, x_k, ..., x_{k-m+1} hasta obtener la precisión deseada.

Los diferentes métodos de esta familia aparecen según el valor de m que se tome:

- Para $m = 1$, resulta nuevamente el método de la secante.
- Para $m = 2$, resulta el método que se trata en el problema 5.4.
- Para $m > 2$, la expresión de $q_m(y)$ no permite un cálculo sencillo de x_{k+1}.

5.2 ECUACIONES POLINOMIALES

5.2.1 Introducción

El teorema fundamental del álgebra nos asegura que, dado un polinomio de grado n con coeficientes complejos

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0 \ ,$$

con $a_n \neq 0$ y $a_i \in \mathbb{C}$ $(i = 0 \div n)$, existen n ceros complejos $\alpha_1, \alpha_2, \ldots, \alpha_n$ (repetidos según su multiplicidad) y, entonces,

$$p(x) = a_n (x - \alpha_1)(x - \alpha_2) \ldots (x - \alpha_n) \ .$$

Es también conocido que, si a_0, a_1, ..., a_n son reales, los ceros complejos de $p(x)$ aparecen en parejas conjugadas.

Para la obtención de los ceros de polinomios de grado 2, 3 ó 4, se conocen fórmulas generales; en cambio, se sabe que estas fórmulas no existen para grados superiores a 4. Así, si exceptuamos los casos triviales en que la factorización del polinomio aparezca como evidente, convendrá utilizar métodos numéricos para la búsqueda de los ceros. Incluso para polinomios de tercer y cuarto grado, el uso de las fórmulas no es sencillo y se prefieren los procedimientos numéricos.

Al usar los métodos numéricos en polinomios, disfrutaremos de unas ventajas particulares: en primer lugar, la facilidad en la evaluación de $p(x)$ y de sus derivadas y, en segundo lugar, la existencia de reglas especiales para la acotación y separación de sus ceros. Además de los métodos estudiados de aproximación de ceros, existen métodos específicos para polinomios. También conviene subrayar que, conocido un cero de un polinomio, podemos buscar los ceros restantes como ceros de un polinomio de grado menor. A continuación estudiamos estos procedimientos específicos para polinomios.

5.2.2 Evaluación y deflación de polinomios

La evaluación de polinomios, y también la de sus derivadas, puede ser llevada a cabo de forma eficiente mediante la regla de Horner, como se explica en el apartado 3.1.2.

La regla de Horner nos ofrece también la posibilidad de obtener el polinomio $p_1(x)$ que aparece al eliminar de $p(x)$ un cero α_1; esto es,

$$p(x) = p_1(x)(x - \alpha_1) \ .$$

El resto de ceros de $p(x)$ lo buscaremos calculando los ceros de $p_1(x)$, cuyo grado es una unidad menor que el de $p(x)$. Gracias a este procedimiento, llamado *deflación de polinomios*, cada vez que obtengamos un cero de un polinomio podremos buscar el resto de ceros trabajando con un polinomio de un grado menor. Si α_2 es un cero del nuevo polinomio $p_1(x)$, podremos realizar una nueva deflación y continuar trabajando con un polinomio $p_2(x)$ de un grado menor, y así sucesivamente.

Ahora bien, dado que los ceros $\alpha_1, \alpha_2, \ldots$ sólo se conocen aproximadamente, $p_1(x)$ está afectado por errores que se propagarán al calcular $p_2(x)$, etc., con lo cual puede ocurrir que los últimos ceros $\ldots, \alpha_{n-1}, \alpha_n$ que calculemos tengan un error muy grande respecto a los ceros exactos del polinomio $p(x)$.

Para que estos errores en los ceros calculados no se amplifiquen considerablemente, convendrá calcular los ceros α_1, α_2, ... de menor a mayor en módulo $|\alpha_1| \leq |\alpha_2| \leq \ldots$; es decir, conviene calcular el cero de módulo menor de cada polinomio intermedio $p_i(x)$.

Este proceso también puede realizarse en sentido contrario; es decir, calculando el cero de mayor valor absoluto de cada polinomio intermedio $p_i(x)$, siempre y cuando la deflación se realice hacia atrás.

Dado el polinomio

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0 \ ,$$

con a_n y a_0 no nulos, la *deflación hacia atrás* parte del polinomio

$$q(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n \ ,$$

que cumple $p(x) = x^n q(\frac{1}{x})$ para $x \neq 0$ y, por lo tanto, $p(\alpha) = 0$ si y sólo si $q(\frac{1}{\alpha}) = 0$. A continuación, en lugar de realizar la deflación estándar, realiza la deflación

$$q_1(x) = \frac{q(x)}{x - \frac{1}{\alpha}} = b_0 x^{n-1} + \ldots + b_{n-1} \ .$$

El polinomio resultante

$$p_1(x) = x^{n-1} q_1(\frac{1}{x}) = b_{n-1} x^{n-1} + \ldots + b_0$$

cumple

$$p_1(x) = \frac{p(x)}{1 - \frac{x}{\alpha}} = -\alpha \frac{p(x)}{x - \alpha} \ .$$

Remarquemos que los procesos de deflación comportan errores numéricos, a menudo importantes, en la obtención de los polinomios intermedios y, por lo tanto, en las aproximaciones de los ceros α_2, α_3, ... encontradas. Es por esto que, después de calcularlas, conviene purificarlas con algún proceso iterativo sobre el polinomio original $p(x)$. Normalmente, son suficientes pocas iteraciones.

5.2.3 Acotación de ceros de polinomios

Se conocen muchos resultados sobre acotación de ceros de polinomios. A continuación se enuncian dos criterios útiles cuando los ceros son reales.

Regla de Laguerre-Thibault

Si, al dividir $p(x)$ por $x-b$ con $b>0$, todos los coeficientes del cociente y del resto son positivos, entonces b es una cota superior de los ceros reales de $p(x)$.

Regla de Newton

Si $p(x)$ es un polinomio de grado n y, para un número real b, $p(b)$, $p'(b)$, ..., $p^{(n)}(b)$ son positivos, entonces b es una cota superior de los ceros reales de $p(x)$.

Estas reglas pueden utilizarse para hallar una cota inferior de los ceros positivos. Dado $p(x)=a_nx^n+a_{n-1}x^{n-1}+\ldots+a_0$, con $a_0\neq 0$, consideremos $q(x)=a_0x^n+a_1x^{n-1}+\ldots+a_n$. Si C es una cota superior de los ceros positivos de $q(x)$, $D=\frac{1}{C}$ será una cota inferior de los ceros positivos de $p(x)$.

Si queremos encontrar la cota superior y la cota inferior de los ceros negativos de $p(x)$, también podemos usar las reglas anteriores buscando la cota inferior y la cota superior, respectivamente, de los ceros positivos del polinomio $r(x)=p(-x)$.

Otros resultados sobre acotación de los ceros α_i, ahora tanto reales como complejos, de un polinomio $p(x)=a_nx^n+a_{n-1}x^{n-1}+\ldots+a_0$ son:

$$
\begin{aligned}
|\alpha_i| &\leq \max\left\{\left|\frac{a_0}{a_n}\right|, 1+\left|\frac{a_1}{a_n}\right|, \ldots, 1+\left|\frac{a_{n-1}}{a_n}\right|\right\}\ , \\
|\alpha_i| &\leq \max\left\{1, \sum_{k=0}^{n-1}\left|\frac{a_k}{a_n}\right|\right\}\ , \\
|\alpha_i| &\leq \max\left\{\left|\frac{a_0}{a_1}\right|, 2\left|\frac{a_1}{a_2}\right|, \ldots, 2\left|\frac{a_{n-1}}{a_n}\right|\right\}\ , \\
|\alpha_i| &\leq \sum_{k=0}^{n-1}\left|\frac{a_k}{a_{k+1}}\right|\ , \\
|\alpha_i| &\leq 2\max\left\{\left|\frac{a_{n-1}}{a_n}\right|, \sqrt{\left|\frac{a_{n-2}}{a_n}\right|}, \sqrt[3]{\left|\frac{a_{n-3}}{a_n}\right|}, \ldots, \sqrt[n]{\left|\frac{a_0}{a_n}\right|}\right\}\ .
\end{aligned}
$$

5.2.4 Separación de ceros reales de polinomios

Sucesiones de Sturm

Una sucesión $\{f_0, f_1, \ldots, f_m\}$ de funciones reales continuas, definidas sobre un intervalo $[a,b]$, es una *sucesión de Sturm para* $f=f_0$ *sobre* $[a,b]$ si verifica que:

1. f_0 es diferenciable en $[a,b]$.
2. f_m no tiene ceros reales en $[a,b]$.
3. Si $f_0(\alpha)=0$, entonces $f_1(\alpha)f_0'(\alpha)>0$.

4. Si $f_i(\alpha) = 0$, entonces $f_{i+1}(\alpha)f_{i-1}(\alpha) < 0$ $(i = 1 \div m - 1)$.

La importancia de las sucesiones de Sturm radica en el *teorema de Sturm* siguiente:

- Sea $\{f_0 = f, f_1, \ldots, f_m\}$ una sucesión de Sturm para f sobre $[a, b]$ con $f(a)f(b) \neq 0$; entonces, el número de ceros reales de f en el intervalo (a, b) es igual a $V(a) - V(b)$, donde $V(x)$ es el número de cambios de signo de $\{f_0(x), f_1(x), \ldots, f_m(x)\}$.

Así, si conocemos una sucesión de Sturm para una función f, podremos separar todos sus ceros reales de manera análoga al método de bisección (véase el problema 5.7). Seguidamente se indica cómo construir una sucesión de Sturm para un polinomio $p(x)$.

Sea $p(x)$ un polinomio real de grado n. Definimos una sucesión de polinomios $(p_i(x))_{i=0,\div m}$, con $m \leq n$, de la manera siguiente:

$$\begin{aligned} p_0(x) &= p(x) \ , \\ p_1(x) &= p'(x) \ , \\ p_{i-1}(x) &= q_i(x)p_i(x) - c_i p_{i+1}(x) \quad (i = 1 \div m - 1) \ , \\ p_{m-1}(x) &= q_m(x)p_m(x) \ , \end{aligned} \tag{5.1}$$

donde $q_i(x)$ y $c_i p_{i+1}(x)$ son el cociente y el resto, éste canviado de signo, de la división de $p_{i-1}(x)$ entre $p_i(x)$, y c_i es una constante positiva arbitraria que generalmente se escoge de manera que no aparezcan coeficientes fraccionarios en $p_{i+1}(x)$. La construcción de esta sucesión acaba cuando el resto es cero; esto es, $p_{m+1}(x) \equiv 0$. Dado que el algoritmo usado es esencialmente el algoritmo de Euclides para el cálculo del máximo común divisor de los polinomios $p_0(x)$ y $p_1(x)$, tendremos que $p_m(x) = \text{m.c.d.}(p_0(x), p_1(x))$.

Si todos los ceros de $p(x)$ son simples, $p_m(x)$ es un polinomio que no se anula sobre la recta real y $\{p_0(x), \ldots, p_m(x)\}$ es una sucesión de Sturm para $p(x)$ sobre cualquier intervalo de la recta real. Si $p(x)$ tiene ceros reales múltiples, dicha sucesión no es de Sturm aunque aún se cumple la tesis del teorema de Sturm (véase el problema 5.7). Así pues, podemos enunciar el resultado siguiente:

- Sean $p_0(x) = p(x)$, $p_1(x)$, $\ldots$, $p_m(x)$, construidos mediante la recurrencia expresada en (5.1), con $p(a)p(b) \neq 0$; entonces, el número de ceros reales diferentes del polinomio $p(x)$ en (a, b) es igual a $V(a) - V(b)$, donde $V(x)$ indica el número de cambios de signo de
$$\{p_0(x), p_1(x), \ldots, p_m(x)\} \ .$$

Como consecuencia de este resultado, calculando $V(x)$ para diferentes valores de x, podremos determinar intervalos de la recta real que contengan un único cero de $p(x)$ cada uno. Este método de separación de ceros de polinomios recibe el nombre de *método de Sturm*.

5.2.5 Métodos numéricos para el cálculo de ceros de polinomios

Al plantearnos el problema de calcular ceros de polinomios, hemos de tener en cuenta las dos ventajas que los polinomios nos ofrecen:

- La posibilidad de evaluación rápida de un polinomio y de su derivada mediante la regla de Horner.
- El uso de la deflación que, conocido un cero de un polinomio, permite buscar los otros ceros trabajando con un polinomio de un grado menor que el inicial.

Los métodos numéricos para el cálculo de ceros de polinomios pueden clasificarse en dos grupos: los métodos generales de búsqueda de ceros y los métodos específicos para polinomios. El primer grupo está formado por los métodos estudiados en el apartado 5.1.2: bisección, secante, Newton (que, en este caso, recibe el nombre de Birge-Vieta), regula falsi, iteración simple, etc., aplicados ahora a funciones polinomiales. Seguidamente se presentan algunos del segundo grupo.

Método de Laguerre

El *método de Laguerre* es un método iterativo cuyo objetivo consiste en generar una sucesión de valores (reales o complejos) x_0, x_1, ..., x_k, ... que converja hacia un cero que puede ser real o complejo. Para la construcción de esta sucesión se utiliza la fórmula siguiente:

$$x_{k+1} = x_k - \frac{np(x_k)}{p'(x_k) \pm \sqrt{H(x_k)}} \ ,$$

donde $p(x)$ es el polinomio sobre el cual trabajamos, n su grado y

$$H(x) = (n-1)[(n-1)(p'(x))^2 - np(x)p^{(2)}(x)] \ .$$

El signo del denominador se escoge de manera que lo haga mayor, en módulo.

Este método es de alto orden de convergencia: para ceros reales es cúbicamente convergente y tiene la ventaja de que, cuando todos los ceros de $p(x)$ son reales, converge sea cual sea la aproximación inicial real z_0 que se haya escogido. En este caso se dice que es *globalmente convergente*.

Método de Bernoulli

Sea

$$\begin{aligned} p(x) &= a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \\ &= a_n(x^n + b_{n-1}x^{n-1} + \ldots + b_1 x + b_0) \ , \end{aligned}$$

con

$$b_i = \frac{a_i}{a_n} \ (i = 0 \div n-1) \ .$$

Una *matriz de Frobenius* asociada a este polinomio (véase el problema 2.10) es la matriz

$$B = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ -b_0 & -b_1 & -b_2 & \ldots & -b_{n-2} & -b_{n-1} \end{pmatrix} \ .$$

Se cumple que

$$\det(B - xI) = (-1)^n(x^n + b_{n-1}x^{n-1} + \ldots + b_1x + b_0) \ .$$

Por tanto, encontrar los ceros de $p(x)$ equivale a encontrar los valores propios de la matriz B. En consecuencia, si aplicamos a la matriz B el método de la potencia expuesto en el capítulo 2, obtendremos el cero de $p(x)$ de módulo máximo, supuesto real y único: $|\alpha_1| > |\alpha_2| \geq \ldots \geq |\alpha_n|$. Así, dados x_0, x_1, ..., x_{n-1} arbitrarios, se cumple que

$$B \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \text{ con } x_n = -b_0x_0 - b_1x_1 - \ldots - b_{n-1}x_{n-1} \ ;$$

es decir,

$$x_n = -\frac{a_0x_0 + a_1x_1 + \ldots + a_{n-1}x_{n-1}}{a_n} \ .$$

Repitiendo sucesivamente este proceso, obtenemos la fórmula iterativa correspondiente al *método de Bernoulli*

$$x_{k+n} = -\frac{a_0x_k + a_1x_{k+1} + \ldots + a_{n-1}x_{k+n-1}}{a_n} \ (k \geq 0) \ .$$

Bajo condiciones análogas a las comentadas en el método de la potencia del capítulo 2, tenemos

$$\lim_{i\to\infty} \frac{x_{i+1}}{x_i} = \alpha_1 \ ,$$

donde α_1 es el cero de $p(x)$ de mayor valor absoluto. La velocidad de convergencia depende esencialmente de $|\frac{\alpha_2}{\alpha_1}|$ y puede ser muy lenta cuando $|\alpha_2| \sim |\alpha_1|$.

Aplicando el método de la potencia inversa a la matriz $B - aI$, podemos hallar el cero real más próximo al valor a, supuesto único en módulo. En el caso $a = 0$, buscaremos el menor cero en valor absoluto y el algoritmo es fácilmente expresable, ya que equivale a buscar el inverso del mayor cero en valor absoluto del polinomio

$$q(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n \ .$$

Así, cuando $a = 0$, dados y_0, ... ,y_{n-1} arbitrarios, calculamos y_{k+n}, $(k \geq 0)$, mediante la fórmula

$$y_{k+n} = -\frac{a_ny_k + a_{n-1}y_{k+1} + \ldots + a_1y_{k+n-1}}{a_0} \ .$$

Bajo condiciones análogas a las del método de la potencia, se cumple ahora que

$$\lim_{i\to\infty} \frac{y_{i+1}}{y_i} = \frac{1}{\alpha_n} \ ,$$

donde α_n es el cero de $p(x)$ de menor módulo.

	$-\frac{a_{n-1}}{a_n}$		0		0	$\dots$	0		0	
0		$\frac{a_{n-2}}{a_{n-1}}$		$\frac{a_{n-3}}{a_{n-2}}$		$\dots$		$\frac{a_0}{a_1}$		0
	$q_1^{(1)}$		$q_0^{(2)}$		$q_{-1}^{(3)}$	$\dots$	$q_{-n+3}^{(n-1)}$		$q_{-n+2}^{(n)}$	
0		$d_1^{(1)}$		$d_0^{(2)}$		$\dots$		$d_{-n+3}^{(n-1)}$		0
	$q_2^{(1)}$		$q_1^{(2)}$		$q_0^{(3)}$	$\dots$	$q_{-n+4}^{(n-1)}$		$q_{-n+3}^{(n)}$	
0		$d_2^{(1)}$		$d_1^{(2)}$		$\dots$		$d_{-n+4}^{(n-1)}$		0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$

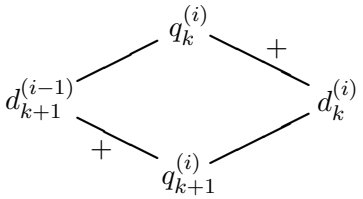
Tabla 5.1: Esquema del método del cociente-diferencia.

Método del cociente-diferencia (Q-D)

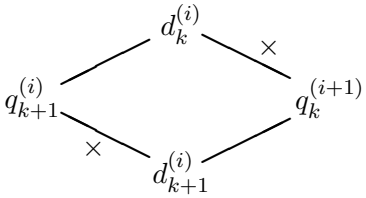
El *método del cociente-diferencia* es una generalización del método de Bernoulli que genera, bajo condiciones adecuadas, sucesiones convergentes a todos los ceros del polinomio $p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0$. El procedimiento se basa en la construcción de la tabla 5.1.

Los elementos $q_k^{(i)}, d_k^{(i)}$ $(k \geq 2-i)$ $(i = 1 \div n)$ de la tabla se van calculando, por filas sucesivas, mediante las llamadas *reglas de los rombos* que se esquematizan a continuación:

- $q_{k+1}^{(i)} = (d_k^{(i)} - d_{k+1}^{(i-1)}) + q_k^{(i)}$, gráficamente:



- $d_{k+1}^{(i)} = \frac{q_k^{(i+1)}}{q_{k+1}^{(i)}} d_k^{(i)}$, gráficamente:



Es decir, para el cálculo de un elemento de la tabla, es suficiente tener en cuenta el rombo formado por él y los tres elementos superiores. Si el rombo está centrado en la columna de las $q^{(i)}$ la suma de los elementos situados en la parte superior-derecha es igual a la de los elementos situados en la parte inferior-izquierda. Si el rombo está centrado en la columna de las $d^{(i)}$ puede afirmarse lo mismo, pero sustituyendo sumas per productos.

Se cumple el resultado siguiente: si el esquema que acaba de describirse se puede construir (esto es, no se anula ningún $q_{k+1}^{(i)}$), entonces

$$\lim_{k\to\infty} q_k^{(i)} = \alpha_i \quad \text{y} \quad \lim_{k\to\infty} d_k^{(i)} = 0$$

para todo índice i tal que $|\alpha_{i+1}|<|\alpha_i|<|\alpha_{i-1}|$.

Hay que notar que la sucesión $(q_k^{(1)})_{k\geq 1}$ es una sucesión de cocientes asociada al método de Bernoulli y que la velocidad de convergencia de estas sucesiones $(q_k^{(i)})_{k\geq 2-i}$ depende de los valores $|\frac{\alpha_i}{\alpha_{i+1}}|$, siendo lenta cuando estos cocientes son próximos a 1. Tal caso hace aconsejable la utilización de este método sólo para obtener una primera aproximación de los ceros del polinomio. Estas aproximaciones se tendrán que purificar después, aplicando un método más rápidamente convergente, como el de Newton o el de la secante, por ejemplo.

La oscilación en los valores de los elementos $d_k^{(i)}$ de una misma columna k indica la existencia de una pareja de ceros complejos conjugados. En tal caso, convendrá tomar las sucesiones $q_{k+1}^{(i)} + q_k^{(i+1)}$ y $q_k^{(i)} q_k^{(i+1)}$. Estas convergen respectivamente a números $-r$ y s tales que $x^2 + rx + s$ es un factor cuadrático del polinomio $p(x)$. Resolviendo la ecuación de segundo grado $x^2 + rx + s = 0$, obtendremos los ceros buscados. El problema 5.8 ilustra estas consideraciones.

Método de Muller-Traub

Este método es utilizable para calcular ceros de cualquier función; ahora bien, su uso más generalizado corresponde al cálculo de ceros de polinomios. Este hecho ha aconsejado que haya sido incluido en este apartado.

El *método de Muller-Traub* está basado en el esquema siguiente: dadas tres aproximaciones x_{k-2}, x_{k-1}, x_k de un cero α de una función f que, en el caso que tratamos, será un polinomio, se calcula el polinomio $p_2(x)$, de grado 2, que interpola a f en x_{k-2}, x_{k-1}, x_k y se escoge x_{k+1} como el cero de $p_2(x)$ más próximo a x_k.

Así pues, en el paso k-ésimo, dados x_{k-2}, x_{k-1}, x_k, calculamos (véase el apartado 3.1.3):

$$\begin{aligned} f[x_k] &= f(x_k)\ , \\ f[x_{k-1}, x_k] &= \frac{f[x_k] - f[x_{k-1}]}{x_k - x_{k-1}}\ , \\ f[x_{k-2}, x_{k-1}, x_k] &= \frac{f[x_{k-1}, x_k] - f[x_{k-2}, x_{k-1}]}{x_k - x_{k-2}}\ , \end{aligned}$$

(obsérvese que los valores $f[x_{k-1}]$ y $f[x_{k-2}, x_{k-1}]$ son ya conocidos del paso anterior); entonces, el polinomio $p_2(x)$ viene dado por

$$p_2(x) = f[x_k] + f[x_{k-1}, x_k](x - x_k) + f[x_{k-2}, x_{k-1}, x_k](x - x_k)(x - x_{k-1})\ ,$$

que podemos expresar como $p_2(x) = a_k(x - x_k)^2 + 2b_k(x - x_k) + c_k$ con

$$\begin{aligned} a_k &= f[x_{k-2}, x_{k-1}, x_k]\ , \\ b_k &= \frac{1}{2}(f[x_{k-1}, x_k] + f[x_{k-2}, x_{k-1}, x_k](x_k - x_{k-1}))\ , \\ c_k &= f[x_k], \end{aligned}$$

y la fórmula iterativa es

$$x_{k+1} = x_k - \frac{c_k}{b_k \pm \sqrt{b_k^2 - a_k c_k}} ,$$

con el signo escogido de manera que el denominador sea máximo en módulo.

Conviene hacer dos observaciones:

1. Si el proceso iterativo no puede continuar porque $a_k = b_k = 0$, hay que volver a empezar, partiendo de otros valores iniciales x_0, x_1, x_2.

2. Aunque se comience con unos valores x_0, x_1, x_2 reales, se pueden obtener, si $b_k^2 < a_k c_k$, iterados complejos. Trabajando con aritmética compleja, este método sirve también para calcular ceros no reales de f.

El orden de este método es al menos $p = 1.84\ldots$, siendo p el único cero positivo del polinomio $x^3 - x^2 - x - 1$. Por tanto, si el cálculo de la sucesión $(x_k)_{k\geq 0}$ es posible y se parte de valores próximos al cero buscado, este método es convergente. El problema 5.6 ofrece un buen ejemplo de su utilización.

Método de Bairstow

Dado un polinomio $p(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$ (supondremos, sin perder generalidad, que $a_n = 1$), el objetivo del *método de Bairstow* consiste en obtener factores cuadráticos de la forma $x^2 + rx + s$ del polinomio $p(x)$. Los ceros de estos factores lo serán también de $p(x)$.

En general, dados r y s, podemos escribir

$$p(x) = (x^2 + rx + s)(x^{n-2} + b_{n-3}x^{n-3} + \ldots + b_0) + lx + m ,$$

donde l y m son funciones de r y s: $l(r,s)$, $m(r,s)$. En estos términos, hay que encontrar r y s verificando el sistema de ecuaciones no lineales

$$\left.\begin{array}{l} l(r,s) = 0 \\ m(r,s) = 0 \end{array}\right\} .$$

Para obtener estos valores será suficiente resolver el sistema anterior. Un método de resolución es el de Newton en dos variables que será estudiado en el apartado 5.3.3 y que, en este caso, adopta la forma especialmente manejable que se describe seguidamente.

A partir de una pareja adecuada de valores iniciales r_0, s_0, se generan sendas sucesiones $(r_k)_{k\geq 0}$, $(s_k)_{k\geq 0}$ convergentes hacia la pareja de valores de r, s buscados.

La construcción de estas sucesiones se realiza de la manera siguiente:

$$\left.\begin{array}{l} r_{k+1} = r_k + \Delta r_k \\ s_{k+1} = s_k + \Delta s_k \end{array}\right\} ,$$

siendo $\Delta r_k, \Delta s_k$ las soluciones del sistema

$$\left.\begin{array}{rcrcl} c_0 \Delta r_k & + & c_1 \Delta s_k & = & b_{-1} \\ (c_{-1} - b_{-1})\Delta r_k & + & c_0 \Delta s_k & = & b_{-2} \end{array}\right\} ,$$

donde los b_j, c_j se generan por las fórmulas recurrentes:

$$\begin{array}{llll} b_{n-1}=0, & b_{n-2}=1, & b_j=a_{j+2}-r_k b_{j+1}-s_k b_{j+2} & (j=n-3\div -2)\ , \\ c_{n-1}=0, & c_{n-2}=1, & c_j=b_j-r_k c_{j+1}-s_k c_{j+2} & (j=n-3\div -1)\ . \end{array}$$

Los valores r_0, s_0 han de escogerse tan próximos como sea posible a los valores buscados, de esta elección dependerá la convergencia del método. Si, por algún procedimiento aproximado, se han obtenido valores reales $\hat{\alpha_1}$, $\hat{\alpha_2}$, próximos a los ceros α_1, α_2, convendrá partir de $r_0=-\hat{\alpha_1}-\hat{\alpha_2}$, $s_0=\hat{\alpha_1}\hat{\alpha_2}$. Si los valores aproximados de que disponemos son dos complejos conjugados $\hat{\alpha}+\hat{\beta}i$, $\hat{\alpha}-\hat{\beta}i$ convendrá, evidentemente, tomar $r_0=-2\hat{\alpha}$ y $s_0=\hat{\alpha}^2+\hat{\beta}^2$.

El proceso iterativo se acabará cuando $\Delta r_k, \Delta s_k$ sean suficientemente pequeÆos. Entonces las raíces de la ecuación $x^2+r_k x+s_k=0$ serán aproximaciones de dos ceros del polinomio $p(x)$. En el problema 5.8 se ofrece un ejemplo de utilización de este método.

Método de Graeffe

En la exposición de este método limitaremos nuestra atención al tratamiento de polinomios con ceros reales simples. No obstante, hay que tener en cuenta que es igualmente útil en el caso más general.

Consideremos un polinomio de grado n, $p(x)=a_n x^n+a_{n-1}x^{n-1}+\ldots+a_0$, con ceros $\alpha_1, \alpha_2, \ldots, \alpha_n$, distintos en módulo y que supondremos ordenados según $|\alpha_1|>|\alpha_2|>\ldots>|\alpha_n|$. Si estos ceros están *muy separados*, en el sentido de que $|\frac{\alpha_k}{\alpha_{k-1}}|$ $(k=2\div n)$ sean suficientemente pequeÆos, entonces

$$\alpha_1 \simeq -\frac{a_{n-1}}{a_n}\ ,\quad \alpha_2 \simeq -\frac{a_{n-2}}{a_{n-1}}\ ,\quad \ldots\ ,\quad \alpha_n \simeq -\frac{a_0}{a_1}\ .$$

Dado un polinomio cualquiera con todos los ceros diferentes en módulo, éstos no estarán muy separados, en general; pero, elevándolos a una potencia conveniente, se puede conseguir separarlos tanto como se desee.

Esta es la idea central del *método de Graeffe*: a partir de un polinomio $p(x)$ de ceros α_1, α_2, $\ldots$, α_n, obtener un nuevo polinomio $p_1(x)$ de ceros $\alpha_1^2, \alpha_2^2, \ldots, \alpha_n^2$. Aplicando reiteradamente este método se irán obteniendo polinomios $p_k(x)$ de ceros $\alpha_1^{2^k}, \alpha_2^{2^k}, \ldots, \alpha_n^{2^k}$, hasta que, considerándolos suficientemente separados, podamos calcular

$$\alpha_i^{2^k} \simeq -\frac{a_{n-i}^{(k)}}{a_{n-i+1}^{(k)}}\quad (i=1\div n)\ ,$$

donde los $a_j^{(k)}$ $(j=0\div n)$ son los coeficientes de $p_k(x)$. A partir de aquí no hay ninguna dificultad en el cálculo de α_i, teniendo en cuenta que la indeterminación del signo se resuelve sustituyendo $+\alpha_i$ y $-\alpha_i$ en el polinomio original $p(x)$. Así, el nucleo del método de Graeffe radica en la construcción recurrente del polinomio $p_{k+1}(x)$, de coeficientes $a_j^{(k+1)}$, a partir del polinomio $p_k(x)$, de coeficientes $a_j^{(k)}$.

Esta construcción se basa en la relación $p_{k+1}(x^2) = (-1)^n p_k(x) p_k(-x)$, de la cual se desprenden las expresiones que nos dan los coeficientes de $p_{k+1}(x)$ a partir de los de $p_k(x)$:

$$\begin{array}{rcl}
a_0^{(k+1)} & = & (-1)^n a_0^{(k)^2}, \\
a_1^{(k+1)} & = & (-1)^{n-1} \left(a_1^{(k)^2} - 2a_0^{(k)} a_2^{(k)} \right) , \\
a_2^{(k+1)} & = & (-1)^{(n-2)} \left(a_2^{(k)^2} - 2a_1^{(k)} a_3^{(k)} + 2a_0^{(k)} a_4^{(k)} \right) , \\
\vdots & \vdots & \vdots \\
a_j^{(k+1)} & = & (-1)^{n-j} \left(a_j^{(k)^2} - 2a_{j-1}^{(k)} a_{j+1}^{(k)} + \ldots + (-1)^l 2a_{j-l}^{(k)} a_{j+l}^{(k)} \right) , \\
\vdots & \vdots & \vdots \\
a_n^{(k+1)} & = & a_n^{(k)^2},
\end{array}$$

donde $l = \min(j, n-j)$ en la expresión del coeficiente $a_j^{(k+1)}$.

5.3 SISTEMAS NO LINEALES

5.3.1 Introducción

En esta sección se exponen, para la resolución de sistemas de ecuaciones no lineales, dos métodos numéricos que son generalizaciones de métodos ya estudiados en el caso de una ecuación no lineal: el método de iteración simple y el de Newton. Por comodidad, trabajaremos únicamente con sistemas no lineales de dos ecuaciones con dos incógnitas que escribiremos como

$$\left.\begin{array}{l} x_1 = g_1(x_1, x_2) \\ x_2 = g_2(x_1, x_2) \end{array}\right\} \quad \text{o} \quad \left.\begin{array}{l} f_1(x_1, x_2) = 0 \\ f_2(x_1, x_2) = 0 \end{array}\right\} ,$$

según convenga.

5.3.2 Método de iteración simple en varias variables

Consideremos el sistema

$$\left.\begin{array}{l} x_1 = g_1(x_1, x_2) \\ x_2 = g_2(x_1, x_2) \end{array}\right\}$$

que, usando la notación vectorial

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} , \quad g(x) = \begin{pmatrix} g_1(x) \\ g_2(x) \end{pmatrix} = \begin{pmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{pmatrix} ,$$

se expresa en la forma $x = g(x)$.

El objetivo del *método de iteración simple en varias variables* radica en la construcción de una sucesión de vectores $x^{(0)}$, $x^{(1)}$, $x^{(2)}$, ... que converja hacia la solución buscada. Partiendo de un $x^{(0)}$ adecuado, la construcción de dicha sucesión se hace según la fórmula recurrente $x^{(k+1)} = g(x^{(k)})$.

Sobre la convergencia de este método se dispone del siguiente resultado:

- Sea R un conjunto cerrado cualquiera de $\mathbb{R}^2$; si $g(x) \in R$, para todo $x \in R$, y existe una constante real $C < 1$ tal que, para alguna norma vectorial $\| \quad \|$, se cumple que

$$\|g(x^{(1)}) - g(x^{(2)})\| \leq C\|x^{(1)} - x^{(2)}\| \ ,$$

para $x^{(1)} \in R$ y $x^{(2)} \in R$; entonces, el sistema $x = g(x)$ tiene una solución única en R y, para cualquier $x^{(0)} \in R$, la sucesión $x^{(0)}$, $x^{(1)}$, $x^{(2)}$, ..., obtenida por la recurrencia $x^{(k+1)} = g(x^{(k)})$ $(k \geq 0)$, converge hacia la solución.

Cuando una función g cumple las hipótesis del teorema anterior, se dice que es *contractiva (de constante C)* dentro de R. De manera análoga al caso de una variable, una condición suficiente para que g sea contractiva en R es que sea continuamente diferenciable y que la matriz jacobiana $Dg(x)$ de g cumpla que

$$\|Dg(x)\| \leq C < 1$$

para todo $x \in R$, siendo $\| \quad \|$ una norma matricial consistente con la norma vectorial considerada.

5.3.3 Método de Newton en varias variables

Consideremos un sistema escrito en la forma

$$\left.\begin{array}{l} f_1(x_1, x_2) = 0 \\ f_2(x_1, x_2) = 0 \end{array}\right\} \ ,$$

donde suponemos que las funciones f_1 y f_2 son diferenciables con continuidad.

Usando la notación vectorial

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \ , \quad f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} \ ,$$

el sistema se expresa en la forma $f(x) = 0$.

Supondremos que la matriz jacobiana es regular en un entorno de la solución,

$$\det\left(Df(x)\right) = \det \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) \end{pmatrix} \neq 0 \ .$$

El objetivo del *método de Newton en varias variables* consiste en generar una sucesión de vectores $x^{(0)}$, $x^{(1)}$, $x^{(2)}$, ..., que converja hacia la solución buscada. Partiendo de un vector $x^{(0)}$ adecuado, la construcción de dicha sucesión se lleva a cabo según la fórmula recurrente

$$x^{(k+1)} = x^{(k)} - (Df(x^{(k)}))^{-1} f(x^{(k)}) \ ,$$

que es una generalización a dimensión dos del método de Newton para una ecuación en una variable.

El método de Newton acostumbra a formularse de la siguiente manera:

$$x^{(k+1)} = x^{(k)} + \Delta x^{(k)} \ ,$$

con $\Delta x^{(k)}$ cumpliendo el sistema lineal

$$Df(x^{(k)})\Delta x^{(k)} = -f(x^{(k)}) \ .$$

Así, en lugar de calcular la inversa de la matriz jacobiana en cada iteración, tan solo hay que resolver un sistema lineal.

Cuando el cálculo de las derivadas parciales de f_1 y de f_2 comporte dificultades (por ejemplo, cuando se desconozcan las expresiones analíticas de f_1 y f_2 y se disponga únicamente de un algoritmo para su cálculo), será necesario aproximarlas por sus cocientes incrementales:

$$\begin{aligned} \frac{\partial f_1}{\partial x_1}(x_1, x_2) &\simeq \frac{f_1(x_1 + h_1, x_2) - f_1(x_1, x_2)}{h_1} \ , \\ \frac{\partial f_1}{\partial x_2}(x_1, x_2) &\simeq \frac{f_1(x_1, x_2 + h_2) - f_1(x_1, x_2)}{h_2} \ , \end{aligned}$$

y, análogamente, para la función f_2 (se pueden usar también otras expresiones basadas en las fórmulas de derivación introducidas en el capítulo precedente). Los valores h_1 y h_2 de las expresiones anteriores deben escogerse de manera conveniente. Hay una variante del método de Newton, que se llama *método de Steffensen*, que consiste en tomar $h_1 = f_1(x^{(k)})$ y $h_2 = f_2(x^{(k)})$ para el cálculo aproximado de $Df(x^{(k)})$ que ha de llevarse a cabo en el paso k-ésimo del proceso iterativo.

COMENTARIOS BIBLIOGRÁFICOS

Una de las referencias más completas sobre resolución de ecuaciones no lineales es [Tra64], que incluye además una amplia bibliografía. Otras referencias generales son [Ost66], [RR78]. El método de Newton es estudiado en prácticamente todos los libros de cálculo numérico y en otros campos: para una base teórica, pueden consultarse [Die68], [Ost66] y, para cuestiones prácticas, [Hen64], [SB80]. Los métodos de la secante y regula falsi se tratan de manera detallada en [Hil74], [Ost66], [RR78]. Para diversos resultados de convergencia global de los métodos de Newton y de iteración simple, destacamos [Hen64], [IK66], [Ost66]. Para la generación de métodos via interpolación, así como para el estudio de sus órdenes de convergencia, pueden consultarse [Dur61], [Hen64], [Hil74], [Hou53], [IK66], [RR78]; [Ost66] para su aceleración. [CdB72] presenta algunos programas en lenguaje `FORTRAN`.

Diversas acotaciones para los ceros de los polinomios se encuentran en [SB80] y, sobre todo, en [Hen74], el cual también estudia muy extensamente la localización y aproximación de ceros complejos. El método de Laguerre se detalla en [RR78], los de Bernoulli, Q-D y Muller-Traub en [Hen64], y los de Bairstow y Graeffe, en [Hil74]. [IK66], [SB80] y, sobre todo, [Wil64] tratan la sensibilidad de los ceros de los polinomios respecto a sus coeficientes.

Para sistemas de ecuaciones no lineales, pueden consultarse, por ejemplo, [IK66], [OR70], [Ost66], [SB80].

PROBLEMAS RESUELTOS

Problema 5.1 *Sea f una función 3 veces diferenciable con continuidad de la cual buscamos un cero simple α.*

a) Deducir la fórmula correspondiente al método iterativo que obtiene x_{k+1} evaluando en 0 el polinomio de Taylor de grado menor o igual que 2 de la función inversa g de f en $f(x_k)$ (método de Chebichev).

b) Probar que el método descrito tiene orden al menos 3 para ceros simples y determinar la constante asintótica del error cuando el orden sea exactamente 3.

SOLUCIÓN:

a) Sea g la función inversa de f en un entorno de α (existe ya que α es un cero simple de f).

Siguiendo el enunciado, consideramos

$$x_{k+1} = \sum_{j=0}^{2} \frac{g^{(j)}(f(x_k))}{j!}(-f(x_k))^j \ .$$

Obsérvese que $g^{(0)}(f(x_k)) = g(f(x_k)) = x_k$, ya que $g(f(x)) = x$.

Derivando dos veces esta última expresión, se obtiene

$$g'(f(x))f'(x) = 1 \ , \quad g^{(2)}(f(x))f'^2(x) + g'(f(x))f^{(2)}(x) = 0 \ ;$$

y, como consecuencia, las expresiones siguientes para las derivadas primera y segunda de la función inversa g en $f(x_k)$ en términos de las derivadas correspondientes de f en x_k:

$$g'(f(x_k)) = \frac{1}{f'(x_k)} \ , \quad g^{(2)}(f(x_k)) = -\frac{f^{(2)}(x_k)}{f'^3(x_k)} \ .$$

Por lo tanto, la fórmula iterativa del método de Chebichev se escribe como

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f^{(2)}(x_k)f^2(x_k)}{2f'^3(x_k)} \ .$$

b) Seguidamente demostraremos que el orden de convergencia de este método, al calcular ceros simples, es al menos 3 y encontraremos la constante asintótica del error en el caso que sea 3. Consideremos la expresión del cero buscado

$$\begin{aligned} \alpha &= g(f(x_k)) - g'(f(x_k))f(x_k) + \frac{g^{(2)}(f(x_k))}{2}f^2(x_k) \\ &\quad -\frac{g^{(3)}(\xi)}{6}f^3(x_k) \ , \quad \xi \in <0, f(x_k)> \ . \end{aligned}$$

Teniendo presente que los tres primeros términos del segundo miembro constituyen la fórmula de iteración, podemos escribir

$$x_{k+1} - \alpha = \frac{g^{(3)}(\xi)}{6} f^3(x_k) \ .$$

Encontraremos ahora las expresiones de $f(x_k)$ y de $g^{(3)}(\xi)$ en función de las derivadas de f.

Por un lado,

$$f(x_k) = f(\alpha) + f'(\eta)(x_k - \alpha) = f'(\eta)(x_k - \alpha)$$

con $\eta \in < \alpha, x_k >$.

Por otro, derivando tres veces la expresión $g(f(x)) = x$, obtenemos que

$$g^{(3)}(f(x))f'^3(x) + 3g^{(2)}(f(x))f'(x)f^{(2)}(x) + g'(f(x))f^{(3)}(x) = 0 \ .$$

Reemplazando las expresiones de $g'(f(x))$ y de $g^{(2)}(f(x))$ obtenidas anteriormente y despejando $g^{(3)}(f(x))$, tenemos

$$g^{(3)}(f(x)) = \frac{3f^{(2)2}(x) - f'(x)f^{(3)}(x)}{f'^5(x)} \ .$$

Sustituyendo estas expresiones en la de $x_{k+1} - \alpha$ y teniendo en cuenta que $\xi = f(g(\xi))$, resulta

$$x_{k+1} - \alpha = \frac{3f^{(2)2}(g(\xi)) - f'(g(\xi))f^{(3)}(g(\xi))}{6f'^5(g(\xi))} f'^3(\eta)(x_k - \alpha)^3$$

y, por tanto,

$$\lim_{x_k \to \alpha} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^3} = \lim_{x_k \to \alpha} \frac{3f^{(2)2}(g(\xi)) - f'(g(\xi))f^{(3)}(g(\xi))}{6f'^5(g(\xi))} f'^3(\eta) \ .$$

Ahora bien, cuando $x_k \to \alpha$, tenemos $\eta \to \alpha$, $f(x_k) \to 0, \xi \to 0$ y $g(\xi) \to \alpha$; entonces,

$$\lim_{x_k \to \alpha} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^3} = \frac{3f^{(2)2}(\alpha) - f'(\alpha)f^{(3)}(\alpha)}{6f'^2(\alpha)} \ .$$

Con esta última expresión queda demostrado que, para el cálculo de ceros simples, el orden de convergencia del método de Chebichev es al menos 3 y que, en el caso de que sea exactamente 3, la constante asintótica del error viene dada por la expresión

$$\frac{3f^{(2)2}(\alpha) - f'(\alpha)f^{(3)}(\alpha)}{6f'^2(\alpha)} \ .$$

Problema 5.2 *a) Determinar el orden de convergencia del método de Steffensen*

$$x_{k+1} = x_k - \frac{f^2(x_k)}{f(x_k + f(x_k)) - f(x_k)} \ ,$$

en el cálculo de ceros simples de funciones dos veces diferenciables con continuidad.

b) Aplicación: Comprobar numéricamente que, al usar este método para el cálculo del cero de $f(x) = \cos x - x$, el orden de convergencia es el encontrado en el apartado anterior.

SOLUCIÓN:

a) Escribimos el método en la forma

$$x_{k+1} = x_k - \frac{f(x_k)}{g(x_k)} \ ,$$

con

$$g(x_k) = \frac{f(x_k + f(x_k)) - f(x_k)}{f(x_k)} \ .$$

Nótese que $g(x_k)$ es una aproximación por derivación numérica de $f'(x_k)$ con paso $f(x_k)$. Usando el desarrollo de Taylor de la función f en un entorno de x_k

$$f(x_k + f(x_k)) = f(x_k) + f'(x_k)f(x_k) + \frac{f^{(2)}(\xi_1)}{2} f^2(x_k) \ ,$$

con $\xi_1 \in < x_k, x_k + f(x_k) >$, se obtiene

$$g(x_k) = f'(x_k) + \frac{1}{2} f^{(2)}(\xi_1) f(x_k) \ .$$

Considerando ahora los desarrollos de f y f' en un entorno del cero simple buscado α:

$$\begin{aligned} f(x_k) &= f'(\xi_2)(x_k - \alpha) \ , \\ f'(x_k) &= f'(\alpha) + f^{(2)}(\xi_3)(x_k - \alpha) \ , \end{aligned}$$

con $\xi_2, \xi_3 \in < \alpha, x_k >$ y sustituyendo en la expresión de $g(x_k)$, obtenemos

$$g(x_k) = f'(\alpha) + f^{(2)}(\xi_3)(x_k - \alpha) + \frac{1}{2} f^{(2)}(\xi_1) f'(\xi_2)(x_k - \alpha) \ .$$

Entonces, teniendo presente que

$$f(x_k) = f'(\alpha)(x_k - \alpha) + \frac{f^{(2)}(\xi_4)}{2} (x_k - \alpha)^2 \ ,$$

con $\xi_4 \in < \alpha, x_k >$, podremos escribir

$$\begin{aligned} x_{k+1} - \alpha &= \frac{(x_k - \alpha) g(x_k) - f(x_k)}{g(x_k)} \\ &= \frac{[f^{(2)}(\xi_3) + \frac{1}{2} f^{(2)}(\xi_1) f'(\xi_2)](x_k - \alpha)^2 - \frac{f^{(2)}(\xi_4)}{2}(x_k - \alpha)^2}{f'(\alpha) + [f^{(2)}(\xi_3) + \frac{1}{2} f^{(2)}(\xi_1) f'(\xi_2)](x_k - \alpha)} \ . \end{aligned}$$

Cuando $k \to \infty, x_k \to \alpha$ y $\xi_i \to \alpha$ para $i = 1, 2, 3, 4$. Por lo tanto,

$$\lim_{k\to\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \frac{f^{(2)}(\alpha)}{2f'(\alpha)}(1 + f'(\alpha)) \ .$$

Así, la convergencia es cuadrática si $f^{(2)}(\alpha) \neq 0$ y $f'(\alpha) \neq -1$. En caso contrario, es al menos cúbica.

b) Consideremos la función concreta $f(x) = \cos x - x$. Podemos escribir la recurrencia dada por el método de Steffensen de la forma $x_{k+1} = G(x_k)$ donde

$$G(x) = x - \frac{(\cos x - x)^2}{\cos(\cos x) - 2\cos x + x}.$$

Tomando $x_0 = 1$ y aplicando la recurrencia descrita, se obtiene

$$\begin{aligned} x_1 &= 0.7280103655 \ , \\ x_2 &= 0.7390669669 \ , \\ x_3 &= 0.739085133167 \ , \\ x_4 &= 0.739085133216 \ , \\ x_5 &= 0.739085133216 \simeq \alpha \ ; \end{aligned}$$

de manera que

$$\begin{aligned} e_0 &= 0.260915 \ , \\ e_1 &= -0.01107477 \ , \\ e_2 &= -0.0000181663 \ , \\ e_3 &= -0.000000000049 \ ; \end{aligned}$$

y, por lo tanto,

$$\frac{e_1}{e_0^2} = -0.163 \ , \quad \frac{e_2}{e_1^2} = -0.148 \ , \quad \frac{e_3}{e_2^2} = -0.148 \ .$$

Nótese que la constante asintótica del error es

$$\frac{f^{(2)}(\alpha)}{2f'(\alpha)}(1 + f'(\alpha)) = -\frac{\alpha \ \mathrm{sen}\, \alpha}{2(1 + \mathrm{sen}\, \alpha)} = -0.1487371708 \ ,$$

valor al cual ha de tender la sucesión

$$\left(\frac{e_{k+1}}{e_k^2}\right)_{k\geq 0} \ .$$

Esta tendencia muestra la convergencia cuadrática del método.

Problema 5.3 *Demostrar que el* método de Newton modificado para ceros múltiples

$$x_{k+1} = x_k - \frac{m\, f(x_k)}{f'(x_k)} \ ,$$

tiene orden de convergencia 2 *al calcular un cero* α *de multiplicidad* m *de* f *siempre que* f *sea* $m + 1$ *veces diferenciable con continuidad y que* $f^{(m+1)}(\alpha) \neq 0$.

SOLUCIÓN:

Si f es $m+1$ veces diferenciable con continuidad, teniendo en cuenta que $f(\alpha) = f'(\alpha) = \ldots = f^{(m-1)}(\alpha) = 0$ y $f^{(m)}(\alpha) \neq 0$, resultan las expresiones:

$$\begin{aligned} f(x) &= \frac{f^{(m)}(\alpha)}{m!}(x-\alpha)^m + \frac{f^{(m+1)}(\xi_1)}{(m+1)!}(x-\alpha)^{m+1} , \\ f'(x) &= \frac{f^{(m)}(\alpha)}{(m-1)!}(x-\alpha)^{m-1} + \frac{f^{(m+1)}(\xi_2)}{m!}(x-\alpha)^m , \end{aligned}$$

con $\xi_1, \xi_2 \in < x, \alpha >$.

Por lo tanto,

$$x_{k+1} - \alpha = x_k - \alpha - \frac{m\left[\frac{f^{(m)}(\alpha)}{m!}(x_k-\alpha)^m + \frac{f^{(m+1)}(\xi_1)}{(m+1)!}(x_k-\alpha)^{m+1}\right]}{\frac{f^{(m)}(\alpha)}{(m-1)!}(x_k-\alpha)^{m-1} + \frac{f^{(m+1)}(\xi_2)}{m!}(x_k-\alpha)^m} ,$$

con $\xi_1, \xi_2 \in < x_k, \alpha >$.

Operando en la expresión anterior, se obtiene

$$\begin{aligned} \frac{x_{k+1}-\alpha}{(x_k-\alpha)^2} &= \frac{\frac{f^{(m+1)}(\xi_2)}{m!} - \frac{mf^{(m+1)}(\xi_1)}{(m+1)!}}{\frac{f^{(m)}(\alpha)}{(m-1)!} + \frac{f^{(m+1)}(\xi_2)}{m!}(x_k-\alpha)} \\ &= \frac{f^{(m+1)}(\xi_2) - \frac{m}{m+1}f^{(m+1)}(\xi_1)}{mf^{(m)}(\alpha) + f^{(m+1)}(\xi_2)(x_k-\alpha)} . \end{aligned}$$

Cuando $k \to \infty$, $x_k \to \alpha$ y $\xi_i \to \alpha$ para $i = 1, 2$. Por lo tanto,

$$\lim_{k\to\infty} \frac{x_{k+1}-\alpha}{(x_k-\alpha)^2} = \frac{f^{(m+1)}(\alpha) - \frac{m}{m+1}f^{(m+1)}(\alpha)}{mf^{(m)}(\alpha)} = \frac{f^{(m+1)}(\alpha)}{m(m+1)f^{(m)}(\alpha)} \neq 0 .$$

Así, el orden de convergencia es 2 y la constante asintótica del error es

$$\frac{f^{(m+1)}(\alpha)}{m(m+1)f^{(m)}(\alpha)} .$$

Problema 5.4 *a) Deducir la fórmula recurrente del método de interpolación inversa iterada de 3 puntos para el cálculo de ceros de una función f.*

b) Encontrar el orden de convergencia y la constante asintótica del error del método anterior al calcular ceros simples para el caso más general.

c) Aplicación: Calcular, con un error menor que 10^{-6}, el cero de la función

$$f(x) = \frac{1}{x} - 2\ln x - 0.5 ,$$

a partir de las aproximaciones iniciales siguientes: $x_0 = 0.5$, $x_1 = 1$, $x_2 = 1.5$.

SOLUCIÓN:

a) Recordemos el algoritmo del método de interpolación inversa iterada de 3 puntos para el cálculo de ceros de una función f:

- Se parte de tres aproximaciones x_0, x_1, x_2.
- Se interpola la función g, inversa de f, mediante un polinomio $q_2(y)$ que pasa por los puntos $(f(x_0), x_0)$, $(f(x_1), x_1)$, $(f(x_2), x_2)$.
- Se toma $x_3 = q_2(0)$ y se repite el proceso con los valores x_1, x_2, x_3 como aproximaciones del cero α buscado.

Seguidamente detallamos el cálculo de x_{k+1} a partir de x_{k-2}, x_{k-1}, x_k. Para simplificar la notación, escribiremos $f_i = f(x_i)$.

Aplicando el método de las diferencias divididas de Newton en los puntos (f_{k-2}, x_{k-2}), (f_{k-1}, x_{k-1}), (f_k, x_k), se obtiene el polinomio interpolador de la función inversa de f

$$q_2(y) = x_k + g[f_k, f_{k-1}](y - f_k) + g[f_k, f_{k-1}, f_{k-2}](y - f_k)(y - f_{k-1}) \ .$$

Tomando $x_{k+1} = q_2(0)$ y operando, se obtiene

$$x_{k+1} = x_k - g[f_k, f_{k-1}]f_k + g[f_k, f_{k-1}, f_{k-2}]f_k f_{k-1} \ ,$$

donde, tal como se ha explicado en el capítulo 3, el cálculo de

$$\begin{aligned} g[f_k, f_{k-1}] &= \frac{x_k - x_{k-1}}{f_k - f_{k-1}} \ , \\ g[f_k, f_{k-1}, f_{k-2}] &= \frac{g[f_k, f_{k-1}] - g[f_{k-1}, f_{k-2}]}{f_k - f_{k-2}} \end{aligned}$$

se realiza a través de un esquema triangular del tipo siguiente:

$$\begin{array}{l|lll} f_k & x_k & & \\ & & g[f_k, f_{k-1}] & \\ f_{k-1} & x_{k-1} & & g[f_k, f_{k-1}, f_{k-2}] \\ & & g[f_{k-1}, f_{k-2}] & \\ f_{k-2} & x_{k-2} & & \end{array} \ .$$

Nótese que, una vez calculado x_{k+1}, cuando se inicie una nueva iteración con x_{k+1}, x_k, x_{k-1}, se podrá aprovechar la parte del esquema triangular que tiene como vértice $g[f_k, f_{k-1}]$, y sólo tendrá que añadirse una nueva fila superior.

b) Para encontrar el orden de convergencia y la constante asintótica del error del método descrito en el apartado anterior, partimos de la expresión del error que se produce al interpolar la función $g(y)$ por el polinomio $q_2(y)$ (consúltese el apartado 3.1.2):

$$g(y) - q_2(y) = \frac{g^{(3)}(\eta_k(y))}{3!}(y - f_k)(y - f_{k-1})(y - f_{k-2}) \ ,$$

donde $\eta_k(y) \in < f_k, f_{k-1}, f_{k-2}, y >$.

Si evaluamos esta expresión para $y = 0$ observando que se satisfacen $g(0) = \alpha$ y $q_2(0) = x_{k+1}$, obtenemos

$$\alpha - x_{k+1} = -\frac{g^{(3)}(\eta_k)}{3!} f_k f_{k-1} f_{k-2} \ ,$$

con $\eta_k = \eta_k(0)$.

Tomando $e_i = x_i - \alpha$ y considerando el polinomio interpolador de Taylor de grado menor o igual a 1 de la función f en α evaluado en los puntos x_k, x_{k-1}, x_{k-2} con los errores correspondientes, se obtiene

$$f_i = f'(\alpha) e_i + \frac{f^{(2)}(\xi_i)}{2} e_i^2 \ ,$$

con $\xi_i \in < x_i, \alpha >$ $(i = k, k-1, k-2)$.

Sustituyendo estas expresiones en la fórmula del error, podemos escribir

$$e_{k+1} = \frac{g^{(3)}(\eta_k)}{3!} \prod_{i=k-2}^{k} \left(f'(\alpha) + \frac{f^{(2)}(\xi_i)}{2} e_i \right) e_i \ .$$

Cuando $k \to \infty$ y supuesta la convergencia $x_k \to \alpha$, tenemos $\xi_i \to \alpha$, $\eta_k \to 0$. Por lo tanto,

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k e_{k-1} e_{k-2}} = \frac{g^{(3)}(0)}{6} f'^3(\alpha) \ .$$

En el problema 5.1 se deduce la expresión de $g^{(3)}(f(x))$ en función de las derivadas hasta orden 3 de f en x

$$g^{(3)}(f(x)) = \frac{3 f^{(2)2}(x) - f'(x) f^{(3)}(x)}{f'^5(x)} \ ,$$

que, para $x = \alpha$, queda

$$g^{(3)}(0) = \frac{3 f^{(2)2}(\alpha) - f'(\alpha) f^{(3)}(\alpha)}{f'^5(\alpha)} \ .$$

Así, podremos escribir

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k e_{k-1} e_{k-2}} = \frac{3 f^{(2)2}(\alpha) - f'(\alpha) f^{(3)}(\alpha)}{6 f'^2(\alpha)} \equiv K \ .$$

Si indicamos el orden de convergencia por p y la constante asintótica del error por L, tendremos

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k^p} = L \ .$$

De manera que, para $k \to \infty$:

$$e_{k+1} \sim L e_k^p \ , \quad e_k \sim L e_{k-1}^p \ , \quad e_{k-1} \sim L e_{k-2}^p \ ;$$

asimismo, para $k \to \infty$,

$$e_{k+1} \sim K e_k e_{k-1} e_{k-2} \quad (k \to \infty) \ .$$

Expresando e_{k+1}, e_{k-1} y e_{k-2} en función de e_k y sustituyendo en la última expresión, obtendremos

$$Le_k^p \sim K \frac{e_k^{1+\frac{1}{p}+\frac{1}{p^2}}}{L^{\frac{2}{p}+\frac{1}{p^2}}} \quad (k \to \infty) \ .$$

Así, será necesario que

$$p = 1 + \frac{1}{p} + \frac{1}{p^2} \ , \quad K = L^{1+\frac{2}{p}+\frac{1}{p^2}}$$

o equivalentemente,

$$p^3 - p^2 - p - 1 = 0 \ , \quad L = K^{\frac{p}{p^2+1}} \ .$$

Así pues, el orden de convergencia del método propuesto será el único cero real del polinomio

$$x^3 - x^2 - x - 1 \ ,$$

que vale, aproximadamente, 1.84; la constante asintótica del error será entonces

$$L = K^{\frac{p}{p^2+1}} \ .$$

Finalmente, obtenemos

$$p \simeq 1.84 \ , \quad L = \left(\frac{3f^{(2)2}(\alpha) - f'(\alpha) f^{(3)}(\alpha)}{6f'^2(\alpha)} \right)^{\frac{p}{p^2+1}} \ .$$

c) Apliquemos ahora el método de interpolación inversa iterada al cálculo del cero de la función $f(x) = \frac{1}{x} - 2\ln x - 0.5$, partiendo de los valores $x_0 = 0.5$, $x_1 = 1$ y $x_2 = 1.5$. En cada iteración realizaremos el cálculo de $g[f_k, f_{k-1}]$ y de $g[f_k, f_{k-1}, f_{k-2}]$ mediante un esquema triangular, teniendo en cuenta que, para su construcción, podremos aprovechar algunos valores del esquema triangular correspondiente a la iteración anterior. Detallamos, a continuación, las sucesivas iteraciones.

En la primera iteración, el esquema triangular de diferencias divididas

f_i	x_i	$g[f_i, f_{i+1}]$	$g[f_i, f_{i+1}, f_{i+2}]$
-0.6442635	1.5		
		-0.4369623	
0.5	1		0.0644183
		-0.2095299	
2.886294	0.5		

nos permite calcular $x_3 \simeq 1.19773$ con $f(x_3) = -0.0259436$.

En la segunda iteración, el esquema triangular de diferencias divididas

f_i	x_i	$g[f_i, f_{i+1}]$	$g[f_i, f_{i+1}, f_{i+2}]$
-0.0259436	1.19773		
		-0.4888569	
-0.6442635	1.5		0.0986696
		-0.4369623	
0.5	1		

nos permite calcular $x_4 \simeq 1.186697$ con $f(x_4) \simeq 0.0003284$.

En la tercera iteración, el esquema triangular de diferencias divididas

f_i	x_i	$g[f_i, f_{i+1}]$	$g[f_i, f_{i+1}, f_{i+2}]$
0.0003284	1.186697		
		-0.4199677	
-0.0259436	1.19773		0.1068726
		-0.4888569	
-0.6442635	1.5		

nos permite calcular $x_5 \simeq 1.186834$ con $f(x_5) \simeq 3.58 \cdot 10^{-7}$.

Veamos ahora que el error de aproximación de x_5 es ya más pequeño que 10^{-6}.

Como

$$f(x_5) = f'(\xi)(x_5 - \alpha) \ , \quad \xi \in < x_5, \alpha > \ ,$$

podemos expresar

$$\epsilon = |e_5| = \left| \frac{f(x_5)}{f'(\xi)} \right| \ .$$

Un breve estudio de la función f nos permite asegurar que $\alpha \in (1,2)$ y que, por lo tanto, $\xi \in (1,2)$ y $1.25 < |f'(\xi)| < 3$.

Resulta, entonces,

$$\epsilon_5 < \frac{3.58 \cdot 10^{-7}}{1.25} < 10^{-6} \ .$$

Así, el cero buscado, con error menor que 10^{-6}, es

$$\alpha \simeq 1.186834 \ .$$

Problema 5.5 *Comparar, según su eficiencia, los métodos de Newton y de la secante al calcular un cero simple α de una función f tal que $f^{(2)}(\alpha) \neq 0$.*

SOLUCIÓN:

Queremos calcular un cero de una función f con una aproximación ϵ usando los métodos de Newton y de la secante a partir de un valor aproximado con error ϵ_0. Consideraremos que el coste de una evaluación de la función f es la unidad e indicaremos por c el coste de una evaluación de la función derivada f'. Despreciaremos los costes correspondientes a las operaciones aritméticas en comparación con los de evaluación de las funciones al aplicar las fórmulas iterativas.

Para aplicar la fórmula iterativa del método de Newton hay que evaluar f y su derivada f'; en cambio, para el método de la secante, sólo es necesario evaluar la función f.

De las características de ambos métodos, se deduce que:

- N iteraciones de Newton tienen un coste de $N(1+c)$ y reducen aproximadamente el error inicial ϵ_0 a $\epsilon_0^{2^N}$, ya que el método de Newton tiene orden 2 en este caso.
- S iteraciones de la secante tienen un coste de S y reducen aproximadamente el error inicial ϵ a $\epsilon_0^{p^S}$, donde $p = \frac{1+\sqrt{5}}{2}$ es el orden del método de la secante en este caso.

Nótese que las constantes asintóticas se han considerado iguales a 1; el resultado sería esencialmente el mismo si no se hiciese esta suposición.

Para obtener, aproximadamente, el mismo error ϵ, se deberá cumplir la condición $N \ln 2 \simeq S \ln p$. Así, el método de Newton será más eficiente que el de la secante cuando el coste total de la aplicación de aquél sea inferior al de la aplicación de éste. Esto pasará para los valores de c que cumplan:

$$N(1+c) \leq S \simeq N \frac{\ln 2}{\ln p} \ ,$$

o sea, $c \leq c_0 = \frac{\ln 2}{\ln p} - 1 \simeq 0.44$.

Por lo tanto, podemos afirmar que el método de Newton es más eficiente que el método de la secante si el coste de una evaluación de f' es menor que c_0 veces el coste de una evaluación de f.

Problema 5.6 *Calcular, por el método de Muller-Traub, todos los ceros del polinomio* $p(x) = 128x^4 - 256x^3 + 160x^2 - 32x + 1$.

SOLUCIÓN:

Tal como se ha expuesto en la parte teórica, el método de Muller-Traub se basa en la construcción de una sucesión de valores $x_0, x_1, x_2, \ldots$ que tienda hacia el cero buscado α. Esta construcción se lleva a cabo a través de los pasos siguientes:

- Se parte de tres abscisas (por ejemplo, $x_0 = -1$, $x_1 = 1$, $x_2 = 0$).
- Se construye el polinomio de grado menor o igual que 2 que pasa por los puntos $(x_0, p(x_0))$, $(x_1, p(x_1))$, $(x_2, p(x_2))$.
- Se escoge x_3 como el cero de este polinomio que sea más próximo a x_2.
- El cálculo de x_4 se realiza de manera similar, pero partiendo de los valores x_1, x_2, x_3.
- Los otros términos de la sucesión se van calculando de la misma manera, partiendo siempre de las tres últimas abscisas calculadas.

- El proceso se interrumpe cuando dos términos consecutivos de la sucesión distan menos de un error admisible prefijado que, en nuestro caso, estableceremos en 10^{-7}.

  Un estudio del polinomio $p'(x)$ en un entorno de cada uno de los ceros buscados nos indica que $1 < |p'(\xi)|$. Así, teniendo en cuenta que

$$|x_k - \alpha| = \left|\frac{p(x_k)}{p'(\xi)}\right| ,$$

  con $\xi \in < x_k, \alpha >$, $|p(x_k)|$ es una cota del error de la aproximación x_k. Usaremos esta cota para estimar la precisión de los resultados α_i que obtengamos.

Aplicando el algoritmo de Muller-Traub al polinomio $p(x)$ obtenemos la sucesión de valores:

$$\begin{aligned} x_0 &= -1 , \\ x_1 &= 1 , \\ x_2 &= 0 , \\ x_3 &= 0.003484363 , \\ x_4 &= 0.032763594 , \\ x_5 &= 0.038144149 , \\ x_6 &= 0.038060056 , \\ x_7 &= 0.038060234 , \\ x_8 &= 0.038060234 . \end{aligned}$$

Así, $\alpha_1 \simeq 0.038060234$, con $| p(\alpha_1) | \simeq 5.3 \cdot 10^{-9}$.

Conocido el cero α_1, podemos proceder, por la regla de Horner, a una deflación del polinomio $p(x)$

$$p(x) = (x - \alpha_1)p_1(x) + p(\alpha_1) ,$$

con

$$p_1(x) = 128x^3 - 251.12829x^2 + 150.442x - 26.274142 .$$

Al efectuar una nueva iteración de Muller-Traub sobre $p_1(x)$ obtenemos:

$$\begin{aligned} x_0 &= 1 , \\ x_1 &= 0.5 , \\ x_2 &= 0 , \\ x_3 &= 0.43103863 , \\ x_4 &= 0.33287561 , \\ x_5 &= 0.31182922 , \\ x_6 &= 0.30862015 , \\ x_7 &= 0.30865827 , \\ x_8 &= 0.30865826 . \end{aligned}$$

El valor $\tilde{\alpha_2} = x_8$ es un cero aproximado del polinomio $p_1(x)$ cuyos coeficientes están afectados de un error debido al proceso de deflación de $p(x)$ que se ha realizado con un cero α_1 que no era exacto. Por esta razón conviene purificar $\tilde{\alpha_2}$, aplicando el método de Muller-Traub al polinomio inicial $p(x)$, tomando como valores de partida los x_6, x_7 y $\tilde{\alpha_2} = x_8$ calculados en el proceso que se ha llevado a cabo sobre $p_1(x)$.

Obtenemos $\alpha_2 \simeq 0.30865828$, con $|p(\alpha_2)| \simeq 3.3 \cdot 10^{-8}$.

Procediendo a una nueva deflación

$$p_1(x) = (x - \alpha_2)p_2(x) + p_1(\alpha_2) \ ,$$

se obtiene

$$p_2(x) = 128x^2 - 211.620030x + 85.123725 \ .$$

Dado que es de segundo grado, el cálculo de los ceros de $p_2(x)$ no tiene ninguna dificultad. Se encuentran: $\tilde{\alpha_3} = 0.69134177$ y $\tilde{\alpha_4} = 0.96193972$.

Nuevamente convendrá purificar $\tilde{\alpha_3}$ y $\tilde{\alpha_4}$ de los errores que se hayan podido acumular a lo largo de las dos deflaciones realizadas. Igual que hemos hecho para $\tilde{\alpha_2}$, este refinamiento se lleva a cabo aplicando el método de Muller-Traub al polinomio inicial $p(x)$. Se obtiene: $\alpha_3 = 0.69134172$, con $|p(\alpha_3)| \simeq 3.3 \cdot 10^{-8}$ y $\alpha_4 = 0.96193977$, con $|p(\alpha_4)| \simeq 7.8 \cdot 10^{-8}$.

Problema 5.7 *a) Indicar cómo puede extenderse el método de Sturm al caso de un polinomio real con ceros reales múltiples.*

b) Aplicación: Separar los ceros reales de

$$p(x) = x^6 - 2x^5 + x^4 - 4x^2 + 8x - 4 \ .$$

SOLUCIÓN:

a) Dado un polinomio real $p(x) = a_n x^n + \ldots + a_0$, construimos la sucesión

$$\begin{aligned} p_0(x) &= p(x) \ , \\ p_1(x) &= p'(x) \ , \\ p_{i-1}(x) &= q_i(x)p_i(x) - c_i p_{i+1}(x) \quad (i = 1 \div m) \ , \end{aligned}$$

con la ayuda del algoritmo de Euclides.

Esta sucesión acaba cuando $p_{m+1}(x) \equiv 0$. Resulta evidente que $m \leq n$, ya que $p_m(x) =$ m.c.d.$(p(x), p'(x))$.

Si todos los ceros reales de $p(x)$ son simples, entonces $\{p_0(x), \ldots, p_m(x)\}$ es una sucesión de Sturm para $p(x)$ sobre cualquier intervalo de la recta real y se cumple, por el teorema de Sturm, que si $p(a)p(b) \neq 0$, el número de ceros reales de $p_0(x)$ en (a, b) es $V(a) - V(b)$, siendo $V(x)$ el número de cambios de signo en la sucesión $\{p_0(x), \ldots, p_m(x)\}$.

Si $p(x)$ tiene ceros reales múltiples, entonces $p_m(x)$ se anula en algunos valores reales. Dado que $p_m(x)$ divide a $p_0(x)$ y $p_1(x)$, es fácil ver por inducción que divide a todos los polinomios $p_i(x)$ $(i = 0 \div m)$. Si consideramos los polinomios

$$f_i(x) = \frac{p_i(x)}{p_m(x)} \ (i = 0 \div m) \ ,$$

resulta que $\{f_0(x), \ldots, f_m(x)\}$ es una sucesión de Sturm para $f_0(x)$ sobre cualquier intervalo de la recta real ya que las condiciones 1, 2 y 4 son inmediatas (véase el apartado 5.2.4) y, para demostrar la condición 3, tan solo tenemos que derivar

$$p_0(x) = f_0(x)p_m(x)$$

para obtener

$$p_0'(x) = f_0'(x)p_m(x) + f_0(x)p_m'(x) \ .$$

Si $f_0(\alpha) = 0$, entonces

$$p_0'(\alpha) = f_0'(\alpha)p_m(\alpha) \ ,$$

$$f_1(\alpha)f_0'(\alpha) = \frac{p_1(\alpha)}{p_m(\alpha)}\frac{p_0'(\alpha)}{p_m(\alpha)} = \left(\frac{p_0'(\alpha)}{p_m(\alpha)}\right)^2 > 0 \ ,$$

ya que, si $p_0'(\alpha) = 0$, también pasa que $p_m(\alpha) = 0$ y, en la fracción, se eliminan los factores anuladores.

Por el teorema de Sturm, si $f_0(a)f_0(b) \neq 0$, el número de ceros reales de $f_0(x)$ en (a, b) es igual a $W(a) - W(b)$, siendo $W(x)$ el número de cambios de signo de la sucesión $\{f_0(x), \ldots, f_m(x)\}$. Nótese que si $p(x) \neq 0$, entonces $p_m(x) \neq 0$ y $V(x) = W(x)$. Por otro lado, dado que $f_0(x)$ tiene los mismos ceros que $p_0(x)$, pero simples, el número de ceros reales de $f_0(x)$ en (a, b) es igual al número de ceros reales diferentes del polinomio $p_0(x)$ en (a, b), si $p_0(a)p_0(b) \neq 0$. Así, se cumple el resultado siguiente:

- Si $p(a)p(b) \neq 0$, el número de ceros reales diferentes de $p(x)$ en (a, b) es igual a $V(a) - V(b)$.

b) Seguidamente aplicamos este resultado al polinomio

$$p(x) = x^6 - 2x^5 + x^4 - 4x^2 + 8x - 4 \ .$$

En primer lugar construimos la sucesión de polinomios asociada. Comenzamos con $p_0(x) = p(x)$, $p_1(x) = p'(x)$. Para el cálculo de $p_2(x)$, dividimos $p_0(x)$ entre $p_1(x)$, obteniendo

$$p_0(x) = (\frac{1}{6}x - \frac{1}{18})p_1(x) - \frac{2}{9}x^4 + \frac{2}{9}x^3 - \frac{24}{9}x^2 + \frac{56}{9}x - \frac{32}{9} \ .$$

Introduciendo $p_2(x) = x^4 - x^3 + 12x^2 - 28x + 16$, resulta que

$$p_0(x) = (\frac{1}{6}x - \frac{1}{18})p_1(x) - \frac{2}{9}p_2(x) \ .$$

El proceso se puede continuar:

- Dividiendo $p_1(x)$ entre $p_2(x)$, se obtendría

$$p_1(x) = (6x - 4)p_2(x) - 72p_3(x) \ ,$$

con $p_3(x) = x^3 - 3x^2 + 3x - 1$.

- Dividiendo $p_2(x)$ entre $p_3(x)$, se obtiene

$$p_2(x) = (x + 2)p_3(x) - 3p_4(x) \ ,$$

con $p_4(x) = -5x^2 + 11x - 6$.

- Dividiendo $p_3(x)$ entre $p_4(x)$, se obtiene

$$p_3(x) = (-\frac{x}{5} + \frac{4}{25})p_4(x) - \frac{1}{25}p_5(x) \ ,$$

con $p_5(x) = -x + 1$.

- Dividiendo $p_4(x)$ entre $p_5(x)$, se obtiene

$$p_4(x) = (5x - 6)p_5(x)$$

y se acaba el proceso.

Obsérvese que, usando las relaciones de recurrencia entre los polinomios $p_i(x)$, para calcular $\{p_0(x), \ldots, p_5(x)\}$ para un valor x dado, son necesarias únicamente 10 sumas y 13 multiplicaciones.

Ya que $p_5(x) = -x + 1 = \text{m.c.d.}(p(x), p'(x))$, resulta que 1 es un cero doble de $p(x)$. Para localizar los otros ceros reales, estudiemos los signos de la sucesión de polinomios para algunos valores x. Se obtiene la siguiente tabla:

x	$-\infty$	-2	-1	0	1	$1.192...$	1.25	1.5	2	$+\infty$
$p_0(x)$	$+$	$+$	$-$	$-$	0	$-$	$-$	$+$	$+$	$+$
$p_1(x)$	$-$	$-$	$-$	$+$	0	$-$	$-$	$+$	$+$	$+$
$p_2(x)$	$+$	$+$	$+$	$+$	0	0	$+$	$+$	$+$	$+$
$p_3(x)$	$-$	$-$	$-$	$-$	0	$+$	$+$	$+$	$+$	$+$
$p_4(x)$	$-$	$-$	$-$	$-$	0	$+$	$-$	$-$	$-$	$-$
$p_5(x)$	$+$	$+$	$+$	$+$	0	$-$	$-$	$-$	$-$	$-$
$V(x)$	4	4	3	3	$-$	2	2	1	1	1

HISTORIA DE LA TABLA:

Hemos comenzado con $-\infty, +\infty$ y, debido a que $V(-\infty) - V(+\infty) = 3$, hay tres ceros reales diferentes, uno de los cuales ya sabemos que es $x = 1$. Hemos ensayado después con 0 y, como $V(-\infty) - V(0) = 1$, hay un cero real en el intervalo $(-\infty, 0)$. Probando con -1 y -2, observamos que este cero real negativo estará en $(-2, -1)$, ya que $V(-2) - V(-1) = 1$. Tanteando con 2, 1.5 y 1.25, detectamos un cero real en $(1.25, 1.5)$. Así quedan separados los tres ceros reales.

Hay que hacer dos observaciones:

1. Al ser $x = 1$ cero múltiple, se cumple que $p_i(1) = 0$ $(i = 0 \div 5)$.

2. Si probamos con $\bar{x} = 1.192143\ldots$, para el cual $p_2(x)$ se anula, aparece un 0 en la columna de signos. En estos casos se ha de ignorar este valor y calcular el número de cambios de signo según los elementos anterior y posterior. Así, si tenemos $-, 0, +$, hay un cambio de signo, pero, si tenemos $-, 0, -$, no hay ningún cambio de signo.

Problema 5.8 *Queremos calcular todos los ceros del polinomio*

$$p(x) = x^4 - 2x^3 - 14x^2 - 2x - 15 \ .$$

Usar el método Q-D para obtener una primera aproximación y los métodos de Birge-Vieta y Bairstow para precisar los resultados.

SOLUCIÓN:

La resolución de este ejercicio se realiza en tres apartados.

a) APROXIMACIÓN DE LOS CEROS POR EL MÉTODO Q-D

El algoritmo descrito en la parte teórica del capítulo nos permite construir la tabla siguiente:

	$q_k^{(1)}$	$d_k^{(1)}$	$q_k^{(2)}$	$d_k^{(2)}$	$q_k^{(3)}$	$d_k^{(3)}$	$q_k^{(4)}$	
	2		0		0		0	
0		7		0.1429		7.5		0
	9		−6.8571		7.3571		−7.5	
0		−5.3333		−0.1533		−7.6456		0
	3.6667		−1.6771		−0.1352		0.1456	
0		2.4394		−0.0124		8.2347		0
	6.1061		−4.1288		8.1118		−8.0890	
0		−1.6495		0.0243		−8.2115		0
	4.4566		−2.4551		−0.1240		0.1225	
0		0.9087		0.0012		8.1120		0
	5.3653		−3.3625		7.9867		−7.9894	
0		−0.5695		−0.0029		−8.1147		0

En esta tabla podemos observar que los valores de las columnas $d^{(1)}$ i $d^{(2)}$ se van haciendo pequeños, lo cual indica que las sucesiones $(q_k^{(1)})_{k\geq 1}$ y $(q_k^{(2)})_{k\geq 0}$ tienden hacia ceros reales de $p(x)$. Podemos considerar los valores 5.3653 y -3.3625 como aproximaciones de estos ceros.

El comportamiento oscilatorio de la columna $d^{(3)}$ nos indica la existencia de un par de ceros complejos conjugados. El cálculo de estos ceros se llevará a cabo a partir de la determinación del correspondiente factor cuadrático de $p(x)$. Este factor será del tipo $x^2 + rx + s$, donde

$$r = -\lim_{k\to\infty}(q_{k+1}^{(3)} + q_k^{(4)}) \ ,$$

$$s = \lim_{k\to\infty} q_k^{(3)} q_k^{(4)} \ ,$$

y se calculará por el método de Bairstow, a partir de un factor con valores aproximados de r y s.

Con los datos que aparecen en la tabla, observamos que:

- Los primeros términos de la sucesión $(q_{k+1}^{(3)} + q_k^{(4)})_{k\geq -2}$ son:

$$-0.142857,\ 0.010417,\ 0.0228,\ -0.0015,\ -0.0027, \ldots$$

- Los primeros términos de la sucesión $(q_k^{(3)} q_k^{(4)})_{k\geq -2}$ son:

$$0,\ 1.0714,\ 1.0938,\ 0.9938,\ 0.9908, \ldots$$

Así, podremos tomar como aproximaciones de r y de s los valores 0.00273 y 0.99081, respectivamente.

b) Refinamiento de los ceros reales por el método de Birge-Vieta

Utilizando la iteración

$$x_{n+1} = x_n - \frac{p(x_n)}{p'(x_n)} \ ,$$

a partir de los valores aproximados calculados por el método Q-D, obtenemos:

$x_0 = 5.3652$	$x_0 = -3.362527$
$x_1 = 5.054466$	$x_1 = -3.070510$
$x_2 = 5.001458$	$x_2 = -3.003374$
$x_3 = 5.000001$	$x_3 = -3.000008$
$x_4 = 5.000000$	$x_4 = -3.000000$

Por tanto, los dos ceros reales de $p(x)$ son 5 y -3.

Ahora podríamos aplicar el método de deflación a $p(x)$ y obtener, sin gran dificultad, los ceros restantes, pero no lo haremos con el fin de mostrar un ejemplo de aplicación del método de Bairstow.

c) Cálculo de los ceros complejos por el método de Bairstow

Partiendo del factor aproximado $x^2 + 0.0027x + 0.9908$ y usando el algoritmo de Bairstow, descrito en el apartado 5.2.5, tenemos $r_0 = 0.0027$ y $s_0 = 0.9908$.

En la primera iteración, los valores:

$$\begin{aligned} b_1 &= -2.0027\ , & c_1 &= -2.00546\ , \\ b_0 &= -14.98534255\ , & c_0 &= -15.97067764\ , \\ b_{-1} &= 0.025234896\ , & c_{-1} &= 2.055864669\ , \\ b_{-2} &= -0.152441640\ , \end{aligned}$$

nos permiten construir el sistema

$$\left.\begin{array}{rcrcrcr} -15.97067764 & \Delta r_0 & - & 2.00546 & \Delta s_0 & = & 0.025234896 \\ 2.030629773 & \Delta r_0 & - & 15.97067764 & \Delta s_0 & = & -0.152441640 \end{array}\right\} \ ,$$

cuya solución es

$$\begin{aligned} \Delta r_0 &= -2.735 \cdot 10^{-3}\ , \\ \Delta s_0 &= 9.197347 \cdot 10^{-3}\ ; \end{aligned}$$

de manera que $r_1 = r_0 + \Delta r_0 = -5 \cdot 10^{-6}$ y $s_1 = s_0 + \Delta s_0 = 1.000007347$.

En la segunda iteración, los valores

$$\begin{aligned} b_1 &= -1.999995\ , & c_1 &= -1.99999\ , \\ b_0 &= -15.00001735\ , & c_0 &= -16.00003470\ , \\ b_{-1} &= -6.5306 \cdot 10^{-5}\ , & c_{-1} &= 1.999859388\ , \\ b_{-2} &= 1.2755967 \cdot 10^{-4}\ , \end{aligned}$$

nos permiten construir el sistema

$$\left.\begin{array}{rcrcrcl} -16.00003470 & \Delta r_1 & - & 1.99999 & \Delta s_1 & = & -6.5306 \cdot 10^{-5} \\ 1.999924694 & \Delta r_1 & - & 16.00003470 & \Delta s_1 & = & 1.2755967 \cdot 10^{-4} \end{array}\right\} ,$$

cuya solución es

$$\begin{array}{rl} \Delta r_1 = & 5 \cdot 10^{-6} \ , \\ \Delta s_1 = & -7.3474 \cdot 10^{-6} \ ; \end{array}$$

de manera que $r_2 = r_1 + \Delta r_1 = 0$ y $s_2 = s_1 + \Delta s_1 = 1$.

Al efectuar una tercera iteración se obtiene $\Delta r_2 = \Delta s_2 = 0$, de manera que los valores definitivos de r y s son, respectivamente, 0 y 1. El factor cuadrático buscado es $x^2 + 1$, que tiene por ceros i, $-i$.

Así, los ceros del polinomio

$$p(x) = x^4 - 2x^3 - 14x^2 - 2x - 15 = (x-5)(x+3)(x^2+1)$$

son $5, -3, i, -i$.

Problema 5.9 *a) Sean $p(x)$ y $q(x)$ dos polinomios y ϵ un número real, pequeño en valor absoluto. ¿Cómo pueden aproximarse los ceros de $\overline{p}(x) = p(x) + \epsilon q(x)$, conocidos los de $p(x)$ y supuestos éstos simples o dobles?*

b) Aplicar la respuesta a la pregunta anterior al cálculo aproximado de los ceros del polinomio $\overline{p}(x) = p(x) + \epsilon q(x)$, con

$$p(x) = x^3 - 9x^2 + 15x + 25 \ , \quad q(x) = -x^2 \ , \quad \epsilon = 10^{-6} \ .$$

SOLUCIÓN:

a) Obsérvese que la cuestión que se plantea es cómo quedan perturbados los ceros de un polinomio al perturbar ligeramente sus coeficientes.

Sea α un cero de $p(x)$ y $\overline{\alpha}$ un cero de $\overline{p}(x)$ y supongamos que $\overline{\alpha}$ proviene de la evolución seguida por α al aumentar el parámetro ϵ, partiendo de 0. Consideraremos siempre que $q(\alpha) \neq 0$ ya que si no α sería también cero de $\overline{p}(x)$ y no estaría afectado por la perturbación ($\overline{\alpha} = \alpha$). Sea $\delta = \overline{\alpha} - \alpha$, nos proponemos aproximar $\delta(\epsilon)$ del cual tan solo conocemos que $\delta(0) = 0$.

Consideramos el desarrollo de Taylor de $\overline{p}(x)$ en un entorno de α, evaluado en $\overline{\alpha}$

$$0 = \overline{p}(\overline{\alpha}) = \overline{p}(\alpha + \delta) = \overline{p}(\alpha) + \overline{p}'(\alpha)\delta + \frac{\overline{p}^{(2)}(\alpha)}{2}\delta^2 + \frac{\overline{p}^{(3)}(\alpha)}{3!}\delta^3 + \cdots \ ,$$

y sustituimos $\overline{p}(x)$ por $p(x) + \epsilon q(x)$

$$\begin{array}{rcl} 0 & = & \epsilon q(\alpha) + (p'(\alpha) + \epsilon q'(\alpha))\delta + (p^{(2)}(\alpha) + \epsilon q^{(2)}(\alpha))\dfrac{\delta^2}{2} \\ & & +(p^{(3)}(\alpha) + \epsilon q^{(3)}(\alpha))\dfrac{\delta^3}{3!} + \cdots \end{array}$$

Ahora conviene distinguir el caso en que α es un cero simple de $p(x)$ del caso en que es doble.

Caso de cero simple

Para obtener una aproximación de δ, tomaremos los dos primeros términos del desarrollo de Taylor,

$$0 \simeq \epsilon q(\alpha) + (p'(\alpha) + \epsilon q'(\alpha))\delta \ ,$$

y despejaremos δ:

$$\delta \simeq -\frac{\epsilon q(\alpha)}{p'(\alpha) + \epsilon q'(\alpha)} = -\frac{\epsilon q(\alpha)}{p'(\alpha)} \left(1 + \epsilon \frac{q'(\alpha)}{p'(\alpha)}\right)^{-1} = -\epsilon \frac{q(\alpha)}{p'(\alpha)} + \mathcal{O}(\epsilon^2) \ .$$

Si hubiésemos tenido en cuenta también los términos en δ^2 en el desarrollo de Taylor, habríamos obtenido el mismo resultado, una vez desarrollada la expresión de δ.

Si el grado de $p(x)$ es 1 y el de $q(x)$ es 0, la expresión obtenida de δ es exacta y se ilustra en la figura 5.6. En esta figura, puesto que

$$\tan\varphi = p'(\alpha) = \frac{\epsilon q(\alpha)}{-\delta} \ ,$$

resulta que

$$\delta = -\frac{\epsilon q(\alpha)}{p'(\alpha)} \ ,$$

como estaba previsto.

Caso de cero doble

Si, en este caso, tomásemos solamente los dos primeros términos del desarrollo de Taylor, no se aportaría ningún dato más para la determinación de δ que la anulación de $p(\alpha)$ y de $p'(\alpha)$, de manera que actuaríamos como si el polinomio $p(x)$ fuese idénticamente nulo y, en realidad, estaríamos buscando un cero de $q(x)$: la expresión obtenida para δ sería igual que el incremento que obtendríamos al aplicar el método de Newton a $q(x)$ a partir de α. Para evitar esta pérdida de información sobre $p(x)$, tendremos en cuenta los tres primeros términos del desarrollo

$$0 \simeq \epsilon q(\alpha) + \epsilon q'(\alpha)\delta + (p^{(2)}(\alpha) + \epsilon q^{(2)}(\alpha))\frac{\delta^2}{2} \ .$$

Así,

$$\delta \simeq \frac{-\epsilon q'(\alpha) \pm H(\alpha)}{p^{(2)}(\alpha) + \epsilon q^{(2)}(\alpha)} \ ,$$

con

$$H(\alpha) = (\epsilon^2 q'^2(\alpha) - 2\epsilon q(\alpha)(p^{(2)}(\alpha) + \epsilon q^{(2)}(\alpha)))^{\frac{1}{2}} \ .$$

Usando desarrollos de Taylor, encontramos las expresiones:

$$H(\alpha) = \sqrt{-2q(\alpha)p^{(2)}(\alpha)}\epsilon^{\frac{1}{2}} + \mathcal{O}(\epsilon^{\frac{3}{2}}) \ ,$$

$$\frac{1}{p^{(2)}(\alpha) + \epsilon q^{(2)}(\alpha)} = \frac{1}{p^{(2)}(\alpha)} - \epsilon \frac{q^{(2)}(\alpha)}{p^{(2)2}(\alpha)} + \mathcal{O}(\epsilon^2) \ ;$$

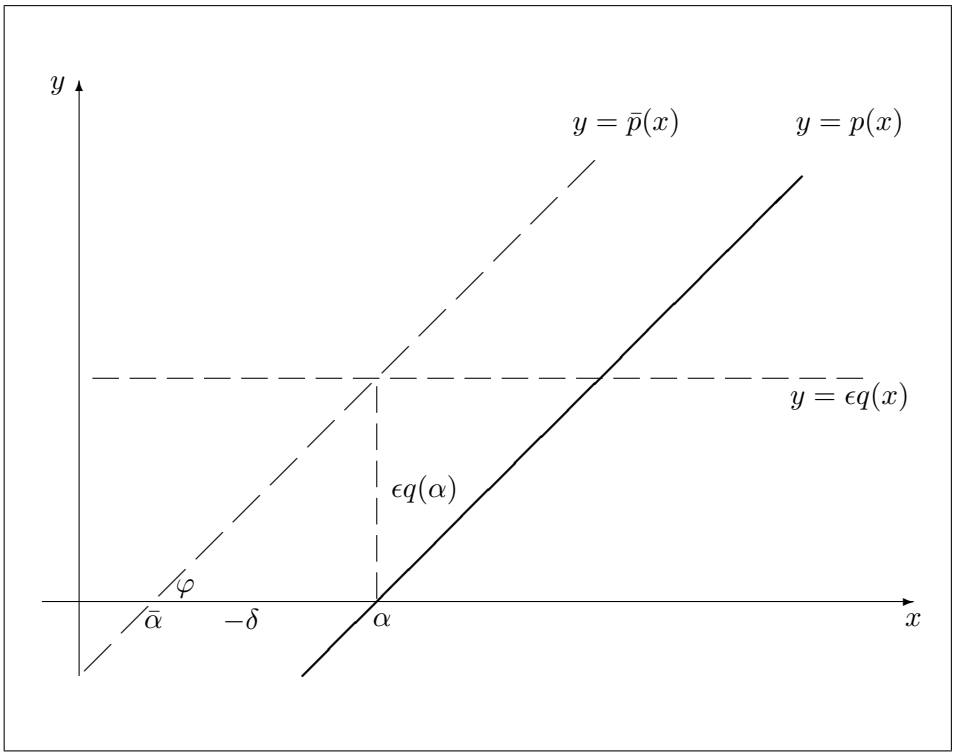


Figura 5.6: Función lineal desplazada por constante.

sustituyéndolas en la expresión de δ, tenemos

$$\delta = \pm\sqrt{\frac{-2q(\alpha)}{p^{(2)}(\alpha)}}\epsilon^{\frac{1}{2}} - \frac{q'(\alpha)}{p^{(2)}(\alpha)}\epsilon + \mathcal{O}(\epsilon^{\frac{3}{2}}) \ .$$

Así, tomando los términos dominantes,

$$\delta \simeq \pm\sqrt{\frac{-2q(\alpha)}{p^{(2)}(\alpha)}}\epsilon^{\frac{1}{2}} - \frac{q'(\alpha)}{p^{(2)}(\alpha)}\epsilon \ .$$

Aquí conviene hacer tres comentarios:

- El doble signo de δ nos indica que el cero doble α de $p(x)$, al ser perturbado, evoluciona (*se bifurca*) hacia dos ceros.
- Consideremos el caso en que el grado de $p(x)$ sea 2 y el de $q(x)$ sea 0. Entonces la expresión encontrada para δ será exacta y podremos realizar un estudio de lo que pasa:
  - $p(x)$ se representará por una parábola con un punto de contacto α con el eje de abscisas, cuya concavidad o convexidad vendrá dada por el signo de la constante $p^{(2)}(x)$.
  - $q(x)$ será una constante que gráficamente perturbará a $p(x)$ trasladando la parábola en sentido vertical según los signos de $q(x)$ y de ϵ.

Si suponemos $\epsilon > 0$ (para $\epsilon < 0$ sería análogo), podemos describir en el cuadro siguiente el comportamiento de los ceros:

signo de $p^{(2)}(\alpha)$	signo de $q(\alpha) = C$	δ es ...	α se bifurca a dos ceros ...	Figura 5.7
+	+	imaginaria	complejos	a)
+	−	real	reales	b)
−	+	real	reales	c)
−	−	imaginaria	complejos	d)

- En el caso de que α fuese un cero de orden k de $p(x)$, se podría reproducir un razonamiento similar y se obtendría una expresión aproximada del tipo

$$0 \simeq \epsilon p_k(\alpha + \delta) + \frac{\delta^k p^{(k)}(\alpha)}{k!} ,$$

donde $p_k(x)$ es el polinomio de Taylor de grado k correspondiente a la función $q(x)$ en un entorno de α:

$$p_k(x) = q(\alpha) + q'(\alpha)(x-\alpha) + \cdots + \frac{q^{(k)}(\alpha)}{k!}(x-\alpha)^k .$$

En este caso, una aproximación de δ, vendría dada por

$$\delta = \sqrt[k]{\frac{-k!q(\alpha)}{p^{(k)}(\alpha)}}\epsilon^{\frac{1}{k}} + \mathcal{O}(\epsilon^{\frac{2}{k}}) .$$

Nótese que, para obtener δ con el mismo orden de precisión en ϵ que en el caso de ceros simples, tendríamos que considerar k términos en su desarrollo.

b) Apliquemos ahora el estudio anterior al cálculo aproximado de los ceros del polinomio $\overline{p}(x) = p(x) + \epsilon q(x)$, con $p(x) = x^3 - 9x^2 + 15x + 25$, $q(x) = -x^2$ y $\epsilon = 10^{-6}$.

Obtenemos:

$$p(x) = (x+1)(x-5)^2 , \quad \overline{p}(x) = x^3 - (9 + 10^{-6})x^2 + 15x + 25 .$$

El cero $\alpha_1 = -1$ de $p(x)$ pasará a un cero $\overline{\alpha}_1 = -1 + \delta_1$ de $\overline{p}(x)$ y el cero doble $\alpha_2 = 5$ de $p(x)$ se bifurcará en dos ceros $\overline{\alpha}_2 = 5 + \delta_2$ y $\overline{\alpha}_3 = 5 + \delta_3$, donde:

$$\delta_1 \simeq -\epsilon\frac{q(\alpha_1)}{p'(\alpha_1)} = -10^{-6}\frac{q(-1)}{p'(-1)} = 2.778 \cdot 10^{-8} ,$$

$$\delta_2 \simeq \epsilon^{\frac{1}{2}}\sqrt{\frac{-2q(\alpha_2)}{p^{(2)}(\alpha_2)}} - \epsilon\frac{q'(\alpha_2)}{p^{(2)}(\alpha_2)} = 2.04207 \cdot 10^{-3} ,$$

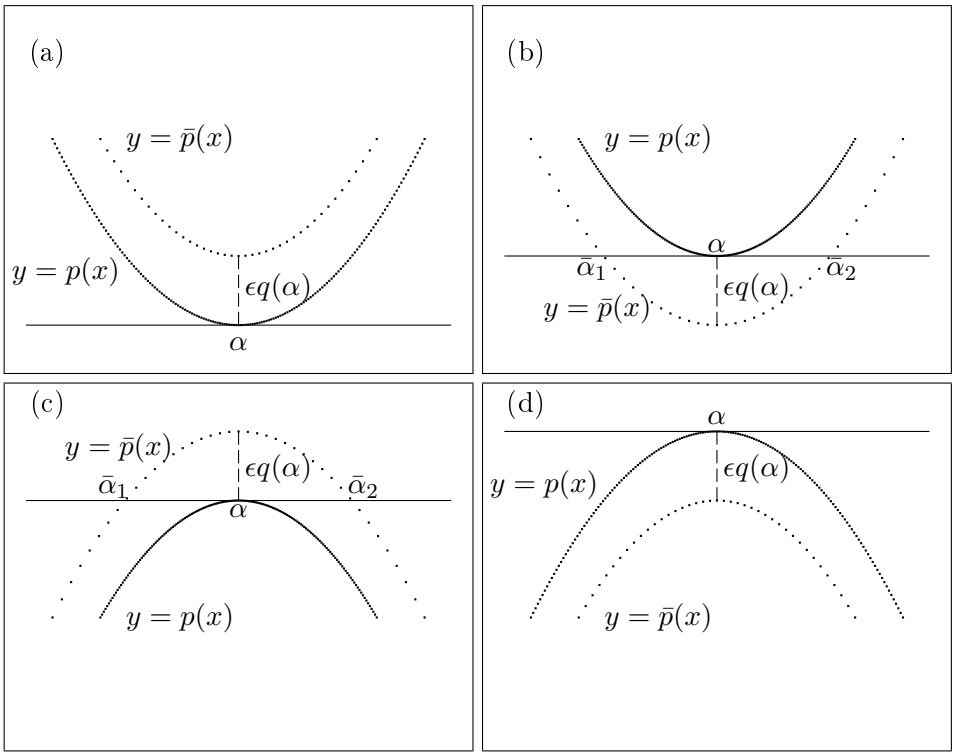


Figura 5.7: Función cuadrática desplazada por una constante.

$$\delta_3 \simeq -2.04207 \cdot 10^{-3} \ .$$

Así:

$$\begin{array}{llll} \overline{\alpha}_1 \simeq & -1 + 2.778 \cdot 10^{-8} \ , & \overline{p}(\overline{\alpha}_1) \simeq & 9 \cdot 10^{-11} \ , \\ \overline{\alpha}_2 \simeq & 5 + 2.04207 \cdot 10^{-3} \ , & \overline{p}(\overline{\alpha}_2) \simeq & 8.5 \cdot 10^{-9} \ , \\ \overline{\alpha}_3 \simeq & 5 - 2.04207 \cdot 10^{-3} \ , & \overline{p}(\overline{\alpha}_3) \simeq & -8.5 \cdot 10^{-9} \ . \end{array}$$

Problema 5.10 *Aplicar el método de Newton con dos variables para calcular la solución del sistema no lineal:*

$$\left.\begin{array}{rl} x_1 = & \operatorname{sen}(x_1 + x_2) \\ x_2 = & \cos(x_1 - x_2) \end{array}\right\}$$

cerca de $x_1 = 1$, $x_2 = 1$. *Acábese el proceso cuando el vector residual, resultante de restar los dos miembros de cada ecuación, sea menor que* 10^{-10}, *en* $\| \ \|_\infty$.

SOLUCIÓN:

Escribimos el sistema en la forma:

$$f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} f_1(x_1,x_2) \\ f_2(x_1,x_2) \end{pmatrix} = \begin{pmatrix} x_1 - \operatorname{sen}(x_1+x_2) \\ x_2 - \cos(x_1-x_2) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} .$$

El método de Newton con dos variables se concreta en la fórmula iterativa:

$$\begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{pmatrix} = \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \end{pmatrix} - \left[Df\begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \end{pmatrix}\right]^{-1} f\begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \end{pmatrix} .$$

En nuestro caso,

$$Df\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1-\cos(x_1+x_2) & -\cos(x_1+x_2) \\ \operatorname{sen}(x_1-x_2) & 1-\operatorname{sen}(x_1-x_2) \end{pmatrix}$$

y el cálculo de la inversa da

$$\left[Df\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right]^{-1} = \frac{1}{\Delta(x_1,x_2)} \begin{pmatrix} 1-\operatorname{sen}(x_1-x_2) & \cos(x_1+x_2) \\ -\operatorname{sen}(x_1-x_2) & 1-\cos(x_1+x_2) \end{pmatrix} ,$$

donde

$$\Delta(x_1,x_2) = 1 - \operatorname{sen}(x_1-x_2) - \cos(x_1+x_2) + 2\operatorname{sen}(x_1-x_2)\cos(x_1+x_2) .$$

Así, la iteración se escribe:

$$\begin{aligned} x_1^{(k+1)} &= x_1^{(k)} - \frac{(1-\operatorname{sen}(x_1^{(k)}-x_2^{(k)}))(x_1^{(k)}-\operatorname{sen}(x_1^{(k)}+x_2^{(k)}))}{\Delta(x_1^{(k)},x_2^{(k)})} \\ &\quad - \frac{\cos(x_1^{(k)}+x_2^{(k)})(x_2^{(k)}-\cos(x_1^{(k)}-x_2^{(k)}))}{\Delta(x_1^{(k)},x_2^{(k)})} , \end{aligned}$$

$$\begin{aligned} x_2^{(k+1)} &= x_2^{(k)} + \frac{\operatorname{sen}(x_1^{(k)}-x_2^{(k)})(x_1^{(k)}-\operatorname{sen}(x_1^{(k)}+x_2^{(k)}))}{\Delta(x_1^{(k)},x_2^{(k)})} \\ &\quad - \frac{(1-\cos(x_1^{(k)}+x_2^{(k)}))(x_2^{(k)}-\cos(x_1^{(k)}-x_2^{(k)}))}{\Delta(x_1^{(k)},x_2^{(k)})} . \end{aligned}$$

Haciendo los cálculos se obtiene la siguiente tabla de valores:

k	$x_1^{(k)}$	$x_2^{(k)}$
0	1	1
1	0.9359511522	1
2	0.9350846651	0.9980207933
3	0.9350820642	0.9980200582
4	0.9350820641	0.9980200582

Se puede comprobar que

$$\left\| f\begin{pmatrix} x_1^{(4)} \\ x_2^{(4)} \end{pmatrix} \right\|_\infty \simeq 10^{-11}$$

y, por tanto, la solución del sistema, con la precisión deseada, es

$$x_1^{(4)} = 0.9350820641 , \quad x_2^{(4)} = 0.9980200582 .$$

PROBLEMAS PROPUESTOS

1. Demostrar que la función
$$f(x) = \frac{x^2}{e^x} - 1$$
tiene un único cero.

2. Demostrar que los polinomios de la forma $x^n + x + 1$ no tienen ningún cero real, si n es par, y tienen exactamente uno, si n es impar.

3. a) Discutir el número de soluciones de la ecuación
$$M = x - e \operatorname{sen} x \quad \textit{(ecuación de Kepler elíptica)} \ ,$$
para todos los valores de $M \in [0, 2\pi)$ y $e \in (0, 1)$.
b) Hacer lo mismo con la ecuación
$$M = e \operatorname{senh} x - x \quad \textit{(ecuación de Kepler hiperbólica)} \ ,$$
para todos los valores de $M > 0$ y $e > 1$.

4. Usando el método de bisección, hallar todas las raíces de las siguientes ecuaciones con un error menor que 10^{-2}:
$$\text{i) } \operatorname{sen} x = \frac{1}{2}x^2 \ , \quad \text{ii) } e^{-x} = x^4 \ , \quad \text{iii) } \ln x = x - 1 \ .$$

5. ¿Qué se obtiene cuando escribimos un número cualquiera en la pantalla de una calculadora y pulsamos reiteradamente la tecla $\boxed{\cos}$ (en radianes)? Responder a la misma pregunta si, en lugar de pulsar $\boxed{\cos}$, pulsamos $\boxed{\arccos}$. Interpretar los resultados.

6. Queremos usar la fórmula de iteración $x_{k+1} = 2^{x_k - 1}$ para resolver la ecuación $2x = 2^x$. Estudiar para qué valores de x_0 converge y, en tal caso, a qué límite lo hace.

7. Explicar el comportamiento en el límite de la iteración $x_{k+1} = g(x_k)$ con $g(x) = 1.4 \cos x$, partiendo de un valor x_0 cualquiera.

8. Consideremos la iteración simple $x_0 = 0$, $x_{k+1} = g(x_k)$ con

$$g(x) = \frac{\epsilon - 2x^2 - x^3}{3} \ ,$$

siendo ϵ suficientemente pequeño.

a) Decidir si es convergente o no y, en caso afirmativo, hacia qué límite.

b) Aplicación: Efectuar los cálculos para $\epsilon = 0.1$ y hallar, en este caso, el orden de convergencia y la constante asintótica del error.

9. Se quiere calcular la solución de la ecuación $e^x = 3x$, usando iteración simple con diferentes funciones de iteración:

$$g_1(x) = \frac{e^x}{3} \ , \quad g_2(x) = \ln(3x) \ , \quad g_3(x) = \frac{e^x - x}{2} \ , \quad g_4(x) = e^x - 2x \ .$$

a) ¿Cuáles de éstas son útiles?

b) ¿Con cuál de ellas son necesarias menos iteraciones para obtener el resultado con una precisión dada, partiendo del mismo valor inicial? Comprobarlo numéricamente, partiendo del valor aproximado $x_0 = 0.6$.

c) Hallar una función de iteración significativamente mejor que las indicadas y usarla para el caso propuesto en b).

10. Sean f_1 y f_2 funciones derivables con continuidad en el intervalo $[a, b]$ y supongamos que una de ellas, por ejemplo f_1, tiene inversa. Queremos encontrar una raíz simple de la ecuación $f_1(x) = f_2(x)$.

Consideremos el método iterativo $f_1(x_{k+1}) = f_2(x_k)$.

a) ¿Bajo qué condiciones es localmente convergente?

b) Aplicación numérica: Calcular la raíz de $(1 + x) \operatorname{sen} x = 1$ perteneciente a $[\frac{\pi}{2}, \pi]$.

11. a) Encontrar funciones de iteración para el cálculo de $\sqrt{a}$ que tengan órdenes de convergencia 1 y 2, respectivamente.

b) Emplearlas para obtener $\sqrt{19}$ con 8 cifras decimales correctas.

12. El *método de la cuerda* está basado en la iteración

$$x_{k+1} = x_k - mf(x_k) \ (k \geq 0)$$

para el cálculo de ceros de una función f dada.

a) ¿Para qué valores de m es localmente convergente hacia un cero simple α de f? ¿Qué pasa si α es un cero múltiple de f?

b) Hallar el orden de convergencia en todos los casos.

c) Indicar qué problemas prácticos hay cuando se quieren conseguir los órdenes más elevados.

13. Calcular por el método de Newton todas las soluciones de las ecuaciones

$$\text{i) } x = \cos x \ , \quad \text{ii) } x^3 = 10 \operatorname{sen} x \ , \quad \text{iii) } x^2 = e^{-x} \ ,$$

con 8 cifras decimales y comprobar la convergencia cuadrática del método en estos casos.

14. a) Usando el método de Newton, hallar los puntos fijos de $g(x) = e^x - 2$ con un error menor que 10^{-6}.

b) ¿En qué intervalos de valores iniciales podemos asegurar que el método converge y hacia qué solución?

15. Sean a, b reales con $b > a$ y $f \in \mathcal{C}^2([a,b])$ cumpliendo $f'(x) \neq 0$, $f^{(2)}(x) \neq 0$ $\forall x \in [a,b]$; supongamos que la función f toma signos diferentes en a y b ($f(a)f(b) < 0$) y que las correcciones que el método de Newton hace de estos puntos son menores que la anchura del intervalo:

$$\left|\frac{f(a)}{f'(a)}\right| < b - a \quad \text{y} \quad \left|\frac{f(b)}{f'(b)}\right| < b - a \ .$$

Demostrar que el método de Newton, aplicado a hallar los ceros de f a $[a,b]$, converge para cualquier $x_0 \in [a,b]$.

16. Dada f de clase $\mathcal{C}^2$ con un único punto de inflexión y tal que

$$\lim_{x \to -\infty} f(x) = -\infty \ , \quad \lim_{x \to \infty} f(x) = \infty \ .$$

¿Qué estrategia hay que seguir, según los diferentes valores de c, para garantizar la convergencia del método de Newton aplicado a la obtención de una solución de $f(x) = c$? ¿Y para obtener todas las soluciones?

17. a) Demostrar que el método de Newton aplicado a la función $f(x) = \frac{1}{x} - a$ permite calcular $\frac{1}{a}$ sin realizar divisiones.

b) ¿Qué relación exacta existe entre $e_{k+1} = x_{k+1} - \frac{1}{a}$ y $e_k = x_k - \frac{1}{a}$ $(k \geq 0)$?

c) Si $a = 0.4$ y $e_0 = -0.2$, ¿para qué valores de k tendremos $|e_k| \leq 10^{-20}$?

d) Dar, en función de a, los valores de x_0 que hacen que el método sea convergente.

18. a) Hallar el procedimiento más simple que, utilizando el método de Newton, calcule iterativamente $\frac{1}{\sqrt[r]{a}}$ para $r \in \mathbb{N}$.

    b) Aplicación: Calcular $\sqrt[3]{0.1}$.

19. a) Demostrar que la función $h(x) = \frac{f(x)}{f'(x)}$ tiene los mismos ceros que f, pero simples.

    b) ¿Qué ventajas e inconvenientes presenta la aplicación del método de Newton a la función h en lugar de su aplicación a f?

    c) Aplicar el método de Newton para hallar el cero de $f(x) = (e^x - \pi)^3$ directamente y usando la función h correspondiente.

20. Determinar para qué valor de $a > 0$ la función $f(x) = \operatorname{sen} x - 2e^{-ax}$ tiene un único cero menor que π de multiplicidad 2. Calcular entonces este cero usando un método iterativo de orden 2.

21. Consideremos la ecuación $f(x) = e^{ax} - x = 0$.

    a) Demostrar que existe a_0 tal que, si $a > a_0$, f no tiene ceros reales y, si $a \leq a_0$, sí que los tiene. Hallar a_0 e indicar cuál es el número de ceros reales en función de a.

    b) Demostrar que, para todo $a < a_0$ y para todo valor inicial real, excepto un punto como máximo, el método de Newton es convergente.

    c) Calcular todos los ceros reales para $a = \frac{1}{4}$.

22. Consideremos la ecuación $ax + b = \operatorname{sen} x$, donde a y b son parámetros reales. Determinar los valores de a para los cuales, cualquiera que sea el valor de b, la ecuación considerada tenga exactamente 3 ceros. Hacer lo mismo para que tenga exactamente 5 y 7 ceros respectivamente, para todo valor de b.

23. Hallar todos los ceros de la función $f(x) = \operatorname{sen} x - 2 - 0.3x$ con un error menor que 10^{-8}, usando el método de Newton.

24. a) Hallar el primer punto de corte positivo α entre las curvas $y = \cos x$ y $y = e^{-x}$.

    b) Calcular el área comprendida entre dichas curvas entre las abscisas 0 y α.

25. a) Modificar el método de Chebichev

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} - \frac{f^{(2)}(x_k) f^2(x_k)}{2f'^3(x_k)} \ ,$$

para que tenga orden al menos 3 cuando lo usamos para hallar ceros de multiplicidad conocida m.

b) Aplicación: Calcular, con un error menor que 10^{-10}, el cero de la función $f(x) = (x \ln x - 1)^2$ que se encuentra cerca de $x_0 = 1.8$ y comprobar numéricamente el orden de convergencia.

26. Sea f una función de la que se quiere calcular un cero simple.

a) Deducir la fórmula correspondiente al método iterativo que obtiene x_{k+1}, a partir de x_k, evaluando en $y = 0$ el polinomio de Taylor de grado menor o igual que 3 asociado a la función inversa local g de f cerca de $f(x_k)$.

b) Hallar el orden mínimo del método y el coeficiente asintótico del error cuando se da este orden mínimo.

27. Consideremos una variante del método de Newton consistente en usar este método, pero utilizando el valor calculado de la derivada q veces antes de calcular una nueva derivada:

$$x_{rq+1} = x_{rq} - \frac{f(x_{rq})}{f'(x_{rq})} \ , \quad x_{rq+i+1} = x_{rq+i} - \frac{f(x_{rq+i})}{f'(x_{rq})} \ (i = 1 \div q - 1) \ .$$

a) Hallar el orden del método iterativo consistente en tomar, como una iteración de éste, q iteraciones del método dado.

b) Comparar su eficiencia con la del método de Newton.

c) Aplicación: Tomar $q = 2$ y comprobar numéricamente el orden encontrado en el cálculo del cero de $f(x) = \cos x - x$.

28. Consideremos el método iterativo dado por la *fórmula de Halley*

$$x_{k+1} = x_k - \frac{2f(x_k)f'(x_k)}{2f'^2(x_k) - f(x_k)f^{(2)}(x_k)} \ .$$

Demostrar que tiene orden de convergencia al menos 3 para ceros simples.

29. Dadas las ecuaciones

$$\text{i) } \operatorname{sen} x = e^{-x} \ , \quad \text{ii) } \tan x = \frac{x}{2} \ ,$$

queremos hallar las soluciones menores que 2π, en valor absoluto. Usar el método de la secante para hallarlas con errores menores que 10^{-6}.

30. a) Usar el método de la secante para calcular el cero de $f(x) = \cos x - x$ con un error menor que 10^{-12}.

    b) Deducir analíticamente cuánto valen

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k e_{k-1}} \ , \quad \lim_{k\to\infty} \frac{e_{k+1}}{e_k^{\frac{1+\sqrt{5}}{2}}} \ .$$

    c) Comprobar numéricamente los límites anteriores.

31. Hallar el orden y la constante asintótica del error del método de la secante, cuando se aplica a la búsqueda de ceros de multiplicidad impar.

32. a) Explicitar el método iterativo consistente en aproximar el término $f'(x_k)$ que aparece en el método de Newton por la derivada en el punto x_k de la parábola que pasa por $(x_k, f(x_k))$, $(x_{k-1}, f(x_{k-1}))$ y $(x_{k-2}, f(x_{k-2}))$.

    b) ¿Cuál es su mínimo orden de convergencia para ceros simples?

    c) Comparar su eficiencia con la del método de Newton, en el caso de buscar ceros simples α con $f^{(2)}(\alpha) \neq 0$.

    d) Aplicarlo a la obtención del cero de $f(x) = \cos x - x$.

33. Diseñamos el siguiente método para obtener ceros simples de la función f: cuando conocemos x_k y x_{k-1}, encontramos el polinomio $q_3(y)$ de interpolación de Hermite a la función inversa local g de f

$$q_3(f(x)) = x \ , \quad q_3'(f(x)) = g'(f(x)) \quad (x = x_k, x_{k-1}) \ ,$$

    después tomamos, como nueva aproximación del cero buscado, $x_{k+1} = q_3(0)$.

    a) Explicitar este método y obtener su orden de convergencia mínimo para la búsqueda de ceros simples.

    b) Generalización: Hallar el orden mínimo (en función de n y m), de un método análogo al anterior, pero haciendo interpolación de Hermite generalizada a g hasta las derivadas de orden n en los puntos $f(x_k)$, ..., $f(x_{k-m})$.

34. Considerando f de clase $\mathcal{C}^3$, queremos construir un método iterativo para hallar ceros simples α de f; para esto, conocidas las aproximaciones x_k y x_{k-1}, encontramos x_{k+1} como el cero más próximo a x_k del polinomio interpolador a f en x_{k-1} y x_k según las condiciones:

$$p_2(x_{k-1}) = f(x_{k-1}) \ , \quad p_2(x_k) = f(x_k) \ , \quad p_2'(x_k) = f'(x_k) \ .$$

    a) Expresar x_{k+1} en función de x_k y x_{k-1}.

b) Hallar una expresión del error $f(x) - p_2(x)$.

c) Denotando $e_k = x_k - \alpha$ y suponiendo que $f^{(3)}(\alpha) \neq 0$, demostrar que

$$\lim_{k\to\infty} \frac{e_{k+1}}{e_k^2 e_{k-1}} = \frac{f^{(3)}(\alpha)}{6f'(\alpha)} \ .$$

d) Hallar su orden y su constante asintótica del error.

35. El método de la secante se puede deducir del de Newton sustituyendo $f'(x_k)$ por la derivada primera en x_k de la recta de interpolación en los puntos de abscisas x_k y x_{k-1}. Tomando como método de partida el de Chebichev y sustituyendo $f^{(2)}(x_k)$ por la derivada segunda en x_k del polinomio interpolador de Hermite en los puntos de abscisas x_k y x_{k-1}, se obtiene un nuevo método iterativo para el cálculo de ceros de funciones.

a) Explicitar la fórmula de iteración de este método.

b) Hallar el orden de convergencia mínimo para ceros simples.

c) Comparar su eficiencia con la del método de la secante y la del método de Chebichev en el caso más general.

36. ¿Cómo varía el cero $\alpha = 2$ de

$$f(x) = e^{6x}(x^3 + 2x^2 - x - 14) + e^{2x}(x^4 - 5x^3 + x^2 + 20)$$

si cambiamos 6 por $6 + \epsilon$ y 20 por $20 + \delta$, con ϵ y δ suficientemente pequeños? Aplicarlo al caso: $\epsilon = 10^{-6}$ y $\delta = 10^{-5}$.

37. Para resolver la ecuación $f(x) = 0$, se aplica el método iterativo

$$x_{k+1} = x_k - g(x_k)f(x_k) \ .$$

a) ¿Qué condiciones ha de cumplir g para que el orden de convergencia hacia un cero simple de f sea lo mayor posible? Comprobar que el método de Newton las verifica.

b) En caso de que sea imposible o muy costosa la evaluación de f' en el método de Newton, podemos aproximar $f'(x_k)$ por

$$\frac{f(x_k + h_k) - f(x_k)}{h_k} \ ,$$

con $h_k > 0$. Si el algoritmo de evaluación de f comete un error relativo acotado por ϵ y $|f^{(2)}(x)| \leq M_2 \ \forall x$, hallar el valor de h_k para el cual la acotación del error cometido en la aproximación de $f'(x_k)$ sea mínima (suponiendo $f(x_k + h_k) \simeq f(x_k)$).

c) Demostrar que, si en la expresión recomendada en a) para $g(x_k)$ se sustituye $f'(x_k)$ por la aproximación hecha en b) con el valor óptimo de h_k encontrado allí, el método iterativo obtenido tiene *convergencia superlineal*; es decir, que

$$\lim_{k\to\infty} \frac{x_{k+1}-\alpha}{x_k-\alpha} = 0 \ .$$

38. El método iterativo $x_{k+1} = g(x_k)$ $(k \geq 0)$, que calcula una solución α de $x = g(x)$, tiene una función de iteración g que se evalúa con un error absoluto acotado por δ.

a) Si $|g'(x)| \leq M < 1$, expresar una cota del error ϵ_k del valor calculado $\mathrm{fl}(x_k)$ de x_k en función del error inicial $\epsilon_0 = |\, x_0 - \alpha \,|$.

b) Estimar el número máximo de iteraciones que convendría realizar al usar el método dado.

39. Aplicamos el método de Newton a un polinomio $p(x)$ con todos los ceros reales con el fin de encontrar su cero α de valor máximo. Si iniciamos las iteraciones con un valor $x_0 > \alpha$, ¿podemos asegurar que es convergente a α? ¿Por qué?

40. Demostrar que, dados un polinomio $p(x)$ de grado n y un valor a tal que $p'(a) \neq 0$, existe como mínimo un cero de $p(x)$ dentro del disco $\mathcal{C}$ definido por

$$\mathcal{C} = \left\{ z \in \mathbb{C} : |z-a| \leq n \left| \frac{p(a)}{p'(a)} \right| \right\} \ .$$

Para la demostración conviene seguir los pasos siguientes:

a) Utilizar el desarrollo de Taylor de $p(x)$ cerca de a.

b) Si α_i $(i = 1 \div n)$ son los ceros de $p(x)$, determinar un polinomio $q(x)$ que tenga los ceros

$$\beta_i = \frac{1}{\alpha_i - a} \quad (i = 1 \div n) \ .$$

c) Ver que, si un polinomio

$$q(x) = b_n x^n + b_{n-1} x^{n-1} + \cdots + b_1 x + b_0$$

tiene los ceros β_i $(i = 1 \div n)$, se cumple que

$$\frac{b_{n-1}}{b_n} = -(\beta_1 + \beta_2 + \cdots + \beta_n) \ .$$

d) Deducir el resultado propuesto inicialmente.

41. La *regla de Descartes* asegura que, si ν es el número de cambios de signo de los coeficientes de un polinomio $p(x)$ y μ es el número de ceros positivos de $p(x)$, entonces $\mu \leq \nu$ y $\nu - \mu$ es par.

    a) Usando la regla de Descartes, demostrar que el polinomio de tercer grado $p(x) = x^3 - x^2 - x - 1$ tiene un único cero real positivo.

    b) Calcularlo por el método de Birge-Vieta.

42. Consideremos la ecuación polinomial $p(x) = 16x^3 + 12x^2 - 8x - 1 = 0$.

    a) Acotar sus raíces con los métodos dados en el apartado 5.2.3.

    b) ¿Cuántas raíces positivas y negativas tiene?

    c) Determinar sus tres raíces.

43. a) Separar las raíces de $p(x) = 36x^4 - 60x^3 - 29x^2 + 9x + 2 = 0$, usando el método de Sturm.

    b) Calcular dichas raíces con 6 cifras decimales correctas por el método de Muller-Traub.

44. Usar el método de Laguerre, partiendo de $x_0 = 1000$, para calcular un cero del polinomio $p(x) = x^4 + 2x^3 + 3x^2 + 4x + 5$.

45. a) Determinar, por el método de Bernoulli, el cero de módulo mínimo del polinomio $p(x) = 32x^3 - 48x^2 + 18x - 1$ con 3 cifras significativas.

    b) Refinarlo hasta 6 cifras significativas usando el método de Birge-Vieta o el de la secante.

46. a) Calcular, con 3 cifras significativas y por el método de Bernoulli, el cero mayor en módulo del polinomio

$$p(x) = x^4 - 4x^3 - 2x^2 - 12x + 9 \ ,$$

    usando las condiciones iniciales $x_k = 0$ $(k = 0 \div 2)$, $x_3 = 1$.

    b) Comprobar que la convergencia en el apartado a) es lenta. Utilizar el método Q-D como método alternativo.

47. Determinar, por el método de Bairstow, los ceros complejos del polinomio $p(x) = x^3 - x - 1$, a partir del factor cuadrático aproximado $x^2 + x$.

48. Calcular, usando el método o la combinación de métodos que se crean más convenientes, todos los ceros de los polinomios siguientes:

    a) $x^3 - 2x - 5$,

    b) $x^3 - 16x^2 - 3$,

    c) $x^4 - 3x^2 + 2x - 1$,

    d) $x^4 + 4x^2 - 3x - 1$,

    e) $x^6 + 5x^3 + 7x + 1$.

49. Determinar el cero de módulo máximo de los polinomios:

    a) $x^3 - 20x^2 - 3x + 18$,

    b) $x^4 - 3x^3 - 60x^2 + 150x + 300$,

    c) $10x^3 - 21x^2 - 40x + 84$.

50. Consideremos el polinomio $p(x) = x^3 - 3x + 2$. ¿Cómo varían sus ceros al perturbarlo sustituyendo el coeficiente -3 por $-3 + 10^{-6}$?

51. Los *puntos 2-periódicos* de una función f son aquellos valores de x que cumplen $x = f(f(x))$. Nótese que los puntos 2-periódicos satisfacen el sistema de ecuaciones

$$\left.\begin{array}{l} x_1 = f(x_2) \\ x_2 = f(x_1) \end{array}\right\} .$$

    Para su resolución, podemos usar un método iterativo de tipo Jacobi

$$x_1^{(k+1)} = f(x_2^{(k)}) \ , \quad x_2^{(k+1)} = f(x_1^{(k)}) \ ,$$

    o de tipo Gauss-Seidel

$$x_1^{(k+1)} = f(x_2^{(k)}) \ , \quad x_2^{(k+1)} = f(x_1^{(k+1)}) \ ,$$

    por analogía con la resolución iterativa de sistemas lineales.

    a) Dar condiciones necesarias de convergencia de los métodos citados, suponiendo conocida la derivada de f en la solución.

    b) ¿Cómo se tendría que variar el método en caso de que no se cumplan tales condiciones?

    c) Aplicación: Encontrar los puntos 2-periódicos de $f(x) = 1.4 \cos x$, partiendo de $x_0 = 1$.

52. Calcular, usando métodos iterativos de tipo Jacobi y Gauss-Seidel a partir de $x_0 = y_0 = 1$, la solución del sistema no lineal

$$\left.\begin{array}{l} x = \operatorname{sen}(x+y) \\ y = \cos(x-y) \end{array}\right\} .$$

53. Dado un sistema no lineal del tipo

$$\left.\begin{array}{l} x_1 = f_1(x_1, x_2, x_3) \\ x_2 = f_2(x_1, x_2, x_3) \\ x_3 = f_3(x_1, x_2, x_3) \end{array}\right\} ,$$

escribimos el sistema como $x = f(x)$, con

$$x = (x_1, x_2, x_3) \ , \quad f = (f_1, f_2, f_3) \ ,$$

y buscamos una solución $\alpha = (\alpha_1, \alpha_2, \alpha_3)$.

a) Obtener condiciones necesarias de convergencia de los métodos iterativos de tipo Jacobi y Gauss-Seidel, en función de los coeficientes de la matriz $A = Df(\alpha)$.

b) Aplicación: Estudiar el caso en que resulta

$$A = \begin{pmatrix} 0.3 & 0.2 & 0.5 \\ -0.3 & 0.4 & -0.2 \\ 0.6 & -0.1 & 0.2 \end{pmatrix} .$$

54. a) Calcular, usando el método de Newton con tres variables, la solución del sistema no lineal

$$\left.\begin{array}{rcl} x &=& 2\cos(-x+y+z) \\ y &=& 2\cos(x-y+z) \\ z &=& 2\cos(x+y-z) \end{array}\right\} ,$$

cerca de $x = y = z = 1$.

b) Comprobar la convergencia cuadrática del método, usando la norma $\| \ \|_\infty$.

BIBLIOGRAFÍA

[Apo57] T.M. Apostol. *Mathematical analysis.* Addison-Wesley, Reading, Mass., 1957. En castellano: Reverté, Barcelona, 1982.

[AS65] M. Abramowitz and I.A. Stegun, editors. *Handbook of mathematical functions.* Dover, N.Y., 1965.

[CdB72] S.D. Conte and C. de Boor. *Elementary numerical analysis, an algorithmic approach.* McGraw-Hill, N.Y., 1972. En castellano: 1974.

[Che66] E.W. Cheney. *Introduction to approximation theory.* McGraw-Hill, N.Y., 1966.

[Cia82] P.G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation.* Masson, Paris, 1982.

[Cra46] H. Cramér. *Mathematical methods of statistics.* Princeton Univ. Press, Princeton, N.J., 1946.

[Dav75] P.J. Davis. *Interpolation and approximation.* Dover, N.Y., 1975.

[DB74] G. Dahlquist and Å. Björck. *Numerical Methods.* Prentice Hall, Englewood Cliffs, N.J., 1974.

[Die68] J. Dieudonné. *Calcul infinitéssimal.* Hermann, Paris, 1968. En castellano: Omega, Barcelona, 1971.

[DM73] B. Demidovitch and I. Maron. *Eléments de calcul numérique.* Mir, Moscou, 1973. En castellano: Paraninfo, Madrid, 1977.

[DR67] P.J. Davis and P. Rabinowitz. *Numerical integration.* Blaisdell, London, 1967.

[Dur61] E. Durand. *Solutions numériques des équations algébriques.* Vol. 1,2. Masson, Paris, 1960, 1961.

[Fel50] W. Feller. *An introduction to the probability theory and its applications.* Wiley, N.Y., 1950.

[Fro69] C.E. Froberg. *Introduction to numerical analysis.* Addison-Wesley, Reading, Mass., 1969. En castellano: Vicens Vives, 1977.

[Gas66] N. Gastinel. *Analyse numérique linéaire.* Hermann, Paris, 1966. En castellano: Reverté, 1975.

[Ham70] S.J. Hammarling. *Latent roots and latent vectors.* Adam Hilger, London, 1970.

[Ham73] R.W. Hamming. *Numerical methods for scientists and engineers.* McGraw-Hill, N.Y., 1973.

[Hen64] P. Henrici. *Elements of numerical analysis.* Wiley, N.Y., 1964. En castellano: Trillas, México, 1968.

[Hen74] P. Henrici. *Applied and computational complex analysis.* Wiley, N.Y., 1974.

[Hil74] F.B. Hildebrand. *Introduction to numerical analysis.* McGraw-Hill, N.Y., 2nd. edition, 1974.

[Hir74] S. Hirsch, M.W.and Smale. *Differential equations, dynamical systems, and linear algebra.* Academic Press, New York, 1974. En castellano: Alianza Universidad, Madrid, 1983.

[Hou53] A.S. Householder. *Principles of numerical analysis.* McGraw-Hill, N.Y., 1953.

[Hou64] A.S. Householder. *The theory of matrices in numerical analysis.* Blaisdell, N.Y., 1964.

[IK66] E. Isaacson and H.B. Keller. *Analysis of numerical methods.* Wiley, N.Y., 1966.

[Jac77] D.A.H. Jacobs, editor. *The state of the art in numerical analysis.* Academic Press, N.Y., 1977.

[Knu69] D.E. Knuth. *The art of computer programming.* Addison-Wesley, Reading, Mass., 1969.

[LH74] C.L. Lawson and R.J. Hanson. *Solving least squares problems.* Prentice-Hall, Englewood Cliffs, N.J., 1974.

[Moo66] R.E. Moore. *Interval analysis.* Prentice-Hall, Englewood Cliffs, N.J., 1966.

[OR70] J. Ortega and W. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* Academic Press, N.Y., 1970.

[Ost66] A.M. Ostrowski. *Solution of equations and systems of equations.* Academic Press, N.Y., 2nd. edition, 1966.

[Par80] B.N. Parlett. *The symmetric eigenvalue problem.* Prentice-Hall, Englewood Cliffs, N.J., 1980.

[Que71] M. Queysanne. *Algebre.* Librairie Armand Colin, Paris, 1971. En castellano: Vicens-Vives, Barcelona, 1973.

[Ral65] A. Ralston. *A first course in numerical analysis.* McGraw-Hill, N.Y., 1965. En castellano: Limusa-Wiley, México, 1970.

[Ric81] J.R. Rice. *Matrix computation and mathematical software.* McGraw-Hill, N.Y., 1981.

[RR78] A. Ralston and P. Rabinowitz. *A first course in numerical analysis.* McGraw-Hill, N.Y., 2nd. edition, 1978.

[RW67] A. Ralston and H.S. Wilf, editors. *Mathematical methods for digital computers.* Wiley, N.Y., 1967.

[SB80] J. Stoer and R. Bulirsch. *Introduction to numerical analysis.* Springer, Berlin, 1980.

[Sch67] F. Scheid. *Numerical analysis, including 775 solved problems.* Schaum, N.Y., 1967. En castellano: 1972.

[SS66] A.H. Stroud and D. Secrest. *Gaussian quadrature formulas.* Prentice-Hall, Englewood Cliffs, N.J., 1966.

[Ste73] G.W. Stewart. *Introduction to matrix computations.* Academic Press, N.Y., 1973.

[Str71] A.H. Stroud. *Approximate calculation of multiple integrals.* Prentice-Hall, Englewood Cliffs, N.J., 1971.

[Sze59] G. Szego. *Orthogonal polynomials.* AMS, N.Y., 1959.

[Tra64] J.F. Traub. *Iterative methods for the solution of equations.* Prentice-Hall, Englewood Cliffs, N.J., 1964.

[Wil64] J.H. Wilkinson. *Rounding errors in algebraic processes.* Prentice-Hall, Englewood Cliffs, N.J., 1964.

[Wil65] J.H. Wilkinson. *The algebraic eigenvalue problem.* Clarendon, Oxford, 1965.

[WR71] J.H. Wilkinson and C. Reinsch. *Handbook for automatic computation. Vol. 2: Linear algebra.* Springer, Berlin, 1971.

[YG72] D.M. Young and R.T. Gregory. *A survey of numerical mathematics.* Addison-Wesley, Reading, Mass., 1972.

[You71] D.M. Young. *Iterative solution of large linear systems.* Academic Press, N.Y., 1971.

ÍNDICE DE SÍMBOLOS Y ALFABÉTICO