UNED

Introducción al Cálculo Numérico

Carlos Moreno González

Introducción al Cálculo Numérico

CARLOS MORENO GONZÁLEZ

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

INTRODUCCIÓN AL CÁLCULO NUMÉRICO

www.uned.es/publicaciones

Todas nuestras publicaciones han sido sometidas a un sistema de evaluación antes de ser editadas

ISBN electrónico: 978-84-362-6634-4

Edición digital: abril de 2014

Prólogo

Modern numerical analysis can be credibly said to begin with the 1947 paper by John von Neumann and Herman Goldstine, "Numerical Inverting of Matrices of High Order"(Bulletin of the AMS, Nov. 1947). It is one of the first papers to study rounding error and include discussion of what today is called scientific computing. Although numerical analysis has a longer and richer history, "modern"numerical analysis, as used here, is characterized by the synergy of the programmable electronic computer, mathematical analysis, and the opportunity and need to solve large and complex problems in applications. The need for advances in applications, such as ballistics prediction, neutron transport, and nonsteady, multidimensional fluid dynamics drove the development of the computer and depended strongly on advances in numerical analysis and mathematical modeling.
SIAM History of Numerical Analysis and Scientific Computing Project 2007.

Las matemáticas suministran un lenguaje que se utiliza en las ciencias e ingenierías para describir con rigor modelos que pretenden representar fenómenos reales. En ellos, se estudian aspectos cualitativos de esos fenómenos y se plantean problemas cuya resolución conduce en último término al desarrollo de tecnologías aplicadas.

Las dificultades que surgen en el estudio de estos problemas, ha estimulado el desarrollo de las matemáticas generando la necesidad de teorías abstractas y contribuyendo de este modo a construir el pensamiento matemático actual. Pero, junto a cada nuevo concepto que permite profundizar en la estructura del conocimiento matemático, surge la necesidad de desarrollar técnicas que permitan utilizar esos conceptos de un modo cuantitativo preciso.

En las matemáticas de la Edad Antigua, los problemas eran de naturaleza aritmética o geométrica. Entonces, una vez comprendido el concepto

de longitud, área y volumen, los métodos *exhaustivos* permitían resolverlos utilizando unidades de medida cada vez más fraccionadas. Posteriormente, el Cálculo Diferencial de Leibniz permitió precisar muchos de estos conceptos y simultáneamente desarrolló los métodos analíticos que permitieron su evaluación exacta. No obstante, las ecuaciones que aparecen en los modelos comienzan a ser más complicadas ya que junto a las ecuaciones numéricas (es decir, aquellas cuyas incógnitas son valores numéricos), empiezan a considerarse ecuaciones cuyas incógnitas son funciones y en ellas aparecen operadores distintos de los algebraicos tales como los integrales y diferenciales.

En la primera parte del siglo XX es ya muy manifiesta la insuficiencia del cálculo simbólico para resolver ecuaciones de la complejidad con la que se presentaban en los modelos de las ciencias e ingenierías. Durante la primera mitad de este siglo permanece un sentimiento de incapacidad por la imposibilidad material de llevar a cabo los cálculos por procedimientos numéricos aproximados. Este sentimiento se mitiga en la segunda mitad del siglo con el desarrollo exponencial de las capacidades de calcular automáticas. Este hecho condiciona de un modo radical el modo de calcular y las técnicas comienzan a adaptarse a este nuevo medio de cálculo.

Los medios de cálculo electrónico permiten realizar una elevada cantidad de operaciones algebraicas en periodos muy reducidos de tiempo. En ese momento surge la necesidad de diseñar algoritmos capaces de aproximar las soluciones a los problemas, mediante operaciones elementales. Los sistemas de ecuaciones numéricas lineales, aun cuando el número de variables fuese muy elevado, podían ya directamente ser resueltos por estas técnicas. También, las ecuaciones diferenciales, que tenían funciones como incógnitas y que involucraban operadores diferenciales, podían ser resueltas eficientemente, si bien necesitaban de un paso intermedio que permitiera convertirlas en ecuaciones numéricas.

El modo más común de pasar de lo continuo a lo discreto es a través de los métodos de aproximación e interpolación que permiten aproximar funciones mediante polinomios u otras funciones discretas. Es bastante razonable pensar que en casi todo el Análisis Numérico, más tarde o más temprano, cualquier análisis de un método constructivo acaba siempre convirtiéndose en una cuestión relacionada con la teoría de la aproximación de funciones. No es de extrañar que los nombres de los que han fundamentado el Análisis Matemático clásico, como Newton, Fourier, Gauss, Legendre, Euler, Chebyshev y otros, se utilicen para denominar métodos y procedimientos del cálculo numérico.

Los objetivos primordiales que pretende cubrir este texto son:

1. La resolución de sistemas de ecuaciones lineales numéricas.

2. El estudio de las técnicas de aproximación e interpolación y su aplicación a la integración aproximada.

3. La resolución de sistemas de ecuaciones no-lineales numéricas.

4. La resolución de problemas de valor inicial y de contorno para ecuaciones diferenciales.

Los dos primeros apartados corresponden al curso *Análisis numérico matricial e interpolación* y los dos siguientes, al curso *Resolución numérica de ecuaciones*, ambos del grado de Matemáticas de la Universidad Nacional de Educación a Distancia (UNED). Deseo expresar mi agradecimiento a aquellas personas que han contribuido con sus observaciones a la mejora de este texto, y particularmente a Alfredo Cano y Juan Bódalo, alumnos del Master de Matemáticas Avanzadas de la UNED, por su contribución a eliminar algunos errores que contenía en sus primeras versiones.

Carlos Moreno
Madrid, mayo de 2011

Índice general

Parte I: Análisis numérico matricial e interpolación

CAPÍTULO 1

Estabilidad y errores en el cálculo numérico

1.1 Introducción

El saber contar y realizar operaciones aritméticas es el resultado de una larga evolución en el pensamiento humano. En todo este proceso, el hombre ha tratado de ayudarse en la realización de los cálculos mediante artilugios que simplificaran su labor. Primero fueron simples guijarros (calculi en latín), después fueron instrumentos mecánicos simples como el ábaco, instrumentos mecánicos articulados como las máquinas de Pascal o de Schickard y finalmente los computadores electrónicos. También, desde la programación en cilindros rotatorios o tarjetas perforadas hasta los desarrollos actuales en *software* basados en los lenguajes modernos de programación, se ha recorrido un largo camino. Todo este conocimiento y tecnología permite actualmente realizar cálculos de gran complejidad.

Sin embargo, un computador produce resultados en respuesta a cálculos programados que posiblemente difieren ligeramente de los valores exactos esperados. Ello es consecuencia de que trabajan con una aritmética discreta que no coincide plenamente con la aritmética exacta de los números enteros o reales. Todavía el ser humano no alcanza a realizar por medios físicos, todos los cálculos que su mente puede concebir, ni siguiera a representar en la memoria de un ordenador más que un subconjunto finito del conjunto de todos los números que puede manejar.

El propósito de este capítulo es estudiar las representaciones más comunes de números en un computador y las aritméticas que usan en sus cálculos. Una vez establecido que los errores respecto a la aritmética real pueden es-

tar presentes, se introduce el concepto de estabilidad numérica que permite seleccionar los algoritmos que son válidos para esta clase de cálculos.

1.2 Representación de números en un computador

La diferencia (en valor absoluto) entre el valor exacto x y el valor obtenido en un cálculo $\bar{x}$, determina el error absoluto cometido $E(x) = |x - \bar{x}|$. Una indicación más precisa del error cometido en el cálculo la da el error relativo, que se define por la siguiente expresión

$$E_R(x) = \frac{|x - \bar{x}|}{|x|}.$$

Si un determinado cálculo produce como resultado $\bar{x} = 1,2345 \times 10^{12}$ siendo el valor exacto esperado $x = 1.2331 \times 10^{12}$, el error absoluto sería $E(x) = 1.4 \times 10^9$. Este error parece muy alto si no se considera el orden de magnitud de los números con los que se realiza el cálculo. Por el contrario, el error relativo

$$E_R(x) = \frac{1.4 \times 10^9}{1.2331 \times 10^{12}} = 1.1353 \times 10^{-3}$$

da una idea más proporcionada de la imprecisión cometida.

Los computadores representan los números en sus memorias asignándoles una cantidad fija de posiciones a las que puede acceder. Esta cantidad puede variar de un computador a otro, de acuerdo con el estándar de representación que se haya atribuido a ese tipo de número. Es obvio que el modo de representación interna de un número en una memoria no tiene que coincidir necesariamente con el modo en que este número se muestra en una pantalla u otro dispositivo terminal.

Con el fin de comprender cómo usan los ordenadores la aritmética discreta, se recuerda el concepto de representación posicional de los números enteros y reales. Se considera el conjunto de números $\mathcal{M}$ que pueden ser representados en la forma

$$\pm 0.d_1 d_2 \cdots d_n \times b^e$$

donde $0 \leq d_i < b$ para $i = 1, 2, \cdots, n$. El número entero positivo b es fijo y corresponde a la base de la representación. El entero e corresponde al exponente y puede variar en un determinado rango. Finalmente, el entero positivo fijo n controla la precisión de la representación. La parte fraccionaria es frecuentemente llamada mantisa. El valor numérico real asignado a esta representación es

$$x = \pm(d_1 b^{-1} + d_2 b^{-2} + \cdots + d_n b^{-n}) \times b^e.$$

Por ejemplo, a la representación 0.1011×2^3 en base 2 le corresponde el valor decimal

$$x = 4 + 1 + \frac{1}{2} = 5.5$$

Idealmente, se puede ampliar el sistema admitiendo que n sea infinito y de este modo, se podrían incluir todos los números reales en esta representación. Obviamente, para un sistema de representación posicional que se pretenda poner en práctica en un computador, se requiere que n sea un número entero fijo y que se ajuste a la capacidad física de su memoria accesible. Por esta razón, todo sistema de representación que utilice un computador, tendrá un conjunto finito de números-máquina.

Habitualmente, para números enteros se emplea una representación en la que el exponente e es fijo e igual a n. De este modo, a la representación $\pm d_1 d_2 \cdots d_n$ le corresponde el valor entero

$$x = \pm(d_1 b^{n-1} + d_2 b^{n-2} + \cdots + d_n).$$

▪ Ejemplo **1** En un sistema binario ($b = 2$), con un número fijo de dígitos $n = 3$, el conjunto de números (que no son negativos) que pueden ser representados de este modo, son

Representación	000	001	010	011	100	101	110	111
Valor real	0	1	2	3	4	5	6	7

Un aspecto relevante de este sistema de representación es que puede producirse un desbordamiento en las operaciones aritméticas debido a que el conjunto es acotado. Por ejemplo, la suma de 3 y 5 produce un número que no puede ser representado en este sistema. ◇

Si el exponente e es un entero variable en un determinado rango, el conjunto de números que pueden ser representados de este modo (llamado de punto flotante) es más amplio, como muestra el siguiente ejemplo

▪ Ejemplo **2** En un sistema binario $b = 2$, si el número de dígitos es $n = 2$, el conjunto de números (que no son negativos) que pueden ser representados en punto flotante con el rango $-2 \leq e \leq 2$, es el siguiente

Representación	$.00 \times 2^e$	$.01 \times 2^e$	$.10 \times 2^e$	$.11 \times 2^e$
Valor	0	$\frac{1}{4} \times 2^e$	$\frac{1}{2} \times 2^e$	$\frac{3}{4} \times 2^e$

En definitiva, el conjunto de números que no son negativos y pueden ser representados de este modo son

$$\{0, \frac{1}{16}, \frac{1}{8}, \frac{3}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{3}{4}, 1, \frac{3}{2}, 2, 3\}.$$

Los aspectos más relevantes de este sistema de representación son los siguientes:

- Los números representados no son equidistantes y se observa una mayor densidad en las proximidades de 0.
- Un número puede tener representaciones distintas. Por ejemplo, el número $\frac{1}{8}$ se puede representar por $.01 \times 2^{-1}$ o por $.10 \times 2^{-2}$.
- Se puede producir un desbordamiento en las operaciones aritméticas debido a que el conjunto es acotado. Por ejemplo, la suma de 1 y 3 produce un número que no puede ser representado en este sistema.
- El resultado de algunas operaciones aritméticas reales con números de este conjunto está fuera del conjunto aunque no se haya producido un desbordamiento del rango. Por ejemplo, la suma real de 2 y $\frac{1}{8}$ no pertenece al conjunto. $\diamond$

Una representación en punto flotante está normalizada si el primer dígito en la parte fraccionaria es necesariamente distinto de 0. De este modo, se evita que un número pueda tener representaciones distintas en un mismo sistema. Actualmente, la representación que usan la mayoría de computadores es la llamada IEEE Standard 754 en punto flotante. Está basada en los tres elementos mencionados anteriormente, el signo, la mantisa y el exponente.

Si la representación es binaria, el primer dígito es 1 (dígito principal) y en la mayoría de las puestas en práctica de este estándar, no es almacenado en memoria (dígito principal implícito). El estándar de representación en punto flotante de IEEE que corresponde a lo que se conoce como precisión simple, utiliza 4 bytes de memoria (32 bits) de los cuales 8 bits son para el exponente E, 23 para la parte fraccionaria F y uno para el signo S. Cada bit almacena un 0 o un 1. El valor asignado a una representación es

$$(-1)^S \times 2^{E-127} \times 1.F$$

El campo del exponente necesita representar a la vez exponentes positivos y negativos. Para conseguirlo se añade un sesgo al valor positivo almacenado para lograr el exponente deseado. En el estándar IEEE 754 para precisión

simple, este valor es 127. De este modo, un valor almacenado E representa un valor real $E - 127$. Para doble precisión el campo del exponente tiene 11 bits y un sesgo de 1023. El bit del signo es 0 para positivos y 1 para negativos.

Por ejemplo, si en memoria está la siguiente información almacenada de acuerdo al estándar IEEE 754 con dígito principal implícito

$$\underbrace{1}_{-} \quad \underbrace{01010011}_{\text{exponente}} \quad \underbrace{10011110001010000101000}_{\text{mantisa}}$$

el número representado es

$$-\left(1+\frac{1}{2}+\frac{1}{2^4}+\frac{1}{2^5}+\frac{1}{2^6}+\frac{1}{2^7}+\frac{1}{2^{11}}+\frac{1}{2^{13}}+\frac{1}{2^{18}}+\frac{1}{2^{20}}\right)\times 2^{83-127}$$

ya que $01010011_2 = 83$.

1.3 Aritmética en un sistema de representación finito

En general, cuando se utiliza un sistema discreto $\mathcal{M}$, inevitablemente aparecen errores al introducir los números reales convirtiéndolos a números-máquina pero también como el resultado de una operación entre números-máquinas que en general no coincide con el resultado que se obtendría en la aritmética real. Conceptualmente, los números reales pueden aproximarse por números del sistema discreto de dos modos ligeramente distintos que se conocen como truncamiento y redondeo.

La puesta en práctica de métodos de truncamiento está basada en la siguiente idea: Un numero real puede ser aproximado por un número de punto flotante con parte fraccionaria infinita en la base b

$$\pm .d_1 d_2 \cdots \times b^e.$$

Consecuentemente, si se trunca esta serie con n dígitos se obtiene una aproximación en el sistema discreto disponible en el computador. Sin embargo, es más usado el llamado método de redondeo que consiste en aproximar un número real positivo por el número-máquina más próximo. Este procedimiento sería el que nos proporcionaría mayor precisión. Para precisar estas ideas se representa la función parte entera de un número real por un corchete [ ] y se define para un número real positivo representado por $x = 0.m \times b^e$ donde m es una cifra con posiblemente una infinidad de dígitos, las siguientes aproximaciones por números máquina de n dígitos

Truncamiento	Redondeo
$\mathcal{T}(x) = [b^n \times 0.m] \times b^{e-n}$	$\mathcal{R}(x) = [b^n \times 0.m + 0.5] \times b^{e-n}$
Ejemplo (3 dígitos, base 10):	Ejemplo(3 dígitos, base 10):
$\mathcal{T}(1/3) = 10^{-3}[10^3 \times 0.333...] = 0.333$	$\mathcal{R}(1/3) = 10^{-3}[10^3 \times 0.333... + 0.5] = 0.333$
$\mathcal{T}(2/3) = 10^{-3}[10^3 \times 0.666...] = 0.666$	$\mathcal{R}(2/3) = 10^{-3}[10^3 \times 0.666... + 0.5] = 0.667$

En lo que sigue se centrará la atención en el método de redondeo. Si $x < 0$ entonces se define

$$\mathcal{R}(x) = -\mathcal{R}(-x).$$

De la definición de redondeo se deduce que

$$|x - \mathcal{R}(x)| \leq \frac{1}{2} b^{e-n}.$$

El número $\frac{1}{2}b^{e-n}$ se conoce como unidad de redondeo o precisión de la máquina.

Una aritmética para un sistema de representación finita estaría disponible si se consigue asignar como resultado de una operación el que corresponde al redondeo de la operación aritmética exacta.

▬ Ejercicio 1 *Sea $x > 0$ un número que se representa exacto en una aritmética finita de punto flotante. Analizar el comportamiento del cociente*

$$\frac{\operatorname{sen}(x+h) - \operatorname{sen} x}{h}$$

cuando se consideran valores de h próximos a 0 y los cálculos se realizan en una aritmética de punto flotante.

Solución: Para h suficientemente pequeño el resultado de redondear $x + h$ coincide con x. Por ello, el numerador es nulo y el valor asignado al cociente es 0. ◇

▬ Ejercicio 2 *Si se usa una aritmética con redondeo, en un sistema de representación decimal y de 3 dígitos de precisión, para realizar la siguiente la operación*

$$\frac{a \times b - c}{b + 2 \times c}, \quad a = 1.34,\ b = 0.712,\ c = -0.355,$$

determinar el error cometido en relación con la aritmética real.

Solución: En la representación finita los números dados son

$$a = 0.134 \times 10^1,\ b = 0,712 \times 10^0,\ c = -0.355 \times 10^0.$$

Consecuentemente, el resultado de las operaciones es el siguiente:

- $a \times b = \mathcal{R}(0.95408) = 0.954 \times 10^0$.
- $a \times b - c = \mathcal{R}(0.954 + 0.355) = 0.131 \times 10^1$.
- $2 \times c = -0.710$.
- $b + 2 \times c = 0.2 \times 10^{-2}$.
- $\frac{a \times b + c}{b + 2 \times c} = 0.655 \times 10^3$.

El resultado exacto es 0.65454×10^3 y consecuentemente el error relativo es

$$E_R = \frac{|655 - 654.54|}{654.54} = 0.00070278\ . \qquad \diamond$$

1.4 Estabilidad en los algoritmos numéricos

Una vez que un entorno de cálculo dispone de una aritmética finita que le permite realizar las operaciones elementales dentro del rango numérico que puede representar, se hace necesario complementarlo con otras funciones elementales que permitan realizar cálculos más complejos. La dificultad está en que estas funciones implican un número elevado de operaciones elementales y puesto que los errores de redondeo respecto a la aritmética exacta son inevitables, el resultado final puede estar muy deteriorado en relación con el exacto.

▪ Ejemplo **3** Si en un entorno que usa el estándar IEEE Double Precision, se realiza el cálculo de e^{-10} mediante el truncamiento de la serie de potencias

$$e^x \approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots \frac{x^{30}}{30!}$$

el resultado aproximado es $9.703415796025504 \times 10^{-4}$. Desafortunadamente, este valor se aleja del valor exacto $4.539992976248485 \times 10^{-5}$. El error de truncamiento de la serie está dado

$$E(-10) = \frac{e^\xi}{31!} < \frac{1}{31!}$$

para algún valor $-10 < \xi < 0$. Consecuentemente, si el cálculo se realizase en una aritmética exacta, el truncamiento de la serie no podría producir un error de ese tamaño. El motivo de la inexactitud está en que los pequeños errores causados por la aritmética finita, han sido magnificados a lo largo del cálculo.

Por otra parte, si se calcula $e^{-10} = \frac{1}{e^{10}}$ usando sólo el desarrollo en serie del denominador, el resultado es $4.539993338712231 \times 10^{-5}$ que es muy próximo al valor exacto. $\diamond$

La razón por la que un procedimiento produce resultados más imprecisos que otro se atribuye a su inestabilidad; Es decir, un error en alguna etapa del procedimiento se propaga de un modo creciente en las siguientes. La necesidad de utilizar algoritmos estables para evaluar una función es obligada por la inevitable presencia de errores de redondeo debidos al uso de aritméticas finitas en los entornos de cálculo automático.

1.5 Ejercicios

▪ Ejemplo **4** En este ejemplo se muestra cómo convertir el número decimal $x = 324.65$ a base 2. Se separa la parte entera 324 de la parte decimal 0.65. Para la parte entera se procede de modo sigue:

$$\begin{aligned} 324 &= 162 \times 2 + 0 \\ 162 &= 81 \times 2 + 0 \\ 81 &= 40 \times 2 + 1 \\ 40 &= 20 \times 2 + 0 \\ 20 &= 10 \times 2 + 0 \\ 10 &= 5 \times 2 + 0 \\ 5 &= 2 \times 2 + 1 \\ 2 &= 1 \times 2 + 0 \end{aligned}$$

En consecuencia se tiene que $324 = 101000100_2$. El dígito más significativo siempre es 1 y los siguientes son los restos de la divisiones por 2 comenzando desde la última división.

En la parte decimal, se procede como sigue

$$\begin{array}{rcl} 0.65 & \times & 2 = 1.3 \\ 0.3 & \times & 2 = 0.6 \\ 0.6 & \times & 2 = 1.2 \\ 0.2 & \times & 2 = 0.4 \\ 0.4 & \times & 2 = 0.8 \\ 0.8 & \times & 2 = 1.6 \\ 0.6 & \times & 2 = 1.2 \\ \dots & \dots & \dots \end{array}$$

En consecuencia se tiene que $0.65 = 0.1010011\cdots_2$. Los dígitos son las partes enteras de las sucesivas multiplicaciones por 2 comenzando por la primera. ◇

▬ EJERCICIO 3 *Hallar la representación en punto flotante (sin dígito principal implícito) de*

1. $2/7$ *en un sistema* $b = 10$, $n = 6$.
2. 327 *en un sistema* $b = 2$, $n = 6$.
3. $-\frac{4}{3}$ *en un sistema* $b = 2$, $n = 6$.

por redondeo.

Solución:

1. $\mathcal{R}(\frac{2}{7}) = 0.285714 \times 10^0$.
2. Puesto que $327 = 2^8 + 2^6 + 2^2 + 2 + 1 = 0.101000111_2 \times 2^9$ entonces se tiene que

$$\mathcal{R}(327) = [2^6 \times 0.101000111_2 + 0.5] \times 2^3 = 0.101001_2 \times 2^9 = 328.$$

3. Puesto que $\frac{4}{3} = 0.10101010...{}_2 \times 2$ entonces se tiene que

$$\mathcal{R}(\frac{4}{3}) = [2^6 \times 0.1010101010_2 + 0.5] \times 2^{-5} = 0.101011_2 \times 2.$$

Consecuentemente $-\frac{4}{3} = -0.101011_2 \times 2$ ◇.

▬ EJERCICIO 4 *Hallar el resultado con la aritmética finita de un sistema de representación $b = 2$, $n = 6$ y $-3 \leq e \leq 3$, sin dígito principal implícito, de las siguientes operaciones*

1. $\frac{1}{3} + \frac{7}{13}$
2. $47 + 347$

Solución:

1. Puesto que $\mathcal{R}(1/3) = 0.101011_2 \times 2^{-1}$ y $\mathcal{R}(7/13) = 0.100010_2$ entonces

$$\begin{aligned} \mathcal{R}(\mathcal{R}(1/3) + \mathcal{R}(7/13)) &= \mathcal{R}(0.1101111_2) = 2^{-6}[110111.1_2 + 0.1_2] \\ &= 2^{-6} \times 111000_2 = 0.111000_2 \ . \end{aligned}$$

2. El valor real máximo que se puede representar en este sistema es

$$0.111111_2 \times 2^3 = 4 + 2 + 1 + 0.5 + 0.25 + 0.125 = 7.875.$$

Consecuentemente, las representaciones de 47 y 347 producen un desbordamiento (overflow) que habitualmente es notificado . ◇

▬ EJERCICIO 5 *Se pretende aproximar el valor de $e^{0.1}$ usando la serie de potencias*

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots$$

en una aritmética basada en el redondeo en base 10 y 7 dígitos de precisión. Si se usan cinco sumandos de la serie de potencias, calcular el error relativo que se comete al realizar la aproximación en la aritmética finita respecto de la que corresponde a la aritmética exacta.

Solución: Los cálculos exactos de los cinco primeros términos de la serie dan las siguientes resultados

n	1	2	3	4	5	$p_5(0.1)$
$\frac{x^n}{n!}$	1	0.1	0.005	$0,0001\hat{6}$	$0.0000041\hat{6}$	$1.1051708\hat{3}$
$\mathcal{R}\left(\frac{x^n}{n!}\right)$	1	0.1	0.005	0.0001667	0.0000042	1.105171

donde p_5 representan los valores de la serie truncada, calculados en aritmética exacta y finita. El sombrero representa el decimal periódico en la expresión decimal.

$$\mathcal{R}(x) = [10^7 \times 0.1105171 + 0.5] \times b^{1-6} = 1.105171$$

En este caso, el resultado no depende del orden en el que se realicen las sumas. El error relativo es

$$E_R(0.1) = \frac{1.105171 - 1.10517083\hat{3}}{1.10517083\hat{3}}. \qquad \diamond$$

CAPÍTULO 2

Sistemas de ecuaciones numéricas lineales

2.1 Introducción

La resolución eficiente de sistemas de ecuaciones numéricas lineales es una necesidad presente en muchos problemas del cálculo científico. El estudio de estructuras elásticas discretas, circuitos eléctricos complejos u otros modelos discretos de la ciencia y de la ingeniería pueden conducir directamente a sistemas lineales de ecuaciones numéricas de gran dimensión. Pero, también la necesidad de resolver grandes sistemas de ecuaciones lineales puede surgir como una etapa intermedia en un procedimiento más amplio como puede ser la resolución numérica de ecuaciones diferenciales o de problemas formulados en modelos estadísticos que usan cantidades importantes de datos. Se debe tener presente que las ecuaciones diferenciales forman parte ineludible de muchos de los modelos matemáticos que se construyen no sólo en Física e Ingeniería sino también en Biología, en Economía y en otras Ciencias Sociales. Desafortunadamente, algunos de los métodos elementales de resolución que se introducen en los cursos básicos de Álgebra Lineal pueden resultar ineficaces cuando el número de incógnitas y ecuaciones crece considerablemente. El objetivo de este capítulo es el estudio de métodos de resolución de sistemas lineales que sean eficientes en problemas reales con media o alta dimensión.

Una breve introducción en las primeras secciones de algunos conceptos básicos del análisis matricial, permitirá posteriormente analizar la estabilidad y la convergencia de los métodos considerados en este capítulo.

Aunque la elección del método más adecuado para resolver un sistema lineal depende fundamentalmente de las propiedades de la matriz de coeficien-

tes, para un sistema lineal sin características especiales, uno de los métodos más eficientes es el clásico método de eliminación de Gauss o alguna de sus variantes. Estos métodos pertenecen a la clase de los métodos directos que conducen a la solución exacta en un número finito de operaciones aunque ello sólo desde un punto de vista teórico, debido a la presencia de errores de redondeo cuando los cálculos se realizan en un ordenador. En otra categoría de métodos, los iterativos permiten generar una sucesión de soluciones aproximadas que convergen a la solución exacta. En este capítulo se estudian métodos de ambas categorías analizando sus propiedades y el modo de ponerlos en práctica.

2.2 Norma matricial subordinada a una norma vectorial

La norma euclídea de un vector (real o complejo) $\mathbf{x}$ de n componentes, cuya componente $i-$ésima es x_i, está definida por la siguiente expresión

$$\| \mathbf{x} \| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}.$$

Más general es la norma p que está definida por

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

para cualquier número positivo p. En el límite, la norma ∞ está definida por

$$\| \mathbf{x} \|_\infty = \max_{i=1,\cdots,n} |x_i|.$$

Una matriz puede representar a una aplicación lineal entre dos espacios euclídeos. Por ello, existen normas matriciales que están definidas en base a esta representación y cuya definición está relacionada con las normas euclídeas en ambos espacios. Así, una norma matricial subordinada a una norma vectorial está definida por

$$\| \mathrm{A} \| = \max_{\|\mathbf{x}\|=1} \| \mathrm{A}\mathbf{x} \|$$

para cualquier matriz A. Obviamente, si la matriz A tiene n filas y m columnas, dos normas vectoriales entrarían en juego. Aunque muchos de los conceptos y resultados que se exponen en este capítulo son extensibles a esta situación, por limitación de objetivos en este texto, se supondrá siempre,

salvo mención en sentido contrario, que la matriz A es cuadrada y la elección de norma para el dominio e imagen es la misma.

En los casos más simples se pueden obtener expresiones que permiten determinar el valor de una norma subordinada de una matriz en términos de sus coeficientes. Algunas de estas situaciones son las siguientes (véase [8, Quarteroni, Sacco, Saleri], páginas 23-24)

- Norma subordinada 1

$$\| \mathrm{A} \|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|,$$

si se considera la norma vectorial

$$\| \mathbf{x} \|_1 = \sum_{j=1}^{n} |x_j|$$

y a_{ij} representa el coeficiente de la matriz A correspondiente a la fila i y la columna j.

- Norma subordinada ∞

$$\| \mathrm{A} \|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|,$$

si se considera la norma vectorial

$$\| \mathbf{x} \|_\infty = \max_{1 \leq j \leq n} |x_j|.$$

- Norma subordinada 2

$$\| \mathrm{A} \|_2 = \rho(\mathrm{A}^t \mathrm{A})^{\frac{1}{2}},$$

si se considera la norma vectorial

$$\| \mathbf{x} \|_2 = \sqrt{\sum_{1}^{n} |x_j|^2},$$

y A^t representa la matriz traspuesta de A y ρ el radio espectral de la matriz (es decir, el máximo de los módulos de los autovalores). En particular, si A es simétrica entonces se cumple que $\| \mathrm{A} \|_2 = \rho(A)$.

La norma subordinada euclídea no coincide con la norma matricial, asociada al producto escalar matricial

$$\mathrm{A} : \mathrm{B} = \mathrm{tr}\ \mathrm{A}^{\mathrm{t}}\mathrm{B},$$

donde tr representa el operador que asocia a una matriz su traza, y que se conoce como norma de Frobenius

$$\| \mathrm{A} \|_F = (\mathrm{A} : \mathrm{A})^{\frac{1}{2}} = \left(\sum_{i,j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}}.$$

Mientras que la norma subordinada euclídea está relacionada con el radio espectral del producto $\mathrm{A}^t\mathrm{A}$, la norma de Frobenius lo está con su traza (es decir, la suma de los elementos de la diagonal principal). Por otra parte, para la matriz identidad I todas las normas matriciales toman el valor 1 pero la norma de Frobenius toma el valor $\sqrt{n}$.

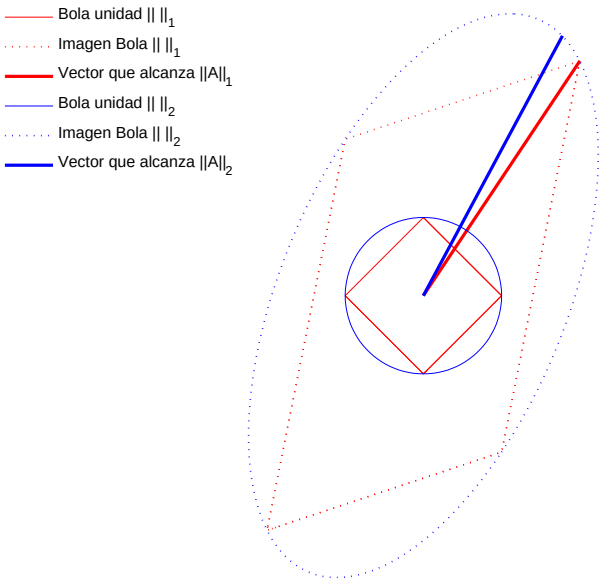


Figura 2.1: Normas subordinadas

▪ Ejemplo **5** En la figura 2.1 se muestran las bolas unidad correspondientes a las normas $\| \ \|_1$ y $\| \ \|_2$ y sus imágenes por la transformación definida por la matriz

$$\mathrm{A} = \begin{pmatrix} 1 & 2 \\ -2 & 3 \end{pmatrix}.$$

También se muestran algunos de los vectores en los que se alcanza el máximo del siguiente conjunto

$$\{\|\mathrm{A}\mathbf{x}\| : \|\mathbf{x}\| = 1\}$$

para cada una de ambas normas. ◇

– EJERCICIO 6 *Determinar la norma subordinada a las normas vectoriales* $\| \ \|_1$, $\| \ \|_2$ *y* $\| \ \|_\infty$ *de la matriz*

$$\mathrm{A} = \begin{pmatrix} 2 & 1 \\ 5 & 4 \end{pmatrix}$$

así como su norma de Frobenius.

Solución: Directamente de la definición se obtiene que

$$\| \mathrm{A} \|_1 = \max\{2+5, 1+4\} = 7,$$

$$\| \mathrm{A} \|_\infty = \max\{2+1, 5+4\} = 9.$$

El producto

$$\mathrm{A}^t\mathrm{A} = \begin{pmatrix} 29 & 22 \\ 22 & 17 \end{pmatrix}$$

tiene como autovalores $\lambda_1 = 0.1965$ y $\lambda_2 = 45.8035$ y como traza $29+17 = 46$. Consecuentemente se tiene que

$$\| \mathrm{A} \|_2 = \sqrt{45.8035} = 6.7678,$$

$$\| \mathrm{A} \|_F = \sqrt{46} = 6.7823 \ .$$

En general, el radio espectral de A no coincide con la norma subordinada a la norma euclídea. De hecho, en el ejemplo de este ejercicio, la matriz A tiene como autovalores $\lambda_1 = 0.5505$ y $\lambda_2 = 5.4495$ y por ello, su radio espectral es 5.4495; valor que no coincide con el de la norma 2. ◇

Una consecuencia inmediata de la definición de norma subordinada es la siguiente desigualdad

$$\| \mathrm{AB} \| \leq \| \mathrm{A} \| \| \mathrm{B} \|$$

válida para dos matrices cualesquiera. Como consecuencia de esta desigualdad se tiene que para toda matriz A

$$\| \mathrm{A}^q \| \leq \| \mathrm{A} \|^q,$$

lo que implica en particular, que la sucesión de potencias sucesivas de una matriz converge a cero si para alguna de sus normas subordinadas se cumple que $\|\mathrm{A}\| < 1$.

Si $\mathbf{x}$ es un vector propio asociado a un autovalor λ de la matriz A, $\mathbf{x}$ verifica que $\mathrm{A}\mathbf{x} = \lambda\mathbf{x}$. De la definición de norma subordinada se deduce que

$$\|\mathrm{A}\mathbf{x}\| \leq \|\mathrm{A}\|\|\mathbf{x}\|$$

En consecuencia

$$|\lambda|\|\mathbf{x}\| \leq \|\mathrm{A}\|\|\mathbf{x}\|$$

y

$$|\lambda| \leq \|\mathrm{A}\|.$$

Puesto que esta desigualdad se verifica para todo autovalor, se concluye que

$$\rho(\mathrm{A}) \leq \|\mathrm{A}\|.$$

El siguiente teorema, cuya demostración se omite (véase [11, Varga] ó [8, Quarteroni, Sacco, Saleri], páginas 25-27), establece de modo más preciso la relación del radio espectral de una matriz con las normas subordinadas

Teorema 1 *Para cualquier matriz cuadrada* A *se cumple*

$$\rho(\mathrm{A}) = \inf\{\|\mathrm{A}\| : \|\ \| \textit{ es una norma subordinada}\}.$$

Es importante tener en cuenta que si A es simétrica, el ínfimo se alcanza en la norma subordinada a la norma euclídea ya que

$$\|\mathrm{A}\|_2 = \rho(\mathrm{A}^2)^{\frac{1}{2}} = \rho(\mathrm{A}).$$

Otra consecuencia del teorema anterior es el siguiente

Corolario 1 *Para una matriz cuadrada arbitraria, las tres afirmaciones siguientes, referidas a las potencias de* A *y el radio espectral, son equivalentes*

1. $\lim_{n\to\infty} \mathrm{A}^n = \mathrm{O}$,
2. $\lim_{n\to\infty} \|\mathrm{A}^n\| = 0$, *para alguna norma subordinada,*
3. $\rho(\mathrm{A}) < 1$.

Es conveniente tener en cuenta que la segunda afirmación podría ampliarse a cualquier norma subordinada ya que todas ellas son equivalentes por ser finita la dimensión de los espacios vectoriales de matrices.

▪ EJEMPLO **6** La existencia de una norma subordinada para la que $\|\mathrm{A}\| > 1$ no impide que $\lim_{n\to\infty} \mathrm{A}^n = \mathrm{O}$. En efecto, para la matriz

$$\mathrm{A} = \begin{pmatrix} \frac{1}{2} & \frac{3}{10} \\ 0 & \frac{4}{5} \end{pmatrix}$$

la norma $\|\mathrm{A}\|_1 = \frac{11}{10} > 1$ pero $\rho(\mathrm{A}) = 0.8 < 1$, lo que garantiza que las potencias de A convergen a O. ◇

2.3 Estabilidad de un sistema de ecuaciones lineales

Cuando se pretende resolver un sistema de ecuaciones numéricas lineales es importante conocer previamente si el sistema es sensible a pequeñas modificaciones en los datos o producidas en el curso del cálculo. Por el momento, se analizará la cuestión de si una pequeña modificación de los datos del sistema $\mathrm{A}\mathbf{x} = \mathbf{b}$ puede generar un importante cambio en la solución. Para ello se considera el sistema perturbado

$$(\mathrm{A} + \delta\mathrm{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$$

donde $\delta\mathrm{A}$, $\delta\mathbf{b}$ son elementos arbitrarios de la misma naturaleza que A y $\mathbf{b}$. El error que introduce esta perturbación está dado por

$$\delta\mathbf{x} = (\mathrm{A} + \delta\mathrm{A})^{-1}(\delta\mathbf{b} - \delta\mathrm{A}\mathbf{x}).$$

Si se toman normas en ambos miembros y se usa para las matrices una norma subordinada, se obtiene que

$$\|\delta\mathbf{x}\| \leq \|(\mathrm{A} + \delta\mathrm{A})^{-1}\|(\|\delta\mathbf{b}\| + \|\delta\mathrm{A}\|\|\mathbf{x}\|).$$

Por otra parte, puesto que $\mathbf{x}$ es solución de la ecuación, se cumple que

$$\|b\| \leq \|\mathrm{A}\|\|\mathbf{x}\|.$$

De las dos últimas desigualdades se deduce que

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|(\mathrm{A} + \delta\mathrm{A})^{-1}\|\|\mathrm{A}\| \left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta\mathrm{A}\|}{\|\mathrm{A}\|} \right)$$

Para completar la estimación del error relativo se usará el siguiente lema de Banach

Lema 1 *Si* D *es una matriz cuadrada tal que* $\|\mathrm{D}\| < 1$ *para una determinada norma subordinada a una norma vectorial entonces* $\mathrm{I}+\mathrm{D}$ *es no-singular y se cumple que*

$$\|(\mathrm{I}+\mathrm{D})^{-1}\| \leq \frac{1}{1-\|\mathrm{D}\|}.$$

Demostración: Para todo vector $\mathbf{x}$ no nulo se tiene que

$$\|(\mathrm{I}+\mathrm{D})\mathbf{x}\| \geq \|\mathbf{x}\| - \|\mathrm{D}\mathbf{x}\| \geq \|\mathbf{x}\|(1-\|\mathrm{D}\|) > 0.$$

Consecuentemente, I+D no es singular. Por otra parte, se cumple la siguiente desigualdad

$$\begin{aligned} 1 &= \|\mathrm{I}\| = \|(\mathrm{I}+\mathrm{D})(\mathrm{I}+\mathrm{D})^{-1}\| \\ &\geq \|(\mathrm{I}+\mathrm{D})^{-1}\| - \|\mathrm{D}\|\|(\mathrm{I}+\mathrm{D})^{-1}\| = (1-\|\mathrm{D}\|)\|(\mathrm{I}+\mathrm{D})^{-1}\|, \end{aligned}$$

lo que prueba el resultado. ◇

Volviendo a la estimación del error, si se usa el lema de Banach se obtiene la siguiente acotación

$$\|(\mathrm{A}+\delta\mathrm{A})^{-1}\| = \|(\mathrm{I}+\mathrm{A}^{-1}\delta\mathrm{A})^{-1}\mathrm{A}^{-1}\| \leq \|\mathrm{A}^{-1}\|\frac{1}{1-\|\mathrm{A}\|\delta\mathrm{A}\|},$$

si se cumple que $\|\mathrm{A}^{-1}\|\|\delta\mathrm{A}\| < 1$. Si se usa esta acotación en la estimación del error relativo de la solución del sistema perturbado, se obtiene que

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathrm{A}\|\|\mathrm{A}^{-1}\|}{1-\|\mathrm{A}^{-1}\|\|\delta\mathrm{A}\|}\left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta\mathrm{A}\|}{\|\mathrm{A}\|}\right).$$

Esta estimación sugiere que una medida de esta sensibilidad es el número de condición de A, definido por

$$\mathrm{cond}(\mathrm{A}) = \|\mathrm{A}\|\|\mathrm{A}^{-1}\|$$

para toda matriz cuadrada no-singular A. De este modo, si $\mathrm{cond}(\mathrm{A})\frac{\|\delta\mathrm{A}\|}{\|\mathrm{A}\|} < 1$ se cumple que

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\mathrm{cond}(\mathrm{A})}{1-\mathrm{cond}(\mathrm{A})\frac{\|\delta\mathrm{A}\|}{\|\mathrm{A}\|}}\left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta\mathrm{A}\|}{\|\mathrm{A}\|}\right).$$

Además, de la relación

$$1 = \|\mathrm{I}\| = \|\mathrm{A}\mathrm{A}^{-1}\| \leq \|\mathrm{A}\|\|\mathrm{A}^{-1}\| = \mathrm{cond}(\mathrm{A})$$

se deduce que el número de condición de una matriz es siempre mayor o igual que 1. Si el número de condición es relativamente pequeño los errores que se generarían en la solución serían pequeños. Por el contrario, para sistemas lineales cuya matriz de coeficientes tenga un número de condición muy alto, no hay garantía de que la solución calculada con una aritmética finita se mantenga próxima a la solución exacta.

▪ EJEMPLO **7** Se considera la matriz

$$\mathrm{A} = \begin{pmatrix} 1+\frac{1}{\epsilon} & -1+\frac{1}{\epsilon} \\ -1+\frac{1}{\epsilon} & 1+\frac{1}{\epsilon} \end{pmatrix}$$

donde $\epsilon > 0$ es un número destinado a tender a 0. La matriz inversa de A es

$$\mathrm{A}^{-1} = \frac{1}{4}\begin{pmatrix} \epsilon+1 & \epsilon-1 \\ \epsilon-1 & \epsilon+1 \end{pmatrix}.$$

Fácilmente se comprueba que

$$\|\mathrm{A}\|_2 = \rho(\mathrm{A}^t\mathrm{A})^{\frac{1}{2}} = \frac{2}{\epsilon}; \quad \|\mathrm{A}^{-1}\|_2 = \rho(\mathrm{A}^{-t}\mathrm{A}^{-1})^{\frac{1}{2}} = \frac{1}{2}$$

y de ello se deduce que $\mathrm{cond}(\mathrm{A}) = \frac{1}{\epsilon}$. La matriz A está mal condicionada cuando $\epsilon \to 0$. ◇

Una matriz está perfectamente condicionada si su número de condición es 1. En particular, las matrices ortogonales, que verifican que $\mathrm{A}^t\mathrm{A} = \mathrm{I}$, están perfectamente condicionadas ya que

$$\|\mathrm{A}\|_2 = \rho(\mathrm{A}^t\mathrm{A})^{\frac{1}{2}} = \rho(\mathrm{I})^{\frac{1}{2}} = 1$$

y

$$\|\mathrm{A}^{-1}\|_2 = \rho(\mathrm{A}\mathrm{A}^t)^{\frac{1}{2}} = \rho(\mathrm{I})^{\frac{1}{2}} = 1.$$

Por otra parte, si A es una matriz simétrica definida positiva entonces se cumple que

$$\|\mathrm{A}\|_2 = \lambda_{max}, \quad \|\mathrm{A}^{-1}\|_2 = \frac{1}{\lambda_{min}}$$

donde λ_{min} and λ_{max} representan respectivamente el mínimo y el máximo autovalor de A. De este modo, se tiene que

$$\mathrm{cond}_2(\mathrm{A}) = \frac{\lambda_{max}}{\lambda_{min}}.$$

2.4 Sistemas lineales de gran dimensión

La resolución eficiente de sistemas lineales es una necesidad presente en muchos problemas del cálculo científico. La dimensión de los problemas que aparecen en el cálculo científico es considerablemente elevada, ya que los sistemas lineales que aparecen en la simulación numérica de procesos físicos pueden tener miles o quizás millones de incógnitas. Esto es algo que no puede ser ignorado salvo que sólo se busque el análisis teórico del problema. Si se ignorase la talla de los problemas se podrían proponer soluciones que fuesen totalmente ineficientes. El cálculo científico tiene en cuenta en la elección de algoritmo de resolución de un sistema lineal, su coste computacional, su estabilidad y cómo se maneja la información de los coeficientes del sistema.

Se puede expresar un sistema lineal compatible y determinado de ecuaciones numéricas en forma matricial como

$$\mathrm{A}\mathbf{x} = \mathbf{b}$$

donde A es una matriz de dimensión $n \times n$ de coeficientes reales o complejos y $\mathbf{b}$ un vector columna de n componentes reales o complejas.

Desde un punto de vista teórico, un sistema lineal de ecuaciones, cuya matriz de coeficientes tenga determinante diferente de cero, puede ser resuelto por la fórmula de Cramer. De este modo, se puede calcular la solución del sistema lineal

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3$$

mediante las fórmulas explícitas

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{12} & a_{13} \\ a_{13} & a_{12} & a_{13} \end{vmatrix}}, \quad x_2 = \frac{\begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{12} & a_{13} \\ a_{13} & a_{12} & a_{13} \end{vmatrix}}, \quad x_3 = \frac{\begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{12} & a_{13} \\ a_{13} & a_{12} & a_{13} \end{vmatrix}}$$

Desde un punto de vista teórico, las fórmulas de Cramer tienen dos estimables cualidades: son explícitas y tienen validez general para cualquier sistema lineal. Sin embargo, desde el punto de vista del cálculo científico, la aplicación de estas fórmulas directamente tiene una dificultad que las hace poco prácticas para el cálculo; requieren el cálculo de determinantes y si la

evaluación directa de un determinante se realiza mediante la regla de Sarrus se requieren $n!$ sumas, cada una de ellas con $n-1$ productos. La función entera $n!$ alcanza pronto valores muy elevados. Las resolución de un sistema lineal de 30 incógnitas es impensable con estas fórmulas. Es preciso buscar algoritmos de resolución que reduzcan el coste computacional, es decir, tales que el número de operaciones elementales que deban realizarse, sea reducido.

La eficiencia de la regla de Cramer mejoraría considerablemente si se utilizan algoritmos más eficientes de cálculo de determinantes que la aplicación directa de la regla de Sarrus. En todo caso, existen métodos más eficaces para ello, como son los que se describen en las siguientes secciones.

2.5 Matrices dispersas

De la experiencia con los sistemas lineales de gran tamaño que surgen en los problemas reales del cálculo científico, se conoce que en su mayoría, el número de coeficientes nulos es muy elevado en proporción a los que no son nulos. Esto nos obliga a considerar formas que no son estándares para representar una matriz en un computador para evitar el tener que almacenar y manejar tanta información inútil.

Generalmente, las matrices son usadas como la representación en una base, de transformaciones lineales que llevan vectores en vectores. Por consiguiente, el índice del vector origen y el índice del vector imagen enlazan a través de una entidad de dos índices lo que sugiere la estructura rectangular de la matriz de coeficientes. Inevitablemente la imagen rectangular de un sistema lineal está siempre en nuestra mente. No obstante, sería un error, si en un problema real, se tratase de almacenar una matriz dispersa en modo rectangular completo. Es decir, si la matriz posee una proporción pequeña de coeficientes que no son nulos, lo realmente eficiente es buscar una estructura que almacene únicamente la información útil. Se puede precisar esta idea con el concepto de densidad de una matriz. Concretamente, se define la densidad de una matriz como el cociente entre el número de coeficientes no nulos y el número total de coeficientes.

El siguiente ejemplo sencillo ilustra esta situación: La matriz de coeficientes y el término independiente de un sistema lineal son

$$\mathrm{A} = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}, \qquad \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}.$$

La densidad de esta matriz es $\frac{3n-2}{n^2}$. Si la dimensión de los vectores es alta, la cantidad de memoria necesaria para almacenar la matriz de coeficientes podría resultar excesiva para la capacidad de almacenamiento del computador. Por ejemplo, si $n = 10000$ la densidad de la matriz es menor que 3×10^{-4}, lo que indica que solamente una proporción ínfima de la matriz contiene información útil.

La primera idea para simplificar el almacenamiento de la matriz es tener en cuenta la estructura simétrica tridiagonal. En otras palabras, la información útil de la matriz está situada en la diagonal principal y las dos diagonales adyacentes. Bastaría con almacenar la diagonal principal y una adyacente para poder reconstruir la matriz entera. De este modo, el sistema lineal está descrito por 3 vectores $(\mathrm{A}_1, \mathrm{A}_2, b)$ que representan las dos diagonales y el término independiente. Existen recursos eficaces para manejar matrices dispersas que almacenan internamente

- Los coeficientes que no son nulos.
- Los índices de filas de los coeficientes que no son nulos.
- Los índices de columnas de los coeficientes que no son nulos.

Por ejemplo, la matriz que se ha creado como llena

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 2 & 4 \\ 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \end{pmatrix}$$

se puede almacenar de modo equivalente en forma dispersa

(1,1)	(4,1)	(2,2)	(3,2)	(4,2)	(2,3)	(1,4)	(2,4)
1	1	1	1	2	2	1	4

La eficacia de un método de resolución está relacionada en general con el hecho de que el método respete o no, estas formas de almacenamiento reducido. La regla de Cramer, aplicada con la regla de Sarrus para la evaluación de determinantes, es en este sentido un procedimiento inadecuado ya que destruye esta estructura reducida de las matrices al considerar las matrices modificadas por el término independiente que dejan de ser tridiagonales. Los entornos de cálculo científico tienen incorporados eficientes modos de operar con matrices dispersas. No está entre los propósitos de este curso el de describir en detalle los modos de almacenamiento disperso ni su relación con los algoritmos de resolución de sistemas lineales pero se recomienda la lectura de bibliografía especializada en el tema.

2.6 Método de eliminación de Gauss

Desde el punto de vista tradicional del Álgebra, el método de eliminación no es tan grato como la regla de Cramer, ya que no está definido por fórmulas explícitas y además no tiene validez general puesto que falla cuando encuentra un coeficiente nulo en la posición seleccionada de la diagonal principal. Sin embargo, desde el punto de vista del cálculo científico, es más útil ya que es el que tiene coste computacional más bajo y se adapta perfectamente a muchas formas de almacenamiento reducido. El método de eliminación, no es un nuevo método que haya surgido con el desarrollo del cálculo científico basado en el cálculo automático. Realmente, la idea en la que se apoya parece muy natural en el proceso de resolución de ecuaciones. De hecho, aunque es un método atribuido a Gauss, según algunas referencias ya fue utilizado en las culturas milenarias de la antigua China.

▪ Ejemplo **8** El proceso de eliminación de eliminación de variables y resolución del sistema

$$\begin{aligned} x + y + z &= 3, \\ 2x - y + z &= 2, \\ x + 2y - 2z &= 1, \end{aligned}$$

se puede llevar a cabo del siguiente modo:

- *Transformación a un sistema triangular*: A la segunda ecuación se le resta la primera multiplicada por 2 y a la tercera ecuación se le resta la primera multiplicada por 1. Como resultado de estas operaciones resulta

$$\begin{aligned} x + y + z &= 3, \\ -3y - z &= -4, \\ y - 3z &= -2. \end{aligned}$$

A la tercera ecuación se le resta la segunda multiplicada por $-\frac{1}{3}$. De ello resulta

$$\begin{aligned} x + y + z &= 3, \\ -3y - z &= -4, \\ -\frac{10}{3}z &= -\frac{10}{3}. \end{aligned}$$

- *Sustitución retrógrada*: De la última ecuación se obtiene $z = 1$. Se sustituye el valor obtenido de z en la penúltima ecuación y se obtiene el valor $y = 1$. Finalmente, se sustituyen los valores de z e y en la anterior ecuación y se obtiene $x = 1$.

Este es el modo tradicional de utilizar el método de eliminación de Gauss. En el proceso anterior se distinguen una primera etapa de triangulación del sistema y una segunda etapa de sustitución retrógrada de las incógnitas. Se pueden organizar los cálculos de un modo matricial. La matriz de coeficientes se descompone como

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & -1 & 1 \\ 1 & 2 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{1}{3} & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & -3 & -1 \\ 0 & 0 & -\frac{10}{3} \end{pmatrix}$$

La primera matriz es triangular inferior y los términos de la diagonal principal son todos iguales a 1. En cada columna, por debajo de la diagonal principal, aparecen los multiplicadores utilizados en la eliminación. La segunda matriz que aparece en la descomposición es la que resulta en el proceso de triangulación.

Se puede ver la eliminación como una secuencia de transformaciones

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & -1 & 1 \\ 1 & 2 & -2 \end{pmatrix} \longmapsto \begin{pmatrix} 1 & 1 & 1 \\ 2 & -3 & -1 \\ 1 & 1 & -3 \end{pmatrix} \longmapsto \begin{pmatrix} 1 & 1 & 1 \\ 2 & -3 & -1 \\ 1 & -\frac{1}{3} & -\frac{10}{3} \end{pmatrix}$$

Para reducir el uso de memoria, las posiciones de la matriz por debajo de la diagonal se van modificando con los multiplicadores (color azul) ya que se supone que en esas posiciones se han producido ceros. En las restantes posiciones de la matriz van apareciendo los nuevos coeficientes de la matriz. Es importante notar que en el paso $k-$ésimo sólo se modifica la submatriz que corresponde a las últimas $4 - k$ filas y columnas.

Se puede comprobar directamente que este procedimiento nos permite construir la factorización

$$\mathrm{A} = \mathrm{LU},$$

donde L es una matriz triangular inferior, cuyos coeficientes coinciden con los que están debajo de la diagonal de la matriz final del proceso y son iguales a 1 en la diagonal principal

$$\mathrm{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{1}{3} & 1 \end{pmatrix}$$

y U es una matriz triangular superior cuyos coeficientes coinciden con los que están en o por encima de la diagonal principal

$$\mathrm{U} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & -3 & -1 \\ 0 & 0 & -\frac{10}{3} \end{pmatrix}.$$

De este modo, la resolución del sistema lineal

$$\mathrm{LU}\mathbf{x} = \mathbf{b}$$

se puede transformar en la resolución consecutiva de los dos siguientes:

$$\mathrm{L}\mathbf{y} = \mathbf{b}, \qquad \text{eliminación progresiva}$$

$$\mathrm{U}\mathbf{x} = \mathbf{y}, \qquad \text{sustitución retrógrada}$$

que es notablemente más simple. $\diamond$

La idea utilizada en el ejemplo precedente puede ser extendida a un sistema de n ecuaciones

$$\mathrm{A}\mathbf{x} = \mathbf{b}.$$

Para realizar la eliminación de los coeficientes de la primera columna se considera la siguiente matriz

$$\mathrm{E}(1) = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{a_{n-1,1}}{a_{11}} & 0 & \cdots & 1 & 0 \\ -\frac{a_{n,1}}{a_{11}} & 0 & \cdots & 0 & 1 \end{pmatrix}$$

Se introduce la siguiente notación $\mathrm{A}(1) = \mathrm{A}$, $\mathrm{A}(2) = \mathrm{E}(1)\mathrm{A}(1)$. La matriz $\mathrm{A}(2)$ tiene ceros en las posiciones que están debajo de la diagonal en la primera columna. Además $\det(\mathrm{A}) = \det(\mathrm{A}(2))$.

Para la eliminación de coeficientes correspondientes a posiciones que están debajo de la diagonal principal, se procede de modo recurrente. Si $\mathrm{A}(j)$ representa la matriz en la que se han eliminado estos coeficientes en las $j-1$ primeras columnas y $\mathrm{E}(j-1)$ la matriz de multiplicadores de la $(j-1)-$ésima columna entonces se construye la matriz de multiplicadores para la columna

$j-$ésima como

$$\mathrm{E}(j) = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 1 & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -\frac{a^j_{j+1,j}}{a^j_{jj}} & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -\frac{a^j_{n-1,j}}{a^j_{jj}} & \cdots & \cdots & 1 & 0 \\ 0 & 0 & \cdots & -\frac{a^j_{n,j}}{a^j_{jj}} & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

Con esta matriz se eliminan los correspondientes coeficientes en la columna $j-$ésima de la matriz $\mathrm{A}(j)$, obteniéndose la matriz

$$\mathrm{A}(j+1) = \mathrm{E}(j)\mathrm{A}(j)$$

que tiene ceros en la posiciones que están debajo de la diagonal principal en las j primeras columnas. Finalmente, se obtiene que

$$\mathrm{A}(n) = \mathrm{E}(n-1)\mathrm{A}(n-1) = \cdots = \mathrm{E}(n-1)\cdots\mathrm{E}(1)\mathrm{A}.$$

Fácilmente se prueba que

$$\mathrm{E}(j)^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 1 & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{a^j_{j+1,j}}{a^j_{jj}} & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{a^j_{n-1,j}}{a^j_{jj}} & \cdots & \cdots & 1 & 0 \\ 0 & 0 & \cdots & \frac{a^j_{n,j}}{a^j_{jj}} & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

y que $\mathrm{L} = \mathrm{E}(1)^{-1} \cdots \mathrm{E}(n-1)^{-1}$ está dado por

$$\mathrm{L} = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ \frac{a^1_{21}}{a^1_{11}} & 1 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 1 & \vdots & \vdots & \vdots & \vdots \\ \frac{a^1_{j1}}{a^1_{11}} & \frac{a^2_{j2}}{a^2_{22}} & \cdots & \frac{a^j_{j+1,j}}{a^j_{jj}} & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{a^1_{n-1,1}}{a^1_{11}} & \frac{a^2_{n-1,2}}{a^2_{22}} & \cdots & \frac{a^j_{n-1,j}}{a_{jj}} & \cdots & \cdots & 1 & 0 \\ \frac{a^1_{n,1}}{a^1_{11}} & \frac{a^2_{n,2}}{a^2_{22}} & \cdots & \frac{a^j_{n,j}}{a_{jj}} & \cdots & \cdots & \frac{a^{n-1}_{n,n-1}}{a^{n-1}_{n-1,n-1}} & 1 \end{pmatrix}$$

Por otra parte, la matriz $\mathrm{U} = \mathrm{A}(n)$ es una matriz triangular superior y se cumple que $\mathrm{A} = \mathrm{LU}$.

Si se cuentan las operaciones necesarias para realizar el cálculo de la factorización LU, se obtiene que son del orden de $\frac{n^3}{3}$, contando sumas, productos y divisiones. Así si $n = 200$, el método requiere más de dos millones de operaciones elementales. En estas circunstancias el efecto de los errores de redondeo puede ser considerable. Un modo de reducir el deterioro de la solución es evitar la división por números pequeños. La técnica, llamada de pivote parcial, en la eliminación Gaussiana utiliza esta idea realizando permutaciones de las filas de modo que el pivote sea el de mayor valor absoluto en la columna considerada. A veces, el uso de estrategias como la de pivote parcial es inevitable. Este es el caso que ocurre si en el cálculo de los multiplicadores para la eliminación de los coeficientes que están debajo de la diagonal, aparece con un cero como pivote. Se ilustra esta situación con el siguiente ejemplo

▪ Ejemplo **9** En el cálculo de la factorización LU de la matriz

$$\mathrm{A} = \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}$$

cuyo determinante no es nulo, el primer pivote es nulo y por consiguiente no puede eliminarse el coeficiente 1 de la segunda fila primera columna. Puesto que parece natural cambiar el orden de las filas y esta operación puede realizarse multiplicando la matriz A por una matriz de permutación P (que se obtiene permutando filas o columnas en la matriz identidad), se obtiene directamente la factorización LU del producto PA

$$\mathrm{PA} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}. \qquad \diamond$$

Para una matriz arbitraria, es evidente que la existencia de la factorización depende de la ausencia de pivotes nulos en la diagonal principal. Si se utiliza una estrategia que altera dos de las $n-k+1$ últimas filas de la matriz $\mathrm{A}(k)$ mediante la multiplicación por una matriz de permutación $\mathrm{P}(k)$, el proceso de eliminación Gaussiana sería el siguiente

$$\mathrm{A}(n) = \mathrm{E}(n-1)\mathrm{P}(n-1)\mathrm{E}(n-2)\mathrm{P}(n-2)\cdots\mathrm{E}(1)\mathrm{P}(1)\mathrm{A}.$$

Sin dificultad se prueba que $\mathrm{P}(i)\mathrm{E}(j) = \hat{\mathrm{E}}(j)\mathrm{P}(i)$ si $i > j$ siendo $\hat{\mathrm{E}}$ es la matriz que resulta de intercambiar en $\mathrm{E}(j)$ las posiciones (i,j) y (k,j), donde $k \geq i > j$ es el índice de la fila que ocupa el nuevo pivote. La razón es que $\mathrm{P}(i)$ sólo afecta a filas y columnas posteriores a la j. De ello se deduce que la factorización anterior puede expresarse como

$$\mathrm{U} = \mathrm{A}(n) = \mathrm{E}(n-1)\hat{\mathrm{E}}(n-2)\cdots\hat{\mathrm{E}}(1)\mathrm{PA} = \mathrm{L}^{-1}\mathrm{PA}$$

donde la matriz P representa la permutación

$$\mathrm{P} = \mathrm{P}(n-1)\mathrm{P}(n-2)\cdots\mathrm{P}(1)$$

y

$$\mathrm{L} = \Big(\mathrm{E}(n-1)\hat{\mathrm{E}}(n-2)\cdots\hat{\mathrm{E}}(1)\Big)^{-1}.$$

De este modo, se obtiene que

$$\mathrm{PA} = \mathrm{LU}.$$

Es decir, el método de eliminación de Gauss permite obtener una factorización LU de la matriz permutada de A. Si se pretendiese resolver un sistema $\mathrm{A}\mathbf{x} = \mathbf{b}$ usando la factorización con pivote, una vez calculada ésta, se aplicaría la técnica de la sustitución retrógrada al sistema $\mathrm{LU}\mathbf{x} = \mathrm{P}\mathbf{b}$.

▬ Ejercicio 7 *Calcular la factorización* LU *de la matriz permutada de*

$$\begin{pmatrix} 2 & 3 & 1 \\ 1 & 1 & 4 \\ 4 & 4 & 1 \end{pmatrix}$$

que resulta al usar la estrategia de pivote parcial.

Solución: *Primera etapa en la eliminación*: El pivote escogido es 4 de la posición $3,1$. La matriz de permutación, la matriz de multiplicadores y la matriz parcialmente transformada correspondientes son

$$\mathrm{P}(1) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathrm{E}(1) = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{pmatrix}, \quad \mathrm{A}(2) = \begin{pmatrix} 4 & 4 & 1 \\ 0 & 0 & \frac{15}{4} \\ 0 & 1 & \frac{1}{2} \end{pmatrix}.$$

Segunda etapa en la eliminación: El pivote escogido es 1 de la posición 3, 2. La matriz de permutación, la matriz de multiplicadores y la matriz parcialmente transformada correspondientes son

$$\mathrm{P}(2) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathrm{E}(2) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathrm{A}(3) = \begin{pmatrix} 4 & 4 & 1 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{15}{4} \end{pmatrix}.$$

De acuerdo con el procedimiento anteriormente indicado, se obtiene la siguiente factorización

$$\mathrm{A}(3) = \mathrm{E}(2)\mathrm{P}(2)\mathrm{E}(1)\mathrm{P}(1)\mathrm{A} = \mathrm{E}(2)\hat{\mathrm{E}}(1)\mathrm{P}(2)\mathrm{P}(1)\mathrm{A}$$

donde

$$\hat{\mathrm{E}}(1) = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{4} & 0 & 1 \end{pmatrix}.$$

Finalmente, la factorización LU buscada es la siguiente

$$\mathrm{L} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{4} & 0 & 1 \end{pmatrix}, \quad \mathrm{U} = \begin{pmatrix} 4 & 4 & 1 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{15}{4} \end{pmatrix}, \quad \mathrm{P} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}. \qquad \diamond$$

El producto de dos matrices triangulares inferiores es triangular inferior. En efecto, si A y B son matrices triangulares inferiores y C = AB entonces

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = \sum_{k=j}^{i} a_{ik} b_{kj}.$$

Consecuentemente, si $j > i$ se tiene que $c_{ij} = 0$ como se quería probar. Además, si A y B son matrices triangulares inferiores con unos en la diagonal principal entonces $c_{ii} = a_{ii} b_{ii} = 1$. Finalmente, sin dificultad se puede también probar que la inversa de una matriz triangular inferior (cuyos elementos de la diagonal principal no se anulen) es triangular inferior. Obviamente, se pueden establecer resultados análogos para matrices triangulares superiores. Con estos resultados se analiza la unicidad de la factorización LU de una matriz. Si se admite que una matriz tiene dos factorizaciones LU

$$\mathrm{A} = \mathrm{LU} = \bar{\mathrm{L}}\bar{\mathrm{U}}$$

sin cambio de pivote, entonces la matriz triangular $\bar{\mathrm{L}}^{-1}\mathrm{L}$ con unos en la diagonal coincide con la matriz triangular superior $\bar{\mathrm{U}}\mathrm{U}^{-1}$. Esto solamente ocurre si ambas matrices son la identidad. Es decir, la factorización LU de una matriz, si existe es única.

▬ EJERCICIO 8 *Determinar la factorización* LU *de la matriz* A *de dimensión* $n \times n$ *definida por*

$$\mathrm{A} = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}.$$

Solución: Si se usa un método de inducción se puede comprobar sin dificultad que la factorización LU de A es

$$\mathrm{A} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\frac{2}{3} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\frac{n-1}{n} & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ 0 & \frac{3}{2} & -1 & \cdots & 0 & 0 \\ 0 & 0 & \frac{4}{3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{n}{n-1} & -1 \\ 0 & 0 & 0 & \cdots & 0 & \frac{n+1}{n} \end{pmatrix}. \diamond$$

Si una matriz arbitraria A tiene una factorización $\mathrm{A} = \mathrm{LU}$ y D representa la matriz diagonal cuyos miembros de la diagonal principal coinciden con los de U entonces la matriz A puede expresarse como $\mathrm{A} = \mathrm{LD}\bar{\mathrm{U}}$ donde $\bar{\mathrm{U}} = \mathrm{D}^{-1}\mathrm{U}$ es una matriz triangular superior con unos en la diagonal principal. Es frecuente usar la siguiente terminología

- $\mathrm{A} = \mathrm{LU}$, factorización de Doolittle,
- $\mathrm{A} = \bar{\mathrm{L}}\bar{\mathrm{U}}$ con $\bar{\mathrm{L}} = \mathrm{LD}$, factorización de Crout,

de modo que la diagonal de unos está en el primer caso, en el primer factor y en el segundo caso, en el segundo.

Puesto que toda matriz de permutación P verifica que $\mathrm{P}^t\mathrm{P} = \mathrm{I}$, realmente la factorización obtenida para la matriz original A es $\mathrm{A} = \mathrm{P}^t\mathrm{LU}$ o lo similar para las restantes factorizaciones. Queda pues una cuestión teórica que debe ser abordada. Es la de examinar bajo que condiciones se puede garantizar que P puede ser la identidad. El siguiente teorema establece una condición basada

en la no anulación de los menores principales de la matriz (es decir, de los determinantes de las matrices formadas por las j primeras filas y columnas de A para $j \leq n$)

Teorema 2 *Si los menores principales de* A *no se anulan entonces admite una factorización* $\mathrm{A} = \mathrm{LU}$.

Demostración: Se utiliza una prueba de inducción. Para $n = 1$, su verificación es trivial. En el caso $n = 2$, para que existiese una factorización $\mathrm{A} = \mathrm{LU}$ de la siguiente forma

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ l_{21} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix},$$

sería necesario que las siguientes operaciones fuesen posibles

$$u_{11} = a_{11}, \;\; u_{12} = a_{12}, \;\; u_{22} = \frac{a_{11}a_{22} - a_{12}a_{21}}{a_{11}}, \;\; l_{21} = \frac{a_{21}}{a_{11}}.$$

Puesto que los dos menores de la matriz A no se anulan

$$a_{11} \neq 0, \;\; a_{11}a_{22} - a_{12}a_{21} \neq 0,$$

en las relaciones anteriores el denominador no se anula y además u_{22} tampoco se anula. Así pues, es posible construir la factorización LU.

Admitiendo que el resultado es cierto para las matrices de dimensión $n-1$ que cumplen las hipótesis del teorema, se probará a continuación que también es cierto para las matrices de dimensión n. Para ello se considera la siguiente descomposición por bloques

$$\mathrm{A} = \begin{pmatrix} \mathrm{A}_{11} & \mathrm{A}_{12} \\ \mathrm{A}_{21} & \mathrm{A}_{22} \end{pmatrix} = \begin{pmatrix} \mathrm{L}_{11} & 0 \\ \mathrm{L}_{21} & \mathrm{L}_{22} \end{pmatrix} \begin{pmatrix} \mathrm{U}_{11} & \mathrm{U}_{12} \\ 0 & \mathrm{U}_{22} \end{pmatrix}$$

donde $\mathrm{A}_{11}, \mathrm{L}_{11}$ y U_{11} representan matrices de dimensión $n-1$ mientras que $\mathrm{A}_{22}, \mathrm{L}_{22}$ y U_{22} son escalares. Esta ecuación matricial equivale a las siguientes

$$\begin{aligned} \mathrm{A}_{11} &= \mathrm{L}_{11}\mathrm{U}_{11} \\ \mathrm{A}_{12} &= \mathrm{L}_{11}\mathrm{U}_{12} \\ \mathrm{A}_{21} &= \mathrm{L}_{21}\mathrm{U}_{11} \\ \mathrm{A}_{22} &= \mathrm{L}_{21}\mathrm{U}_{12} + \mathrm{L}_{22}\mathrm{U}_{22} \end{aligned}$$

Puesto que A_{11} es de dimensión $n-1$ y sus menores principales (que lo son también de A) no son nulos existe la factorización $\mathrm{A}_{11} = \mathrm{L}_{11}\mathrm{U}_{11}$. Directamente U_{12} se define como $\mathrm{U}_{12} = \mathrm{L}_{11}^{-1}\mathrm{A}_{12}$ (la inversa de L_{11} existe ya que su determinante es igual a 1). De mismo modo, se calcula $\mathrm{L}_{21} = \mathrm{A}_{21}\mathrm{U}_{11}^{-1}$. Finalmente, $\mathrm{L}_{22} = 1$ por ser un elemento de la diagonal de L y $\mathrm{U}_{22} = \mathrm{A}_{22} - \mathrm{L}_{21}\mathrm{U}_{12}$. ◇

2.7 Métodos especiales para matrices simétricas

Otra posibilidad es factorizar la matriz en la forma $\mathrm{A} = \tilde{\mathrm{L}}\tilde{\mathrm{U}}$ donde $\tilde{\mathrm{L}} = \mathrm{L}\mathrm{D}^{\frac{1}{2}}$ y $\tilde{\mathrm{U}} = \mathrm{D}^{-\frac{1}{2}}\mathrm{U}$ (la matriz diagonal $\mathrm{D}^{\frac{1}{2}}$ tiene en la diagonal principal las raíces cuadradas de los elementos de D en las mismas posiciones). La matriz $\tilde{\mathrm{L}}$ es triangular inferior y $\tilde{\mathrm{U}}$ es triangular superior y ambas tienen la misma diagonal principal. Además, de la unicidad de la factorización LU se deduce la unicidad de la factorización $\tilde{\mathrm{L}}\tilde{\mathrm{U}}$.

Si la matriz A es simétrica entonces $\mathrm{A} = \tilde{\mathrm{L}}\tilde{\mathrm{U}} = \tilde{\mathrm{U}}^t\tilde{\mathrm{L}}^t$. De la unicidad de la factorización $\tilde{\mathrm{L}}\tilde{\mathrm{U}}$ se deduce $\tilde{\mathrm{L}} = \tilde{\mathrm{U}}^t$, es decir, que existe una factorización $\mathrm{A} = \tilde{\mathrm{L}}\tilde{\mathrm{L}}^t$ donde $\tilde{\mathrm{L}}$ es una matriz triangular inferior. Esta factorización se conoce como factorización de Cholesky de una matriz simétrica.

Teorema 3 *Una matriz simétrica* A *es definida positiva si y sólo si admite una factorización de Cholesky* $\mathrm{A} = \mathrm{L}\mathrm{L}^t$ *donde* L *es una matriz triangular inferior invertible.*

Demostración: Si $\mathrm{A} = \mathrm{L}\mathrm{L}^t$ entonces

$$\mathbf{x}^t\mathrm{A}\mathbf{x} = \mathbf{x}^t\mathrm{L}\mathrm{L}^t\mathbf{x} = \|\mathrm{L}^t\mathbf{x}\|^2 > 0 \quad \text{para todo } \mathbf{x} \neq 0.$$

Recíprocamente, si A es definida positiva entonces todos los menores principales son positivos. Del teorema 2 de la sección anterior se deduce que A admite una factorización LU sin permutación. Razonando como anteriormente se prueba la existencia de una factorización de Cholesky para A. ◇

Si una matriz A admite una factorización de Cholesky $\mathrm{A} = \mathrm{L}\mathrm{L}^t$ entonces

$$a_{ij} = \sum_{k=1}^{n} l_{ik}l_{jk} = \sum_{k\leq \min\{i,j\}} l_{ik}l_{jk}.$$

Para $j = 1$ se tiene que $a_{i1} = l_{i1}l_{11}$ para $i = 1, \cdots, n$, de donde se deduce que

$$l_{11} = \sqrt{a_{11}}, \qquad l_{i1} = \frac{a_{i1}}{l_{11}}.$$

Para $j > 1$ e $i > j$, se tiene que

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}, \quad l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}}{l_{jj}}.$$

Este modo de realizar el cálculo de L se adapta bien a las formas de almacenamiento reducido. Por ejemplo, si la matriz A es tridiagonal, la matriz L es tridiagonal.

En la siguiente matriz, se representa una etapa intermedia en el cálculo. Los coeficientes en colores azul ó rojo, representan valores calculados. El coeficiente en verde va a ser sustituido por l_{43}, implicando en el cálculo a sí mismo y a los coeficientes en rojo.

$$\begin{pmatrix} l_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1n} \\ l_{21} & l_{22} & a_{23} & a_{24} & \cdots & a_{2n} \\ l_{31} & l_{32} & l_{33} & a_{34} & \cdots & a_{3n} \\ l_{41} & l_{42} & a_{43} & a_{44} & \cdots & a_{4n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & a_{n3} & a_{n4} & \cdots & a_{nn} \end{pmatrix}.$$

▬ EJERCICIO 9 *Determinar la factorización de Cholesky de la matriz*

$$\mathrm{A} = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}.$$

Usar este resultado para calcular el determinante de A. *Indicación: Los coeficientes de la factorización deben ser expresados de modo exacto y no mediante una aproximación decimal.*

Solución: Los coeficientes de la matriz L están dados por

$$\begin{aligned} l_{11} &= \sqrt{a_{11}} = \sqrt{2} \\ l_{21} &= \frac{a_{21}}{l_{11}} = -\sqrt{\frac{1}{2}}, \quad l_{22} = \sqrt{a_{22} - l_{21}^2} = \sqrt{\frac{3}{2}} \\ l_{31} &= 0 \quad l_{32} = -\sqrt{\frac{2}{3}} \quad l_{33} = \sqrt{\frac{4}{3}} \\ l_{41} &= 0 \quad l_{42} = 0 \quad l_{43} = -\sqrt{\frac{3}{4}} \quad l_{44} = \sqrt{\frac{5}{4}} \end{aligned}$$

La matriz L es

$$\mathrm{L} = \begin{pmatrix} \sqrt{2} & 0 & 0 & 0 \\ -\sqrt{\frac{1}{2}} & \sqrt{\frac{3}{2}} & 0 & 0 \\ 0 & -\sqrt{\frac{2}{3}} & \sqrt{\frac{4}{3}} & 0 \\ 0 & 0 & -\sqrt{\frac{3}{4}} & \sqrt{\frac{5}{4}} \end{pmatrix}.$$

En consecuencia, el determinante de la matriz A se puede calcular como

$$|\mathrm{A}| = |\mathrm{L}|^2 = 2\,\frac{3}{2}\,\frac{4}{3}\,\frac{5}{4} = 5. \qquad \diamond$$

2.8 Factorización QR

Las matrices ortogonales conservan las longitudes de los vectores y están perfectamente condicionadas, de modo que en su uso, no se espera que se produzcan crecientes notables de errores de redondeo. El siguiente resultado se puede ver como una variante del que establece la factorización LU de una matriz con la modificación de que uno de los factores es una matriz ortogonal

Teorema 4 *Toda matriz invertible* A *admite una factorización* A = QR *donde* Q *es una matriz ortogonal y* R *una matriz triangular superior.*

Demostración: Para probar este teorema basta tener en cuenta cualquiera de las construcciones de la factorización QR que se describen a continuación. ◇

Se puede usar una factorización QR de la matriz A para resolver el sistema lineal $\mathrm{A}\mathbf{x} = \mathbf{b}$. En efecto, una vez calculada la factorización bastaría resolver el sistema triangular $\mathrm{R}\mathbf{x} = \mathrm{Q}^t\mathbf{b}$ por sustitución retrógrada. Si bien es cierto que en la norma subordinada a la norma euclídea, el número de condición de A y R coinciden, lo cierto es que evita la sustitución progresiva y consecuentemente limita la generación de errores de redondeo. Se puede considerar como un método más adecuado para resolver sistemas mal condicionados que los métodos basados en la factorización LU. Otras aplicaciones interesantes de esta factorización, se estudiarán en capítulos posteriores.

2.8.1. Método de ortogonalización de Gram-Schmidt

La idea básica de este método aplicado a una matriz invertible A es la siguiente:

- Se consideran los n vectores columna $\{\mathbf{a}^i : i = 1, 2, \cdots, n\}$ de A.
- A cada uno de ellos se le restan las componentes respecto a los vectores que le preceden.
- Finalmente se normalizan estos vectores.

No obstante, se pueden separar las etapas de ortogonalización y normalización del siguiente modo: Se introduce la nueva base

$$\begin{aligned}
\mathbf{p}^1 &= \mathbf{a}^1, \\
\mathbf{p}^2 &= \mathbf{a}^2 - \frac{\mathbf{a}^2 \cdot \mathbf{p}^1}{\|\mathbf{p}^1\|^2}\mathbf{p}^1, \\
&\cdots \\
\mathbf{p}^n &= \mathbf{a}^n - \sum_{i=1}^{n-1} \frac{\mathbf{a}^n \cdot \mathbf{p}^i}{\|\mathbf{p}^i\|^2}\mathbf{p}^i.
\end{aligned}$$

Si P representa la matriz de columnas $\{\mathbf{p}^i : i = 1, 2, \cdots, n\}$, las ecuaciones anteriores pueden expresarse en forma matricial como

$$\mathrm{P} = \mathrm{A} - \mathrm{P} \begin{pmatrix} 0 & m_{12} & \cdots & m_{1n} \\ 0 & 0 & \cdots & m_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_{n-1,n} \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

con $m_{ij} = \frac{\mathbf{a}^j \cdot \mathbf{p}^i}{\|\mathbf{p}^i\|^2}$ para $j > i$, de donde se deduce que

$$\mathrm{A} = \mathrm{P} \begin{pmatrix} 1 & m_{12} & \cdots & m_{1,n-1} & m_{1n} \\ 0 & 1 & \cdots & m_{2,n-1} & m_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & m_{n-1,n} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Si D es la matriz diagonal definida por $d_{ii} = \|\mathbf{p}^i\|$, las matrices $\mathrm{Q} = \mathrm{PD}^{-1}$ y

$$\mathrm{R} = \mathrm{D} \begin{pmatrix} 1 & m_{12} & \cdots & m_{1,n-1} & m_{1n} \\ 0 & 1 & \cdots & m_{2,n-1} & m_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & m_{n-1,n} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix},$$

forman una factorización QR de la matriz A.

▪ Ejemplo **10** Sea A la matriz definida por

$$\mathrm{A} = \begin{pmatrix} 2 & -1 & 0 \\ 0 & 0 & -2 \\ 0 & 2 & -1 \end{pmatrix}.$$

Los vectores columna de esta matriz son

$$\mathbf{a}^1 = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{a}^2 = \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix}, \ \mathbf{a}^3 = \begin{pmatrix} 0 \\ -2 \\ -1 \end{pmatrix}.$$

Los vectores $\mathbf{p}^i$ son

$$\mathbf{p}^1 = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{p}^2 = \mathbf{a}^2 + \frac{1}{2}\mathbf{p}^1 = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}, \ \mathbf{p}^3 = \mathbf{a}^3 - 0\mathbf{p}^1 + \frac{1}{2}\mathbf{p}^2 = \begin{pmatrix} 0 \\ -2 \\ 0 \end{pmatrix}.$$

La matriz D que resulta es la siguiente

$$\mathrm{D} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Finalmente, una descomposición QR de la matriz A es la siguiente

$$\mathrm{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & 2 \end{pmatrix}. \qquad \diamond$$

El proceso de Gram-Schmidt requiere que los vectores columnas sean linealmente independientes ya que si uno de ellos es combinación lineal de los precedentes, el vector resultante es nulo. De hecho el método se vuelve inestable cuando los vectores columnas están próximos a ser linealmente dependientes.

2.8.2. Método de Householder

Una transformación de Householder es una aplicación lineal que lleva un vector en su reflejado respecto a un eje. La matriz asociada a una transformación de Householder tiene la siguiente forma

$$\mathrm{H} = \mathrm{I} - 2\mathbf{e}\mathbf{e}^t$$

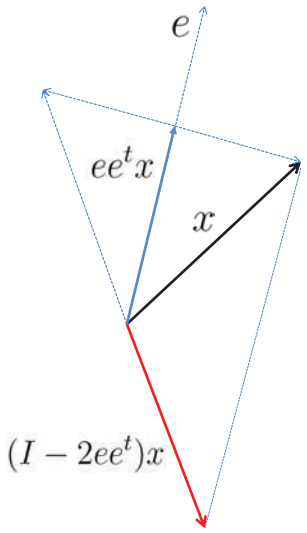


Figura 2.2: Transformación de Householder

donde **e** es un vector de norma 1 que representa al eje. Se prueba sin dificultad que H es una matriz simétrica ortogonal.

Siempre se puede encontrar una aplicación lineal que transforme un vector **x** fijado en otro vector **y** de la misma norma, también fijado. Para construirla, basta tomar

$$\mathbf{e} = \pm \frac{\mathbf{x} - \mathbf{y}}{\|\mathbf{x} - \mathbf{y}\|}.$$

En efecto, de la identidad

$$\| \mathbf{x} \|^2 - \mathbf{x} \cdot \mathbf{y} = \frac{1}{2} \left(\| \mathbf{x} \|^2 - \| \mathbf{y} \|^2 \right) + \frac{1}{2} \| \mathbf{x} - \mathbf{y} \|^2,$$

teniendo en cuenta que $\| \mathbf{x} \| = \| \mathbf{y} \|$, se deduce que

$$\mathrm{H}\mathbf{x} = \mathbf{x} - 2 \frac{\mathbf{x} - \mathbf{y}}{\|\mathbf{x} - \mathbf{y}\|^2} (\|\mathbf{x}\|^2 - \mathbf{x} \cdot \mathbf{y}) = \mathbf{x} - (\mathbf{x} - \mathbf{y}) = \mathbf{y}.$$

También es evidente que la transformación de Householder $\mathrm{H}^* = \mathrm{I} - 2\bar{\mathbf{e}}\bar{\mathbf{e}}^t$ para

$$\bar{\mathbf{e}} = \pm \frac{\mathbf{x} + \mathbf{y}}{\|\mathbf{x} + \mathbf{y}\|}.$$

transforma **x** en $-\mathbf{y}$ y verifica que $\mathrm{H}^* = 2\mathbf{e}\mathbf{e}^t - \mathrm{I}$. En la figura 2.8.2 se comprueba que H^* representa la simetría respecto al eje definido por **e** mientras que H es una reflexión sobre el mismo eje.

▬ EJERCICIO 10 *Sea θ un ángulo fijado. Construir una transformación de Householder* H *que lleve el vector* $(1,0)^t$ *en el vector* $(\cos\theta, \operatorname{sen}\theta)^t$. *¿Coincide* H *con la rotación de ángulo θ?.*

Solución: Si $\mathbf{u} = (1-\cos\theta, -\operatorname{sen}\theta)^t$ entonces se tiene que

$$|\mathbf{u}|^2 = 2(1-\cos\theta).$$

Se elige el vector $\mathbf{e}$ como

$$\mathbf{e} = \frac{1}{\sqrt{2(1-\cos\theta)}}(1-\cos\theta, -\operatorname{sen}\theta)^t.$$

La matriz de Householder resultante será

$$\mathrm{H} = \mathrm{I} - 2ee^t = \begin{pmatrix} \cos\theta & \operatorname{sen}\theta \\ \operatorname{sen}\theta & -\cos\theta \end{pmatrix}.$$

H es una simetría ya que $\det \mathrm{H} = -1$ y la rotación tiene determinante positivo. La rotación de ángulo θ también transforma $(1,0)^t$ en $(\cos\theta, \operatorname{sen}\theta)^t$ pero no es una simetría respecto a un eje. ◇

Se examina ahora cómo pueden usarse las transformaciones de Householder para construir una factorización QR de una matriz A. Si se representan por $\mathbf{a}^i$ los vectores columna de la matriz A y P es una matriz arbitraria de la mismas dimensiones que A, entonces

$$\mathrm{PA} = (\mathrm{P}\mathbf{a}^1, \mathrm{P}\mathbf{a}^2, \cdots, \mathrm{P}\mathbf{a}^n)$$

donde las comas indican la concatenación de los vectores columna $\mathrm{P}\mathbf{a}^i$. El método de Householder utiliza la siguiente construcción: En la primera etapa, se construye la transformación de Householder P(1) que transforma $\mathbf{a}^1$ en el vector $\|\mathbf{a}^1\|(1,0,\cdots,0)^t$. La construcción de esta transformación es posible ya que en este caso ambos vectores tienen la misma norma. Si se representa $\mathrm{A}(1) = \mathrm{A}$ y se define la matriz $\mathrm{A}(2) = \mathrm{P}(1)\mathrm{A}(1)$, esta última matriz tiene ceros debajo del primer elemento de la diagonal principal. Se descompone la matriz A(2) como sigue

$$\mathrm{A}(2) = \left(\begin{array}{c|c} a_{11}(2) & \mathrm{A}_{12}(2) \\ \hline 0 & \mathrm{A}_{22}(2) \end{array}\right),$$

donde $\mathrm{A}_{22}(2)$ es una matriz cuadrada de dimensión $n-1$. Se construye una matriz de Householder $\mathrm{P}_{22}(2)$ de dimensión $n-1$ que produzca ceros debajo del primer elemento de la diagonal y con ella se construye la matriz de dimensión n

$$\mathrm{P}(2) = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & \mathrm{P}_{22}(2) \end{array}\right).$$

Con esta matriz ortogonal se construye $\mathrm{A}(3) = \mathrm{P}(2)\mathrm{A}(2) = \mathrm{P}(2)\mathrm{P}(1)\mathrm{A}$ que tiene ceros debajo de la diagonal en las dos primeras columnas. Si se reitera el procedimiento hasta construir la matriz triangular superior

$$\mathrm{R} = \mathrm{P}(n-1)\cdots\mathrm{P}(2)\mathrm{P}(1)\mathrm{A}.$$

Si se toma

$$\mathrm{Q} = \mathrm{P}(1)\mathrm{P}(2)\cdots\mathrm{P}(n-1)$$

se obtiene una matriz ortogonal tal que $\mathrm{A} = \mathrm{QR}$.

▪ Ejemplo **11** Sea A la matriz definida por

$$\mathrm{A} = \begin{pmatrix} 2 & -1 & 0 \\ 0 & 0 & -2 \\ 0 & 2 & -1 \end{pmatrix}.$$

Según la construcción de Householder que se acaba de describir, la matriz $\mathrm{P}(1)$ es la identidad $\mathrm{P}(1) = \mathrm{I}$ ya que $\|\mathbf{x}\| = \|(2,0,0)^t\| = 2$ y por lo tanto $\mathbf{x} = \mathbf{y}$ y $\mathrm{A}(2) = \mathrm{A}$.

Para la segunda etapa, se prescinde de la primera fila y de la primera columna. De ello resulta una matriz 2×2 que tiene en la primera columna el vector $(0,2)^t$ de norma 2. De acuerdo con el método QR se va a transformar este vector en otro que sea de la forma $2(1,0)^t$. Para realizar esta transformación, según la construcción de Householder, se usa el vector unitario

$$\mathbf{e} = \frac{(0,2)^t - (2,0)^t}{\|(0,2)^t - (2,0)^t\|} = \frac{1}{\sqrt{2}}(-1,1)^t$$

y las matrices de Householder correspondientes son

$$\mathrm{P}_{22}(2) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathrm{P}(2) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Finalmente se obtiene

$$\mathrm{R} = \mathrm{P}(2)\mathrm{P}(1)\mathrm{A} = \begin{pmatrix} 2 & -1 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & -2 \end{pmatrix}.$$

$$\mathrm{Q} = \mathrm{P}(1)\mathrm{P}(2) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Es posible utilizar una tercera transformación ortogonal con el fin de conseguir coeficientes positivos en la diagonal principal de R. Bastaría usar la simetría

$$\mathrm{P}(3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

para producir la factorización final

$$\mathrm{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 \\ 0 & 2 & -1 \\ 0 & 0 & 2 \end{pmatrix}. \qquad \diamond$$

Ahora se analiza la unicidad de la factorización QR. Supongamos que A tiene dos descomposiciones QR,

$$\mathrm{A} = \mathrm{QR} = \hat{\mathrm{Q}}\hat{\mathrm{R}}.$$

Se considera la matriz $\mathrm{T} = \hat{\mathrm{Q}}^t\mathrm{Q} = \hat{\mathrm{R}}\mathrm{R}^{-1}$. En consecuencia $\mathrm{T} = \hat{\mathrm{Q}}^t\mathrm{Q}$ es ortogonal y también $\mathrm{T} = \hat{\mathrm{R}}\mathrm{R}^{-1}$ es triangular superior. Puesto que la inversa de una matriz ortogonal es su traspuesta, la matriz T tiene que ser diagonal con coeficientes en valor absoluto igual a l. De la igualdad $\hat{\mathrm{R}} = \mathrm{TR}$ se deduce que si impone la condición de que los coeficientes de la diagonal principal en R son positivos entonces la factorización QR es única. Por otra parte, de las construcciones de los apartados anteriores se deduce toda matriz no singular A admite una única factorización $\mathrm{A} = \mathrm{QR}$ donde Q es una matriz ortogonal y R una matriz triangular superior con coeficientes positivos en la diagonal principal.

2.9 Métodos iterativos

Como ya se ha indicado anteriormente, una dificultad que puede presentarse al resolver un sistema lineal de gran dimensión es el volumen de memoria que un computador puede requerir para almacenar la propia matriz y las que surgen en su factorización. En estas circunstancias, algunos métodos iterativos que generan una sucesión que se aproxima a la solución, pueden ser una buena alternativa.

La idea básica de esta clase de métodos es la de considerar la solución $\mathbf{x}$ como un punto fijo de una transformación afín

$$\mathrm{T}\mathbf{x} = \mathrm{H}\mathbf{x} + \mathbf{c}.$$

La aplicación reiterada de esta transformación

$$\mathbf{x}^{(k+1)} = \mathrm{H}\mathbf{x}^{(k)} + \mathbf{c}$$

puede conducir en el límite a la solución si se cumplen determinadas condiciones de convergencia.

Un resultado fundamental para establecer la convergencia de esta sucesión es el que proporciona el siguiente

Teorema 5 *Sean* H *una matriz cuadrada y* $\mathbf{c}$ *un vector tales que la ecuación* $\mathbf{x} = \mathrm{H}\mathbf{x} + \mathbf{c}$ *tiene una solución única* $\mathbf{x}$. *La sucesión* $\{\mathbf{x}^{(k)}\}$ *generada por la transformación*

$$\mathbf{x}^{(k+1)} = \mathrm{H}\mathbf{x}^{(k)} + \mathbf{c}$$

con un vector de partida arbitrario $\mathbf{x}^{(0)}$, *converge a* $\mathbf{x}$ *si y sólo si* $\rho(\mathrm{H}) < 1$.

Demostración: Si $\mathbf{x}$ es la solución del sistema lineal $\mathbf{x} = \mathrm{H}\mathbf{x} + \mathbf{c}$, cumple que

$$\mathbf{x}^{(k+1)} - \mathbf{x} = \mathrm{H}(\mathbf{x}^{(k)} - \mathbf{x}) = \cdots = \mathrm{H}^{k+1}(\mathbf{x}^{(0)} - \mathbf{x}).$$

De ello, se deduce que

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}\| \leq \|\mathrm{H}^{k+1}\|\|\mathbf{x}^{(0)} - \mathbf{x}\| \leq \|\mathrm{H}\|^{k+1}\|\mathbf{x}^{(0)} - \mathbf{x}\| \tag{2.1}$$

para cualquier norma vectorial, en el supuesto de que la norma matricial esté subordinada a ella. De este modo, si $\rho(\mathrm{H}) < 1$, para cualquier número real ϵ tal que $0 < \epsilon < 1 - \rho(\mathrm{H})$, del teorema 1 de la página 20 se desprende que existe una norma vectorial tal que la norma subordinada $\|\ \|$ correspondiente verifica que

$$\rho(\mathrm{H}) < \|\mathrm{H}\| < \rho(\mathrm{H}) + \epsilon < 1.$$

En consecuencia, se tiene que

$$\lim_{k\to\infty} \|\mathbf{x}^{(k+1)} - \mathbf{x}\| = 0.$$

Recíprocamente, supongamos que $\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}$ para todo vector inicial $\mathbf{x}^{(0)}$. Si se toma en particular $\mathbf{x}^{(0)} = \mathbf{x} + \mathbf{e}^i$ para $\mathbf{e}^i = (0, \cdots, 0, 1, 0, \cdots, 0)^t$ (el 1 en la posición i), entonces $\mathbf{x}^{(k)} - \mathbf{x} = \mathrm{H}^k\mathbf{e}^i$. Pero, $\mathrm{H}^k\mathbf{e}^i$ es la i-ésima columna de H^k. Puesto que $\mathbf{x}^{(k)} - \mathbf{x}$ tiende a 0, se cumple que $\|\mathrm{H}^k\mathbf{e}^i\|_1 \to 0$ para cualquier i, o equivalentemente, $\|\mathrm{H}^k\|_1 \to 0$ lo que implica que que $\rho(\mathrm{H}) < 1$. La unicidad de solución que se exige al sistema lineal en el enunciado del lema, se debe a que gracias al lema de Banach, si $\rho(\mathrm{H}) < 1$ la matriz $I - \mathrm{H}$ no es singular. ◇

2.10 Métodos iterativos clásicos

Obviamente, hay muchos modos de interpretar la solución de un sistema lineal $\mathrm{A}\mathbf{x} = \mathbf{b}$ como un punto fijo de una transformación $\mathbf{x} = \mathrm{H}\mathbf{x} + \mathbf{c}$. No obstante, el llamado método de relajación agrupa a algunos de ellos y puede utilizarse incluso en el caso no-lineal. La idea de este método es la siguiente: Se considera una matriz M fácilmente invertible que se conoce como pre-acondicionador (preconditioner). Con ayuda de esta matriz se construye la ecuación de punto fijo

$$\mathbf{x} = \mathbf{x} - \mathrm{M}^{-1}(\mathrm{A}\mathbf{x} - \mathbf{b})$$

que se acomoda a la situación de la sección anterior si se toma $\mathrm{H} = \mathrm{I} - \mathrm{M}^{-1}\mathrm{A}$ y $\mathbf{c} = \mathrm{M}^{-1}b$. De acuerdo con el teorema 5 de la página 45, la dificultad está en encontrar pre-acondicionadores que verifiquen

$$\rho(\mathrm{I} - \mathrm{M}^{-1}\mathrm{A}) < 1$$

y que la resolución del sistema lineal

$$\mathrm{M}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = \mathbf{b} - \mathrm{A}\mathbf{x}^{(k)}$$

pueda realizarse de un modo muy simple.

Un grupo de métodos iterativos clásicos puede ser incluido en esta clase, si se descompone la matriz A como

$$\mathrm{A} = \mathrm{D} - \mathrm{L} - \mathrm{U} \tag{2.2}$$

donde

- D es una matriz diagonal verificando que

$$d_{ii} = a_{ii} \qquad \text{para } i = 1, \ldots, n$$

- L es una matriz triangular inferior verificando que

$$l_{ij} = -a_{ij} \qquad \text{para } i, j = 1, \ldots, n, i > j$$

- U es una matriz triangular superior verificando que

$$u_{ij} = -a_{ij} \qquad \text{para } i, j = 1, \ldots, n, i < j$$

Por ejemplo, la siguiente matriz está descompuesta de este modo

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Con ayuda de estas matrices se definen los siguientes pre-acondicionadores para la relajación del sistema

1. Método de Jacobi M = D. La iteración resultante es

$$\mathrm{D}\mathbf{x}^{(k+1)} = (\mathrm{L}+\mathrm{U})\mathbf{x}^{(k)} + \mathbf{b}.$$

2. Método de Gauss-Seidel M = D − L. La iteración resultante es

$$(\mathrm{D}-\mathrm{L})\mathbf{x}^{(k+1)} = \mathrm{U}\mathbf{x}^{(k)} + \mathbf{b}.$$

3. Método de sobre-relajación (Successive-Over-Relaxation) $\mathrm{M} = \frac{1}{\omega}\mathrm{D} - \mathrm{L}$. La iteración resultante es

$$\left(\frac{1}{\omega}\mathrm{D}-\mathrm{L}\right)\mathbf{x}^{(k+1)} = \left(\left(\frac{1}{\omega}-1\right)\mathrm{D}+\mathrm{U}\right)\mathbf{x}^{(k)} + \mathbf{b}.$$

   El parámetro de relajación ω permite escalar adecuadamente la diagonal del sistema frente a la parte inferior a la diagonal.

▪ Ejemplo **12** Se considera el sistema lineal

$$\begin{aligned} 2x + y + z &= 1 \\ x - 2y + z &= 1 \\ x + y + 2z &= 1. \end{aligned}$$

En este sistema la iteración de Jacobi resulta

$$\begin{aligned} 2x^{(k+1)} + y^{(k)} + z^{(k)} &= 1 \\ x^{(k)} - 2y^{(k+1)} + z^{(k)} &= 1 \\ x^{(k)} + y^{(k)} + 2z^{(k+1)} &= 1. \end{aligned}$$

Si se parte de un vector de componentes $x^{(0)} = y^{(0)} = z^{(0)} = 1$, de la primera ecuación se deduce que

$$x^{(1)} = \frac{1}{2}\left(1 - y^{(0)} - z^{(0)}\right) = -\frac{1}{2}.$$

De la segunda ecuación se obtiene

$$y^{(1)} = -\frac{1}{2}\left(1 - x^{(0)} - z^{(0)}\right) = \frac{1}{2}$$

y finalmente, de la tercera ecuación se obtiene

$$z^{(1)} = \frac{1}{2}\left(1 - x^{(0)} - y^{(0)}\right) = -\frac{1}{2}. \qquad \diamond$$

Se puede observar en el ejemplo precedente que en la puesta en práctica del método de Jacobi hay que tener cuidado en no mezclar las componentes $x_i^{(k+1)}$ con las componentes $x_i^{(k)}$. Es necesario no almacenar el valor de $x_i^{(k+1)}$ en la posición que ocupa $x_i^{(k)}$ ya que este valor se utilizará en el cálculo en los nodos adyacentes. Por lo tanto es preciso mantener simultáneamente dos vectores para poder realizar la iteración completa.

La diferencia esencial entre los métodos directos e iterativos estriba en que los métodos directos alcanzan la solución exacta, salvo errores de redondeo, y tienen un número predecible de operaciones mientras que los métodos iterativos realizan un número de operaciones que depende del criterio de parada que se utilice para finalizar las iteraciones.

– EJERCICIO 11 *Se considera el siguiente sistema de ecuaciones lineales*

$$x_{i+1} - 2x_i + x_{i-1} = 1, \quad \textit{para } i = 1, 2, 3, 4,$$

$$x_0 = x_5 = 0.$$

- *Expresar el sistema en forma matricial.*
- *Determinar si el método iterativo de Jacobi es convergente para este sistema lineal.*

Solución: La forma matricial de expresar el sistema lineal es la siguiente

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}.$$

El método de Jacobi corresponde a la elección

$$\mathrm{D} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} \qquad \mathrm{L} + \mathrm{U} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

en la descomposición (2.2). Consecuentemente, se tiene que

$$\mathrm{H} = \mathrm{D}^{-1}(\mathrm{L} + \mathrm{U}) = \frac{1}{2} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Desafortunadamente

$$\| \mathrm{H} \|_\infty = \| \mathrm{H} \|_1 = 1$$

y por ello se debe afinar el análisis calculando el radio espectral de H. El polinomio característico de la matriz simétrica H es

$$\lambda^4 - \frac{3}{4}\lambda^2 + \frac{1}{16}$$

y sus raíces son

$$\lambda = \pm\sqrt{\frac{3 \pm \sqrt{5}}{8}}.$$

Puesto que valor máximo de estas raíces es menor que 1 en valor absoluto, entonces se concluye que el método de Jacobi es convergente . ◇

El cálculo del radio espectral de H para verificar la condición necesaria y suficiente para la convergencia, no es en general sencillo. Sin embargo, el teorema 5 de la página 45 permite establecer algunas condiciones suficientes para la convergencia, que pueden ser fácilmente verificables. En concreto, las condiciones

- $\| \mathrm{H} \|_\infty < 1$,
- $\| \mathrm{H} \|_1 < 1$

garantizan la convergencia del método.

A fin de expresar de un modo más simple estas condiciones en el caso de los métodos clásicos se introducen las siguientes definiciones

- A es estrictamente diagonal dominante por filas si $\sum_{j \neq i} |a_{ij}| < |a_{ii}|$ para cada valor de i.
- A es estrictamente diagonal dominante por columnas si $\sum_{i \neq j} |a_{ij}| < |a_{jj}|$ para cada valor de j

que permiten formular el siguiente

Teorema 6 *Si* A *es una matriz estrictamente diagonal dominante por filas entonces los métodos de Jacobi y Gauss-Seidel son convergentes.*

Demostración: En el caso del método de Jacobi, la matriz que define la iteración es

$$\mathrm{H}_J = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & -\frac{a_{n-1,n}}{a_{n-1,n-1}} \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \cdots & 0 \end{pmatrix}.$$

Consecuentemente se tiene que $\| \mathrm{H}_J \|_\infty < 1$ si y sólo si A es estrictamente diagonal dominante.

En el caso del método de Gauss-Seidel, la matriz que define la iteración es $\mathrm{H}_{GS} = (\mathrm{D}-\mathrm{L})^{-1}\mathrm{U}$. Sea λ un autovalor de esta matriz y $\mathbf{x}$ un vector propio asociado tal que $\| \mathbf{x} \|_\infty = 1$. De la definición de la norma se deduce que existe un índice k tal que $|x_k| = 1$ y $|x_i| \leq 1$ para todo $i \neq k$. Consecuentemente se tiene

$$\lambda(\mathrm{D} - \mathrm{L})\mathbf{x} = \mathrm{U}\mathbf{x}.$$

Si se selecciona la $k-$ésima componente de los vectores de esta igualdad, se obtiene

$$\lambda(a_{kk}x_k - \sum_{i<k} a_{ki}x_i) = \sum_{i>k} a_{ki}x_i.$$

De esta igualdad se deduce

$$|\lambda| \leq \frac{\sum_{i>k} |a_{ki}||x_i|}{|a_{kk}||x_k| - \sum_{i<k} |a_{ki}||x_i|} \leq \frac{\sum_{i>k} |a_{ki}|}{|a_{kk}| - \sum_{i<k} |a_{ki}|}.$$

Finalmente de la definición de matriz estrictamente diagonal dominante se deduce que $|\lambda| < 1$, lo que prueba que $\rho(\mathrm{H}_{GS}) < 1$ $\diamond$.

Teorema 7 *Si* A *es una matriz simétrica definida positiva, el método de Gauss-Seidel es convergente.*

Demostración: Sea λ un autovalor de H_{GS}, que puede ser complejo y $\mathbf{x} \in \mathbb{C}^n$ un vector propio asociado. Puesto A es simétrica, $\mathrm{U} = \mathrm{L}^t$ y consecuentemente

$$\mathrm{L}^t\mathbf{x} = \lambda(\mathrm{D} - \mathrm{L})\mathbf{x}.$$

Si se multiplica esta igualdad por $\mathbf{x}^*$ (traspuesto y conjugado de $\mathbf{x}$) se obtiene

$$\mathbf{x}^*\mathrm{L}^t\mathbf{x} = \lambda\mathbf{x}^*(\mathrm{D} - \mathrm{L})\mathbf{x}$$

de donde se deduce que

$$\mathbf{x}^*\mathrm{L}^t\mathbf{x} + \lambda\mathbf{x}^*\mathrm{L}\mathbf{x} = \lambda\mathbf{x}^*\mathrm{D}\mathbf{x}. \tag{2.3}$$

Por otra parte, si λ es distinto de 0, se cumple que

$$\mathbf{x}^*\mathrm{A}\mathbf{x} = \mathbf{x}^*\mathrm{D}\mathbf{x} - \mathbf{x}^*\mathrm{L}\mathbf{x} - \mathbf{x}^*\mathrm{L}^t\mathbf{x} = \left(\frac{1}{\lambda} - 1\right)\mathbf{x}^*\mathrm{L}^t\mathbf{x}$$

y consecuentemente $\lambda \neq 1$ ya que $\mathbf{x}^*\mathrm{A}\mathbf{x} > 0$. Además, se tiene

$$\mathbf{x}^*\mathrm{L}^t\mathbf{x} = \frac{\lambda}{1-\lambda}\mathbf{x}^*\mathrm{A}\mathbf{x}.$$

Si se trasponen y conjugan ambos miembros de la anterior desigualdad, se obtiene

$$\mathbf{x}^*\mathrm{L}\mathbf{x} = \frac{\bar{\lambda}}{1-\bar{\lambda}}\mathbf{x}^*\mathrm{A}\mathbf{x}.$$

Si se sustituyen estas expresiones en la igualdad 2.3, se llega a

$$\frac{1-|\lambda|^2}{|1-\lambda|^2}\mathbf{x}^*\mathrm{A}\mathbf{x} = \mathbf{x}^*\mathrm{D}\mathbf{x}.$$

Por ser A definida positiva, los coeficientes de la diagonal principal son positivos. De este modo, los números $\mathbf{x}^*\mathrm{A}\mathbf{x}$ y $\mathbf{x}^*\mathrm{D}\mathbf{x}$ son positivos, por lo que se tiene que cumplir que

$$\frac{1-|\lambda|^2}{|1-\lambda|^2} > 0,$$

lo que implica que $|\lambda| < 1$ como se quería demostrar. ◇

▬ EJERCICIO 12 *Sea α una constante real conocida. Se considera el sistema lineal*

$$\begin{aligned} 4x_1 + x_2 + \alpha x_3 &= 1 \\ x_1 + 4x_2 + x_3 &= 2 \\ \alpha x_1 + x_2 + 4x_3 &= 3. \end{aligned}$$

Analizar la convergencia del método Gauss-Seidel cuando se aplica a este sistema lineal, en términos de los posibles valores de α, usando los criterios disponibles para matrices diagonalmente dominantes y para matrices definidas positivas.

Solución: La matriz de coeficientes

$$A = \begin{pmatrix} 4 & 1 & \alpha \\ 1 & 4 & 1 \\ \alpha & 1 & 4 \end{pmatrix}$$

es estrictamente diagonal dominante si y sólo si $|\alpha| < 3$. En este rango de valores de la constante α, el método de Gauss-Seidel es convergente. Por otra parte, la matriz A es definida positiva si y sólo si

$$\det\,(\mathrm{A}) = -4\alpha^2 + 2\alpha + 56 > 0$$

. Consecuentemente, si $\alpha \in (-\frac{7}{2}, 4)$, el método de Gauss-Seidel es convergente. $\diamond$

TEOREMA 8 *Si $\omega \notin (0,2)$ el método SOR no es convergente. Si* A *es una matriz simétrica definida positiva, el método SOR es convergente si y sólo si $0 < \omega < 2$.*

Demostración: Sea

$$\mathrm{H}_{SOR} = \left(\frac{1}{\omega}\mathrm{D} - \mathrm{L}\right)^{-1} \left(\left(\frac{1}{\omega} - 1\right)\mathrm{D} + \mathrm{U}\right)$$

la matriz que define la iteración en el método SOR. Si se calcula el determinante de esta matriz se obtiene

$$\begin{aligned} \det\left(\mathrm{H}_{\mathrm{SOR}}\right) &= \det\left(\left(\frac{1}{\omega}\mathrm{D} - \mathrm{L}\right)^{-1}\right) \det\left(\left(\frac{1}{\omega} - 1\right)\mathrm{D} + \mathrm{U}\right) \\ &= \omega^n \left(\frac{1}{\omega} - 1\right)^n = (1-\omega)^n. \end{aligned}$$

Puesto que el determinante de una matriz coincide con el producto de sus autovalores, se deduce que

$$|1-\omega|^n = \prod_{i=1}^{n} |\lambda_i| \leq \rho(\mathrm{H}_{SOR})^n$$

de donde finalmente se obtiene

$$\rho(\mathrm{H}_{SOR}) \geq |1-\omega|$$

lo que demuestra que una condición necesaria para la convergencia del método es que $\omega \in (0,2)$.

Por otra parte, si $\mathrm{S} = \left(\frac{1}{\omega}\ \mathrm{D} - \mathrm{L}\right)^{-1}$ entonces

$$\frac{1}{\omega}\mathrm{D} = \mathrm{L} + \mathrm{S}^{-1}.$$

Con estas igualdades equivalentes se puede eliminar el parámetro ω de la expresión que define H_{SOR} y de este modo se obtiene

$$\mathrm{H}_{SOR} = \mathrm{I} - \mathrm{SA}.$$

Se introduce una nueva matriz $\mathrm{N} = \mathrm{A} - \mathrm{H}_{SOR}^t \mathrm{A} \mathrm{H}_{SOR}$. Puesto que A es simétrica, se cumple que

$$\begin{aligned} \mathrm{N} &= \mathrm{ASA} + \mathrm{A}^t\mathrm{S}^t\mathrm{A} - \mathrm{A}^t\mathrm{S}^t\mathrm{ASA} \\ &= \mathrm{A}^t\mathrm{S}^t(\mathrm{S}^{-t} + \mathrm{S}^{-1} - \mathrm{A})\mathrm{SA} = \mathrm{A}^t\mathrm{S}^t \left(\frac{2}{\omega} - 1\right) \mathrm{DSA}. \end{aligned}$$

Esta matriz es definida positiva si y sólo si $\omega \in (0, 2)$. Además, si esta matriz es definida positiva y $\mathbf{x}$ es un vector propio real asociado a un autovalor λ de H_{SOR} se cumple que

$$0 < \mathbf{x}^t N \mathbf{x} = \mathbf{x}^t \mathrm{A} \mathbf{x} - (\mathrm{H}_{SOR}\mathbf{x})^t \mathrm{A} \mathrm{H}_{SOR} \mathbf{x} = (1 - |\lambda|^2) \mathbf{x}^t \mathrm{A} \mathbf{x}$$

de donde se deduce que $|\lambda| < 1$. ◇

En el espacio vectorial de las matrices reales se considera la siguiente relación de orden parcial: una matriz D verifica la relación $\mathrm{D} \geq 0$ si y sólo si $d_{ij} \geq 0$ para $i, j = 1, \cdots, n$. Las siguientes definiciones se apoyan en esta relación de orden

- $\mathrm{A} = \mathrm{B} - \mathrm{C}$ es una descomposición regular de A si $\mathrm{B}^{-1} \geq 0$ y $\mathrm{C} \geq 0$.
- A es una M-matriz si $a_{ij} \leq 0$ para $i \neq j$ y $\mathrm{A}^{-1} \geq 0$.

▪ Ejemplo **13** La matriz A definida por

$$\mathrm{A} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

cuya inversa es

$$\mathrm{A}^{-1} = \begin{pmatrix} 3/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 3/4 \end{pmatrix}$$

es una M-matriz ya que los coeficientes fuera de la diagonal son negativos mientras que los coeficientes de su inversa no lo son. La descomposición $\mathrm{A} = \mathrm{B} - \mathrm{C}$ con

$$\mathrm{B} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad \mathrm{C} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

es regular ya que ni los coeficientes de la inversa de B ni los de C son negativos. La matriz A no es estrictamente diagonal dominante.

El siguiente teorema muestra la utilidad de estos conceptos

Teorema 9

- *Si* $\mathrm{A}^{-1} \geq 0$ *y* $\mathrm{A} = \mathrm{B} - \mathrm{C}$ *es una descomposición regular entonces* $\rho(\mathrm{B}^{-1}\mathrm{C}) < 1$.
- *Si* A *es una M-matriz, entonces las descomposiciones de Jacobi y Gauss-Seidel son regulares. En este caso, ambos métodos son convergentes.*

Ejercicio 13 *Se considera el sistema lineal*

$$\begin{aligned} x_1 - 4x_2 &= -3 \\ 8x_1 - x_2 &= 7. \end{aligned}$$

Determinar si se puede utilizar el método iterativo de Gauss-Seidel para aproximar su solución. En caso desfavorable, proponer alguna idea simple para que pueda utilizarse.

Solución: En forma matricial el sistema lineal resulta

$$\begin{pmatrix} 1 & -4 \\ 8 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -3 \\ 7 \end{pmatrix}.$$

La matriz $\mathrm{H}_{GS} = (\mathrm{D} - \mathrm{L})^{-1}\mathrm{U}$ asociada al método de Gauss-Seidel es

$$\mathrm{H}_{GS} = \begin{pmatrix} 0 & 4 \\ 0 & 32 \end{pmatrix}$$

cuyo radio espectral es 32. Así pues, no es convergente el método de Gauss-Seidel aplicado directamente al sistema lineal. No obstante, es inmediato ver que alterando el orden de las ecuaciones resulta

$$\begin{pmatrix} 8 & -1 \\ 1 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 7 \\ -3 \end{pmatrix}.$$

En este caso, la matriz de coeficientes es estrictamente diagonal dominante y consecuentemente el método es convergente. ◇

Como se pone de manifiesto en el ejercicio anterior, una etapa fundamental en la aplicación de un método iterativo es la adecuación previa de la matriz. Un sistema lineal $\mathrm{A}\mathbf{x} = \mathbf{b}$ no modifica sus soluciones si se multiplican ambos miembros de la igualdad por una matriz no-singular. Las elecciones más convenientes de estas matrices corresponden a matrices de permutación que alteran el orden de las ecuaciones o de las incógnitas o matrices diagonales que modifican las escalas de los coeficientes.

2.11 Ejercicios

▬ EJERCICIO 14 *Calcular* $\|\mathrm{A}\|_1, \|\mathrm{A}\|_2$ *y* $\|\mathrm{A}\|_\infty$ *siendo* A *la matriz definida por*

$$\mathrm{A} = \frac{1}{27}\begin{pmatrix} -11 & 20 & 8 \\ 20 & 16 & -8 \\ 8 & -8 & -5 \end{pmatrix}$$

Solución: De la simetría de la matriz se deduce que

$$\|\mathrm{A}\|_1 = \max_{1\le j\le 3}\left\{\sum_{i=1}^{3}|a_{ij}|\right\} = \|\mathrm{A}\|_\infty = \max_{1\le i\le 3}\left\{\sum_{j=1}^{3}|a_{ij}|\right\} = \frac{44}{27}.$$

Por otra parte, se tiene que

$$\mathrm{A}^t\mathrm{A} = \mathrm{A}^2 = \frac{1}{81}\begin{pmatrix} 65 & 4 & -32 \\ 4 & 80 & 8 \\ -32 & 8 & 17 \end{pmatrix}.$$

Los autovalores de esta matriz son las raíces de su polinomio característico

$$\begin{vmatrix} \frac{65}{81}-\lambda & \frac{4}{81} & -\frac{32}{81} \\ \frac{4}{81} & \frac{80}{81}-\lambda & \frac{8}{81} \\ -\frac{32}{81} & \frac{8}{81} & \frac{17}{81}-\lambda \end{vmatrix} = -\lambda(\lambda-1)^2.$$

Consecuentemente se tiene que

$$\|\mathrm{A}\|_{=}\rho(\mathrm{A}^t\mathrm{A})^{\frac{1}{2}} = 1 \qquad \diamond$$

▬ EJERCICIO 15 *La matriz de Hilbert* H_n*, de dimensión* $n \times n$*, tiene como coeficientes*

$$h_{ij} = \frac{1}{i+j-1}$$

para $i, j = 1, 2, \cdots, n$*. Determinar el siguiente límite*

$$\lim_{n\to\infty}\|\mathrm{H}_n\|$$

usando las normas $\|\ \|_1$ *y* $\|\ \|_\infty$*.*

Solución: Puesto que la matriz de Hilbert es simétrica, el valor del límite es independiente de la norma (1 ó ∞) usada. Además, se tiene que

$$\|H_n\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} \frac{1}{i+j-1} = \sum_{i=1}^{n} \frac{1}{i}.$$

Ya que la serie armónica $\sum_{n=1}^{\infty} \frac{1}{n}$ es divergente, se cumple que

$$\lim_{n\to\infty} \|\mathrm{H}_n\| = \infty. \qquad \diamond$$

Ejercicio 16 *Se considera la matriz*

$$\mathrm{A} = \begin{pmatrix} a & 1+a \\ 0 & a \end{pmatrix}$$

donde a representa una constante real positiva. Se pide:

1. *Calcular el número de condición de A en la norma subordinada a la norma* $\|\ \|_\infty$.
2. *Estimar el error relativo de la solución del sistema lineal perturbado*

$$(\mathrm{A}+\delta\mathrm{A})(\mathbf{x}+\delta x) = \mathbf{b}+\delta\mathbf{b}$$

en términos de los errores relativos de A *y* $\mathbf{b}$.

Solución: La matriz inversa de A está dada por

$$\mathrm{A}^{-1} = \frac{1}{a}\begin{pmatrix} 1 & -1-\frac{1}{a} \\ 0 & 1 \end{pmatrix}.$$

Las normas de A y A^{-1} están dadas por

$$\|\mathrm{A}\|_\infty = 1+2a, \quad \|\mathrm{A}^{-1}\|_\infty = \frac{1}{a^2}(1+2a).$$

Consecuentemente, el número de condición de la matriz A está dado por

$$\mathrm{cond}_\infty(\mathrm{A}) = \left(\frac{1+2a}{a}\right)^2.$$

La siguiente acotación

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \left(\frac{(1+2a)^2}{a^2-(1+2a)^2\frac{\|\delta\mathrm{A}\|}{\|\mathrm{A}\|}}\right)\left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}+\frac{\|\delta\mathrm{A}\|}{\|\mathrm{A}\|}\right).$$

permite estimar el error relativo de la solución del sistema lineal perturbado para toda perturbación que verifique

$$\frac{\|\delta \mathrm{A}\|}{\|\mathrm{A}\|} < \frac{1}{\mathrm{cond}_\infty(\mathrm{A})} = \left(\frac{a}{1+2a}\right)^2. \qquad \diamond$$

▬ EJERCICIO 17 *Calcular una factorización* LU *de una permutación de la matriz*

$$\mathrm{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & \cos\alpha & -\operatorname{sen}\alpha \\ 0 & 0 & \operatorname{sen}\alpha & \cos\alpha \end{pmatrix}$$

para $0 < \alpha < \frac{\pi}{4}$*, usando la estrategia de pivote parcial.*

Solución: Las únicas matrices diferentes de la identidad que intervienen en la factorización son

$$\mathrm{P}^{(1)} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathrm{E}^{(3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\tan\alpha & 1 \end{pmatrix}.$$

Únicamente hay cambio de pivote entre la primera y segunda fila ya que $\cos\alpha > \operatorname{sen}\alpha$ para $0 < \alpha < \frac{\pi}{4}$. Por lo tanto, la factorización de Doolittle PA = LU obtenida corresponde a $\mathrm{P} = \mathrm{P}^{(1)}$ y

$$\mathrm{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \tan\alpha & 1 \end{pmatrix}, \quad \mathrm{U} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\alpha & -\operatorname{sen}\alpha \\ 0 & 0 & 0 & \frac{1}{\cos\alpha} \end{pmatrix}.$$

La factorización de Crout PA = LU corresponde a $\mathrm{P} = \mathrm{P}^{(1)}$ y

$$\mathrm{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\alpha & 0 \\ 0 & 0 & \operatorname{sen}\alpha & \frac{1}{\cos\alpha} \end{pmatrix}, \quad \mathrm{U} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -\tan\alpha \\ 0 & 0 & 0 & 1 \end{pmatrix}. \qquad \diamond$$

▬ EJERCICIO 18 *Sea* L *la matriz cuyos coeficientes son* $l_{ij} = \max\{i-j+1, 0\}$ *para todo* $i, j = 1, \cdots, n$*. Se sabe que* L *es la matriz triangular inferior que aparece en la factorización de Cholesky de una matriz* A *de* n *filas y* n *columnas, de la que únicamente se conoce que su coeficiente* a_{n2} *es 14. Determinar la dimensión de* A.

Solución: Puesto que

$$a_{n2} = \sum_{k=1}^{n} l_{nk} l_{2k} = l_{n1} l_{21} + l_{n2} l_{22} = 2n + n - 1 = 3n - 1.$$

y se conoce que $a_{n2} = 14$, se concluye que $n = 5$. $\diamond$

– EJERCICIO 19 *Determinar en qué rango de valores de α, la matriz*

$$\mathrm{A} = \begin{pmatrix} 2 & 1 & \alpha \\ 1 & 2 & 1 \\ \alpha & 1 & 2 \end{pmatrix}$$

es invertible y admite una factorización de Cholesky.

Solución: El determinante de la matriz A está dado por la siguiente expresión

$$\det \mathrm{A} = -2(\alpha + 1)(\alpha - 2).$$

Consecuentemente, A es invertible si y sólo si α es diferente de -1 y 2. Además, A es definida positiva si y sólo si $-2(\alpha + 1)(\alpha - 2) > 0$ ya que solamente en este supuesto, los menores principales son positivos. Luego, A admite una factorización de Cholesky si y sólo si $-1 < \alpha < 2$.

Un modo alternativo consiste en intentar construir la factorización y determinar cuando falla.

$$\begin{array}{lll} l_{11} = \sqrt{2} & & \\ l_{21} = \frac{\sqrt{2}}{2} & l_{22} = \sqrt{\frac{3}{2}} & \\ l_{31} = \frac{\sqrt{2}}{2}\alpha & l_{32} = \sqrt{\frac{2}{3}}\left(1 - \frac{\alpha}{2}\right) & l_{33} = \sqrt{\frac{4+2\alpha-2\alpha^2}{3}} \end{array}.$$

La factorización existe y la matriz es invertible si y sólo si $4 + 2\alpha - 2\alpha^2 > 0$. Esta última condición equivale a que $-1 < \alpha < 2$ $\diamond$.

– EJERCICIO 20 *Determinar la factorización de Cholesky de la matriz* A *de dimensión $n \times n$ cuyos únicos coeficientes que no son nulos, son*

$$\begin{array}{rll} a_{11} & = 1, & \\ a_{ii} & = 1 + i^2, & \textit{para } i = 2, 3, \cdots, n, \\ a_{i,i+1} & = a_{i+1,i} = -i, & \textit{para } i = 1, 2, \cdots, n-1 \end{array}$$

Calcular el determinante de A.

Solución: La matriz A tiene como coeficientes

$$\mathrm{A} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 5 & -2 & \cdots & 0 & 0 \\ 0 & -2 & 10 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+(n-1)^2 & -n+1 \\ 0 & 0 & 0 & \cdots & -n+1 & 1+n^2 \end{pmatrix}.$$

La primera columna de la matriz L de la factorización de Cholesky de A tiene como coeficientes no nulos los siguientes

$$l_{11} = 1, \quad l_{21} = -1.$$

Para la segunda columna se obtienen los siguientes coeficientes no nulos

$$l_{22} = 2, \quad l_{32} = -1.$$

Por inducción del resultado

$$l_{i-1,i-1} = i-1, \quad l_{i,i-1} = -1$$

se deduce que

$$l_{ii} = \sqrt{a_{ii} - l_{i,i-1}^2} = \sqrt{1+i^2-1} = i,$$

$$l_{i+1,i} = \frac{a_{i+1,i}}{l_{ii}} = -1.$$

Puesto que la matriz L está dada por

$$\mathrm{L} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 2 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & n-1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & n \end{pmatrix}$$

se obtiene que

$$\det \mathrm{A} = (\det \mathrm{L})^2 = (\mathrm{n}!)^2. \qquad \diamond$$

▬ EJERCICIO 21 *Construir una transformación de Householder que transforme el vector* $\mathbf{x} = (\cos\alpha, -\operatorname{sen}\alpha, 1, 1)^t$ *en el vector* $\mathbf{y} = (1, 1, \operatorname{sen}\alpha, \cos\alpha)^t$. *Encontrar la matriz asociada a esta transformación para* $\alpha = \frac{\pi}{2}$.

Solución: Puesto que $\|\mathbf{x}\| = \|\mathbf{y}\| = \sqrt{3}$ la transformación de Householder asociada al vector

$$\mathbf{e} = \frac{1}{\sqrt{6 - 4\cos\alpha}}(\cos\alpha - 1, -\operatorname{sen}\alpha - 1, 1 - \operatorname{sen}\alpha, 1 - \cos\alpha)^t$$

transforma el vector $\mathbf{x}$ en el vector $\mathbf{y}$. Para el caso $\alpha = \frac{\pi}{2}$ la matriz asociada es

$$\mathrm{H} = \mathrm{I} - 2\mathbf{e}\mathbf{e}^t = \begin{pmatrix} \frac{2}{3} & -\frac{2}{3} & 0 & \frac{1}{3} \\ -\frac{2}{3} & -\frac{1}{3} & 0 & \frac{2}{3} \\ 0 & 0 & 1 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & \frac{2}{3} \end{pmatrix}. \qquad \diamond$$

Ejercicio 22

Sean A *la matriz y* $\mathbf{b}$ *el vector definidos por*

$$\mathrm{A} = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 2 \\ 0 & 2 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Aplicar el método de Gram-Schmidt para orto-normalizar el conjunto de vectores $\{\mathbf{b}, \mathrm{A}\mathbf{b}, \mathrm{A}^2\mathbf{b}\}$.

Solución: Puesto que la matriz construida con los vectores $\mathbf{b}, \mathrm{A}\mathbf{b}$ y $\mathrm{A}^2\mathbf{b}$

$$\begin{pmatrix} 1 & -1 & 5 \\ -1 & 3 & -1 \\ 1 & -1 & 5 \end{pmatrix}$$

tiene determinante nulo, basta orto-normalizar $\mathbf{b}$ y $\mathrm{A}\mathbf{b}$ ya que el tercer vector será nulo. La ortogonalización conduce a

$$\mathbf{p}^1 = \mathbf{b}; \quad \mathbf{p}^2 = \mathrm{A}\mathbf{b} - \frac{\mathrm{A}\mathbf{b}\cdot\mathbf{b}}{\|\mathbf{b}\|^2}\mathbf{b} = \frac{2}{3}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}.$$

Si se divide cada uno de los vectores $\mathbf{p}^1$ y $\mathbf{p}^2$ por su norma se obtienen los vectores

$$\mathbf{e}^1 = \frac{\sqrt{3}}{3}\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \quad e^2 = \frac{\sqrt{6}}{6}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}. \qquad \diamond$$

▬ EJERCICIO 23 *Se considera el sistema lineal* $\mathrm{A}\mathbf{x} = \mathbf{b}$ *donde*

$$\mathrm{A} = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Hallar el valor de β *que haga mínimo el radio espectral de* $\mathrm{I} - \beta\mathrm{A}$. *Analizar la convergencia del método de relajación que corresponde a la elección del pre-acondicionador* $\mathrm{M} = \alpha\mathrm{I}$ *(método de Richardson) donde* α *representa un parámetro real conocido. Encontrar la elección óptima del parámetro* α.

Solución: Existe una matriz invertible P tal que $\mathrm{A} = \mathrm{PJP}^{-1}$ donde J representa la matriz de Jordan de la matriz A. Consecuentemente se tiene que

$$\mathrm{I} - \beta\mathrm{A} = \mathrm{P}(\mathrm{I} - \beta\mathrm{J})\mathrm{P}^{-1},$$

lo que prueba que λ es un autovalor de A si y sólo si $1 - \beta\lambda$ es un autovalor de $\mathrm{I} - \beta\mathrm{A}$.

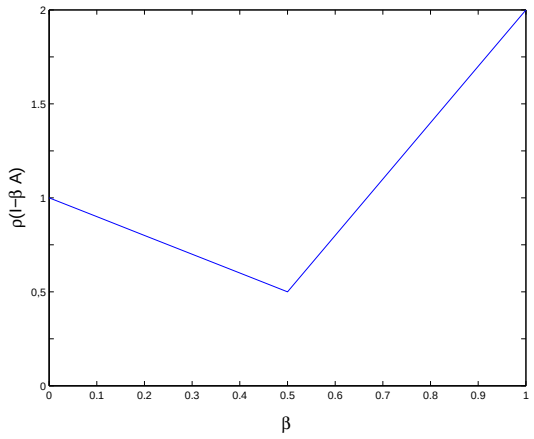


Figura 2.3: Gráfica de $\rho(I - \beta\mathrm{A})$

Los autovalores de A ordenados en orden creciente son $\lambda_1 = \lambda_2 = 1$ y $\lambda_3 = 3$. Además se tiene que

$$\rho(\mathrm{I} - \beta\mathrm{A}) = \max_{i=1,2,3} |1 - \lambda_i\beta| = \begin{cases} |1 - \beta|, & \text{si } \beta \leq \frac{1}{2}, \\ |1 - 3\beta|, & \text{si } \beta \geq \frac{1}{2}. \end{cases}$$

En el segundo tramo de la expresión anterior, para $\beta \geq \frac{1}{2}$ la función es monótona creciente y consecuentemente tiene el mínimo en $\beta = \frac{1}{2}$. De

modo similar se justifica para el primer tramo que la función es monótona decreciente y tiene el mínimo en $\beta = \frac{1}{2}$. De ello se deduce que

$$\min_{\beta} \rho(\mathrm{I} - \beta \mathrm{A}) = \frac{1}{2}$$

y el mínimo se alcanza en $\beta = \frac{1}{2}$.

El método de Richardson es convergente si y sólo si

$$\rho(\mathrm{I} - \frac{1}{\alpha}\mathrm{A}) = \max\{\left|1 - \frac{1}{\alpha}\right|, \left|1 - \frac{3}{\alpha}\right|\} < 1$$

lo que equivale a que $\alpha > \frac{3}{2}$. De la desigualdad 2.1 se desprende que en general el método será óptimo cuando el radio espectral de $\mathrm{A} - \frac{1}{\alpha}\mathrm{I}$ sea mínimo. Es decir, el valor óptimo buscado es $\alpha = 2$. ◇

▬ EJERCICIO 24 *Se considera el sistema lineal*

$$x_{i+1} = x_i + x_{i-1}, \qquad \textit{para } i = 1, 2, 3$$

$$x_0 = 1, \;\; x_4 = 5.$$

Analizar la convergencia del método de Jacobi cuando se aplica a la resolución de este sistema de ecuaciones.

Solución: En forma matricial, el sistema resulta

$$\begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 5 \end{pmatrix}$$

La matriz H de Jacobi asociada a este sistema es

$$\mathrm{H} = \mathrm{D}^{-1}(\mathrm{L} + \mathrm{U}) = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}.$$

Los autovalores de esta matriz son $\pm\sqrt{2}\mathbf{i}$. Consecuentemente, se tiene que

$$\rho(\mathrm{H}) = \sqrt{2} > 1,$$

lo que implica que el método de Jacobi no es convergente para este sistema lineal. ◇

▬ EJERCICIO 25 *Analizar la convergencia de los métodos de Jacobi, Gauss-Seidel y SOR cuando se aplican al sistema lineal:*

$$\begin{aligned} 4x_1 + x_2 &= 1 \\ x_1 + 4x_2 + x_3 &= 2 \\ &\vdots \quad \vdots \quad \vdots \\ x_{n-2} + 4x_{n-1} + x_n &= n-1 \\ x_{n-1} + 4x_n &= n. \end{aligned}$$

Solución: La matriz de coeficientes del sistema es la siguiente

$$\mathrm{A} = \begin{pmatrix} 4 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 4 & 1 & \cdots & 0 & 0 \\ 0 & 1 & 4 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 4 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 4 \end{pmatrix}.$$

La matriz es simétrica, estrictamente diagonalmente dominante por filas o columnas. Del teorema 6 de la página 49 se deduce que los métodos de Jacobi y Gauss-Seidel son convergentes. Para estudiar la convergencia del método SOR, previamente se analiza si A es una matriz definida positiva. En primer lugar se observa que el determinante en los casos $n = 1$ y $n = 2$ es positivo y crece con n. Por inducción se probará que si el determinante es positivo para $n-2$ y $n-1$ y es creciente en n hasta $n-1$, también lo será para n. Se representa por A_n el determinante de A. Si se desarrolla el determinante por la última fila, se obtiene la recurrencia

$$\mathrm{A}_n = 4\mathrm{A}_{n-1} - \mathrm{A}_{n-2}.$$

Consecuentemente, puesto $\mathrm{A}_{n-1} > \mathrm{A}_{n-2} > 0$ se concluye que $\mathrm{A}_n > 0$. esto prueba que A es definida positiva. Del teorema 8 de la página 52 se deduce que el método SOR es convergente si y sólo si $0 < \omega < 2$. ◇

CAPÍTULO 3

Aproximación de autovalores

3.1 Autovalores y vectores propios

Un aspecto de gran relevancia en la comprensión de la estructura de una aplicación lineal A es la determinación de vectores que la aplicación transforma en vectores paralelos a si mismos. Un vector propio es un vector que se transforma en un múltiplo de sí mismo y la constante de proporcionalidad del transformado de un vector propio respecto al original, se conoce como autovalor. De este modo, si λ es un autovalor real de una matriz real A, existe un vector $\mathbf{x}$ (vector propio) que no se anula, tal que $\mathrm{A}\mathbf{x} = \lambda\mathbf{x}$. Además de los autovalores reales, una matriz puede tener autovalores complejos cuando se extiende A como transformación lineal de $\mathbb{C}^n$ en $\mathbb{C}^n$.

Salvo en los casos de baja dimensión o de algunas matrices especiales, no existen fórmulas explícitas que permitan calcular de un modo directo los autovalores de una matriz que deben ser aproximados por métodos iterativos. Existen dos clases de métodos para aproximar los autovalores de una matriz:

- Un primer grupo que contempla al autovalor como una raíz de un polinomio y usa técnicas especializadas en resolución de ecuaciones escalares numéricas. La condición de autovalor se puede expresar como

$$\det\left(\mathrm{A} - \lambda\mathrm{I}\right) = 0.$$

  El primer miembro de esta ecuación es el polinomio característico de grado menor o igual que n en la variable λ, que tiene como raíces, n autovalores complejos, si cada uno se cuenta tantas veces como indica

su multiplicidad algebraica (exponente de $(\lambda - \lambda_i)$ en la factorización compleja del citado polinomio).

- Un segundo grupo que enfoca el problema como matricial y no implica directamente el concepto de polinomio característico.

Se retrasa el estudio de los métodos del primer grupo a los capítulos dedicados a la resolución de ecuaciones numéricas no-lineales y se dedica este capítulo a los métodos basados en el análisis matricial.

Si A está representada por una matriz triangular como la siguiente

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,n-1} & a_{1n} \\ 0 & a_{22} & \cdots & a_{2,n-1} & a_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & \cdots & 0 & a_{nn} \end{pmatrix}$$

los autovalores son $\{a_{11}, a_{22}, \cdots, a_{nn}\}$. Una matriz cuadrada cualquiera puede transformarse en una matriz de esta forma, mediante un cambio de base, como establece el siguiente teorema clásico del Álgebra Lineal

Teorema 10 (de Schur) *Para toda matriz real cuadrada* A *existe una matriz ortogonal* P *tal* $\mathrm{T} = \mathrm{P}^t\mathrm{AP}$ *es una matriz triangular por bloques, siendo los bloques, matrices* 1×1 *ó* 2×2*. En cada caja* 1×1*, aparece un autovalor real y en cada caja* 2×2*, aparecen dos autovalores complejos conjugados.*

Puesto que la matriz inversa de una matriz ortogonal coincide con su traspuesta, se tiene que T y $\mathrm{A} = \mathrm{PTP}^{-1}$ son semejantes y por consiguiente, tienen los mismos autovalores.

Ejemplo 14 La matriz

$$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

puede ser triangularizada por

$$\mathrm{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

para obtener

$$\mathrm{T} = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Puesto que el determinante de una matriz diagonal por bloques es igual al producto de los determinantes de cada bloque, se obtiene que los autovalores de la primera caja son $1 \pm i$ y el correspondiente a la segunda caja es 1. $\diamond$

Desafortunadamente, este teorema clarifica la estructura de las matrices pero no ayuda en el cálculo de los autovalores ya que la dificultad de calcular la descomposición de Schur es equivalente a la de calcular los autovalores.

3.2 Sucesiones de Krylov

En el capítulo anterior se establecieron algunas relaciones entre las potencias sucesivas de una matriz y sus autovalores. Para profundizar en estas relaciones se introduce el concepto de sucesión de Krylov, definida como la acción de las potencias sucesivas de una matriz sobre un vector fijado. De un modo más preciso, se llama sucesión de Krylov asociada al vector $\mathbf{x}$ y a la matriz cuadrada A, a la siguiente sucesión

$$\{\mathbf{x}, \mathrm{A}\mathbf{x}, \mathrm{A}^2\mathbf{x}, \cdots, \mathrm{A}^n\mathbf{x}, \cdots\}.$$

Sea $p_n(\lambda) = \sum_{i=0}^{n} a_i \lambda^i$ el polinomio característico de A. De acuerdo con el teorema de Cayley-Hamilton la matriz A es solución de la ecuación característica matricial. Es decir, A verifica que

$$(-1)^n \mathrm{A}^n + a_{n-1}\mathrm{A}^{n-1} + \cdots + a_1 \mathrm{A} + a_0 \mathrm{I} = \mathrm{O},$$

donde O representa la matriz cero. Si se multiplican ambos miembros por un vector arbitrario $\mathbf{x}$ se obtiene la expresión

$$a_{n-1}\mathrm{A}^{n-1}\mathbf{x} + \cdots + a_1 \mathrm{A}\mathbf{x} + a_0 \mathbf{x} = -(-1)^n \mathrm{A}^n \mathbf{x}$$

que establece que el $(n+1)-$ésimo vector de la sucesión de Krylov es linealmente dependiente de los que le preceden. Esta relación puede ser vista como un sistema lineal de incógnitas $a = (a_0, a_1, \cdots, a_{n-1})^t$

$$(\mathbf{x}|\mathrm{A}\mathbf{x}|\cdots|\mathrm{A}^{n-1}\mathbf{x})a = (-1)^{n+1}\mathrm{A}^n\mathbf{x}$$

que podría ser utilizado para calcular el polinomio característico.

– Ejercicio 26 *Aplicar el método de las sucesiones de Krylov para determinar el polinomio característico de la matriz*

$$\mathrm{A} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 2 \\ 3 & -1 & 0 \end{pmatrix}$$

partiendo del vector $\mathbf{x} = (1, 0, 0)^t$.

Solución: La sucesión de Krylov del vector $\mathbf{x}$ tiene los siguientes términos

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathrm{A}\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \quad \mathrm{A}^2\mathbf{x} = \begin{pmatrix} 4 \\ 6 \\ 3 \end{pmatrix}, \quad \mathrm{A}^3\mathbf{x} = \begin{pmatrix} 13 \\ 18 \\ 6 \end{pmatrix} \cdots$$

obtenida multiplicando sucesivamente por A el vector $\mathbf{x}$.

Consecuentemente, si se resuelve el sistema de ecuaciones

$$\begin{pmatrix} 1 & 1 & 4 \\ 0 & 0 & 6 \\ 0 & 3 & 3 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \mathrm{A}^3\mathbf{x} = \begin{pmatrix} 13 \\ 18 \\ 6 \end{pmatrix}$$

se obtiene que el polinomio característico es

$$p(\lambda) = -\lambda^3 + 3\lambda^2 - \lambda + 2. \qquad \diamond$$

Sin embargo, una dificultad podría estar en que el conjunto de los n primeros vectores de la sucesión de Krylov no fuese un conjunto de vectores independientes, en cuyo caso el sistema lineal podría ser singular. El polinomio q de grado mínimo que verifique la ecuación matricial $q(\mathrm{A}) = \mathrm{O}$ se conoce como polinomio mínimo de la matriz A. Obviamente, el grado del polinomio mínimo es menor o igual que n.

Las propiedades más relevantes del polinomio mínimo de una matriz están recogidos en el siguiente

Teorema 11 *El polinomio mínimo de una matriz* A *verifica las siguientes propiedades:*

- *El polinomio mínimo existe y es único.*
- *El polinomio mínimo es invariante frente a transformaciones de semejanza.*
- *El polinomio mínimo tiene las mismas raíces que el polinomio característico aunque posiblemente con multiplicidad inferior.*

La prueba de este teorema puede encontrarse en muchos textos básicos de Álgebra Lineal.

Se llama grado de Krylov del vector $\mathbf{x}$ respecto a la matriz A al número máximo de vectores linealmente independientes en la sucesión de Krylov asociada a $\mathbf{x}$. El grado de Krylov de cualquier vector está entre 1 (este es el caso de un vector propio) y el grado del polinomio mínimo de la matriz.

▪ EJEMPLO **15** Se considera la matriz definida por

$$\mathrm{A} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

El grado de Krylov de vector $\mathbf{x} = (-1, 0, 1)^t$ es 1 ya que $\mathrm{A}\mathbf{x} = 2\mathbf{x}$. El grado de Krylov de vector $\mathbf{x} = (1, 1, 1)^t$ es 2 ya que los vectores $\mathbf{x}$, $\mathrm{A}\mathbf{x} = (1, 0, 1)^t$ y $\mathrm{A}^2\mathbf{x} = (2, -2, 2)^t$ son linealmente dependientes. El grado de Krylov de vector $\mathbf{x} = (1, 1, -1)^t$ es 3 ya que los tres vectores $\{\mathbf{x}, \mathrm{A}\mathbf{x}, \mathrm{A}^2\mathbf{x}\}$ son linealmente independientes. Consecuentemente, los coeficientes del polinomio característico forman la solución del sistema lineal

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 6 \\ -1 & -3 & -8 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \mathrm{A}^3\mathbf{x} = \begin{pmatrix} -6 \\ 20 \\ -22 \end{pmatrix}.$$

Si se resuelve el sistema, se obtiene que el polinomio característico es el siguiente

$$p(x) = -\lambda^3 + 6\lambda^2 - 10\lambda + 4. \qquad \diamond$$

Una vez determinado el polinomio característico de una matriz, sus autovalores podrían ser calculados como las raíces de este polinomio utilizando las técnicas específicas de resolución de ecuaciones escalares numéricas. Existe la posibilidad de recorrer el camino inverso. Es decir, dado un polinomio $p(x) = \sum_{i=0}^{n} a_i x^i$ de una variable se busca una matriz que tenga este polinomio como el característico. Obviamente, esta matriz no es única ya que todas las matrices equivalentes tienen el mismo polinomio característico. Una de estas matrices es la llamada matriz de compañía del polinomio, que está definida como

$$C(p) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\frac{a_0}{a_n} & -\frac{a_1}{a_n} & -\frac{a_2}{a_n} & \cdots & -\frac{a_{n-1}}{a_n} \end{pmatrix}$$

o su traspuesta. Más precisamente, la matriz de compañía $C(p)$ tiene como polinomio característico el polinomio $(-1)^n \frac{p(x)}{a_n}$ y consecuente el polinomio p tiene como raíces, los autovalores de la matriz $C(p)$.

3.3 Método de la potencia iterada

Sea A una matriz diagonalizable mediante una matriz P y cuyos autovalores puedan ser ordenados como

$$|\lambda_1| > |\lambda_2| \geq \cdots |\lambda_n|$$

de modo que la primera desigualdad sea estricta. La sucesión de Krylov asociada a un vector $\mathbf{x}^{(0)}$ arbitrario que no se anula

$$\mathbf{x}^{(k)} = \mathrm{A}\mathbf{x}^{(k-1)} = \mathrm{A}^k\mathbf{x}^{(0)} = \mathrm{P}\mathrm{D}^k\mathrm{P}^{-1}\mathbf{x}^{(0)}$$

está formada por vectores de norma creciente hacia el infinito salvo que el radio espectral de A sea menor que 1, en cuyo caso tienden al vector 0. De modo más preciso

$$\frac{1}{\lambda_1^k}\mathrm{D}^k = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \left(\frac{\lambda_2}{\lambda_1}\right)^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(\frac{\lambda_n}{\lambda_1}\right)^k \end{pmatrix}$$

que tiene en la diagonal principal potencias cuya base tiene módulo menor o igual que 1 y consecuentemente el límite

$$\lim_{k\to\infty} \frac{1}{\lambda_1^k}\mathrm{D}^k = \mathrm{D}_0 \tag{3.1}$$

es una matriz con ceros en la diagonal principal salvo en la posición $(1,1)$. Es decir

$$\lim_{k\to\infty} \frac{\mathbf{x}^{(k)}}{\lambda_1^k} = \lim_{k\to\infty} \mathrm{P}\frac{1}{\lambda_1^k}\mathrm{D}^k\mathrm{P}^{-1}\mathbf{x}^{(0)} = \mathrm{P}\mathrm{D}_0\mathrm{P}^{-1}\mathbf{x}^{(0)} = \mathbf{y}^{(0)}.$$

Este resultado tiene en principio poco interés ya que el límite $\mathbf{y}^{(0)}$ es un vector que depende linealmente de $\mathbf{x}^{(0)}$ través de una matriz que implica la matriz P y que se supone que es uno de los objetivos de nuestro cálculo. Sin embargo, esta dependencia desaparece si se calcula el límite de los cocientes de primeras componentes del vector $\mathbf{x}^{(k)}$. En efecto, si $y_1^{(0)} \neq 0$ se tiene que

$$\lim_{k\to\infty} \frac{x_1^{(k)}}{x_1^{(k-1)}} = \lambda_1 \lim_{k\to\infty} \frac{\frac{x_1^{(k)}}{\lambda_1^k}}{\frac{x_1^{(k-1)}}{\lambda_1^{k-1}}} = \lambda_1.$$

Este límite no depende ni de $\mathbf{y}^{(0)}$ ni del vector de partida $\mathbf{x}^{(0)}$. El resultado sería aún válido si el cociente se efectúa entre las componentes $i-$ésima de los vectores de la sucesión en vez del cociente de las primeras componentes. Es importante tener en cuenta que $y_1^{(0)} = 0$ si la primera componente de $\mathrm{P}^{-1}\mathbf{x}^{(0)}$ es distinta de 0. Puesto que el primer vector columna de P es un vector propio asociado a λ_1, una condición necesaria y suficiente para que no se produzca una forma indeterminada en el límite anterior, es que $x^{(0)}$ tenga componente distinta de 0 en la dirección un vector propio asociado a λ.

▪ EJEMPLO **16** Se pretende calcular el autovalor dominante de la matriz

$$\mathrm{A} = \begin{pmatrix} 0 & -1 & 1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

mediante el método de la potencia, partiendo del vector $\mathbf{x}^{(0)} = (1, -1, 2)^t$. Los vectores iterantes son

$x^{(1)} = \mathrm{A}x^{(0)} =$	$\begin{matrix} 3 \\ -3 \\ 4 \end{matrix}$	$x^{(2)} = \mathrm{A}x^{(1)} =$	$\begin{matrix} 7 \\ -7 \\ 8 \end{matrix}$	$x^{(3)} = \mathrm{A}x^{(2)} =$	$\begin{matrix} 15 \\ -15 \\ 16 \end{matrix}$

$x^{(4)} = \mathrm{A}x^{(3)} =$	$\begin{matrix} 31 \\ -31 \\ 32 \end{matrix}$	$x^{(5)} = \mathrm{A}x^{(4)} =$	$\begin{matrix} 63 \\ -63 \\ 64 \end{matrix}$	$\begin{matrix} \cdots \\ \cdots \\ \cdots \end{matrix}$

Por consiguiente, los cocientes de componentes obtenidos son

$\frac{x_i^{(1)}}{x_i^{(0)}} =$	$\begin{matrix} 3 \\ 3 \\ 2 \end{matrix}$	$\frac{x_i^{(2)}}{x_i^{(1)}} =$	$\begin{matrix} 2,3333 \\ 2,3333 \\ 2 \end{matrix}$	$\frac{x_i^{(3)}}{x_i^{(2)}} =$	$\begin{matrix} 2,1428 \\ 2,1428 \\ 2 \end{matrix}$	$\frac{x_i^{(4)}}{x_i^{(3)}} =$	$\begin{matrix} 2,0666 \\ 2,0666 \\ 2 \end{matrix}$	$\frac{x_i^{(5)}}{x_i^{(4)}} =$	$\begin{matrix} 2,0322 \\ 2,0322 \\ 2 \end{matrix}$

Las tres sucesiones de cocientes convergen al autovalor dominante $\lambda = 2$.

El método de la potencia iterada se basa en este cálculo. La principal crítica que merece este método es que cuando el número de iteraciones es elevado, nos encontramos con el cociente de dos números o bien muy pequeños o bien muy grandes lo que produce el consiguiente deterioro del cálculo.

No obstante, esta idea puede ponerse en práctica salvando esta dificultad, si en cada iteración se normaliza el resultado con una norma $\| \ \|$

$$\mathbf{w}^{(k)} = \frac{\mathrm{A}\mathbf{w}^{(k-1)}}{\|\mathrm{A}\mathbf{w}^{(k-1)}\|}.$$

En este caso, la sucesión se mantendrá siempre en el conjunto de los vectores de norma 1. Además

$$\lim_{k\to\infty} \mathbf{w}^{(k)} = \lim_{k\to\infty} \frac{\mathrm{A}^k \mathbf{w}^{(0)}}{\|\mathrm{A}\mathbf{w}^{(0)}\|} = \frac{\mathbf{y}^{(0)}}{\|\mathbf{y}^{(0)}\|}$$

$$\lim_{k\to\infty} \mathrm{A}\mathbf{w}^{(k)} = \lim_{k\to\infty} \frac{\mathrm{A}^{k+1} \mathbf{w}^{(0)}}{\|\mathrm{A}\mathbf{w}^{(0)}\|} = \lambda_1 \frac{\mathbf{y}^{(0)}}{\|y^{(0)}\|}$$

Consecuentemente se tiene que

$$\lim_{k\to\infty} \frac{(\mathrm{A}\mathbf{w}^{(k)})_1}{w_1^{(k)}} = \lambda_1$$

Esta es la idea básica del método de la potencia iterada con normalización que permite calcular el autovalor dominante de una matriz.

La cuestión que ahora se plantea es cómo se pueden calcular los restantes autovalores. La siguiente idea da una orientación sobre el modo en que esto podría llevarse a cabo: Los autovalores de la matriz $\mathrm{A} - \alpha I$ para cualquier escalar α son $\{\lambda - \alpha : \lambda \text{ es un autovalor de A}\}$. En otras palabras, se pueden desplazar los autovalores de una matriz perturbando la matriz con un múltiplo de la identidad. Sin embargo, con un desplazamiento un autovalor real intermedio no puede ser convertido en el de máximo valor absoluto. No obstante, si se puede convertir en el de menor valor absoluto. Para calcular el autovalor de menor valor absoluto, se puede utilizar un razonamiento, en cierto modo inverso que el utilizado en el de la potencia. En efecto, si los autovalores de una matriz están ordenados como

$$|\lambda_1| \geq |\lambda_2| \geq \cdots |\lambda_{n-1}| > |\lambda_n| > 0$$

entonces los autovalores de A^{-1} están ordenados como

$$\frac{1}{|\lambda_n|} > \frac{1}{|\lambda_{n-1}|} \geq \cdots \geq \frac{1}{|\lambda_1|}.$$

En estas condiciones se puede utilizar el método de la potencia a la matriz A^{-1}. De ese modo se podría aproximar el autovalor de menor valor absoluto mediante el cociente $\frac{w_1^{(k-1)}}{x_1^{(k)}}$ asociado a la iteración

$$\mathrm{A}\mathbf{x}^{(k)} = \mathbf{w}^{(k-1)}, \qquad \mathbf{w}^{(k)} = \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|}$$

para $k > 0$. Este método se conoce como el de la potencia inversa.

El método de la potencia no tiene una validez general ya que requiere que exista un único autovalor dominante. Este no es el caso de matrices como las ortogonales cuyos autovalores tienen módulo 1. En este caso el límite que aparece en la ecuación 3.1 puede no existir. De hecho, si A es una matriz que tiene como autovalores $\lambda_1 = 1$ y $\lambda_2 = -1$, el límite de la sucesión de potencias

$$\mathrm{D}^k = \begin{pmatrix} 1 & 0 \\ 0 & (-1)^k \end{pmatrix}$$

no existe.

3.4 Método QR

¿Se puede construir una matriz triangular que *casi* sea semejante a una matriz dada?. Existen algunos métodos que permiten hacerlo. Uno de ellos, llamado método QR para el cálculo de los autovalores para una matriz A, está basado en la factorización QR como sugiere su nombre. Para ponerlo en práctica, se construye la siguiente sucesión: $\mathrm{A}^{(1)} = \mathrm{A}$. Con la factorización $\mathrm{A}^{(1)} = Q^{(1)}\mathrm{R}^{(1)}$ se construye $\mathrm{A}^{(2)} = \mathrm{R}^{(1)}\mathrm{Q}^{(1)}$. Es fundamental constatar que $\mathrm{A}^{(1)}$ y $\mathrm{A}^{(2)}$ tienen los mismos autovalores ya que son semejantes

$$\mathrm{A}^{(1)} = \mathrm{Q}^{(1)}\mathrm{A}^{(2)}\left(\mathrm{Q}^{(1)}\right)^t.$$

Reiterando el procedimiento se construye la sucesión de matrices

$$\{\mathrm{A}^{(1)}, \mathrm{A}^{(2)}, \cdots, \mathrm{A}^{(k)}, \cdots\}.$$

▬ Teorema 12 *Si los autovalores de una matriz* A *verifican que*

$$|\lambda_1| > \cdots > |\lambda_n| > 0$$

entonces la sucesión de matrices equivalentes construidas por el algoritmo QR converge a una matriz triangular superior.

▪ Ejemplo **17** Las 3 primeras iteraciones que produce el algoritmo QR para la matriz

$$\mathrm{A} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

son las siguientes

$$\mathrm{Q}^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \mathrm{R}^{(1)} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \mathrm{A}^{(2)} = \mathrm{R}^{(1)}\mathrm{Q}^{(1)} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$\mathrm{Q}^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}, \mathrm{R}^{(2)} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & \sqrt{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & \frac{\sqrt{2}}{2} \end{pmatrix}, \mathrm{A}^{(3)} = \begin{pmatrix} 1 & \sqrt{2} & 0 \\ 0 & 1,5 & 0,5 \\ 0 & 0,5 & -0,5 \end{pmatrix}$$

$$\mathrm{Q}^{(3)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{3\sqrt{10}}{10} & \frac{\sqrt{10}}{10} \\ 0 & \frac{\sqrt{10}}{10} & -\frac{3\sqrt{10}}{10} \end{pmatrix}, \mathrm{R}^{(3)} = \begin{pmatrix} 1 & \sqrt{2} & 0 \\ 0 & \frac{\sqrt{10}}{2} & \frac{\sqrt{10}}{10} \\ 0 & 0 & \frac{\sqrt{10}}{5} \end{pmatrix}, \mathrm{A}^{(4)} = \begin{pmatrix} 1 & \frac{3\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \\ 0 & 1.6 & 0.2 \\ 0 & 0.2 & -0.6 \end{pmatrix}$$

Los autovalores de A son $\{-0.618, 1, 1.618\}$.

3.5 Ejercicios

▬ EJERCICIO 27 *Determinar las primeras iteraciones generadas por el método de la potencia con normalización con norma infinito cuando se aplica a la matriz*

$$\begin{pmatrix} 0 & -1 & 1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

y se utiliza como vector inicial $x^{(0)} = (1, -1, 2)^t$.

Solución: Los datos de las dos primeras iteraciones son los siguientes

$x^{(0)} =$	$\begin{matrix}1\\-1\\2\end{matrix}$	$x^{(1)} = Aw^{(0)} =$	$\begin{matrix}1.5\\-1.5\\2\end{matrix}$	$x^{(2)} = Aw^{(1)} =$	$\begin{matrix}1.75\\-1.75\\2\end{matrix}$
$\|x^{(0)}\|_\infty =$	2	$\|x^{(1)}\|_\infty =$	2	$\|x^{(2)}\|_\infty =$	2
$w^{(0)} =$	$\begin{matrix}0.5\\-0.5\\1\end{matrix}$	$w^{(1)} =$	$\begin{matrix}0.75\\-0.75\\1\end{matrix}$	$w^{(2)} =$	$\begin{matrix}0.875\\-0.875\\1\end{matrix}$
		$\frac{Aw_1^{(0)}}{w_1^{(0)}} =$	3	$\frac{Aw_1^{(1)}}{w_1^{(1)}} =$	2.3333

◇

▬ EJERCICIO 28 *Realizar una iteración (sin utilizar decimales) con el método* QR *para el cálculo de los autovalores de la matriz*

$$\mathrm{A} = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}$$

usando el método de Householder para la factorización.

Solución: Para la eliminación en la primera columna, se transformará el vector $x = (1, 2, 2)^t$ en el vector $(3, 0, 0)^t$. La transformación de Householder necesaria para ello es la siguiente

$$\mathrm{P}(1) = \mathrm{I} - \frac{2}{12}\begin{pmatrix} -2 \\ 2 \\ 2 \end{pmatrix}(-2, 2, 2) = \frac{1}{3}\begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ 2 & -2 & 1 \end{pmatrix}$$

A continuación se elminan los ceros de la primera columna mediante el producto

$$\mathrm{R} = \mathrm{P}(1)\mathrm{A} = \begin{pmatrix} 3 & 1 & \frac{2}{3} \\ 0 & 1 & -\frac{2}{3} \\ 0 & 0 & \frac{1}{3} \end{pmatrix}.$$

Como R es ya triangular no es necesario realizar más eliminaciones y se toma $\mathrm{Q} = \mathrm{P}(1)$. De este modo se obtiene la siguiente factorización QR

$$\mathrm{A} = \mathrm{QR} = \frac{1}{3}\begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ 2 & -2 & 1 \end{pmatrix}\begin{pmatrix} 3 & 1 & \frac{2}{3} \\ 0 & 1 & -\frac{2}{3} \\ 0 & 0 & \frac{1}{3} \end{pmatrix}.$$

Así pues, la primera iteración en el método QR conduce a la siguiente matriz

$$\mathrm{A}(2) = \mathrm{RQ} = \frac{1}{9}\begin{pmatrix} 19 & 17 & 14 \\ 2 & 7 & -8 \\ 2 & -2 & 1 \end{pmatrix}. \qquad \diamond$$

▬ Ejercicio 29 *Sea* A *la matriz definida por*

$$\mathrm{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\operatorname{sen}\theta \\ 0 & \operatorname{sen}\theta & \cos\theta \end{pmatrix}$$

donde θ *es un ángulo dado. Determinar el término n-ésimo de la sucesión de Krylov asociada a la matriz* A *y al vector* $\mathbf{x} = (1, \cos\theta, -\operatorname{sen}\theta)^t$. *Por medio de esta sucesión de Krylov determinar el polinomio característico de* A.

Solución: En primer lugar, se calcula la acción de las potencias sucesivas de A sobre el vector x

$$\begin{aligned} \mathrm{A}\mathbf{x} &= \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \\ \mathrm{A}^2\mathbf{x} &= \mathrm{A}(\mathrm{A}\mathbf{x}) = \begin{pmatrix} 1 \\ \cos\theta \\ \operatorname{sen}\theta \end{pmatrix}, \\ &\dots \\ \mathrm{A}^n\mathbf{x} &= \mathrm{A}(\mathrm{A}^{n-1}\mathbf{x}) = \begin{pmatrix} 1 \\ \cos(n-1)\theta \\ \operatorname{sen}(n-1)\theta \end{pmatrix} \end{aligned}$$

Los coeficientes del polinomio característico son soluciones del siguiente sistema lineal

$$\begin{pmatrix} 1 & 1 & 1 \\ \cos\theta & 1 & \cos\theta \\ -\operatorname{sen}\theta & 0 & \operatorname{sen}\theta \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 \\ \cos 2\theta \\ \operatorname{sen} 2\theta \end{pmatrix}.$$

Si se resuelve el sistema se obtiene que el polinomio característico buscado es

$$p(\lambda) = -\lambda^3 + (1 + 2\cos\theta)\lambda^2 - (1 + 2\cos\theta)\lambda + 1. \qquad \diamond$$

▁ Ejercicio 30 *Determinar las raíces del polinomio*

$$p(x) = x^3 + 3x^2 + 2x$$

usando el método de la potencia iterada con $\infty-$normalización en la matriz de compañía de p partiendo del punto $(1,1,1)^t$.

Solución: La matriz de compañía del polinomio es la siguiente

$$C(p) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -2 & -3 \end{pmatrix}.$$

Los datos de las dos primeras iteraciones son los siguientes

$x^{(0)} =$	$\begin{matrix}1\\1\\1\end{matrix}$	$x^{(1)} = Aw^{(0)} =$	$\begin{matrix}1\\1\\-5\end{matrix}$	$x^{(2)} = Aw^{(1)} =$	$\begin{matrix}\frac{1}{5}\\-1\\\frac{13}{5}\end{matrix}$
$\|x^{(0)}\|_\infty =$	1	$\|x^{(1)}\|_\infty =$	5	$\|x^{(2)}\|_\infty =$	$\frac{13}{5}$
$w^{(0)} =$	$\begin{matrix}1\\1\\1\end{matrix}$	$w^{(1)} =$	$\begin{matrix}\frac{1}{5}\\\frac{1}{5}\\-1\end{matrix}$	$w^{(2)} =$	$\begin{matrix}\frac{1}{13}\\\frac{-5}{13}\\1\end{matrix}$
		$\frac{(Aw^{(0)})_1}{(w^{(0)})_1} =$	1	$\frac{(Aw^{(1)})_1}{(w^{(1)})_1} =$	1

.

Los valores obtenidos $\{1, 1\}$, son solamente los dos primeros términos de una sucesión que debe converger al autovalor $\lambda = -2$. Para entender como esta sucesión se va produciendo sería necesario calcular algunos términos más. ◇

Ejercicio 31 *Si α es una constante , aproximar el autovalor de módulo máximo de la matriz*

$$\mathrm{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2\cos^2\alpha & 2\cos\alpha\,\mathrm{sen}\,\alpha \\ 0 & 2\cos\alpha\,\mathrm{sen}\,\alpha & -2\sin^2\alpha \end{pmatrix}$$

usando el método de la potencia (sin normalización) tomando como vector inicial $\mathbf{v}^{(0)} = (1, 1, 1)^t$, probando con las diferentes componentes.

Solución: Sin dificultad se prueba que las potencias sucesivas de A son las siguientes matrices:

$$\mathrm{A}^n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (-2)^n\cos^2\alpha & -(-2)^n\,\mathrm{sen}\,\alpha\cos\alpha \\ 0 & -(-2)^n\,\mathrm{sen}\,\alpha\cos\alpha & (-2)^n\,\mathrm{sen}^2\,\alpha \end{pmatrix}.$$

De acuerdo con el método de la potencia se obtiene la siguiente sucesión de vectores

$$\mathbf{v}^{(n)} = \begin{pmatrix} 1 \\ (-2)^n(\cos^2\alpha - \mathrm{sen}\,\alpha\cos\alpha) \\ (-2)^n(\mathrm{sen}^2\,\alpha - \mathrm{sen}\,\alpha\cos\alpha) \end{pmatrix}$$

Obviamente, las sucesiones generadas por los cocientes de las componentes son las siguientes:

$$\{1, 1, 1, \cdots\},$$
$$\{-2(\cos^2\alpha - \operatorname{sen}\alpha\cos\alpha), -2, -2, \cdots\},$$
$$\{-2(\operatorname{sen}^2\alpha - \operatorname{sen}\alpha\cos\alpha), -2, -2, \cdots\}.$$

La razón de que la primera sucesión no converja al autovalor de módulo máximo es que el vector $(1, 0, 0)$ es un vector propio asociado al autovalor $\lambda = 1$. Fácilmente se comprueba que los autovalores de la matriz A son $\{-2, 0, 1\}$. $\diamond$

CAPÍTULO 4

Aproximación de funciones

4.1 Introducción

Cuando se quiere evaluar una función de cierta complejidad, es conveniente pensar que el cálculo automático que realiza un computador solamente emplea las operaciones aritméticas básicas junto con las más simple operaciones de comparación de números en el rango de la máquina. Esta limitación implica que en estas circunstancias, debe introducirse algún tipo de aproximación por funciones simples que puedan ser programadas directamente mediante un número finito de instrucciones. En general, la respuesta más adecuada a estas dificultades es la aproximación de la función cuya evaluación en un punto dado no sea sencilla, por funciones cuya evaluación únicamente requiera operaciones aritméticas básicas, tales como los polinomios o las funciones racionales. De este modo se puede representar la función en un computador con mínima información y se puede reducir el coste computacional de su evaluación en un punto. Evidentemente, el cómputo del valor de una función analítica en un punto como la suma de una serie numérica infinita, no puede ejecutarse de un modo real salvo que se incluya algún criterio de parada que de por válida la precisión obtenida en una etapa del proceso de suma.

Con este objetivo, los primeros propósitos en este capítulo son los siguientes:

1. Analizar el modo más sencillo de evaluar un polinomio.
2. Precisar el sentido que se pretende dar en cada momento a una aproximación ya que no existe un único modo de entender lo que se quiere

decir cuando se usa este término.

4.2 Evaluación de polinomios

La evaluación por sustitución directa de la variable x en el polinomio

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

requiere

- $(1+2+\cdots+n-1) = \frac{n(n-1)}{2}$ multiplicaciones para calcular las potencias.
- n multiplicaciones de las potencias calculadas por los coeficientes.
- n sumas.

En total, requiere $\frac{n^2+n}{2}$ multiplicaciones $+n$ sumas.

Sin embargo, los polinomios admiten la siguiente representación anidada

$$p(x) = a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + a_n x)\cdots)),$$

que podría ser utilizada para una evaluación más eficiente.

Si se representa por q_i el polinomio definido por

$$q_i(x) = a_{i-1} + a_i x$$

para $i = 1, \cdots, n$, se puede evaluar el polinomio p como

$$p(x) = q_1(q_2(\cdots(q_n(x)))$$

para cualquier valor de x.

Este modo de evaluar iterativamente las potencias, que se conoce como algoritmo de Horner, permite reducir el número de operaciones a $n-1$ multiplicaciones y $n-1$ sumas. Se puede probar que para la evaluación directa de un polinomio, este procedimiento es óptimo en el sentido de que no es posible encontrar otro procedimiento que implique menos multiplicaciones. Se sugieren los libros de Knuth [7] para un estudio más profundo de las cuestiones relacionadas con la complejidad computacional de los algoritmos.

Tradicionalmente se llama regla de Ruffini al uso de la representación anidada en forma de cuadro

a_n	a_{n-1}	a_{n-2}	$\cdots$	a_0
	xa_n	$x(a_{n-1}+xa_n)$	$\cdots$	$x(a_1+x(a_2+\cdots+x(a_{n-1}+a_nx)))$
a_n	$a_{n-1}+xa_n$	$a_{n-2}+x(a_{n-1}+xa_n)$	$\cdots$	$p(x)$

que se utiliza para ilustrar el uso de la evaluación de un polinomio mediante representación anidada.

▪ Ejemplo **18** Para evaluar el polinomio $p(x) = x^3 - x^2 + 2x - 5$ en el punto $x = 1$ se puede utilizar la tabla

	1	-1	2	-5
1		1	0	2
	1	0	2	-3

◇

4.3 Aproximación de funciones

Sea V un espacio vectorial de funciones de una variable real, dotado de una norma $\|\ \|$. Es decir, el espacio V tiene asociada una aplicación $\|\ \| : V \to \mathbb{R}^+$ que verifica las siguientes propiedades

1. $\|f\| = 0 \Leftrightarrow f = 0$,
2. $\|\lambda f\| = |\lambda| \|f\|$,
3. $\|f + g\| \leq \|f\| + \|g\|$

para todas las elecciones de $\lambda \in \mathbb{R}$ y $f, g \in V$. Si se fija un subespacio vectorial $U \subset V$ de dimensión finita, se puede considerar el siguiente problema de aproximación

Dada una función $f \in V$ arbitraria, hallar $g \in U$ tal que

$$\|f - g\| \leq \|f - h\|$$

para toda función $h \in U$. Una función g que minimiza la distancia a f, se denomina mejor aproximación ó proyección de f en U con la norma $\|\ \|$.

En muchas ocasiones en este capítulo, el espacio de funciones V considerado será el espacio de las funciones continuas en un intervalo $[a, b]$. En este

espacio, uno de los modos más simples de medir la distancia entre elementos es mediante la siguiente norma

$$\| f \|_2 = \left(\int_a^b f(x)^2 dx \right)^{\frac{1}{2}}.$$

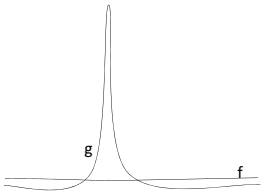
Más general es la norma$\| \|_p$ definida por

$$\| f \|_p = \left(\int_a^b |f(x)|^p \right)^{\frac{1}{p}}.$$

Como caso límite, la norma$\| \|_\infty$ está definida por

$$\| f \|_\infty = \max_{x \in [a,b]} |f(x)|.$$

Típicamente, la aproximación con la norma $\| \|_2$ se llama de mínimos cuadrados mientras que la aproximación con la norma $\| \|_\infty$ se llama aproximación uniforme o min-max.



En la figura se muestran las gráficas de dos funciones f y g que parecen próximas con la norma $\| \|_2$ pero no con la norma $\| \|_\infty$.

La mejor aproximación a una función dada puede ser bastante diferente, dependiendo de la norma usada ya que normas distintas pueden expresar diferentes criterios para medir la proximidad entre dos funciones.

▬ EJERCICIO 32 *Dadas las funciones* $g_1(x) = \operatorname{sen} x$ *y* $g_2(x) = 1 - \cos x$ *definidas sobre el intervalo* $[0, 2\pi]$*, determinar cuál de ellas es la más próxima a la función* $f(x) = x$ *de acuerdo a cada una de las siguientes normas:*

1. *La norma uniforme* $\| \|_\infty$*,*

2. *La norma cuadrado-integral* $\| \|_2$*.*

Solución: Puesto que

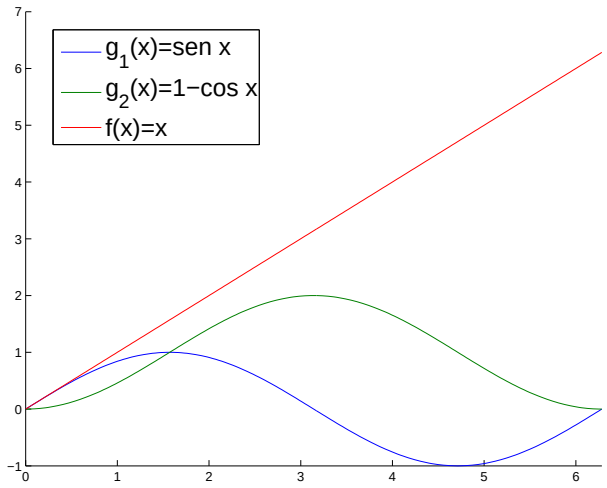


Figura 4.1: Proximidad y normas

$$\begin{aligned}
\|f-g_1\|_2 &= \left(\int_0^{2\pi}(\operatorname{sen} x - x)^2\,dx\right)^{\frac{1}{2}} = \left(5\pi + \frac{8}{3}\pi^3\right)^{\frac{1}{2}} \approx 9.91,\\
\|f-g_2\|_2 &= \left(\int_0^{2\pi}(1-\cos x - x)^2\,dx\right)^{\frac{1}{2}}\\
&= \left(\frac{\pi\,(8\,\pi^2 - 12\,\pi + 9)}{3}\right)^{\frac{1}{2}} \approx 7.25,\\
\|f-g_1\|_\infty &= \max_{0\le x\le 2\pi}|\operatorname{sen} x - x| = 2\pi,\\
\|f-g_2\|_\infty &= \max_{0\le x\le 2\pi}|1-\cos x - x| = 2\pi,
\end{aligned}$$

de estos cálculos se deduce que g_2 está más próxima a f que g_1 en la norma integral pero ambas están a la misma distancia con la norma uniforme. ◇

En relación con la aproximación de funciones, es esencial el siguiente resultado

▬ Teorema 13 *Sea V un espacio normado de funciones. Para toda función $f \in V$ y todo subespacio vectorial $U \subset V$ de dimensión finita, existe al menos una función $g \in U$ que realiza la mejor aproximación a f en U. Además, si la norma es estrictamente convexa, es decir, si se verifica la desigualdad*

$$\|tf_1 + (1-t)f_2\| < t\|f_1\| + (1-t)\|f_2\| = \|f_1\| = \|f_2\|$$

para todas las funciones $f_1, f_2 \in V$ tales $\|f_1\| = \|f_2\|$ y los escalares $t \in (0,1)$, entonces existe una única función que realiza la mejor aproximación.

Demostración: Si $\{\phi_0, \phi_1, \cdots, \phi_n\}$ es una base de U, el problema de encontrar la mejor aproximación se puede expresar como un problema de optimización en $\mathbb{R}^{n+1}$

$$\text{Hallar} \quad \min_{\boldsymbol{\alpha} \in \mathbb{R}^{n+1}} J(\boldsymbol{\alpha})$$

donde

$$J(\boldsymbol{\alpha}) = \|f - \sum_{0}^{n} \alpha_i \phi_i\|^2.$$

El número real J_m definido por

$$J_m = \inf_{\boldsymbol{\alpha} \in \mathbb{R}^{n+1}} J(\boldsymbol{\alpha})$$

siempre existe ya que J está acotada inferiormente por 0. Para cada entero positivo k se escoge un vector $\boldsymbol{\alpha}^k \in \mathbb{R}^{n+1}$ tal que

$$J(\boldsymbol{\alpha}^k) \leq J_m + \frac{1}{k}. \tag{4.1}$$

Si no existiese $\boldsymbol{\alpha}^k$ para algún k, J_m no podría ser el ínfimo de J ya que en este caso $J(\boldsymbol{\alpha}) > J_m + \frac{1}{k}$ para todo $\boldsymbol{\alpha} \in \mathbb{R}^{n+1}$. Es habitual en este tipo de argumentaciones, referirse a $\{\boldsymbol{\alpha}^k\}$ como una sucesión minimizante de J.

El conjunto A definido por

$$A = \{\boldsymbol{\beta} \in \mathbb{R}^{n+1} : J(\boldsymbol{\beta}) \leq J_m + 1\}.$$

es cerrado ya que la función J es continua. Además, es fácil verificar que la función J verifica la siguiente propiedad de crecimiento en el infinito

$$\lim_{\|\boldsymbol{\alpha}\| \to \infty} J(\boldsymbol{\alpha}) = \infty.$$

De esta propiedad se deduce que A es acotado. De ello se desprende que la sucesión minimizante admite al menos un punto límite $\boldsymbol{\alpha}$ y además existe una subsucesión $\{\boldsymbol{\alpha}^{k_p} : p = 1, 2, \cdots\}$ tal que $\lim_{p \to \infty} \boldsymbol{\alpha}^{k_p} = \boldsymbol{\alpha}$. Puesto que J es una función continua si se toman límites en la desigualdad 4.1, se obtiene que

$$J(\boldsymbol{\alpha}) \leq J_m$$

y consecuentemente que $J(\boldsymbol{\alpha}) = J_m$. Esto prueba que la función $g = \sum_{i=0}^{n} \alpha_i \phi_i$ realiza la mejor aproximación a f.

Si la norma es estrictamente convexa y existen dos funciones $g_1, g_2 \in U$ distintas que realizan la mejor aproximación a f, puesto que

$$\|f - g_1\| = \|f - g_2\| = \min_{g \in U} \|f - g\|$$

entonces

$$\begin{aligned} \|f - \frac{1}{2}(g_1 + g_2)\| &= \|\frac{1}{2}(f - g_1) + \frac{1}{2}(f - g_2)\| \\ &< \frac{1}{2}\|f - g_1\| + \frac{1}{2}\|f - g_2\| \end{aligned}$$

lo cual es contradictorio con el hecho de que g_1 y g_2 realicen la mejor aproximación a la función f. $\diamond$

En la mayoría de las situaciones particulares consideradas en este capítulo, V será el espacio de las funciones continuas en un intervalo cerrado y acotado y el subespacio vectorial U será el espacio de los polinomios de una variable de grado menor ó igual que n, para un entero n fijado previamente. Se usará la notación $\mathcal{P}_n$ para representar este subespacio de dimensión $n + 1$.

4.4 Aproximación por mínimos cuadrados

El concepto de producto escalar generaliza el concepto de producto escalar euclídeo y es particularmente útil para estudiar la mejor aproximación a una función cuando se usan normas que provienen de él. Se define como una aplicación $\langle\ ,\ \rangle : V \times V \to \mathbb{R}^+$ que verifica las siguientes propiedades

1. $\langle f, f\rangle = 0 \Leftrightarrow f = 0$,
2. $\langle f, g\rangle = \langle g, f\rangle$,
3. $\langle \alpha f + \mu g, h\rangle = \alpha \langle f, h\rangle + \mu \langle g, h\rangle$

para toda elección de $\lambda, \mu \in \mathbb{R}$ y $f, g, h \in V$.

Todo producto escalar permite definir una norma mediante la expresión

$$\|f\| = \langle f, f\rangle^{\frac{1}{2}}$$

para todo $f \in V$. En particular, el producto escalar

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

define la norma $\|\ \|_2$.

No es la intención de este curso indagar exhaustivamente sobre las propiedades de productos escalares y normas, que uno puede consultar en un

curso básico de Análisis Matemático. Únicamente, se destacará la relevante desigualdad de Cauchy-Schwarz

$$| \langle f, g \rangle | \leq \|f\| \ \|g\|$$

para todas las funciones $f, g \in V$. La igualdad se alcanza si y sólo si existen escalares λ y μ, tales que $\lambda f + \mu g = 0$ y $|\lambda| + |\mu| > 0$.

La norma definida por un producto escalar es estrictamente convexa. En efecto, si f y g son distintos y tienen norma igual a 1 y además no son linealmente dependientes (si lo son $f = -g$) entonces se cumple que

$$\begin{aligned} \|\lambda f + (1-\lambda)g\|^2 &= \langle \lambda f + (1-\lambda)g, \lambda f + (1-\lambda)g \rangle \\ &= \lambda^2 + (1-\lambda)^2 + 2\lambda(1-\lambda) \langle f, g \rangle . \end{aligned}$$

Si se usa la desigualdad de Cauchy-Schwarz estricta se obtiene que

$$\|\lambda f + (1-\lambda)g\|^2 < \lambda^2 + (1-\lambda)^2 + 2\lambda(1-\lambda) = 1.$$

Si $g = -f$ la desigualdad se prueba directamente. El razonamiento se extiende sin dificultad al caso en el que f y g no son unitarios aunque tengan la misma norma. Consecuentemente, toda norma asociada a un producto escalar es estrictamente convexa y el problema de mejor aproximación en un espacio de dimensión finita siempre tiene una única solución.

El vector $\boldsymbol{\alpha}$ que realiza el mínimo de J, anula su gradiente. Es decir, se cumple que

$$\frac{\partial J}{\partial \alpha_j} = 2 \left\langle f - \sum_0^n \alpha_i \phi_i, \phi_j \right\rangle = 0$$

para $j = 0, 1, \cdots, n$.

Estas relaciones pueden organizarse en forma de lo que se conoce como las ecuaciones normales del problema de aproximación

$$G\boldsymbol{\alpha} = \bar{\mathbf{f}}$$

donde $\bar{\mathbf{f}}$ es el vector de componentes $\bar{f}_i = \langle f, \phi_i \rangle$ y G la matriz de Gram definida por

$$G_{ij} = \langle \phi_i, \phi_j \rangle$$

para $i, j = 0, 1, \cdots, n$.

Fácilmente se comprueba que la matriz de Gram asociada a la base, es simétrica y definida positiva. En efecto, G es definida positiva ya que

$$\boldsymbol{\alpha}^t G \boldsymbol{\alpha} = \left\langle \sum_0^n \alpha_i \phi_i, \sum_0^n \alpha_i \phi_i \right\rangle = \| \sum_0^n \alpha_i \phi_i \|^2 > 0$$

si $\boldsymbol{\alpha} \neq \mathbf{0}$.

En definitiva, se puede encontrar la solución al problema de mejor aproximación con una norma que proviene de un producto escalar, resolviendo un sistema de ecuaciones lineales.

▬ EJERCICIO 33 *Determinar la mejor aproximación a la función $f(x) = e^x$ en el subespacio $U = \mathcal{P}_2$, usando la norma asociada al siguiente producto escalar*

$$\langle f, g\rangle = \int_0^1 \left(f(x)g(x) + f'(x)g'(x)\right) dx$$

en el espacio V de las funciones de clase $C^1([0,1])$.

Solución: La matriz de Gram de la base $\{1, x, x^2\}$ de $\mathcal{P}_2$, es

$$G = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{4}{3} & \frac{5}{4} \\ \frac{1}{3} & \frac{5}{4} & \frac{23}{15} \end{pmatrix}$$

el vector de términos independientes por

$$\bar{\mathbf{f}} = \begin{pmatrix} e-1 \\ e \\ e \end{pmatrix}.$$

Se han calculado los coeficientes de esta matriz y de este vector, usando directamente la definición. Por ejemplo, el coeficiente $(2,3)$ de la matriz de Gram se ha obtenido como

$$G_{23} = \int_0^1 \left(x^3 + 2x\right) dx = \frac{5}{4}.$$

Si se resuelve el sistema lineal $G\boldsymbol{\alpha} = \bar{f}$ se obtiene que la mejor aproximación es

$$p(x) = \frac{675e - 1041 + (756 - 24e)x + 390(e-1)x^2}{793}. \qquad \diamond$$

4.5 Aproximación discreta por mínimos cuadrados

En determinadas circunstancias, el interés por la proximidad de dos funciones puede reducirse a observarla en un conjunto relevante de puntos y no en todo el intervalo. Esta proximidad recortada puede medirse utilizando

seminormas. Este término se refiere a una aplicación que cumple la segunda y la tercera condición de norma pero no necesariamente la primera.

Se considera un conjunto finito de puntos (nodos)

$$x_0 < x_1 < \cdots < x_{n-1} < x_n$$

en el intervalo de interés. Con ayuda de estos puntos y las normas usadas en las secciones anteriores, se pueden definir seminormas para medir la proximidad discreta entre dos funciones. Así, relacionada con la norma $\| \ \|_2$, se puede definir la seminorma

$$|f|_2 = \left(\sum_{i=0}^{n} f(x_i)^2 \right)^{\frac{1}{2}}.$$

Más general es la seminorma p definida por

$$|f|_p = \left(\sum_{i=0}^{n} |f(x_i)|^p \right)^{\frac{1}{p}}.$$

Como caso límite, la semi-norma ∞ discreta está definida por

$$|f|_\infty = \max_{0 \le i \le n} |f(x_i)|.$$

Se comprende bien que estas seminormas no verifican la primera propiedad que se exigía a las normas. Una función que se anule en todos los puntos del conjunto $\{x_i : i = 0, 1, \cdots, n\}$, no tiene forzosamente que ser nula en los demás puntos y sin embargo, el valor de su seminorma es 0. Es frecuente referirse a la mejor aproximación con las seminormas que se han introducido previamente, como aproximación discreta.

El hecho de que los conjuntos de nivel de una seminorma no son necesariamente compactos, impide argumentar de mismo modo que en la sección anterior cuando se pretende establecer resultados de existencia de mejor aproximación discreta. Es decir, aunque se pruebe la existencia de una sucesión minimizante, acotada con una seminorma, no se puede concluir la existencia de una subsucesión convergente a una función que realice el mínimo.

▪ Ejemplo **19** Se considera la seminorma

$$|f|_2 = \left(\sum_{i=1}^{3} f(i)^2 \right)^{\frac{1}{2}}$$

construida sobre los puntos $\{1, 2, 3\} \subset I = [0, 4]$ de observación de la función y el subespacio $U = \mathcal{P}_3$. La sucesión $\{f_n\}$ definida por

$$f_n(x) = n(x-1)(x-2)(x-3)$$

verifica que $|f_n|_2 = 0$ para todo n. La sucesión $\{f_n\}$ está acotada con la seminorma pero no es posible encontrar una subsucesión convergente en esa seminorma en U.

Sin embargo, si se restringe el subespacio U a $\mathcal{P}_2$, la seminorma considerada es una verdadera norma en U ya que si un polinomio de grado menor o igual que 2 se anulase en 3 puntos distintos entonces sería necesariamente el polinomio 0. En ese caso, el argumento de compacidad podría utilizarse. ◇

Si una seminorma definida en el espacio de las funciones continuas es una verdadera norma en el subespacio U en el que se busca la aproximación, aunque no lo sea en todo el espacio, los razonamientos del teorema 13 de la página 83 siguen siendo válidos y la existencia de mejor aproximación está garantizada. Si la seminorma proviene de un producto escalar degenerado (que puede que no cumpla la primera condición de producto escalar), la mejor aproximación es única, las ecuaciones normales son válidas y la matriz de Gram no es singular si y sólo si la seminorma considerada es una norma en U.

▬ EJERCICIO 34 *Obtener la mejor aproximación mediante un polinomio de grado menor o igual que* 1*, de la función*

$$f(x) = \frac{1}{2}\cos \pi x + \frac{1}{3}\operatorname{sen} \frac{\pi}{2} x.$$

por mínimos cuadrados en $I = [-1, 1]$ *y la mejor aproximación discreta por mínimos cuadrados en* $\{-1, 0, 1\}$.

Solución: Se considera la base $\{1, x\}$ del espacio $\mathcal{P}_1$. El sistema lineal que determina la mejor aproximación es

$$\begin{pmatrix} \langle 1,1\rangle & \langle 1,x\rangle \\ \langle x,1\rangle & \langle x,x\rangle \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \langle f,1\rangle \\ \langle f,x\rangle \end{pmatrix}$$

donde $\langle\ ,\ \rangle$ representa alternativamente a los productos

$$\begin{aligned} \langle f,g\rangle &= \int_{-1}^{1} f(x)g(x)\ dx, \\ \langle f,g\rangle &= f(-1)g(-1) + f(0)g(0) + f(1)g(1) \end{aligned}$$

Las ecuaciones normales que resultan son, en el primer caso

$$\begin{pmatrix} 2 & 0 \\ 0 & \frac{2}{3} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{8}{3\pi^2} \end{pmatrix},$$

y en el segundo

$$\begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ \frac{2}{3} \end{pmatrix}.$$

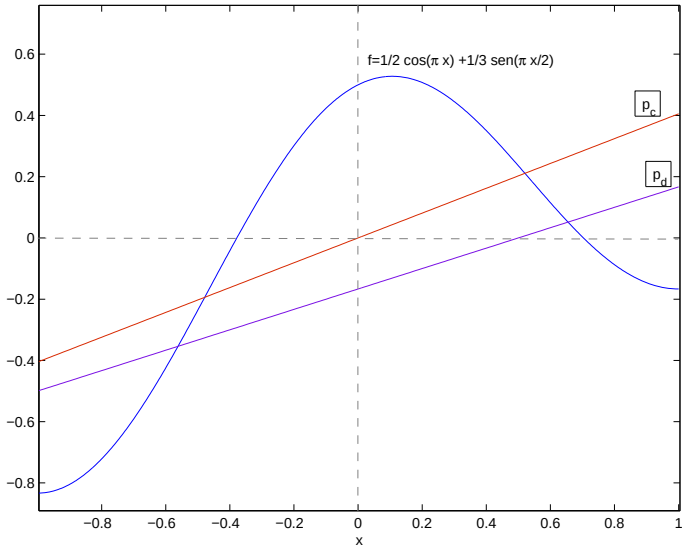


Figura 4.2: Aproximación por mínimos cuadrados continua y discreta

En consecuencia, las mejores aproximaciones son en cada uno de los casos, las siguientes (véase figura 4.2):

$$p_c(x) = \frac{4x}{\pi^2}, \quad p_d(x) = -\frac{1}{6} + \frac{x}{3}. \qquad \diamond$$

▪ EJEMPLO **20** *(Regresión lineal).* Con los datos $\{(x_i, y_i) : i = 0, 1, \cdots, n\}$ se busca la recta de regresión $y = mx+b$ definida como la mejor aproximación a una función $y = y(x)$ tal que $y_i = y(x_i)$, en el sentido de mínimos cuadrados discretos. Con el producto

$$\langle f, g \rangle = \sum_{i=0}^{n} f(x_i) g(x_i),$$

se calcula la matriz de Gram y el vector independiente de las ecuaciones normales asociados a la base $\{1, x\}$, obteniéndose

$$G = \begin{pmatrix} \langle 1,1\rangle & \langle 1,x\rangle \\ \langle x,1\rangle & \langle x,x\rangle \end{pmatrix} = \begin{pmatrix} n+1 & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i & \sum_{i=0}^n x_i^2 \end{pmatrix},$$

$$\bar{\mathbf{f}} = \begin{pmatrix} \langle 1,y\rangle \\ \langle x,y\rangle \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^n y_i \\ \sum_{i=0}^n x_i y_i \end{pmatrix}.$$

Si se resuelve este sistema se obtienen las conocidas fórmulas de la regresión lineal

$$m = \frac{\sum_{i=0}^n x_i \sum_{i=0}^n y_i - (n+1)\sum_{i=0}^n x_i y_i}{\left(\sum_{i=0}^n x_i\right)^2 - (n+1)\sum_{i=0}^n x_i^2} = \frac{\sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^n (x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_x^2},$$

$$b = \bar{y} - m\bar{x}.$$

donde $\bar{x}$ y $\bar{y}$ representan las medias de x e y respectivamente, σ_x la desviación típica de x y σ_{xy} la covarianza de x e y. $\diamond$

▪ Ejemplo **21** *(Aproximación de Taylor).* Sea V el espacio vectorial de las funciones indefinidamente diferenciables y $U = \mathcal{P}_n$. Se considera un punto x_0, alrededor del cual se desea que la aproximación sea precisa. A este punto se le asocia el siguiente producto escalar degenerado

$$\langle f, g\rangle = \sum_{i=0}^n f^{(i)}(x_0) g^{(i)}(x_0).$$

Para la base $\{1, x - x_0, \cdots, (x - x_0)^n\}$ de $\mathcal{P}_n$, se construyen las ecuaciones normales de la aproximación. Puesto que

$$\frac{d^i}{dx^i}(x - x_0)^j|_{x=x_0} = \begin{cases} j!, & \text{si } j = i, \\ 0, & \text{si } j \neq i, \end{cases}$$

se tiene que

$$\left\langle (x - x_0)^j, (x - x_0)^k \right\rangle = \begin{cases} j!k!, & \text{si } j = k \leq n \\ 0, & \text{en otro caso.} \end{cases}$$

De modo similar se obtiene que

$$\left\langle (x - x_0)^j, f \right\rangle = j! f^{(j)}(x_0).$$

En consecuencia las ecuaciones normales asociadas al problema de aproximación de Taylor son

$$\begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 2! & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & n! \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f'(x_0) \\ f''(x_0) \\ \vdots \\ f^{(n)}(x_0) \end{pmatrix}.$$

Finalmente, se deduce que la mejor aproximación de Taylor de f es

$$p_n(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n. \qquad \diamond$$

▪ Ejemplo **22** Se considera la base $\{1, x, \cdots, x^n\}$ de $\mathcal{P}_n$ en el intervalo $[-1, 1]$. La matriz de Gram de esta base con respecto al producto escalar definido por

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

está dada por

$$g_{ij} = \langle x^i, x^j \rangle = \int_{-1}^{1} x^{i+j}dx = \frac{1}{i+j+1} \begin{cases} 0, & \text{si } i+j \text{ impar} \\ 2, & \text{en otro caso.} \end{cases}$$

Esta matriz, que se conoce como la matriz de Hilbert. Desafortunadamente está mal condicionada y para dimensión elevada, la resolución del sistema lineal puede resultar afectada por la presencia de errores de redondeo. $\diamond$

La resolución de las ecuaciones normales se vuelve particularmente simple si la base es ortonormal, es decir, si $\langle \phi_i, \phi_j \rangle = 0$ para $i \neq j$ y $\langle \phi_i, \phi_i \rangle = 1$ para $i, j = 0, 1, \cdots, n$. Esta condición equivale a que la matriz de Gram asociada a esta base sea la matriz identidad. En este caso, la mejor aproximación a f es

$$g = \sum_{i=0}^{n} \langle f, \phi_i \rangle \, \phi_i.$$

En una situación general, se espera que la base no sea ortonormal. Sin embargo, conviene tener en cuenta que una base de un espacio vectorial con producto escalar, siempre puede ser transformada en otra ortonormal, mediante un proceso de orto-normalización como el de Gram-Schmidt.

▪ EJEMPLO **23** Una sucesión de polinomios ortogonales construidos con el proceso de Gram-Schmidt a partir de la base $\{1, x, \cdots, x^n\}$ es la siguiente:

$$\begin{aligned} p_0(x) &= 1, \\ p_1(x) &= x - \frac{\langle p_0, x\rangle}{\langle p_0, p_0\rangle} p_0 = x, \\ p_2(x) &= x^2 - \frac{\langle p_0, x^2\rangle}{\langle p_0, p_0\rangle} p_0 - \frac{\langle p_1, x^2\rangle}{\langle p_1, p_1\rangle} p_1 = x^2 - \frac{1}{3}, \\ \cdots & \quad \cdots \quad \diamond \end{aligned}$$

Obviamente, existen otros modos de construir bases ortogonales de polinomios que no están basados directamente el procedimiento de Gram-Schmidt. Además, en el caso de aproximación con seminormas, la aplicación de este procedimiento puede implicar divisiones por cero ya que existen polinomios con seminorma nula que no son el polinomio 0.

▬ EJERCICIO 35 *Construir los tres primeros polinomios de una sucesión de polinomios ortogonales con coeficiente principal igual a 1 respecto al producto escalar degenerado*

$$\langle p, q\rangle = \sum_{i=0}^{3} p(x_i) q(x_i)$$

donde x_i es el $(i+1)-$ésimo punto del conjunto $\{-1, 0, 1, 2\}$. Hallar la mejor aproximación de la función $f(x) = \operatorname{sen} \frac{\pi}{2} x$ en el espacio de los polinomios de grado 2 con respecto a la seminorma inducida por este producto escalar.

Solución: Habitualmente se llama coeficiente principal al coeficiente de la potencia de mayor grado del polinomio. Por esta razón, los polinomios serán de la forma

$$\begin{aligned} p_0(x) &= 1, \\ p_1(x) &= x + a, \\ p_2(x) &= x^2 + bx + c \end{aligned}$$

y deben cumplir:

$$\langle p_0, p_1\rangle = 0, \quad \langle p_0, p_2\rangle = 0, \quad \langle p_1, p_2\rangle = 0.$$

De ello se desprenden las siguientes ecuaciones

$$\begin{aligned} 4a+2 &= 0, \\ 2b+4c &= -6, \\ 2ab+4ac+6a+6b+2c &= -8. \end{aligned}$$

La resolución de estas ecuaciones permite responder a a la primera cuestión del ejercicio

$$\begin{aligned} p_0(x) &= 1, \\ p_1(x) &= x-\frac{1}{2}, \\ p_2(x) &= x^2-x-1. \end{aligned}$$

El cálculo de la mejor aproximación a f se puede organizar con la siguiente tabla

$f(x)p_i(x)$	-1	0	1	2	$< f, p_i >$
$\operatorname{sen} \frac{\pi}{2}x$	-1	0	1	0	0
$\left(x-\frac{1}{2}\right)\operatorname{sen} \frac{\pi}{2}x$	$\frac{3}{2}$	0	$\frac{1}{2}$	0	2
$(x^2-x-1)\operatorname{sen} \frac{\pi}{2}x$	-1	0	-1	0	$-2.$

Consecuentemente, la mejor aproximación es la función

$$\begin{aligned} g(x) &= \sum_{i=0}^{2} \langle f, p_i \rangle \frac{p_i(x)}{\|p_i\|^2} = \frac{2}{5}\left(x-\frac{1}{2}\right) - \frac{1}{2}\left(x^2-x-1\right) \\ &= -\frac{1}{2}x^2+\frac{9}{10}x+\frac{3}{10} \end{aligned}$$

ya que $\|p_1\|^2 = 5$ y $\|p_2\|^2 = 4$. Finalmente, la mejor aproximación es única ya que el producto escalar no es degenerado en $\mathcal{P}_2$. En efecto, si un polinomio de grado menor o igual a 2, se anula en 3 puntos distintos es necesariamente el polinomio nulo. $\diamond$

Como ya se ha señalado anteriormente, el método de ortogonalización de Gram-Schmidt suministra un procedimiento para generar sucesiones de polinomios ortogonales respecto a un producto escalar. Sin embargo, el siguiente resultado permite calcularlas de un modo más simple

– TEOREMA 14 *Si* $\{p_n\}$ *es una sucesión de polinomios ortogonales respecto al producto escalar*

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

donde ω *es una función positiva en* (a, b)*. Además, se supone que* p_n *es de grado* n *y tiene coeficiente principal igual a 1 (coeficiente del monomio de mayor grado). En estas condiciones se tiene que*

$$\begin{aligned} p_0(x) &= 1, \\ p_1(x) &= x - a_1, \\ p_2(x) &= (x - a_2)p_1(x) - b_2p_0(x), \\ \cdots &\quad \cdots \\ p_n(x) &= (x - a_n)p_{n-1} - b_np_{n-2}(x) \\ \cdots &\quad \cdots \end{aligned}$$

donde

$$a_n = \frac{\langle xp_{n-1}, p_{n-1}\rangle}{\langle p_{n-1}, p_{n-1}\rangle}, \quad b_n = \frac{\langle xp_{n-1}, p_{n-2}\rangle}{\langle p_{n-2}, p_{n-2}\rangle}.$$

Además, la sucesión definida por esta recurrencia de tres términos es la única sucesión ortogonal de grados distintos y coeficiente principal igual a 1.

Demostración: Los polinomios $\{p_0, p_1, \cdots, p_n\}$ forman una base de $\mathcal{P}_n$. En efecto, la matriz de cambio a la base formada por los monomios $\{1, x, x^2, \cdots, x^n\}$ es una matriz triangular inferior con unos en la diagonal principal.

El polinomio $xp_{n-1}(x)$ es de grado n y su coeficiente principal es 1. Consecuente, se puede expresar de modo único como combinación lineal de los polinomios $\{p_0, p_1, \cdots, p_n\}$

$$xp_{n-1}(x) = \sum_{i=0}^{n} \alpha_i p_i(x),$$

con $\alpha_n = 1$. Si se multiplica escalarmente esta igualdad por p_i, se obtiene

$$\langle xp_{n-1}, p_i\rangle = \alpha_i \langle p_i, p_i\rangle$$

para $i = 0, 1, \cdots, n-1$.

Por otra parte, el polinomio p_{n-1} es ortogonal al espacio de polinomios generado por $\{p_0, p_1, \cdots, p_{n-2}\}$ y en particular, a todos los polinomios de la forma xp_i para $i = 0, 1, \cdots, n-3$. De ello se deduce que

$$\langle xp_{n-1}, p_i \rangle = \langle p_{n-1}, xp_i \rangle = 0$$

para $i = 0, 1 \cdots, n-3$ y consecuentemente se tiene

$$p_n(x) = (x - \alpha_{n-1})p_{n-1}(x) - \alpha_{n-2}p_{n-2}(x).$$

Finalmente, se llega al resultado si se elige $a_n = \alpha_{n-1}$ y $b_n = \alpha_{n-2}$. ◇

Si el intervalo es simétrico respecto al origen y la función peso es par (es decir, verifica que $\omega(-x) = \omega(x)$ para todo $x \in [a, b]$), todos los coeficientes a_n en el teorema anterior, son nulos. En este caso, todos los polinomios p_n de orden par son funciones pares y los polinomios p_n de orden impar son funciones impares (es decir, verifican que $p(-x) = -p(x)$ para todo $x \in [a, b]$). Sin dificultad, se pueden probar estos resultados usando un método de inducción, teniendo en cuenta que la integral en un intervalo simétrico de una función impar, es nula.

▪ Ejemplo **24** Los polinomios de Legendre forman una sucesión de polinomios ortogonales respecto al producto escalar

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

que puede ser generada por la recurrencia de tres términos. Puesto que el intervalo es simétrico, los coeficientes a_i son nulos. Finalmente, los polinomios ortogonales están dados en la siguiente tabla

a_n	b_n	$p_n(x)$
.	.	1
0	.	x
0	$\frac{1}{3}$	$x^2 - \frac{1}{3}$
0	$\frac{4}{15}$	$x^3 - \frac{3}{5}x$
0	$\frac{9}{35}$	$x^4 - \frac{6}{7}x^2 + \frac{3}{35}$
$\vdots$	$\vdots$	$\vdots$

◇

Es conveniente notar que la demostración del teorema anterior permanece válida para cualquier producto escalar que cumpla

$$\langle xf, g\rangle = \langle f, xg\rangle = 0$$

para todas la funciones $f, g \in V$ (por xf se entiende la función definida en $[a, b]$ por $h(x) = xf(x)$). En ocasiones resulta interesante ponderar en la definición del producto escalar, la relevancia de una parte del intervalo $[a, b]$ frente a otras. Por ejemplo, si se pretende ponderar más la proximidad de dos funciones en los extremos del intervalo $[-1, 1]$ que en el centro del intervalo se puede considerar el siguiente producto escalar ponderado

$$\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)\omega(x)\, dx, \qquad \text{con} \quad \omega(x) = \frac{1}{\sqrt{1-x^2}}. \tag{4.2}$$

En la siguiente sección se introduce una sucesión de polinomios ortogonales respecto a este producto escalar con peso.

4.6 Polinomios de Chebyshev

Las funciones definidas por

$$T_n(x) = \cos(n \arccos(x)),$$

para $n = 0, 1, \cdots$, se llaman polinomios de Chebyshev en el intervalo $[-1, 1]$.

No es evidente al observar esta definición, que se trate de una sucesión de polinomios, ya que, en principio, intervienen funciones trigonométricas. Para $n = 0, 1$ directamente se comprueba que

$$T_0(x) = 1, \;\; T_1(x) = x.$$

Para comprobar que la función T_n es un polinomio para cualquier n, se considera la identidad

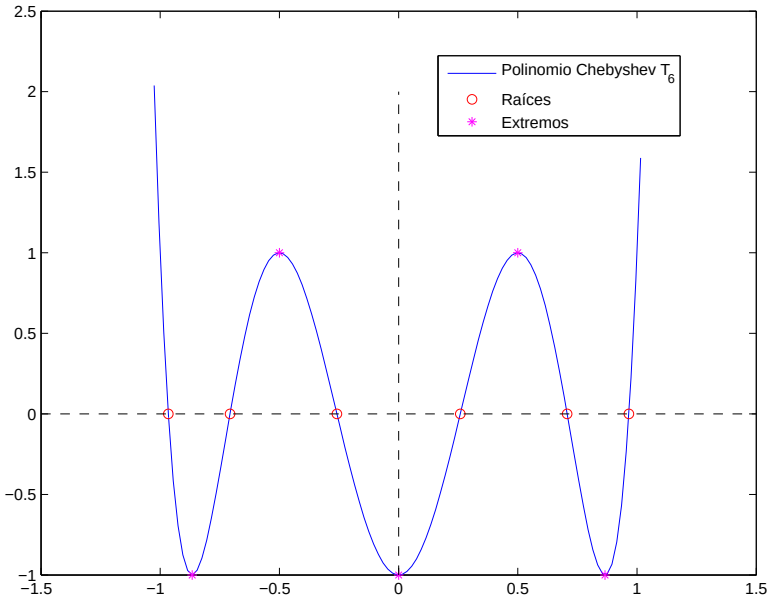
$$\cos(a + b) + \cos(a - b) = 2 \cos a \cos b$$

que aplicada a $a = n \arccos x$ y $b = \arccos x$ lleva a la siguiente relación

$$\cos((n + 1) \arccos x) + \cos((n - 1) \arccos x) = 2x \cos(n \arccos x)).$$

Si se usa la definición de polinomio de Chebyshev en esta igualdad, se obtiene que

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{4.3}$$



Mediante esta recurrencia de tres términos y los polinomios de bajo grado T_0 y T_1, se pueden construir las expresiones polinomiales de los elementos de la sucesión T_n. Así pues, T_n es un polinomio de grado n.

Directamente de la definición primitiva de los polinomios de Chebyshev se deduce que las raíces de T_n son

$$x_i = \cos\left(\frac{2i-1}{2n}\pi\right), \quad i = 1, \cdots, n$$

y las de su derivada

$$\bar{x}_i = \cos\left(\frac{i}{n}\pi\right), \quad i = 1, \cdots, n-1.$$

En las raíces del polinomio T_n' se cumple que

$$T_n(\bar{x}_i) = \cos i\pi = (-1)^i$$

y puesto que

$$|T_n(x)| \leq 1, \qquad \text{para todo } x \in [0,1],$$

se deduce que el polinomio de Chebyshev T_n alcanza sus máximos y mínimos de modo alternativo en los $n-1$ puntos $\{\bar{x}_i : i = 1, \cdots, n-1\}$.

Sin duda, una propiedad relevante de estos polinomios es que para n fijo, el conjunto $\{T_0, T_1, \cdots, T_n\}$ es una base ortogonal del espacio vectorial de

los polinomios de grado menor o igual que n con respecto al producto escalar definido por 4.2. En efecto, para probar que se anula el producto

$$\langle T_n, T_m \rangle = \int_{-1}^{1} \frac{\cos(n \arccos(x)) \cos(m \arccos(x))}{\sqrt{1-x^2}} dx$$

para $n \neq m$, se introduce el cambio de variable $\theta = \arccos x$. Puesto que

$$d\theta = -\frac{1}{\sqrt{1-x^2}} dx$$

se obtiene

$$\begin{aligned} \langle T_n, T_m \rangle &= \int_0^{\pi} \cos(n\theta) \cos(m\theta) d\theta \\ &= \frac{1}{2} \int_0^{\pi} (\cos((n+m)\theta) + \cos((n-m)\theta)) d\theta \\ &= \begin{cases} 0, & \text{si } m \neq n, \\ \frac{\pi}{2}, & \text{si } m = n \neq 0, \\ \pi, & \text{si } m = n = 0. \end{cases} \end{aligned}$$

Así pues, la sucesión $\{\sqrt{\frac{1}{\pi}}, \sqrt{\frac{2}{\pi}} T_n : n > 0\}$ es una sucesión de polinomios ortonormales respecto al producto escalar de Chebyshev.

Sea a_n el coeficiente de la potencia x^n del polinomio T_n (en adelante, coeficiente principal). De la fórmula de recurrencia 4.3 se deduce directamente que

$$a_{n+1} = 2a_n = 2^n.$$

Es frecuente estandarizar la sucesión de polinomios de Chebyshev dividiendo cada polinomio por su coeficiente principal. En este caso, se obtiene la sucesión de polinomios determinada por la recurrencia

$$\begin{aligned} \hat{T}_0 &= 1, \\ \hat{T}_1 &= x, \\ \hat{T}_2(x) &= x\hat{T}_1(x) - \frac{1}{2}\hat{T}_0(x) \\ \cdots &\quad \cdots \\ \hat{T}_{n+1}(x) &= x\hat{T}_n(x) - \frac{1}{4}\hat{T}_{n-1}(x) \end{aligned}$$

para $n \geq 2$.

Los polinomios estandarizados de Chebyshev son los polinomios de coeficiente principal 1 de menor norma uniforme como se prueba en el siguiente

▬ TEOREMA 15 *El polinomio estandarizado de Chebyshev* $\hat{T}_n = \frac{1}{2^{n-1}} T_n$ *verifica la siguiente desigualdad*

$$\max_{x\in[-1,1]} |\hat{T}_n(x)| \leq \max_{x\in[-1,1]} |p(x)|,$$

para todo polinomio $p \in \mathcal{P}_n$ *de coeficiente principal igual a 1.*

Demostración: Supongamos que p es un polinomio de grado n con coeficiente principal igual a 1 que verifica que

$$\max_{x\in[-1,1]} |p(x)| < \max_{x\in[-1,1]} \frac{1}{2^{n-1}} |T_n(x)|$$

El polinomio $q(x) = \frac{1}{2^{n-1}} T_n(x) - p(x)$ es de grado $n-1$ a lo sumo y toma valores que cambian de signo en los $n+1$ extremos $\bar{x}_k$ del polinomio de Chebyshev. De acuerdo con el teorema de Rolle, el polinomio q tiene n raíces y puesto tiene como grado máximo posible, $n-1$ tiene que ser nulo. Esto es contradictorio con el supuesto inicial. ◇

4.7 Aproximación trigonométrica

La mayoría de los subespacios U usados en la aproximación han sido espacios de polinomios. Pero debe quedar claro que esto es solamente una conveniencia y no una exigencia. Es decir, los teoremas de existencia y unicidad de mejor aproximación son válidos si se utilizan otras subespacios funcionales. Obviamente, el requisito de que sean funciones de fácil evaluación es indispensable para que tenga sentido práctico la aproximación.

▬ EJERCICIO 36 *Se considera el espacio de las funciones continuas en* $[-\pi, \pi]$ *con el producto escalar definido por*

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

y el subespacio generado por las funciones $\{1, \cos x, \operatorname{sen} x\}$*. Calcular la mejor aproximación de la función* $f(x) = x$ *en este subespacio.*

Solución: Es inmediato verificar que las funciones de la base son ortogonales y sus normas son

$$\langle 1, 1\rangle^{\frac{1}{2}} = 1,$$

$$\langle \cos x, \cos x\rangle^{\frac{1}{2}} = \langle \operatorname{sen} x, \operatorname{sen} x\rangle^{\frac{1}{2}} = \frac{1}{\sqrt{2}}.$$

Así pues, la base

$$\left\{1, \sqrt{2}\cos x, \sqrt{2}\,\mathrm{sen}\, x\right\}$$

es una base ortonormal. La mejor aproximación de la función f está dada por

$$g = \langle f, 1\rangle + \left\langle f, \sqrt{2}\cos x\right\rangle \sqrt{2}\cos x + \left\langle f, \sqrt{2}\,\mathrm{sen}\, x\right\rangle \sqrt{2}\,\mathrm{sen}\, x = 2\,\mathrm{sen}\, x.$$

Es importante tener en cuenta, durante la realización de estos cálculos, que una función impar solamente puede tener componentes no nulas en la función $\mathrm{sen}\, x$ ya que la integral de una función impar en un intervalo simétrico respecto al origen siempre es nula. ◇

De modo más general, se considera el subespacio U de los polinomios trigonométricos generados por las funciones trigonométricas

$$\{1, \cos x, \cdots, \cos nx, \mathrm{sen}\, x, \cdots, \mathrm{sen}\, nx\}$$

para algún $n > 0$, en el intervalo $[-\pi, \pi]$ y el producto escalar usado en el ejemplo anterior. Si se tienen en cuenta las siguientes igualdades trigonométricas para $k, j \geq 0$

$$\begin{aligned}
\int_{-\pi}^{\pi} \cos kx\,\mathrm{sen}\, jx\; dx &= 0,\\
\int_{-\pi}^{\pi} \cos kx\cos jx\; dx &= \int_{-\pi}^{\pi} \mathrm{sen}\, kx\,\mathrm{sen}\, jx\; dx = \begin{cases} 0, & \text{si } k \neq j,\\ \pi, & \text{si } k = j > 1\\ 2\pi, & \text{si } k = j = 0 \end{cases}
\end{aligned}$$

se puede comprobar que la mejor aproximación a una función f en U es el siguiente polinomio trigonométrico

$$g_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n}(a_k\cos kx + b_k\,\mathrm{sen}\, kx)$$

donde

$$\begin{aligned}
a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos kx\; dx,\\
b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,\mathrm{sen}\, kx\; dx.
\end{aligned}$$

Esta función se conoce como serie de Fourier finita de la función f.

▪ EJEMPLO **25** Los coeficientes de la serie de Fourier de la función $f(x) = e^x$ en el intervalo $[-\pi, \pi]$ son los siguientes

$$\begin{aligned}
a_0 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\ dx = 2a, & b_1 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,\mathrm{sen}\,x\ dx = a,\\
a_1 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos x\ dx = -a, & b_2 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,\mathrm{sen}\,2x\ dx = -\frac{4}{5}a,\\
a_2 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos 2x\ dx = \frac{2}{5}a, & b_3 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,\mathrm{sen}\,3x\ dx = \frac{3}{5}a,\\
a_3 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos 3x\ dx = -\frac{1}{5}a, & & \\
\ldots & \ldots & &
\end{aligned}$$

donde

$$a = \sinh \pi = \frac{e^{\pi} - e^{-\pi}}{2\pi}.$$

Consecuentemente, la mejor aproximación trigonométrica de orden 3 a f es

$$g_3(x) = \frac{e^{\pi} - e^{-\pi}}{2\pi}\left(1 - \cos x + \mathrm{sen}\,x + \frac{2}{5}\cos 2x - \frac{4}{5}\,\mathrm{sen}\,2x - \frac{1}{5}\cos 3x + \frac{3}{5}\,\mathrm{sen}\,3x\right).$$

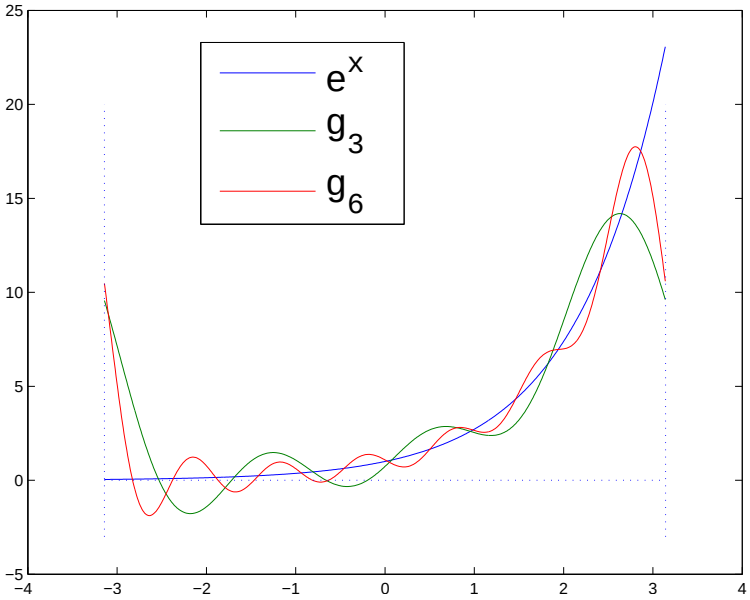


Figura 4.3: Aproximaciones trigonométricas de órdenes 3 y 6 a la función e^x

En la figura 4.3 se observa que la precisión de la aproximación trigonométrica se deteriora bruscamente en las proximidades de los extremos del

intervalo. De hecho, este deterioro se incrementa cuando se aumenta el grado de la aproximación. La aproximación de grado 6 parece que se aproxima mejor a la función f lejos de los extremos pero empeora en sus proximidades. ◇

La razón de este comportamiento hay que buscarla en el siguiente hecho: El valor de una aproximación trigonométrica en los extremos es siempre la suma de la serie finita alternada

$$g_n(-\pi) = g_n(\pi) = \frac{a_0}{2} + \sum_{k=1}^{n} (-1)^k a_k.$$

Es decir, la aproximación trigonométrica es siempre una función de periodo 2π. Este hecho deteriora la calidad de la aproximación en los extremos del intervalo. Consecuentemente, la aproximación trigonométrica resulta adecuada para aproximar funciones periódicas. De un modo más formal, el espacio V debiera ser

$$V = \{f \in C([-\pi, \pi]) : f(-\pi) = f(\pi)\}.$$

Algo sorprendentemente, si se relaja la condición de continuidad de la función y se considera un espacio más amplio de funciones que admitan discontinuidades de salto, el mismo fenómeno (que se conoce como fenómeno de Gibbs) aparece de nuevo, esta vez alrededor del punto donde la función tiene el salto.

De modo similar al caso de polinomios algebraicos, se puede desarrollar también una versión discreta de la aproximación de mínimos cuadrados trigonométrica. Para ello, se considera una partición del intervalo $[-\pi, \pi]$ en sub-intervalos igualmente espaciados, definida por los nodos

$$x_k = \left(\frac{k}{m} - 1\right) \pi$$

para $k = 0, 1, \cdots, 2m$. Asociado a esta partición se considera el producto discreto

$$\langle f, g \rangle = \sum_{k=0}^{2m-1} f(x_k) g(x_k). \tag{4.4}$$

Un hecho relevante, que simplificará los cálculos, se establece en el siguiente

▬ Teorema 16 *Para $n < m$, los polinomios trigonométricos*

$$\{\cos kx, \operatorname{sen} kx : k = 0, 1, \cdots, n\}$$

son ortogonales respecto al producto escalar discreto (4.4). La aproximación trigonométrica discreta de la función f está dada por

$$g(x) = \frac{a_0}{2} + \sum_{j=1}^{n} (a_j \cos jx + b_j \operatorname{sen} jx) \tag{4.5}$$

donde

$$\begin{aligned} a_j &= \frac{1}{m} \sum_{k=0}^{2m-1} f(x_k) \cos jx_k, \\ b_j &= \frac{1}{m} \sum_{k=0}^{2m-1} f(x_k) \operatorname{sen} jx_k. \end{aligned}$$

Demostración: Muchas técnicas de la aproximación trigonométrica se fundamentan en resultados del Análisis Complejo. La demostración de este lema puede considerarse como un ejemplo de ello. Se representa la unidad imaginaria por $\mathtt{i} = \sqrt{-1}$. De la fórmula de Euler

$$e^{\mathtt{i}x} = \cos x + \mathtt{i} \operatorname{sen} x,$$

se deducen la fórmulas

$$\begin{aligned} \cos x &= \frac{e^{\mathtt{i}x} + e^{-\mathtt{i}x}}{2}, \\ \operatorname{sen} x &= \frac{e^{\mathtt{i}x} - e^{-\mathtt{i}x}}{2\mathtt{i}}, \end{aligned}$$

que permiten relacionar los polinomios trigonométricos reales e imaginarios.

Para calcular la matriz de Gram de esta base respecto al anterior producto escalar se usará el siguiente

Lema 2 *Para cualquier número entero M tal que $-2m \leq M \leq 2m$, se cumple*

$$\sum_{k=0}^{2m-1} e^{iMx_k} = \begin{cases} 2m & \text{si } M = 0 \text{ ó } M = \pm 2m; \\ 0 & \text{en otro caso.} \end{cases} \tag{4.6}$$

Demostración: En efecto, de la fórmula de la suma de los términos de una progresión geométrica se deduce que

$$\begin{aligned}\sum_{k=0}^{2m-1} e^{\mathtt{i}Mx_k} &= e^{-\mathtt{i}M\pi}\sum_{k=0}^{2m-1} e^{\mathtt{i}M(x_k+\pi)} = (-1)^M \sum_{k=0}^{2m-1} e^{\mathtt{i}M\frac{k}{m}\pi} \\ &= (-1)^M(1+a+a^2+\cdots+a^{2m-1}) \\ &= (-1)^M\frac{1-a^{2m}}{1-a},\end{aligned}$$

considerando $a = e^{\mathtt{i}\frac{M\pi}{m}}$. Sin embargo, puesto que M es número entero, a es una raíz $2m-$ésima de la unidad, como muestra la siguiente igualdad

$$a^{2m} = \cos\left(2m\frac{M\pi}{m}\right) + \mathtt{i}\,\mathrm{sen}\left(2m\frac{M\pi}{m}\right) = 1.$$

Si $M = 2m$ el cociente que aparece en la expresión anterior es una forma indeterminada. No obstante, si se usa la regla de L'Hôpital, se obtiene

$$\sum_{k=0}^{2m-1} e^{\mathtt{i}2mx_k} = \lim_{M\to 2m}(-1)^M\frac{1-a^{2m}}{1-a} = (-1)^{2m}\frac{-2ma^{2m-1}}{-1} = 2m.$$

El mismo argumento se puede aplicar al caso $M = -2m$.

Finalmente, si $M = 0$, directamente se verifica que

$$\sum_{k=0}^{2m-1} e^{\mathtt{i}Mx_k} = 2m. \qquad \diamond$$

Si se utilizan las anteriores relaciones trigonométricas en la siguiente forma

$$\begin{aligned}\langle \cos px, \cos qx\rangle &= \frac{1}{4}\sum_{k=0}^{2m-1}\left(e^{\mathtt{i}(p+q)x_k} + e^{-\mathtt{i}(p+q)x_k} + e^{\mathtt{i}(p-q)x_k} + e^{-\mathtt{i}(p-q)x_k}\right) \\ &= \begin{cases} m, & \text{si } p = q; \\ 0, & \text{en otro caso.}\end{cases} \\ \langle \mathrm{sen}\, px, \mathrm{sen}\, qx\rangle &= \frac{-1}{4}\sum_{k=0}^{2m-1}\left(e^{\mathtt{i}(p+q)x_k} + e^{-\mathtt{i}(p+q)x_k} - e^{\mathtt{i}(p-q)x_k} - e^{-\mathtt{i}(p-q)x_k}\right) \\ &= \begin{cases} m, & \text{si } p = q; \\ 0, & \text{en otro caso.}\end{cases} \\ \langle \cos px, \mathrm{sen}\, qx\rangle &= \frac{1}{4\mathtt{i}}\sum_{k=0}^{2m-1}\left(e^{\mathtt{i}(p+q)x_k} - e^{-\mathtt{i}(p+q)x_k} - e^{\mathtt{i}(p-q)x_k} + e^{-\mathtt{i}(p-q)x_k}\right) = 0.\end{aligned}$$

para todo $p, q = 0, 1, \cdots, n$, de donde se obtiene el resultado buscado.

Es importante notar que la limitación de $n < m$ se debe a que si $n = m$ la matriz de Gram no tendría determinante distinto de cero ya que

$$\langle \operatorname{sen} mx, \operatorname{sen} mx \rangle = 0.$$

y el producto escalar sería degenerado en el espacio generado por los polinomios trigonométricos de orden menor o igual a m. No obstante, es posible tomar $n = m$ si se excluye el polinomio trigonométrico $\operatorname{sen} nx$. ◇

▁ EJERCICIO 37 *Determinar la mejor aproximación trigonométrica de la función*

$$f(x) = 8\left(\frac{x}{\pi}\right)^3 + 4\left(\frac{x}{\pi}\right)^2 - 8\frac{x}{\pi},$$

para el producto escalar discreto

$$\langle f, g \rangle = \sum_{k=0}^{3} f(x_k) g(x_k).$$

basado en los puntos $x_k = \left(\frac{k}{2} - 1\right)\pi$ *para* $k = 0, 1, 2, 3$.

Solución: Se construye la siguiente tabla para $m = 2$

x	1	$\cos x$	$\operatorname{sen} x$	$\cos 2x$	$\operatorname{sen} 2x$	f
$-\pi$	1	-1	0	1	0	4
$-\pi/2$	1	0	-1	-1	0	4
0	1	1	0	1	0	0
$\pi/2$	1	0	1	-1	0	-2
	3	-2	-3	1	0	
	a_0	a_1	b_1	a_2	b_2	

de la que se deduce que la aproximación de orden $n = 1$ es

$$g_1(x) = \frac{3}{2} - 2\cos x - 3\operatorname{sen} x.$$

Podría usarse la aproximación de orden $n = 2$ (excluyendo el término seno). En este caso se obtiene

$$g_2(x) = \frac{3}{2} - 2\cos x - 3\operatorname{sen} x + \cos 2x.$$

◇

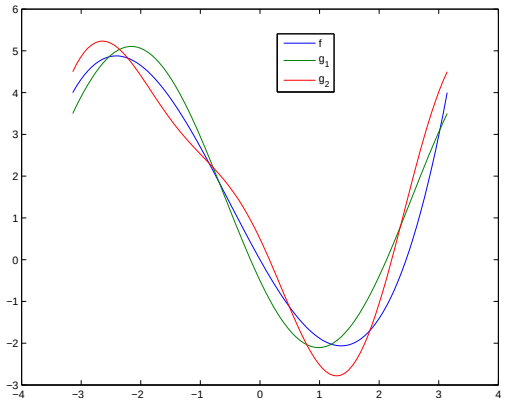


Figura 4.4: Aproximaciones trigonométricas de órdenes 1 y 2 a la función $f(x) = 8\left(\frac{x}{\pi}\right)^3 + 4\left(\frac{x}{\pi}\right)^2 - \frac{8x}{\pi}$.

4.8 Aproximación uniforme

La mejor aproximación uniforme a una función es en cierto sentido más severa que la aproximación por mínimos cuadrados ya que esta proximidad debe mantenerse en todos los puntos del intervalo y una alteración de la función en un entorno de un punto la alejaría notablemente de su mejor aproximación. Antes de entrar en el análisis de la mejor aproximación uniforme en un espacio de polinomios de grado limitado, es importante destacar que para toda función continua existe un polinomio tan próximo a la función como se desee, en el sentido uniforme. Se describe este resultado, de modo más preciso, en el siguiente

– TEOREMA 17 *(de Weierstrass) Si f es una función continua en un intervalo cerrado y acotado I, para todo número real positivo ε existe un polinomio p tal que $\|f - p\|_\infty < \varepsilon$.*

Demostración: La sucesión de polinomios que se construye a continuación, conocida como sucesión de polinomios de Bernstein, tiene por límite uniforme la función f. Por razones de sencillez, se supone que $I = [0, 1]$ pero un sencillo cambio de variable podría extender el razonamiento a cualquier intervalo cerrado y acotado.

Para $n \geq 0$, se define el polinomio de Bernstein de grado n como

$$B_{n,f}(x) = \sum_{i=0}^{n} f\left(\frac{i}{n}\right)\binom{n}{i}x^i(1-x)^{n-i}.$$

Puesto que

$$\sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} = (x+1-x)^n = 1$$

$B_{n,f}$ puede interpretarse como un valor promediado de los valores de f en los puntos $\frac{i}{n}$. Para demostrar el teorema se probará que la sucesión de polinomios de Bernstein $\{B_{n,f}\}$ converge uniformemente a f.

En primer lugar, se establecen las dos siguientes propiedades de los polinomios de Bernstein:

1. El operador B_n es lineal y positivo en el espacio de las funciones continuas:

   - $B_{n,\alpha f+\beta g} = \alpha B_{n,f} + \beta B_{n,g}$,
   - Si $f(x) \geq 0$ para todo $x \in I$ se cumple que $B_{n,f}(x) \geq 0$ para todo $x \in I$

   cuya prueba es trivial.

2. La sucesión $\{B_{n,f}\}$ converge uniformemente a f si f es un polinomio de grado menor o igual que 2. En efecto, si se deriva dos veces la relación

   $$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

   respecto de la variable x, se obtienen las relaciones

   $$x(x+y)^{n-1} = \sum_{i=0}^{n} \frac{i}{n} \binom{n}{i} x^i y^{n-i}, \tag{4.7}$$

   $$x^2(x+y)^{n-2} = \sum_{i=0}^{n} \frac{i}{n} \frac{i-1}{n-1} \binom{n}{i} x^i y^{n-i}. \tag{4.8}$$

   En particular, si se elige $y = 1-x$ en las últimas igualdades, se obtiene

   $$x = \sum_{i=0}^{n} \frac{i}{n} \binom{n}{i} x^i (1-x)^{n-i}, \tag{4.9}$$

   $$x^2 = \sum_{i=0}^{n} \frac{i}{n} \frac{i-1}{n-1} \binom{n}{i} x^i (1-x)^{n-i} \tag{4.10}$$

de donde se deduce que $x = B_{n,x}$ y si además se utiliza la igualdad 4.10, junto con la siguiente igualdad

$$\frac{i}{n}\,\frac{i-1}{n-1} = \frac{n}{n-1}\,\frac{i^2}{n^2} - \frac{1}{n-1}\frac{i}{n},$$

se obtiene

$$x^2 = \frac{n}{n-1}B_{n,x^2} - \frac{1}{n-1}x.$$

Finalmente, de la relación

$$B_{n,x^2} = \frac{n-1}{n}x^2 + \frac{1}{n}x \tag{4.11}$$

se deduce que el límite uniforme de B_{n,x^2} es x^2. De la linealidad del operador B_n se desprende que este resultado se conserva para todo polinomio de grado menor o igual que 2.

En una segunda etapa de la demostración, dada una función continua f en $I = [0, 1]$, se probará que para todo $\varepsilon > 0$ existe un entero positivo n_0 tal que para todo $n > n_0$ se cumple que

$$\|f - B_{n,f}\|_\infty < \varepsilon.$$

Puesto que f es uniformemente continua en I existe $\delta > 0$ tal que

$$|x_1 - x_2| < \delta \Rightarrow |f(x_1) - f(x_2| < \frac{\varepsilon}{4}$$

en I. Para un punto arbitrario $\mathbf{x}_0 \in I$ se considera el polinomio

$$q(x) = f(x_0) + \frac{\varepsilon}{4} + 2\|f\|_\infty \frac{(x - x_0)^2}{\delta^2}.$$

Si $|x - x_0| < \varepsilon$ se tiene

$$q(x) \geq f(x_0) + \frac{\varepsilon}{4} > f(x_0) + |f(x) - f(x_0)| \geq f(x).$$

Si $|x - x_0| \geq \varepsilon$ se tiene

$$q(x) \geq f(x_0) + \frac{\varepsilon}{4} + 2\|f\|_\infty \geq f(x) + \frac{\varepsilon}{4} > f(x).$$

Consecuentemente, en cualquier caso el polinomio de segundo grado q toma en todos los puntos, valores iguales ó superiores a los que toma f.

Por otra parte, de acuerdo con los resultados de la primera etapa, se puede escoger n_0 tal que para $n > n_0$ se cumpla que

$$\|q - B_{n,q}\|_\infty < \frac{3\varepsilon}{4}.$$

También está garantizado que $B_{n,f}(x) \leq B_{n,q}(x)$ para todo $x \in I$ debido a que el operador B_n es positivo. De ello se deduce que

$$B_{n,f}(x_0) \leq B_{n,q}(x_0) < q(x_0) + \frac{3\varepsilon}{4} = f(x_0) + \varepsilon.$$

Si se razona de modo similar con el polinomio

$$q(x) = f(x_0) - \frac{\varepsilon}{4} - 2\|f\|_\infty \frac{(x - x_0)^2}{\delta^2}$$

se obtiene la desigualdad

$$B_{n,f}(x_0) > f(x_0) - \varepsilon.$$

Puesto que x_0 es arbitrario se obtiene finalmente que

$$\|B_{n,f} - f\|_\infty < \varepsilon. \qquad \diamond$$

▬ EJERCICIO 38 *Hallar los tres primeros polinomios $B_{n,f}$ de Bernstein en la aproximación uniforme a la función $f(x) = |x|$ en el intervalo $[-1, 1]$.*

Solución: El polinomio de Bernstein de orden n relativo a una función g en el intervalo $[0, 1]$, tiene la siguiente expresión analítica

$$B_{n,g}(t) = \sum_{i=0}^{n} \binom{n}{i} g(\frac{i}{n}) t^i (1 - t)^{n-i}.$$

Puesto que el intervalo considerado en el planteamiento del ejercicio es $[-1, 1]$ se realiza el siguiente cambio de variable $x = 2t - 1$. De este modo se obtiene el polinomio de Bernstein de la aproximación a $g(t) = |2t - 1|$ en $[0, 1]$ es

$$B_{n,g}(t) = \sum_{i=0}^{n} \binom{n}{i} \left|\frac{2i}{n} - 1\right| t^i (1 - t)^{n-i}.$$

En particular, se tiene

- Para $n = 1$

$$B_{1,g}(t) = \binom{1}{0}(1 - t) + \binom{1}{1} t = 1.$$

- Para $n = 2$

$$B_{2,g}(t) = \binom{2}{0}(1-t)^2 + \binom{2}{2}t^2 = 2t^2 - 2t + 1.$$

- Para $n = 3$

$$\begin{aligned} B_{3,g}(t) &= \binom{3}{0}(1-t)^3 + \binom{3}{1}\frac{1}{3}t(1-t)^2 \\ &+ \binom{3}{2}\frac{1}{3}t^2(1-t) + \binom{3}{3}t^3 = 2t^2 - 2t + 1. \end{aligned}$$

Retornando a la variable x, se obtienen los polinomios buscados

$$\{1, \frac{1+x^2}{2}, \frac{1+x^2}{2}\}. \qquad \diamond$$

Si se examina en detalle la sucesión de polinomios de Bernstein dada en 4.11 se observa que para $f = x^2$

$$\|B_{n,f} - f\|_\infty = \max_{x\in[0,1]} \frac{|x^2 - x|}{n} = \frac{1}{4n}.$$

Es decir, para obtener una aproximación con error a 10^{-4} se necesitaría el polinomio de grado $n = 2500$. A pesar de que f ya es polinomio de grado 2, es necesario ir a un polinomio de Bernstein de grado muy alto para aproximarse de modo preciso. Esto implica que, pese a que los polinomios de Bernstein dan aproximaciones explícitas para aproximar funciones continuas, no suministran una técnica muy eficiente debido a su lenta convergencia y su interés es más conceptual que práctico. En el siguiente ejemplo se intenta mostrar de un modo más visual estas deficiencias de la aproximación mediante polinomios de Bernstein.

- Ejemplo **26** Se considera la función f definida por

$$f(x) = e^{3x} \operatorname{sen} \pi x$$

en $I = [0, 1]$. Esta función es regular y no evidencia ninguna particularidad por la que la aproximación por polinomios de Bernstein pueda ser de mala calidad. Sin embargo, la figura 4.5 muestra que es necesario alcanzar valores muy altos en el grado n, para aproximarse adecuadamente a la función.

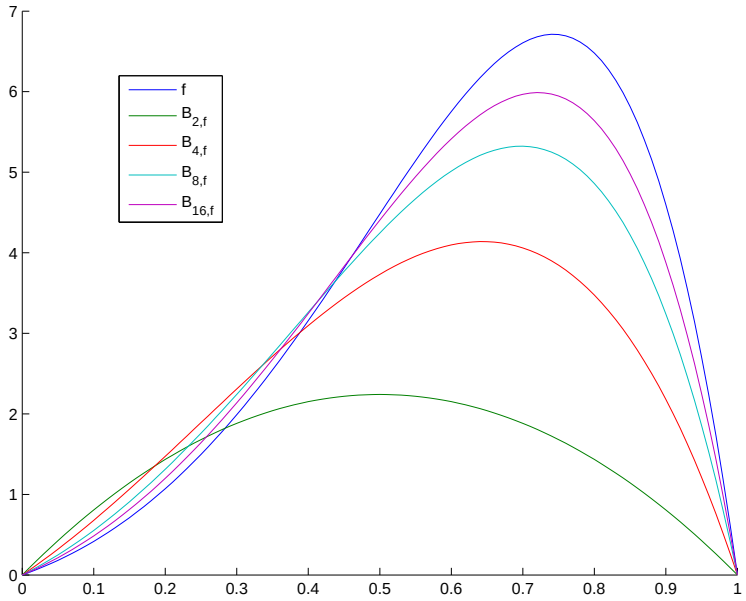


Figura 4.5: Aproximaciones uniformes con polinomios de Bernstein

Los ejemplos anteriores ponen de manifiesto que la estrategia de aproximar una función continua por el polinomio de Bernstein de un grado determinado, podría ser mejorada buscando la mejor aproximación a la función dentro del espacio de los polinomios de grado menor o igual que el dado. Los siguientes resultados permiten caracterizar la mejor aproximación uniforme.

TEOREMA 18 *(Criterio de Kolmogorov). Sea U un subespacio vectorial de dimensión finita del espacio de las funciones continuas en un intervalo $[a,b]$ y g un elemento de U. Entonces g es una mejor aproximación de f en U si y sólo si ningún elemento de U tiene el mismo signo que $f-g$ en todos los puntos del conjunto*

$$A_{f-g} = \{x \in [a,b] : |f(x) - g(x)| = \|f - g\|_\infty\}.$$

Demostración: En primer lugar se prueba que si la función g verifica la condición del enunciado del teorema entonces es una mejor aproximación de f en U.

Si $q \in U$ entonces $q - g \in U$. Si se aplica la condición de Kolmogorov a $q - g$, se deduce que existe un punto $x_0 \in A_{f-g}$ tal que

$$(f(x_0) - g(x_0))(g(x_0) - q(x_0)) \geq 0.$$

En consecuencia, se deduce que

$$\begin{aligned}\|f-q\|_\infty^2 &\geq (f(x_0)-q(x_0))^2 = (f(x_0)-g(x_0)+g(x_0)-q(x_0))^2\\ &= (f(x_0)-g(x_0))^2+(g(x_0)-q(x_0))^2+2(f(x_0)-g(x_0)(g(x_0)-q(x_0))\\ &\geq (f(x_0)-g(x_0))^2 = \|f-g\|_\infty^2\end{aligned}$$

lo que prueba que g es una mejor aproximación de f en U.

Recíprocamente, se supone que existe una función $q \in U$ que tiene el mismo signo que $f-g$ en A_{f-g} y se probará que g no es una mejor aproximación de f en U. Sean p y h la funciones definidas por

$$p = \frac{q}{\|q\|_\infty}, \qquad h = p(f-g).$$

Puesto que h es continua y positiva en el conjunto compacto $A_{f-g} \subset [a,b]$ se tiene que $\delta = \min_{x\in A_{f-g}} h(x) > 0$ ya que el mínimo se alcanza en algún valor de A_{f-g}. Además el conjunto

$$S = \{x \in [a,b] : h(x) \leq \frac{\delta}{2}\}$$

es compacto. Finalmente, se define

$$M = \max_{x\in S}(f(x)-g(x))$$

y se escoge ε tal que

$$0 < \varepsilon < \min\{\delta, \|f-g\|_\infty - M\}.$$

Si $x \in [a,b]$ pero $x \notin S$ entonces

$$\begin{aligned}|f(x)-g(x)-\varepsilon p(x)|^2 &= (f(x)-g(x))^2+\varepsilon^2 p(x)^2-2\varepsilon h(x)\\ &< (f(x)-g(x))^2+\varepsilon^2-\varepsilon\delta\\ &= (f(x)-g(x))^2+\varepsilon(\varepsilon-\delta) < (f(x)-g(x))^2.\end{aligned}$$

Si $x \in S$ entonces

$$|f(x)-g(x)-\varepsilon p(x)| \leq M+\varepsilon \leq \|f-g\|_\infty.$$

En definitiva, se tiene

$$\|f-g-\varepsilon p\|_\infty \leq \|f-g\|_\infty$$

lo que prueba que $g-\varepsilon p \in U$ mejora la aproximación a f de g. $\diamond$

▪ EJEMPLO **27** La función $g(x) = \frac{3}{4}x$ es la mejor aproximación uniforme de la función $f(x) = x^3$ en $[-1, 1]$ para $U = \mathcal{P}_2$. Fácilmente, se comprueba que

$$A_{f-g} = \{-1, -\frac{1}{2}, \frac{1}{2}, 1\}$$

y $\|f - g\|_\infty = \frac{1}{4}$ (véase figura 4.6). La función $f - g$ toma en A_{f-g} los valores $\{-\frac{1}{4}, \frac{1}{4}, -\frac{1}{4}, \frac{1}{4}\}$. Un polinomio $q(x) = a + bx + cx^2$ que tome valores con los

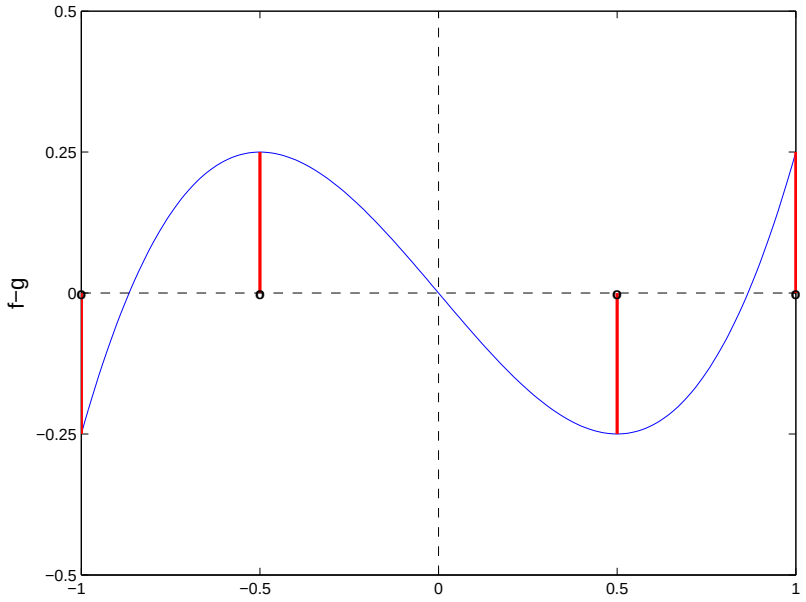


Figura 4.6: Gráfica de $f(x) - g(x) = x^3 - 3/4x$

mismos signos en los puntos de A_{f-g} debe verificar que

$$\begin{aligned} a - b + c &< 0, \\ a - \frac{b}{2} + \frac{c}{4} &> 0, \\ a + \frac{b}{2} + \frac{c}{4} &< 0, \\ a + b + c &> 0. \end{aligned}$$

Puesto que estas condiciones son incompatibles, queda probado que g es una mejor aproximación de f ◇

▬ TEOREMA 19 *(de la alternancia de Chebyshev) Un polinomio g es la mejor aproximación uniforme en $\mathcal{P}_n$ de una función f continua en $[a, b]$ si y sólo*

si $|f-g|$ *alcanza su máximo en* $n+2$ *puntos distintos con signo alternante en* $f-g$.

Demostración: Si $|f-g|$ alcanza su valor máximo en $n+2$ puntos con alternancia de signo y un polinomio p tiene el mismo signo que $f-g$ en esos puntos, entonces p tiene $n+1$ raíces lo que es imposible si p es un polinomio de grado n que no es nulo.

Recíprocamente, si se supone que $f-g$ toma los valores $\|f-g\|_\infty$ con signos alternantes en $m \leq n+1$ puntos y se probará que g no es la mejor aproximación de f. Se definen los conjuntos

$$A^{\pm}_{f-g} = \{x \in [a,b] : f(x) - g(x) = \pm\|f-g\|_\infty\}.$$

Se escoge $\varepsilon > 0$ y un conjunto $\{y_1, y_2, \cdots, y_m\}$ tal que los intervalos $(y_i - \varepsilon, y_i + \varepsilon)$ sean disjuntos y en ellos $f-g$ conserve el signo. A continuación se escogen m puntos $\{z_1, z-2, \cdots, z_n\}$ tales que

$$y_1 < z_1 < y_2 < \cdots < y_{m-1} < z_m < y_m.$$

Con ayuda de estos puntos se construye el polinomio

$$q(x) = (x-z_1)(x-z_2)\cdots(x-z_m)$$

que mantiene el signo en cada intervalo (z_i, z_{i+1}) para $i = 1, \cdots, m-1$ y consecuentemente también lo hace en los intervalos $(y_i - \varepsilon, y_i + \varepsilon)$ y alterna los signos en intervalos adyacentes. Del teorema de Kolmogorov se deduce que g no es la mejor aproximación a f en U $\diamond$.

▪ Ejemplo **28** El polinomio $g = 0$ es la mejor aproximación de la función $f(x) = \operatorname{sen} 3x$ en el intervalo $[0, 2\pi]$ en $\mathcal{P}_4$. En efecto, la diferencia $f-g$ alterna entre el máximo valor absoluto y su opuesto en 6 puntos. Por el contrario, el polinomio nulo no es una mejor aproximación de f en el espacio de los polinomios de grado menor o igual que 5 $\diamond$.

Una primera consecuencia del teorema de alternancia de Chebyshev es la unicidad de la aproximación uniforme.

▬ Teorema 20 *Existe una única mejor aproximación uniforme de una función continua en* $I = [a,b]$, *en* $\mathcal{P}_n$.

Demostración: Se supone que g_1 y g_2 son dos mejores aproximaciones de la función f en $\mathcal{P}_n$ y se representa por d la mínima distancia $d = \|f - g_1\|_\infty = \|f - g_2\|_\infty$. De la desigualdad triangular de la norma se deduce que $\frac{g_1+g_2}{2}$ también es una mejor aproximación de f. Del teorema de alternancia de Chebyshev se deduce la existencia de $n+2$ puntos $\{x_0, x_1, \cdots, x_n, x_{n+1}\}$ en los que $f - \frac{g_1+g_2}{2}$ toma alternativamente los valores extremos $\pm d$. Además, para $i = 0, 1, \cdots, n+1$ se tiene que

$$d = \left| f(x_i) - \frac{g_1(x_i) + g_2(x_i)}{2} \right| \leq \frac{1}{2}|f(x_i) - g_1(x_i)| + \frac{1}{2}|f(x_i) - g_2(x_i)| \leq d$$

lo que implica que esta desigualdad es realmente una igualdad y por lo tanto se cumple una de las tres siguientes posibilidades

- $g_1(x_i) = g_2(x_i)$,
- $g_1(x_i) = f(x_i) + d$, $g_2(x_i) = f(x_i) - d$,
- $g_2(x_i) = f(x_i) + d$, $g_1(x_i) = f(x_i) - d$.

Sin embargo, las dos últimas posibilidades conducen a que $d = 0$ y consecuentemente se deduce que la primera siempre ocurre. Pero, dos polinomios de grado menor o igual que n, que coinciden en $n+2$ puntos, son necesariamente iguales. Esto concluye la prueba. ◇

▪ Ejemplo **29** El polinomio $x^n - \frac{1}{2^{n-1}}T_n$ de grado $n-1$ es la mejor aproximación uniforme de x^n en $\mathcal{P}_{n-1}$ como establece el teorema 15. Además, del teorema anterior se deduce que $\frac{1}{2^{n-1}}T_n$ es el único polinomio de grado n y coeficiente principal igual a 1, que cumple el teorema 15. ◇

Una segunda consecuencia del teorema de alternancia de Chebyshev es el procedimiento para construir la mejor aproximación uniforme, que se conoce como método de intercambio de Remez y que se describe a continuación:

1. Seleccionar $n+2$ puntos $x_0 < x_1 < \cdots < x_n < x_{n+1}$ en $I = [a, b]$ arbitrariamente.

2. Calcular el polinomio $p_n \in \mathcal{P}_n$ y el parámetro d tales que

$$f(x_i) - p_n(x_i) = (-1)^{i+1} d \tag{4.12}$$

   para $i = 0, 1, \cdots, n, n+1$. Si se escoge una base de $\mathcal{P}_n$, las $n+1$ componentes del polinomio en esa base y valor de d son las $n+2$ incógnitas de un sistema lineal.

3. Determinar los puntos en los que la función $|f - p_n|$ alcanza su máximo y reemplazar con ellos, todos o parte de los x_i utilizados.

4. Volver a la etapa 2.

Se referencia [2] para justificar las siguientes afirmaciones

- La sucesión formada por los polinomios p_n obtenidos por el algoritmo de Remez es convergente a la mejor aproximación uniforme. La convergencia es cuadrática si f es diferenciable.
- Una buena elección de los puntos de partida es la de la raíces del polinomio de Chebyshev de grado $n + 2$.

▬ EJERCICIO 39 *Calcular la mejor aproximación uniforme de la función* $f(x) = x^4$ *en* $\mathcal{P}_2$ *en el intervalo* $[-1, 1]$*, usando el algoritmo de Remez (Dos iteraciones son suficientes).*

Solución: Puesto que $n = 2$ se escogen como 4 puntos de arranque, los extremos del intervalo y las raíces del polinomio de Chebyshev de grado 2

$$-1 < -\frac{\sqrt{2}}{2} < \frac{\sqrt{2}}{2} < 1.$$

El polinomio buscado en la primera iteración se expresa como

$$p(x) = a_0 + a_1 x + a_2 x^2$$

en la base de los monomios $\{1, x, x^2\}$. El sistema lineal 4.12 se convierte en

$$\begin{array}{cccccc} a_0 & -a_1 & +a_2 & -d & = & 1 \\ a_0 & -\frac{\sqrt{2}}{2}a_1 & +\frac{1}{2}a_2 & +d & = & \frac{1}{4} \\ a_0 & +\frac{\sqrt{2}}{2}a_1 & +\frac{1}{2}a_2 & -d & = & \frac{1}{4} \\ a_0 & +a_1 & +a_2 & +d & = & 1 \end{array}$$

Si se resuelve el sistema se obtiene que el polinomio correspondiente a la primera iteración de Remez es

$$p(x) = -\frac{1}{2} + \frac{3}{2}x^2$$

y $d = 0$. Para proceder en la fase de intercambio, se calculan los extremos de $f - p$ que corresponden a $x = 0$ y $x = \pm\frac{\sqrt{3}}{2}$. De acuerdo con el teorema de alternancia de Chebyshev se necesitan cuatro puntos en los que la función

de error tome alternativamente sus valores extremos. Consecuentemente el polinomio p no es la mejor aproximación uniforme.

Existen varias variantes del algoritmo de Remez que van desde intercambiar un solo punto extremo hasta intercambiar todos ellos. En este caso, a fin de preservar la simetría se procede a cambiar los nodos $x = \pm\frac{\sqrt{2}}{2}$ por los extremos simétricos $x = \pm\frac{\sqrt{3}}{2}$. En este caso, el sistema lineal que se obtiene es

$$\begin{array}{rrrrcl} a_0 & -a_1 & +a_2 & -d & = & 1 \\ a_0 & -\frac{\sqrt{3}}{2}a_1 & +\frac{3}{4}a_2 & +d & = & \frac{9}{16} \\ a_0 & +\frac{\sqrt{3}}{2}a_1 & +\frac{3}{4}a_2 & -d & = & \frac{9}{16} \\ a_0 & +a_1 & +a_2 & +d & = & 1 \end{array}$$

Si se resuelve el sistema se obtiene que el polinomio correspondiente a la

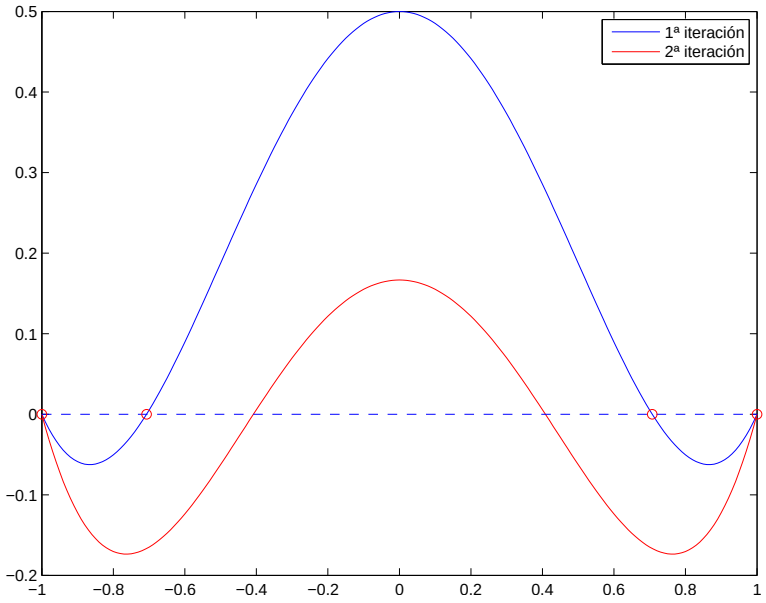


Figura 4.7: Primera y segunda iteración en el algoritmo de Remez

segunda iteración de Remez es

$$p(x) = -\frac{3}{4} + \frac{7}{4}x^2$$

y $d = 0$. Se calculan los extremos de $f - p$ que corresponden a $x = 0$ y $x = \pm\sqrt{\frac{7}{8}}$. De nuevo el número de extremos es insuficiente. $\diamond$

4.9 Ejercicios

▬ EJERCICIO 40 *Determinar la mejor aproximación de x^3 en $\mathcal{P}_2$ usando el producto escalar de Chebyshev.*

Solución: La sucesión $\{\sqrt{\frac{2}{\pi}}T_n\}$ es ortonormal respecto al producto escalar de Chebyshev. Consecuentemente, los polinomios

$$\{\sqrt{\frac{1}{\pi}}, \sqrt{\frac{2}{\pi}}x, \sqrt{\frac{2}{\pi}}(2x^2-1)\}$$

forman una base ortonormal de $\mathcal{P}_2$. La mejor aproximación de cualquier función continua f está dada por

$$g = \frac{1}{\pi}\langle 1, f\rangle + \frac{2}{\pi}\langle x, f\rangle x + \frac{2}{\pi}\left\langle 2x^2-1, f\right\rangle (2x^2-1) = \frac{2}{\pi}\left\langle x, x^3\right\rangle x$$

por ser el peso, una función par y el intervalo de integración, simétrico respecto al origen. Finalmente, se obtiene

$$g = \left(\frac{2}{\pi}\int_{-1}^{1}\frac{x^4}{\sqrt{1-x^2}}dx\right)x = \frac{3x}{4} \qquad \diamond$$

▬ EJERCICIO 41 *Aplicar el procedimiento de orto-normalización de Gram-Schmidt a la base $\{1, x, x^2\}$ de $\mathcal{P}_2$ respecto al siguiente producto escalar*

$$\langle f, g\rangle = \frac{1}{2}\int_0^1 (1+3x^2)f(x)g(x)\ dx.$$

Solución: Puesto que

$$\langle 1, 1\rangle = \frac{1}{2}\int_0^1 (1+3x^2)\ dx = 1,$$

el primer término de la base ortonormal es

$$p_0(x) = 1.$$

Puesto que

$$\langle p_0, x\rangle = \frac{1}{2}\int_0^1 (1+3x^2)x\ dx = \frac{5}{8},$$

$$\left\langle x-\frac{5}{8}, x-\frac{5}{8}\right\rangle = \frac{73}{960},$$

el segundo término de la base ortonormal es

$$p_1(x) = \sqrt{\frac{960}{73}}\left(x - \frac{5}{8}\right).$$

Si se usa un argumento similar, se obtiene que

$$p_2(x) = \frac{1}{2}\sqrt{\frac{7}{10877}}\left(1095x^2 - 1200x + 239\right). \qquad \diamond$$

– EJERCICIO 42 *Hallar la recta que mejor aproxima la gráfica de la función*

$$y = \frac{x}{1+x^2}$$

con la norma inducida por el producto escalar

$$\langle f, g\rangle = \int_0^5 f(x)g(x)\ dx,$$

usando una base de polinomios ortogonales.

Solución: Para que los polinomios 1 y $x - a$ sean ortogonales tiene que verificarse que

$$0 = \langle 1, x-a\rangle = \int_0^5 (x-a)\ dx = \frac{1}{2}\left((5-a)^2 - a^2\right)$$

de donde se deduce que $a = \frac{5}{2}$. La mejor aproximación a la función

$$f(x) = \frac{x}{1+x^2}$$

es

$$g_c(x) = \frac{\langle 1, f\rangle}{\langle 1, 1\rangle} + \frac{\left\langle x - \frac{5}{2}, f\right\rangle}{\left\langle x - \frac{5}{2}, x - \frac{5}{2}\right\rangle}(x - \frac{5}{2}).$$

Puesto que

$$\langle 1, 1\rangle = 5, \ \ \langle 1, f\rangle = \frac{\ln 26}{2},$$

$$\left\langle x - \frac{5}{2}, x - \frac{5}{2}\right\rangle = \frac{125}{12}, \ \ \left\langle x - \frac{5}{2}, f\right\rangle = 5 - \arctan 5 - \frac{5}{4}\ln 26$$

se deduce

$$g_c(x) = \frac{\ln 26}{10} + \frac{60 - 12\arctan 5 - 15\ln 26}{125}(x - \frac{5}{2}) = 0.4329 - 0.0428x. \qquad \diamond$$

▬ EJERCICIO 43 *Hallar la recta que mejor aproxima la gráfica de la función*

$$y = \frac{x}{1+x^2}$$

con el producto discreto

$$\langle f, g\rangle = f(0)g(0) + f(2)g(2) + f(3)g(3) + f(5)g(5)$$

usando una base de polinomios ortogonales. ¿Cuál de las dos aproximaciones (obtenidas en este ejercicio y en el anterior) es mejor en el sentido de la norma uniforme?.

Solución: Para que los polinomios 1 y $x - a$ sean ortogonales tiene que verificarse que

$$0 = \langle 1, x-a\rangle = 10 - 4a$$

de donde se deduce que $a = \frac{5}{2}$. Así pues, el conjunto $\{1, x - \frac{5}{2}\}$ es una base ortogonal del subespacio de los polinomios de grado menor o igual que 1.

La mejor aproximación a la función $f(x) = \frac{x}{1+x^2}$ es

$$g_d(x) = \frac{\langle 1, f\rangle}{\langle 1, 1\rangle} + \frac{\left\langle x - \frac{5}{2}, f\right\rangle}{\left\langle x - \frac{5}{2}, x - \frac{5}{2}\right\rangle}(x - \frac{5}{2}).$$

Puesto que

$$\langle 1, 1\rangle = 4, \ \ \langle 1, f\rangle = \frac{58}{65},$$

$$\left\langle x - \frac{5}{2}, x - \frac{5}{2}\right\rangle = 13, \ \ \left\langle x - \frac{5}{2}, f\right\rangle = \frac{28}{65}$$

se deduce

$$g_d(x) = \frac{237}{1690} + \frac{28}{845}x.$$

Fácilmente se comprueba que (véase figura 4.8)

$$\|f - g_c\|_\infty = 0.4329, \quad \|f - g_d\|_\infty = 0.3277 \qquad \diamond$$

▬ EJERCICIO 44 *Construir los cuatro primeros polinomios de grado creciente, de coeficiente principal igual a 1, que sean ortogonales respecto al producto escalar*

$$\langle f, g\rangle = \int_{-1}^{1} |x| f(x)g(x)\ dx.$$

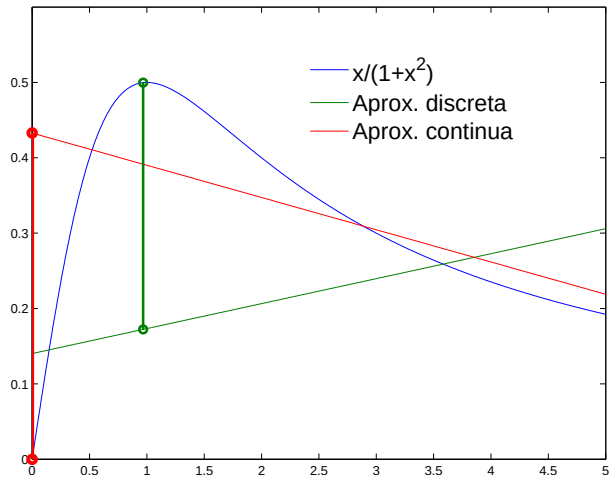


Figura 4.8: Comparación de mejor aproximación continua y discreta

Solución: Antes de comenzar con los cálculos es preciso notar que

$$\int_{-1}^{1} |x| f(x)\, dx = 0$$

para toda función impar en $[-1, 1]$. Así pues, si se utiliza la fórmula de recurrencia de tres términos introducida en este capítulo, se tiene que

$$a_n = \frac{\langle x p_{n-1}, p_{n-1} \rangle}{\langle p_{n-1}, p_{n-1} \rangle} = 0.$$

De este modo se obtiene

a_n	b_n	$p_n(x)$
.	.	1
0	.	x
0	$\frac{1}{2}$	$x^2 - \frac{1}{2}$
0	$\frac{1}{6}$	$x^3 - \frac{2}{3}x.$

$\diamond$

– EJERCICIO 45 *Determinar la mejor aproximación a la función $f(x) = e^x$ en $\mathcal{P}_2$ usando la norma asociada al siguiente producto escalar*

$$\langle f, g \rangle = \int_0^1 (f(x)g(x) + f'(x)g'(x))\, dx.$$

Solución: La matriz de Gram de la base $\{1, x, x^2\}$ de $\mathcal{P}_2$ es

$$G = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{4}{3} & \frac{5}{4} \\ \frac{1}{3} & \frac{5}{4} & \frac{23}{15} \end{pmatrix}$$

el vector de términos independientes por

$$\bar{f} = \begin{pmatrix} e-1 \\ e \\ e \end{pmatrix}.$$

Si se resuelve el sistema lineal $G\alpha = \bar{f}$ se obtiene que la mejor aproximación es

$$p(x) = 1.0011 + 0.8711x + 0.8451x^2. \qquad \diamond$$

▬ EJERCICIO 46 *Determinar si el polinomio $g(x) = 1 - x^2$ es la mejor aproximación uniforme a la función $f(x) = x^3$ en el espacio de los polinomios generados por $\{1, x^2\}$ en $[-1, 1]$ usando directamente el criterio de Kolmogorov.*

Solución: Se comprueba que

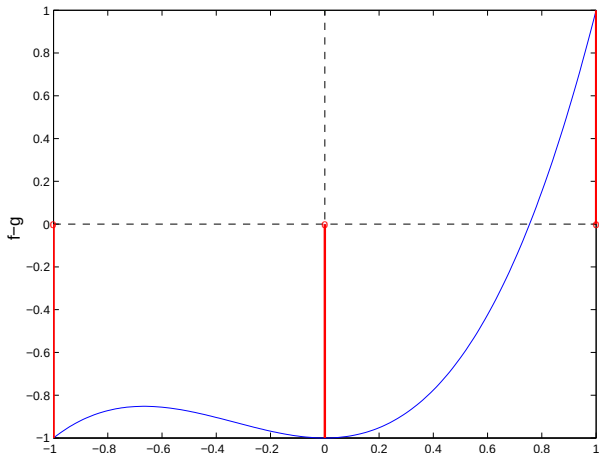


Figura 4.9: Gráfica de $f(x) - g(x) = x^3 + x^2 - 1$

$$A_{f-g} = \{-1, 0, 1\}$$

buscando los puntos donde se alcanzan los valores extremos de

$$f - g = x^3 + x^2 - 1.$$

Si un polinomio $q(x) = a + bx^2$ tiene el mismo signo que $f - g$ en A_{f-g} entonces

$$\begin{aligned} a + b &< 0, \\ a &< 0, \\ a + b &> 0 \end{aligned}$$

Puesto que estas relaciones son incompatibles, se concluye que g es la mejor aproximación uniforme a f. Es importante destacar que el espacio de aproximaciones U no es el espacio de polinomios de grado menor o igual que n para algún n, y por ello no está justificado el uso del teorema de Chebyshev. De hecho, no hay alternancia de signo en los valores extremos de $\|f-g\|_\infty$. $\diamond$

CAPÍTULO 5

Interpolación de funciones

5.1 Introducción

¿Qué ocurre cuando el ajuste de un polinomio a una función en el sentido discreto de mínimos cuadrados es perfecto?. Para precisar esta cuestión, se considera una mejor aproximación p_n de una función f en $U = \mathcal{P}_n$, en el sentido de mínimos cuadrados discretos en el siguiente conjunto de nodos

$$x_0 < x_1 < \cdots < x_{m-1} < x_m.$$

Es conveniente separar tres situaciones distintas:

1. El número de nodos menos 1 es menor el máximo grado de los polinomios de U, es decir, $m < n$. En este caso, la seminorma $|\ |_2$ inducida por el producto discreto, no es una norma ya que el polinomio $q \in \mathcal{P}_n$ definido por

   $$q(x) = (x - x_0)(x - x_1) \cdots (x - x_m)$$

   verifica que $|q|_2 = 0$ y no es nulo. Por lo tanto, la mejor aproximación p_n de f en $\mathcal{P}_n$ no es única. De hecho, cualquier polinomio de la forma $p_n + aq$ para una constante arbitraria a, es una mejor aproximación.

Esto es consecuencia de que

$$\begin{aligned} |f-(p_n+aq)|_2 &= \left(\sum_{i=0}^{m}(f(x_i)-(p_n(x_i)+aq(x_i)))^2\right)^{\frac{1}{2}} \\ &= \left(\sum_{i=0}^{m}(f(x_i)-p_n(x_i))^2\right)^{\frac{1}{2}} = |f-p_n|_2. \end{aligned}$$

2. El número de nodos menos 1 es igual el máximo grado de los polinomios de U, es decir, $m = n$. La seminorma inducida por el producto es una norma y la mejor aproximación es única. En este caso, se cumple que la aproximación es perfecta

   $$|f-p_n|_2 = 0.$$

   La justificación de este resultado se hará en la siguiente sección. De este hecho se desprende que $f(x_i) = p_n(x_i)$ para $i = 0, 1, \cdots, n$.

3. El número de nodos menos 1 es mayor el máximo grado de los polinomios de U, es decir, $m > n$. La seminorma inducida por el producto es una norma y la mejor aproximación es única pero $|f-p_n|_2$ no es necesariamente 0.

Este capítulo está dedicado al análisis de la situación $m = n$. En este caso, el polinomio p_n de grado menor o igual que n, que es la mejor aproximación en el sentido de mínimos cuadrados discreta, alcanza la distancia 0, ya que

$$p_n(x_i) = f(x_i)$$

para $i = 0, 1, \cdots, n$ y se conoce como polinomio de interpolación de la función f en los nodos $\{x_i : i = 0, 1, \cdots, n\}$. El hecho de que el número de nodos de interpolación y la dimensión del espacio de polinomios coincidan, es esencial para que esto ocurra.

Se puede considerar el polinomio de interpolación como una combinación lineal de los valores de f en los nodos

$$p_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i) \tag{5.1}$$

en cada punto x. Los coeficientes de la combinación lineal dependerán del punto considerado x. Es frecuente referirse a ellos como las funciones de forma de la aproximación. La propiedad más deseable de las funciones de forma es que no dependan de la función f.

En la determinación de las funciones de forma, se puede utilizar el hecho de que la aproximación es exacta cuando la propia función a interpolar es un polinomio de grado menor o igual que n. Antes de entrar en la teoría general, para ilustrar esta idea se analiza en detalle un ejemplo muy simple.

▪ EJEMPLO **30** Si se quiere usar un método de interpolación para que una función f conocida en los puntos $x_0 < x_1$, pueda ser aproximada por el siguiente polinomio

$$p(x) = l_0(x)f(x_0) + l_1(x)f(x_1),$$

de modo que esta aproximación sea exacta para todos los polinomios de grado menor o igual que 1, debe cumplirse que

$$\begin{aligned} l_0(x) + l_1(x) &= 1 \\ l_0(x)x_0 + l_1(x)x_1 &= x \end{aligned}$$

ya que si es exacta para la base $\{1, x\}$ lo será para cualquier polinomio de grado menor o igual que 1. En forma matricial, estas ecuaciones resultan ser las siguientes

$$\begin{pmatrix} 1 & 1 \\ x_0 & x_1 \end{pmatrix} \begin{pmatrix} l_0(x) \\ l_1(x) \end{pmatrix} = \begin{pmatrix} 1 \\ x \end{pmatrix}.$$

Este sistema lineal de ecuaciones tiene como solución única la siguiente

$$l_0(x) = \frac{x - x_1}{x_0 - x_1}, \qquad l_1(x) = \frac{x - x_0}{x_1 - x_0}.$$

Es decir, el problema de interpolación tiene solución única y el polinomio de interpolación asociado es el siguiente

$$p(x) = \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1). \qquad \diamond$$

5.2 Interpolación de Lagrange

Los ideas utilizadas en el ejemplo de la sección anterior pueden ser generalizadas al caso de un conjunto de $n + 1$ puntos $\{x_i : \ i = 0, 1, ..., n\}$. De hecho, las funciones de forma definidas por 5.1, son únicas como se prueba en el siguiente

Teorema 21 *Existe un único conjunto $L = \{l_0, l_1, ..., l_n\}$ de funciones de forma, correspondientes al conjunto de $n+1$ nodos $\{x_i : i = 0, 1, ..., n\}$, definidas por*

$$l_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}$$

Además, L constituye una base de $\mathcal{P}_n$ que se conoce como base de Lagrange.

Demostración: Puesto que el polinomio de interpolación correspondiente a un polinomio p de $\mathcal{P}_n$ es el propio polinomio, las componentes de p en esta base, son los valores del polinomio en los nodos. Es decir, su expresión en términos de las funciones de L, es

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

lo que prueba que L es un sistema de generadores de $\mathcal{P}_n$. Puesto que hay $n+1$ funciones de forma y la dimensión de $\mathcal{P}_n$ es $n+1$, se concluye que L es una base de $\mathcal{P}_n$. De la unicidad de las componentes de un vector en una base se desprende la unicidad de L.

Si se expresan los elementos de la base $\{1, x, ..., x^n\}$ de $\mathcal{P}_n$ en términos de la base de Lagrange, se obtiene

$$x^j = \sum_{i=0}^{n} l_i(x) x_i^j.$$

En forma matricial, las relaciones anteriores se escriben como

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} l_0 \\ l_1 \\ \vdots \\ l_n \end{pmatrix} = \begin{pmatrix} 1 \\ x \\ \vdots \\ x^n \end{pmatrix} \tag{5.2}$$

La matriz de coeficientes A de este sistema es la matriz de Vandermonde del conjunto $\{x_0, x_1, ..., x_n\}$ y por consiguiente es invertible. Los elementos de la base de Lagrange están dados por

$$l_i(x) = \frac{|A_i|}{|A|}$$

donde A_i representa la matriz de coeficientes en la que se ha sustituido la columna i-ésima por el término independiente. Así pues, l_i verifica

$$l_i(x_j) = \delta_{ij}$$

donde δ_{ij} representa la delta de Kronecker (1 si $i = j$ y 0 si $i \neq j$). Estas relaciones pueden ser consideradas en si mismas como un problema de interpolación y consecuentemente determinan unívocamente las funciones básicas. De hecho, la siguiente expresión de las funciones básicas

$$l_i(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}$$

puede comprobarse directamente probando que verifica esta propiedad. ◇

El conocimiento explícito de la base de Lagrange permite, dada una función cualquiera f, construir un polinomio de grado menor o igual que n

$$p(x) = \sum_{i=0}^{n} l_i(x) f(x_i)$$

que interpole a f en $n + 1$ puntos $x_0, x_1, ..., x_n$. Conviene recordar que la solución del problema de interpolación es la solución del problema de mejor aproximación de una función en el espacio de los polinomios de grado n respecto a la seminorma definida por el producto discreto

$$\langle f, g \rangle = \sum_{i=0}^{n} f(x_i) g(x_i).$$

La base de Lagrange es ortonormal respecto a $\langle\ ,\ \rangle$. En efecto, se cumple que

$$\langle l_i, l_j \rangle = \sum_{k=0}^{n} \delta_{ik} \delta_{kj} = \delta_{ij}.$$

5.3 Método de Newton

La base $\{1, x, x^2, \cdots, x^n\}$ de $\mathcal{P}_n$ no es adecuada para el análisis y cálculo de los polinomios de interpolación porque no implica cómo los nodos de la interpolación están distribuidos en el intervalo. En la base de Lagrange, todos los polinomios básicos son del mismo grado n y el coste computacional para evaluarla en un conjunto de puntos es más elevado que en el caso de los monomios. Otro modo alternativo de construir el polinomio de Lagrange es el que se basa en el uso de la base

$$\{1, x - x_0, (x - x_0)(x - x_1), \cdots, (x - x_0)(x - x_1) \cdots (x - x_{n-1})\}$$

de $\mathcal{P}_n$ formada por los polinomios

$$\omega_i(x) = (x - x_0)(x - x_1) \cdots (x - x_{i-1})$$

para $i = 1, \cdots, n$ y el polinomio constante $\omega_0(x) = 1$. Esta base incorpora la información de cómo se distribuyen los nodos de la interpolación (salvo el último) y su grado es ascendente. Se conoce en este contexto, como la base de Newton.

Es importante destacar que un polinomio

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

expresado en la base de Newton, admite la siguiente representación anidada

$$p(x) = a_0 + (x - x_0)(a_1 + (x - x_1)(a_2 + \cdots + (x - x_{n-2})(a_{n-1} + a_n(x - x_{n-1}) \cdots))).$$

Así pues, para calcular el valor de p en un punto x, se podría utilizar una regla de Ruffini adaptada a esta base. Por ejemplo, en el caso $n = 2$, se podría proceder mediante la tabla

$x - x_1$	$x - x_0$	
a_2	a_1	a_0
	$(x - x_1)a_2$	$(x - x_0)(a_1 + a_2(x - x_1))$
a_2	$a_1 + a_2(x - x_1)$	$a_0\omega_0(x) + a_1\omega_1(x) + a_2\omega_2(x).$

para evaluar el polinomio de Newton, con mínimo coste computacional.

Para estudiar cómo se puede expresar el polinomio de interpolación en términos de esta base, se hace uso de la propiedad de exactitud

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

del polinomio de interpolación sobre los polinomios p de grado menor o igual que n. En particular, si se utiliza esta condición sobre los elementos básicos ω_i para $i = 0, 1, \cdots, n$ y se tiene en cuenta que $\omega_i(x_j) = 0$ si $j < i$, se obtiene el sistema lineal

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & \omega_1(x_1) & \cdots & \omega_1(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n(x_n) \end{pmatrix} \begin{pmatrix} l_0 \\ l_1 \\ \vdots \\ l_n \end{pmatrix} = \begin{pmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_n \end{pmatrix} \tag{5.3}$$

que relaciona la base Lagrange con la de Newton. La traspuesta de la matriz triangular superior D asociada a este sistema es la matriz de cambio de base

que permite relacionar las componentes de cualquier polinomio $p \in \mathcal{P}_n$ en ambas bases. En particular, las componentes del polinomio de interpolación en ambas bases están relacionadas por el sistema lineal

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & \omega_1(x_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_1(x_n) & \cdots & \omega_n(x_n) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

donde a_i representan las componentes del polinomio de interpolación en la base de Newton. Las componentes a_i se llaman diferencias divididas (ó cociente) de la función f en los nodos $\{x_i : i = 0, 1, \cdots, n\}$ y la expresión del polinomio de interpolación se conoce como polinomio de Newton. Es habitual representar la componente a_i del polinomio de interpolación de Newton p_n por $f[x_0, x_1, \cdots, x_i]$, de modo que

$$p_n(x) = f[x_0] + f[x_0, x_1]\omega_1(x) + \cdots + f[x_0, x_1, \cdots, x_n]\omega_n(x).$$

Puesto que las componentes del polinomio de interpolación son únicas en cualquier base, las diferencias divididas son simétricas. En efecto, ya que el polinomio de interpolación es único, independientemente de la ordenación de los datos, si se efectúa una permutación j de los índices naturales, se cumple que

$$f[x_0, x_1, \cdots, x_n] = f[x_{j_0}, x_{j_1}, \cdots, x_{j_n}].$$

La resolución del sistema lineal anterior puede llevarse a cabo de un modo muy próximo al método de eliminación de Gauss y que a continuación se describe para $n = 3$

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & x_1 - x_0 & 0 \\ 1 & x_2 - x_0 & (x_2 - x_0)(x_2 - x_1) \end{pmatrix} \begin{pmatrix} f[x_0] \\ f[x_0, x_1] \\ f[x_0, x_1, x_2] \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & x_1 - x_0 & 0 \\ 0 & x_2 - x_1 & (x_2 - x_0)(x_2 - x_1) \end{pmatrix} \begin{pmatrix} f[x_0] \\ f[x_0, x_1] \\ f[x_0, x_1, x_2] \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 - f_0 \\ f_2 - f_1 \end{pmatrix}$$

Eliminación restando filas consecutivas

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & x_2 - x_0 \end{pmatrix} \begin{pmatrix} f[x_0] \\ f[x_0, x_1] \\ f[x_0, x_1, x_2] \end{pmatrix} = \begin{pmatrix} f_0 \\ \frac{f_1 - f_0}{x_1 - x_0} \\ \frac{f_2 - f_1}{x_2 - x_1} \end{pmatrix}$$

Normalización de pivotes

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & x_2 - x_0 \end{pmatrix} \begin{pmatrix} f[x_0] \\ f[x_0, x_1] \\ f[x_0, x_1, x_2] \end{pmatrix} = \begin{pmatrix} f_0 \\ \frac{f_1 - f_0}{x_1 - x_0} \\ \frac{f_2 - f_1}{x_2 - x_1} - \frac{f_1 - f_0}{x_1 - x_0} \end{pmatrix}$$

Eliminación restando filas consecutivas

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f[x_0] \\ f[x_0, x_1] \\ f[x_0, x_1, x_2] \end{pmatrix} = \begin{pmatrix} f_0 \\ \frac{f_1 - f_0}{x_1 - x_0} \\ \frac{\frac{f_2 - f_1}{x_2 - x_1} - \frac{f_1 - f_0}{x_1 - x_0}}{x_2 - x_0} \end{pmatrix}$$

Normalización de pivotes. ◇

Este ejemplo parece justificar la notación y el nombre de diferencias divididas.

En el caso general, se puede probar que las diferencias divididas de orden j están dadas por

$$f[x_0, x_1, \cdots, x_j] = \frac{f[x_1, x_2, \cdots, x_j] - f[x_0, x_1, \cdots, x_{j-1}]}{x_j - x_0}$$

para $j = 1, 2 \cdots, n$.

Tradicionalmente, se calculan las diferencias en tablas organizadas como sigue:

$$\begin{array}{lll} x_0 & f(x_0) & \\ & & \frac{f(x_1)-f(x_0)}{x_1-x_0} \\ x_1 & f(x_1) & \\ & & \frac{f(x_2)-f(x_1)}{x_2-x_1} \\ x_2 & f(x_2) & \\ & \vdots & \\ \vdots & & \\ & \vdots & \\ x_n & f(x_n) & \end{array} \qquad \begin{array}{l} \\ \\ \frac{\frac{f(x_2)-f(x_1)}{x_2-x_1} - \frac{f(x_1)-f(x_0)}{x_1-x_0}}{x_2-x_0} \\ \\ \vdots \\ \\ \vdots \end{array}$$

▬ EJERCICIO 47 *Dada la siguiente tabla de valores*

x	0.1	0.2	0.3	0.4	0.5	0.6
$f(x)$	0.7	0.8	1	1.15	1.25	1.3

se pide:

1. *Calcular la tabla de diferencias divididas.*
2. *Utilizando la fórmula de Newton, calcular el valor interpolado de* $f(0.55)$.

Solución: La tabla de las diferencias dividas es la siguiente:

$$
\begin{array}{lllllll}
0.1 & 0.7 & & & & & \\
 & & 1 & & & & \\
0.2 & 0.8 & & 5 & & & \\
 & & 2 & & -25 & & \\
0.3 & 1 & & -2.5 & & 62.5 & \\
 & & 1.5 & & 0 & & -125 \\
0.4 & 1.15 & & -2.5 & & 0 & \\
 & & 1 & & 0 & & \\
0.5 & 1.25 & & -2.5 & & & \\
 & & 0.5 & & & & \\
0.6 & 1.3 & & & & &
\end{array}
$$

El polinomio de interpolación de Newton viene dado por

$$
\begin{aligned}
p_5(x) &= 0.7 + (x - 0.1) \\
&+ 5(x - 0.1)(x - 0.2) \\
&- 25(x - 0.1)(x - 0.2)(x - 0.3) \\
&+ 62.5(x - 0.1)(x - 0.2)(x - 0.3)(x - 0.4) \\
&- 125(x - 0.1)(x - 0.2)(x - 0.3)(x - 0.4)(x - 0.5).
\end{aligned}
$$

La evaluación de este polinomio para un valor de x determinado puede organizarse en un algoritmo similar al de Horner

x	$x - x_4$	$x - x_3$	$x - x_2$	$x - x_1$	$x - x_0$	
0.55	0.05	0.15	0.25	0.35	0.45	
	$f[x_0, \cdots, x_5]$	$f[x_0, \cdots, x_4]$	$f[x_0, \cdots, x_3]$	$f[x_0, x_1, x_2]$	$f[x_0, x_1]$	$f[x_0]$
	−125	62.5	−25	5	1	0.7
		−6.25	8.4375	−4.1406	0.3007	0.5853
	−125	56.25	−16.5625	0.8593	1.3007	1.2853

De la tabla se deduce que el valor aproximado de $f(0.55)$ es 1.28535156 ◇

Se examina ahora el caso en el que los nodos están uniformemente distribuidos. Si $h = x_{i+1} - x_i$ para $i = 0, 1, \cdots, n-1$ entonces

$$\omega_i(x_j) = \begin{cases} \frac{j!}{(j-i)!}h^i, & \text{si } i \leq j; \\ 0, & \text{si } i > j. \end{cases} = i!h^i \begin{cases} \binom{j}{i}, & \text{si } i \leq j; \\ 0, & \text{si } i > j. \end{cases}$$

para $i, j = 0, 1, \cdots, n$. En este caso el sistema que permite calcular las diferencias divididas es el siguiente

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & n & \frac{n(n-1)}{2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \Delta^0 f(x_0) \\ \Delta^1 f(x_0) \\ \vdots \\ \Delta^n f(x_0) \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

donde $\Delta^i f(x_0) = f[x_0, x_1, \cdots, x_i]h^i i!$ se llama diferencia finita de orden i. Si se resuelve este sistema lineal se obtiene

$$\Delta^i f(x_0) = \sum_{k=0}^{i} \binom{i}{k} (-1)^{i+k} f(x_k)$$

lo que permite interpretar la diferencia finita como la aplicación i-veces reiterada del operador $\Delta f(x_j) = f(x_{j+1}) - f(x_j)$.

Por otra parte, el polinomio de interpolación se puede expresar en términos de la diferencias finitas como

$$p_n(x) = \sum_{i=0}^{n} \frac{\Delta^i f}{i!h^i} \omega_i(x).$$

5.4 Error en la interpolación de Lagrange

El siguiente teorema permitirá obtener estimaciones del error que se comete cuando se aproxima el valor de una función en un punto por el valor de su polinomio de interpolación de Lagrange en ese punto, en el caso de una función suficientemente regular.

– Teorema 22 *El error que se comete al aproximar una función f de clase $n+1$ en el intervalo $[a, b]$ por su polinomio de interpolación de Lagrange de grado menor o igual que n en los nodos*

$$a = x_0 < x_1 < \cdots < x_n = b$$

está dado para cualquier valor de $x \in [a,b]$*, por la siguiente fórmula*

$$E(x) = f(x) - p_n(x) = \frac{(x-x_0)\cdots(x-x_n)}{(n+1)!} f^{(n+1)}(\xi)$$

para algún número $\xi \in (a,b)$ *(que depende de* x*).*

Demostración: Fijado un punto x diferente de los nodos, se considera la función auxiliar g definida por

$$g(t) = E(t) - \frac{\omega_{n+1}(t)}{\omega_{n+1}(x)} E(x)$$

para todo $t \in [a,b]$. La función g se anula en los $n+1$ nodos de interpolación y en x. Del teorema de Rolle se deduce que g' se anula en $n+1$ puntos distintos. reiterando el razonamiento se deduce que $g^{(n+1)}$ se anula en al menos un punto ξ en (a,b). Entonces se tiene

$$E^{(n+1)}(\xi) - \frac{(n+1)!}{\omega_{n+1}(x)} E(x) = 0$$

lo que conduce a

$$E(x) = \frac{\omega_{n+1}(x)}{(n+1)!} E^{(n+1)}(\xi) = \frac{\omega_{n+1}(x)}{(n+1)!} f^{(n+1)}(\xi)$$

ya que $\frac{d^{n+1}p_n}{dx^n} = 0$. $\diamond$

Obviamente el polinomio de interpolación de grado $n+1$ a f en los puntos $x_0, x_1, \cdots, x_n$ y x toma el mismo valor que f en x. Es decir

$$f(x) = p_n(x) + f[x_0, x_1, \cdots, x_n, x]\omega_{n+1}(x). \tag{5.4}$$

De esta igualdad y de la fórmula del error de interpolación se desprende

$$f[x_0, x_1, \cdots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

En particular se tiene que

$$f[x_0, x_1, \cdots, x_n] = \frac{f^{(n)}(\xi)}{n!}$$

para algún $\xi \in [a,b]$.

▬ EJERCICIO 48 *Estimar del modo más fino que sea posible el error máximo que se comete al calcular* $\sqrt{x}$*, interpolando el valor de la función en los puntos* x_0 *y* x_1*, contenidos en el intervalo* $[1,2]$.

Solución: El error de interpolación de la función $f(x) = \sqrt{x}$ está dado por la fórmula

$$E(x) = \frac{f''(\xi)}{2}(x - x_0)(x - x_1)$$

siendo ξ un punto desconocido del intervalo $[x_0, x_1]$. Puesto que x_0 y x_1 son arbitrarios en el intervalo $[1, 2]$, se trata de buscar una cota del error que sea independiente de estos dos puntos. Esto se puede obtener usando la siguiente acotación

$$|E(x)| \leq \frac{\max_{t\in[x_0,x_1]} |f''(t)|}{2} \max_{t\in[x_0,x_1]} |(t - x_0)(t - x_1)|.$$

Se puede acotar separadamente cada uno de los factores en el segundo miembro de la desigualdad anterior. Así se obtiene que para $t \in [x_0, x_1]$

$$|f''(t)| = \frac{t^{-3/2}}{4} \leq \frac{x_0^{-3/2}}{4} \leq \frac{1}{4}.$$

Puesto que el máximo del polinomio $(t - x_0)(x_1 - t)$ en el intervalo $[x_0, x_1]$ se alcanza en $\frac{x_0+x_1}{2}$, se deduce que

$$\max_{t\in[x_0,x_1]} |(t - x_0)(t - x_1)| = \frac{(x_1 - x_0)^2}{4} \leq \frac{1}{4}.$$

En consecuencia, una estimación del error es la siguiente

$$|E(x)| \leq \frac{1}{32}$$

para todo $x \in [x_0, x_1]$ $\diamond$.

Ahora se plantea la siguiente cuestión: ¿cómo colocar estratégicamente los nodos de interpolación para mejorar la aproximación en un sentido continuo?. El ejercicio anterior muestra que quizás podría refinarse la precisión con la que el polinomio de interpolación se aproxima a la función tomando adecuadamente los nodos. En este sentido parece razonable escoger los nodos x_i de modo que minimicen el máximo de $|(x - x_0)(x - x_1) \cdots (x - x_n)|$ en $[a, b]$. Esto se puede conseguir del modo siguiente: Mediante un cambio de variable lineal se transforma el problema de interpolación en el intervalo $[a, b]$ en otro equivalente en el intervalo $[-1, 1]$. En este intervalo, de acuerdo con el teorema 15 la elección de los nodos óptimos corresponde a las raíces del polinomio de Chebyshev de grado $n + 1$ en cuyo caso se tiene que

$$\max_{x\in[-1,1]} |(x - x_0)(x - x_1) \cdots (x - x_n)| = \frac{1}{2^n}.$$

5.5 Algoritmos de Aitken y Neville

Las situaciones en las que se usan técnicas de interpolación polinomial son muy variadas. En algunas de estas situaciones, la interpolación es parte de un procedimiento más amplio que la utiliza como aproximación de una función, beneficiándose del hecho de que un polinomio se puede fácilmente derivar o integrar. En esos casos, el cálculo de las funciones básicas o las diferencias divididas parece el objetivo básico. Sin embargo, es fácil imaginar situaciones más sencillas en las que se disponen de algunos datos sobre el valor de la función en algunos puntos y lo único que se pretende es obtener un valor aproximado de la función en un punto en el que no se dispone de información. Uno podría razonablemente pensar en evitar cálculos innecesarios y realizar las operaciones imprescindibles para obtener esta aproximación en el referido punto. Del mismo modo que la evaluación de un polinomio arbitrario no se hace por simple sustitución de la variable por su valor, calculando separadamente cada una de las potencias, ya que ello implica un coste computacional alto y que existen otros procedimientos elementales como el algoritmo de Horner que permiten reducirlo considerablemente, también el coste computacional de la evaluación de polinomio de interpolación en un valor de la variable x, puede reducirse considerablemente usando los algoritmos de Aitken y Neville.

Para comprender la base teórica en la que estos procedimientos se sustentan, se analiza la siguiente situación: Se conoce el valor de una función en $n+1$ puntos $X = \{x_0, x_1, \cdots, x_n\}$ distintos. Se supone que se ha construido el polinomio de interpolación p_{n-1} de grado $n-1$ correspondiente en los puntos $\{x_0, x_1, \cdots, x_{n-1}\}$, directamente se comprueba que

$$p_n(x) = p_{n-1}(x) + (f_n - p_{n-1}(x_n))l_n(x)$$

es el polinomio de interpolación correspondiente a todos los datos (l_n es la función básica de Lagrange asociada al nodo x_n) ya que es único y $p_n(x_i) = f_i$ para todo $i = 0, 1, \cdots, n$.

Esta consideración es aún válida si el punto separado del conjunto X es arbitrario y no necesariamente el último. Por esta razón, se puede diseñar el siguiente procedimiento: Del conjunto X se extrae un punto x_j y se construye el polinomio de interpolación p^j_{n-1} con los puntos restantes. A continuación, se repone el punto x_j a X y se repite el procedimiento para otro punto distinto x_k, obteniéndose un segundo polinomio de interpolación p^k_{n-1} de grado menor o igual que $n-1$. De acuerdo con la relación anterior se tiene que

$$p_n(x) = p^j_{n-1}(x) + (f_j - p^j_{n-1}(x_j))l_j(x)$$
$$p_n(x) = p^k_{n-1}(x) + (f_k - p^k_{n-1}(x_k))l_k(x)$$

Si se multiplican, la primera igualdad por $x - x_j$, y la segunda por $x - x_k$, se deduce que

$$(x_k - x_j)p_n(x) = (x - x_j)p_{n-1}^j(x) - (x - x_k)p_{n-1}^k(x) + q(x)$$

donde q es el polinomio definido por

$$q(x) = (x - x_j)(f_j - p_{n-1}^j(x_j))l_j(x) - (x - x_k)(f_k - p_{n-1}^k(x_k))l_k(x)$$

El polinomio q es grado n ya que es la diferencia de dos polinomios de grado n. Además se anula en todos los puntos de X. Consecuentemente, el polinomio q es el nulo y por ello se obtiene la fórmula

$$p_n(x) = \frac{(x - x_j)p_{n-1}^j(x) - (x - x_k)p_{n-1}^k(x)}{x_k - x_j}.$$

Esta fórmula permite justificar los algoritmos de Aitken y Neville que se usan para evaluar el polinomio de interpolación en un punto sin necesidad de calcular sus coeficientes o evaluar las funciones de forma de la interpolación. La diferencia entre ellos está en qué puntos x_j y x_k se seleccionan. La idea de Aitken es utilizar siempre x_{n-1} y x_n mientras que la de Neville es utilizar x_0 y x_n. Por ejemplo, para calcular el polinomio de interpolación $p(x; x_0, x_1, x_2, x_3)$, construido en los puntos $\{x_0, x_1, x_2, x_3\}$, mediante dos polinomios construidos usando tres puntos, el algoritmo de Aitken usa los polinomios $p(x; x_0, x_1, x_2)$ y $p(x; x_0, x_1, x_3)$ mientras que el algoritmo de Neville usa los polinomios $p(x; x_0, x_1, x_2)$ y $p(x; x_1, x_2, x_3)$.

Se pueden organizar los cálculos de modo recurrente en alguna de las disposiciones siguientes:

Algoritmo de Aitken

x_0	$x - x_0$	f_0			
x_1	$x - x_1$	f_1	$p(x; x_0, x_1)$		
x_2	$x - x_2$	f_2	$p(x; x_0, x_2)$	$p(x; x_0, x_1, x_2)$	
x_3	$x - x_3$	f_3	$p(x; x_0, x_3)$	$p(x; x_0, x_1, x_3)$	$p(x; x_0, x_1, x_2, x_3)$

En esta tabla, los polinomios $p(x; x_0, x_1, x_2)$ y $p(x; x_0, x_1, x_2, x_3)$ se calculan del modo siguiente

$$p(x; x_0, x_1, x_2) = \frac{(x - x_1)p(x; x_0, x_2) - (x - x_2)p(x; x_0, x_1)}{x_2 - x_1}.$$

$$p(x; x_0, x_1, x_2, x_3) = \frac{(x - x_2)p(x; x_0, x_1, x_3) - (x - x_3)p(x; x_0, x_1, x_2)}{x_3 - x_2}.$$

Algoritmo de Neville

x_0	$x-x_0$	f_0			
x_1	$x-x_1$	f_1	$p(x;x_0,x_1)$		
x_2	$x-x_2$	f_2	$p(x;x_1,x_2)$	$p(x;x_0,x_1,x_2)$	
x_3	$x-x_3$	f_3	$p(x;x_2,x_3)$	$p(x;x_1,x_2,x_3)$	$p(x;x_0,x_1,x_2,x_3)$

En esta tabla, los polinomios $p(x;x_0,x_1,x_2)$ y $p(x;x_0,x_1,x_2,x_3)$ se calculan del modo siguiente

$$p(x;x_0,x_1,x_2)=\frac{(x-x_0)p(x;x_1,x_2)-(x-x_2)p(x;x_0,x_1)}{x_2-x_0}.$$

$$p(x;x_0,x_1,x_2,x_3)=\frac{(x-x_0)p(x;x_1,x_2,x_3)-(x-x_3)p(x;x_0,x_1,x_2)}{x_3-x_0}.$$

EJERCICIO 49 *Completar la siguiente tabla*

x_i	$x-x_i$	$f(x_i)$	$p(x;x_i,x_{i+1})$	$p(x;x_i,x_{i+1},x_{i+2})$	$p(x;x_0,x_1,x_2,x_3)$
-2	2.5	*-1*			
-1	1.5	-0.5	*0.25*		
0	0.5	0	0.25	*0.25*	
1	?	?	?	?	*0.25*

que pone en práctica el algoritmo de Neville para interpolar una función en un punto x desconocido. Determinar el punto x en el que se ha interpolado la función.

Solución: Los valores desconocidos de la tabla son

x_i	$x-x_i$	$f(x_i)$	$p(x;x_i,x_{i+1})$	$p(x;x_i,x_{i+1},x_{i+2})$	$p(x;x_i,x_j,x_k,x_p)$
-2	2.5	-1			
-1	1.5	-0.5	0.25		
0	0.5	0	0.25	0.25	
1	$x-1$	$f(1)$	$p(x;0,1)$	$p(x;-1,0,1)$	0.25

En primer lugar, puesto que $2.5=x+2$, se tiene que $x=0.5$. Por otra parte, las fórmulas que generan los tres últimos coeficientes de la última fila de la tabla son

$$p(x;0,1)=0.5f(1),$$

$$p(x;-1,0,1) = \frac{1.5p(x;0,1) + 0.125}{2},$$

$$0.25 = \frac{2.5p(x;-1,0,1) + 0.125}{3}.$$

Consecuentemente, de la última ecuación se deduce que $p(x;-1,0,1) = 0.25$. De la anterior se deduce que $p(x;0,1) = 0.25$ y finalmente se obtiene que $f(1) = 0.5$.

Consecuentemente, la tabla completa es

x_i	$x - x_i$	$f(x_i)$	$p(x;x_i,x_j)$	$p(x;x_i,x_j,x_k)$	$p(x;x_i,x_j,x_k,x_p)$
-2	2.5	-1			
-1	1.5	-0.5	0.25		
0	0.5	0	0.25	0.25	
1	-0.5	0.5	0.25	0.25	0.25

◇

5.6 Interpolación compuesta

El control del error de interpolación está en la $(n+1)$ derivada de la función a interpolar. No obstante, funciones con expresiones analíticas sencillas pueden presentar fuertes oscilaciones en la derivadas sucesivas. Por ejemplo, el polinomio de interpolación de Lagrange para la función de Runge $f(x) = \frac{1}{1+x^2}$ definida sobre el intervalo $[-5,5]$ presenta una convergencia muy pobre como muestra la figura 5.1.

Parece razonable plantearse la siguiente cuestión: ¿aumentar el grado del polinomio de interpolación es siempre la mejor estrategia para mejorar la precisión?. No es una cuestión simple dar una respuesta rotunda a esta cuestión. Hay situaciones que evidencian que muchas veces es preferible, a fin de agilizar el cálculo o conseguir una mayor estabilidad en la aproximación, agrupar los datos en pequeños grupos e interpolar independientemente en cada uno de ellos. En la situación extrema, con cada dato se construye un polinomio constante, que se considera como aproximación de la función hasta estar próximos al siguiente dato, en donde se toma como constante el nuevo dato. Se construye de este modo un polinomio constante a trozos de modo similar a como se construye la función redondeo o la función parte entera (interpolación por el punto más próximo).

Un procedimiento más lento, pero más preciso, consiste en considerar los datos a pares y construir el polinomio de primer grado que interpola a ambos datos. Se construye de este modo un polinomio lineal a trozos.

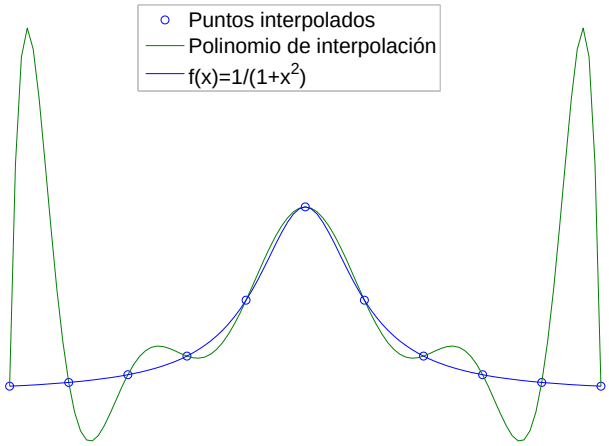


Figura 5.1: Función de Runge

5.7 Interpolación de Hermite.

La interpolación compuesta de Lagrange garantiza la continuidad del polinomio de interpolación aunque no la de sus derivadas. Para obtener continuidad en las derivadas y evitar de este modo que la gráfica tenga puntos angulosos es preciso utilizar la interpolación de Hermite que usa, no sólo el valor de la función en los puntos, sino también el valor de sus derivadas.

▪ Ejemplo **31** De una función f se conocen los datos que están en la siguiente tabla de valores:

x	0	1
$f(x)$	0	1
$f'(x)$	1	-1

.

Si se busca el polinomio de interpolación de Lagrange a estos datos se obtiene que $p_1(x) = x$. Es decir, una recta que pasa por el origen, ajusta perfectamente estos datos. Sin embargo, los datos para la derivada no se ajustan en el punto 1. Si se usa la interpolación de Hermite, es necesario ajustar también los valores de las derivadas. Para ello se necesita usar una clase más amplia de polinomios. Parece razonable, ya que se usan 2 datos más, utilizar polinomios de $\mathcal{P}_3$. Por ello, se intentará determinar los coeficientes del polinomio

$$p_3(x) = a + bx + cx^2 + dx^3$$

para que se cumplan las 4 condiciones de interpolación. Esto conduce al sistema lineal

$$\begin{aligned} a &= 0, \\ a+b+c+d &= 1 \\ b &= 1 \\ b+2c+3d &= -1 \end{aligned}$$

Si se resuelve el sistema se obtiene que polinomio de interpolación es

$$p_3(x) = x + 2x^2 - 2x^3. \qquad \diamond$$

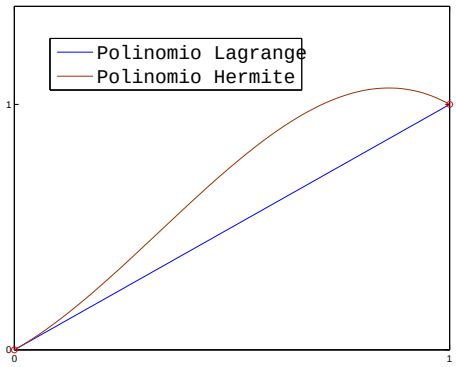


Figura 5.2: Polinomio de Hermite $p_3(x) = x + 2x^2 - 2x^3$

De un modo más general se puede plantear un problema de interpolación de Hermite del modo siguiente: Dados los valores de una función y de su derivada en los nodos $a = x_0 < \cdots < x_n = b$ hallar un polinomio p de grado $2n+1$ tal que

$$p(x_i) = f_i, \;\; p'(x_i) = f_i'$$

para $i = 0, 1, \cdots, n$. Se puede interpretar el polinomio de interpolación de Hermite como la mejor aproximación con respecto a la norma inducida por el producto escalar

$$\langle f, g \rangle = \int_a^b \left(f(x)g(x) + f'(x)g'(x) \right) dx$$

correspondiente al caso en el que la distancia es nula.

A fin de obtener un resultado de existencia y unidad de solución a este problema, así como para desarrollar métodos eficientes de construcción, se consideran los $2n+1$ polinomios de Newton

$$\begin{array}{c|c} \omega_0 = 1 & \\ \omega_1 = x - x_0 & \omega_2 = (x-x_0)^2 \\ \omega_3 = (x-x_0)^2(x-x_1) & \omega_4 = (x-x_0)^2(x-x_1)^2 \\ \vdots & \vdots \\ \omega_{2n} = (x-x_0)^2\cdots(x-x_{n-1})^2 & \omega_{2n+1} = (x-x_0)^2\cdots(x-x_{n-1})^2(x-x_n) \end{array}$$

que generan el espacio de los polinomios de grado menor o igual que $2n+1$. Se busca un polinomio de interpolación que en cada punto dependa linealmente de los datos

$$p(x) = \sum_{i=0}^{n} (l_{2i}(x) f_i + l_{2i+1}(x) f_i').$$

Se impone la condición de que esta fórmula de interpolación sea exacta en el espacio de los polinomios de grado menor o igual que $2n+1$; Es decir, que el polinomio de interpolación de Hermite de un polinomio sea él mismo. Si se impone la exactitud en la base de Newton se obtiene el siguiente sistema lineal

$$\sum_{i=0}^{n} (l_{2i}(x)\omega_j(x_i) + l_{2i+1}(x)\omega_j'(x_i)) = \omega_j(x),$$

para cada $x \in [a,b]$ y para $j = 0, \cdots, 2n+1$. La matriz de coeficientes D de este sistema resulta ser

$$D = \begin{pmatrix} \omega_0(x_0) & \omega_0'(x_0) & \cdots & \omega_0(x_n) & \omega_0'(x_n) \\ \omega_1(x_0) & \omega_1'(x_0) & \cdots & \omega_1(x_n) & \omega_1'(x_n) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \omega_{2n}(x_0) & \omega_{2n}'(x_0) & \cdots & \omega_{2n}(x_n) & \omega_{2n}'(x_n) \\ \omega_{2n+1}(x_0) & \omega_{2n+1}'(x_0) & \cdots & \omega_{2n+1}(x_n) & \omega_{2n+1}'(x_n) \end{pmatrix},$$

el vector de incógnitas, $l = (l_0, l_1, \cdots, l_{2n+1})^t$ y el de términos independientes, $\omega = (\omega_0, \omega_1, \cdots, \omega_{2n+1})^t$.

Si se tiene en cuenta que

$$\begin{array}{lcll} \omega_{2j}(x_i) & = & \omega_j(x_i)^2 = 0, & \text{para} \quad j > i \\ \omega_{2j+1}(x_i) & = & \omega_{2j}(x_i)(x_i - x_j) = 0, & \text{para} \quad j \geq i \\ \omega_{2j}'(x_i) & = & 2\omega_j'(x_i)\omega_j(x_i) = 0, & \text{para} \quad j > i \\ \omega_{2j+1}'(x_i) & = & 2\omega_j'(x_i)\omega_j(x_i)(x_i - x_j) + \omega_j(x_i)^2 = 0, & \text{para} \quad j > i \end{array},$$

se comprueba que la matriz D es triangular superior. Además, puesto que

$$\omega_i(x_i) > 0, \;\; \omega_i'(x_i) > 0$$

para todo $i = 0, 1, \cdots, n$ el determinante de D es positivo.

La resolución de este sistema proporciona las funciones de forma o funciones básicas de Hermite. De modo similar a como se procedió en el caso de la interpolación de Lagrange se pueden calcular las componentes del polinomio de interpolación en la base $\{\omega_i : i = 0, \cdots, 2n+1\}$ resolviendo el sistema

$$D^t \begin{pmatrix} f[x_0] \\ f[x_0, x_0] \\ f[x_0, x_0, x_1] \\ \vdots \\ f[x_0, x_0, \cdots, x_n, x_n] \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f'(x_0) \\ f(x_1) \\ \vdots \\ f'(x_n) \end{pmatrix}.$$

Los cálculos pueden organizarse en tablas como la siguiente

$$\begin{array}{cccccc}
x_0 & f(x_0) & & & & \\
 & & f[x_0, x_0] = f'(x_0) & & & \\
x_0 & f(x_0) & & f[x_0, x_0, x_1] = \frac{f[x_0, x_1] - f[x_0, x_0]}{x_1 - x_0} & & \\
 & & f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} & & & \\
x_1 & f(x_1) & & f[x_0, x_1, x_1] = \frac{f[x_1, x_1] - f[x_0, x_1]}{x_1 - x_0} & \cdots & \\
 & & f[x_1, x_1] = f'(x_1) & & & \\
x_1 & f(x_1) & & & & \\
\vdots & \vdots & \vdots & \vdots & & \\
x_n & f(x_n) & & & &
\end{array}$$

La regla para la construcción de la tabla es la misma que en el caso de la interpolación de Lagrange, excepto que en las dos primeras columnas se repiten datos y en la tercera se toman las derivadas cuando se anula el denominador.

▁ Ejercicio 50 *Dada la siguiente tabla de valores:*

x	$f(x)$	$f'(x)$
1	-1	-3
2	-3	-1
5	3	2

Se pide:

1. *Hallar la tabla de diferencias divididas.*

2. Calcular el polinomio que interpola los datos de la tabla.

Solución: La tabla de diferencias divididas correspondiente a estos datos es la siguiente:

$$\begin{array}{cccccc}
1 & -1 & & & & \\
 & & -3 & & & \\
1 & -1 & & 1 & & \\
 & & -2 & & 0 & \\
2 & -3 & & 1 & & 0 \\
 & & -1 & & 0 & & -\frac{1}{48} \\
2 & -3 & & 1 & & -\frac{1}{12} & \\
 & & 2 & & -\frac{1}{3} & & \\
5 & 3 & & 0 & & & \\
 & & 2 & & & & \\
5 & 3 & & & & &
\end{array}$$

y el polinomio de interpolación pedido es

$$p(x) = -1 - 3(x-1) + (x-1)^2 - \frac{1}{48}(x-1)^2(x-2)^2(x-5). \qquad \diamond$$

▁ Ejercicio 51 *Dada la siguiente tabla de valores:*

x	$f(x)$	$f'(x)$
-3	1	1
1		2
3	1	

determinar si existe un polinomio de grado menor o igual que 3 *que interpole estos datos.*

Solución: Si se impone a la fórmula de interpolación

$$p(x) = l_0(x)f(-3) + l_1(x)f'(-3) + l_2(x)f'(1) + l_3(x)f(3)$$

ser exacta para los polinomios $1, x+3, (x+3)^2$ y $(x+3)^2(x-1)$ se obtiene el sistema lineal

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 6 \\ 0 & 0 & 8 & 36 \\ 0 & 0 & 16 & 72 \end{pmatrix} \begin{pmatrix} l_0(x) \\ l_1(x) \\ l_2(x) \\ l_3(x) \end{pmatrix} = \begin{pmatrix} 1 \\ x+3 \\ (x+3)^2 \\ (x+3)^2(x-1) \end{pmatrix}$$

que es singular ya que el determinante de la matriz de coeficientes es 0. En consecuencia, el problema de interpolación no tiene solución. $\diamond$

5.8 Interpolación por esplines cúbicos

En esta sección se aborda la siguiente cuestión: ¿cómo se interpola a trozos conservando una cierta regularidad?. Si no se dispone de información sobre las derivadas de la función pero se necesita regularidad en la interpolación compuesta, se pueden utilizar las funciones esplines que son polinómicas a trozos y conservan algunos ordenes de regularidad en los nodos de interpolación. La idea básica consiste en la unión de los trozos de los polinomios construidos en sub-intervalos contiguos, de modo que las derivadas laterales coincidan en el nodo común.

Se considera el siguiente conjunto de nodos de interpolación

$$a = x_0 < x_1 < \cdots < x_n = b.$$

Una función s de clase $C^2([a,b])$ es un esplín cúbico en $[a,b]$ si s es un polinomio de grado menor o igual que 3 en cada intervalo $[x_i, x_{i+1}]$. Se dice que es un esplín cúbico de interpolación a los datos $\{(x_i, y_i) : i = 0, 1, \cdots, n\}$ si $s(x_i) = y_i$ para $i = 0, 1, \cdots, n$. En la figura 5.3 se observa como un esplín interpola en 5 nodos los valores de la función $f(x) = \operatorname{sen} x$, manteniendo la regularidad hasta la segunda derivada al pasar por los referidos nodos.

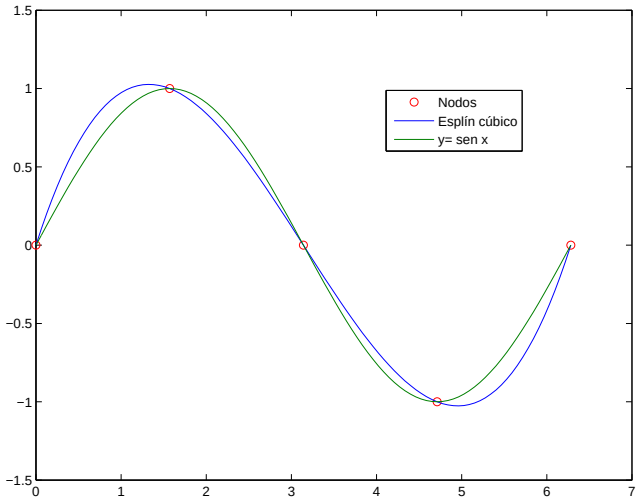


Figura 5.3: Interpolación por esplines

Para poder utilizar una expresión analítica, se consideran las restricciones s_i del esplín s a cada intervalo $[x_i, x_{i+1}]$ para $i = 0, 1, \cdots, n-1$. De este modo se puede representar el esplín mediante los n polinomios de tercer grado

$$s_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i$$

para cada $i = 0, \cdots, n-1$. La dificultad reside en que los coeficientes a_i, b_i, c_i y d_i no pueden ser determinados separadamente para cada $i = 0, \cdots, n-1$, ya que en cada intervalo solamente se dispone de las dos ecuaciones

$$s_i(x_i) = y_i, \;\; s_i(x_{i+1}) = y_{i+1}$$

que proporciona la condición de interpolación. El hecho de que el esplín tenga derivada primera y segunda continua en los nodos de interpolación nos proporciona dos nuevas ecuaciones

$$\begin{aligned} s_i'(x_i) &= s_{i-1}'(x_i), \qquad \text{si } i > 0, \\ s_i''(x_i) &= s_{i-1}''(x_i), \qquad \text{si } i > 0, \\ s_i'(x_{i+1}) &= s_{i+1}'(x_{i+1}), \qquad \text{si } i < n-1, \\ s_i''(x_{i+1}) &= s_{i+1}''(x_{i+1}), \qquad \text{si } i < n-1 \end{aligned}$$

en cada uno de los extremos de cada intervalo. Pero, queda claro que no pueden resolverse de modo independiente en cada intervalo ya que involucran a los polinomios s_{i-1} y s_{i+1}. Esto pone en claro la naturaleza global del esplín.

Para determinar esos coeficientes se introducen las siguientes variables auxiliares

$$z_i := s''(x_i)$$

para $i = 0, \cdots, n$. Aunque el objetivo final es el cálculo de los coeficientes a_i, b_i, c_i y d_i para cada uno de los sub-intervalos, como cálculo intermedio se determinarán los valores de z_i.

De la propia definición de las variables z_i se desprende que si se utiliza la expresión del esplín en cada sub-intervalo para evaluar la derivada segunda en x_i, se obtienen n relaciones entre las variables z_i y los coeficientes de los polinomios. Puesto que

$$\begin{aligned} s_i'(x) &= 3a_i(x - x_i)^2 + 2b_i(x - x_i) + c_i, \\ s_i''(x) &= 6a_i(x - x_i) + 2b_i, \end{aligned}$$

de la definición de las variables z_i, se obtiene que

$$b_i = \frac{z_i}{2} \quad \text{para } i = 0, 1, \ldots, n-1.$$

Por otra parte, si se emplean para evaluar s'' en x_i, las expresiones que le corresponden en los intervalos $[x_{i-1}, x_i]$ y $[x_i, x_{i+1}]$ para $i = 1, \cdots, n-1$, el resultado debe ser el mismo. Esto conduce a las siguientes relaciones

$$6a_{i-1}(x_i - x_{i-1}) + 2b_{i-1} = 2b_i$$

para $i = 1, \cdots, n-1$. De estas relaciones se deduce que

$$a_i = \frac{z_{i+1} - z_i}{6h_i} \tag{5.5}$$

donde $h_i = x_{i+1} - x_i$ para $i = 0, \cdots, n-2$.

Si se imponen las condiciones de interpolación en los extremos del intervalo $[x_i, x_{i+1}]$, se obtienen las ecuaciones siguientes

$$\begin{aligned} d_i &= y_i, \\ a_i h_i^3 + b_i h_i^2 + c_i h_i + d_i &= y_{i+1} \end{aligned}$$

para $i = 0, \cdots, n-1$. Si se usan las expresiones de a_i, b_i y d_i en la segunda de estas relaciones se obtiene que

$$c_i = \frac{y_{i+1} - y_i}{h_i} - \frac{2z_i + z_{i+1}}{6} h_i \tag{5.6}$$

para $i = 0, \cdots, n-2$.

De la continuidad de las derivadas en el punto x_{i+1} se deduce que

$$3a_i h_i^2 + 2b_i h_i + c_i = c_{i+1}.$$

para $i = 0, 1, \cdots, n-2$. Si se insertan en esta relación las expresiones que dan a_i, b_i y c_i en términos de x_i, y_i y h_i se obtiene

$$\frac{h_i}{6} z_i + \frac{h_i + h_{i+1}}{3} z_{i+1} + \frac{h_{i+1}}{6} z_{i+2} = \frac{y_{i+2} - y_{i+1}}{h_{i+1}} - \frac{y_{i+1} - y_i}{h_i}$$

para $i = 0, 1, \cdots, n-2$. Si se define $r_i = \frac{h_i}{h_i + h_{i+1}}$, la anterior relación se convierte en

$$r_i z_i + 2z_{i+1} + (1 - r_i) z_{i+2} = 6 f[x_i, x_{i+1}, x_{i+2}]$$

para $i = 0, 1, \cdots, n-2$. Los segundos miembros de estas relaciones son diferencias divididas que pueden calcularse sin dificultad. De hecho, estas relaciones podrían ser interpretadas como un sistema lineal de $n-1$ ecuaciones de incógnitas z_i para $i = 0, 1, \cdots, n$

$$\begin{pmatrix} r_0 & 2 & 1-r_0 & \cdots & 0 & 0 & 0 \\ 0 & r_1 & 2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & 1-r_{n-3} & 0 \\ 0 & 0 & 0 & \cdots & r_{n-2} & 2 & 1-r_{n-2} \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_n \end{pmatrix} = 6 \begin{pmatrix} f[x_0, x_1, x_2] \\ f[x_1, x_2, x_3] \\ \vdots \\ f[x_{n-2}, x_{n-1}, x_n] \end{pmatrix}.$$

En cualquier caso, se trata de un sistema de ecuaciones indeterminado ya que el número de incógnitas excede en 2 al de ecuaciones. Consecuentemente, el problema, tal y como se ha planteado, admite una infinidad de soluciones.

Existen varias posibilidades para completar este sistema de ecuaciones. Las más habituales son las siguientes:

- Esplín natural
$$s_0''(x_0) = s_{n-1}''(x_n) = 0,$$
- Esplín con pendiente fijada en los extremos (End Slope Spline)
$$s_0'(x_0) = y_0', \quad s_{n-1}'(x_n) = y_n',$$
- Esplín periódico
$$s_0'(x_0) = s_{n-1}'(x_n), \quad s_0''(x_0) = s_{n-1}''(x_n),$$
- Esplín sin nudo (Not-a-Knot Spline)
$$s_0'''(x_1) = s_1'''(x_1), \quad s_{n-2}'''(x_{n-1}) = s_{n-1}'''(x_{n-1}).$$

A modo de ejemplo, se examina a continuación el caso del esplín natural. Se pueden incorporar las condiciones de esplín natural suprimiendo la primera y la última de las columnas de sistema lineal

$$\begin{pmatrix} 2 & 1-r_0 & \cdots & 0 & 0 \\ r_1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 2 & 1-r_{n-3} \\ 0 & 0 & \cdots & r_{n-2} & 2 \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_{n-1} \end{pmatrix} = 6 \begin{pmatrix} f[x_0, x_1, x_2] \\ f[x_1, x_2, x_3] \\ \vdots \\ f[x_{n-2}, x_{n-1}, x_n] \end{pmatrix}.$$

La matriz de coeficientes de este sistema es estrictamente diagonal dominante y consecuentemente es no singular. Es decir, calculadas las diferencias divididas a partir de los datos, la resolución de este sistema nos permite conocer el valor de las variables z_i y a partir de ello, calcular los coeficientes del esplín en cada sub-intervalo a partir de las fórmulas 5.5 y 5.6 para $i = 0, \cdots, n-2$. Imponiendo las condiciones de interpolación para s_{n-1} y s_{n-1}'' en los extremos del intervalo $[x_{n-1}, x_n]$ se prueba la validez de las fórmulas 5.5 y 5.6 para $i = n-1$.

▬ Ejercicio 52 *Dada la siguiente tabla de valores:*

x	0	0.25	0.50	0.75	1
$f(x)$	0	0.7071	1.0000	0.7071	0

calcular los valores que toma la derivada segunda del esplín cúbico natural asociado a estos datos en los puntos $0.25, 0.5$ *y* 0.75. *Hallar la expresión de* $s(x)$ *en el intervalo* $[0.25, 0.50]$.

Solución: En primer lugar, se construye la tabla de diferencias divididas

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$
0	0		
		2.8284	
0.25	0.7071		−3.3136
		1.1716	
0.5	1		−4.6864
		−1.1716	
0.75	0.7071		−3.3136
		−2.8284	
1	0		

En segundo lugar, se resuelve el sistema

$$\begin{pmatrix} 2 & 0.5 & 0 \\ 0.5 & 2 & 0.5 \\ 0 & 0.5 & 2 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} -19.8816 \\ -28.1184 \\ -19.8816 \end{pmatrix}$$

con $z_0 = z_4 = 0$. La solución es $z_1 = -7.344, z_2 = -10.3872, z_3 = -7.344$.

Finalmente, la fórmulas que dan los coeficientes nos proporcionan directamente los coeficientes del polinomio

$$s_1(x) = a_1(x - 0.25)^3 + b_1(x - 0.25)^2 + c_1(x - 0.25) + d_1$$

que resultan

$$a_1 = -2.0288 \quad b_1 = -3.672 \quad c_1 = 2.2164 \quad d_1 = 0.7071$$

Se puede verificar directamente este resultado imponiendo a s_1 que verifique las cuatro condiciones

$$s_1(x_1) = y_1, \;\; s_1(x_2) = y_2, \;\; s_1''(x_1) = z_1, \;\; s_1''(x_2) = z_2. \qquad \diamond$$

5.9 Ejercicios

EJERCICIO 53 *Estimar el error que se comete al calcular $e^{\sqrt{x}}$, interpolando el valor de la función en dos puntos x_0 y x_1 arbitrarios en el intervalo $[1, 2]$.*

Solución: El error de interpolación de la función $f(x) = e^{\sqrt{x}}$ puede ser estimado por la fórmula

$$E(x) = \frac{f''(\xi)}{2}(x - x_0)(x - x_1)$$

siendo ξ un punto desconocido del intervalo $[x_0, x_1]$. Puesto que x_0 y x_1 son arbitrarios en el intervalo $[1, 2]$, se trata de buscar una cota del error que sea independiente de estos dos puntos. Esto se puede obtener usando la siguiente acotación

$$|E(x)| \leq \frac{\max_{t\in[x_0,x_1]} |f''(t)|}{2} \max_{t\in[x_0,x_1]} |(t-x_0)(t-x_1)|.$$

Además, se tiene que para $t \in [x_0, x_1]$

$$|f''(t)| = \left|\frac{1}{4t}\left(1 - \frac{1}{\sqrt{t}}\right)\right| e^{\sqrt{t}}.$$

Puesto que f'' es monótona creciente en $[1, 2]$, su máximo se alcanza en x_1 y puesto que x_1 es arbitrario en $[1, 2]$, se puede usar la siguiente acotación

$$\max_{t\in[x_0,x_1]} |f''(t)| \leq |f''(2)| = \frac{1}{8}\left(1 - \frac{1}{\sqrt{2}}\right) e^{\sqrt{2}} < 0.1507.$$

Por otra parte, se tiene

$$\max_{t\in[x_0,x_1]} |(t-x_0)(t-x_1)| \leq \frac{(x_1-x_0)^2}{4} \leq \frac{1}{4}.$$

En consecuencia, una estimación del error es la siguiente

$$|E(x)| \leq 0.019$$

para todo $x \in [x_0, x_1]$. $\diamond$

– Ejercicio 54 *Se considera un conjunto de puntos de interpolación*

$$a = x_0 < x_1 < \cdots < x_n = b$$

en el intervalo $[a, b]$. *Se pide calcular el polinomio de interpolación de Lagrange en estos nodos de las siguientes funciones*

1. $f(x) = \omega_{n+1}(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$,
2. $f(x) = x^{n+1}$,
3. $f(x) = \omega_n(x)x^3$.

Solución:

1. Puesto $f(x_i) = \omega_{n+1}(x_i) = 0$, el polinomio de Lagrange es

$$p_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i) = 0$$

.

2. De la fórmula del error se deduce que

$$x^{n+1} = p_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x) = p_n(x) + \omega_{n+1}(x)$$

de donde se obtiene que $p_n(x) = x^{n+1} - \omega_{n+1}(x)$.

3. De la fórmula de Lagrange se deduce que

$$p_n(x) = \sum_{i=0}^{n} l_i(x)\omega_n(x_i)x_i^3 = \omega_n(x_n)x_n^3 l_n(x) = x_n^3\omega_n(x). \qquad \diamond$$

Ejercicio 55 *Construir el polinomio de interpolación de la función de Heaviside*

$$f(x) = \begin{cases} 0, & si\ x < 0 \\ 1, & si\ x \geq 0 \end{cases}$$

en los nodos

$$-1 < -\frac{1}{2} < -\frac{1}{3} < 0 < \frac{1}{3} < \frac{1}{2} < 1.$$

Solución:

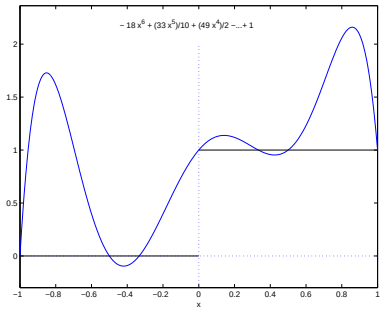


Figura 5.4: Gráfica del polinomio de interpolación: Ejercicio 55

Se construye la tabla de diferencias

$$
\begin{array}{cccccccc}
-1 & 0 & & & & & & \\
 & & 0 & & & & & \\
-\frac{1}{2} & 0 & & 0 & & & & \\
 & & 0 & & 6 & & & \\
-\frac{1}{3} & 0 & & 6 & & -\frac{279}{20} & & \\
 & & 3 & & -\frac{63}{5} & & \frac{213}{10} & \\
0 & 1 & & -\frac{9}{2} & & 18 & & -18 \\
 & & 0 & & \frac{27}{5} & & -\frac{147}{10} & \\
\frac{1}{3} & 1 & & 0 & & -\frac{81}{20} & & \\
 & & 0 & & 0 & & & \\
\frac{1}{2} & 1 & & 0 & & & & \\
 & & 0 & & & & & \\
1 & 1 & & & & & &
\end{array}
$$

En consecuencia, el polinomio de interpolación es

$$p(x) = -18\,x^6 + \frac{33\,x^5}{10} + \frac{49\,x^4}{2} - \frac{115\,x^3}{24} - 7\,x^2 + \frac{239\,x}{120} + 1. \qquad \diamond$$

▬ Ejercicio 56 *La función $y = f(x)$ es conocida solamente en los siguientes valores:*

$$f(0) = 1, \quad f'(0.1) = -\frac{10}{3}, \quad f'(0.2) = -\frac{10}{3}, \quad f(0.3) = 0.001\ .$$

Esta función tiene un punto de inflexión en el intervalo $0 < x < 0.3$. Calcular una aproximación de la abscisa de dicho punto usando interpolación mediante polinomios.

Solución: Se considera la base de los polinomios de grado menor o igual 3 formada por $\{1, x, x^2, x^3\}$. Si se imponen a un polinomio de tercer grado $p_3(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ las cuatro condiciones de interpolación, el sistema lineal que resulta es

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.2 & 0.03 \\ 0 & 1 & 0.4 & 0.12 \\ 1 & 0.3 & 0.09 & 0.027 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1. \\ -\frac{10}{3} \\ -\frac{10}{3} \\ 0.001 \end{pmatrix}.$$

Puesto que el determinante de D es distinto de 0, el problema de interpolación tiene solución única, por lo que se puede calcular los coeficientes del polinomio de interpolación mediante este sistema lineal.

Si se resuelve el sistema de modo exacto, se obtiene la solución

$$p_3(x) = 1 - \frac{83}{25}x - \frac{1}{10}x^2 + \frac{2}{9}x^3.$$

Puesto que el punto de inflexión del polinomio de interpolación está dado por $x = -\frac{a_2}{3a_3}$, un valor aproximado del punto de inflexión de la función f es $\frac{3}{20}$. Los cálculos han sido realizados de modo exacto pero es interesante verificar los resultados cuando se utiliza una precisión limitada. Para valores de x pequeños la parte más relevante del polinomio es $1 - \frac{83}{25}x$ cuya gráfica es una recta. El ejemplo podría servir para ilustrar la inestabilidad algunos problemas de interpolación. ◇

▬ Ejercicio 57 *Para interpolar por el método de Hermite una función f se observan los valores de la función y su primera derivada en los puntos $\{0, 1\}$. Una vez calculadas las diferencias divididas se tiene que*

$$f[0] = f[0, 0] = f[0, 0, 1] = f[0, 0, 1, 1] = 1$$

Se pide:

1. *Determinar los valores de $f(0), f(1), f'(0), f'(1)$.*
2. *Calcular el polinomio de interpolación de Hermite de f en los puntos indicados.*

Solución: La tabla de diferencias divididas contiene ya la siguiente información

$$\begin{array}{llllll}
0 & 1 & & & & \\
 & & 1 & & & \\
0 & f(0) & & 1 & & \\
 & & f[0,1] & & 1 & \\
1 & f(1) & & f[0,1,1] & & \\
 & & f[1,1] & & & \\
1 & f(1) & & & &
\end{array}$$

De la última columna se deduce que

$$\frac{f[0, 1, 1] - 1}{1} = 1$$

y consecuentemente $f[0, 1, 1] = 2$. De modo similar, de

$$\frac{f[0, 1] - 1}{1} = 1$$

se deduce $f[0,1] = 2$. También se tiene que

$$\frac{f[1,1] - 2}{1} = 2$$

y consecuentemente $f[1,1] = 4$. De la tercera columna se deduce que

$$f'(0) = 1, \quad \frac{f(1) - f(0)}{1} = 2, \quad f'(1) = 4,$$

lo que permite completar la tabla

$$\begin{array}{cccccc}
0 & 1 & & & & \\
 & & 1 & & & \\
0 & 1 & & 1 & & \\
 & & 2 & & 1 & \\
1 & 3 & & 2 & & \\
 & & 4 & & & \\
1 & 3 & & & &
\end{array}$$

De acuerdo con la fórmula de Newton, se deduce que el polinomio de interpolación de Hermite es

$$p(x) = 1 + x + x^2 + x^2(x-1) = 1 + x + x^3. \qquad \diamond$$

– Ejercicio 58 *Dada la siguiente tabla de valores:*

x	0	1	2	3	4
$f(x)$	0	0	2	4	6

calcular el esplín cúbico natural de interpolación a estos datos en el intervalo $[0,1]$.

Solución: En primer lugar se construye la tabla de diferencias divididas

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$
0	0		
		0	
1	0		1
		2	
2	2		0
		2	
3	4		0
		2	
4	6		

Puesto que los valores de las abscisas están igualmente espaciadas se tiene que $r_i = \frac{1}{2}$. En segundo lugar se resuelve el sistema

$$\begin{pmatrix} 2 & 0.5 & 0 \\ 0.5 & 2 & 0.5 \\ 0 & 0.5 & 2 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix}$$

con $z_0 = z_4 = 0$. La solución es $z_1 = \frac{45}{14}, z_2 = -\frac{6}{7}, z_3 = \frac{3}{14}$.

Finalmente, la fórmulas que dan los coeficientes nos proporcionan directamente los coeficientes del polinomio

$$s_0(x) = a_0x^3 + b_0x^2 + c_0x + d_0$$

No obstante se pueden calcular directamente imponiendo a s_1 que verifique las cuatro condiciones

$$s_0(0) = 0, \;\; s_0(1) = 0, \;\; s_0''(0) = 0, \;\; s_0''(1) = \frac{45}{14}.$$

de donde se deduce que $s_0(x) = \frac{15}{28}(x^3 - x)$. $\diamond$

CAPÍTULO 6

Derivación e integración numérica

6.1 Introducción

El cálculo de derivadas o integrales no es en general simple salvo que la función a derivar o integrar, sea elemental. En lo que se refiere a la integración, incluso en casos tan sencillos como el siguiente

$$\int_0^1 e^{-x^2} dx,$$

en el que no es posible realizar la evaluación con técnicas analíticas, será necesario recurrir a aproximaciones numéricas. Aunque el cálculo automático permite potenciar estos cálculos simbólicos al máximo, lo cierto es que sus posibilidades son limitadas.

Las ideas que soportan las técnicas de derivación o cuadratura numérica que se emplearán en este capítulo no serán nuevas. Simplemente, los procedimientos de aproximación de funciones introducidos en el capítulo anterior serán aplicados directamente al derivando o al integrando que será sustituido por una función elemental, posiblemente polinomial. Una vez aproximada la función, se integra o deriva directamente el polinomio que la aproxima.

En este capítulo, después de una breve introducción a la derivación numérica, el énfasis se hará en el desarrollo de métodos de cuadratura. Además, aunque cualquiera de los procedimientos de aproximación descritos en los capítulos anteriores, podría ser empleado, en este capítulo se considerarán únicamente los métodos basados en la interpolación.

6.2 Fórmulas de derivación numérica

Sea f una función definida en un intervalo $[a, b]$ y derivable en un punto $c \in (a, b)$. Si la evaluación de la derivada de f en c es complicada parece razonable construir un polinomio de interpolación en este intervalo y después derivarlo en el punto $x = c$. Si solamente se utilizan los extremos del intervalo $x_0 = a < x_1 = b$ el polinomio de interpolación construido con las funciones básicas de Lagrange es

$$p(x) = \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1)$$

y su derivada en el punto $x = c$ está dada por

$$p'(x) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Esta expresión no es otra que la que corresponde a un cociente incremental cuyo límite cuando el tamaño del intervalo tiende a 0 es el valor exacto de la derivada de f en x_0. Sin embargo la puesta en práctica de una idea tan simple, tiene algunas dificultades cuando los cálculos de los cocientes incrementales se realizan con una aritmética finita. En ejemplo 1 del capítulo 1 se ponía de manifiesto que cuando el incremento de la variable independiente es muy pequeño, el numerador podría anularse y por lo tanto también el cociente incremental aun cuando el valor teórico del límite fuese distinto de 0. Cuando se pretende calcular una derivada de alto orden de este modo, las dificultades se incrementan considerablemente.

▪ Ejemplo **32** Se desea aproximar el valor de la derivada tercera en un punto x_0 de una función f que no se conoce pero cuya evaluación en un punto arbitrario no es costosa. Lo más natural parece escoger un conjunto de puntos igualmente espaciados $x_0, x_0 + h, x_0 + 2h, x_0 + 3h$ y derivar tres veces consecutivas el polinomio de interpolación construido sobre esos nodos.

El polinomio de interpolación de Lagrange construido sobre los nodos equidistantes $x_0, x_0 + h, x_0 + 2h, x_0 + 3h$ por la fórmula progresiva de Newton es

$$\begin{aligned} p_3(x) &= f(x_0) + \frac{\Delta f(x_0)}{h}(x - x_0) + \frac{\Delta^2 f(x_0)}{2h^2}(x - x_0)(x - x_1) \\ &+ \frac{\Delta^3 f(x_0)}{6h^3}(x - x_0)(x - x_1)(x - x_2). \end{aligned}$$

Si se deriva 3 veces consecutivamente se obtiene

$$p_3'''(x_0) = \frac{\Delta^3 f(x_0)}{h^3}.$$

Para ilustrar el cálculo se consideran como datos particulares $f(x) = \sqrt{x}$, el punto $x_0 = 1$ y el tamaño de paso entre nodos $h = 0.05$. Se construye la tabla de diferencias finitas

$$\begin{array}{cccccc}
1 & & 1 & & & \\
 & & & 0.493901754 & & \\
1.05 & 1.02469508 & & & -0.11626508 & \\
 & & 0.482275246 & & & 0.052324133 \\
1.1 & 1.04880885 & & -0.10841646 & & \\
 & & 0.4714336 & & & \\
1.15 & 1.07238053 & & & &
\end{array} .$$

Consecuentemente, el polinomio de interpolación es

$$\begin{aligned} p(x) &= 1 + 0.4939016(x-1) - 0.116262(x-1)(x-1.05) \\ &+ 0.052293(x-1)(x-1.05)(x-1.1) \end{aligned}$$

y el resultado buscado

$$p_3'''(x_0) = 0.31371803.$$

Por otra parte, ya que

$$f'''(x_0) = \frac{3}{8} x_0^{-\frac{5}{2}},$$

el valor exacto es $f'''(1) = 0.375$ $\diamond$

El cálculo efectivo de los coeficientes de una fórmula de derivación numérica pueden llevarse a cabo del modo siguiente. Se construye el polinomio de interpolación de la función a derivar f

$$p_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i)$$

donde l_i representan las funciones básicas de Lagrange. La fórmula de derivación se construye como

$$D(f)(\bar{x}_0) = \sum_{i=0}^{n} l_i'(\bar{x}_0) f(x_i).$$

Es importante observar que los coeficientes $a_i = l_i'(\bar{x}_0)$ de la fórmula de derivación puede ser calculados por el sistema lineal

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ n\bar{x}_0^{n-1} \end{pmatrix} \tag{6.1}$$

obtenido derivando ambos miembros del sistema de ecuaciones 5.2 de la página 128. En principio, el punto de derivación $\bar{x}_0$ no tiene porque coincidir con ninguno de los nodos de interpolación.

Si lo que que se pretende es construir una fórmula de derivación de mayor orden, el procedimiento es el mismo. Así para la derivada r-ésima, la formula tendrá los siguientes coeficientes

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ n(n-1)\cdots(n-r)\bar{x}_0^{n-r} \end{pmatrix}$$

EJERCICIO 59 *Hallar a_1, a_2 y a_3 para que la fórmula*

$$f''(2) \approx a_1 f(1) + a_2 f(2) + a_3 f(3)$$

esté basada en la interpolación (de Lagrange) y dar una expresión del error para el caso en el que $f \in C^5([1,3])$.

Solución: Si la fórmula está basada en la interpolación, proviene de sustituir la función a derivar por un polinomio de interpolación. Consecuentemente, si la interpolación de Lagrange es exacta para los polinomios de grado menor o igual que 2 también lo será la fórmula de derivación numérica y por ello se cumplirá que

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 4 & 9 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}.$$

Consecuentemente la fórmula de derivación numérica buscada es

$$f''(2) \approx D(f) = f(1) - 2f(2) + f(3).$$

Si se toma $f(x) = x^3$ se obtiene que $f''(2) = 12$ y este número coincide con el valor aproximado por la fórmula. Se puede verificar que esto no ocurre para el polinomio x^4. Así pues, la fórmula es exacta hasta polinomios de grado menor o igual que 3.

Para estimar el error se consideran los desarrollos de Taylor alrededor de $x = 2$

$$f(1) = f(2) - f'(2) + \frac{1}{2}f''(2) - \frac{1}{6}f'''(2) + \frac{1}{4!}f^{(4)}(2) - \frac{1}{5!}f^{(5)}(\xi),$$

$$f(3) = f(2) + f'(2) + \frac{1}{2}f''(2) + \frac{1}{6}f'''(2) + \frac{1}{4!}f^{(4)}(2) + \frac{1}{5!}f^{(5)}(\bar{\xi}),$$

donde $1 \leq \xi \leq 2$ y $2 \leq \bar{\xi} \leq 3$ son desconocidos. De ello se desprende que

$$f''(2) - D(f) = -\frac{1}{12} f^{(4)}(2) + \frac{1}{5!} f^{(5)}(\xi) - \frac{1}{5!} f^{(5)}(\bar{\xi}). \qquad \diamond$$

6.3 Método de Extrapolación de Richardson

En la sección previa se ha discutido acerca de la inestabilidad que encierra el cálculo aproximado de una derivada cuando los puntos usados en el procedimiento de interpolación, se aproximan al punto donde se deriva. Si este conjunto de puntos depende de un parámetro h destinado a tender a 0, es posible que por debajo de un determinado umbral para h, la precisión de la aproximación buscada, se deteriore drásticamente. Un modo de mejorar la precisión, que limita este umbral del parámetro h, es utilizar los métodos de extrapolación de Richardson.

La idea básica de estos métodos es la siguiente: Si se representa por $N(h)$ la aproximación correspondiente al parámetro h, y se efectúan varias aproximaciones con distintos valores de h en el rango estable, con estos valores se puede reconstruir aproximadamente la función $N(h)$ y a continuación extrapolar el valor $N(0)$. Por ejemplo, cuando se está utilizando la siguiente fórmula de derivación numérica

$$D_h f(x_0) = \frac{f(x_0 + h) - f(x_0 - h)}{2h}$$

para aproximar el valor de la derivada de f en el punto x_0, se está considerando una familia de puntos $\{x_0 - h, x_0 + h\}$ que depende de un parámetro h destinado a tender a 0. La función $N(h)$ es en este ejemplo, la que asigna a cada valor de h el número $D_h f(x_0)$. Se utiliza el término extrapolación para indicar el punto donde se interpola, $h = 0$ está fuera del mínimo intervalo que contiene a los nodos usados.

La cuestión es ahora cómo distribuir los valores de h, que se usan como nodos de la interpolación, para que la precisión conseguida con la extrapolación mejore la precisión obtenida con cada uno de ellos. Se puede encontrar una respuesta a esta cuestión cuando teóricamente se dispone de una fórmula que establezca la precisión de la aproximación, del siguiente tipo

$$N(h) = N(0) + \alpha h^m + O(h^{m+1}).$$

En este fórmula se utiliza una O grande de Landau para representar una función para la que existe un h_0 tal que la función $\frac{O(h^{m+1})}{h^{m+1}}$ está acotada en el

intervalo $(0, h_0)$. En esta terminología, se dice que $N(h)$ es una aproximación a $N(0)$ de orden m.

Si en la fórmula anterior se sustituye h por rh para un número $0 < r < 1$ arbitrario, se obtiene que

$$N(rh) = N(0) + \alpha h^m r^m + O(h^{m+1}).$$

Si se combinan estas relaciones se obtiene

$$\frac{r^m N(h) - N(rh)}{r^m - 1} = N(0) + O(h^{m+1}).$$

Es decir, el cociente que está a la izquierda de la anterior igualdad

$$N_1(h) = \frac{r^m N(h) - N(rh)}{r^m - 1}$$

es una aproximación a $N(0)$ de orden $m+1$ mientras que $N(h)$ solamente es una aproximación de orden m. De este modo, combinando el resultado de la fórmula para h con el de rh, se puede construir una aproximación mejor al valor buscado.

Esta idea se puede utilizar de modo recurrente para aumentar el orden de aproximación en una unidad en cada etapa. Se puede organizar el cálculo del modo siguiente: Se calculan $N(h), N(rh), N(r^2h), \cdots$ de modo que $r^i h$ se mantenga en el rango estable. Con cada pareja $N(r^{i-1}h), N(r^i h)$ se construye la aproximación de orden $m+1$

$$N_1(r^{i-1}h) = \frac{r^m N(r^{i-1}h) - N(r^i h)}{r^m - 1}$$

para $i = 1, 2, \ldots$. Si se usa el mismo procedimiento para el conjunto $N_1(h), N_1(rh), N_1(r^2h), \cdots$ se obtiene un conjunto de aproximaciones de orden $m+2$

$$N_2(r^{i-1}h) = \frac{r^{m+1} N_1(r^{i-1}h) - N_1(r^i h)}{r^{m+1} - 1}$$

para $i = 1, 2, \ldots$ y así sucesivamente.

Se pueden organizar los cálculos en la siguiente tabla

$$\begin{array}{cccc} h & N(h) & N_1(h) & N_2(h) \\ rh & N(rh) & N_1(rh) & N_2(rh) \\ r^2h & N(r^2h) & N_1(r^2h) & \cdots \\ r^3h & N(r^3h) & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

– EJERCICIO 60 *Aproximar el valor de la derivada de la función* $f(x) = \operatorname{sen} 2\pi x$ *en el punto* $x_0 = 0$*, usando un método de extrapolación de Richardson sobre la fórmula de derivación numérica*

$$f'(x_0) \approx D_h f(x_0) = \frac{f(x_0+h) - f(x_0-h)}{2h}$$

hasta un error de orden 4*, partiendo del valor* $h = 0.5$ *y con un factor de decrecimiento* $r = \frac{1}{2}$*.*

Solución: Si se usan los desarrollos de Taylor de la función f alrededor del punto x_0

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{1}{2}f''(x_0)h^2 + \frac{1}{6}f'''(x_0)h^3 + O(h^4),$$
$$f(x_0-h) = f(x_0) - f'(x_0)h + \frac{1}{2}f''(x_0)h^2 - \frac{1}{6}f'''(x_0)h^3 + O(h^4),$$

se obtiene la fórmula del error

$$D_h f(x_0) = f'(x_0) + \frac{h^2}{6}f'''(x_0) + O(h^3)$$

Si se aplica el método de Richardson a la fórmula

$$N(h) = D_h f(0) = \frac{\operatorname{sen} 2\pi h - \operatorname{sen}(-2\pi h)}{2h} = \frac{\operatorname{sen} 2\pi h}{h},$$

que tiene orden de aproximación $m = 2$, y se toma $r = \frac{1}{2}$, se obtiene la siguiente tabla

i	r^i	$r^i h$	$N(r^i h)$	$N_1(r^i h)$	$N_2(r^i h)$
0	1	$\frac{1}{2}$	0	$\frac{16}{3}$	$\frac{16(8\sqrt{2}-3)}{21}$
1	$\frac{1}{2}$	$\frac{1}{4}$	4	$\frac{4(4\sqrt{2}-1)}{3}$	
2	$\frac{1}{4}$	$\frac{1}{8}$	$4\sqrt{2}$		
3	$\frac{1}{8}$				

ya que

$$\begin{aligned} N_1(h) &= \frac{r^2 N(h) - N(rh)}{r^2 - 1}, \\ N_1(rh) &= \frac{r^2 N(rh) - N(r^2 h)}{r^2 - 1} \\ N_2(h) &= \frac{r^3 N_1(h) - N_1(rh)}{r^3 - 1}. \end{aligned}$$

Se ha escogido arbitrariamente el valor $r = \frac{1}{2}$ en el intervalo $(0, 1)$. Para otra elección la aproximación podría variar ligeramente. Desafortunadamente, en general, no se puede anticipar para que valor de r el resultado será el mejor. ◇

6.4 Cuadratura basada en la interpolación

Sea f una función definida en un intervalo $[a, b]$ y

$$a \leq x_0 < x_1 < \cdots < x_n \leq b$$

un conjunto de nodos que serán usados para construir un polinomio de interpolación de Lagrange p_n de la función f. Es razonable pensar en aproximar la integral de la función f en $[a, b]$ por la integral de p_n en el mismo intervalo. De este modo, si se utiliza la expresión de Lagrange para el polinomio de interpolación, se obtiene

$$Q(f) = \int_a^b p_n(x)\ dx = \sum_{i=0}^{n} \left(\int_a^b l_i(x)\ dx \right) f(x_i) = \sum_{i=0}^{n} \alpha_i f(x_i).$$

Los números $\alpha_i = \int_a^b l_i(x)\ dx$ son conocidos como los pesos de la fórmula de cuadratura. En principio, no hay ninguna restricción que obligue a que los extremos x_0 y x_n, coincidan con los extremos del intervalo $[a, b]$. Es frecuente llamar fórmulas de cuadratura cerradas a aquellos en que esto ocurre y abiertas en otro caso.

La primera consideración que es preciso hacer sobre una fórmula construida así, es que es exacta al menos en la misma clase de polinomios en que lo era la fórmula de interpolación. En este sentido, puesto que la interpolación de Lagrange era exacta para polinomios de grado menor o igual que n, la fórmula de cuadratura también lo será.

▪ Ejemplo **33** Se desea construir una fórmula de cuadratura del tipo

$$\int_0^1 f(x)\ dx \approx Q(f) = \alpha_0 f(0) + \alpha_1 f(1)$$

para integrar funciones f en el intervalo $[0, 1]$ con el máximo grado de precisión.

Si se impone la condición de que la fórmula de cuadratura sea exacta para los polinomios $\{1, x\}$, se obtienen las ecuaciones

$$\alpha_0 + \alpha_1 = \int_0^1 dx = 1,$$

$$\alpha_1 = \int_0^1 x \; dx = \frac{1}{2},$$

de donde se deduce que la fórmula buscada es

$$\int_0^1 f(x) \; dx \approx Q(f) = \frac{1}{2} f(0) + \frac{1}{2} f(1).$$

Esta fórmula no es exacta para polinomios de grado 2 ya que

$$\int_0^1 x^2 \; dx = \frac{1}{3} \neq Q(f) = \frac{1}{2} 0^2 + \frac{1}{2} 1^2 = \frac{1}{2}. \qquad \diamond$$

En general, el cálculo de los pesos α_i puede realizarse mediante el sistema lineal

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \int_a^b 1 \; dx \\ \int_a^b x \; dx \\ \vdots \\ \int_a^b x^n \; dx \end{pmatrix} = \begin{pmatrix} b-a \\ \frac{b^2-a^2}{2} \\ \vdots \\ \frac{b^{n+1}-a^{n+1}}{n+1} \end{pmatrix}. \quad (6.2)$$

Este sistema se obtiene al imponer la condición de exactitud a los polinomios $\{1, x, x^2, \cdots, x^n\}$ o lo que es lo mismo, al integrar ambos miembros en la ecuación 5.2 de la página 128. En definitiva, una fórmula de cuadratura

$$Q(f) = \sum_{i=0}^{n} \alpha_i f(x_i)$$

que utiliza evaluaciones del integrando en $n+1$ puntos está basada en la interpolación de Lagrange en esos puntos si y sólo si es exacta para todo polinomio de grado menor o igual que n.

Alternativamente, se puede utilizar el sistema triangular superior de Newton

$$\mathrm{D} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \int_a^b \; dx \\ \int_a^b (x - x_0) \; dx \\ \vdots \\ \int_a^b \omega_n(x) \; dx \end{pmatrix}$$

que se obtiene integrando en la ecuación 5.3 de la página 130.

– EJERCICIO 61 *Construir la fórmula de cuadratura en* $[0, 4]$ *basada en la interpolación de Lagrange en los puntos* $1, 2, 3$.

Solución: El sistema lineal de Newton asociado al problema es

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \int_0^4 1 \, dx \\ \int_0^4 (x-1) \, dx \\ \int_0^4 (x-1)(x-2) \, dx \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \\ \frac{16}{3} \end{pmatrix},$$

de donde se deduce que la fórmula de cuadratura buscada es

$$Q(f) = \frac{8}{3} f(1) - \frac{4}{3} f(2) + \frac{8}{3} f(3). \qquad \diamond$$

– TEOREMA 23 *El error de cuadratura*

$$e = \int_a^b f(x) \, dx - Q(f)$$

de una fórmula basada en la interpolación, puede estimarse del modo siguiente

$$|e| \leq \frac{\max_{\xi \in [a,b]} f^{(n+1)}(\xi)}{(n+1)!} \int_a^b |\omega_{n+1}(x)| \, dx.$$

Demostración: Si se expresa el error de interpolación mediante una diferencia dividida

$$f(x) = \sum_{i=0}^{n} l_i(x) f(x_i) + f[x_0, x_1, \cdots, x_n, x] \omega_{n+1}(x).$$

y se integran ambos miembros de esta igualdad se obtiene

$$\begin{aligned} \int_a^b f(x) \, dx &= \sum_{i=0}^{n} \alpha_i f(x_i) + \int_a^b f[x_0, x_1, \cdots, x_n, x] \omega_{n+1}(x) \, dx \\ &= \sum_{i=0}^{n} \alpha_i f(x_i) + \int_a^b \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \omega_{n+1}(x) \, dx \end{aligned} \tag{6.3}$$

siendo $\xi : [a,b] \to [a,b]$ una función desconocida. Del hecho de que $f^{(n+1)} \circ \xi$ sea continua y por lo tanto acotada en $[a,b]$, se obtiene el resultado buscado. $\diamond$

▬ EJERCICIO 62 *Construir una fórmula de cuadratura numérica basada en la interpolación de Lagrange usando como nodos las raíces del polinomio de Chebyshev de grado 2 en el intervalo* $[-1,1]$. *Estimar el error cometido con esa fórmula respecto a la integral exacta de la siguente función*

$$f(x) = \operatorname{sen} \pi x$$

.

Solución: Las raíces del polinomio de Chebyshev de grado 2 son $x = \pm\frac{\sqrt{2}}{2}$. De acuerdo con la fórmula 6.2 de la página 165, los coeficientes de la fórmula de cuadratura son las soluciones del siguiente sistema lineal

$$\begin{pmatrix} 1 & 1 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

Así pues, la fórmula buscada es

$$Q(f) = f\left(-\frac{\sqrt{2}}{2}\right) + f\left(\frac{\sqrt{2}}{2}\right).$$

La estimación del error que establece el teorema 23 es la siguiente

$$\begin{aligned} |e| &\leq \frac{\pi^2}{2}\int_{-1}^{1}\left|x^2-\frac{1}{2}\right|dx \\ &= \frac{\pi^2}{2}\left(\int_{-1}^{-\frac{\sqrt{2}}{2}}\left(x^2-\frac{1}{2}\right)dx - \int_{-\frac{\sqrt{2}}{2}}^{\frac{\sqrt{2}}{2}}\left(x^2-\frac{1}{2}\right)dx + \int_{\frac{\sqrt{2}}{2}}^{1}\left(x^2-\frac{1}{2}\right)\right)dx \\ &= \frac{2\sqrt{2}-1}{6}\pi^2. \qquad \diamond \end{aligned}$$

6.5 Fórmulas cerradas de Newton-Cotes

Se considera ahora el caso en el que los nodos de interpolación están igualmente espaciados. Sea $h = x_{i+1} - x_i$ para $i = 0, 1, \cdots, n-1$. Inicialmente, se considera el intervalo de integración $[0,1]$. El sistema lineal que determina los pesos de la fórmula de cuadratura verifican

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & \cdots & n \\ 0 & 1 & 2^2 & \cdots & n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 2^n & \cdots & n^n \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{n}{2} \\ \frac{n^2}{3} \\ \vdots \\ \frac{n^n}{n+1} \end{pmatrix}.$$

La resolución de los sistemas de Vandermonde correspondientes a $n = 0, 1, 2, 3$ conduce a la siguiente tabla de pesos

n	nombre	pesos
0	Extremo izquierdo	$\alpha_0 = 1$
1	Trapecios	$\alpha_0 = \alpha_1 = \frac{1}{2}$
2	Simpson	$\alpha_0 = \alpha_2 = \frac{1}{6},\ \alpha_1 = \frac{2}{3}$
3		$\alpha_0 = \alpha_3 = \frac{1}{8},\ \alpha_1 = \alpha_2 = \frac{3}{8}$

En el caso general de un intervalo $[a, b]$, la fórmula del cambio de variable permite probar directamente que los pesos de una fórmula de cuadratura son el resultado de multiplicar los pesos de la fórmula normalizada al intervalo $[0, 1]$ por la longitud del intervalo. De este modo, las fórmulas de cuadratura del trapecio y de Simpson resultan ser las siguientes:

$$\int_a^b f(x)\ dx \approx Q_T(f) = \frac{b-a}{2}(f(a) + f(b)) \text{ Trapecios,}$$

$$\int_a^b f(x)\ dx \approx Q_S(f) = \frac{b-a}{6}(f(a) + 4f((a+b)/2) + f(b)) \text{ Simpson.}$$

La fórmula del error de cuadratura establecida en la sección precedente puede ser refinada en el caso de las fórmulas de Newton-Cotes. Con este fin se consideran los siguientes resultados del Cálculo Diferencial

– TEOREMA 24 *1. (Valor medio integral) Si f es una función continua en $[a, b]$ y w es una función integrable en $[a, b]$ tal que su signo no cambia en este intervalo entonces*

$$\int_a^b f(x)\omega(x)\ dx = f(\xi)\int_a^b \omega(x)\ dx$$

para algún $\xi \in (a, b)$

2. (Valor medio discreto) Sean

- *f una función continua en $[a, b]$,*
- *$\{x_i : i = 0, \cdots, n\}$, $n + 1$ puntos distintos en $[a, b]$*
- *$\{\rho_i : i = 0, \cdots, n\}$ $n + 1$ valores teniendo el mismo signo,*

entonces, existe $\eta \in [a,b]$ *tal que*

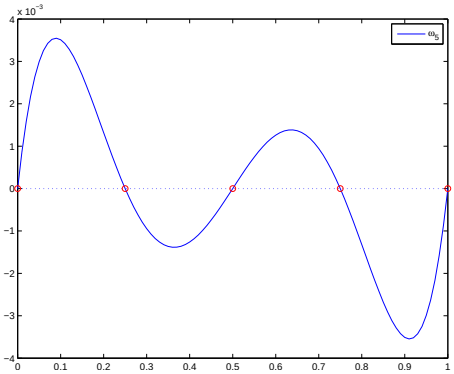
$$\sum_{i=0}^{n} \rho_i f(x_i) = \left(\sum_{i=0}^{n} \rho_i \right) f(\eta).$$

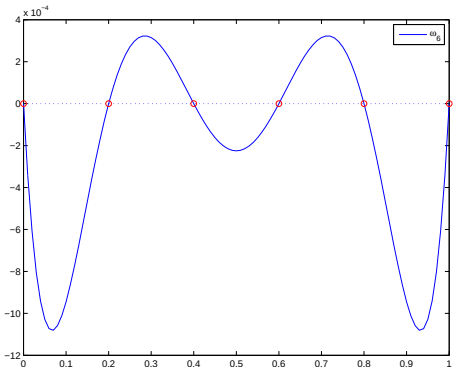
También es de interés el siguiente

Lema 3 *Para cualquier conjunto de puntos igualmente espaciados*

$$\{x_i : i = 0, 1, \cdots, n,\ x_0 = a, x_n = b\}$$

se cumple que la función $\bar{\omega}_{n+1}(x) = \omega_{n+1}\left(x + \frac{a+b}{2}\right)$ *es una función par si* $n+1$ *es par o impar si* $n+1$ *es impar.*





Demostración: Puesto que

$$\frac{a+b}{2} = x_0 + \frac{nh}{2}$$

se tiene que

$$\begin{aligned} \bar{\omega}(x) &= \omega\left(x + \frac{a+b}{2}\right) = \omega\left(x + x_0 + \frac{nh}{2}\right) \\ &= \prod_{i=0}^{n}\left(x + x_0 + \frac{nh}{2} - x_0 - ih)\right) = \prod_{i=0}^{n}\left(x + \frac{n-2i}{2}h\right). \end{aligned}$$

En este producto, puede cambiarse el índice i por $n-i$ ya que el recorrido de ambos es el mismo. De este modo se obtiene

$$\bar{\omega}(x) = \prod_{i=0}^{n}\left(x - \frac{n-2i}{2}h\right) = \prod_{i=0}^{n}\left(x - (x_0 + \frac{hk}{2}) + x_i\right) = (-1)^{n+1}\bar{\omega}(-x)$$

como se quería demostrar. ◇

En el caso de la fórmula de cuadratura de los trapecios, de la primera igualdad en la expresión 6.3 se deduce

$$e_1(f) = \int_a^b f[a,b,x]\omega_2(x)\ dx = \int_a^b \frac{f''(\xi(x))}{2}\omega_2(x)\ dx$$

donde $\xi = \xi(x)$ es una función desconocida que toma valores en $[a,b]$. Puesto que $\omega_2(x) = (x-a)(x-b)$ conserva el signo en $[a,b]$ del teorema del valor medio se deduce que

$$e_1(f) = \frac{f''(\eta)}{2}\int_a^b \omega_2(x)\ dx = -\frac{f''(\eta)}{12}(b-a)^3.$$

para algún $\eta \in (a,b)$.

En el caso de la fórmula de cuadratura de Simpson, de la primera igualdad en la expresión 6.3 se deduce

$$\begin{aligned} e_2(f) &= \int_a^b f[a,\frac{a+b}{2},b,x]\omega_3(x)\ dx \\ &= \int_a^{\frac{a+b}{2}} f[a,\frac{a+b}{2},b,x]\omega_3(x)\ dx + \int_{\frac{a+b}{2}}^b f[a,\frac{a+b}{2},b,x]\omega_3(x)\ dx \\ &= \int_a^{\frac{a+b}{2}} f[a,\frac{a+b}{2},b,x]\omega_3(x)\ dx - \int_a^{\frac{a+b}{2}} f[a,\frac{a+b}{2},b,a+b-x]\omega_3(x)\ dx \\ &= 2\int_a^{\frac{a+b}{2}} f\left[a,\frac{a+b}{2},b,x,a+b-x\right](x-\frac{a+b}{2})\omega_3(x)\ dx \end{aligned}$$

ya que $\omega_3(a+b-x) = -\omega_3(x)$ para $x \in [\frac{a+b}{2},b]$.

Por otra parte, $\omega_3(x) = (x-a)\left(x-\frac{a+b}{2}\right)(x-b)$ es positiva en $[a,\frac{a+b}{2}]$, del teorema del valor medio se deduce

$$e_2(f) = -\frac{2f^{(4)}(\eta)(b-a)}{4!}\int_a^{\frac{a+b}{2}}(x-\frac{a+b}{2})\omega_3(x)\ dx = -\frac{f^{(4)}(\eta)}{90}\left(\frac{b-a}{2}\right)^5.$$

para algún $\eta \in (a,b)$.

– Ejercicio 63 *Aplicar las fórmulas de cuadratura de Newton-Cotes de órdenes $n = 1,2,3,4$ para aproximar*

$$\int_{-5}^{5}\frac{dx}{1+x^2}$$

Solución: Las ecuaciones

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -5 & -\frac{5}{2} & 0 & \frac{5}{2} & 5 \\ 25 & \frac{25}{4} & 0 & \frac{25}{4} & 25 \\ -125 & -\frac{125}{8} & 0 & \frac{125}{8} & 125 \\ 625 & \frac{625}{16} & 0 & \frac{625}{16} & 625 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 0 \\ \frac{250}{3} \\ 0 \\ 1250 \end{pmatrix}$$

determinan la fórmula de cuadratura

$$Q_4(f) = \frac{1}{9}\left(7(f(-5) + f(5)) + 32(f(-2.5) + f(2.5)) + 12f(0)\right)$$

en el intervalo $[-5, 5]$. De modo similar se calculan las fórmulas para $n = 1, 2, 3$.

$$\begin{aligned} Q_1(f) &= 5(f(-5) + f(5)) \\ Q_2(f) &= 10(\frac{1}{6}f(-5) + \frac{2}{3}f(0) + \frac{1}{6}f(5)) \\ Q_3(f) &= 10(\frac{1}{8}(f(-5) + f(5)) + \frac{3}{8}(f(-5/3) + f(5/3))) \end{aligned}$$

Estas fórmulas y las correspondientes a grados superiores, aplicadas a la función $\frac{1}{1+x^2}$ generan los siguientes resultados

$$Q_1(f) = 0.3846, \quad Q_2(f) = 6.7949,$$

$$Q_3(f) = 2.0814, \quad Q_4(f) = 2.3740,$$
$$Q_5(f) = 2.3077, \quad Q_6(f) = 3.8704,$$
$$Q_7(f) = 2.8990, \quad Q_4(f) = 1.5005,$$

que muestran que aumentando el grado de la fórmula no siempre se alcanza mayor precisión (el resultado exacto es 2.7468) cuando la función tiene oscilaciones fuertes en las derivadas sucesivas. ◇

6.6 Cuadratura compuesta

El ejercicio 63 se pone de manifiesto que en ocasiones la mejora de la precisión no pasa por aumentar el grado de exactitud polinomial. En estas situaciones, suele ser más eficiente para calcular la integral de una función

f sobre un intervalo $[a,b]$, dividir primero $[a,b]$ en un número fijado de sub-intervalos, y aplicar en cada uno de ellos una fórmula de Newton-Cotes de bajo grado. Las fórmulas resultantes se conocen como fórmulas de Newton-Cotes compuestas.

- *Fórmula de cuadratura de los trapecios compuesta*: Si se utiliza la fórmula de los trapecios en cada uno de los m sub-intervalos en que se divide el intervalo $[a,b]$ se obtiene

$$\begin{aligned} Q_{1,m}(f) &= \sum_{i=0}^{m-1} \frac{h}{2}\left(f(x_i)+f(x_{i+1})\right) \\ &= \frac{h}{2}\left(f(x_0)+2\sum_{i=1}^{m-1} f(x_i)+f(x_m)\right) \end{aligned} \tag{6.4}$$

  donde $h=\frac{b-a}{m}$. Además el error está dado por

$$E_{1,m}(f) = -\sum_{i=0}^{m-1} \frac{f''(\eta_i)}{12}h^3$$

  para $\eta_i \in [x_i, x_{i+1}]$. Del teorema del valor medio discreto se deduce que existe $\eta \in [a,b]$ tal que

$$E_{1,m}(f) = -\frac{f''(\eta)}{12}mh^3 = -\frac{f''(\eta)}{12}(b-a)h^2. \tag{6.5}$$

- *Fórmula de cuadratura de Simpson compuesta*: Se divide el intervalo $[a,b]$ en $2m$ sub-intervalos $\{[x_j, x_{j+1}] : j = 0, \cdots 2m-1\}$ y se utiliza la fórmula de Simpson en cada uno de los intervalos $[x_{2i}, x_{2(i+1)}]$ para $i = 0, 1, \cdots, m-1$. De este modo se obtiene

$$Q_{2,m}(f) = \sum_{i=0}^{m-1} \frac{h}{6}\left(f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2(i+1)})\right). \tag{6.6}$$

donde $h = \frac{b-a}{m}$. También se puede expresar la fórmula de cuadratura del modo siguiente:

$$\begin{aligned} Q_{2,m}(f) &= \frac{h}{6}\left(f(x_0) + 2\sum_{i=1}^{m-1} f(x_{2i}) + 4\sum_{i=0}^{m-1} f(x_{2i+1}) + f(x_{2m})\right) \\ &= \frac{h}{3}\left(f(x_0) + 2\sum_{i=1}^{m-1} f(x_{2i}) + 2\sum_{i=0}^{m-1} f(x_{2i+1}) + f(x_{2m})\right) \\ &- \frac{h}{6}\left(f(x_0) + 2\sum_{i=1}^{m-1} f(x_{2i}) + f(x_{2m})\right) \\ &= \frac{4}{3}Q_{1,2m}(f) - \frac{1}{3}Q_{1,m}(f) \end{aligned} \tag{6.7}$$

que relaciona las fórmulas de cuadratura compuesta de Simpson y de los trapecios.

▪ Ejemplo **34** Si se utiliza la fórmula de los trapecios compuesta con 2 y 4 subintervalos para calcular $\int_0^1 x^4\,dx$, se obtiene

$$\begin{aligned} Q_{1,2}(x^4) &= \frac{1}{4}\left(0 + 2\frac{1}{2^4} + 1\right) = \frac{9}{32}, \\ Q_{1,4}(x^4) &= \frac{1}{8}\left(0 + 2\frac{1}{4^4} + 2\frac{1}{2^4} + 2\frac{3^4}{4^4} + 1\right) = \frac{113}{512}, \end{aligned}$$

y si se usa la fórmula de Simpson compuesta con 4 subintervalos, se obtiene

$$Q_{2,2}(x^4) = \frac{1}{12}\left(0 + \frac{4}{4^4} + \frac{1}{2^4} + \frac{1}{2^4} + 4\frac{3^4}{4^4} + 1\right) = \frac{77}{384}$$

Los valores obtenidos verifican la relación 6.7. ◇

Para analizar el error de la fórmula de cuadratura compuesta de Simpson se tendrá en cuenta que sumando los errores en cada sub-intervalo, el error total está dado por

$$E_{2,m}(f) = -\sum_{i=0}^{m-1} \frac{f^{(4)}(\eta_i)}{90}\left(\frac{h}{2}\right)^5$$

para $\eta_i \in [x_i, x_{i+1}]$. Del teorema del valor medio discreto se deduce que existe $\eta \in [a, b]$ tal que

$$E_{2,m}(f) = -\frac{f^{(4)}(\eta)}{90} m \left(\frac{h}{2}\right)^5 = -\frac{f^{(4)}(\eta)}{180}(b-a)\left(\frac{h}{2}\right)^4. \tag{6.8}$$

▬ EJERCICIO 64 *Calcular el número de sub-intervalos en los que hay que dividir el intervalo* $[1,3]$ *para que se pueda aproximar el valor de* $\ln 3$ *con error menor que* 10^{-4} *usando el método de los trapecios compuesto para evaluar la integral*

$$\ln 3 = \int_1^3 \frac{dt}{t} \quad .$$

Solución: La derivada segunda de la función $f(x) = \frac{1}{x}$ admite la siguiente acotación

$$|f''(x)| = \left|\frac{2}{x^3}\right| \leq 2$$

para $x \in [1,3]$. De acuerdo con la fórmula 6.5, para asegurar que el error es inferior a 10^{-4} basta tomar m verificando

$$\frac{2}{12} 2 \left(\frac{2}{m}\right)^2 < 10^{-4},$$

lo que equivale a $m > \frac{200}{\sqrt{3}}$. ◇

▬ EJERCICIO 65 *Calcular*

$$\int_0^1 e^x \, dx$$

por la fórmula compuesta de Simpson, calculando previamente la longitud h *de cada uno de los sub-intervalos, para garantizar cinco decimales exactos.*

Solución: Basta tomar h tal que

$$\frac{h^4}{2880} \max_{x\in[0,1]} e^x < 10^{-5}.$$

Es decir, si se toma

$$h < \left(\frac{2880 \times 10^{-5}}{e}\right)^{\frac{1}{4}} = 0.3208$$

se obtiene un error menor que 10^{-5}. Esta condición se cumple si $m = \frac{1}{h} = 4$. Consecuentemente, la integral puede ser aproximada mediante

$$\begin{aligned} \int_0^1 e^x \, dx &\approx \frac{1}{24}\left(1 + e + 2(e^{1/4} + e^{1/2} + e^{3/4}) + 4(e^{1/8} + e^{3/8} + e^{5/8} + e^{7/8})\right) \\ &= 1.71828 \qquad \diamond \end{aligned}$$

6.7 Fórmulas de Gauss

Cuando se intenta mejorar la precisión de una fórmula de cuadratura, es interesante saber si una elección estratégica de los puntos de interpolación puede ayudar en este sentido. En la fórmula 6.3 se establece que el error de cuadratura de una fórmula basada en la interpolación de Lagrange es

$$e = \int_a^b f[x_0, x_1, \cdots, x_n, x]\omega_{n+1}(x)\ dx = \langle f[x_0, x_1, \cdots, x_n, x], \omega_{n+1}(x)\rangle\ .$$

En el estudio de los polinomios de Chebyshev se probó que la elección de los nodos como las raíces de un polinomio de esta familia minimiza ω_{n+1} en la norma uniforme. Obviamente, parece que minimizando el factor en el que se puede influir directamente, en la fórmula del error, se podría mejorar la precisión de la aproximación.

Otro modo de analizar esta cuestión es maximizar el grado de los polinomios para los que la cuadratura es exacta, situando adecuadamente los nodos. En este sentido, la respuesta se encuentra de nuevo en las raíces de los polinomios ortogonales aunque no necesariamente de los de Chebyshev. En la argumentación será esencial el siguiente

Lema 4 *Si $f(x) = x^p$ con $p > n$, la diferencia dividida $f[x_0, x_1, \cdots, x_n, x]$ es un polinomio de grado menor o igual $p - n - 1$. Si $p \leq n$, se cumple que $f[x_0, x_1, \cdots, x_n, x] = 0$.*

Demostración: Para $p > n$ se utilizará un razonamiento de inducción en n. Para $n = 0$ se tiene que

$$f[x_0, x^p] = \frac{x^p - x_0^p}{x - x_0}$$

es obviamente un polinomio de grado $p - 1$. Se supone ahora que es cierto para $n - 1$. Puesto que $f[x_0, x_1, \cdots, x_{n-1}, x]$ es un polinomio de grado $p - n$ por inducción y $f[x_0, x_1, \cdots, x_{n-1}, x] - f[x_0, x_1, ..., x_n]$ tiene como raíz x_n entonces

$$f[x_0, x_1, \cdots ., x_n, x] = \frac{f[x_0, x_1, \cdots, x_{n-1}, x] - f[x_0, x_1, ..., x_n]}{x - x_n}$$

es un polinomio de grado $p - n - 1$.

Para $p \leq n$, se cumple que $f^{(n+1)} = 0$. De la fórmula 5.4 se deduce el resultado. $\diamond$

Las fórmulas de cuadratura basadas en la interpolación que utilizan como nodos de interpolación las raíces de un polinomio ortogonal se llaman genéricamente fórmulas de Gauss. El siguiente teorema de Gauss establece una propiedad excepcional de este tipo de cuadratura.

▬ TEOREMA 25 *Si se escogen como nodos de interpolación, las raíces del término de grado $n+1$ en una sucesión de polinomios ortogonales, la fórmula de cuadratura basada en esta interpolación es exacta para todos los polinomios de grado menor o igual que $2n+1$.*

Demostración: De acuerdo con las notaciones de las secciones anteriores, el polinomio ω_{n+1} es el término $n+1$ de la sucesión de polinomios ortogonales con coeficiente principal 1 y consecuentemente ω_{n+1} es ortogonal a todo polinomio de grado menor o igual que n. Si f es el polinomio x^p con $p > n$ entonces $f[x_0, x_1, \cdots, x_n, x]$ es un polinomio de grado menor o igual que $p-n-1$. Consecuentemente, el error e se anula para toda función f que sea un polinomio de grado p tal que $p-n-1 \leq n$ y la fórmula de cuadratura es exacta para polinomios de grado menor o igual que $2n+1$. ◇

En el cálculo de integrales existen situaciones en las que en el integrando es posible distinguir un factor ω que es una función integrable, toma valores positivos y que permanece fijo y un segundo factor que podría cambiar según los datos del problema. Es decir, en estas situaciones

$$\int_a^b f(x)\omega(x)\ dx$$

se distingue entre el integrado f (que se aproximará por interpolación) y el peso ω. En esta situación, se pueden utilizar los argumentos anteriores sin modificaciones esenciales, y de este modo, la fórmula del error de la cuadratura se convertiría en

$$e = \int_a^b f[x_0, x_1, \cdots, x_n, x]\omega_{n+1}(x)\omega(x)\ dx = \langle f[x_0, x_1, \cdots, x_n, x], \omega_{n+1}(x)\rangle_\omega .$$

Las fórmulas de cuadratura Gaussianas con peso, tendrían exactitud $2n+1$ si utilizan como nodos de interpolación las raíces del polinomio ortogonal de grado n respecto al producto escalar

$$\langle f, g\rangle_\omega = \int_a^b f(x)g(x)\omega(x)\ dx.$$

▬ EJERCICIO 66 *Calcular*

$$\int_{0.2}^{1.5} e^{-x^2} dx$$

utilizando una fórmula de Gauss de tres términos.

Solución: El cambio de variable $x = 0.65t + 0.85$ transforma el intervalo $[-1, 1]$ en el intervalo $[0.2, 1.5]$. Este cambio de variable no cambia el grado de un polinomio. Además, si p y q son polinomios ortogonales respecto al producto escalar

$$\langle p, q\rangle = \int_{0.2}^{1.5} p(x)q(x)\ dx$$

los polinomios $\hat{p}(t) = p(x(t))$ y $\hat{q}(t) = q(x(t))$ son ortogonales respecto al producto escalar

$$\langle p, q\rangle = \int_{-1}^{1} \hat{p}(t)\hat{q}(t)\ dt.$$

En definitiva, una sucesión de polinomios ortogonales de Legendre en $[-1, 1]$ puede transformarse en una sucesión de polinomios ortogonales de Legendre en $[0.2, 1.5]$ usando este cambio de variable.

En el ejemplo 24 se mostró que el polinomio ortogonal de grado 3 en $[-1, 1]$ es $p_3(t) = t^3 - \frac{3}{5}t$. Así pues, los nodos de Gauss son 0 y $\pm\sqrt{\frac{3}{5}}$. De acuerdo con la fórmula 6.2, los pesos de la cuadratura de Gauss son solución del sistema lineal

$$\begin{pmatrix} 1 & 1 & 1 \\ -\sqrt{0.6} & 0 & \sqrt{0.6} \\ 0.6 & 0 & 0.6 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ \frac{2}{3} \end{pmatrix}$$

Si se resuelve el sistema se obtiene, $\alpha_0 = \alpha_2 = \frac{5}{9}$ y $\alpha_1 = \frac{8}{9}$.

Los nodos de la fórmula de cuadratura en $[0.2, 1.5]$ se obtienen mediante el cambio de variable

$$x_0 = -0.65\sqrt{0.6} + 0.85,\ \ x_1 = 0.85,\ \ x_2 = 0.65\sqrt{0.6} + 0.85$$

y los polinomios básicos de Lagrange l_i asociados a los t_i se transforman en $\hat{l}_i$ correspondientes a los x_i. Por otra parte, del teorema del cambio de variable se obtiene

$$\hat{\alpha}_i = \int_{0.2}^{1.5} \hat{l}_i dx = \int_{-1}^{1} 0.65 l_i(t)\ dt = 0.65\alpha_i.$$

Finalmente, usando esta fórmula de cuadratura de Gauss se obtiene

$$\begin{aligned} \int_{0.2}^{1.5} e^{-x^2} dx &= \frac{0.65}{9}\left(5e^{-(-0.65\sqrt{0.6}+0.85)^2} + 5e^{-(0.65\sqrt{0.6}+0.85)^2} + 8e^{-0.85^2}\right) \\ &\approx 0.65860208567047 \qquad \diamond \end{aligned}$$

– EJERCICIO 67 *Determinar los polinomios de grado 2 de Chebyshev y de Legendre en el intervalo* $[-1, 1]$. *Construir las fórmulas de cuadratura en*

$[-1,1]$ *basadas en la interpolación de Lagrange en las raíces de ambos polinomios. Calcular mediante ambas fórmulas el valor aproximado de la integral*

$$\int_{-1}^{1} x^n \, dx.$$

Determinar el grado máximo de exactitud de ambas fórmulas.

Solución:

Polinomios de Chebyshev

$$\begin{aligned} p_0(x) &= 1 \\ p_1(x) &= x \\ p_2(x) &= 2x^2 - 1 \end{aligned}$$

Las raíces son $x = \pm\frac{\sqrt{2}}{2}$. Los pesos de la cuadratura verifican

$$\begin{pmatrix} 1 & 1 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

La fórmula de cuadratura es

$$Q_C(f) = f\left(-\frac{\sqrt{2}}{2}\right) + f\left(\frac{\sqrt{2}}{2}\right).$$

El valor aproximado de la integral es

$$Q_C(x^n) = \begin{cases} 2^{1-\frac{n}{2}}, & \text{si } n \text{ es par;} \\ 0, & \text{si } n \text{ es impar.} \end{cases}$$

Polinomios de Legendre

$$\begin{aligned} p_0(x) &= 1 \\ p_1(x) &= x \\ p_2(x) &= x^2 - \frac{1}{3} \end{aligned}$$

Las raíces son $x = \pm\frac{\sqrt{3}}{3}$. Los pesos de la cuadratura verifican

$$\begin{pmatrix} 1 & 1 \\ -\frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

La fórmula de cuadratura es

$$Q_L(f) = f\left(-\frac{\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right).$$

El valor aproximado de la integral es

$$Q_L(x^n) = \begin{cases} 2 \times 3^{-\frac{n}{2}}, & \text{si } n \text{ es par;} \\ 0, & \text{si } n \text{ es impar.} \end{cases}$$

El valor exacto de la integral es

$$\int_{-1}^{1} x^n dx = \begin{cases} \frac{2}{n+1}, & \text{si } n \text{ es par;} \\ 0, & \text{si } n \text{ es impar.} \end{cases}$$

Consecuentemente, el grado de exactitud de la cuadratura de Gauss-Legendre es 3 mientras que la de Chebyshev es 1. ◇

6.8 Ejercicios

– Ejercicio 68 *Se considera la siguiente fórmula de derivación numérica*

$$f'(x_0) \approx D(f;x_0) \equiv \frac{-5f(x_0) + 9f(x_0+2h) - 4f(x_0+3h)}{6h}$$

para un punto arbitrario x_0 y un tamaño de discretización h positivo. Calcular el error de aproximación de la fórmula si f es de clase C^4 y determinar si está basada en la interpolación.

Solución: Si se utilizan desarrollos de Taylor de f alrededor del punto x_0 se obtiene

$$f(x_0+2h) = f(x_0) + 2f'(x_0)h + 2f''(x_0)h^2 + \frac{4}{3}f'''(x_0)h^3 + \frac{2}{3}f^{(4)}(\xi_1)h^4.$$

$$f(x_0+3h) = f(x_0) + 3f'(x_0)h + \frac{9}{2}f''(x_0)h^2 + \frac{9}{2}f'''(x_0)h^3 + \frac{27}{8}f^{(4)}(\xi_2)h^4$$

para algunos puntos ξ_1 y ξ_2 en el intervalo $[x_0, x_0+3h]$. En consecuencia se obtiene

$$D(f;x_0) = f'(x_0) - f'''(x_0)h^2 + O(h^3).$$

Si f es un polinomio de grado menor o igual que 2, de acuerdo con la relación anterior, la fórmula es exacta. Puesto que se utilizan 3 puntos de evaluación, la fórmula está basada en la interpolación. ◇

▬ Ejercicio 69 *Se considera la siguiente fórmula de derivación numérica*

$$f'(x_0) \approx D(f;x_0) \equiv \frac{-2f(x_0+2h) + 16f(x_0+h) - 16f(x_0-h) + 2f(x_0-2h)}{24h}$$

para un punto arbitrario x_0 y un tamaño de discretización h positivo. Calcular el error de aproximación de la fórmula si f es de clase C^4 y determinar si está basada en la interpolación.

Solución: Si se utilizan desarrollos de Taylor de f alrededor del punto x_0 se obtiene

$$f(x_0+h) - f(x_0-h) = 2f'(x_0)h + \frac{1}{3}f'''(x_0)h^3 + \frac{1}{4!}\left(f^{(4)}(\xi_1) - f^{(4)}(\xi_2)\right)h^4.$$

$$f(x_0+2h) - f(x_0-2h) = 4f'(x_0)h + \frac{8}{3}f'''(x_0)h^3 + \frac{2^4}{4!}\left(f^{(4)}(\xi_3) - f^{(4)}(\xi_4)\right)h^4$$

para algunos puntos ξ_1, ξ_2, ξ_3 y ξ_4 en el intervalo $[x_0-2h, x_0+2h]$. En consecuencia se obtiene

$$\begin{aligned} D(f;x_0) &= f'(x_0) + \frac{2}{3}\left(f^{(4)}(\xi_1) - f^{(4)}(\xi_2)\right)h^3 \\ &- \frac{4}{3}\left(f^{(4)}(\xi_3) - f^{(4)}(\xi_4)\right)h^3 \\ &= f'(x_0) + O(h^3). \end{aligned}$$

Si f es un polinomio de grado menor o igual que 3, de acuerdo con la relación anterior, la fórmula es exacta. Puesto que se utilizan 4 puntos de evaluación, la fórmula está basada en la interpolación. $\diamond$

▬ EJERCICIO 70 *Calcular x_1 y x_2 para que la fórmula aproximada*

$$\int_{-1}^{1} f(x)\ dx \approx \frac{1}{3}(f(-1) + 2f(x_1) + 3f(x_2))$$

sea exacta para un polinomio del grado máximo posible.

Solución: La fórmula es exacta para el polinomio $p(x) = 1$. Si se impone la condición de que la fórmula de cuadratura sea exacta para los polinomios $\{x, x^2\}$ se obtienen las ecuaciones

$$\begin{aligned} 2x_1 + 3x_2 &= 1, \\ 2x_1^2 + 3x_1^2 &= 1, \end{aligned}$$

de las que se deducen los pares de soluciones

$$x_1 = \frac{1 \pm \sqrt{6}}{5}, \quad x_2 = \frac{3 \mp 2\sqrt{6}}{15}.$$

La fórmula resultante

$$\int_{-1}^{1} f(x)\ dx \approx \frac{1}{3}\left(f(-1) + 2f(\frac{1 \pm \sqrt{6}}{5}) + 3f(\frac{3 \mp 2\sqrt{6}}{15})\right)$$

no es exacta para polinomios de grado 3 ya que

$$\begin{aligned} \int_{-1}^{1} f(x)\ dx &= 0 \neq \frac{1}{3}\left(-1 + 2\left(\frac{1 \pm \sqrt{6}}{5}\right)^3 + 3\left(\frac{3 \mp 2\sqrt{6}}{15}\right)^3\right) \\ &= \begin{cases} -0.1165, \\ -0.2035 \end{cases} . \qquad \diamond \end{aligned}$$

▬ EJERCICIO 71 *Demostrar que existe un número $c > 0$ en el intervalo $[0, 1]$ tal que la fórmula*

$$\int_{-1}^{1} f(x)\ dx \approx f(c) + f(-c)$$

es exacta para todo polinomio de grado menor o igual que 3. Generalizando la fórmula anterior, demostrar que existen dos constantes c_1 y c_2 en $[a, b]$ tal que la fórmula

$$\int_{a}^{b} f(x)\ dx \approx \frac{b-a}{2}(f(c_1) + f(c_2))$$

es exacta para todo polinomio de grado menor e igual que 3. *Expresar* c_1 *y* c_2 *en función de* a *y* b.

Solución: La fórmula es siempre exacta para los polinomios $1, x$ y x^3. Para que sea exacta para el polinomio x^2 es necesario que $c = \frac{\sqrt{3}}{3}$. No es exacta para x^4 ya que

$$\int_{-1}^{1} x^4\, dx = \frac{2}{5} \neq 2\left(\frac{\sqrt{3}}{3}\right)^4 = \frac{2}{9}.$$

La fórmula de cambio de variable

$$\bar{x} = \frac{b-a}{2}x + \frac{b+a}{2}$$

transforma el intervalo $[-1, 1]$ en el intervalo $[a, b]$ y no cambia el grado de los polinomios. Consecuentemente la fórmula de cuadratura es generalizable siendo

$$c_1 = -\frac{b-a}{2}\frac{\sqrt{3}}{3} + \frac{b+a}{2}, \quad c_2 = \frac{b-a}{2}\frac{\sqrt{3}}{3} + \frac{b+a}{2}. \qquad \diamond$$

– Ejercicio 72 *Calcular*

$$\int_{0.2}^{1.5} e^{-x^2} dx$$

utilizando la fórmula de Gauss de tres términos

nodos	*pesos*
$-\sqrt{0.6}$	5/9
0	8/9
$\sqrt{0.6}$	5/9

para el intervalo $[-1, 1]$.

Solución: El cambio de variable $x = 0.65t + 0.85$ transforma el intervalo $[-1, 1]$ en el intervalo $[0.2, 1.5]$. Si se usa este cambio de variable se obtiene

$$\int_{0.2}^{1.5} e^{-x^2} dx = \int_{-1}^{1} 0.65 e^{-(0.65t+0.85)^2}\, dt.$$

Finalmente, usando la fórmula de cuadratura de Gauss se obtiene

$$\begin{aligned}\int_{0.2}^{1.5} e^{-x^2} dx &\approx \frac{0.65}{9}\left(5e^{-(-0.65\sqrt{0.6}+0.85)^2} + 5e^{-(0.65\sqrt{0.6}+0.85)^2} + 8e^{-0.85^2}\right) \\ &= 0.65860208567047 \qquad \diamond\end{aligned}$$

Ejercicio 73 *Determinar los pesos y nodos de la fórmula de cuadratura*

$$Q(f) = \alpha_0 f(x_0) + \alpha_1 f(x_1)$$

que aproxime el valor de la integral

$$\int_{-1}^{1} |x| f(x)\ dx$$

con el máximo orden de exactitud. Indicación: Resolver el problema mediante

1. *Imponiendo directamente las condiciones de exactitud sobre los elementos de la base formada por los monomios de coeficiente 1.*
2. *Construyendo la sucesión de polinomios ortogonales respecto al producto escalar*
$$\langle f, g\rangle = \int_{-1}^{1} |x| f(x) g(x)\ dx$$

Ejercicio 74 *Calcular la longitud del arco de la curva $y = x - \frac{x^2}{2}$ desde $x = 0$ hasta $x = 1$ con error inferior a 10^{-2}. Indicación: La longitud del arco de la curva $y = y(x)$ entre $x = a$ y $x = b$ está dado por la fórmula*

$$l = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\ dx.$$

Solución: La longitud del arco de la curva entre 0 y 1 está dado por

$$l = \int_0^1 \sqrt{1 + (1-x)^2}\ dx.$$

Lo más simple para evaluar esta integral parece usar la fórmula compuesta del trapecio

$$\int_a^b f(x)\ dx \approx \frac{h}{2}\left(f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right)$$

con $x_0 = a < x_1 < \cdots < x_n = b$ y $h = \frac{b-a}{n}$. El error que introduce esta fórmula es

$$R_n(f) = -\frac{b-a}{12} h^2 f''(\xi)$$

para algún $\xi \in [a, b]$. Se trata de encontrar n tal que $|R_n(f)| \leq 10^{-2}$. En este caso, esta desigualdad se convierte en

$$\frac{1}{12n^2}\,|f''(\xi)| < 10^{-2}$$

siendo

$$f''(x) = -\frac{1}{((x-1)^2+1)^{3/2}}.$$

El máximo de $|f''|$ en el intervalo $[0,1]$ se alcanza en $x = 1$ y en consecuencia es

$$M = \max_{x\in[0,1]} |f''(x)| = 1.$$

Bastará tomar $n = 3 > \sqrt{10^2/12}$ para garantizar la precisión deseada. Finalmente el valor aproximado buscado es

$$l \approx \frac{1}{6}\left(f(0) + 2(f(1/3) + f(2/3)) + f(1)\right) = \frac{11}{9} = 1.1543$$

mientras que el valor exacto es 1.1478. ◇

▬ EJERCICIO 75 *Determinar los coeficientes a_0, a_1 y a_2 para que la fórmula de cuadratura*

$$Q(f) = a_0 f(0) + a_1 f(\frac{1}{2}) + a_2 f(1)$$

que aproxima el valor de la integral

$$\int_0^1 f(x)x^3 \, dx,$$

tenga grado máximo de exactitud.

Solución: La condición de exactitud

$$Q(x^i) = \int_0^1 x^i x^3 \, dx = \frac{1}{i+4}$$

para $i = 0, 1, 2$, conduce al sistema lineal

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & \frac{1}{2} & 1 \\ 0 & \frac{1}{4} & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{5} \\ \frac{1}{6} \end{pmatrix}.$$

Si se resuelve este sistema lineal se obtiene la fórmula de cuadratura

$$Q(f) = -\frac{1}{60}f(0) + \frac{2}{15}f(\frac{1}{2}) + \frac{2}{15}f(1)$$

Puesto que

$$Q(x^3) = \frac{2}{15}\frac{1}{2^3} + \frac{2}{15} = \frac{3}{20} \neq \int_0^1 x^6 dx = \frac{1}{7}$$

el grado máximo de exactitud es 2. ◇

Parte II: Resolución numérica de ecuaciones

CAPÍTULO 7

Resolución numérica de ecuaciones no-lineales escalares

7.1 Introducción

Las raíces de una ecuación $f(x) = 0$, definida por una función real f que no es un polinomio de grado bajo, no pueden ser determinadas por técnicas analíticas salvo en situaciones muy particulares. Es decir, en general no es posible que las incógnitas de una ecuación, que no es lineal, puedan ser despejadas mediante expresiones fácilmente evaluables. En el caso de las ecuaciones definidas por las funciones elementales más simples, los polinomios, se hizo un gran esfuerzo por encontrar expresiones, admitiendo radicales, para sus raíces. En este sentido, pudo resultar alentador el éxito de Ferrari cuando encontró en el siglo XVI, una expresión con radicales para la ecuación general polinómica de cuarto orden. No obstante, la discusión finaliza con los resultados de Ruffini y Abel que prueban la imposibilidad de resolver la polinómica de quinto orden. En general, las soluciones de las ecuaciones numéricas que no son lineales, deben ser aproximadas por sucesiones que converjan a ellas y el modo más común de generarlas es a través de algoritmos iterativos. Por otra parte, muchas veces pierde sentido buscar expresiones analíticas cuya evaluación puede resultar muy complicada, para resolver ecuaciones que no son lineales cuando un simple algoritmo puede permitir construir, de un modo eficiente, una sucesión que aproxima la solución.

Del mismo modo que ocurre con otros problemas que es necesario resolver numéricamente, no existe el *mejor* modo de resolver una ecuación si la

cuestión se plantea de un modo muy general. La elección del método más adecuado para resolver una ecuación que no es lineal, depende de la forma, de la regularidad y del coste de evaluación de la función que la define, así como de otros factores. Por esta razón, en este capítulo se presenta una amplia colección de métodos en la que pueda ser seleccionado el método adecuado, cuando se quiera resolver un problema concreto. En este capítulo se tratará de la resolución de ecuaciones escalares de una sola variable y en el próximo, de la resolución de sistemas de ecuaciones que no son lineales.

7.2 Método de dicotomía o bisección

La aplicación de un algoritmo numérico a la resolución de una ecuación, debe estar precedida de un análisis de la existencia y de la localización de la raíz que se quiere determinar. Para garantizar que existe una solución de una ecuación en un intervalo, el procedimiento más simple es el basado en el conocido

TEOREMA 26 *(de Bolzano). Si f es una función continua en un intervalo $[a,b]$ tal que $f(a)f(b)<0$ entonces existe al menos una raíz en (a,b).*

Evidentemente, la condición $f(a)f(b)<0$ implica la existencia de una solución pero no garantiza que sea única. De hecho, en estas condiciones, la ecuación puede tener un número infinito de raíces. En la figura 7.1 se representan las gráficas de las funciones $f(x)=1.5\cos(2\pi x)-x-1$ y

$$f(x)=\begin{cases} x\,\mathrm{sen}(1/|x|), & \text{si } x\neq 0\\ 0, & \text{si } x=0\end{cases}$$

que tienen 3 e infinitas raíces, respectivamente.

Una aplicación directa del teorema de Bolzano a la resolución de ecuaciones que no son lineales, es el llamado método de dicotomía o bisección. La idea básica de este método consiste en dividir el intervalo en dos, en cada iteración, seleccionando uno de los dos sub-intervalos con el criterio de que la función f tome valores de signo opuesto en sus extremos. De este modo, siempre se tiene garantizada la existencia de solución en su interior. Obviamente, si en alguno de estos extremos, la función alcanza el valor cero, el algoritmo se para.

A continuación se describe el método con más detalle. Se considera la sucesión de intervalos encajados generada del modo siguiente:

- Sea $I^{(0)}=[a^{(0)},b^{(0)}]=I$.

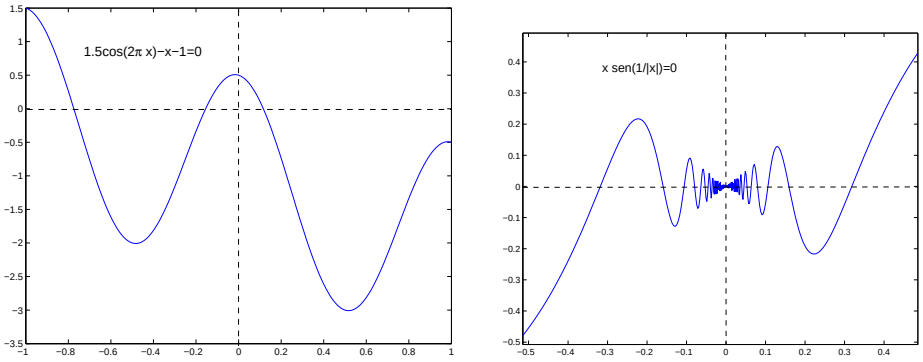


Figura 7.1: Ecuaciones con un número impar o infinito de raíces

- Construido $I^{(k-1)} = [a^{(k-1)}, b^{(k-1)}]$ tal que $f(a^{(k-1)})f(b^{(k-1)}) < 0$, se evalúa $f(c^{(k-1)})$ para

$$c^{(k-1)} = \frac{a^{(k-1)} + b^{(k-1)}}{2}$$

y de acuerdo con el resultado se considera el nuevo sub-intervalo $I^{(k)} = [a^{(k)}, b^{(k)}]$ definido por

$$a^{(k)} = \begin{cases} a^{(k-1)}, & \text{si } f(c^{(k-1)})f(a^{(k-1)}) < 0 \\ c^{(k-1)}, & \text{si } f(c^{(k-1)})f(a^{(k-1)}) > 0 \end{cases}$$

$$b^{(k)} = \begin{cases} b^{(k-1)}, & \text{si } f(c^{(k-1)})f(b^{(k-1)}) < 0 \\ c^{(k-1)}, & \text{si } f(c^{(k-1)})f(b^{(k-1)}) > 0 \end{cases}$$

Si $f(c^{(k-1)}) = 0$ el algoritmo se detiene ya que se ha encontrado una raíz $x = c^{(k-1)}$.

Si el algoritmo no se detiene en ninguna etapa, las sucesiones $\{a^{(k)}\}$ y $\{b^{(k)}\}$ son monótonas y acotadas y consecuentemente, convergentes. Además se cumple que

$$b^{(k)} - a^{(k)} = \frac{b-a}{2^k}, \qquad f(a^{(k)})f(b^{(k)}) < 0$$

para todo $k > 0$. Consecuentemente, el valor

$$\alpha = \lim_{k\to\infty} a^{(k)} = \lim_{k\to\infty} b^{(k)} = \lim_{k\to\infty} c^{(k)}$$

verifica que $f(\alpha)^2 \leq 0$, lo que implica que $f(\alpha) = 0$.

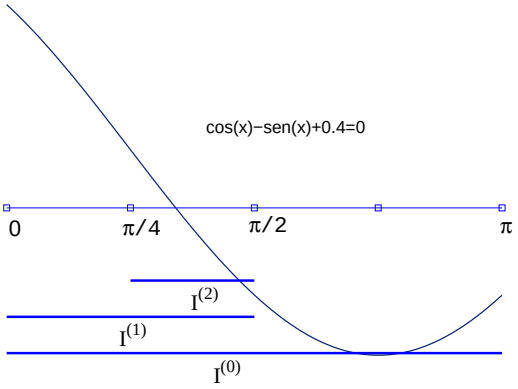


Figura 7.2: Bisección en la ecuación $\cos x - \operatorname{sen} x + 0.4 = 0$

Finalmente, para obtener una estimación del error de aproximación basta tener en cuenta que

$$|c^{(k)} - \alpha| \leq \frac{b-a}{2^{k+1}}.$$

Así, para obtener una precisión ε en la aproximación, basta escoger

$$k \geq \frac{\ln((b-a)/\varepsilon)}{\ln 2} - 1 \;\Rightarrow\; |c^{(k)} - \alpha| < \varepsilon. \tag{7.1}$$

▬ EJERCICIO 76 *Utilizar el método de dicotomía para encontrar una raíz de la ecuación* $xe^x = 1$ *en el intervalo* $[0,1]$*, con una precisión de dos decimales.*

Solución: La ecuación tiene una solución en $[0,1]$ puesto que $f(x) = xe^x - 1$ verifica que $f(0)f(1) < 0$. En estas condiciones, el método de dicotomía genera una sucesión convergente a una raíz en el intervalo $[0,1]$.

De la acotación 7.1 se deduce que el número mínimo de iteraciones a realizar con el método de dicotomía para asegurar la convergencia con error menor a $\varepsilon = 10^{-2}$ en la sucesión $\{c^{(k)}\}$, es

$$k > 2\frac{\ln 10}{\ln 2} - 1 = 5.6439\ .$$

La sucesión de intervalos encajados generada por el algoritmo es la siguiente:

k	0	1	2	3	4	5	6
$a^{(k)}$	0	0.5	0.5	0.5	0.5625	0.5625	0.5625
$b^{(k)}$	1	1	0.75	0.625	0.625	0.5938	0.5781

De la tabla se deduce que $\alpha = 0.5703$ es un valor aproximado de la raíz con dos decimales exactos. $\diamond$

Una ventaja del método de bisección es que únicamente se requiere conocer el signo de la función en los extremos de los intervalos y es muy simple a programar. Precisamente ya que no hace uso del valor absoluto que la función toma en los extremos, no distingue entre aquellos en los que valor es bajo y consecuentemente estarán próximos a la raíz y aquellos que toman valores altos y que razonablemente estarán más alejados. Esto hace que el método pueda resultar lento para aproximar la solución.

7.3 Métodos de punto fijo

Un punto x que verifica que $x = g(x)$ se dice que es un punto fijo de la función g. Raíces de ecuaciones y puntos fijos de funciones son conceptos relacionados aunque no de un modo canónico. Si x es una raíz de una ecuación $f(x) = 0$, es también un punto fijo de funciones tales como

$$g(x) = x - \lambda f(x)$$

donde λ es un escalar arbitrario. Recíprocamente, si x es un punto fijo de la función g, es también una raíz de la ecuación

$$f(x) \equiv x - g(x) = 0.$$

Así pues, es razonable pensar que se pueden utilizar las técnicas de aproximación de puntos fijos para resolver ecuaciones que no son lineales.

El conocido siguiente resultado establece condiciones que garantizan la existencia de un punto fijo de una función

Teorema 27 (de Brouwer) *Si g es una función continua en un intervalo $I = [a, b]$, que transforma I en un subconjunto de I, entonces existe al menos un punto fijo α de g en el intervalo I.*

Demostración: Se considera la función f definida por $f(x) = x - g(x)$ para todo $x \in I = [a, b]$. En los extremos f cumple que

$$f(a) = a - g(a) \leq 0, \quad f(b) = b - g(b) \geq 0,$$

ya que $g(a) \geq a$ y $g(b) \leq b$. Del teorema de Bolzano se deduce que existe un punto $\xi \in I$ tal que $f(\xi) = 0$. Finalmente se concluye que ξ es un punto fijo de g. $\diamond$

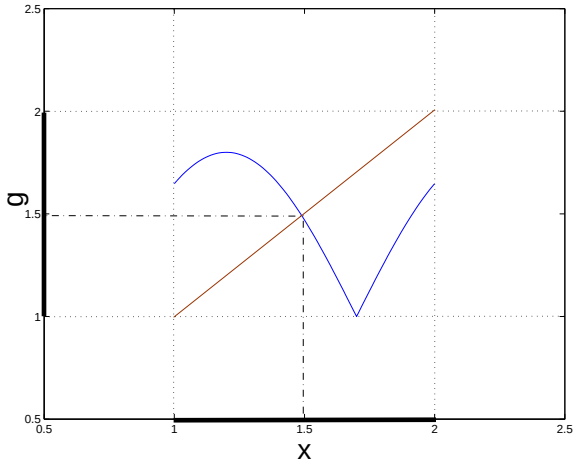


Figura 7.3: Ilustración del teorema de Brouwer

▪ EJEMPLO **35** Las raíces de la ecuación $e^{x+1} - 2x = 3$ son puntos fijos

$$x = \ln(2x + 3) - 1$$

de la función $g(x) = \ln(2x+3)-1$ y recíprocamente. Puesto que g es monótona creciente, es inmediato verificar que transforma el intervalo $I = [0, 1]$ en el intervalo $[\ln 3 - 1, \ln 5 - 1] \subset I$. De la continuidad de g en $[0, 1]$ y del teorema de Brouwer, se deduce que g tiene un punto fijo en el intervalo $[0, 1]$. De ello se sigue que la ecuación tiene una raíz en ese intervalo. ◇

El siguiente resultado establece condiciones para garantizar no sólo la existencia sino también la unicidad de un punto fijo de una función

▬ TEOREMA 28 (DE LA CONTRACCIÓN, DE BANACH) . *Si g es una función de clase C^1 en un intervalo $I = [a, b]$, tal que transforma I en un subconjunto de I y que su derivada verifica la siguiente acotación*

$$|g'(x)| < 1, \quad \textit{para todo } x \in [a, b]$$

entonces g posee en I un único punto fijo α. La sucesión generada por el algoritmo $x^{(n)} = g(x^{(n-1)})$ para $x^{(0)} \in I$ arbitrario, es convergente a α. Además, se cumple que

$$\lim_{k\to\infty} \frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} = g'(\alpha) \tag{7.2}$$

Demostración: La existencia de punto fijo se desprende directamente del teorema de Brouwer. Por otra parte, del teorema del valor medio se deduce que g es contractiva en I. Es decir, para todo $x, y \in I$ se tiene que

$$|g(x) - g(y)| \leq \max_{\xi \in I} |g'(\xi)||x - y| = \rho|x - y|$$

donde

$$\rho = \max_{\xi \in I} |g'(\xi)| < 1$$

por ser g' una función continua en el compacto I. Si existieran dos puntos fijos α y β en I, puesto que g es contractiva, se cumpliría que

$$|\alpha - \beta| = |g(\alpha) - g(\beta)| \leq \rho|\alpha - \beta| < |\alpha - \beta|$$

lo cual es imposible.

Por otra parte, si α es punto fijo de g, se cumple que

$$|x^{(k)} - \alpha| = |g(x^{(k-1)}) - g(\alpha)| \leq \rho|x^{(k-1)} - \alpha|.$$

Si se reitera este razonamiento, se obtiene

$$|x^{(k)} - \alpha| \leq \rho^k |x^{(0)} - \alpha| \tag{7.3}$$

lo que prueba que la sucesión generada por aproximaciones sucesivas es convergente al punto fijo.

Finalmente, del teorema del valor medio se deduce que

$$\lim_{k \to \infty} \frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} = \lim_{k \to \infty} g'(\xi^{(k)}) = g'(\alpha)$$

para alguna sucesión $\xi^{(k)}$ de puntos intermedios a α y $x^{(k)}$ para todo k. ◇

Se puede extender este resultado a intervalos cerrados que no están acotados. Los argumentos utilizados para probar el teorema seguiría siendo válidos, salvo que la ausencia de compacidad en el intervalo I requería una condición más fuerte para garantizar que g fuese una contracción. No obstante, bastaría imponer la condición de la existencia de $\rho \in (0, 1)$ tal que

$$|g'(x)| < \rho < 1, \quad \text{para todo } x \in I.$$

▪ Ejemplo **36** La función $g(x) = \ln(2x + 3) - 1$, definida en el intervalo $[0, 1]$, verifica que

$$\max_{x \in [0,1]} |g'(x)| = \max_{x \in [0,1]} \frac{2}{2x + 3} = \frac{2}{3} < 1.$$

Por otra parte, g es creciente y por lo tanto transforma el intervalo $[0,1]$ en el intervalo

$$g([0,1]) = [g(0), g(1)] = [\ln 3 - 1, \ln 5 - 1] \subset [0,1].$$

Puesto que g es de clase C^1, del teorema de la contracción se deduce que tiene un único punto fijo en el intervalo $[0,1]$ que puede ser aproximado por iteraciones sucesivas con la función g. $\diamond$

▬ EJERCICIO 77 *La ecuación* $f(x) = x^2 - 2x - 3 = 0$ *puede expresarse de las siguientes formas:*

$$a)\ x = \sqrt{2x+3}, \quad b)\ x = \frac{3}{x-2}, \quad c)\ x = \frac{x^2-3}{2}.$$

Estudiar el comportamiento del método iterativo asociado a cada una de esas expresiones, partiendo de $x_0 = 4$ *e ilustrarlo gráficamente.*

Solución:

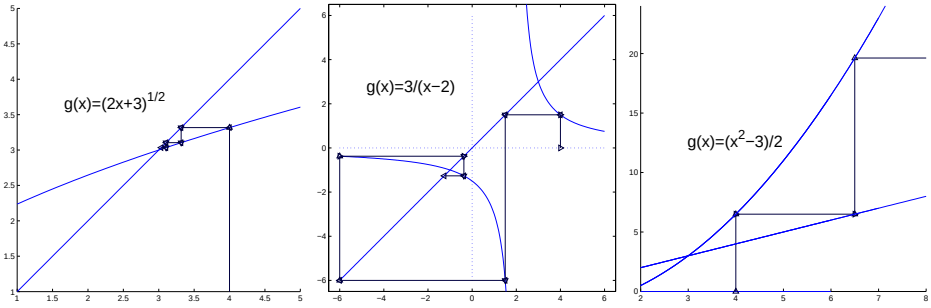


Figura 7.4: Resolución de la ecuación $x^2 - 2x - 3$

Sean $g_1(x) = \sqrt{2x+3}$, $g_2(x) = \frac{3}{x-2}$ y $g_3(x) = \frac{x^2-3}{2}$. La función g_1 es contractiva en $I = [0, \infty)$ ya

$$\max_{x \in I} |g'(x)| = \max_{x \in I} \left| \frac{1}{\sqrt{2x+3}} \right| = \frac{1}{\sqrt{3}}.$$

Además g_1 transforma I en un subconjunto de I. La sucesión generada por las aproximaciones sucesivas, partiendo de $x^{(0)} = 4$, es convergente a la única raíz que la ecuación tiene en I.

La función g_2 es contractiva en $I = (-\infty, 0]$ ya

$$\max_{x\in I} |g'(x)| = \max_{x\in I} \left| \frac{-3}{(x-2)^2} \right| = \frac{3}{4}.$$

Además g_2 transforma I en un subconjunto de I. La sucesión generada por las aproximaciones sucesivas, partiendo de $x^{(0)} = 4$, es convergente a la única raíz que la ecuación tiene en I ya que, si bien $x^{(0)} = 4$ no pertenece a I, a partir de $x^{(2)} = -6$ ya se encuentra en I.

La función g_3 no es contractiva en ningún entorno de las raíces $\alpha = -1$ o $\alpha = 3$. La sucesión generada por las aproximaciones sucesivas, partiendo de $x^{(0)} = 4$, es divergente hacia ∞. ◇

▬ TEOREMA 29 *(Estimación del error) Bajo las mismas hipótesis y notaciones que en el teorema anterior, si ρ representa la constante definida por*

$$\rho = \max_{x\in I} |g'(x)|,$$

para $x^{(0)} \in I$, se tiene la siguiente estimación

$$|x^{(k)} - \alpha| < \frac{\rho^k}{1-\rho} |x^{(1)} - x^{(0)}|.$$

Demostración: Para todo k y m, tal que $k < m$, de la desigualdad triangular del valor absoluto se deduce que

$$\begin{aligned} |x^{(k)} - x^{(m)}| &\leq |x^{(k)} - x^{(k+1)}| + |x^{(k+1)} - x^{(k+2)}| + \cdots + |x^{(m-1)} - x^{(m)}| \\ &\leq \rho^k |x^{(0)} - x^{(1)}| + \rho^{k+1} |x^{(0)} - x^{(1)}| + \cdots + c^{m-1} |x^{(0)} - x^{(1)}| \\ &= \rho^k \frac{\rho^{m-k} - 1}{\rho - 1} |x^{(0)} - x^{(1)}|. \end{aligned}$$

Si se hace tender $m \to \infty$ manteniendo k fijo y se tiene en cuenta que $\rho^{m-k} \to 0$ y $x^{(m)} \to \alpha$, se obtiene el resultado buscado. ◇

▬ EJERCICIO 78 *Determinar un número suficiente de iteraciones que hay que realizar en*

$$x^{(k+1)} = \frac{\left(x^{(k)}\right)^2 + 3}{2x^{(k)}}$$

para que converja a un punto fijo con error menor de 10^{-5}, partiendo de un punto arbitrario del intervalo $I = \left[\frac{3}{2}, 1\right]$.

Solución: Puesto que

$$\rho = \max_{x\in I} |g'(x)| = \max_{x\in I} \left|\frac{x^2-3}{2x^2}\right| = \frac{1}{3} < 1,$$

y $g(I) \subset I$, un condición suficiente para que el error sea inferior a 10^{-5} es la siguiente

$$\frac{\frac{1}{3^k}}{1-\frac{1}{3}} \left|\frac{\left(x^{(0)}\right)^2+3}{2x^{(0)}} - x^{(0)}\right| < 10^{-5}.$$

Puesto que

$$\left|\frac{\left(x^{(0)}\right)^2+3}{2x^{(0)}} - x^{(0)}\right| > \frac{1}{4}$$

para todo $x^{(0)} \in I$, se tiene que la condición

$$k > 1 + \frac{5\ln 10 - \ln 8}{\ln 3} \approx 9.5867$$

es suficiente para garantizar la precisión deseada. ◇

7.4 Velocidad de convergencia

La convergencia de un método no es el único aspecto que se analiza cuando se está decidiendo sobre el procedimiento a utilizar para resolver una ecuación no-lineal. Evidentemente, la rapidez de un método parece una propiedad muy deseable, si bien, la afirmación de que un método es rápido ó lento es algo que resulta muy impreciso y que sería necesario matizar y cuantificar. Es posible establecer algunos criterios con los que comparar la rapidez de los algoritmos en base a los que se podrían después ordenar las preferencias por un método. Estos criterios se basan en el análisis del comportamiento de la sucesión de errores

$$\{|e^{(k)}|\} = \{|x^{(k)} - \alpha|\}.$$

Se dice que una sucesión $\{\varepsilon^{(k)}\}$ de números positivos converge a 0 con orden de convergencia p, si verifica que existen una constante positiva C y un entero positivo k_0, tales que

$$\frac{\varepsilon^{(k+1)}}{\left(\varepsilon^{(k)}\right)^p} < C,$$

para todo $k \geq k_0$. Basada en esta definición se introduce la siguiente: La sucesión $\{x^{(k)}\}$ generada por un algoritmo es convergente con orden p a una

raíz α si la sucesión de errores converge a 0 con orden p. Algunos autores llaman a $\{x^{(k)}\}$ sucesión $R-$convergente de orden p si la sucesión de errores está acotada por una sucesión $\{\varepsilon^{(k)}\}$ convergente a 0 con orden p.

Por ejemplo, la sucesión de errores en el método de dicotomía está acotada por la sucesión $\frac{b-a}{2^k}$ que tiene convergencia de orden 1 y por lo tanto, la sucesión de errores es R-convergente de orden 1.

En el caso particular $p = 1$, si el siguiente límite

$$\mu = \lim_{k\to\infty} \frac{|x^{(k+1)} - \alpha|}{|x^{(k)} - \alpha|}$$

existe, entonces

- si $\mu \in (0, 1)$ se dice que la convergencia es lineal;
- si $\mu = 0$ se dice que la convergencia es superlineal y
- si $\mu = 1$ se dice que la convergencia es sublineal.

En esta situación, se define la velocidad asintótica de convergencia como el número real $R = -\ln \mu$.

El teorema de la contracción garantiza que las sucesiones generadas por las aproximaciones sucesivas son de orden al menos 1 y su convergencia es al menos lineal y su velocidad asintótica es $-\ln |g'(\alpha)|$.

▪ Ejemplo **37** En el ejercicio 77, el método iterativo definido por la función g_1 tiene velocidad asintótica de convergencia igual a $\ln 3$ cuando se aproxima a la raíz $\alpha = 3$. También la velocidad asintótica de convergencia del método definido por g_2 cuando se aproxima a $\alpha = -1$, es $\ln 3$. ◇

▬ Teorema 30 *Si g es una función de clase C^{p+1} en un intervalo I tal que*

$$\begin{aligned} \frac{d^i g}{dx^i}(\alpha) &= 0, \quad \text{para } i = 1, 2, \cdots, p, \\ \frac{d^{p+1} g}{dx^{p+1}}(\alpha) &\neq 0, \end{aligned}$$

para un punto fijo α de g, entonces toda sucesión convergente generada el método de punto fijo asociado a g, tiene $p + 1$ como orden de convergencia.

Demostración: Si se desarrolla por Taylor la función g alrededor del punto α se obtiene

$$x^{(k+1)} = g(x^{(k)}) = \alpha + \frac{1}{(p+1)!} \frac{d^{p+1} g}{dx^{p+1}}(\xi^{(k)})(x^k - \alpha)^{p+1}$$

siendo $\xi^{(k)}$ un punto intermedio entre α y $x^{(k)}$. De esta relación se deduce

$$\lim_{k\to\infty}\frac{x^{(k+1)}-\alpha}{(x^{(k)}-\alpha)^{p+1}}=\lim_{k\to\infty}\frac{1}{(p+1)!}\frac{d^{p+1}g}{dx^{p+1}}(\xi^{(k)})=\frac{1}{(p+1)!}\frac{d^{p+1}g}{dx^{p+1}}(\alpha). \qquad \diamond$$

▪ EJEMPLO **38** La siguiente función

$$g(x)=(x-1)^2\,\mathrm{sen}(\pi x)+1$$

tiene un punto fijo en $x=1$. Puesto que

$$\begin{aligned} g'(x) &= 2(x-1)\,\mathrm{sen}(\pi x)+\pi\cos(\pi x)(x-1)^2,\\ g''(x) &= 2\,\mathrm{sen}(\pi x)+4\pi\cos(\pi x)(x-1)-\pi^2\,\mathrm{sen}(\pi x)(x-1)^2,\\ g'''(x) &= 6\pi\cos(\pi x)-\pi^3\cos(\pi x)(x-1)^2-6\pi^2\,\mathrm{sen}(\pi x)(x-1), \end{aligned}$$

se tiene que

$$\begin{aligned} g'(1) &= 0,\\ g''(1) &= 0,\\ g'''(1) &= -6\pi. \end{aligned}$$

Consecuentemente, el orden de convergencia a $x=1$ de la sucesión generada por aproximaciones sucesivas es 3. $\diamond$

Si un algoritmo produce sucesiones convergentes (de orden p) para un punto de partida $x^{(0)}$ arbitrario en un intervalo I, se dice que es globalmente convergente en I, mientras que se dice que es localmente convergente (de orden p) si se tiene la garantía de que esto sólo ocurre para los puntos $x^{(0)}$ de un entorno $[\alpha-\delta,\alpha+\delta]$ de la raíz α, sin precisar su radio δ. En este sentido, se podría debilitar el teorema de la contracción como muestra el siguiente

▬ TEOREMA 31 (DE OSTROWSKI) *Si g es una función de clase C^1 en un entorno de un punto fijo α de g y verifica que*

$$|g'(\alpha)|<1$$

entonces el método de punto fijo asociado a g es localmente convergente a α.

Demostración: Puesto que g' es una función continua, existe un intervalo $I=[\alpha-\varepsilon,\alpha+\varepsilon]$ tal que

$$\max_{x\in I}|g'(x)|\le\rho<1.$$

Si $\{x^{(k)}\}$ es la sucesión generada por aproximaciones sucesivas, partiendo de un punto de I, entonces se cumple que

$$|\alpha - x^{(k)}| = |g(\alpha) - g(x^{(k-1)})| \leq \rho|\alpha - x^{(k-1)}| \leq \cdots \leq \rho^k|\alpha - x^{(0)}|$$

lo que prueba que $\{x^{(k)}\} \to \alpha$. $\diamond$

7.5 Método de la secante

Para mejorar la convergencia de los métodos de punto fijo

$$x^{(k+1)} = x^{(k)} - \lambda f(x^{(k)})$$

obtenidos por relajación de la ecuación $f(x) = 0$, se pueden considerar aquellos en los que el parámetro de relajación se modifica en cada iteración

$$x^{(k+1)} = x^{(k)} - \lambda^{(k)} f(x^{(k)}).$$

Un método de esta clase corresponde a la elección

$$\lambda^{(k)} = \frac{1}{f[x^{(k-1)}, x^{(k)}]} = \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}.$$

De este modo $x^{(k+1)}$ es el punto de corte de la secante a la gráfica de f en los puntos $x^{(k-1)}$ y $x^{(k)}$, con el eje OX. Es necesario destacar que la puesta en práctica de este método requiere el conocimiento de dos valores iniciales $x^{(0)}$ y $x^{(1)}$.

Un resultado de convergencia local es el que establece el siguiente

– Teorema 32 *Sean f una función de clase C^2 en un intervalo cerrado I tal que $f'(x) \neq 0$ en I y α una raíz de la ecuación $f(x) = 0$ en I. Si los datos iniciales $x^{(0)}$ y $x^{(1)}$ están suficientemente próximos a α, entonces la sucesión generada por el método de la secante converge a α con orden $p = \frac{1+\sqrt{5}}{2}$.*

Demostración: Directamente se comprueba la igualdad

$$x^{(k+1)} - \alpha = \frac{f[\alpha, x^{(k-1)}, x^{(k)}]}{f[x^{(k-1)}, x^{(k)}]}(x^{(k-1)} - \alpha)(x^{(k)} - \alpha)$$

en la que se ha utilizado la notación de las diferencias divididas. Si se representa por $e^{(k)} = x^{(k)} - \alpha$ el error en cada iteración y se tiene en cuenta la relación existente entre derivadas y diferencias cocientes se puede probar que

$$e^{(k+1)} = \frac{f''(\xi^{(k)})}{2f'(\eta^{(k)})} e^{(k)} e^{(k-1)} \tag{7.4}$$

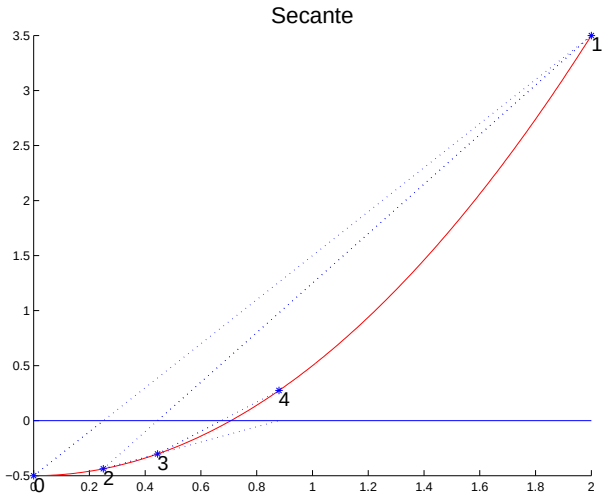


Figura 7.5: Método de la secante

para algún $\xi^{(k)}$ y algún $\eta^{(k)}$ pertenecientes al mínimo intervalo que contiene a $\alpha, x^{(k)}$ y $x^{(k-1)}$.

Sea

$$C = \frac{\max_{x\in I} |f''(x)|}{2\min_{x\in I} |f'(x)|}$$

y δ un número positivo menor que $\frac{1}{C}$, tal que $[\alpha-\delta, \alpha+\delta] \subset I$. De la relación 7.4 se deduce la siguiente desigualdad

$$C|e^{(k+1)}| \leq C|e^{(k)}|C|e^{(k-1)}| \tag{7.5}$$

para todo $k > 1$.

Si se toman $x^{(0)}$ y $x^{(1)}$ tales

$$\max\{|e^{(0)}|, |e^{(1)}|\} < \delta$$

entonces se cumple que

$$\begin{aligned}
C|e^{(0)}| &< K = C\delta < 1,\\
C|e^{(1)}| &< K,\\
C|e^{(2)}| &< K^2,\\
C|e^{(3)}| &< K^3,\\
C|e^{(4)}| &< K^5,\\
&\vdots \quad \vdots\\
C|e^{(k+1)}| &< K^{p_{k+1}},
\end{aligned}$$

donde p_k es la llamada sucesión de Fibonacci, definida por

$$p_{k+1} = p_k + p_{k-1}, \;\; p_0 = p_1 = 1.$$

Puesto que $\lim_{k\to\infty} p_k = \infty$ entonces

$$\lim_{k\to\infty} e^{(k)} = 0$$

como se quería demostrar.

Por otra parte, en un capítulo posterior se probará que el término general de la sucesión de Fibonacci está dado explícitamente por

$$p_k = \frac{1}{\sqrt{5}}\left(r_1^{k+1} + r_2^{k+1}\right)$$

siendo

$$r_1 = \frac{1+\sqrt{5}}{2}, \quad r_2 = \frac{1-\sqrt{5}}{2}.$$

De la desigualdad 7.5 se deduce que

$$\frac{|e^{(k+1)}|}{|e^{(k)}|^{r_1}} \leq C^{r_1} K^{p_k(1-r_1)} K^{p_{k-1}} = C^{r_1} K^{p_{k+1}-r_1 p_k} = K^{\frac{r_2^k}{\sqrt{5}}(r_2-r_1)}.$$

Puesto que $|r_2| < 1$ se tiene que

$$\lim_{k\to\infty} \frac{|e^{(k+1)}|}{|e^{(k)}|^{r_1}} = 0$$

lo que prueba que el orden de convergencia es r_1 (sección de oro). $\diamond$

Aunque se parta de $x^{(0)}$ y $x^{(1)}$ tales que $f(x^{(0)})f(x^{(1)}) < 0$ no está garantizado que la raíz α pertenezca al intervalo $[x^{(k-1)}, x^{(k)}]$. Una variante del método de la secante, conocida como método de *regula falsi* utiliza en cada iteración, el término $x^{(k)}$ y uno de los precedentes $x^{(k')}$ (que no es necesariamente $x^{(k-1)}$) que cumple $f(x^{(k)})f(x^{(k')}) < 0$. Es decir, la iteración es

$$x^{(k+1)} = x^{(k)} - \frac{1}{f[x^{(k')}, x^{(k)}]} f(x^{(k)}).$$

Esencialmente, el método de *regula falsi* combina aspectos del método de la secante y del método de bisección. De un modo más preciso, se describe el método del modo siguiente: Se supone que la ecuación $f(x) = 0$ tiene una

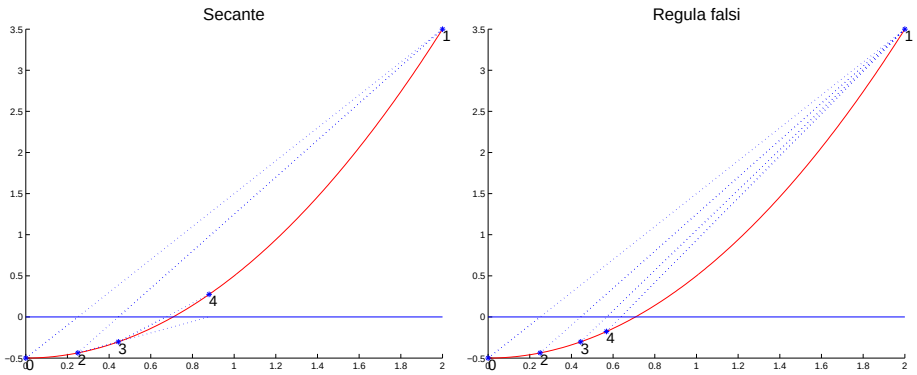


Figura 7.6: Diferente comportamiento de los métodos de la secante y *regula falsi* en la ecuación $x^2 - 0.5 = 0$

raíz en un intervalo $I = [a^{(0)}, b^{(0)}]$ y $f(a^{(0)})f(b^{(0)}) < 0$. De modo similar al método de la secante se define

$$x^{(1)} = b^{(0)} - \frac{1}{f[a^{(0)}, b^{(0)}]} f(b^{(0)}).$$

Si $x^{(1)}$ no es una raíz, se escogen $a^{(1)}$ y $b^{(1)}$ como

$$a^{(1)} = \begin{cases} a^{(0)} & \text{si } f(a^{(0)})f(x^{(1)}) < 0 \\ x^{(1)}, & \text{en otro caso} \end{cases}, \quad b^{(1)} = \begin{cases} b^{(0)} & \text{si } f(b^{(0)})f(x^{(1)}) < 0 \\ x^{(1)}, & \text{en otro caso} \end{cases}.$$

Si se repite la construcción sucesivas veces, se obtiene una sucesión $\{x^{(k)}\}$ que se aproxima a la raíz, como se justifica en el siguiente

– TEOREMA 33 *Si f es de clase $C^2(I)$ y su derivada segunda no se anula en I, entonces el método de* regula falsi *es linealmente convergente. Además, una de las sucesiones $\{a^{(k)}\}$ o $\{b^{(k)}\}$ es una sucesión constante.*

Demostración: Se supone que $f(a^{(0)}) < 0$, $f(b^{(0)}) > 0$ y $f'(x) \geq 0$ para todo $x \in I$ (como en la figura 7.6). El argumento que se utilizará en esta situación podrá ser utilizado en los otros casos separadamente.

Sea $p^{(k)}$ el polinomio que interpola a f en los extremos del intervalo $[a^{(k)}, b^{(k)}]$ y que corta a eje OX en el punto $x^{(k+1)}$. El error de interpolación en cualquier punto del intervalo está dado por

$$f(x) - p^{(k)}(x) = \frac{(x - a^{(k)})(x - b^{(k)})}{2} f''(\xi)$$

para algún punto $\xi \in (a^{(k)}, b^{(k)})$ (que depende de x). De esta relación en el punto $x^{(k)}$ se deduce

$$f(x^{(k)}) = f(x^{(k)}) - p^{(k)}(x^{(k)}) \leq 0,$$

lo que implica que $a^{(k+1)} = x^{(k)}$ y $b^{(k+1)} = b^{(k)}$. Así pues, el extremo derecho del intervalo es fijo e igual a $b^{(0)}$ y la iteración podría expresarse como

$$x^{(k+1)} = b^{(0)} - \frac{1}{f[x^{(k)}, b^{(0)}]} f(b^{(0)}) = x^{(k)} - \frac{1}{f[x^{(k)}, b^{(0)}]} f(x^{(k)}).$$

Puesto que $f(b^{(0)}) > 0$ y $f(x^{(k)}) < 0$, se tiene que $x^{(k+1)} > x^{(k)}$. Ya que la sucesión es monótona creciente y acotada es convergente a un límite α.

Si g es la función definida por

$$g(x) = b^{(0)} - \frac{x - b^{(0)}}{f(x) - f(b^{(0)})} f(b^{(0)}) = \frac{b^{(0)} f(x) - x f(b^{(0)})}{f(x) - f(b^{(0)})}$$

las iteraciones de *regula falsi* coinciden con la aproximaciones sucesivas mediante la función g. Puesto que g es continua en $(a^{(0)}, b^{(0)})$ se cumple que

$$g(\alpha) = \lim_{k \to \infty} g(x^{(k)}) = \lim_{k \to \infty} x^{(k+1)} = \alpha$$

lo que prueba que α es una raíz.

Ahora se calcula la derivada

$$g'(\alpha) = \frac{f(b^{(0)}) + f'(\alpha)(\alpha - b^{(0)})}{f(b^{(0)}))}.$$

Del teorema del valor medio se deduce la existencia de un punto $\xi \in (\alpha, b^{(0)})$ tal que

$$f(b^{(0)}) = f'(\xi)(b^{(0)} - \alpha)$$

y consecuentemente se tiene que

$$g'(\alpha) = 1 - \frac{f'(\alpha)}{f'(\xi)}.$$

Ya que f'' no es negativa, f' es monótona creciente y en consecuencia se deduce que $f'(\alpha) \leq f'(\xi)$ y $g'(\alpha) \leq 1$. $\diamond$

7.6 Método de Müller

La idea básica del método de la secante es la de resolver en cada iteración, una ecuación lineal formada por interpolación de la función en los dos últimos valores aproximados de la raíz que se han obtenido. Un paso adelante en la mejora de los algoritmos, consiste en aproximar la ecuación mediante interpolación cuadrática basada en las tres últimas aproximaciones. El método de Müller consiste en resolver en cada etapa la ecuación de segundo grado

$$f(x^{(k)})+f[x^{(k)},x^{(k-1)}](x-x^{(k)})+f[x^{(k)},x^{(k-1)},x^{(k-2)}](x-x^{(k)})(x-x^{(k-1)})=0$$

conocidos $x^{(k-2)}, x^{(k-1)}$ y $x^{(k)}$.

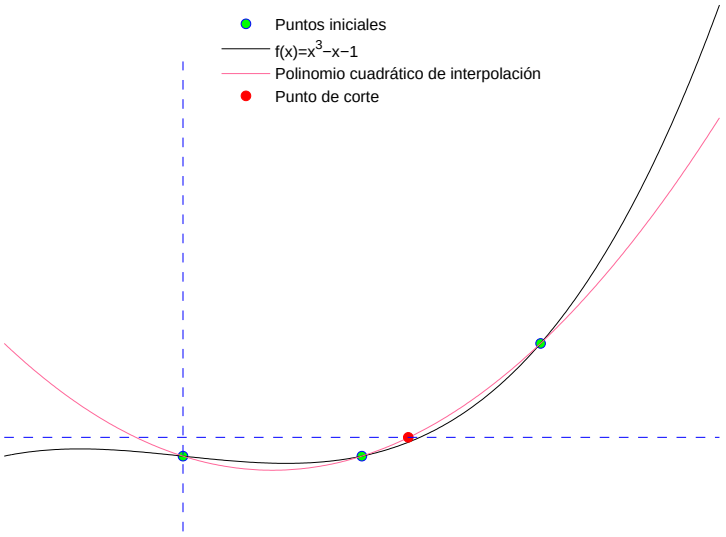


Figura 7.7: Método de Müller

Para hacer más cómoda la resolución secuencial de esta ecuación en cada etapa se introduce el número

$$\delta_k = f[x^{(k)},x^{(k-1)}] + (x^{(k)}-x^{(k-1)})f[x^{(k)},x^{(k-1)},x^{(k-2)}]$$

que depende de las tres últimas iteraciones. Se transforma la ecuación que define la iteración de Muller del modo siguiente

$$\begin{aligned} & f(x^{(k)}) + \left(f[x^{(k)},x^{(k-1)}] + f[x^{(k)},x^{(k-1)},x^{(k-2)}](x^{(k)}-x^{(k-1)})\right)(x-x^{(k)}) \\ + \quad & f[x^{(k)},x^{(k-1)},x^{(k-2)}](x-x^{(k)})^2 = 0 \end{aligned}$$

Con la notación δ_k, esta igualdad se transforma en

$$f[x^{(k)}, x^{(k-1)}, x^{(k-2)}](x - x^{(k)})^2 + \delta_k(x - x^{(k)}) + f(x^{(k)}) = 0$$

Si se resuelve esta ecuación de segundo grado y se racionaliza el numerador se obtiene

$$x = x^{(k)} - \lambda_k f(x^{(k)})$$

siendo

$$\lambda_k = \frac{2}{\delta_k \mp \sqrt{\delta_k^2 - 4f(x^{(k)})f[x^{(k)}, x^{(k-1)}, x^{(k-2)}]}}.$$

La elección del signo que haga mayor el denominador en valor absoluto, ayuda a estabilizar el cálculo.

– EJERCICIO 79 *Obtener una aproximación de la raíz real del polinomio* $p(x) = x^3 - x - 1$ *en el intervalo* $[0, 2]$*, aplicando el método de Müller y tomando como valores iniciales, los extremos del intervalo y su punto medio.*

Solución: Los cálculos para la puesta en práctica del método de Müller se integran en la siguiente tabla:

x	$f(x)$	$f[x,y]$	$f[x,y,z]$	δ
0	-1			
		0		
1	-1		3	9
		6		
2	5		4.26376262	3.98547975
		7.12462118		
1.26376262	-0.24541246		4.58550544	4.28033778
		4.01446922		
1.32174283	-0.0126527		3.91019497	
		4.25270532		
$\boxed{1.32468953}$	-0.00012124			

En la primera columna se encuentran las aproximaciones sucesivas a la solución y en la segunda los residuos de la ecuación. ◇

7.7 Método de Newton

El método de Newton (ó método de Newton-Raphson) es otro método de relajación con parámetro variable

$$x^{(k+1)} = x^{(k)} - \frac{1}{f'(x^{(k)})} f(x^{(k)}).$$

que corresponde a la elección

$$\lambda^{(k)} = \frac{1}{f'(x^{(k)})}.$$

De este modo $x^{(k+1)}$ es el punto de corte de la tangente a la gráfica de f en el punto $x^{(k)}$, con el eje OX.

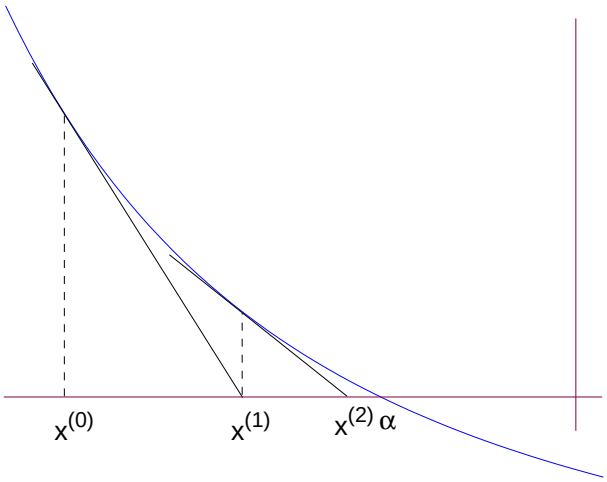


Figura 7.8: Método de Newton

El método de Newton está definido por la función

$$g(x) = x - \frac{f(x)}{f'(x)}$$

que tiene como derivada

$$g'(x) = \frac{f(x)f''(x)}{(f'(x))^2}$$

Un resultado de convergencia local es el que establece el siguiente

▬ TEOREMA 34 *Sean f una función de clase C^2 en un intervalo cerrado I que contiene una raíz α de la ecuación $f(x) = 0$. Si*

$$\frac{\max_{x\in I} |f''(x)|}{\min_{x\in I} |f'(x)|} < \infty$$

entonces la sucesión generada por el método de Newton es localmente cuadráticamente convergente a α.

Demostración: La convergencia local es consecuencia del teorema de Ostrowski. Por otra parte, puesto que $g'(\alpha) = 0$ del teorema 30 se deduce que el orden de convergencia es 2. ◇

Para garantizar la convergencia del método de Newton para todo punto inicial en un determinado intervalo I se pueden utilizar diferentes argumentos de monotonía y convexidad. Se puede reunir algunas de estas técnicas en el siguiente

▬ TEOREMA 35 *Sea f una función de clase $C^1(I)$ en el intervalo $I = [a, b]$ que es estrictamente convexa, es decir, que verifica que*

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

para todo $x, y \in I$, tales que $x \neq y$ y $0 < \lambda < 1$. Se representa por $\{x^{(k)}\}$ la sucesión generada por el método de Newton partiendo de un punto $x^{(0)} \in I$ tal que $f'(x^{(0)}) \neq 0$. Entonces, se tiene que

1. *La ecuación $f(x) = 0$ tiene a lo sumo dos raíces en I.*

2. *Si $\min_{x\in I} f(x) = f(a) < 0 \leq f(b)$, existe una única raíz α y la sucesión $\{x^{(k)}\}$ es monótona decreciente y convergente a la raíz si $x^{(0)} > \alpha$. Por el contrario, si $x^{(0)} < \alpha$ entonces $x^{(1)} > \alpha$.*

3. *Si $\min_{x\in I} f(x) = f(b) < 0 \leq f(a)$, existe una única raíz α y la sucesión $\{x^{(k)}\}$ es monótona creciente y convergente a la raíz si $x^{(0)} < \alpha$. Por el contrario, si $x^{(0)} > \alpha$ entonces $x^{(1)} < \alpha$.*

4. *Si $\min_{x\in I} f(x) < 0 \leq \min\{f(a), f(b)\}$, existe dos raíces $\alpha < \beta$ y la sucesión $\{x^{(k)}\}$ es monótona creciente y convergente a la raíz α si $x^{(0)} < \alpha$ y es monótona decreciente y convergente a la raíz β si $x^{(0)} > \beta$.*

Demostración:

1. Si $x < y < z$ son tres raíces de la ecuación, y se puede expresar como

$$y = \lambda x + (1 - \lambda)z,$$

siendo

$$\lambda = \frac{z - y}{z - x}.$$

Puesto que f es estrictamente convexa, se cumple que

$$0 = f(y) < \lambda f(x) + (1 - \lambda)f(z) = 0$$

lo cual es imposible.

2. En este caso, la función es monótona creciente estrictamente y consecuente sólo existe una raíz. Puesto que f' es positiva en I, de la igualdad

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}$$

se deduce que $x^{(k+1)} \leq x^{(k)}$ si $x^{(k)} \geq \alpha$. Además, por ser f convexa y diferenciable verifica que

$$f(y) - f(x) \geq f'(x)(y - x)$$

para todo $x, y \in [a, b]$. En particular, se tiene

$$f(x^{(k+1)}) = f\left(x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}\right) \geq f(x^{(k)}) - \frac{f(x^{(k)})}{f'(x^{(k)})} f'(x^{(k)}) = 0.$$

Consecuentemente $\{x^{(k)}\}$ es una sucesión monótona decreciente acotada por la raíz α. Esta sucesión tiene un límite que necesariamente debe coincidir con la raíz como se demuestra tomando límites en la expresión que define la iteración de Newton y usando la continuidad de f y f'. Finalmente, puesto que

$$f(\alpha) \geq f(x^{(0)}) + (\alpha - x^{(0)})f'(x^{(0)})$$

se tiene que

$$x^{(1)} = x^{(0)} - \frac{f(x^0)}{f'(x^{(0)})} \geq \alpha.$$

la recta tangente a la gráfica de f

3. El razonamiento es similar al del caso anterior.

4. Si divide el intervalo I en $[a, \eta]$ y $[\eta, b]$ siendo η el punto donde f alcanza el mínimo, y se aplican los apartados anteriores a cada uno de ellos, se obtiene el resultado buscado. $\diamond$

▪ EJEMPLO **39** Se pretende estudiar la convergencia del método de Newton cuando se aplica a la siguiente ecuación

$$f(x) \equiv 2x^3 - 4x - \operatorname{sen}(\pi x) = 0.$$

Puesto que la función es impar, $x = 0$ es una raíz y las restantes raíces positivas y negativas son simétricas entre sí. Las derivadas sucesivas de la función f son las siguientes

$$\begin{aligned} f'(x) &= 6x^2 - 4 - \pi \cos(\pi x), \\ f''(x) &= 12x + \pi^2 \operatorname{sen}(\pi x), \end{aligned}$$

La derivada segunda es positiva en $(0, \infty)$ y

$$\lim_{x \to \pm\infty} f(x) = \pm\infty.$$

Puesto que la función f es estrictamente convexa en $[0, \infty)$ tiene a lo sumo dos puntos de corte con el eje $y = 0$ y por consiguiente f tiene a lo sumo una raíz positiva β. Las tres únicas raíces de la ecuación so $\{-\beta, 0, \beta\}$.

De acuerdo con el teorema anterior si $x^{(0)} > \beta$, la sucesión generada por el método de Newton converge a β como una sucesión monótona decreciente. Por simetría, si $x^{(0)} < -\beta$, la sucesión generada por el método de Newton converge a $-\beta$ como una sucesión monótona creciente. Si x_m representa el valor positivo de x en el que se alcanza el mínimo, entonces para $x_m < x^{(0)} < \beta$ se tiene que $x^{(1)} > \beta$ y las siguientes iteraciones forman una sucesión monótona decreciente a β. Una situación similar se produce en la parte simétrica negativa.

La duda está en cómo se comporta cuando $x^{(0)}$ está en el intervalo $(-x_m, x_m)$ (si $x^{(0)} = \pm x_m$, la derivada f se anula y la iteración no está definida). En este caso, tres situaciones distintas son posibles

- $x^{(1)}$ es expulsado del intervalo $[-x_m, x_m]$,
- $\{x^{(k)}\}$ es convergente como una sucesión alternante a 0, o
- $\{x^{(k)}\}$ forma un ciclo tal que $x^{(k+1)} = -x^{(k)}$ para todo k. De hecho se puede probar que existen dos posible elecciones de $x^{(0)}$ que conducen a esta situación.

$\diamond$.

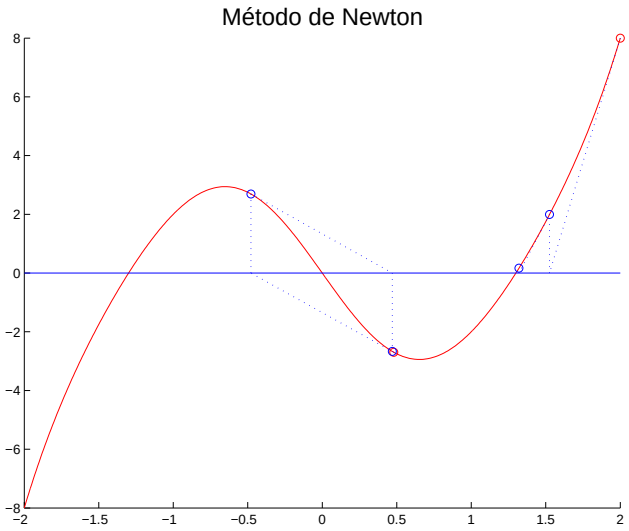


Figura 7.9: Convergencia monótona y ciclo

7.8 Método de Newton para raíces múltiples

Sea f una función de clase C^m en un intervalo que contiene una raíz α de la ecuación $f(x) = 0$. Una raíz α de la ecuación $f(x) = 0$ tiene multiplicidad m si $f'(\alpha) = \cdots = f^{m-1)}(\alpha) = 0$ y $f^{m)}(\alpha) \neq 0$. Si se usa la regla de L'Hôpital se obtiene que

$$\lim_{x\to\alpha} \frac{f(x)}{(x-\alpha)^m} = \frac{f^{m)}(\alpha)}{m!}.$$

$$\lim_{x\to\alpha} \frac{f'(x)}{(x-\alpha)^{m-1}} = \frac{m f^{m)}(\alpha)}{m!}.$$

$$\lim_{x\to\alpha} \frac{f''(x)}{(x-\alpha)^{m-2}} = \frac{m(m-1) f^{m)}(\alpha)}{m!}.$$

En consecuencia la función $g(x) = x - \frac{f(x)}{f'(x)}$ verifica

$$g'(\alpha) = \lim_{x\to\alpha} \frac{f(x)f''(x)}{(f'(x))^2} = \lim_{x\to\alpha} \frac{\frac{f(x)}{(x-\alpha)^m}\frac{f''(x)}{(x-\alpha)^{m-2}}}{\left(\frac{f'(x)}{(x-\alpha)^{m-1}}\right)^2} = 1 - \frac{1}{m}.$$

De acuerdo con el teorema 30, el orden de convergencia se reduce a 1. Si la multiplicidad de la raíz es conocida *a priori* el método de Newton puede ser modificado de modo que la iteración resultante sea

$$x^{(k+1)} = x^{(k)} - m\frac{f(x^{(k)})}{f'(x^{(k)})}$$

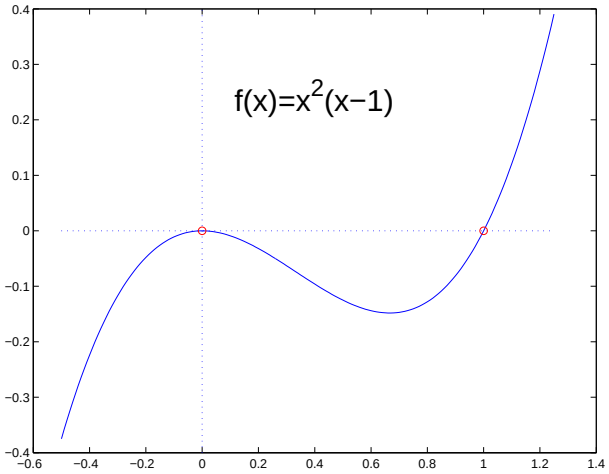


Figura 7.10: Raíces simples y dobles

para $k \geq 0$. En este caso, con el mismo razonamiento anterior, se prueba que la convergencia es cuadrática.

7.9 Raíces de ecuaciones polinómicas

Una alternativa al uso de métodos matriciales para la aproximación de los autovalores, es la resolución de la ecuación característica mediante alguno de los métodos iterativos que se han descrito en este capítulo. Así pues el estudio de ecuaciones no-lineales definidas por una función polinómica merece una especial atención.

Como se ha puesto de manifiesto en las secciones precedentes una parte importante del problema de encontrar las raíces de una ecuación es su localización previa en un intervalo en el que alguno de los métodos iterativos pueda ser utilizado. En el caso de funciones polinómicas, los siguientes resultados del Análisis Matemático Clásico puede servir de ayuda en este propósito

Teorema 36 *(Regla de los signos de Descartes) Sea r el número de cambios de signo que hay en el conjunto formado por los coeficientes de un polinomio p de grado n y k el número de raíces positivas de p. Entonces, $k \leq r$ y $r - k$ es un número par.*

Teorema 37 *(de Cauchy) Todas las raíces de un polinomio $p_n(x) = \sum_{i=0}^{n} a_i x^i$ están en el círculo de plano complejo*

$$\Gamma = \left\{ z \in \mathbb{C} : |z| < 1 + \max_{0 \leq i \leq n-1} \left| \frac{a_i}{a_n} \right| \right\}.$$

Demostración: Sea

$$\eta = \max_{0\leq i\leq n-1} \left|\frac{a_i}{a_n}\right|.$$

Para una raíz arbitraria z del polinomio se tiene que

$$\begin{aligned} |z|^n &= \left|\frac{a_{n-1}}{a_n}z^{n-1} + \cdots + \frac{a_1}{a_n}z + \frac{a_0}{a_n}\right| \\ &\leq \eta\left(|z|^{n-1} + \cdots + |z| + 1\right) = \eta\frac{|z|^n - 1}{|z| - 1}. \end{aligned}$$

Si $|z| \geq 1 + \eta$ entonces se cumple

$$|z|^n \leq |z|^n - 1$$

lo cual es imposible. ◇

▪ EJEMPLO **40** La regla de Descartes indica que a lo sumo, cero o dos de las cuatro raíces de la ecuación

$$f(x) \equiv x^4 + 2x^3 - 7x^2 + 3 = 0,$$

son positivas. Del teorema de Cauchy se deduce que la raíces positivas, si existieran, estarían en el intervalo $[0, 8]$. Además, de la tabla de signos

x	0	1	2	3	4	5	6	7	8
$signo f(x)$	+	−	+	+	+	+	+	+	+

se deduce que una raíz positiva está en $[0, 1]$ y la otra en $[1, 2]$.

Las raíces negativas de la ecuación anterior son las raíces positivas de la siguiente

$$f(-x) \equiv x^4 - 2x^3 - 7x^2 + 3 = 0.$$

Con un razonamiento similar, se deduce que existen dos raíces negativas en $[-2, -1]$ y $[-1, 0]$. ◇

La regla de Descartes no resuelve con certeza el problema de determinar el número exacto de raíces reales que tiene un polinomio de coeficientes reales. El problem fue resuelto en 1829 por Sturm que diseña un procedimiento para contar el número de raíces reales que pertenecen a un determinado intervalo, observando los signos de una determinada sucesión de polinomios en los extremos. A continuación se describe cómo se construye esa sucesión.

Sea p un polinomio de grado n p' su derivada, el algoritmo de Euclides permite calcular el polinomio máximo común divisor de p y p' mediante la siguiente secuencia de polinomios, de grado decreciente, formada con los restos de las divisiones entre dos términos consecutivos

$$\begin{aligned} p_0(x) &= p(x), \\ p_1(x) &= p'(x), \\ p_2(x) &= -\text{resto}(p_0, p_1) = p_1(x)q_0(x) - p_0(x), \\ p_3(x) &= -\text{resto}(p_1, p_2) = p_2(x)q_1(x) - p_1(x), \\ &\cdots \\ 0 &= -\text{resto}(p_{m-1}, p_m). \end{aligned}$$

El último polinomio p_m es el máximo común divisor de p y su derivada.

Si p sólo tiene raíces simples p_m es una constante. En este caso, directamente se comprueba que la sucesión de polinomios p_i tiene las siguientes propiedades: Para cualquier intervalo $I = [a, b]$,

1. p_0 no tiene raíces múltiples en $I = [a, b]$.

2. p_m no se anula en I.

3. Si α es una raíz de p_j en el intervalo I para $0 < j < m$ entonces

$$p_{j-1}(\alpha)p_{j+1}(\alpha) < 0.$$

4. Si α es una raíz de p_0 entonces $p_1(\alpha)p_0'(\alpha) > 0$.

En general, cualquier conjunto de funciones $\{f_0, f_1, \cdots, f_m\}$ que verifique estas cuatro propiedades se conoce como sucesión de Sturm asociada al polinomio p. En particular, con ayuda del algoritmo de Euclides se ha construido una sucesión de polinomios de Sturm en cualquier intervalo para un polinomio que no tenga raíces múltiples. En otro caso, basta con dividir el polinomio p por el máximo común divisor de p y su derivada p', para obtener un polinomio con las mismas raíces pero con multiplicidad 1.

El interés de la sucesiones de Sturm se desprende del siguiente

– TEOREMA 38 *Si $\{f_0, f_1, \cdots, f_m\}$ es una sucesión de Sturm para el polinomio p en $[a, b]$ y $f_0(a)f_0(b) \neq 0$ entonces el número de raíces del polinomio p contenidas en $[a, b]$ es igual a la diferencia entre el número de cambios de signos que hay en $\{f_0(a), f_1(a), \cdots, f_m(a)\}$ y $\{f_0(b), f_1(b), \cdots, f_m(b)\}$.*

▬ EJERCICIO 80 *Aplicar el método de las sucesiones de Sturm para localizar todas las raíces reales de la siguiente ecuación*

$$x^4 - 2.4x^3 + 1.03x^2 + 0.6x - 0.32 = 0.$$

Solución: Con el algoritmo de Euclides se obtiene la siguiente sucesión de Sturm

$$\begin{aligned} p_0(x) &= x^4 - 2.4x^3 + 1.03x^2 + 0.6x - 0.32, \\ p_1(x) &= 4x^3 - 7.2x^2 + 2.06x + 0.6, \\ p_2(x) &= 0.5650x^2 - 0.7590x + 0.2300, \\ p_3(x) &= 2.0220x - 1.3436, \\ p_4(x) &= 0.0249. \end{aligned}$$

Del teorema de Cauchy se deduce que las raíces reales están en el intervalo

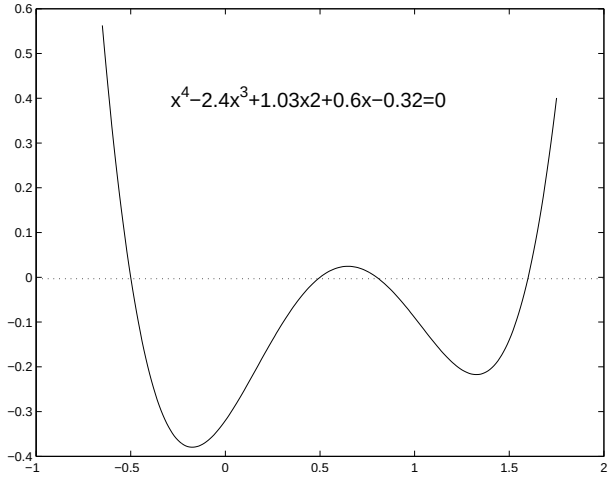


Figura 7.11: Método de Sturm

$[-4, 4]$. Los signos de la sucesión de Sturm en los valores enteros del intervalo están dados en la tabla

x	p_0	p_1	p_2	p_3	p_4	nº de cambios
-4	$+$	$-$	$+$	$-$	$+$	4
-3	$+$	$-$	$+$	$-$	$+$	4
-2	$+$	$-$	$+$	$-$	$+$	4
-1	$+$	$-$	$+$	$-$	$+$	4
0	$-$	$+$	$+$	$-$	$+$	3
1	$-$	$-$	$+$	$+$	$+$	1
2	$+$	$+$	$+$	$+$	$+$	0

Consecuentemente, el número de raíces en $[-1, 0]$ es 1, en $[0, 1]$ es 2 y en $[1, 2]$ es 1. La inspección se realiza solamente hasta $x = 2$ ya que las cuatro raíces has sido localizadas.

7.10 Ejercicios

▬ EJERCICIO 81 *Hallar por el método de bisección la menor raíz positiva de* $e^x - 3x = 0$ *calculando el error máximo de cada iteración.*

Solución: Puesto que $f(x) = e^x - 3x$ es estrictamente monótona decreciente hasta $\ln 3$ y estrictamente monótona creciente a partir de $\ln 3$,

$$\lim_{x \to \pm\infty} f(x) = \infty,$$

$f(0) > 0$ y $f(\ln 3) < 0$ entonces tiene únicamente dos raíces. La menor raíz positiva está separada en el intervalo $[0, \ln 3]$.

La sucesión de intervalos encajados generada por el algoritmo es la siguiente

n	0	1	2	3	4
$a^{(k)}$	0	$\frac{\ln 3}{2}$	$\frac{\ln 3}{2}$	$\frac{\ln 3}{2}$	$\frac{9 \ln 3}{16}$
$b^{(k)}$	$\ln 3$	$\ln 3$	$\frac{3 \ln 3}{4}$	$\frac{5 \ln 3}{8}$	$\frac{5 \ln 3}{8}$
error $<$	$\ln 3$	$\frac{\ln 3}{2}$	$\frac{\ln 3}{4}$	$\frac{\ln 3}{8}$	$\frac{\ln 3}{16}$

◇

▬ EJERCICIO 82 *En el siglo segundo, Ptolomeo construyó una tabla con los valores de la función seno para los ángulos espaciados en un grado entre* 0^o *y* 90^o*. Ptolomeo conocía los valores para* 72^o *y* 60^o *de los cálculos para pentágonos y hexágonos regulares y usando las fórmulas trigonométricas de senos de diferencia de ángulos y de ángulo mitad pudo obtener el valor del seno de gran parte de los ángulos que correspondían a valores enteros en grados. Sin embargo, Ptolomeo podía calcular* $\operatorname{sen} 3^o$, $\operatorname{sen} \frac{3}{2}^o$ *y* $\operatorname{sen} \frac{3}{4}^o$ *pero no* $\operatorname{sen} 1^o$*. Finalmente recurrió a un método de interpolación para aproximar el valor de* $\operatorname{sen} 1^o$ *y así pudo completar la tabla. En el año 1400, en la ciudad de Samarkanda, el astrónomo al-Kashi encontró un método más preciso para calcular* $\operatorname{sen} 1^o$*. Su idea era que conocido* $\operatorname{sen} 3^o$ *se podría calcular* $\operatorname{sen} 1^o$ *usando la relación trigonométrica*

$$\operatorname{sen} 3\alpha = 3 \operatorname{sen} \alpha - 4 \operatorname{sen}^3 \alpha,$$

válida para cualquier ángulo α. Aproximar por un método de punto fijo, el valor de $\operatorname{sen} 1^o = \operatorname{sen} \frac{\pi}{180}$ *usando la relación anterior y conociendo que*

$$\operatorname{sen} 3^o = \operatorname{sen} \frac{\pi}{60} \approx 0.052335956242944.$$

Solución: El valor $\operatorname{sen} 1^o$ es solución de la ecuación polinómica

$$4x^3 - 3x + a = 0$$

siendo $a = \operatorname{sen} 3^o$. Las raíces de esta ecuación son puntos fijos de la función

$$g(x) = \frac{a + 4x^3}{3}.$$

Puesto que la función seno es creciente en un entorno de 0, el valor $\operatorname{sen} 1^o$ está en el intervalo $[0, a]$ en el que se verifica que

$$|g'(x)| = |4x^2| \leq 4a^2 < 1$$

Además g es monótona creciente y por lo tanto, transforma $[0, a]$ en el intervalo $[\frac{a}{3}, \frac{a+4a^3}{3}] \subset [0, a]$. Del teorema de la aplicación contractiva se deduce que la sucesión generada por aproximaciones sucesivas con la función g es convergente a la única solución de la ecuación en $[0, a]$.

– Ejercicio 83 *¿Existe alguna raíz de la ecuación*

$$x + \tan 3x + \operatorname{sen} x = 0$$

en el intervalo $(\pi/6, \pi/2)$?. En caso afirmativo, aproximar la raíz usando algún método de aproximaciones sucesivas a un punto fijo de alguna función.

Solución: La función que define la ecuación tiene como límite $\pm\infty$ en los extremos del intervalo y es continua y monótona creciente. Se puede reducir el intervalo incrementando el extremo inferior o disminuyendo el extremo superior y comprobando que el nuevo extremo mantiene el signo. La idea es conseguir que las condiciones del teorema de la contracción se verifiquen. En este sentido, se puede aún seguir garantizando que la ecuación posee al menos una raíz en $(\pi/6, \pi/3]$. Se puede eliminar la singularidad en $\pi/6$, multiplicando la ecuación por $\cos 3x$. De este modo, la ecuación equivalente en $(\pi/6, \pi/3)$ es

$$f(x) \equiv (x + \operatorname{sen} x) \cos 3x + \operatorname{sen} 3x = 0.$$

Sea λ un número arbitrario y g la función definida por $g(x) = x + \lambda f(x)$ que tiene como punto fijo la raíz de la ecuación.

En el intervalo $[\pi/6, \pi/3]$ se cumple que

$$\begin{array}{rcl} -\dfrac{1}{2} & \leq & -\cos 2x \leq \dfrac{1}{2} \\ -1 & \leq & \cos 3x \leq 0 \\ -1 & \leq & \cos 4x \leq -\dfrac{1}{2} \\ -1 & \leq & -\operatorname{sen} 3x \leq 0. \end{array}$$

De ello se deduce que la función

$$f'(x) = 4\cos 3x - \cos 2x + 2\cos 4x - 3x \operatorname{sen} 3x$$

verifica $-\gamma_1 = -10 \leq f'(x) \leq -\gamma_2 = -\frac{1}{2}$. (una representación gráfica de la función indica que $-\gamma_1 = -7, -\gamma_2 = -3$ son cotas más finas de f'). Una estimación de g', es

$$1 - \gamma_1\lambda \leq g'(x) \leq 1 - \gamma_2\lambda$$

para todo $x \in [\pi/6, \pi/3]$.

Así pues, si se escoge $\lambda < \frac{1}{\gamma_1}$, se cumple que

$$0 \leq g'(x) < 1$$

para todo $x \in [\pi/6, \pi/3]$. Consecuentemente, la función g es monótona creciente. Además

$$g(\pi/6) = \frac{\pi}{6} + \lambda, \quad g(\pi/3) = \frac{\pi}{3} - \left(\frac{\pi}{3} + \frac{\sqrt{3}}{2}\right)\lambda.$$

y g transforma el intervalo $I = [\pi/6, \pi/3]$ en un subconjunto de I. En este caso, puesto que g verifica en I todas las condiciones del teorema de la contracción, la iteración $x^{(k+1)} = g(x^{(k)})$ para cualquier $x^{(0)} \in I$ es convergente a la única raíz que la ecuación tiene en ese intervalo. $\diamond$

– EJERCICIO 84 *Analizar la convergencia de los siguientes algoritmos iterativos:*

1. $x_{n+1} = \frac{3}{x_n}$,
2. $x_{n+1} = x_n + \frac{(x_n)^2 - 3}{2}$,
3. $x_{n+1} = \frac{(x_n)^2 + 3}{2x_n}$

hacia un punto fijo positivo, para valores iniciales x_0 positivos.

Solución: Los puntos fijos de las transformaciones g que definen los algoritmos son $\pm\sqrt{3}$.

1. No es convergente para ningún valor inicial x_0 positivo que sea distinto de $\sqrt{3}$ ya que
$$x_{n+2} = \frac{3}{x_{n+1}} = \frac{3}{\frac{3}{x_n}} = x_n.$$
Es decir, la sucesión generada zigzaguea alrededor de $\sqrt{3}$.

2. No es convergente para ningún $x_0 > 0$ distinto de $\sqrt{3}$ ya que
$$x_{n+1} = \begin{cases} < x_n, & \text{si } x_n < \sqrt{3} \\ > x_n, & \text{si } x_n > \sqrt{3} \end{cases}$$
Consecuentemente, la sucesión siempre se aleja de $\sqrt{3}$.

3. Puesto que $|g'(x)| = \frac{1}{2}\left|1 - \frac{3}{x^2}\right|$ y
$$\frac{1}{2}\left|1 - \frac{3}{x^2}\right| < 1 \quad \Leftrightarrow \quad -1 < \frac{3}{x^2} < 3$$
g es contractiva en $(1, \infty)$. Por otra parte, g tiene un mínimo $x = \sqrt{3}$ y por esta razón, g transforma $I = [\sqrt{3}, \infty)$ en un subconjunto de I. Además, si $0 < x_0 < \sqrt{3}$ entonces
$$x_1 = g(x_0) > g(\sqrt{3}) = \sqrt{3}.$$
Todo ello prueba que el algoritmo es convergente para cualquier valor inicial positivo. $\diamond$

– EJERCICIO 85 *Se considera la ecuación*

$$x = \ln(x + 2).$$

Determinar si la sucesión generada por el método de Newton-Raphson es convergente si se toma $x^{(0)} = 5$ *como valor inicial.*

Solución: La función $f(x) = x - \ln(x + 2)$ tiene como derivada segunda

$$f''(x) = \frac{1}{(x+2)^2}.$$

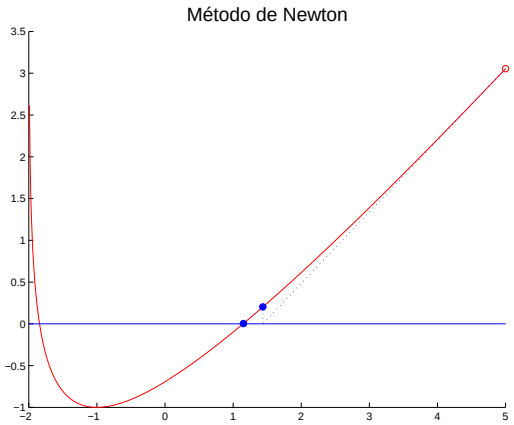


Figura 7.12: Método de Newton para la ecuación $x - \ln(x+2) = 0$

Consecuentemente, la función es estrictamente convexa en su dominio de definición $(-2, \infty)$. Puesto que

$$\lim_{x \to -2} f(x) = \infty, \quad f(-1) < 0, \quad \lim_{x \to \infty} f(x) = \infty.$$

Así pues, de acuerdo con el teorema 35 de la página 207 la ecuación tiene dos soluciones. El valor aproximado $x^{(n+1)}$ está dado por

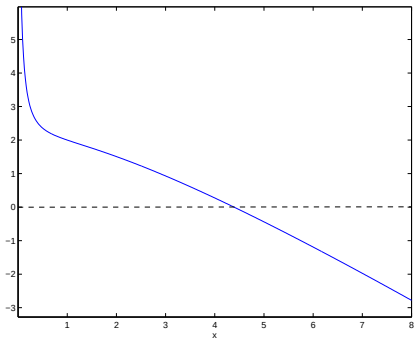
$$x^{(n+1)} = x^{(n)} - \frac{x^{(n)} - \ln(x^{(n)}+2)}{1 - \frac{1}{x^{(n)}+2}} = \frac{(x^{(n)}+2)\ln(x^{(n)}+2) - x^{(n)}}{x^{(n)}+1}.$$

Puesto que $f(5) > 0$, necesariamente la raíz máxima β es menor que 5. Consecuentemente la sucesión generada por el método de Newton con $x^{(0)} = 5$ como punto de partida, es monótona decreciente y convergente a la raíz β. $\diamond$

▬ EJERCICIO 86 *Calcular la raíz positiva de la ecuación*

$$f(x) \equiv \frac{3}{\sqrt{x}} + 2\ln x - x = 0$$

mediante un método iterativo. Realizar un análisis de la convergencia del método iterativo utilizado. En la figura de la derecha se muestra la gráfica de f.



Solución: Sin dificultad se prueba que f es monótona decreciente y

$$\lim_{x\to 0} \frac{3}{\sqrt{x}} + 2\ln x - x = \infty, \qquad \lim_{x\to\infty} \frac{3}{\sqrt{x}} + 2\ln x - x = -\infty.$$

Consecuentemente, la ecuación tiene una raíz positiva. Parece natural considerar la raíz de la ecuación como punto fijo de la función g definida por

$$g(x) = \frac{3}{\sqrt{x}} + 2\ln x$$

en el intervalo $[3, 8]$. Puesto que

$$0 < g'(x) = -\frac{3}{2\sqrt{x^3}} + \frac{2}{x} < 1$$

en $[3, 8]$, la función g es creciente y contractiva. Así pues,

$$g([3, 8]) = [g(3), g(8)] \subset [3, 8].$$

La sucesión $x^{(k)}$ que se aproxima a la raíz, definida por $x^{(k+1)} = g(x^{(k)})$ y que parte de $x^{(0)} = 3$, es

$$\{3, 3.929, 4.250, 4.349, 4.379, 4.387, 4.390, 4.390, 4.391, 4.391, 4.391, \cdots\}$$

▬ EJERCICIO 87 *Existen muchas variantes del método de Newton para la resolución de ecuaciones no-lineales escalares. Una de estas variantes es el llamado método de Newton-Halley (o de Bayley), que está definido por*

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}\left(1 - \frac{f(x^{(n)})f''(x^{(n)})}{2(f'(x^{(n)}))^2}\right)^{-1}.$$

Analizar la convergencia de este método cuando se aplica la búsqueda de raíces positivas de la ecuación $x^2 - 3 = 0$.

Solución: Si $f = x^2 - 3$ el método de Newton-Halley toma la forma

$$x^{(n+1)} = g(x^{(n)})$$

donde g es la función definida por

$$g(x) = \frac{x^3 + 9x}{3(1 + x^2)}$$

cuya derivada es

$$g'(x) = \left(\frac{x^2 - 3}{\sqrt{3}(1 + x^2)}\right)^2.$$

Inmediatamente se comprueba que g es contractiva en el conjunto

$$\{x : x > \sqrt{2\sqrt{3}-3}\}$$

y es monótona creciente. Consecuentemente, g verifica las hipótesis del teorema de la aplicación contractiva en el intervalo $[1, \infty)$. $\diamond$

▬ EJERCICIO 88 *El polinomio*

$$p(x) = x^5 + x + 1$$

tiene una raíz en el intervalo $[-1, 0]$. *Usar el punto medio de este intervalo como punto inicial y aplicar el método de Newton. ¿Es la sucesión que así se genera, convergente a la raíz?. Justificar rigurosamente la respuesta.*

Solución: Puesto que

$$\begin{aligned} p'(x) &= 5x^4 + 1 > 0, \\ p''(x) &= 20x^3 \leq 0 \end{aligned}$$

para todo $x \in [-1, 0]$, el polinomio es creciente y cóncavo en el intervalo $[-1, 0]$. La iteración que corresponde a la aplicación del método de Newton es en este caso, la siguiente

$$x^{(k+1)} = x^{(k)} - \frac{p(x^{(k)})}{p'(x^{(k)})} = g(x^{(k)})$$

siendo

$$g(x) = \frac{4x^5 - 1}{5x^4 + 1}.$$

Si se considera como valor inicial $x^{(0)} = 0.5$, se tiene que

$$x^{(1)} = g(-0.5) = -\frac{20}{21}.$$

Puesto en este punto $p(x^{(1)}) < 0$, se tiene que $x^{(2)} > x^{(1)}$. Además para toda función cóncava diferenciable se verifica que

$$p(y) - p(x) \leq p'(x)(y - x)$$

para todo x, y, en particular se tiene que

$$p(x^{(2)}) \leq p(x^{(1)}) + p'(x^{(1)})(x^{(2)} - x^{(1)}) = p(x^{(1)}) - p(x^{(1)}) = 0.$$

En consecuencia, $x^{(2)} \leq \alpha$ por ser p monótona creciente. Razonando de modo recurrente, se deduce que la sucesión generada por el método de Newton es monótona creciente y acotada. Ello implica que es convergente y como g es continua, necesariamente lo es a una raíz de la ecuación.

Por otra parte, los primeros términos de la sucesión generada por el método de Newton son

$$\{-0.6667, -0.7681, -0.7552, -0.7549, -0.7549, \cdots .\} \qquad \diamond$$

CAPÍTULO 8

Resolución de sistemas de ecuaciones no lineales

8.1 Introducción

Resolver problemas de optimización de funciones tiene un gran interés ya que numerosos problemas científicos se formulan de esta manera. Esencialmente hay dos niveles de dificultad en los problemas de optimización. En el primer nivel se sitúan los problemas de optimización diferenciable sin restricciones. El cálculo diferencial en varias variables permite transformar esta clase de problemas en la resolución de un sistema de ecuaciones numéricas, que posiblemente no sean lineales. En el segundo nivel, se consideran restricciones en la optimización ó se elimina la regularidad de las funciones objetivo de la optimización. Este segundo nivel se escapa de las intenciones de este curso y no será tratado en este texto.

Las primeras secciones de este capítulo están dedicadas a la resolución de sistemas de ecuaciones numéricas que no son necesariamente lineales. En las siguientes secciones se considera la situación particular en la que el sistema de ecuaciones está definido mediante el gradiente de la función objetivo en un problema de optimización.

Sea F una función de n variables, que tiene n componentes, definida en un conjunto $\Omega \subset \mathbb{R}^n$. Se considera la ecuación $F(\mathbf{x}) = \mathbf{0}$ asociada a esta función y se intenta encontrar sus soluciones. Los métodos numéricos que se pueden emplear para resolver esta ecuación no son esencialmente muy distintos a los empleados para resolver ecuaciones no-lineales escalares aunque pueden

presentar algunas dificultades adicionales debido al incremento de dimensión. Si las derivadas de la función F son fácilmente evaluables, de nuevo, el método de Newton emerge como el más eficiente. En los problemas de optimización, F aparece como el gradiente de una función escalar f.

8.2 Métodos de punto fijo

Como en el caso escalar, se pueden reformular las soluciones de sistemas de n ecuaciones de n incógnitas como puntos fijos de funciones G de $\mathbb{R}^n$ en $\mathbb{R}^n$. De este modo se puede considerar la posibilidad de utilizar el método de aproximaciones sucesivas asociado a esta aplicación

$$\mathbf{x}^{(k+1)} = G(\mathbf{x}^{(k)})$$

partiendo de un $\mathbf{x}^{(0)}$ dado. La convergencia de esta clase de métodos se apoya frecuentemente en la siguiente versión n-dimensional del teorema de Banach

Teorema 39 *Sea G una aplicación contractiva en un conjunto cerrado $B \subset \mathbb{R}^n$, es decir, una función para la que existe una constante $0 \leq \rho < 1$ tal que*

$$\|G(\mathbf{x}) - G(\mathbf{y})\| \leq \rho \|\mathbf{x} - \mathbf{y}\|$$

para todo $\mathbf{x}, \mathbf{y} \in B$. Si $G(B) \subset B$, entonces existe un único punto fijo α de G en B.

En el caso escalar, el modo más simple de establecer que la función que define las aproximaciones sucesivas es una contracción es el uso del teorema de valor medio. En dimensión n, no es posible disponer de equivalente directo de este teorema pero si se puede bajar al caso de una variable, considerando el segmento que une dos puntos. Para ello se necesita que B sea un conjunto convexo, es decir, que para todos los puntos $\mathbf{x}, \mathbf{y} \in B$, el segmento que los une, esté en B. De un modo más preciso, si G es una función continuamente diferenciable en el conjunto cerrado convexo B y se aplica el teorema del valor medio a la función

$$g(t) = G(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$$

en el intervalo $[0, 1]$, se obtiene que

$$g(1) - g(0) = G(\mathbf{y}) - G(\mathbf{x}) = g'(\xi)$$

para algún $\xi \in [0, 1]$. Si se aplica la regla de la cadena, se obtiene que

$$g'(\xi) = G'(\mathbf{x} + \xi(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})$$

donde $G'(\mathbf{z})$ representa la matriz jacobiana de G el punto $\mathbf{z}$. Si se usa una norma matricial subordinada a una norma en $\mathbb{R}^n$, se deduce que

$$\|G(\mathbf{y}) - G(\mathbf{x})\| = \|G'(\mathbf{x} + \xi(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})\| \leq \max_{\mathbf{z}\in\Omega} \|G'(\mathbf{z})\|\|\mathbf{x} - \mathbf{y}\|$$

para todo $\mathbf{x}, \mathbf{y} \in B$. Consecuentemente, si

$$\rho = \max_{\mathbf{z}\in B} \|G'(\mathbf{z})\| < 1,$$

G es una aplicación contractiva de constante ρ en B.

Por otra parte, puesto que el radio espectral $\rho(\mathrm{A})$ de una matriz A verifica que

$$\rho(\mathrm{A}) = \inf\{\|\mathrm{A}\| : \| \ \| \text{ es una norma subordinada}\}$$

entonces, si

$$\rho = \max_{\mathbf{z}\in B} \rho(G'(\mathbf{z})) < 1,$$

G es una contracción en B en alguna norma subordinada.

El modo más simple de construir una aplicación G que tenga como puntos fijos las raíces de una función F, es usar una técnica de relajación como se ilustra con el siguiente:

▬ Ejercicio 89 *Usar un método de punto fijo para aproximar la intersección de las siguientes curvas*

$$\begin{aligned} x^2 - y &= 1, \\ x + (y-1)^2 &= 6. \end{aligned}$$

Solución: Los puntos buscados tienen coordenadas (x, y) que son soluciones de la siguiente ecuación

$$F(x, y) \equiv (x^2 - y - 1, x + (y-1)^2 - 6) = (0, 0).$$

También son puntos fijos de la función $G(x, y) = (x, y) - \lambda F(x, y)$ para cualquier valor de λ, distinto de cero.

La matriz jacobiana de G en un punto (x, y) es la siguiente

$$G'(x, y) = \mathrm{I} - \lambda \begin{pmatrix} 2x & -1 \\ 1 & 2(y-1) \end{pmatrix}.$$

Si se usa la norma $\| \ \|_1$, se tiene que

$$\|G'(x, y)\|_1 = \max\{|1 - 2\lambda x| + |\lambda|, |\lambda| + |1 - 2\lambda(y-1)|\}.$$

Si se toma $\lambda = \frac{1}{4}$, se obtiene que en el cuadrado

$$B = [1,3] \times [2,4]$$

se cumple que

$$\max_{(x,y)\in B} \|G'(x,y)\|_1 = \frac{1}{4} + \frac{1}{2} \max \left\{ \max_{1\le x\le 3} |2-x|, \max_{2\le y\le 4} |3-y| \right\} = \frac{3}{4}.$$

Así pues, en este cuadrado la función G es contractiva con constante $\rho = \frac{3}{4}$. Además, directamente se prueba que G transforma el dominio convexo cerrado $B = [1,3] \times [2,4]$ en un subconjunto de B. Del teorema de la contracción se deduce que existe una única solución en B que puede ser aproximada por iteraciones sucesivas con la función G. $\diamond$

El teorema de Ostrowski introducido en el capítulo anterior, admite la siguiente generalización al caso de dimensión n

▬ TEOREMA 40 (OSTROWSKI) *Si G es una función de clase C^1 en un entorno de un punto fijo $\boldsymbol{\alpha}$ y verifica que*

$$\rho_{G'(\boldsymbol{\alpha})} < 1,$$

entonces la sucesión generada por aproximaciones sucesivas con esta función, es localmente convergente a $\boldsymbol{\alpha}$.

Es importante resaltar que el teorema de Ostrowski asegura la convergencia local mientras que el teorema de la aplicación contractiva garantiza la global. Esto quiere decir que con el apoyo del teorema de Ostrowski no podremos tener la garantía de que el punto inicial escogido es adecuado ya que no podemos precisar si se está suficientemente cerca. No obstante, si partiendo de un punto la iteración converge y G es continua, entonces lo hace a un punto fijo de G. Otra cuestión es cómo garantizar la condición de Ostrowski en un punto que no conocemos y queremos calcular. Obviamente, tendremos que probar que se cumple la condición en un entorno en el que tengamos localizada una raíz.

▪ EJEMPLO **41** Se considera la transformación no-lineal G definida por

$$\begin{aligned} x &= x^2 - 3y^2 + 3, \\ y &= 2x^3 + 3y^2 - 4. \end{aligned}$$

La matriz jacobiana de

$$G(x,y) = (x^2 - 3y^2 + 3, 2x^3 + 3y^2 - 4)^t$$

en el punto fijo $(1,1)$, está dada por

$$G'(1,1) = \begin{pmatrix} 2 & -6 \\ 6 & 6 \end{pmatrix}$$

cuyos autovalores son $4(1 \pm \sqrt{2}i)$. El radio espectral de la matriz jacobiana en $(1,1)$ es $4\sqrt{3}$. Puesto que es mayor que 1, el teorema de Ostrowski no puede aplicarse en el punto fijo $(1,1)$. ◇

8.3 Método de Newton

Sea F una función diferenciable en un dominio Ω y $\mathbf{x}^{(0)} \in \Omega$ un punto arbitrario. La idea básica del método de Newton es la linealización de la ecuación $F(\mathbf{x}) = 0$ mediante el desarrollo de Taylor de orden 1 alrededor de $\mathbf{x}^{(0)}$

$$F(\mathbf{x}^{(0)}) + F'(\mathbf{x}^{(0)})(\mathbf{x} - \mathbf{x}^{(0)}) = \mathbf{0}.$$

La solución $\mathbf{x}^{(1)}$ de este sistema lineal se utiliza para una nueva linealización. De este modo, conocido $\mathbf{x}^{(k)}$ se calcula $\mathbf{x}^{(k+1)}$ como la solución del sistema lineal

$$F'(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)}) = -F(\mathbf{x}^{(k)}).$$

▪ Ejemplo **42** Se pretende resolver el sistema de ecuaciones

$$\begin{aligned} F_1(x,y) \equiv x - y + x^2y^2 - 1 &= 0, \\ F_2(x,y) \equiv y + x^2 + 2xy^2 - 2 &= 0. \end{aligned}$$

Este sistema de ecuaciones no lineales puede ser resuelto por aproximaciones sucesivas generadas por

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

o por el método de Newton

$$\begin{pmatrix} 1 + 2x^{(k)}(y^{(k)})^2 & -1 + 2(x^{(k)})^2y^{(k)} \\ 2x^{(k)} + 2(y^{(k)})^2 & 1 + 4x^{(k)}y^{(k)} \end{pmatrix} \begin{pmatrix} x^{(k+1)} - x^{(k)} \\ y^{(k+1)} - y^{(k)} \end{pmatrix} = - \begin{pmatrix} F_1(x^{(k)}, y^{(k)}) \\ F_2(x^{(k)}, y^{(k)}) \end{pmatrix}.$$

Aparentemente, el número de iteraciones que requiere el método de Newton para alcanzar la misma precisión es menor que las que requiere el método de punto fijo escogido. ◇

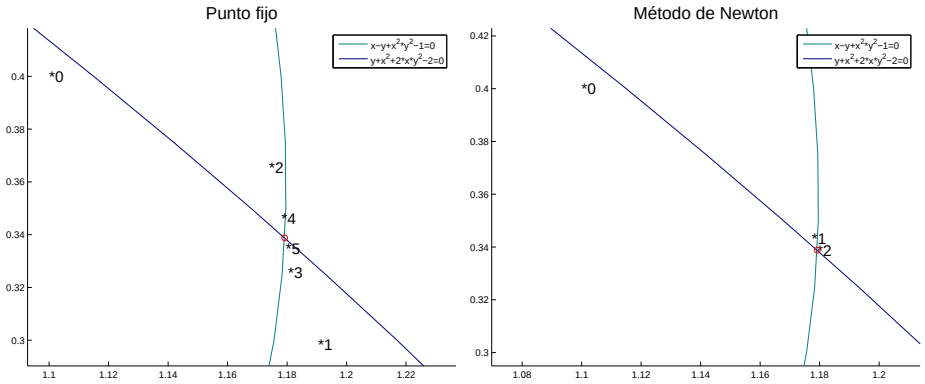


Figura 8.1: Método de Newton

Cuando el número de incógnitas es elevado, la resolución del sistema lineal requiere un esfuerzo computacional adicional. Existen numerosas modificaciones del método de Newton orientadas a simplificar la resolución del sistema lineal. Algunas de ellas pasan por considerar la descomposición de Jacobi de la matriz jacobiana

$$F'(\mathbf{x}^{(k)}) = \mathrm{D}^{(k)} - \mathrm{L}^{(k)} - \mathrm{U}^{(k)}$$

que usa la parte diagonal, la parte inferior y la superior a la diagonal de la matriz. La idea es utilizar un método de Newton inexacto que considere la iteración de Jacobi con el sistema diagonal

$$\mathrm{D}^{(k)}(\mathbf{x} - \mathbf{x}^{(k)}) = -F(\mathbf{x}^{(k)})$$

ó la iteración de Gauss-Seidel con el sistema triangular inferior

$$(\mathrm{D}^{(k)} - \mathrm{L}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)}) = -F(\mathbf{x}^{(k)}).$$

▁ EJERCICIO 90 *Describir la primera iteración por el método de Newton inexacto con la iteración de Gauss-Seidel para resolver el siguiente sistema de ecuaciones*

$$x = 0.5 \operatorname{sen} x + 0.2 \cos y \tag{8.1}$$
$$y = 0.5 \cos x - 0.2 \operatorname{sen} y \tag{8.2}$$

partiendo de $x^{(0)} = y^{(0)} = \frac{\pi}{6}$.

Solución: La matriz jacobiana de la función

$$F(x, y) = (x - 0.5 \operatorname{sen} x - 0.2 \cos y, y - 0.5 \cos x + 0.2 \operatorname{sen} y)$$

es

$$F'(x,y) = \begin{pmatrix} 1 - 0.5\cos x & 0.2\,\mathrm{sen}\, y \\ 0.5\,\mathrm{sen}\, x & 1 + 0.2\cos y \end{pmatrix}$$

El paso de $(x^{(0)}, y^{(0)})$ a $(x^{(1)}, y^{(1)})$ se realiza resolviendo el sistema lineal triangular inferior

$$\begin{pmatrix} 1 - 0.5\cos x^{(0)} & 0 \\ 0.5\,\mathrm{sen}\, x^{(0)} & 1 + 0.2\cos y^{(0)} \end{pmatrix} \begin{pmatrix} x^{(1)} - x^{(0)} \\ y^{(1)} - y^{(0)} \end{pmatrix} = - \begin{pmatrix} F_1(x^{(0)}, y^{(0)}) \\ F_2(x^{(0)}, y^{(0)}) \end{pmatrix}.$$

Si se resuelve el sistema, se obtiene $(x^{(1)}, y^{(1)}) = (0.3465, 0.3989)$. $\diamond$

8.4 Método de Broyden

En aquellas ocasiones en las que el cálculo de las derivadas parciales de la función F encierra muchas dificultades, es recomendable recurrir a métodos como el de Broyden, que evitan su cálculo, siguiendo una idea que, en cierto modo, generaliza el método de la secante para ecuaciones de una sola variable.

El método de Broyden involucra en cada iteración a tres términos $\mathbf{x}^{(k-1)}$, $\mathbf{x}^{(k)}$ y $\mathbf{x}^{(k+1)}$ y no utiliza exactamente la matriz jacobiana de F. Si en una iteración se ha calculado $\mathbf{x}^{(k)}$ mediante el sistema lineal

$$J_{k-1}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) = -F(\mathbf{x}^{(k-1)})$$

se trata ahora de construir una matriz J_k para la siguiente iteración y que conserve alguna similitud con la matriz jacobiana. En este sentido, parece razonable imponer a J_k la siguiente condición

$$J_k(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) = F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-1)}). \tag{8.3}$$

No obstante, con esta única condición es imposible determinar J_k por lo que es preciso añadir otras condiciones adicionales. En el método de Broyden, se impone la condición de que J_k se comporte como J_{k-1} en el espacio ortogonal a $\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$, es decir, que se cumpla

$$(J_k - J_{k-1})\mathbf{z} = \mathbf{0}, \quad \text{para todo } \mathbf{z} \in \mathbb{R}^n \quad \text{tal que } \mathbf{z} \cdot (\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) = 0.$$

Se considera una base ortogonal formada por $\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$ y $n-1$ vectores en el subespacio $\{\mathbf{z} : \mathbf{z} \cdot (\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) = 0\}$. Para todo $\mathbf{x} \in \mathbb{R}^n$ se cumple que

$$(J_k - J_{k-1})\mathbf{x} = (J_k - J_{k-1}) \left(\frac{\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}}{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|} \cdot \mathbf{x} \right) \frac{\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}}{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}.$$

Si se utiliza la condición 8.3 y se elimina $\mathbf{x}$, se obtiene la fórmula

$$J_k = J_{k-1} + \frac{(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-1)}) - J_{k-1}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}))(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})^t}{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2}.$$

y $\mathbf{x}^{(k+1)}$ se define como la única solución del sistema lineal

$$J_k(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) = -F(\mathbf{x}^{(k)}).$$

Una elección adecuada de matriz inicial puede ser $J_0 = F'(\mathbf{x}^{(0)})$. Por otra parte, puesto que la matriz del sistema lineal de ecuaciones que se resuelve en cada iteración solamente se modifica en una dirección, resulta muy adecuada la utilización de la fórmula de Sherman-Morrison para su inversión que establece que si una matriz A no es singular y $\mathbf{u}$ y $\mathbf{v}$ son vectores tales $1 + \mathbf{v}^t \mathrm{A}^{-1}\mathbf{u} \neq 0$, entonces $\mathrm{A} + \mathbf{u}\mathbf{v}^t$ no es singular y

$$(\mathrm{A} + \mathbf{u}\mathbf{v}^t)^{-1} = \mathrm{A}^{-1} - \frac{\mathrm{A}^{-1}\mathbf{u}(\mathrm{A}^{-1}\mathbf{v})^t}{1 + \mathbf{v}^t\mathrm{A}^{-1}\mathbf{u}}.$$

Si se escoge en el método de Broyden

$$\begin{aligned} \mathbf{u}^{(k)} &= \frac{F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-1)}) - J_{k-1}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})}{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}, \\ \mathbf{v}^{(k)} &= \frac{\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}}{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|} \end{aligned}$$

y $A = J_{k-1}$ se obtiene que

$$J_k^{-1} = J_{k-1}^{-1} - \frac{J_{k-1}^{-1}\mathbf{u}^{(k)}\left(J_{k-1}^{-1}\mathbf{v}^{(k)}\right)^t}{1 + \mathbf{v}^{(k)} \cdot J_{k-1}^{-1}\mathbf{u}^{(k)}}.$$

– EJERCICIO 91 *Aproximar una solución del sistema de ecuaciones*

$$\begin{aligned} x^2 + y^2 &= 1, \\ x + y &= 1, \end{aligned}$$

usando el método de Broyden con punto de partida $(x^{(0)}, y^{(0)}) = (1, \frac{1}{2})$ *(basta con calcular la primera iteración).*

Solución: La matriz jacobiana de

$$F(x,y) = (x^2 + y^2 - 1, x + y - 1)^t$$

es

$$F'(x,y) = \begin{pmatrix} 2x & 2y \\ 1 & 1 \end{pmatrix}$$

y en particular

$$J_0^{-1} = F'(1,1/2)^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$$

De este modo se obtiene

$$\begin{pmatrix} x^{(1)} \\ y^{(1)} \end{pmatrix} = \begin{pmatrix} x^{(0)} \\ y^{(0)} \end{pmatrix} - J_0^{-1} \begin{pmatrix} \left(x^{(0)}\right)^2 + \left(y^{(0)}\right)^2 - 1 \\ x^{(0)} + y^{(0)} - 1 \end{pmatrix} = \begin{pmatrix} \frac{5}{4} \\ -\frac{1}{4} \end{pmatrix}.$$

A continuación se calculan los vectores de error

$$\begin{aligned} \mathbf{v}^{(1)} &= \frac{1}{\sqrt{10}} \begin{pmatrix} 1 \\ -3 \end{pmatrix}, \\ \mathbf{u}^{(1)} &= \frac{5}{2\sqrt{10}} \begin{pmatrix} -1 \\ 2 \end{pmatrix} \end{aligned}$$

y finalmente, se obtiene

$$J_1^{-1} = \begin{pmatrix} -\frac{1}{7} & \frac{1}{2} \\ \frac{3}{7} & -\frac{1}{2} \end{pmatrix}. \qquad \diamond$$

8.5 Raíces complejas de un polinomio

Los métodos que se han utilizado para calcular las raíces de un polinomio, permitían aproximar las raíces reales y podían ser puestos en práctica en un computador con aritmética real. Se pueden utilizar las mismas ideas para aproximar las raíces complejas en un entorno de cálculo que use una aritmética compleja. Por otra parte, un polinomio de coeficientes reales puede tener raíces complejas que pueden ser clasificadas en pares de raíces conjugadas. El método de Bairstow evita utilizar aritmética compleja en el cálculo de las raíces de un polinomio de coeficientes reales, calculando cada uno de los polinomios reales de segundo grado que tienen cada par de raíces conjugadas.

Sea $p = \sum_{i=0}^{n} a_i x^i$ un polinomio de grado n. Si se divide p por un polinomio arbitrario $x^2 + rx + s$ se obtiene un polinomio cociente q y un resto $A(r,s)x + B(r,s)$ de modo que

$$p(x) = q(x)(x^2 + rx + s) + A(r,s)x + B(r,s). \tag{8.4}$$

Si se pretende que $x^2 + rx + s$ tenga como raíces, dos de las del polinomio p, los parámetros r y s deben verificar las siguientes ecuaciones no lineales

$$A(r, s) = 0,$$

$$B(r, s) = 0.$$

Si se expresa $q(x) = \sum_{i=0}^{n-2} b_i x^i$, identificando coeficientes en la ecuación 8.4 se obtiene que

$$b_i = a_{i+2} - rb_{i+1} - sb_{i+2} \tag{8.5}$$

para $i = n-2, \ldots, 0$. Además

$$A(r, s) = a_1 - rb_0(r, s) - sb_1(r, s), \tag{8.6}$$
$$B(r, s) = a_0 - sb_0(r, s). \tag{8.7}$$

Se puede utilizar el método de Newton para resolver este sistema de dos ecuaciones y dos incógnitas. Es importante notar que el cálculo de la matriz jacobiana de la función $F(r, s) = (A(r, s), B(r, s))^t$ implica el cálculo de las derivadas parciales de b_0 y b_1, lo que puede hacerse con ayuda de las relaciones 8.5.

▪ Ejemplo **43** Para hallar las raíces complejas del polinomio

$$p(x) = x^4 + 3x^2 - 4$$

usando el método de Bairstow, se consideran las relaciones de recurrencia 8.5 que en este caso se convierten en

$$\begin{aligned} b_2 &= a_4 - rb_3 - sb_4 = 1, \\ b_1 &= a_3 - rb_2 - sb_3 = -r, \\ b_0 &= a_2 - rb_1 - sb_2 = 3 + r^2 - s. \end{aligned}$$

De 8.6 se deduce que las ecuaciones no lineales a resolver son

$$\begin{aligned} A(r, s) &= -r(3 + r^2 - s) + sr = 0, \\ B(r, s) &= -4 - s(3 + r^2 - s) = 0. \end{aligned}$$

En este caso tan simple, este sistema tiene como soluciones $r = 0$ y $s = \{-1, 4\}$, lo que conduce a la descomposición del polinomio

$$p(x) = x^4 + 3x^2 - 4 = (x^2 + 4)(x^2 - 1). \qquad \diamond$$

8.6 Optimización sin restricciones

Sea $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ una función diferenciable en el conjunto abierto Ω. Se considera el problema que consiste en minimizar f en Ω. Si f alcanza el mínimo en un punto crítico $x \in \Omega$, verifica la ecuación no-lineal

$$\nabla f(\mathbf{x}) = 0,$$

donde el operador ∇ representa el operador gradiente. De este modo, la resolución de un problema de optimización sin restricciones conduce a la resolución de un sistema de ecuaciones numéricas no-lineales. Es necesario precisar que no todo punto crítico de la función objetivo es necesariamente un punto que optimice la función. Si existen más de un punto crítico sería necesario separar previamente el que realice el mínimo. En ocasiones, algunas condiciones sobre la función objetivo, tales como la convexidad estricta, permiten garantizar la unicidad de solución al problema.

Los métodos numéricos desarrollados en las secciones precedentes, son aplicables en esta situación. No obstante, en esta sección, se analizan algunos métodos que están especialmente diseñados para este tipo particular de sistemas no lineales. También, se puede hacer un planteamiento inverso asociando a un sistema no-lineal un problema de optimización e intentando aplicar sobre este problema las técnicas numéricas específicas de optimización. No obstante, hay que tener en cuenta en este caso, que no toda función no-lineal es un gradiente de una función escalar y consecuentemente, que de modo general este procedimiento no es válido.

Si la función objetivo f es una función cuadrática, es decir, es una función definida de la siguiente forma

$$f(\mathbf{x}) = \frac{1}{2}\mathrm{A}\mathbf{x} \cdot \mathbf{x} - \mathbf{b} \cdot \mathbf{x}$$

donde A representa una matriz $n \times n$, simétrica y definida positiva, entonces, $\mathbf{x}$ realiza el mínimo de esta función en $\mathbb{R}^n$ si y sólo si $\mathbf{x}$ es el vector solución de la ecuación $\mathrm{A}\mathbf{x} = \mathbf{b}$.

▪ Ejemplo **44** Se considera el siguiente sistema de ecuaciones lineales

$$x_{i+1} - 2x_i + x_{i-1} = 1, \quad \text{para } i = 1, 2, 3, 4,$$

$$x_0 = x_5 = 0.$$

Si

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix}.$$

$\mathrm{A}\mathbf{x} = \mathbf{b}$ es la forma matricial de expresar el sistema lineal. Este sistema de ecuaciones es equivalente a la resolución del problema de minimización sin restricciones

$$\min_{\mathbf{x}\in R^4} \frac{1}{2}\mathrm{A}\mathbf{x}\cdot\mathbf{x} - \mathbf{b}\cdot\mathbf{x} = \min_{\mathbf{x}\in R^4} \sum_{i=1}^{4}(x_i^2 - x_i x_{i+1} + x_i)$$

con $x_0 = x_5 = 0$. $\diamond$

Para resolver los problemas de minimización sin restricciones, los métodos de descenso utilizan iteraciones para aproximar el punto óptimo, definidas del siguiente modo: Dado un vector inicial $\mathbf{x}^{(0)}$, calculada la iteración $k-$ésima, se introduce

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \rho_k \mathbf{d}^{(k)}$$

donde $\mathbf{d}^{(k)}$ es un vector de dirección convenientemente escogido y ρ_k un escalar llamado tamaño de paso. Razonablemente, la dirección $\mathbf{d}^{(k)}$ debe ser escogido de modo

$$f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}).$$

En este sentido, puesto la dirección en el que el decrecimiento de la función objetivo, medido a través de su derivada direccional, es más fuerte es la opuesta a la del gradiente de la función, parece adecuado escoger direcciones que formen ángulo con $-\nabla f(\mathbf{x}^{(k)})$ menor de $\frac{\pi}{2}$ radianes. De modo más preciso, se definen las direcciones de descenso $\mathbf{d}^{(k)}$ como aquellas direcciones no nulas que verifiquen que

$$\mathbf{d}^{(k)} \cdot \nabla f(\mathbf{x}^{(k)}) < 0.$$

Sea g la función escalar de variable t definida por

$$g(t) = f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)}).$$

Si f es continuamente diferenciable y $\mathbf{d}^{(k)}$ es una dirección de descenso en el punto $\mathbf{x}^{(k)}$, la función g es estrictamente decreciente en un entorno de $\mathbf{0}$ ya que

$$g'(0) = \nabla f(\mathbf{x}^{(k)}) \cdot \mathbf{d}^{(k)} < 0.$$

Es decir, existe un ρ_k suficientemente pequeño tal que

$$f(\mathbf{x}^{(k)} + \rho_k \mathbf{d}^{(k)}) < f(\mathbf{x}^{(k)}).$$

Un modo adecuado de calcular ρ_k es resolviendo el problema en una variable:

$$\rho_k = \text{Argumento mínimo}_{\ t>0} \quad f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)})$$

o equivalentemente, la ecuación escalar no-lineal

$$\nabla f(\mathbf{x}^{(k)} + t\mathbf{d}^{(k)}) \cdot \mathbf{d}^{(k)} = 0.$$

La estrategia de elección de dirección de descenso más básica corresponde a la siguiente

$$\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})$$

y se conoce como método del gradiente. Si se utiliza el anteriormente mencionado método de elección del paso, el método se dice de máximo descenso (steepest descent).

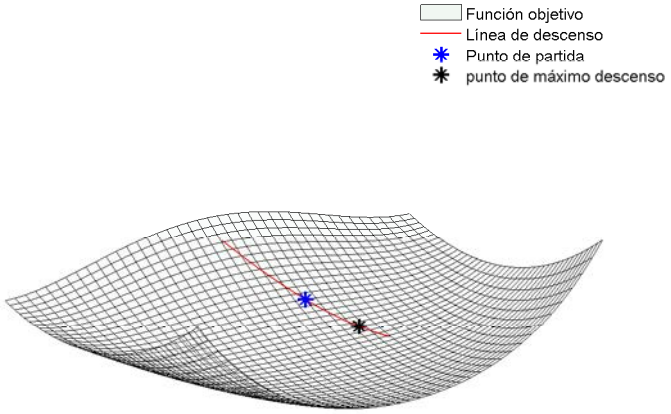


Figura 8.2: Método de máximo descenso

▬ Ejercicio 92 *Describir la primera iteración del método del gradiente con máximo descenso para encontrar el mínimo de la función*

$$f(x, y) = 100(y - x^2)^2 + (1 - x)^2$$

que se conoce como la función banana (ó valle parabólico de Rosenbrock), partiendo del punto $(0.5, 0.25)$.

Solución: En este caso, la dirección de descenso está dada por

$$\mathbf{d}^{(0)} = -\nabla f(x^{(0)}, y^{(0)}) = (1, 0)^t.$$

Para encontrar el tamaño del paso ρ_0, se busca el mínimo de la función

$$g(t) = f(x^{(0)}+td_1^{(0)}, y^{(0)}+td_2^{(0)}) = f(0.5+t, 0.25) = 100t^2(1+t)^2+\left(t-\frac{1}{2}\right)^2.$$

Puesto que

$$g'(t) = 400t^3 + 600t^2 + 202t - 1$$

el valor del paso de máximo descenso $\rho_0 = 0.0049$ puede ser aproximado por un método de resolución de ecuaciones en una variable. Finalmente se obtiene que $(x^{(1)}, y^{(1)}) = (0.5049, 0.25)$. $\diamond$

El método de Newton para la optimización sin restricciones se basa en la aproximación de una función f de clase C^2 por su desarrollo en serie de Taylor de orden 2 en un punto próximo a una solución. De este modo, dado un aproximado punto $\mathbf{x}^{(0)}$ aproximado a la solución, el problema de minimizar la función f es sustituida por el problema de minimización cuadrática aproximada

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}^{(0)}) + \nabla f(\mathbf{x}^{(0)}) \cdot (\mathbf{x}-\mathbf{x}^{(0)}) + \frac{1}{2}\nabla^2 f(\mathbf{x}^{(0)})(\mathbf{x}-\mathbf{x}^{(0)}) \cdot (\mathbf{x}-\mathbf{x}^{(0)})$$

donde $\nabla^2 f(\mathbf{x}^{(0)})$ representa la matriz Hessiana de f en $\mathbf{x}^{(0)}$ cuyas componentes son las derivadas parciales segundas de la función. Consecuentemente, el mínimo se alcanza en la solución del sistema lineal

$$\nabla^2 f(\mathbf{x}^{(0)})(\mathbf{x}-\mathbf{x}^{(0)}) = -\nabla f(\mathbf{x}^{(0)}).$$

De modo reiterado, si se conoce $\mathbf{x}^{(k)}$ se calcula $\mathbf{x}^{(k+1)}$ como solución de

$$\nabla^2 f(\mathbf{x}^{(k)})(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}) = -\nabla f(\mathbf{x}^{(k)}).$$

Evidentemente, este método no es otro que el método de Newton, introducido en las secciones anteriores aplicado a la condición de gradiente nulo

$$\nabla f(\mathbf{x}) = 0.$$

Desde un punto de vista puramente conceptual, podría expresarse esta igualdad como en los métodos de descenso

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \nabla^2 f(\mathbf{x}^{(k)})^{-1}\nabla f(\mathbf{x}^{(k)}).$$

Obviamente,la inversión del Hessiano no es recomendable para poner en práctica el método, salvo en baja dimensión. También puede incluirse un tamaño de paso ρ_k

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \rho_k\nabla^2 f(\mathbf{x}^{(k)})^{-1}\nabla f(\mathbf{x}^{(k)})$$

que pude ser determinado por un criterio de máximo descenso. Si el Hessiano es una matriz definida positiva, la dirección

$$\mathbf{d}^{(k)} = -\nabla^2 f(\mathbf{x}^{(k)})^{-1} \nabla f(\mathbf{x}^{(k)})$$

es de descenso ya que

$$\mathbf{d}^{(k)} \cdot \nabla f(\mathbf{x}^{(k)}) = -\nabla^2 f(\mathbf{x}^{(k)})^{-1} \nabla f(\mathbf{x}^{(k)}) \cdot \nabla f(\mathbf{x}^{(k)}) \leq 0.$$

Esta situación se produce cuando la función objetivo es una función estrictamente convexa.

8.7 Ejercicios

▬ Ejercicio 93 *Invertir la siguiente matriz*

$$\mathrm{B} = \begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{pmatrix}$$

usando la fórmula de Sherman-Morrison.

Solución: Se aplica la fórmula de Sherman-Morrison con

$$\mathrm{A} = \mathrm{I}, \quad \mathbf{u} = \mathbf{v} = (1, \cdots, 1)^t.$$

De este modo, se obtiene

$$\mathrm{B}^{-1} = \mathrm{I} - \frac{1}{1+n} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \frac{1}{1+n} \begin{pmatrix} n & -1 & \cdots & -1 \\ -1 & n & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & n \end{pmatrix}. \qquad \diamond$$

▬ Ejercicio 94 *Calcular la raíces complejas del siguiente polinomio*

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

usando el método de Bairstow.

Solución: Si se identifican coeficientes en la siguiente igualdad

$$(b_0 + b_1 x)(x^2 + rs + s) + Ax + B = x^3 - x^2 + x - 1$$

se obtienen las recurrencias

$$\begin{aligned} b_1 &= 1, \\ b_1 r + b_0 &= -1, \end{aligned}$$

y las expresiones que definen A y B

$$\begin{aligned} b_1 s + b_0 r + A &= 1, \\ b_0 s + B &= -1. \end{aligned}$$

El sistema de ecuaciones no-lineales a resolver es

$$\begin{aligned} 1 - s + (1 + r)r &= 0, \\ -1 + (1 + r)s &= 0. \end{aligned}$$

Si se suman ambas ecuaciones, se obtiene que

$$r(s + 1 + r) = 0$$

de donde se deduce que $r = 0$ ó $s = -1 - r$. En el primer caso, el polinomio buscado es $x^2 + 1$. En el segundo caso, no existe ningún número real r tal que

$$(1 + r)^2 = -1.$$

Consecuentemente, las raíces complejas del polinomio son $\pm i$. ◇

▬ EJERCICIO 95 *Describir la primera iteración del método del gradiente con máximo descenso para encontrar el mínimo de la función*

$$f(x, y) = (y + x^2 - 1)^2 + (x + y^2 - 1)^2$$

partiendo del punto $(0, 0)$.

Solución: En este caso, la dirección de descenso está dada por

$$\mathbf{d}^{(0)} = -\nabla f(x^{(0)}, y^{(0)}) = (2, 2)^t.$$

Para encontrar el tamaño del paso ρ_0, se busca el mínimo de la función

$$g(t) = f(x^{(0)} + td_1^{(0)}, y^{(0)} + td_2^{(0)}) = f(2t, 2t) = 2(4t^2 + 2t - 1)^2.$$

Puesto que el mínimo de g se alcanza en el mismo punto que

$$h(t) = |4t^2 + 2t - 1|$$

el valor del paso de máximo descenso $\rho_0 = \frac{-1+\sqrt{5}}{4}$. Finalmente se obtiene que

$$(x^{(1)}, y^{(1)}) = (\frac{-1+\sqrt{5}}{2}, \frac{-1+\sqrt{5}}{2}). \qquad \diamond$$

CAPÍTULO 9

Ecuaciones en diferencias finitas

9.1 Introducción

Una ecuación en diferencias finitas está definida por una relación entre términos consecutivos de una sucesión, que pretende ser una solución. Por ejemplo, los términos x_n de la sucesión

$$\{0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \cdots, \},$$

conocida como sucesión de Fibonacci, verifican la ecuación en diferencias

$$x_{n+1} = x_n + x_{n-1}, \quad x_0 = 0, \ x_1 = 1.$$

El orden de una ecuación en diferencias finitas es la diferencia entre el mayor y el menor de los índices que aparecen en la relación que define la ecuación. De acuerdo con esta definición, la ecuación de Fibonacci es de orden 2. Las ecuaciones en diferencias aparecen en diversas áreas de la ciencia pero son particularmente importantes en algunos métodos de resolución de ecuaciones diferenciales como se pondrá de manifiesto en capítulos posteriores. El término *ecuación en diferencias finitas* suele ser utilizado en estas situaciones pero no es el único utilizado para este tipo de relaciones. En algunos de los capítulos anteriores se han estudiado *recurrencias* y *algoritmos iterativos* que son conceptos relacionados. De hecho, se podría expresar la ecuación que define la sucesión de Fibonacci en forma vectorial como

$$\mathbf{x}^{(n+1)} = \mathrm{A}\mathbf{x}^{(n)}$$

donde $\mathbf{x}^{(n)} = (x_n, x_{n+1})^t$ y A es la matriz de coeficientes

$$\mathrm{A} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

y el vector inicial $\mathbf{x}^{(0)} = (x_0, x_1)^t$ tiene como componentes a todos los datos iniciales. De este modo, la sucesión generada por la recurrencia es una sucesión de Krylov asociada a la matriz A.

En este sentido, el análisis de las ecuaciones en diferencias no parece aportar nada relevante en relación con el estudio de métodos iterativos de resolución de ecuaciones realizado en capítulos precedentes. En el caso no-lineal, la relación

$$\mathbf{x}^{(n+1)} = F(\mathbf{x}^{(n)}) = F \circ F \circ \overset{n+1)}{\cdots} \circ F(\mathbf{x}^{(0)})$$

se conoce muchas veces como sistema dinámico discreto, fundamentalmente, cuando el interés del estudio se centra en aspectos más cualitativos de la propia sucesión, más que en el cálculo del límite cuando $n \to \infty$.

Así pues, este capítulo no ofrece novedades conceptuales. Es en cierto sentido subsidiario de los capítulos posteriores dedicados a la resolución numérica de ecuaciones diferenciales. La resolución de ecuaciones en diferencias sencillas permitirá ilustrar algunos de los conceptos que aparecerán en esos capítulos. Además, este propósito está limitado al estudio de las ecuaciones lineales en diferencias, con coeficientes constantes, que pueden ponerse en la siguiente forma

$$x_{n+1} + a_{k-1}x_n + a_{k-2}x_{n-1} + \cdots + a_0 x_{n-k+1} = b_n.$$

En este caso, el cálculo de la sucesión asociada a la ecuación en diferencias es muy simple si se conocen los k primeros términos de la sucesión. Por ejemplo, en la sucesión de Fibonacci, puesto que $x_0 = 0$ y $x_1 = 1$, se pueden calcular los siguientes términos del modo siguiente

$$x_2 = x_1 + x_0 = 1, \quad x_3 = x_2 + x_1 = 2, \quad x_4 = x_3 + x_2 = 3, \quad \cdots$$

No obstante, en ocasiones es deseable conocer el valor del término $n-$ésimo sin necesidad de conocer todos los que le preceden. Es decir, se trata de encontrar una expresión de x_n en términos exclusivamente de n y de los k primeros términos. Esto es lo que se conoce comúnmente como resolver la ecuación en diferencias.

9.2 Ecuaciones lineales homogéneas en diferencias con coeficientes constantes

La ecuación lineal homogénea en diferencias

$$x_{n+1} + a_{k-1}x_n + a_{k-2}x_{n-1} + \cdots + a_0 x_{n-k+1} = 0, \quad a_0 \neq 0, \quad n \geq k-1$$

puede expresarse como

$$x_{n+k} + a_{k-1}x_{n+k-1} + a_{k-2}x_{n+k-2} + \cdots + a_1 x_{n+1} + a_0 x_n = 0, \quad a_0 \neq 0, \quad n \geq 0.$$

En esta forma es más fácil expresarla en forma vectorial como

$$\mathbf{x}^{(n+1)} = \mathrm{A}\mathbf{x}^{(n)} = \mathrm{A}^n \mathbf{x}^{(0)},$$

donde $\mathbf{x}^{(n)} = (x_n, \cdots, x_{n+k-1})^t$ y A es la matriz de compañía

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{k-1} \end{pmatrix},$$

cuya ecuación característica en la incógnita λ es

$$\lambda^k + a_{k-1}\lambda^{k-1} + \cdots + a_1\lambda + a_0 = 0.$$

Puesto que el determinante de la matriz de compañía es $(-1)^{n-1}a_0$, la matriz es invertible y todos sus autovalores son distintos de 0.

El cálculo de las potencias sucesivas de A puede ser simplificado si se utiliza su factorización canónica de Jordan. De acuerdo con el teorema de representación de Jordan, existe una matriz invertible P tal que

$$\mathrm{A} = \mathrm{PJP}^{-1},$$

donde la matriz canónica de Jordan J es una matriz diagonal por bloques

$$\mathrm{J} = \begin{pmatrix} \mathrm{J}_1 & \mathrm{O} & \cdots & \mathrm{O} \\ \mathrm{O} & \mathrm{J}_2 & \cdots & \mathrm{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{O} & \mathrm{O} & \cdots & \mathrm{J}_m \end{pmatrix}$$

en la que cada bloque tiene la siguiente forma

$$\mathrm{J}_i = \begin{pmatrix} \lambda_i & 1 & \cdots & \mathrm{O} & \mathrm{O} \\ \mathrm{O} & \lambda_i & \cdots & \mathrm{O} & \mathrm{O} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{O} & \mathrm{O} & \cdots & \lambda_i & 1 \\ \mathrm{O} & \mathrm{O} & \cdots & \mathrm{O} & \lambda_i \end{pmatrix}$$

para $i = 1, 2, \cdots, m$ y la suma de las dimensiones de los bloques de Jordan es k.

De este modo se tiene que

$$\mathrm{A}^n = \mathrm{P}\mathrm{J}^n\mathrm{P}^{-1}$$

y

$$\mathrm{J}^n = \begin{pmatrix} \mathrm{J}_1^n & \mathrm{O} & \cdots & \mathrm{O} \\ \mathrm{O} & \mathrm{J}_2^n & \cdots & \mathrm{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{O} & \mathrm{O} & \cdots & \mathrm{J}_m^n \end{pmatrix}.$$

La $n-$ésima potencia de cada bloque está dada por

$$\mathrm{J}_i^n = \begin{pmatrix} \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} & \cdots & \binom{n}{m_i-2}\lambda_i^{n-m_i+2} & \binom{n}{m_i-1}\lambda_i^{n-m_i+1} \\ \mathrm{O} & \lambda_i^n & \cdots & \binom{n}{m_i-3}\lambda_i^{n-m_i+3} & \binom{n}{m_i-2}\lambda_i^{n-m_i+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{O} & \mathrm{O} & \cdots & \lambda_i^n & \binom{n}{1}\lambda_i^{n-1} \\ \mathrm{O} & \mathrm{O} & \cdots & \mathrm{O} & \lambda_i^n \end{pmatrix}.$$

Por ejemplo, si

$$\mathrm{J} = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

entonces

$$\mathrm{J}^n = \begin{pmatrix} 2^n & \binom{n}{1}2^{n-1} & 0 & 0 & 0 \\ 0 & 2^n & 0 & 0 & 0 \\ 0 & 0 & 3^n & \binom{n}{1}3^{n-1} & \binom{n}{2}3^{n-2} \\ 0 & 0 & 0 & 3^n & \binom{n}{1}3^{n-1} \\ 0 & 0 & 0 & 0 & 3^n \end{pmatrix}.$$

Consecuentemente, la solución de la ecuación en diferencias homogénea es combinación lineal de las siguientes funciones en la variable n

$$\left\{\lambda_i^n, \binom{n}{1}\lambda_i^{n-1}, \cdots, \binom{n}{m_i-1}\lambda_i^{n-m_i+1} : \ i=1,\cdots,m\right\}.$$

– EJERCICIO 96 *El matemático francés Edouard Lucas (1842-1891) fue quién dio el nombre de sucesión de Fibonacci a la sucesión de números* $\{x_n\}$ *anteriormente descrita, mientras estudiaba otra sucesión* $\{2,1,3,4,7,11,18,\cdots\}$ *de propiedades similares y que hoy se conoce como la sucesión de los números de Lucas. Probar que la siguiente relación*

$$x_{2n} = x_n z_n$$

entre los términos z_n *de la sucesión de Lucas y los términos* x_n *de la sucesión de Fibonacci es válida.*

Solución: Las sucesiones de Fibonacci y de Lucas son soluciones de la ecuación en diferencias

$$x_{n+1} = x_n + x_{n-1}$$

con distintas condiciones iniciales. Esta ecuación en diferencias tiene como ecuación característica

$$\lambda^2 - \lambda - 1 = 0$$

cuyas raíces son

$$\lambda_1 = \frac{1-\sqrt{5}}{2}, \ \ \lambda_2 = \frac{1+\sqrt{5}}{2}.$$

Consecuentemente λ_1^n y λ_2^n son soluciones de la ecuación en diferencias. La solución general de esta ecuación es

$$x_n = c_1\left(\frac{1+\sqrt{5}}{2}\right)^n + c_2\left(\frac{1-\sqrt{5}}{2}\right)^n. \tag{9.1}$$

Si se imponen los dos pares de condiciones iniciales, se obtienen las soluciones

$$x_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$$

$$z_n = \left(\frac{1+\sqrt{5}}{2}\right)^n + \left(\frac{1-\sqrt{5}}{2}\right)^n.$$

Directamente se prueba que $x_n z_n = x_{2n}$. ◇

Obviamente, la situación es más complicada cuando la ecuación característica presenta raíces múltiples.

▪ Ejemplo **45** Para hallar la solución general de la ecuación en diferencias

$$x_{n+1} - 6x_n + 12x_{n-1} - 8x_{n-2} = 0.$$

se analiza la ecuación característica asociada

$$\lambda^3 - 6\lambda^2 + 12\lambda - 8 = 0,$$

que en este caso tiene como raíz triple a $\lambda = 2$. Ahora, la cuestión es decidir cuál de las tres siguientes matrices

$$\mathrm{J} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad \mathrm{J} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad \mathrm{J} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

corresponde a la matriz de Jordan de la matriz de compañía

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 8 & -12 & 6 \end{pmatrix}$$

asociada a la ecuación en diferencias. Para ello basta tener en cuenta que los rangos de las matrices $A - \lambda\mathrm{I}$ y $\mathrm{J} - \lambda\mathrm{I}$ coinciden. Así, si el rango de $\mathrm{A} - \lambda\mathrm{I}$ es 0, la elección corresponde a la primera matriz, si el rango es 1 corresponde a la segunda y si es 2, corresponde a la tercera. En este caso, el rango de $A - 2\mathrm{I}$ es 2, por lo que la matriz de Jordan es

$$\mathrm{J} = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

La solución general de la ecuación en diferencias es

$$x_n = c_1 2^n + c_2 n 2^{n-1} + c_3 n(n-1)2^{n-2}$$

donde c_1, c_2 y c_3 son constantes arbitrarias. ◇

Es importante destacar que la discusión realizada en el ejemplo anterior conduce siempre al mismo resultado en lo que se refiere al rango, a causa de que para una matriz de compañía A de dimensión k el rango de $\mathrm{A} - \lambda\mathrm{I}$, con $\lambda \neq 0$, tiene siempre rango $k - 1$ ya que el menor principal

$$\begin{vmatrix} -\lambda & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda & 1 \\ 0 & 0 & 0 & \cdots & 0 & -\lambda \end{vmatrix}$$

de la matriz $\mathrm{A} - \lambda \mathrm{I}$, es distinto de 0, mientras que el determinante de la matriz es 0, por ser λ un autovalor. Es decir, cada uno de los autovalores de una matriz de compañía está en un único bloque de Jordan.

Otra situación importante a considerar corresponde al caso en el que existen raíces complejas. Si los coeficientes de la ecuación lineal en diferencias son constantes y reales, y $\lambda = |\lambda|(\cos\theta + i\,\mathrm{sen}\,\theta)$ es una raíz compleja del polinomio característico entonces $\bar{\lambda} = |\lambda|(\cos\theta - i\,\mathrm{sen}\,\theta)$ también es una raíz. Una combinación de las dos soluciones λ^n y $\bar{\lambda}^n$, conduce a la solución

$$c_1\lambda^n + c_2\bar{\lambda}^n = (c_1 + c_2)|\lambda|^n \cos n\theta + i(c_1 - c_2)|\lambda|^n \,\mathrm{sen}\, n\theta.$$

Si se escoge

$$c_1 = \frac{1}{2}(C_1 - iC_2), \quad c_2 = \frac{1}{2}(C_1 + iC_2)$$

para C_1 y C_2 arbitrarios, se obtiene que

$$C_1|\lambda|^n \cos n\theta + C_2|\lambda|^n \,\mathrm{sen}\, n\theta$$

es solución de la ecuación en diferencias para todo C_1 y C_2. Es decir, $|\lambda|^n \cos n\theta$ y $|\lambda|^n \,\mathrm{sen}\, n\theta$ son también soluciones reales de la ecuación en diferencias. Se puede utilizar un argumento similar para el caso de raíces complejas de multiplicidad mayor que 1.

– Ejercicio 97 *Hallar la solución de la ecuación en diferencias*

$$x_{n+1} + 2x_n + 2x_{n-1} = 0$$

tal que $x_0 = x_{102} = 1$.

Solución: La ecuación característica

$$\lambda^2 + 2\lambda + 2 = 0$$

tiene como raíces los números complejos $\lambda = -1 \pm i = \sqrt{2}(\cos\frac{3\pi}{4} \pm i\,\mathrm{sen}\,\frac{3\pi}{4})$. La solución general de la ecuación en diferencias es

$$x_n = 2^{\frac{n}{2}}\left(c_1 \cos\frac{3\pi n}{4} + c_2\,\mathrm{sen}\,\frac{3\pi n}{4}\right).$$

Si se imponen las condiciones adicionales, se obtiene

$$c_1 = 1, \qquad 2^{51}\left(c_1 \cos\frac{153\pi}{2} + c_2\,\mathrm{sen}\,\frac{153\pi}{2}\right) = 1$$

de donde se deduce que la solución buscada es

$$x_n = 2^{\frac{n}{2}} \left(\cos \frac{3\pi n}{4} + 2^{-51} \operatorname{sen} \frac{3\pi n}{4} \right). \qquad \diamond$$

Es importante destacar que si en el ejercicio anterior la condición *de contorno* fuese $x_{100} = 1$, el problema no tendría solución. Es decir, un problema de valores iniciales para una ecuación lineal en diferencias siempre tiene solución, pero un problema de contorno puede no tenerla.

Con toda esta argumentación se ha probado que el conjunto de soluciones de una ecuación en diferencias homogénea es combinación de k soluciones, linealmente independientes, construidas con los autovalores de la matriz de compañía. A fin de profundizar en la estructura algebraica del conjunto de las soluciones de una ecuación en diferencias, se consideran k soluciones arbitrarias de la ecuación homogénea y se representa por $\mathrm{C}^{(n)}$ la matriz cuyas columnas están formadas por los vectores $\mathbf{x}^{(n)}$, correspondientes a cada una de la soluciones. Esta matriz se conoce como matriz de Casorati asociada a las k soluciones de la ecuación lineal en diferencias, y a su determinante $c^{(n)} = \det \mathrm{C}^{(n)}$, como casoratiano.

▪ EJEMPLO **46** La ecuación en diferencias de Fibonacci tiene como soluciones básicas

$$x_n = \lambda_1^n, \;\; y_n = \lambda_2^n$$

donde λ_1 y λ_2 son los autovalores $\frac{1 \pm \sqrt{5}}{2}$. La matriz casotariana de estas soluciones es

$$\mathrm{C}^{(n)} = \begin{pmatrix} \lambda_1^{n-1} & \lambda_2^{n-1} \\ \lambda_1^n & \lambda_2^n \end{pmatrix}$$

y el casoratiano

$$c^{(n)} = \lambda_1^{n-1}\lambda_2^n - \lambda_1^n\lambda_2^{n-1} = (\lambda_1\lambda_2)^{n-1}(\lambda_2 - \lambda_1) = (-1)^{n-1}\sqrt{5}.$$

Si se consideran como soluciones las de Fibonacci y de Lucas la matriz de Casorati de estas soluciones puede expresarse como

$$\mathrm{C}^{(n)} = \begin{pmatrix} \lambda_1^{n-1} & \lambda_2^{n-1} \\ \lambda_1^n & \lambda_2^n \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ 1 & 1 \end{pmatrix}$$

y el casoratiano $c^{(n)} = 2(-1)^{n-1}$. $\diamond$

En general, la matriz de Casorati verifica que $C^{(n)} = \mathrm{A}^n C^{(0)}$ y consecuentemente

$$c^{(n)} = \det\,\mathrm{C}^{(\mathrm{n})} = (\det\,\mathrm{A})^{\mathrm{n}} \det\,\mathrm{C}^{(0)} = \left((-1)^{\mathrm{k}}\mathrm{a}_0\right)^{\mathrm{n}}\,\mathrm{c}^{(0)}.$$

De todo ello se deduce el siguiente

Teorema 41 *Para k soluciones de la ecuación lineal homogénea, las tres siguientes afirmaciones son equivalentes*

1. *Las k soluciones son linealmente dependientes.*
2. *El casoratiano $c^{(n)} = 0$ para algún n.*
3. *El casoratiano $c^{(n)} = 0$ para todo $n \geq 0$.*

9.3 Ecuaciones lineales en diferencias con coeficientes constantes

Ahora se considera la siguiente ecuación en diferencias con coeficientes constantes

$$x_{n+1} + a_{k-1}x_n + a_{k-2}x_{n-1} + \cdots + a_0 x_{n-k+1} = b_n.$$

Si el término independiente no cumple la condición $b_n = 0$ para todo $n > k$, la ecuación no es homogénea. La ecuación se puede expresarse en forma vectorial como

$$\mathbf{x}^{(n+1)} = \mathrm{A}\mathbf{x}^{(n)} + \mathbf{b}^{(n)}$$

donde $\mathbf{b}^{(n)} = (0, \cdots, 0, b_{n+k-1})^t$. Evidentemente, para todo dato inicial $\mathbf{x}^{(0)}$ la ecuación tiene solución única.

Si se conoce una solución particular $\mathbf{x}_0^{(n)}$ de esta ecuación y $\mathbf{z}^{(n)}$ es la solución general de la ecuación homogénea asociada

$$\mathbf{z}^{(n+1)} = \mathrm{A}\mathbf{z}^{(n)},$$

entonces $\mathbf{x}_0^{(n)} + \mathbf{z}^{(n)}$ es la solución general de la ecuación completa. De este modo, una estrategia para resolver la ecuación no homogénea consiste en encontrar alguna solución particular y completarla con la solución general de la ecuación homogénea.

Aunque no hay un procedimiento general que permita calcular con sencillez una solución particular de la ecuación completa, en algunas situaciones

especiales existen métodos que pueden resultar eficaces. En este sentido, si b_n tiene la siguiente forma

$$b_n = d^n p(n)$$

donde d es una constante y p un polinomio en n, entonces se puede calcular una solución particular de la ecuación completa por el llamado método de los coeficientes indeterminados. Para ello, basta probar con una solución de la forma

$$x_{0n} = d^n q(n)$$

donde $q(n)$ es un polinomio en n a determinar.

▬ EJERCICIO 98 *Calcular la solución general de la ecuación lineal en diferencias*

$$x_{n+1} - 5x_n + 6x_{n-1} = 5^n n.$$

Solución: En primer lugar se calcula la solución general de la homogénea

$$x_{n+1} - 5x_n + 6x_{n-1} = 0.$$

Para ello, se buscan las raíces de la ecuación característica

$$\lambda^2 - 5\lambda + 6 = 0.$$

de ello resulta que la solución general de la homogénea es

$$z_n = c_1 2^n + c_2 3^n.$$

Se considera una solución particular de la completa

$$x_n^0 = 5^n(r_1 n + r_0).$$

Si se sustituye esta expresión en la ecuación completa se obtiene $r_0 = -\frac{95}{36}$ y $r_1 = \frac{5}{6}$.

Consecuentemente, la solución general de la completa es

$$x_n = 5^n \left(\frac{5}{6}n - \frac{95}{36}\right) + c_1 2^n + c_2 3^n. \qquad \diamond$$

▬ EJERCICIO 99 *Hallar una función f que verifique las condiciones*

$$\begin{aligned} f(n+1) - 3f(n) - n &= 0, \quad \textit{para todo número entero } n > 0, \\ f(0) &= 0. \end{aligned}$$

Solución: Se considera la siguiente ecuación en diferencias

$$x_{n+1} - 3x_n = n.$$

La ecuación característica es $\lambda - 3 = 0$ y por lo tanto, la solución general de la ecuación homogénea asociada es $z_n = c3^n$ donde c es una constante arbitraria. Para calcular una solución particular de la completa se usará el método de identificación de parámetros, probando con la siguiente función

$$x_n^0 = r_1 n + r_0.$$

De este modo se obtiene que

$$-2r_1 n + r_1 - 2r_0 = n,$$

de donde se deduce que

$$x_n^0 = -\frac{1}{2}n - \frac{1}{4}.$$

La solución general de la completa es

$$x_n = c3^n - \frac{1}{2}n - \frac{1}{4}$$

y si se impone la condición en $n = 0$ se obtiene que $c = \frac{1}{4}$. De este modo se llega al siguiente resultado

$$f(n) = \frac{1}{4}\left(3^n - 2n - 1\right). \qquad \diamond$$

9.4 Estabilidad

Una ecuación en diferencias se dice estable si todas sus soluciones permanecen acotadas cuando n tiende a infinito. Una ecuación en diferencias se dice fuertemente estable si todas sus soluciones tienden a 0 cuando n tiende a infinito.

TEOREMA 42 *Una ecuación lineal homogénea en diferencias con coeficientes constantes es estable si y sólo si todas las raíces del polinomio característico tienen módulo menor o igual que uno y aquellas cuyo módulo es 1, son raíces simples. La ecuación es fuertemente estable si y sólo si todas sus raíces tienen módulo menor que 1.*

Demostración: Una solución de una ecuación lineal en diferencias con coeficientes constantes se puede expresar como una combinación lineal de términos de la forma $n^i\lambda^n$ para i tal que $0 \leq i \leq r-1$, siendo r la multiplicidad de la raíz λ del polinomio característico de la ecuación. Obviamente, la ecuación en diferencias es estable (ó fuertemente estable) si y sólo si lo son cada una de las soluciones elementales. La solución elemental $n^i\lambda^n$ será acotada si y sólo si $|\lambda| \leq 1$ y $i = 0$ si $\lambda = 1$. La solución elemental $n^i\lambda^n$ tiende a 0 cuando $n \to \infty$ si y sólo si $|\lambda| < 1$. También podría haberse empleado el corolario 1 de la página 20 para justificar la segunda afirmación. ◇

▪ EJEMPLO **47** Para estudiar la estabilidad de la ecuación en diferencias

$$2x_{n+1} - 5x_n + 2x_{n-1} = 0$$

se considera su ecuación característica

$$\lambda^2 - \frac{5}{2}\lambda + 1 = 0$$

cuyas raíces son $\lambda_1 = 2$ y $\lambda_2 = \frac{1}{2}$. De acuerdo con el teorema precedente, la ecuación en diferencias no es estable. La solución general de esta ecuación es

$$x_n = c_1 2^n + c_2 2^{-n}.$$

En particular, para $c_1 = 0$ y $c_2 = 1$, la solución $x_n = 2^{-n}$ es acotada y tiende a cero cuando $n \to \infty$, pero la solución correspondiente a $c_1 = 1$ y $c_2 = 0$, $x_n = 2^n$ no es acotada.

▪ EJEMPLO **48** La ecuación característica de la ecuación en diferencias

$$x_{n+1} + 2x_n + 2x_{n-1} = 0$$

es

$$\lambda^2 + 2\lambda + 2 = 0$$

y tiene como raíces los números complejos $\lambda = -1 \pm i = \sqrt{2}(\cos\frac{3\pi}{4} \pm i \operatorname{sen}\frac{3\pi}{4})$. Consecuentemente la ecuación es inestable. La solución de la ecuación correspondiente a los datos iniciales $x_0 = 0, x_1 = 1$ presenta oscilaciones de magnitud creciente como muestra la figura 9.1. ◇

▬ EJERCICIO 100 *¿Se pueden encontrar valores iniciales x_0, x_1 y x_2 tales que la correspondiente solución de la ecuación en diferencias*

$$4x_{n+2} - 8x_{n+1} + 5x_n - x_{n-1} = 0$$

tienda a ∞ cuando $n \to \infty$?.

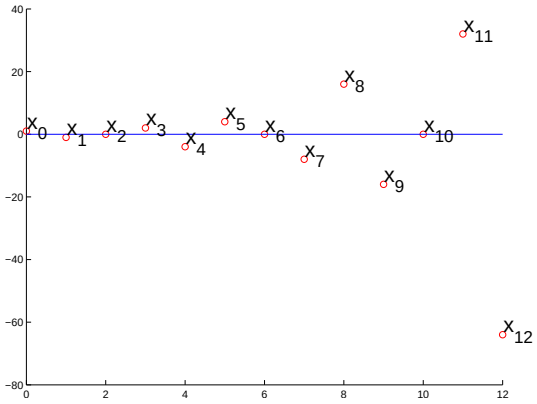


Figura 9.1: Solución de la ecuación $x_{n+1} + 2x_n + 2x_{n-1} = 0$ para $x_0 = 0$ y $x_1 = 1$

Solución: La ecuación característica asociada a esta ecuación es

$$4\lambda^3 - 8\lambda^2 + 5\lambda - 1 = 0$$

cuyas raíces son $\lambda = 1$ (simple) y $\lambda = \frac{1}{2}$ (doble). Consecuentemente, la ecuación en diferencias es estable y sus soluciones son acotadas para datos iniciales cualesquiera. ◇

9.5 Ejercicios

▁ EJERCICIO 101 *Determinar las soluciones acotadas de la ecuación en diferencias*

$$x_{n+1} - x_n - x_{n-1} + x_{n-2} = 0$$

que verifiquen las condiciones iniciales $x_0 = 1,\ x_1 = 3$.

Solución: La ecuación característica asociada a la ecuación en diferencias es la siguiente

$$\lambda^3 - \lambda^2 - \lambda + 1 = 0.$$

Las raíces de esta ecuación son -1 y 1 (doble). En consecuencia, la solución general de la ecuación en diferencias homogénea es

$$x_n = c_1 + c_2 n + c_3(-1)^n$$

donde c_1, c_2 y c_3 representa constantes arbitrarias. Puesto que el primer y tercer sumando en la expresión anterior están acotados, para que la solución

sea acotada es necesario y suficiente que $c_2 = 0$. Por otra parte, si se imponen las condiciones iniciales, se obtiene

$$c_1 + c_3 = 1,$$

$$c_1 - c_3 = 3,$$

de donde se deduce que $c_1 = 2$ y $c_3 = -1$. En definitiva, la única solución que verifica las condiciones impuestas es

$$x_n = 2 - (-1)^n. \qquad \diamond$$

▬ EJERCICIO 102 *Hallar la solución de la ecuación en diferencias*

$$x_{n+1} + 2x_n + x_{n-1} = 3^{n-1}$$

que cumple $x_0 = x_1 = 1$.

Solución: En primer lugar se calcula la solución general de la ecuación homogénea en diferencias asociada

$$x_{n+1} + 2x_n + x_{n-1} = 0.$$

Para ello se considera la ecuación característica

$$\lambda^2 + 2\lambda + 1 = 0$$

que tiene una raíz doble $\lambda = -1$. En consecuencia, la solución general de la homogénea es

$$z_n = c_1(-1)^n + c_2 n(-1)^n.$$

Ahora, se busca una solución particular de la completa en la forma

$$x_n^0 = 3^{n-1}(r_1 n + r_0).$$

Ello conduce a $r_1 = 0$ y $r_0 = \frac{3}{16}$. La solución general de la completa es

$$x_n = c_1(-1)^n + c_2 n(-1)^n + \frac{3^n}{16}.$$

Si se imponen las condiciones iniciales se obtiene

$$x_n = \frac{15}{16}(-1)^n - \frac{7}{4}(-1)^n n + \frac{3^n}{16}. \qquad \diamond$$

▬ EJERCICIO 103 *Determinar el término x_{10} de una sucesión generada por una ecuación lineal homogénea en diferencias de orden 3 sabiendo que sus primeros términos son*

$$\{1, 1, 2, 5, 12, 27, \cdots\}.$$

Solución: Si se impone a la ecuación

$$x_{n+1} + \alpha_2 x_n + \alpha_1 x_{n-1} + \alpha_0 x_{n-2} = 0$$

que los términos $\{5, 12, 27\}$ se obtiene un sistema lineal

$$\begin{pmatrix} -1 & -1 & -2 \\ -1 & -2 & -5 \\ -2 & -5 & -12 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 12 \\ 27 \end{pmatrix}$$

cuya solución conduce a la ecuación en diferencias

$$x_{n+1} - 4x_n + 5x_{n-1} - 2x_{n-2} = 0$$

La ecuación característica asociada

$$\lambda^3 - 4\lambda^2 + 5\lambda - 2 = 0$$

tiene como autovalores $\lambda_1 = 1$ (doble) y $\lambda_2 = 2$. La solución general de la ecuación en diferencias es

$$x_n = c_1 + c_2 n + c_3 2^n$$

y la particular es

$$x_n = -n + 2^n.$$

Finalmente, se obtiene que $x_{10} = 1014$. ◇

▬ EJERCICIO 104 *Hallar la solución de la ecuación en diferencias*

$$x_{n+1} + 2x_{n-1} + x_{n-3} = 0$$

que verifica las condiciones iniciales $x_0 = x_1 = x_2 = x_3 = 1$.

Solución: La ecuación característica asociada a la ecuación en diferencias

$$\lambda^4 + 2\lambda^2 + 1 = 0$$

tiene como raíces a $\lambda = \pm i$ con multiplicidad 2. Consecuentemente, la solución general de esta ecuación es

$$x_n = (c_1 + c_2 n)\cos\frac{n\pi}{2} + (c_3 + c_4 n)\,\mathrm{sen}\,\frac{n\pi}{2}.$$

Para seleccionar la solución que verifica las condiciones iniciales, se resuelve el sistema

$$\begin{aligned} c_1 &= 1, \\ c_3 + c_4 &= 1, \\ -c_1 - 2c_2 &= 1, \\ -c_3 - 3c_4 &= 1. \end{aligned}$$

De ello se deduce que la solución buscada es

$$x_n = (1-n)\cos\frac{n\pi}{2} + (2-n)\,\mathrm{sen}\,\frac{n\pi}{2}. \qquad \diamond$$

▬ EJERCICIO 105 *Calcular el determinante de la matriz $n \times n$ tridiagonal*

$$\mathrm{A}_n = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}$$

Solución: Se representa por x_n el determinante de la matriz A_n. Si se desarrolla el determinante por la primera fila, directamente se comprueba que x_n es solución de la ecuación en diferencias

$$x_n = 2x_{n-1} - x_{n-2}$$

La ecuación característica asociada a esta ecuación es

$$\lambda^2 - 2\lambda + 1 = 0$$

Esta ecuación tiene una raíz doble $\lambda = 1$. Consecuentemente la solución general de la ecuación es

$$x_n = c_1 + c_2 n.$$

Si se imponen las condiciones iniciales $x_1 = 2$ y $x_2 = 3$ se obtiene que $x_n = n + 1$. $\diamond$

Ejercicio 106 *Sean α y β constantes positivas tales que $\alpha > 2\beta$. Calcular el determinante de la matriz $n \times n$*

$$\mathrm{A}_n = \begin{pmatrix} \alpha & -\beta & 0 & \cdots & 0 & 0 \\ -\beta & \alpha & -\beta & \cdots & 0 & 0 \\ 0 & -\beta & \alpha & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha & -\beta \\ -\beta & 0 & 0 & \cdots & -\beta & \alpha \end{pmatrix}$$

Solución: Se representa por x_n el determinante de la matriz A_n y por y_n el determinante de

$$\mathrm{B}_n = \begin{pmatrix} \alpha & -\beta & 0 & \cdots & 0 & 0 \\ -\beta & \alpha & -\beta & \cdots & 0 & 0 \\ 0 & -\beta & \alpha & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha & -\beta \\ 0 & 0 & 0 & \cdots & -\beta & \alpha \end{pmatrix}$$

para $n > 0$.

Si se desarrolla el determinante por la primera columna, directamente se comprueba que x_n e y_n son soluciones de las siguientes ecuaciones en diferencias

$$\begin{aligned} x_n - \alpha y_{n-1} + \beta^2 y_{n-2} &= \beta^n, \\ y_n - \alpha y_{n-1} + \beta^2 y_{n-2} &= 0. \end{aligned}$$

La ecuación característica asociada a la segunda ecuación es

$$\lambda^2 - \alpha\lambda + \beta^2 = 0.$$

Esta ecuación tiene dos raíces reales simples

$$\lambda_1 = \frac{\alpha - \sqrt{\alpha^2 - 4\beta^2}}{2}, \quad \lambda_2 = \frac{\alpha + \sqrt{\alpha^2 - 4\beta^2}}{2}.$$

La solución general de la segunda ecuación es

$$y_n = c_1\lambda_1^n + c_2\lambda_2^n.$$

Los coeficientes c_1 y c_2 se determinan imponiendo las condiciones iniciales $y_0 = \alpha$ y $y_1 = \alpha^2 - \beta^2$

$$\begin{aligned} c_1\lambda_1 + c_2\lambda_2 &= \alpha, \\ c_1\lambda_1^2 + c_2\lambda_2^2 &= \alpha^2 - \beta^2. \end{aligned}$$

Si se usa la ecuación característica se puede simplificar este sistema lineal para obtener

$$\begin{aligned} c_1\lambda_1 + c_2\lambda_2 &= \alpha, \\ c_1 + c_2 &= 1. \end{aligned}$$

Consecuentemente, la solución de la ecuación en diferencias de variable y_n es

$$y_n = \frac{\lambda_2 - \alpha}{\lambda_2 - \lambda_1}\lambda_1^n + \frac{\alpha - \lambda_1}{\lambda_2 - \lambda_1}\lambda_2^n.$$

Finalmente, el determinante de la matriz A_n viene dado por la expresión

$$x_n = \alpha y_{n-1} - \beta^2 y_{n-2} - \beta^n. \qquad \diamond$$

– EJERCICIO 107 *(Iteración de Bernoulli) Calcular el autovalor de módulo máximo de la matriz*

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

Solución: El polinomio característico de la matriz es

$$p(\lambda) = \begin{vmatrix} 1-\lambda & 0 & 1 \\ 1 & 1-\lambda & 0 \\ 1 & 1 & 1-\lambda \end{vmatrix} = -\lambda^3 + 3\lambda^2 - 2\lambda + 1.$$

Una raíz real positiva λ_1 se puede localizar en el intervalo $[1, \infty)$. Además, el producto de las tres raíces (dos raíces son imaginarias conjugadas) es igual al determinante de la matriz que es uno. Consecuentemente, la raíz positiva es la de mayor módulo.

Se asocia a este polinomio la siguiente ecuación en diferencias

$$x_{n+1} = 3x_n - 2x_{n-1} + x_{n-2}$$

cuya solución general es

$$x_n = c_1\lambda_1^n + |\lambda|^n(c_2 \cos n\theta + c_3 \operatorname{sen} n\theta)$$

donde λ y θ representan el módulo y el argumento de la raíces conjugadas. Consecuentemente, si $c_1 \neq 0$ se cumple que

$$\lim_{n\to\infty} \frac{x_{n+1}}{x_n} = \lambda_1.$$

Con este resultado se puede aproximar el autovalor de módulo máximo generando la sucesión que parta de $x_0 = 1, x_1 = 1$ y $x_2 = 1$

$$\{1, 1, 1, 2, 5, 12, 28, 65, 151, 351, 816, \cdots\}$$

y calculando los cocientes

$$\{1, 1, 2, 2.5, 2.4, 2.3333, 2.3214, 2.3231, 2.3245, 2.3248, 2.3248, \cdots\}$$

se obtiene una sucesión convergente al autovalor de módulo máximo. Es preciso observar que la elección de los datos iniciales es arbitraria, sin más limitación que produzcan una sucesión de términos distintos de 0. ◇

▬ EJERCICIO 108 *Calcular la probabilidad de que al lanzar n veces una moneda al aire no salgan dos caras seguidas.*

Solución: Se representa por x_n el número de casos favorables en n lanzamientos. Si en el último lanzamiento sale una cruz, el número de casos posibles se reduce a x_{n-1} ya que no influye este hecho en los anteriores lanzamientos. Si en el último lanzamiento sale una cara, el número de casos posibles se reduce a x_{n-2} ya que fuerza a que el penúltimo sea cruz. Consecuentemente, el número de casos favorables verifica la ecuación de Fibonacci

$$x_n = x_{n-1} + x_{n-2}.$$

La solución general de la ecuación de Fibonacci está dada por la fórmula 9.1 de la página 245. Si se imponen las siguientes condiciones iniciales $x_1 = 2$ y $x_2 = 3$ se obtienen las ecuaciones

$$\begin{aligned} c_1\lambda_1 + c_2\lambda_2 &= 2, \\ c_1\lambda_1^2 + c_2\lambda_2^2 &= 3. \end{aligned}$$

(λ_1 y λ_2 representan las raíces de la ecuación característica). Puesto que el número de casos posibles es 2^n, se tiene que la probabilidad del suceso es la siguiente

$$p_n = \left(2 - \frac{1}{\lambda_1}\right)\left(\frac{\lambda_1}{2}\right)^n + \left(-2 + \frac{3}{\lambda_2}\right)\left(\frac{\lambda_2}{2}\right)^n. \qquad \diamond$$

EJERCICIO 109 *Determinar el término general de la sucesión*

$$\left\{\frac{1}{2}, \frac{4}{3}, \frac{7}{5}, \frac{10}{9}, \dots\right\}$$

sabiendo que numerador y denominador siguen sendas ecuaciones lineales en diferencias de segundo orden homogéneas.

Solución: El numerador y denominador del término general de la sucesión de fracciones $\left\{\frac{y_n}{x_n}\right\}$ verifican las ecuaciones en diferencias

$$y_{n+1} + \alpha_1 y_n + \alpha_0 y_{n-1} = 0,$$

$$x_{n+1} + \beta_1 x_n + \beta_0 x_{n-1} = 0.$$

Si se imponen las dos condiciones iniciales se obtienen los sistemas

$$\begin{pmatrix} 1 & 4 \\ 4 & 7 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} -7 \\ -10 \end{pmatrix}, \quad \begin{pmatrix} 2 & 3 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} -5 \\ -9 \end{pmatrix},$$

de donde se deduce que las ecuaciones en diferencias buscadas son

$$y_{n+1} - 2y_n + y_{n-1} = 0,$$

$$x_{n+1} - 3x_n + 2x_{n-1} = 0.$$

Las raíces de la primera ecuación característica son 1 (doble) y las de la segunda, 1 y 2. Las soluciones generales de ambas ecuaciones son

$$y_n = c_1 n + c_2, \quad x_n = c_1 2^n + c_2$$

Si se imponen las condiciones iniciales, se obtiene que el término general buscado es

$$\frac{y_n}{x_n} = \frac{3n+1}{2^n+1}. \qquad \diamond$$

CAPÍTULO 10

Problemas de valor inicial para ecuaciones diferenciales

10.1 Introducción

En los capítulos precedentes, las ecuaciones que se pretendían resolver, tenían como incógnitas un escalar, un vector o una sucesión numérica. En este capítulo, el interés se centra en la resolución de ecuaciones cuya incógnita es una función de una variable real y que involucran operadores diferenciales. Como se ha señalado anteriormente, los métodos analíticos que permiten obtener una expresión explícita de la solución en términos de funciones elementales, pueden ser aplicados en situaciones muy restringidas. En situaciones generales, posiblemente son los métodos numéricos la única alternativa válida para resolver el problema.

Los métodos numéricos transforman las ecuaciones diferenciales en ecuaciones numéricas mediante técnicas basadas en la aproximación de funciones. Una característica esencial de las ecuaciones numéricas que se generan en el proceso de discretización de las ecuaciones diferenciales es el elevado número de incógnitas que implican. De [5] se han extraído los siguientes comentarios; Uno de los más importantes cálculos de la historia de la ciencia es la predicción hecha por Clairaut, Lalande y Lepaute en 1748 sobre el retorno del cometa Halley. Sobre ello, Lalande escribió *Durante seis meses, calculábamos de la mañana a la noche, a veces incluso en las comidas; como consecuencia de lo cual, contraje una enfermedad que cambió mi constitución para siempre. La ayuda prestada por Madame Lalande fue tal que sin ella, no hubiésemos*

podido llevar a cabo una tarea que implicaba el cálculo de la distancia de Júpiter y Saturno al cometa, separadamente por cada grado, durante 150 años. Sin duda, muchas cosas han cambiado desde entonces. Ahora, los cálculos pueden ser llevados a cabo por computadoras en tiempos reducidos. Métodos cuya aplicación práctica parecía descartada, se revalorizan a la vista de las nuevas posibilidades de cálculo. El tiempo de cálculo, que antes era el primer condicionante, ahora puede ser menos relevante en la elección de un método que la estabilidad del esquema numérico empleado en la discretización de la ecuación diferencial.

En relación con las ecuaciones diferenciales, se pueden plantear dos problemas de naturaleza muy diferente: Los problemas de valor inicial (ó de Cauchy) que fijan el valor de la solución en un punto inicial y los problemas de contorno que usan información sobre la solución, en los extremos del intervalo de interés. Las técnicas que se emplean son diferentes ya que unos son problemas de naturaleza local mientras que los otros, lo son de naturaleza global. Este capítulo está dedicado exclusivamente la resolución numérica de problemas de valor inicial.

10.2 Método de Euler

Asociado a una función $f : D = [a, b] \times [c, d] \subset \mathbb{R}^2 \to \mathbb{R}$ y un punto $(t_0, x_0) \in D$, se considera el siguiente problema de valor inicial: Hallar una función $x : I \subset [a, b] \to \mathbb{R}$ tal que $t_0 \in I$ y

$$x(t_0) = x_0, \quad \frac{dx}{dt} = f(t, x(t)), \quad \text{para todo } t \in I.$$

El planteamiento de problemas de valor inicial para ecuaciones diferenciales ordinarias o en derivadas parciales es un instrumento esencial en la construcción de modelos que tratan de representar la evolución temporal de alguna magnitud. Por esta razón, se usará la letra t para representar esta variable a la que en ocasiones nos referiremos como tiempo sin menoscabo de que pueda tener otras interpretaciones. La continuidad de la función f permite garantizar la existencia local de solución (teorema de Cauchy-Peano) y una condición de Lipschitz en la segunda variable

$$|f(t, x_1) - f(t, x_2)| < K|x_1 - x_2| \quad \text{para todo } x_1, x_2 \in [c, d]$$

para alguna constante $K > 0$, permite asegurar la existencia global en $[a, b]$ y la unicidad de solución (teorema de Picard-Lipschitz).

La prueba estándar de la existencia de solución al problema de valor inicial para la ecuación diferencial en el teorema de Cauchy-Peano, es constructiva

en el sentido de que permite generar una sucesión de funciones discretas (es decir, determinadas por un número finito de valores) que se aproxima a la solución. La ecuación diferencial establece que el valor de la pendiente de la solución en t viene dado por el valor de la función f en el punto $(t, x(t))$. La idea en la que se apoya esta prueba consiste en imponer esa condición únicamente en un número finito de puntos de $[a, b]$ y reemplazar la función incógnita por un polinomio de interpolación lineal a trozos, asociado a la partición del intervalo que define ese conjunto de puntos y cuya gráfica se conoce como poligonal de Euler.

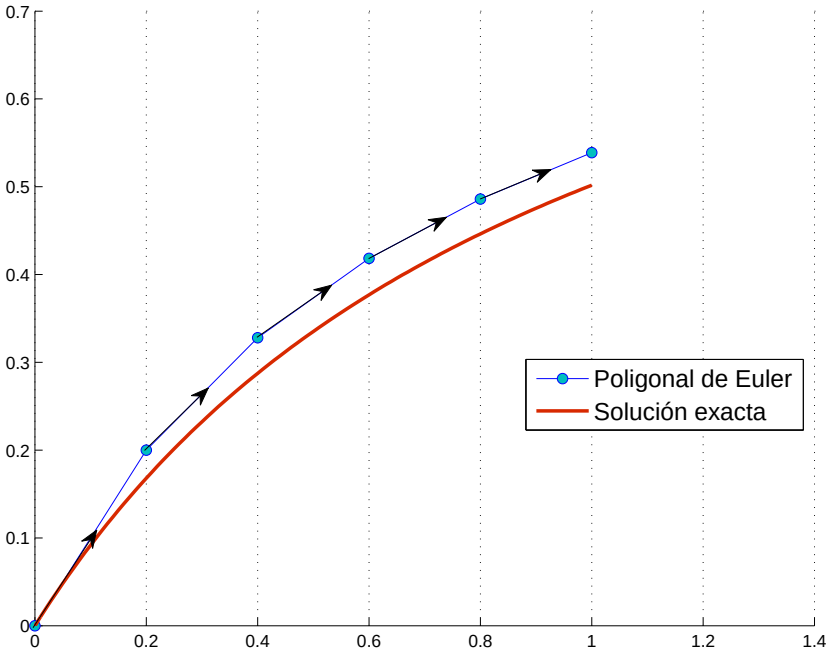


Figura 10.1: Método explícito de Euler para $\frac{dx}{dt} = (1-x)^2,\ x(0) = 0$

Sea $t_0 < t_1 < \cdots < t_n < \ldots$ una colección de puntos en los que se quiere aproximar la solución. Se representa por $h_n = t_{n+1} - t_n$ el tamaño de paso de tiempo en t_n y por $x_0, x_1, \cdots$ los valores aproximados de la solución en los tiempos t_n para $n = 0, 1, \cdots, N_h = \max\{n : t_n \in I\}$, con el método de Euler. El valor de la derivada lateral izquierda del polinomio de interpolación en t_n es $\frac{x_{n+1}-x_n}{h_n}$. Consecuentemente, la ecuación diferencial en los puntos (t_n, x_n) se convierte en la relación numérica

$$\frac{x_{n+1} - x_n}{h_n} = f(t_n, x_n)$$

para $n = 0, 1, \cdots, N_h - 1$, que se conoce como esquema explícito de Euler.

▪ Ejemplo **49** En el problema de valor inicial para la ecuación diferencial

$$x(t_0) = x_0, \quad \frac{dx}{dt} = x, \quad \text{para todo } a \leq t \leq b$$

el esquema de Euler conduce a

$$x_{n+1} = (1 + h_n)x_n.$$

En el caso de paso constante $h_n = h$, el valor x_n se puede expresar en términos del valor inicial

$$x_n = (1 + h)^n x_0.$$

Es oportuno recordar que la solución exacta de esta ecuación diferencial es

$$x(t_n) = e^{t_n - t_0} x_0 = e^{nh} x_0$$

en el tiempo $t_n = t_0 + nh$.

El error global de discretización en el instante t_n es

$$\begin{aligned} x_n - x(t_n) &= \left[(1+h)^n - e^{nh}\right] x_0 \\ &= \left[1 + nh + \frac{n(n-1)}{2}h^2 + \cdots + h^n - 1 - nh - \frac{n^2h^2}{2} - \cdots\right] x_0 \\ &= -\frac{n}{2}h^2 + O(h^3) = -\frac{t_n - t_0}{2}x_0 h + O(h^3) = O(h) \end{aligned}$$

ya que $|t_n - t_0|$ está acotado por la longitud $b - a$ del intervalo. En la expresión anterior, $O(h)$ representa una O grande de Landau en h.

También resulta interesante conocer la precisión con la que la solución exacta verifica la ecuación en diferencias asociada al esquema y que se puede medir por el error local de truncamiento T_n definido por

$$\begin{aligned} T_n &= x(t_{n+1}) - x(t_n) - hf(t_n, x(t_n)) \\ &= e^{nh}(e^h - 1 - h)x_0 = e^{nh}\left(\frac{1}{2}h^2 + \cdots\right) x_0 = O(h^2) \end{aligned}$$

ya que e^{nh} está acotado por ser el intervalo $[a, b]$ acotado. ◇

10.3 Esquemas lineales multipaso

Un modo adecuado para diseñar esquemas en diferencias finitas que permitan aproximar las soluciones a los problemas de valor inicial para una ecuación diferencial es representar la ecuación en forma integral

$$x(t) - x(t_0) = \int_{t_0}^{t} f(s, x(s))\, ds$$

y posteriormente aplicar fórmulas de cuadratura para aproximar la integral. Sin embargo, puesto que las motivaciones de los esquemas pueden ser muy variadas, a continuación se organiza su análisis, clasificándolos por su forma, mejor que por su motivación.

En esta sección se consideran los esquemas de k pasos (de tamaño de paso h constante) definidos por la expresión

$$\sum_{i=0}^{k} a_i x_{\hat{i}+n} + h \sum_{i=0}^{k} b_i f(t_{\hat{i}+n}, x_{\hat{i}+n}) = 0 \tag{10.1}$$

con la notación de índices retrasados $\hat{i} = i - k + 1$ y donde a_i y b_i son coeficientes que deben ser elegidos adecuadamente. En esta expresión

- n es un índice relacionado con la evolución del tiempo t_n.
- k es un índice independiente del tiempo, característico del esquema.
- i e $\hat{i}$ son índices relativos al sumatorio.

Con esta notación, el esquema explícito de Euler (1 paso) podría expresarse como

$$x_n - x_{n+1} + hf(t_n, x_n) = x_{\hat{0}+n} - x_{\hat{1}+n} + hf(t_{\hat{0}+n}, x_{\hat{0}+n}) = 0$$

donde $\hat{0} = 0$ y $\hat{1} = 1$. En este caso, $a_0 = 1, a_1 = -1$ y $b_0 = 1, b_1 = 0$.

Este grupo de esquemas se conoce como el de los métodos lineales multipaso. El adjetivo lineal que se utiliza, solamente se refiere al modo en que se combinan los valores x_i y $f_{\hat{i}+n} = f(t_{\hat{i}+n}, x_{\hat{i}+n})$. De hecho, la ecuación en diferencias 10.1 no es lineal si $b_k \neq 0$ y la función f no es lineal. En este caso, el esquema se dice implícito. Por el contrario, si $b_k = 0$, el esquema se dice explícito. Habitualmente, los esquemas multipaso se expresan en forma normalizada con $a_k = -1$

$$x_{n+1} = \sum_{i=1}^{k-1} a_i x_{\hat{i}+n} + h \sum_{i=0}^{k} b_i f_{\hat{i}+n}. \tag{10.2}$$

▪ EJEMPLO **50** Uno de los esquemas más simples de dos pasos es el de Simpson

$$x_{n+1} = x_{n-1} + \frac{h}{3}(f_{n+1} + 4f_n + f_{n-1}).$$

En las siguientes igualdades se ilustra el modo en que se numeran los coeficientes

$$x_{n+1} = \underbrace{0}_{a_1} x_n + \underbrace{1}_{a_0} x_{n-1} + h \left(\underbrace{\frac{1}{3}}_{b_2} f_{n+1} + \underbrace{\frac{4}{3}}_{b_1} f_n + \underbrace{\frac{1}{3}}_{b_0} f_{n-1} \right),$$

$$\begin{aligned} x_{n+1} &= \sum_{i=0}^{1} a_i x_{i+n-1} + h \sum_{i=0}^{2} b_i f_{i+n-1}, \\ &= \sum_{i=0}^{1} a_i x_{\hat{i}+n} + h \sum_{i=0}^{2} b_i f_{\hat{i}+n}. \end{aligned}$$

Es importante destacar que para arrancar un método multipaso, es preciso primero determinar los valores iniciales $x_1, \cdots, x_{k-1}$ por otro método que requiera menos pasos previos ya que el único dato que proviene del problema exacto es x_0 y el primer valor calculable con el esquema es

$$x_k = a_0 x_0 + \cdots + a_{k-1} x_{k-1} + h(b_0 f_0 + b_1 f_1 + \cdots + b_k f_k). \tag{10.3}$$

▬ EJERCICIO 110 *Aproximar el valor de la solución del problema de valor inicial*

$$x(0) = 1, \quad \frac{dx}{dt} = x, \quad \textit{para todo } t \geq 0$$

en $t = 2$, usando el esquema de Simpson y arrancando con el esquema explícito de Euler con el tamaño de paso $h = \frac{1}{2}$.

Solución: Si se usa un esquema explícito de Euler para arrancar, los primeros términos de x_n son

$$x_0 = 1, \;\; x_1 = x_0 + hx_0 = 1 + h = \frac{3}{2}.$$

Para $n \geq 2$, el esquema de Simpson se aplica mediante la ecuación

$$x_{n+1} = x_{n-1} + \frac{1}{6}(x_{n+1} + 4x_n + x_{n-1}),$$

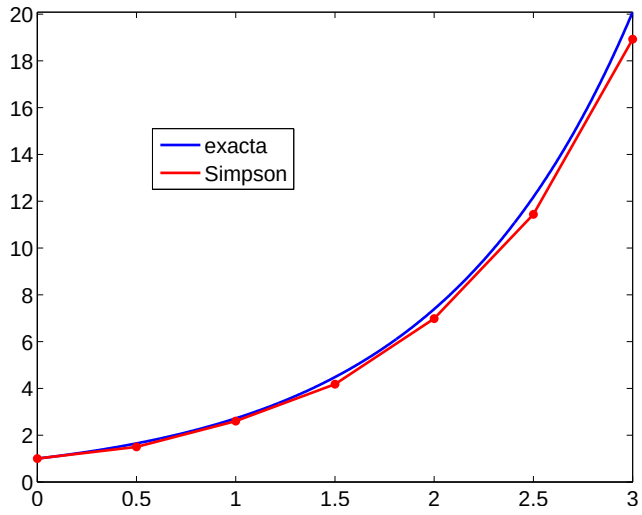


Figura 10.2: Solución aproximada por el método de Simpson

o equivalentemente

$$x_{n+1} = \frac{1}{5}(7x_{n-1} + 4x_n).$$

La solución generada

$$\left\{1, \frac{3}{2}, \frac{13}{5}, \frac{209}{50}, \frac{873}{125}, \frac{14299}{1250}, \frac{59153}{3125}\right\}$$

está representada en la figura 10.2. ◇

Si se denota por $\mathbf{x}^{(k)} = (x_0, x_1, \cdots, x_k)^t$ y $\mathbf{f}^{(k)} = (f_0, f_1, \cdots, f_k)^t$, la igualdad 10.3 podría escribirse como

$$\mathbf{a} \cdot \mathbf{x}^{(k)} + h\mathbf{b} \cdot \mathbf{f}^{(k)} = 0$$

donde $\mathbf{a}$ y $\mathbf{b}$ son los vectores que tienen de componentes los $k+1$ coeficientes a_i y b_i, respectivamente. En general, si se se denota por

$$\mathbf{x}^{(n+1)} = (x_{\hat{0}+n}, x_{\hat{1}+n}, \cdots, x_{\hat{k}+n})^t \quad \text{y} \quad \mathbf{f}^{(n+1)} = (f_{\hat{0}+n}, f_{\hat{1}+n}, \cdots, f_{\hat{k}+n})^t,$$

el esquema podría escribirse en forma compacta como

$$\mathbf{a} \cdot \mathbf{x}^{(n+1)} + h\mathbf{b} \cdot \mathbf{f}^{(n+1)} = 0.$$

Por otra parte, mientras que en un esquema explícito, el cálculo de x_{n+1} solamente implica la evaluación de los restantes términos que aparecen en

Nombre	Esquema
Euler (1 etapa, explícito)	$x_{n+1} = x_n + hf_n$
Euler (1 etapa, implícito)	$x_{n+1} = x_n + hf_{n+1}$
Punto medio (2-etapas, explícito)	$x_{n+1} = x_{n-1} + 2hf_n$
Trapecios (1-etapa, implícito)	$x_{n+1} = x_n + \frac{h}{2}(f_{n+1} + f_n)$
Simpson (2-etapas, implícito)	$x_{n+1} = x_{n-1} + \frac{h}{3}(f_{n+1} + 4f_n + f_{n-1})$
Adams-Bashforth (2-etapas, explícito)	$x_{n+1} = x_n + \frac{h}{2}(3f_n - f_{n-1})$
Adams-Moulton (2-etapas, implícito)	$x_{n+1} = x_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1})$

Figura 10.3: Esquemas lineales multipaso

el esquema usando los valores precedentes de la solución, en un esquema implícito es preciso resolver una ecuación no-lineal para poder determinarlo si la función f no es lineal en la variable x.

En la tabla de la figura 10.3 se muestran los principales esquemas lineales multipaso. Por ejemplo, el método de Adams-Bashforth, que en forma vectorial se expresa como

$$(0, 1, -1) \begin{pmatrix} x_{n-1} \\ x_n \\ x_{n+1} \end{pmatrix} + h \left(-\frac{1}{2}, \frac{3}{2}, 0 \right) \begin{pmatrix} f_{n-1} \\ f_n \\ f_{n+1} \end{pmatrix} = 0,$$

es un esquema explícito ya que $b_2 = 0$, mientras que el método de Adams-Moulton

$$(0, 1, -1) \begin{pmatrix} x_{n-1} \\ x_n \\ x_{n+1} \end{pmatrix} + h \left(-\frac{1}{12}, \frac{8}{12}, \frac{5}{12} \right) \begin{pmatrix} f_{n-1} \\ f_n \\ f_{n+1} \end{pmatrix} = 0.$$

Si f es una función de Lipschitz de constante K en la variable x en $\mathbb{R}$, la ecuación en diferencias 10.2 tiene solución en el caso implícito si $h < \frac{1}{Kb_k}$. En efecto, x_{n+1} es un punto fijo de la aplicación contractiva

$$g(x) = hb_k f(t_{n+1}, x) + r_n$$

donde el término r está dado por una expresión que depende del esquema y que es independiente de x. Por ejemplo, en el método de Adams-Moulton,

en la etapa n la función g es

$$g(x) = \frac{5h}{12} f(t_{n+1}, x) + \underbrace{x_n + \frac{h}{12}(8f_n - f_{n-1})}_{r_n}.$$

Asociado a un esquema en diferencias multipaso se considera el operador diferencial

$$L(x, h, t) = \mathbf{a} \cdot \mathbf{x}(t + h) + h\mathbf{b} \cdot \frac{d\mathbf{x}}{dt}(t + h)$$

donde se conviene que $\mathbf{x}$ representa la función vectorial definida por

$$\mathbf{x}(t) = (x(t - kh), x(t - (k - 1)h), \cdots, x(t))^t,$$

para cualquier función x de clase $C^1([a, b])$

Se define el error local de truncamiento de la solución x del problema de valor inicial en el tiempo t como

$$\tau_h = L(x, h, t).$$

El adjetivo local utilizado hace referencia a lo siguiente: Si se supone que el error en las k etapas previas del cálculo de la solución x del problema de valor inicial, es nulo, es decir, si $x_{n-i} = x(t_{n-i})$ para $i = 0, 1, \cdots, k - 1$ entonces se tiene

$$\begin{aligned} \tau_h &= L(x, h, t_n) = \mathbf{a} \cdot (\mathbf{x}(t_{n+1}) - \mathbf{x}^{(n+1)}) + h\mathbf{b} \cdot (\frac{d\mathbf{x}}{dt}(t_{n+1}) - \mathbf{f}^{(n+1)}) \\ &= a_k(x(t_{n+1}) - x_{n+1}) + hb_k(f(t_{n+1}, x(t_{n+1})) - f(t_{n+1}, x_{n+1})). \end{aligned}$$

Así, en el caso de un esquema explícito ($b_k = 0$), el error local de truncamiento puede interpretarse como el error que se comete en una etapa suponiendo que el error en las etapas previas es nulo. En el caso implícito, se tiene la acotación

$$|L(x, h, t_n)| < (1 + hb_k K)|x(t_{n+1}) - x_{n+1}|. \tag{10.4}$$

Un esquema lineal de k pasos se dice que es consistente si

$$\lim_{h \to 0} \max_{k \le n \le N_h} \frac{L(x, h, t_n)}{h} = 0$$

para toda función x de clase $C^1([a, b])$. Un esquema lineal de k pasos tiene orden p si existen constantes positivas h_0 y C tales que

$$\max_{n \ge k} \left| \frac{L(x, h, t_n)}{h} \right| \le Ch^p$$

para $h < h_0$ para cualquier función $x \in C^{p+1}([a,b])$.

En principio, la verificación de que un esquema es consistente y la determinación de su orden no parece sencilla si se usan directamente las definiciones. No obstante, la aplicación del teorema de Taylor puede simplificar esta tarea como se establece en el siguiente resultado

Teorema 43 *Se representa por* $\mathbf{i}^m$ *el vector definido por*

$$\mathbf{i}^m = (k^m, (k-1)^m, \cdots, 1^m, 0^m)^t$$

para cualquier entero $m \geq 0$ *(con la determinación* $0^0 = 1$*). Un esquema multipaso definido por 10.2 es consistente si y sólo si*

$$\mathbf{a} \cdot \mathbf{i}^0 = 0, \quad -\mathbf{a} \cdot \mathbf{i}^1 + \mathbf{b} \cdot \mathbf{i}^0 = 0. \tag{10.5}$$

Un esquema multipaso consistente, definido por 10.2, tiene orden p si y sólo si

$$-\mathbf{a} \cdot \mathbf{i}^m + m\mathbf{b} \cdot \mathbf{i}^{m-1} = 0 \tag{10.6}$$

para $m = 1, \cdots, p$.

Demostración: Si se aplica el teorema de Taylor a una función x de clase $C^{p+1}([a,b])$ para aproximar el valor de $x(t_n+ih)$ y el de $\frac{dx}{dt}(t_n+ih)$, mediante el desarrollo alrededor del punto t_n, para i arbitrario, se obtiene que

$$x(t_{n+i}) = x(t_n) + ih\frac{dx}{dt}(t_n) + \cdots + \frac{(ih)^p}{p!}\frac{d^p x}{dt^p}(t_n) + \frac{(ih)^{p+1}}{(p+1)!}\frac{d^{p+1}x}{dt^{p+1}}(\xi_{ni}),$$

$$\frac{dx}{dt}(t_{n+i}) = \frac{dx}{dt}(t_n) + ih\frac{d^2x}{dt^2}(t_n) + \cdots + \frac{(ih)^{p-1}}{(p-1)!}\frac{d^p x}{dt^p}(t_n) + \frac{(ih)^p}{p!}\frac{d^{p+1}x}{dt^{p+1}}(\varsigma_{ni}),$$

para algunos valores ξ_{ni}, ς_{ni} pertenecientes a los intervalos de extremos t_n y t_{n+i}. Si se aplican estos desarrollos a $i = -k, -k+1, \cdots, 0$ y se expresan las igualdades resultantes en forma vectorial, se obtiene

$$\mathbf{x}(t_n) = x(t_n)\mathbf{i}^0 - h\frac{dx}{dt}(t_n)\mathbf{i}^1 + \cdots + \frac{(-h)^{p+1}}{(p+1)!}\frac{d^{p+1}x}{dt^{p+1}}(t_n)\mathbf{i}^{p+1} + (-h)^{p+1}\mathrm{R}_n^1,$$

$$\frac{d\mathbf{x}}{dt}(t_n) = \frac{dx}{dt}(t_n)\mathbf{i}^0 - h\frac{d^2x}{dt^2}(t_n)\mathbf{i}^1 + \cdots + \frac{(-h)^p}{p!}\frac{d^{p+1}x}{dt^{p+1}}(t_n)\mathbf{i}^p + (-h)^p\mathrm{R}_n^2,$$

donde las i-ésimas componentes de los vectores R_n^1 y R_n^2 son

$$R_{ni}^1 = \frac{i^{p+1}}{(p+1)!}\left(\frac{d^{p+1}x}{dt^{p+1}}(\xi_{ni}) - \frac{d^{p+1}x}{dt^{p+1}}(t_n))\right),$$

$$R^2_{ni} = \frac{i^p}{p!}\left(\frac{d^{p+1}x}{dt^{p+1}}(\varsigma_{ni}) - \frac{d^{p+1}x}{dt^{p+1}}(t_n)\right).$$

Si se utiliza este desarrollo en la expresión que define el operador L, se deduce

$$\begin{aligned} L(x,h,t_n) &= x(t_n)\mathbf{a}\cdot\mathbf{i}^0 - h\frac{dx}{dt}(t_n)\mathbf{a}\cdot\mathbf{i} \\ &+ \cdots + \frac{(-h)^{p+1}}{(p+1)!}\frac{d^{p+1}x}{dt^{p+1}}(t_n)\mathbf{a}\cdot\mathbf{i}^{p+1} \\ &+ h\left(\frac{dx}{dt}(t_n)\mathbf{b}\cdot\mathbf{i}^0 - h\frac{d^2x}{dt^2}(t_n)\mathbf{b}\cdot\mathbf{i}\right. \\ &+ \left.\cdots + \frac{(-h)^p}{p!}\frac{d^{p+1}x}{dt^{p+1}}(t_n)\mathbf{b}\cdot\mathbf{i}^p\right) \\ &+ (-h)^{p+1}\sum_{i=1}^{k}\left(a_{k-i}R^1_{ni} + b_{k-i}R^2_{ni}\right). \end{aligned}$$

Puesto que $\frac{d^{p+1}x}{dt^{p+1}}$ es uniformemente continua en I, el último sumando de la igualdad anterior puede ser sustituido por $O(h^{p+1})$. En consecuencia, se tiene

$$\begin{aligned} \frac{L(x,h,t_n)}{h^{p+1}} &= \frac{x(t_n)}{h^{p+1}}\mathbf{a}\cdot\mathbf{i}^0 \\ &+ \frac{\frac{dx}{dt}(t_n)}{h^p}\left(-\mathbf{a}\cdot\mathbf{i} + \mathbf{b}\cdot\mathbf{i}^0\right) \\ &+ \cdots \\ &+ (-1)^p\frac{\frac{1}{p!}\frac{d^px}{dt^p}(t_n)}{h}\left(-\frac{1}{p+1}\mathbf{a}\cdot\mathbf{i}^{p+1} + \mathbf{b}\cdot\mathbf{i}^p\right) \\ &+ \frac{O(h^{p+1})}{h^{p+1}}. \end{aligned}$$

Puesto que x es una función arbitraria, los valores de $x(t_n)$ y $\frac{dx}{dt}(t_n)$ son arbitrarios. En consecuencia, para que $\frac{L(x,h,t_n)}{h^{p+1}}$ esté acotado (que el esquema sea de orden p) es necesario y suficiente es que se cumplan las condiciones 10.6. En particular para $p = 1$ se obtienen las condiciones de consistencia. ◇

– EJERCICIO 111 *Determinar las constantes α y β para que el esquema*

$$x_{n+1} = x_n + \frac{h}{2}(\alpha f_{n+1} + \beta f_n)$$

sea consistente y su orden sea máximo.

Solución: Con el fin de organizar los cálculos, se organiza la tabla

$\mathbf{a}$	$\mathbf{i}^0$	$\mathbf{i}^1$	$\mathbf{i}^2$
1	1	1	1
-1	1	0	0
$\mathbf{a}\cdot\mathbf{i}^m$	0	1	1

$\mathbf{b}$	$\mathbf{i}^0$	$\mathbf{i}^1$	$\mathbf{i}^2$
$\frac{\beta}{2}$	1	1	1
$\frac{\alpha}{2}$	1	0	0
$\mathbf{b}\cdot\mathbf{i}^m$	$\frac{\alpha+\beta}{2}$	$\frac{\beta}{2}$	$\frac{\beta}{2}$

De acuerdo con el teorema anterior, una condición necesaria y suficiente para que el esquema 1-etapa sea consistente es que los coeficientes α y β verifiquen la siguiente igualdad

$$-1+\frac{\alpha+\beta}{2}=0.$$

Para que el esquema sea de orden 2 es necesario y suficiente que se cumpla

$$-1+\beta=0.$$

Consecuentemente, se tiene

$$\alpha=\beta=1.$$

Para que el esquema fuese de orden 3 sería necesario y suficiente que se cumpliese

$$-1+\frac{3}{2}\beta=0,$$

lo que no ocurre. El esquema obtenido es el que corresponde al método de los trapecios (ó de Crank-Nicolson). ◇

Otra interpretación distinta del concepto de consistencia, la da el siguiente

▁ TEOREMA 44 *Un esquema lineal multipaso es consistente si y solo si es exacto cuando se aplica a los siguientes problemas de valor inicial*

$$\begin{aligned}\frac{dx}{dt} &= 0, \quad x(0)=1,\\ \frac{dx}{dt} &= 1, \quad x(0)=0.\end{aligned}$$

Demostración: Puesto que la solución del primer problema de valor inicial es $x(t)=1$, el esquema produce una solución que coincide con ella si y solo si

$$\mathbf{a}\cdot\mathbf{x}^{(n+1)}+h\mathbf{b}\cdot\mathbf{f}^{(n+1)}=\mathbf{a}\cdot\mathbf{i}^0=0.$$

Puesto que la solución del segundo problema de valor inicial es $x(t)=t$, si el esquema produce una solución que coincide con ella, entonces

$$\mathbf{x}^{(n+1)}=(n+1)h\mathbf{i}^0-h\mathbf{i}^1.$$

En consecuencia el esquema genera la solución exacta del segundo problema de valor inicial si y solo si

$$\mathbf{a} \cdot \mathbf{x}^{(n+1)} + h\mathbf{b} \cdot \mathbf{f}^{(n+1)} = -h\mathbf{a} \cdot \mathbf{i}^1 + h\mathbf{b} \cdot \mathbf{i}^0 = 0. \qquad \diamond$$

10.4 Estabilidad de los métodos multipaso

Un esquema en diferencias transforma un problema de valor inicial para una ecuación diferencial, en un problema de valor inicial para una ecuación en diferencias. Es razonable pensar que una propiedad que debe conservar la ecuación en diferencias, es la estabilidad que pudiera tener la ecuación diferencial. En este sentido, la estabilidad de un esquema numérico para la aproximación de la solución de un problema de valor inicial, se entiende como la estabilidad de la ecuación en diferencias asociada para valores iniciales arbitrarios. No obstante, la relación entre el tamaño del paso h y el índice general n, hace más complejo el concepto de estabilidad y por ello se introducirán algunos matices. En un intervalo acotado, n puede crecer indefinidamente manteniendo $t_n = a + nh$ constante si h tiende acompasadamente a 0. En un intervalo que no está acotado n puede crecer indefinidamente acompasadamente con $t_n = a + nh$ manteniendo h constante. En esta sección se considerará únicamente el caso acotado.

Un esquema multipaso definido por 10.2, se dice que es cero-estable en el problema de valor inicial

$$\frac{dx}{dt} = f(t, x(t)), \qquad x(t_0) = x_0,$$

si para todo $\epsilon > 0$ existen constantes positivas C y h_0 tales para todo $0 < h < h_0$ se cumple que

$$|z_n - x_n| \leq C\epsilon$$

para todo $n \leq N_h$ siendo $N_h = \max\{n : t_n \in I\}$ y donde z_n representa la solución del sistema perturbado

$$z_{n+1} = \sum_{i=0}^{k-1} a_i z_{\hat{i+n}} + h \sum_{i=0}^{k} b_i f_{\hat{i+n}} + h\delta_{n+1},$$

con las condiciones iniciales perturbadas

$$z_j = x_j + \delta_j,$$

para $j = 0, \cdots, k-1$, siempre que las perturbaciones verifiquen que $|\delta_n| \leq \epsilon$ para todo $n \geq 0$.

Como se ha comentado anteriormente, un esquema multipaso no genera los valores $x_1, x_2, \cdots, x_{k-1}$ y ello obliga a aproximarlos por otros esquemas de menos pasos. El concepto de cero-estabilidad permite determinar si los errores que se producen en estas aproximaciones (que se reflejan en δ_j, $j = 0, \cdots, k-1$) y los que se producen en el curso de los cálculos (que se reflejan en δ_j, $j \geq k$), vuelven inestables los cálculos posteriores.

▪ EJEMPLO **51** Sea α una constante positiva. Se considera el siguiente esquema

$$x_{n+1} - (1-\alpha)x_n - \alpha x_{n-1} = (1+\alpha)hf_n \tag{10.7}$$

para aplicarlo a la resolución del problema de valor inicial trivial

$$\frac{dx}{dt} = 0, \quad x(0) = 0. \tag{10.8}$$

Como el método tiene dos pasos, es preciso usar un método de un paso para aproximar x_1. Si se supone que el error cometido en esta aproximación es ϵ entonces la resolución del problema de valor inicial

$$x_{n+1} - (1-\alpha)x_n - \alpha x_{n-1} = 0, \qquad x_0 = 0, \ x_1 = \epsilon$$

proporciona la solución aproximada por el esquema. Fácilmente se comprueba que se puede expresar la solución como

$$x_n = \frac{\epsilon}{1+\alpha}\left(1 - \epsilon(-\alpha)^n\right).$$

En la figura 10.4 se comprueba que para un valor de $\alpha = 1.05$, el esquema no es cero-estable. ◇

Como ha ocurrido con el concepto de consistencia, el concepto de cero-estabilidad es difícil de utilizar manejando directamente su definición. Afortunadamente, también aquí es posible desarrollar una técnica algebraica que permite garantizar en determinadas circunstancias la cero-estabilidad de un esquema.

Un esquema multipaso definido por 10.2 y aplicado al ejemplo de prueba

$$\frac{dx}{dt} = \lambda x, \quad x(0) = x_0. \tag{10.9}$$

produce la ecuación lineal en diferencias finitas

$$x_{n+1} = \sum_{i=0}^{k-1} a_i x_{\hat{i}+n} + h\lambda \sum_{i=0}^{k} b_i x_{\hat{i}+n}. \tag{10.10}$$

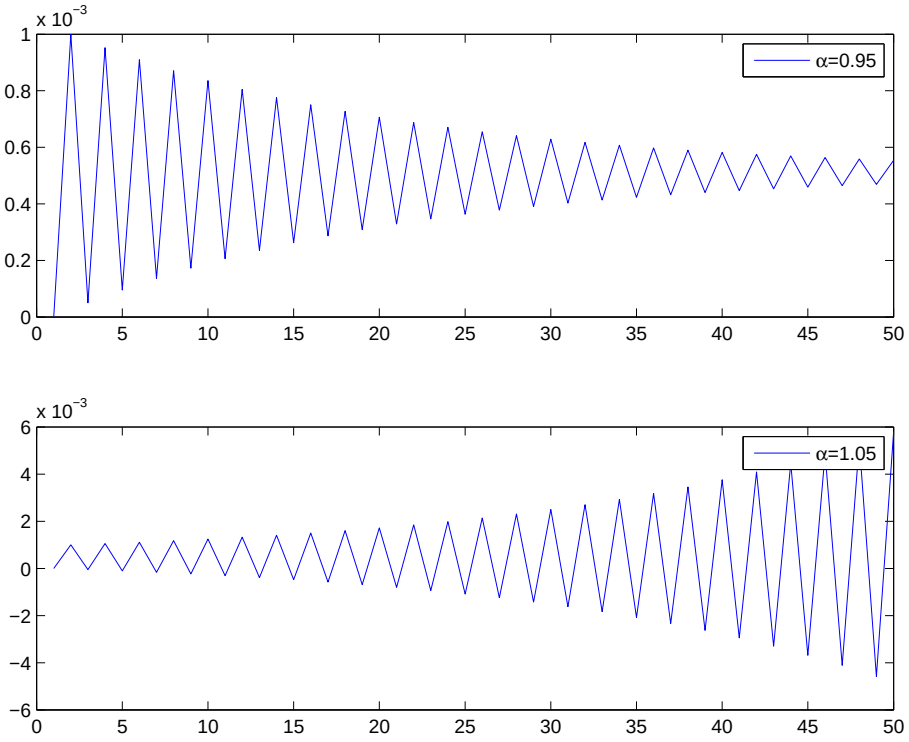


Figura 10.4: Esquema 10.7 para $\frac{dx}{dt} = 0,\ x(0) = 0$

El polinomio característico asociado a esta ecuación en diferencias

$$\Pi(r) = -\sum_{i=0}^{k} \left(a_i r^i + h\lambda b_i r^i\right)$$

puede expresarse como

$$\Pi(r) = \rho(r) - h\lambda\sigma(r),$$

donde

$$\begin{aligned} \rho(r) &= -\sum_{i=0}^{k} a_i r^i = r^k - \sum_{i=0}^{k-1} a_i r^i, \\ \sigma(r) &= \sum_{i=0}^{k} b_i r^i. \end{aligned}$$

Los polinomios ρ y σ se conocen respectivamente como primer y segundo polinomio característico del esquema multipaso.

Se dice que el esquema multipaso considerado cumple la condición de la raíz si la ecuación lineal en diferencias que se genera cuando se aplica el esquema al ejemplo modelo con $\lambda = 0$ es estable, es decir, si las raíces del primer polinomio característico tienen módulo menor o igual que 1 y aquellas que tienen módulo 1, son simples. Nótese que si el esquema es consistente, el primer polinomio característico siempre tiene a 1 como raíz.

Las soluciones exactas del ejemplo modelo con $\lambda = 0$ son las soluciones constantes $x(t) = x_0$. De este modo, un esquema que cumpla la condición de la raíz, produce soluciones acotadas cuando se aplica a la ecuación diferencial $\frac{dx}{dt} = 0$.

▪ EJEMPLO **52** El siguiente esquema en diferencias

$$x_{n+1} = \frac{x_n}{2} + x_{n-1} + \frac{3}{2}hf_n$$

tiene como primer polinomio característico

$$\rho(r) = r^2 - \frac{r}{2} - 1,$$

cuyas raíces son

$$r = \frac{1}{4}(1 \pm \sqrt{17}).$$

Consecuentemente, no cumple la condición de la raíz. De hecho, si se aplica este esquema al problema de valor inicial

$$x(t_0) = 0, \qquad \frac{dx}{dt} = 0$$

genera una ecuación en diferencias

$$x_{n+1} - \frac{x_n}{2} - x_{n-1} = 0$$

cuya solución general es

$$x_n = c_1 \frac{1}{4^n}(1 + \sqrt{17})^n + c_2 \frac{1}{4^n}(1 - \sqrt{17})^n.$$

Si se impone la condición inicial se obtiene que

$$x_n = c_1 \frac{1}{4^n}(1 + \sqrt{17})^n - c_1 \frac{1}{4^n}(1 - \sqrt{17})^n.$$

Es decir, como todos los métodos de más de un paso no determinan completamente la solución que queda dependiendo del parámetro arbitrario c_1

hasta que se añada un método de un paso que permita determinar x_1 y de este modo el esquema pueda arrancar. Si se comete un error ϵ en la aproximación de x_1 por un esquema de un paso, la solución del esquema multipaso sería

$$x_n = \frac{2\epsilon}{\sqrt{17}}\left(\left(\frac{1+\sqrt{17}}{4}\right)^n - \left(\frac{1-\sqrt{17}}{4}\right)^n\right).$$

Si bien el segundo sumando permanece acotado, por el contrario el primer sumando tiende a infinito y la sucesión no está acotada. ◇

La condición de la raíz puede parecer una condición débil, pero no lo es si se combina con la condición de consistencia. En el caso de un intervalo acotado, el hecho de que h tienda a 0, hace que la contribución del segundo polinomio sea pequeña y que la condición de consistencia permita controlar el error. Este resultado como se establece en el siguiente

▬ TEOREMA 45 *Para un esquema lineal multipaso consistente con el problema de valor inicial*

$$x(t_0) = x_0, \quad \frac{dx}{dt} = f(t, x(t)), \quad \textit{para todo } t \in I,$$

la condición de la raíz es equivalente a la cero-estabilidad.

La prueba de este resultado no es simple. En la referencia ([8, Quarteroni,Sacco,Saleri], página 496) se puede encontrar una demostración rigurosa de este teorema.

▪ EJEMPLO **53** Se considera el siguiente ejemplo de prueba

$$\frac{dx}{dt} = \lambda x, \quad x(0) = x_0, \tag{10.11}$$

en el intervalo $I = [0, 1]$. Si se aplica el esquema del punto medio a esta ecuación diferencial se obtiene la siguiente ecuación en diferencias

$$x_{n+1} - 2\lambda h x_n - x_{n-1} = 0.$$

El primer polinomio característico del esquema es

$$\rho^2 - 1 = 0.$$

Las dos raíces de este polinomio son ± 1 y por lo tanto, aunque tienen valor absoluto igual a 1, son simples. Por ello, el esquema cumple la condición de la raíz. Por otra parte, el esquema es consistente con el problema de valor inicial (véase ejercicio 115). De acuerdo con el teorema anterior, el esquema es cero-estable. ◇

Más exigente es la condición que se define a continuación. Se dice que el esquema multipaso cumple la condición fuerte de la raíz si las raíces del primer polinomio característico tienen módulo menor que 1 salvo una de ellas que es simple e igual a 1. Un esquema que cumpla la condición fuerte de la raíz, produce soluciones que convergen a x_0 si $n \to \infty$, cuando se aplica a la ecuación diferencial $\frac{dx}{dt} = 0$.

10.5 Convergencia de los métodos multipaso

▪ EJEMPLO **54** Se pretende calcular una solución aproximada del problema de valor inicial

$$x(t_0) = 1, \qquad \frac{dx}{dt} = -x, \quad \text{para todo } t \in [0,1]$$

usando el esquema de Adams-Bashforth. En este caso el esquema genera la siguiente ecuación en diferencias

$$x_{n+1} - \left(1 - \frac{3h}{2}\right) x_n - \frac{h}{2} x_{n-1} = 0.$$

La solución general de esta ecuación es

$$x_n = c_1 \lambda_1^n + c_2 \lambda_2^n$$

donde las raíces λ_1 y λ_2, ordenadas de mayor a menor, se corresponden con los valores de la siguiente expresión

$$\frac{1}{2}\left(1 - \frac{3h}{2} \pm \sqrt{\left(1 - \frac{3h}{2}\right)^2 + 2h}\right).$$

Si se impone la condición inicial, se obtiene que las soluciones que produce este método de dos pasos, son todas las que tienen la forma

$$x_n = c_1 \lambda_1^n + (1 - c_1)\lambda_2^n$$

o equivalentemente

$$x_n = \frac{x_1 - \lambda_2}{\lambda_1 - \lambda_2}(\lambda_1^n - \lambda_2^n) + \lambda_2^n.$$

La ecuación en diferencias genera una familia uniparamétrica de soluciones mientras que la solución exacta $x(t) = e^{-t}$ es única. De hecho, si se escoge

$x_1 = \lambda_2$ la solución $x_n = \lambda_2^n$ no se aproxima a ella mientras que si se escoge $x_1 = \lambda_1$, la solución $x_n = \lambda_1^n$ si que converge a la exacta. Así pues, la convergencia de un esquema multipaso está condicionada por la elección de los r primeros pasos. Es decir, la elección del esquema secundario que permita obtener los primeros pasos que permitan arrancar al método multipaso, puede condicionar la convergencia del método, como se manifiesta en este ejemplo.

Por otra parte, es conveniente precisar en qué sentido la solución aproximada x_n converge a la solución exacta. Parece natural pensar que para un valor del tiempo $t \in (0, 1]$, si se escoge un tamaño de paso h tal que $t = nh$ entonces $\lim_{n\to\infty} x_n = x(t)$. En el ejemplo en consideración se cumple que

$$\lim_{n\to\infty} \lambda_1^n = \lim_{n\to\infty} \frac{1}{2^n}\left(1 - \frac{3x}{2n} + \sqrt{\left(1 - \frac{3x}{2n}\right)^2 + 2\frac{x}{n}}\right) = e^{-x}.$$

para todo $x \in (0, 1]$. Alternativamente, la condición de convergencia se puede expresar como

$$\lim_{h\to 0} \max_{0\le n\le N_h} |e^{-t_n} - \lambda_1^n| = 0. \qquad \diamond$$

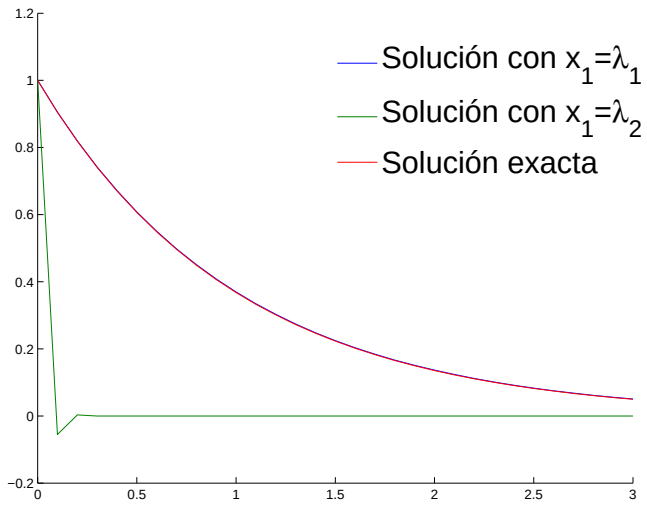


Figura 10.5: Soluciones aproximadas por el método de Adams-Bashforth

Si x es la solución exacta de un problema de valor inicial

$$x(t_0) = x_0, \qquad \frac{dx}{dt} = f(t, x(t)), \quad \text{para todo } t \in I,$$

se puede comparar con la solución aproximada $\{x_n\}$ generada por el esquema multipaso, si se evalúa en los puntos de la red $\{t_n\}$ construida con el tamaño de paso h. Se representa por $e_n^h = x(t_n) - x_n$ el error que se produce al aproximar la solución x por la solución x_n que produce el esquema multipaso definido por 10.2. El esquema se dice convergente si el error tiende a 0, cuando se usa un método de arranque convergente. De un modo más preciso, el esquema se dice convergente si

$$\lim_{h\to 0} \max_{0\le i\le k} |e_i^h| = 0 \quad \Leftrightarrow \quad \lim_{h\to 0} \max_{0\le i\le N_h} |e_i^h| = 0$$

y el esquema se dice convergente con orden q si

$$\lim_{h\to 0} \max_{0\le i\le k} |e_i^h| = O(h^q) \quad \Leftrightarrow \quad \lim_{h\to 0} \max_{0\le i\le N_h} |e_i^h| = O(h^q).$$

De la desigualdad 10.4 se deduce que si un esquema es convergente entonces

$$\lim_{h\to 0} \max_{k\le n\le N_h} \left|\frac{L(x,h,t_n)}{h}\right| \le \lim_{h\to 0} \max_{k\le n\le N_h} (1 + hb_kK)|e_{n+1}^h| = 0$$

donde K es una constante de Lipschitz de f respecto a la variable x. Es decir, una condición necesaria para que un esquema sea convergente, es que sea consistente.

La convergencia de un método multipaso puede ser garantizada simplemente verificando algunas propiedades relativas a las raíces de los polinomios característicos como establece el siguiente

▬ Teorema 46 *Un método multipaso definido por 10.2 es convergente si y sólo si es consistente y satisface la condición de la raíz. Además, si el esquema es de orden q, es convergente con orden q.*

En la referencia ([8, Quarteroni,Sacco,Saleri],página 498) se puede encontrar una demostración rigurosa de este teorema.

▬ Ejercicio 112 *Estudiar la convergencia del siguiente esquema numérico:*

$$x_{n+1} - x_{n-1} = \frac{h}{5}(3f(t_n, x_n) - 2f(t_{n-1}, x_{n-1})).$$

El primer polinomio característico es

$$\rho(r) = r^2 - 1$$

cuyas raíces son $r = \pm 1$. Así pues, el esquema cumple la condición de la raíz. Por otra parte, para analizar la consistencia se construye la tabla

$\mathbf{a}$	$\mathbf{i}^0$	$\mathbf{i}^1$	$\mathbf{i}^2$
1	1	2	4
0	1	1	1
-1	1	0	0
$\mathbf{a}\cdot\mathbf{i}^m$	0	2	4

$\mathbf{b}$	$\mathbf{i}^0$	$\mathbf{i}^1$	$\mathbf{i}^2$
$-\frac{2}{5}$	1	2	4
$\frac{3}{5}$	1	1	1
0	1	0	0
$\mathbf{b}\cdot\mathbf{i}^m$	$\frac{1}{5}$	$-\frac{1}{5}$	-1

Este esquema verifica

$$\mathbf{a}\cdot\mathbf{i}^0 = 0, \quad -\mathbf{a}\cdot\mathbf{i}^1 + \mathbf{b}\cdot\mathbf{i}^0 = -\frac{9}{5}$$

lo que prueba que el esquema no consistente y por lo tanto no es convergente. ◇

Otra cuestión importante en orden a seleccionar un método multipaso para resolver un problema de valor inicial es la de determinar el esquema convergente con el máximo orden cuando se se fija el número de pasos. Si se cuentan coeficientes y condiciones de orden se puede pensar que el orden más alto posible para un método lineal de k pasos es $2k$. Desafortunadamente, este hecho es incompatible con la estabilidad para $k > 1$. El siguiente teorema establece que cuando esto es posible

– TEOREMA 47 *(Primera barrera de Dahlquist) El orden más alto de un método de k pasos 0-estable es $k+1$ si k es impar y $k+2$ si k es par.*

10.6 Estabilidad en intervalos que no están acotados

Se considera la situación en la que el intervalo $I = [a, \infty)$ sobre el que se quiere aproximar la solución, no está acotado. Obviamente, el tamaño de paso no podrá tomarse excesivamente pequeño si lo que se busca es aproximar la solución para un valor elevado de t. En este caso, aunque el valor de h se fije razonablemente pequeño, el valor de N_h crece a infinito. En esta situación, el concepto de cero-estabilidad no parece el más adecuado.

Se dice que el método multipaso cumple la condición absoluta de la raíz si existe $h_0 > 0$ tal que todas las raíces de polinomio característico Π tienen módulo menor que 1 para $h < h_0$. De acuerdo con los resultados de estabilidad del capítulo precedente, la condición absoluta de estabilidad equivale a que la ecuación en diferencias que el esquema produce para el ejemplo de prueba sea fuertemente estable con independencia del método auxiliar utilizado para el arranque del método multipaso.

▬ EJERCICIO 113 *Estudiar la estabilidad absoluta del esquema de Simpson.*

Solución: El primer y segundo polinomio de estabilidad del esquema son

$$\rho(r) = r^2 - 1, \quad \sigma(r) = \frac{1}{3}(r^2 + 4r + 1)$$

y el polinomio característico es

$$\Pi(r) = \left(\frac{\bar{h}}{3} - 1\right) r^2 + \frac{4}{3}\bar{h}r + \left(\frac{\bar{h}}{3} + 1\right)$$

para $\bar{h} = \lambda h$. Las raíces del polinomio característico son

$$r = \frac{2\bar{h} \pm \sqrt{9 + 3\bar{h}^2}}{3 - \bar{h}}.$$

La condición $|r| < 1$ no se verifica para ambas raíces cualquiera que sea $\bar{h} = \lambda h < 3$. Consecuentemente, el esquema no tiene ningún intervalo de estabilidad absoluta. ◇

10.7 Ejercicios

▬ EJERCICIO 114 *Se considera el siguiente problema de valor inicial:*

$$\frac{dx}{dt} = tx, \quad x(0) = 1.$$

Comparar la solución obtenida en $t = 0.5$ al usar un método de Euler implícito con paso $h = 0.1$ y la solución exacta.

Solución: La solución exacta del problema de valor inicial es $x(t) = e^{\frac{t^2}{2}}$. El valor aproximado mediante el método de Euler implícito se calcula por el esquema

$$\frac{x_{n+1} - x_n}{h} = t_{n+1} x_{n+1}$$

hasta $n = 4$. Si se tiene en cuenta que $t_{n+1} = (n+1)h$, se obtiene que

$$x_{n+1} = \frac{1}{1 - (n+1)h^2} x_n, \quad x_0 = 1,$$

Consecuentemente

$$x_5 = \frac{1}{(1 - 5h^2)(1 - 4h^2)(1 - 3h^2)(1 - 2h^2)(1 - h^2)} = 1.1651$$

mientras que el valor exacto es $x(0.5) = e^{0.125} = 1.1331$. ◇

▬ EJERCICIO 115 *Determinar los valores de α para los que el siguiente esquema*

$$x_{n+1} - (1-\alpha)x_n - \alpha x_{n-1} = (1+\alpha)hf_n$$

sea consistente con el siguiente problema de valor inicial

$$x(t_0) = x_0, \quad \frac{dx}{dt} = f(t, x(t)), \quad \textit{para todo } t \in I.$$

Solución: La tabla de orden asociada a este esquema es

$\mathbf{a}$	$\mathbf{i}^0$	$\mathbf{i}^1$	$\mathbf{i}^2$
α	1	2	4
$1-\alpha$	1	1	1
-1	1	0	0
$\mathbf{a}\cdot\mathbf{i}^m$	0	$1+\alpha$	$1+3\alpha$

$\mathbf{b}$	$\mathbf{i}^0$	$\mathbf{i}^1$
0	1	2
$1+\alpha$	1	1
0	1	0
$\mathbf{b}\cdot\mathbf{i}^m$	$1+\alpha$	$1+\alpha$

El esquema es consistente para todo valor de α y es de orden 2 en el caso $\alpha = 1$. ◇

▬ EJERCICIO 116 *Determinar entre todos los esquemas lineales de un paso que pueden usarse para resolver un problema de valor inicial, aquellos que tengan orden máximo.*

Solución: El número de coeficientes de un esquema de un paso es 4 (aunque se sabe que el coeficiente $a_1 = -1$). Es posible pensar que los coeficientes verifican las ecuaciones

$$\begin{aligned} \mathbf{a}\cdot\mathbf{i}^0 &= 0, \\ -\mathbf{a}\cdot\mathbf{i}^1 + \mathbf{b}\cdot\mathbf{i}^0 &= 0, \\ -\mathbf{a}\cdot\mathbf{i}^2 + 2\mathbf{b}\cdot\mathbf{i}^1 &= 0, \\ -\mathbf{a}\cdot\mathbf{i}^3 + 3\mathbf{b}\cdot\mathbf{i}^2 &= 0 \end{aligned}$$

que garantizarían que los esquemas obtenidos son de orden 3. Si estas ecuaciones se expresan en forma matricial se obtiene

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 1 \\ -1 & 0 & 2 & 0 \\ -1 & 0 & 3 & 0 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Puesto que $a_1 = -1$, el sistema se reduce a

$$\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 1 \\ -1 & 2 & 0 \\ -1 & 3 & 0 \end{pmatrix} \begin{pmatrix} a_0 \\ b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

El sistema lineal es incompatible. Consecuentemente, si excluye la última ecuación (con lo que el orden se reduce a 2), se obtiene un sistema compatible determinado cuya única solución es $a_0 = 1$, $b_0 = b_1 = \frac{1}{2}$. ◇

▬ EJERCICIO 117 *Estudiar la consistencia, cero-estabilidad y convergencia de los siguientes esquemas*

1. $$x_{n+1} - x_n = \frac{h}{2}(3f_{n+1} - f_n),$$

2. $$x_{n+1} - x_n = \frac{h}{5}(3f_{n+1} - 2f_n),$$

3. $$x_{n+1} + x_n - 2x_{n-1} = \frac{h}{4}(9f_{n+1} + 3f_n)$$

para resolver el problema de valor inicial

$$x(t_0) = x_0, \quad \frac{dx}{dt} = f(t, x(t)), \quad \textit{para todo } t \in I.$$

Solución: Si se usan los criterios algebraicos de consistencia y estabilidad, así como el teorema 46, se deduce

	Consistencia	Cero-estabilidad	Convergencia
1)	sí	sí	sí
2)	no	Criterio de la raíz: sí	no
3)	sí	no	no

◇

▬ EJERCICIO 118 *Los métodos de Adams-Moulton de k pasos cumplen que*

$$a_k = -1, a_{k-1} = 1, a_{k-2} = \cdots = a_0 = 0$$

y son de orden $k + 1$. Determinar los coeficientes del método de Adams-Moulton de 3 pasos.

Solución: La tabla de orden asociada a este esquema es

$\mathbf{a}$	$\mathbf{i}^0$	$\mathbf{i}^1$	$\mathbf{i}^2$	$\mathbf{i}^3$	$\mathbf{i}^4$
0	1	3	9	27	81
0	1	2	4	8	16
1	1	1	1	1	1
-1	1	0	0	0	0
$\mathbf{a}\cdot\mathbf{i}^m$	0	1	1	1	1

$\mathbf{b}$	$\mathbf{i}^0$	$\mathbf{i}^1$	$\mathbf{i}^2$	$\mathbf{i}^3$
b_0	1	3	9	27
b_1	1	2	4	8
b_2	1	1	1	1
b_3	1	0	0	0
$\mathbf{b}\cdot\mathbf{i}^m$	$b_0+b_1+b_2+b_3$	$3b_0+2b_1+b_2$	$9b_0+4b_1+b_2$	$27b_0+8b_1+b_2$

Si se resuelve el sistema lineal

$$\begin{aligned} b_0+b_1+b_2+b_3 &= 1\\ 3b_0+2b_1+b_2 &= \frac{1}{2}\\ 9b_0+4b_1+b_2 &= \frac{1}{3}\\ 27b_0+8b_1+b_2 &= \frac{1}{4} \end{aligned}$$

se obtiene

$$b_0=\frac{1}{24},\quad b_1=-\frac{5}{24},\quad b_2=\frac{19}{24},\quad b_3=\frac{3}{8}.\qquad \diamond$$

CAPÍTULO 11

Problemas de contorno para ecuaciones diferenciales

11.1 Introducción

Una solución de una ecuación diferencial puede quedar determinada por datos relativos al instante inicial $t = 0$. No obstante, en los llamados problemas de contorno, se selecciona una solución imponiendo una o varias condiciones en ambos extremos del intervalo de interés. No parece sencillo hacer un planteamiento general de esta clase de problemas, razón por la que el estudio de estas cuestiones en este capítulo se limitará al siguiente problema: Hallar $u : [a, b] \to \mathbb{R}$ tal que

$$-\frac{d}{dx}\left(\alpha(x)\frac{du}{dx}\right) + \beta(x)\frac{du}{dx} = f(u, x)$$

y verifique las condiciones de contorno $u(a) = u_a$ y $u(b) = u_b$. Las funciones $\alpha, \beta : [a, b] \to \mathbb{R}$ y $f : \mathbb{R} \times [a, b] \to \mathbb{R}$ son funciones continuas que forman parte de los datos conocidos del problema.

El operador diferencial asociado a esta ecuación

$$L = -\frac{d}{dx}\left(\alpha(x)\frac{d}{dx}\right) + \beta(x)\frac{d}{dx}$$

es lineal aunque la ecuación completa puede ser no-lineal si la función f no es un polinomio de primer orden en u.

El cambio de notaciones respecto al capítulo anterior no es caprichoso, sino que se debe al interés en usar la variable independiente x por su significado espacial en muchos modelos sobre los que se formula el problema de contorno. Esto contrasta con los problemas de valor inicial para los que se ha usado la variable t por su habitual significado temporal. Muchos de los modelos sobre los que se plantea un problema de contorno representan un equilibrio estático frente las dinámicas que representan los modelos sobre los que se formulan los problemas de valor inicial.

En los problemas de valor inicial para ecuaciones diferenciales ordinarias, la regularidad de las funciones que intervienen en la ecuación es suficiente para garantizar la existencia y unicidad de la solución. No es esta la situación que ocurre en los problemas de contorno para ecuaciones diferenciales. Para comprender que la naturaleza de ambos problemas es distinta, se puede hacer el siguiente análisis: Sea $v(x;s)$ la única solución del problema lineal

$$-\frac{d}{dx}\left(\alpha(x)\frac{dv}{dx}\right)+\beta(x)\frac{dv}{dx}+\gamma(x)v=f(x)$$

verificando las condiciones iniciales $v(a;s)=1$ y $v'(a;s)=s$. Por otra parte, sea $w(x)$ la única solución de la ecuación diferencial homogénea con las condiciones iniciales $w(a)=0$ y $w'(a)=1$. A continuación se intenta construir una solución del problema de contorno $u(a)=u_a$ y $u(b)=u_b$, como una combinación lineal $u=u_av+sw$. Del hecho de que la ecuación sea lineal se desprende que esta combinación lineal es una solución del problema de contorno si se verifica que

$$u_av(b;s)+sw(b)=u_b.$$

Obviamente, puede ocurrir que esta ecuación numérica de variable s no tenga solución o tenga una infinidad de ellas.

▪ Ejemplo **55** Se considera el problema de contorno

$$u''+\lambda^2u=0,\qquad u(0)=u(1)=1.$$

La solución general de la ecuación diferencial es

$$u(x)=c_1\cos(\lambda x)+c_2\,\mathrm{sen}(\lambda x).$$

Sin más que imponer las condiciones de contorno, se puede determinar la solución del problema de contorno. No obstante se usará este ejemplo para ilustrar el procedimiento teórico descrito anteriormente, que se basa en la resolución de dos problemas de valor inicial y la resolución de una ecuación numérica en la variable s.

La solución del problema de valor inicial $u(0) = u_0$ y $u'(0) = u'_0$ es

$$u(x) = u_0 \cos(\lambda)x) + \frac{u'_0}{\lambda} \operatorname{sen}(\lambda x).$$

De este modo, las soluciones v y w están dadas por

$$\begin{aligned} v(x;s) &= \cos(\lambda x) + \frac{s}{\lambda} \operatorname{sen}(\lambda x), \\ w(x) &= \frac{1}{\lambda} \operatorname{sen}(\lambda x). \end{aligned}$$

Ahora, se busca una solución de la forma

$$u(x) = v(x;s) + sw(x) = \cos(\lambda x) + \frac{2s}{\lambda} \operatorname{sen}(\lambda x).$$

Para encontrar la solución al problema de contorno basta resolver la ecuación

$$u(1) = v(1;s) + sw(1) = \cos\lambda + \frac{2s}{\lambda} \operatorname{sen}\lambda = 1$$

despejando la variable s. De ello se deduce que para cualquier λ, la solución del problema de contorno es

$$u(x) = \cos(\lambda x) + \frac{1 - \cos\lambda}{\operatorname{sen}\lambda} \operatorname{sen}(\lambda x).$$

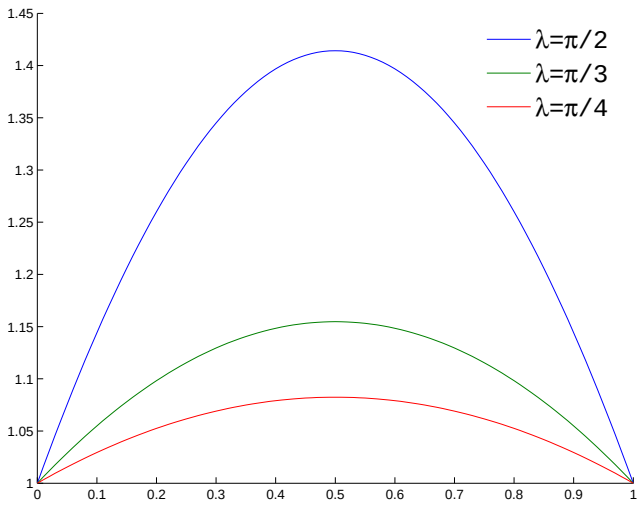


Figura 11.1: Soluciones de $u'' + \lambda^2 u = 0, \quad u(0) = u(1) = 1$

Si se considera el problema de contorno

$$u'' + \lambda^2 u = 0, \qquad u(0) = u(1) = 0$$

con condiciones de contorno homogéneas, se busca una solución de la forma

$$u(x) = sw(x) = \frac{s}{\lambda}\,\mathrm{sen}(\lambda x).$$

Para encontrar la solución al problema de contorno basta resolver la ecuación

$$u(1) = \frac{s}{\lambda}\,\mathrm{sen}\,\lambda = 0$$

despejando la variable s. Para cualquier valor de λ siempre existe la solución trivial $u = 0$. Pero, si $\mathrm{sen}\,\lambda = 0$, s puede ser escogido arbitrariamente. En otras palabras, si $\lambda = k\pi$ para algún numero entero k, la función

$$u(x) = \mathrm{sen}(k\pi x)$$

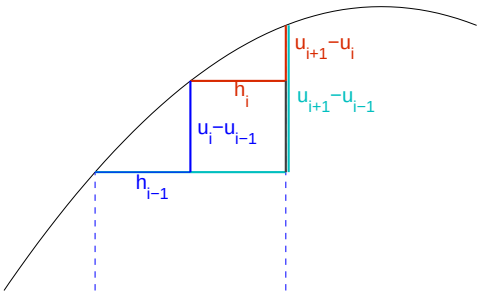
(ó cualquier múltiplo de ella) es solución del problema de contorno homogéneo. ◇

11.2 Métodos de diferencias finitas

La idea básica del método es similar a la utilizada en los problemas de valor inicial. Sea

$$a = x_0 < x_1 < \cdots < x_{N(h)} = b$$

una partición del intervalo $[a, b]$. Se representa por $h_i = x_{i+1} - x_i$, el tamaño de paso en x_i y por $u_0, u_1, \cdots, u_N$, los valores aproximados de la solución en los puntos x_i. Las derivadas que aparecen en el operador diferencial pueden ser aproximadas por diferencias divididas de alguno de los siguientes tipos



- Progresivas $\frac{u_{i+1}-u_i}{h_i}$,
- Centradas $\frac{u_{i+1}-u_{i-1}}{h_i+h_{i-1}}$,
- Retrógradas $\frac{u_i-u_{i-1}}{h_{i-1}}$.

Por ejemplo, el operador diferencial considerado en la sección anterior podría ser aproximado por el operador discreto

$$L_h u = -\frac{\alpha(x_{i+\frac{1}{2}})\frac{u_{i+1}-u_i}{h} - \alpha(x_{i-\frac{1}{2}})\frac{u_i-u_{i-1}}{h}}{h} + \beta(x_i)\frac{u_{i+1}-u_{i-1}}{2h}$$

para paso constante $h = \frac{b-a}{N}$. En esta fórmula se ha usado la siguiente notación

$$x_{i+\frac{1}{2}} = x_i + \frac{h}{2}, \quad x_{i-\frac{1}{2}} = x_i - \frac{h}{2}.$$

Esta discretización conduce al sistema de ecuaciones

$$-\frac{\alpha(x_{i+\frac{1}{2}})(u_{i+1}-u_i) - \alpha(x_{i-\frac{1}{2}})(u_i-u_{i-1})}{h^2} + \beta(x_i)\frac{u_{i+1}-u_{i-1}}{2h} = f(u_i, x_i)$$

para $i = 1, 2, \cdots, N-1$. Se trata de un sistema de $N-1$ ecuaciones con $N-1$ incógnitas $u_1, u_2, \cdots, u_{N-1}$ (u_0 y u_N son ya conocidas por las condiciones de contorno del problema) que no es lineal.

Es importante destacar la naturaleza global del problema. Es decir, no es posible determinar de modo recurrente u_{i+1} a partir de u_i como en las aproximaciones por esquemas en diferencias finitas para problemas de valor inicial. En este caso, es necesario resolver conjuntamente el sistema para determinar el valor de las incógnitas u_i.

▬ EJERCICIO 119 *Aproximar la solución del siguiente problema de contorno*

$$-\frac{d^2u}{dx^2} = 1, \qquad 0 \le x \le 1, \quad u(0) = u(1) = 0$$

usando un método de diferencia finitas con paso constante $h = \frac{1}{5}$.

Solución: Si se usa la aproximación por diferencias finitas

$$u''(x) \approx \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}$$

en

u	u_0	u_1	u_2	u_3	u_4	u_5
x	0	0.2	0.4	0.6	0.8	1

se obtiene el sistema lineal

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = 1, \qquad i = 1, \cdots, 4, \quad u_0 = u_5 = 0,$$

que en forma matricial resulta

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} \frac{1}{25} \\ \frac{1}{25} \\ \frac{1}{25} \\ \frac{1}{25} \end{pmatrix}$$

Si se usa el método de Cholesky para resolver el sistema lineal se obtienen dos sistemas lineal

$$Ly = \frac{1}{25}(1,1,1,1)^t, \qquad L^t u = y.$$

De la resolución retrógrada y progresiva de ambos sistemas triangulares se obtiene la solución

$$u_1 = \frac{2}{25},\ u_2 = \frac{3}{25},\ u_3 = \frac{3}{25},\ u_4 = \frac{2}{25}.$$

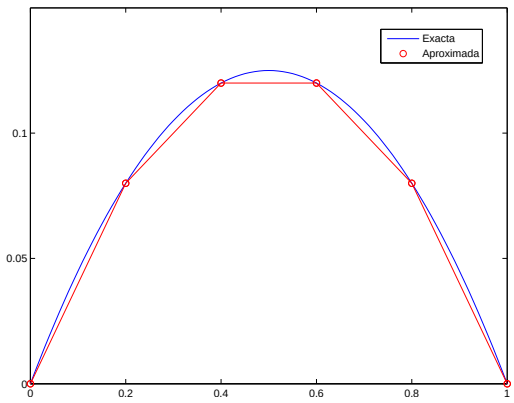


Figura 11.2: Soluciones exacta y aproximada

Como alternativa a este método matricial se podría calcular la solución aproximada del problema de contorno resolviendo la ecuación en diferencias mediante los métodos expuestos en capítulos anteriores. En este caso, puesto que la ecuación característica asociada a la ecuación en diferencias es

$$-\lambda^2 + 2\lambda - 1 = 0$$

la solución general de la homogénea es $c_1 + c_2 n$. Para encontrar una solución particular de la no-homogénea se prueba con rn^2 (aunque el grado del

segundo miembro es 0, se incrementa el grado del polinomio de prueba en 2 debido a que el operador en diferencias se anula en los polinomios de grado menor o igual que 1). De este modo se encuentra que

$$-r[(n+1)^2 - 2n^2 + (n-1)^2] = \frac{1}{25}$$

de donde se deduce que $r = -\frac{1}{50}$ y consecuentemente la solución general de la ecuación en diferencias es

$$x_n = c_1 + c_2 n - \frac{1}{50} n^2.$$

Si se imponen la condiciones de contorno se obtiene

$$x_n = \frac{n}{10}\left(1 - \frac{n}{5}\right) = \frac{1}{2} t_n (1 - t_n), \qquad n = 0, 1, \cdots, 5.$$

En este caso, la solución aproximada coincide con la exacta en los nodos. Naturalmente, este hecho es excepcional y es debido a que la derivación numérica coincide con la exacta cuando se trata de un polinomio de grado 2. Es importante destacar que el método matricial empleado en este ejercicio es mucho más eficiente que el método basado en la resolución de las ecuaciones en diferencias finitas que no resultaría simple de utilizar si los coeficientes y término independiente en la ecuación diferencial, no fuesen constantes. ◇

11.3 Análisis de la convergencia

Se considera únicamente el siguiente problema modelo: Se busca una función $u : [a, b] \to \mathbb{R}$ tal que

$$Lu \equiv -\frac{d^2 u}{dx^2} + \beta u = f$$

en (a, b) y verifique las condiciones de contorno $u(a) = u_a$ y $u(b) = u_b$. Se supone que $\beta(x) \geq 0$ para todo $x \in [a, b]$.

La solución exacta de este problema de contorno, se puede comparar con la solución aproximada $\mathbf{u}_h = \{u_i : i = 0, \cdots, N\}$ generada por el esquema en diferencias finitas,

$$L_h \mathbf{u}_h \,|_i \equiv -\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + \beta(x_i) u_i = f(x_i)$$

en los puntos de la red $\{x_i : i = 1, \cdots, N-1\}$ construida con el tamaño de paso constante h.

Con el propósito de analizar la convergencia del método, se define el vector error como

$$\mathbf{e}_h = \{u(x_i) - u_i : i = 0, 1, \cdots, N\}.$$

El esquema en diferencias finitas aplicado a la resolución del problema de contorno, se dice convergente si $\lim_{h\to 0} \|\mathbf{e}_h\|_\infty = 0$. El análisis de la convergencia de un esquema en diferencias finitas para un problema de contorno se basa generalmente en la comparación del error global de la solución y el error de discretización del operador diferencial en la solución, definido como

$$\boldsymbol{\tau}_h = L_h\mathbf{u} - L\mathbf{u}$$

donde $\mathbf{u}$ representa el vector de componentes $\{u(x_i) : i = 0, 1, \cdots, N\}$.

Si la solución u es de clase $C^4([a, b])$, se consideran sus desarrollos de Taylor alrededor de un punto x_i

$$u(x_{i-1}) = u(x_i) - \frac{du}{dx}(x_i)h + \frac{d^2u}{dx^2}(x_i)\frac{h^2}{2} - \frac{d^3u}{dx^3}(x_i)\frac{h^3}{6} + \frac{d^4u}{dx^4}(\xi_i)\frac{h^4}{24},$$

$$u(x_{i+1}) = u(x_i) + \frac{du}{dx}(x_i)h + \frac{d^2u}{dx^2}(x_i)\frac{h^2}{2} + \frac{d^3u}{dx^3}(x_i)\frac{h^3}{6} + \frac{d^4u}{dx^4}(\varsigma_i)\frac{h^4}{24},$$

para algunos puntos ξ_i y ς_i tales que $x_{i-1} \le \xi_i \le x_i$ y $x_i \le \varsigma_i \le x_{i+1}$, para todo $i = 1, 2, \cdots, N-1$.

Si se utilizan esto desarrollos en la expresión que definen el error de discretización, se obtiene

$$\begin{aligned}
\boldsymbol{\tau}_h(x_i) &= -\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} + \beta(x_i)u_i - f(x_i) \\
&= -\frac{d^2u}{dx^2}(x_i) + \beta(x_i)u_i - f(x_i) + \left(\frac{d^4u}{dx^4}(\xi_i) + \frac{d^4u}{dx^4}(\varsigma_i)\right)\frac{h^2}{24} \\
&= \left(\frac{d^4u}{dx^4}(\xi_i) + \frac{d^4u}{dx^4}(\varsigma_i)\right)\frac{h^2}{24}
\end{aligned}$$

ya que u es solución de la ecuación diferencial. De ello se deduce que

$$\|\boldsymbol{\tau}_h\|_\infty \le \frac{h^2}{12}\max_{x\in[a,b]}\left|\frac{d^4u}{dx^4}(x)\right|. \tag{11.1}$$

Por otra parte, se tiene que

$$\begin{aligned}
\boldsymbol{\tau}_h(x_i) &= -\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} + \beta(x_i)u_i - f(x_i) \\
&= -\frac{e_{i+1} - 2e_i + e_{i-1}}{h^2} + \beta(x_i)e_i
\end{aligned}$$

con la condición de contorno $e_0 = e_n = 0$, donde e_i representan las componentes de $\mathbf{e}_h$. Si se expresan estas ecuaciones en forma matricial, se obtiene

$$(\mathrm{A}^h + \mathrm{D}^h)\mathbf{e}_h = \boldsymbol{\tau}_h$$

donde A^h es la matriz de dimensiones $(N-1)\times(N-1)$, de componentes

$$\mathrm{A}^h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 2 \end{pmatrix}$$

y D^h la matriz diagonal que tiene a $\beta(x_i)$ como componente $i-$ésima en la diagonal principal, para $i = 1, \cdots, N-1$.

Para realizar el análisis de este sistema lineal se utilizará el siguiente

Lema 5 *Los autovalores de la matriz*

$$A_n = \begin{pmatrix} b & c & 0 & \cdots & 0 & 0 \\ a & b & c & \cdots & 0 & 0 \\ 0 & a & b & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b & c \\ 0 & 0 & 0 & \cdots & a & b \end{pmatrix}$$

de dimensión $n \times n$, son

$$\lambda_j = b + 2\sqrt{ac}\,\cos\frac{j\pi}{n+1} \tag{11.2}$$

para $j = 1, \cdots, n$ si a y c tienen el mismo signo.

Demostración: El polinomio característico de la matriz A_n se representará por p_n. Si se desarrolla el determinante que lo define, por la primera fila, se comprueba que p_n es solución de la ecuación en diferencias para cada λ

$$p_n(\lambda) = (b-\lambda)p_{n-1}(\lambda) - ac\; p_{n-2}(\lambda)$$

con los valores iniciales $p_1(\lambda) = b - \lambda$ y $p_2(\lambda) = (b-\lambda)^2 - ac$. La ecuación característica asociada a esta ecuación en diferencias, para cada valor de λ, es la siguiente

$$\rho^2 - (b-\lambda)\rho + ac = 0.$$

Sus autovalores son

$$\rho = \frac{b - \lambda \pm \sqrt{(b-\lambda)^2 - 4ac}}{2}.$$

En una primera etapa, se buscan los autovalores de A_n tales que

$$(b-\lambda)^2 \leq 4ac.$$

En este caso, se introduce el cambio de variable $\lambda = \lambda(\theta)$ definido por

$$\lambda = b + 2\sqrt{ac}\cos\theta$$

Si se sustituye λ en la expresión que da los autovalores de la ecuación en diferencias, se obtiene

$$\rho = -\sqrt{ac}(\cos\theta \pm i\,\mathrm{sen}\,\theta)$$

donde i representa la unidad imaginaria. La solución general de la ecuación en diferencias es

$$p_n \;=\; (-1)^n (ac)^{\frac{n}{2}} (c_1 \cos n\theta + c_2\,\mathrm{sen}\, n\theta).$$

Si se imponen las condiciones iniciales a la ecuación en diferencias se obtiene

$$p_n = (ac)^{\frac{n}{2}} (-1)^n \frac{\mathrm{sen}\ (n+1)\theta}{\mathrm{sen}\,\theta}.$$

Consecuentemente, para los valores de $\theta = \frac{j\pi}{n+1}$ para $j = 1, \cdots, n$, el polinomio característico se anula. Si se deshace el cambio de variable se obtiene la fórmula 11.2, lo que concluye la prueba ya que todos los autovalores son distintos. Es decir, no es posible encontrar autovalores que cumplan

$$(b-\lambda)^2 - 4ac > 0. \qquad \diamond$$

Si se utiliza este lema a la matriz A^h que aparece en el análisis de la convergencia del método de diferencias finitas, se deduce que sus autovalores son

$$\lambda_j^h = \frac{2}{h^2}\left(1 + \cos\frac{j\pi}{N}\right)$$

para $j = 1, \cdots, N-1$. El autovalor mínimo de A^h es

$$\lambda_{N-1}^h = \frac{2}{h^2}\left(1 + \cos\frac{(N-1)\pi}{N}\right) > 0.$$

Además, si se tiene en cuenta que $h = \frac{b-a}{N}$ y se utiliza el desarrollo de Taylor de la función cos alrededor de 0, se obtiene

$$\lambda_{N-1}^h = \frac{2}{h^2}(1 - \cos\frac{\pi}{N}) = \frac{2}{h^2}\left(\frac{\pi^2 h^2}{2(b-a)^2} + O(h^4)\right) = \frac{\pi^2}{(b-a)^2} + O(h^2). \tag{11.3}$$

y por lo tanto existen dos constantes h_0 y C, tales que

$$\frac{1}{\lambda_{N-1}^h} \leq C$$

para todo $h < h_0$.

De todo ello, se deduce que A^h es una matriz simétrica definida positiva

$$\mathrm{A}^h\mathbf{x} \cdot \mathbf{x} \geq \lambda_{N-1}^h \|\mathbf{x}\|^2$$

para todo $\mathbf{x} \in \mathbb{R}^{N-1}$.

Por otra parte D^h es una matriz semi-definida positiva que verifica

$$\mathrm{D}^h\mathbf{x} \cdot \mathbf{x} \geq 0$$

para todo $\mathbf{x} \in \mathbb{R}^{N-1}$. Finalmente se obtiene que $\mathrm{A}^h + \mathrm{D}^h$ es una matriz definida positiva que verifica

$$(\mathrm{A}^h + \mathrm{D}^h)\mathbf{x} \cdot \mathbf{x} \geq \lambda_{N-1}^h \|\mathbf{x}\|^2$$

para todo $\mathbf{x} \in \mathbb{R}^{N-1}$.

Si se aplica la desigualdad de Cauchy-Schwarz, se obtiene para la norma matricial subordinada

$$\|\mathbf{x}\| \leq \frac{1}{\lambda_{N-1}^h}\|(\mathrm{A}^h + \mathrm{D}^h)\mathbf{x}\|$$

para todo $\mathbf{x} \in \mathbb{R}^{N-1}$. Si se aplica esta desigualdad a $\mathbf{x} = \mathbf{e}_h$ se obtiene que

$$\|e_h\| \leq \frac{1}{\lambda_{N-1}^h}\|\boldsymbol{\tau}_h\| \leq C\|\boldsymbol{\tau}_h\|$$

Finalmente, si se usa la estimación del error de discretización 11.1, se prueba que

$$\|e_h\| \leq \bar{C}h^2$$

para alguna constante positiva $\bar{C}$ independiente de h, lo que prueba el siguiente

– TEOREMA 48 *El esquema en diferencias finitas aplicado al ejemplo modelo es convergente.*

11.4 Estabilidad, consistencia y convergencia

El argumento utilizado en la prueba del último teorema de la sección previa se apoya en el análisis de la ecuación $(\mathrm{A}^h + \mathrm{D}^h)\mathbf{e}_h = \tau$. Se obtuvo una acotación $\|e_h\| \leq C\|\boldsymbol{\tau}_h\|$ utilizando el hecho de que $\mathrm{A}^h + \mathrm{D}^h$ era una matriz simétrica definida positiva.

El argumento se puede utilizar en ampliar a otros problemas de contorno y a otros esquemas en diferencias finitas si $\mathrm{A}^h + \mathrm{D}^h$ no tiene determinante nulo, $\| \left(A^h + \mathrm{D}^h\right)^{-1} \|\|$ está acotada uniformemente en h y se dispone de una estimación para el error de discretización como $O(h^\beta)$ con $\beta > 0$. La idea es aplicar la propiedad principal de norma subordinada a la igualdad $\mathbf{e}_h = \left(A^h + \mathrm{D}^h\right)^{-1} \tau$. De este modo, se obtiene que

$$\|e_h\| \leq \| \left(A^h + \mathrm{D}^h\right)^{-1} \|\|\boldsymbol{\tau}_h\|$$

y el resto de la argumentación se puede seguir aplicando para establecer la convergencia del esquema en diferencias finitas.

Esto motiva la siguiente definición: Si un esquema en diferencias finitas conduce el problema de contorno a un sistema lineal de la forma $(A^h + \mathrm{D}^h)\mathbf{u}_h = \mathbf{f}^h$, entonces se dice que el esquema es estable si existen un constante positiva C y una cota para el tamaño de paso h_0 tales que para todo $h < h_0$ se cumpla

$$\| \left(A^h + \mathrm{D}^h\right)^{-1} \| \leq C.$$

Se introduce también la noción de consistencia de un esquema respecto a un problema de contorno. Un esquema es consistente con un problema de contorno lineal si

$$\lim_{h \to 0} \boldsymbol{\tau}_h \to 0.$$

Con estos conceptos se puede ampliar el resultado de la sección anterior hasta el siguiente

– Teorema 49 *Si un esquema para un problema de contorno lineal es estable y consistente con el problema entonces es convergente.*

– Ejercicio 120 *Estudiar la convergencia del siguiente esquema en diferencias finitas*

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + 2(u_{i+1} + u_{i-1}) = 1 \tag{11.4}$$

para $i = 1, \cdots, N-1$ *y*

$$u_0 = u_N = 1$$

para aproximar la solución del siguiente problema de contorno

$$-\frac{d^2u}{dx^2} = 1 - 4u, \quad 0 \leq x \leq 1, \quad u(0) = u(1) = 1$$

Solución: Si se aplica este esquema al problema de contorno se obtiene un sistema de ecuaciones lineales cuya matriz de coeficientes es la siguiente

$$\mathrm{A}^h + \mathrm{D}^h = \frac{1}{h^2} \begin{pmatrix} 2 & 2h^2-1 & 0 & \cdots & 0 & 0 \\ 2h^2-1 & 2 & 2h^2-1 & \cdots & 0 & 0 \\ 0 & 2h^2-1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & 2h^2-1 \\ 0 & 0 & 0 & \cdots & 2h^2-1 & 2 \end{pmatrix}.$$

Si se usa el lema 5 se deduce que los autovalores de la matriz $\mathrm{A}^h + \mathrm{D}^h$ son

$$\lambda_j^h = \frac{2}{h^2}\left(1 + (1-2h^2)\cos\frac{j\pi}{N}\right)$$

para $j = 1, \cdots, N-1$. Fácilmente, se prueba que

$$\lambda_{N-1}^h = 4 + \pi^2 + O(h^2)$$

lo que prueba que $\mathrm{A}^h + \mathrm{D}^h$ no tiene determinante nulo. Además, ya que es simétrica

$$\| \left(A^h + \mathrm{D}^h\right)^{-1} \| = \frac{1}{\lambda_{N-1}^h}$$

y consecuentemente está acotada. Así pues, el esquema es estable.

Por otra parte, se tiene

$$\begin{aligned} \boldsymbol{\tau}_h(x_i) &= -\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} + 2(u(x_{i+1} + u_{i-1}) - 1 \\ &= -\frac{d^2u}{dx^2}(x_i) + 4u(x_i) - 1 + O(h^2) = O(h^2) \end{aligned}$$

para $i = 1, 2, \cdots, N-1$, lo que prueba que el esquema es consistente con el problema de contorno. Si se utiliza el teorema anterior se deduce que el esquema es convergente. $\diamond$

11.5 Otras condiciones de contorno

En las secciones precedentes se han considerado únicamente condiciones de contorno que fijan los valores de la solución en los extremos. Este tipo de condiciones se conoce como de Dirichlet. Se pueden considerar otros tipos de condiciones en los extremos para seleccionar una solución. Particularmente importantes son las llamadas condiciones de Neumann que fijan los valores de la primera derivada de la función en un extremo.

▬ EJERCICIO 121 *Hallar una solución al siguiente problema de contorno*

$$u'' + \lambda^2 u = 0, \quad u(0) = u'(1) = 0.$$

Solución: Como en el primer ejemplo de este capítulo, se busca una solución de la forma

$$u(x) = sw(x) = \frac{s}{\lambda}\,\mathrm{sen}(\lambda x).$$

Para encontrar la solución al problema de contorno basta resolver la ecuación

$$u'(1) = s\cos\lambda = 0,$$

despejando la variable s. Para cualquier valor de λ siempre existe la solución trivial $u = 0$. Pero, si $\cos\lambda = 0$, s puede ser escogido arbitrariamente. En otras palabras, si $\lambda = \left(k + \frac{1}{2}\right)\pi$ para algún numero entero k, la función

$$u(x) = \mathrm{sen}\left(\left(k + \frac{1}{2}\right)\pi x\right)$$

(ó cualquier múltiplo de ella) es solución del problema de contorno homogéneo de Neumann. ◇

La aproximación numérica de un problema de contorno de Neumann sigue las mismas pautas que en el caso de los problemas de Dirichlet. La única diferencia está en cómo se aproxima la condición de contorno de Neumann. El modo más simple consiste en aproximar la derivada de la solución mediante una diferencia dividida progresiva en el caso del extremo lateral izquierdo o una retrógrada, en el caso del extremo lateral derecho.

▬ EJERCICIO 122 *Aproximar las soluciones del problema de contorno*

$$-u'' = 0, \quad u(0) = 1,\ u'(1) = 0$$

mediante un método de diferencias finitas que use una diferencia dividida retrógrada para aproximar la condición de contorno en el extremo lateral derecho. Determinar si la solución del problema aproximado es única.

Solución: Si se utiliza el esquema usual para discretizar la derivada segunda, se obtiene el siguiente sistema lineal

$$-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}=0, \qquad i=1,2,\cdots,N-1, \quad u_0=1, u_N=u_{N-1}.$$

Para analizar el problema discreto, se expresa en forma matricial $\mathrm{A}\mathbf{x}=\mathbf{b}$ donde

$$\mathrm{A}=\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}, \qquad \mathbf{b}=\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}.$$

Si se desarrolla el cálculo del determinante de A por la última fila, se comprueba que está dado por la fórmula

$$\det\,(\mathrm{A})=\det\,(\mathrm{A_{N-2}})-\det\,(\mathrm{A_{N-3}}),$$

donde A_j representa la matriz formada por las j primeras filas y columnas de la matriz A para $j=1,2,\cdots,N-2$. De acuerdo con el ejercicio 105, se tiene que

$$\det\,(\mathrm{A})=\mathrm{N}-1-(\mathrm{N}-2)=1>0,$$

lo que prueba que el sistema tiene una única solución. ◇

11.6 Ejercicios

EJERCICIO 123 *Determinar si el siguiente problema de contorno*

$$-u''+\lambda^2 u=0, \qquad u(0)=u(1)=1,$$

tiene solución y en caso afirmativo, encontrarla.

Solución: La solución general de la ecuación diferencial es

$$u(x)=c_1e^{\lambda x}+c_2e^{-\lambda x}.$$

La solución del problema de valor inicial $u(0)=u_0$ y $u'(0)=u_0'$ es

$$u(x)=\frac{u_0+u_0'/\lambda}{2}e^{\lambda x}+\frac{u_0-u_0'/\lambda}{2}e^{-\lambda x}.$$

De este modo, las soluciones v y w están dadas por

$$\begin{aligned} v(x;s) &= \frac{1}{2\lambda}\left((\lambda+s)e^{\lambda x}+(\lambda-s)e^{-\lambda x}\right), \\ w(x) &= \frac{1}{2\lambda}\left(e^{\lambda x}-e^{-\lambda x}\right). \end{aligned}$$

Ahora, se busca una solución de la forma

$$u(x) = v(x;s) + sw(x) = \frac{1}{2\lambda}\left((\lambda+2s)e^{\lambda x}+(\lambda-2s)e^{-\lambda x}\right)$$

Para encontrar la solución al problema de contorno basta resolver la ecuación

$$v(1;s) + sw(1) = \frac{1}{2\lambda}\left((\lambda+2s)e^{\lambda}+(\lambda-2s)e^{-\lambda}\right) = 1$$

despejando la variable s. De ello se deduce que para cualquier λ, la solución del problema de contorno es

$$u(x) = \frac{e^{\lambda x}+e^{\lambda(1-x)}}{1+e^{\lambda}}. \qquad \diamond$$

▬ EJERCICIO 124 *Expresar en forma matricial el sistema de ecuaciones lineales que resulta al aproximar la solución del siguiente problema de contorno*

$$-\frac{d^2u}{dx^2} = 1-4u, \quad 0 \le x \le 1, \quad u(0) = u(1) = 1$$

usando un método estándar de diferencia finitas. Determinar si está bien condicionado cuando el número de nodos empleado es elevado.

Solución: Si se usa la aproximación por diferencias finitas

$$\frac{d^2u}{dx^2}(x) \approx \frac{u(x+h)-2u(x)+u(x-h)}{h^2}$$

en el problema de contorno se obtiene el sistema de ecuaciones

$$-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + 4u_i = 1 \tag{11.5}$$

para $i = 1, 2, \cdots, N-1$ y $u_0 = u_N = 1$. En forma matricial, este sistema resulta

$$\frac{1}{h^2}\begin{pmatrix} 2+4h^2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2+4h^2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2+4h^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2+4h^2 & -1 \\ -1 & 0 & 0 & \cdots & -1 & 2+4h^2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{pmatrix} = \begin{pmatrix} 1+\frac{1}{h^2} \\ 1 \\ \vdots \\ 1+\frac{1}{h^2} \end{pmatrix}.$$

Los autovalores de la matriz de coeficientes A^h son

$$\lambda_j^h = \frac{2}{h^2}\left(1 + 2h^2 + \cos\frac{j\pi}{N}\right).$$

Puesto que la matriz es simétrica definida positiva, su número de condición coincide con el cociente entre el máximo y mínimo autovalor. En consecuencia se tiene que

$$\mathrm{cond}(\mathrm{A}^h) = \frac{1 + \frac{2}{N^2} + \cos\frac{\pi}{N}}{1 + \frac{2}{N^2} + \cos\frac{(N-1)\pi}{N}} = \frac{2 + N^2(1 + \cos\frac{\pi}{N})}{2 + N^2(1 - \cos\frac{\pi}{N})}.$$

Así pues, cuando $N \to \infty$ el número de condición tiende a ∞. ◇

– EJERCICIO 125 *Aproximar la solución del siguiente problema de contorno*

$$-\frac{d^2u}{dx^2} + 5u = 0, \quad 0 \le x \le 1, \quad u(0) = 1,\ u(1) = 8$$

usando un método de diferencia finitas con paso $h = \frac{1}{10}$.

Solución: Si se usa la aproximación por diferencias finitas

$$\frac{d^2u}{dx^2}(x) \approx \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}$$

en la ecuación diferencial, se obtiene el sistema de ecuaciones

$$u_{i+1} - (2 + 5h^2)u_i + u_{i-1} = 0 \tag{11.6}$$

para $i = 1, \cdots, 9$ y $u_0 = u_{10} = 8$. La ecuación característica de la ecuación en diferencias es

$$\lambda^2 - \frac{41}{20}\lambda + 1 = 0.$$

que tiene como raíces $\lambda_1 = \frac{4}{5}$ y $\lambda_1 = \frac{5}{4}$. La solución general de la ecuación en diferencias es

$$u_n = c_1\frac{4^n}{5^n} + c_2\frac{5^n}{4^n}.$$

Si se imponen las condiciones de contorno, se obtiene

$$\begin{aligned} c_1 + c_2 &= 1, \\ c_1\frac{4^{10}}{5^{10}} + c_2\frac{5^{10}}{4^{10}} &= 8 \end{aligned}$$

de donde se deduce que

$$u_n = \frac{99}{694}\frac{4^n}{5^n} + \frac{595}{694}\frac{5^n}{4^n}.$$ ◇

Bibliografía

[1] D.N. Arnold, *A Concise Introduction to Numerical Analysis*. Pre-Publicación, 2001.

[2] E.W. Cheney, *Introduction to Approximation Theory*, 2nd ed. Providence, RI: Amer. Math. Soc., 1999.

[3] K. Eriksson, D. Estep, P. Hansbo y C. Johnson, *Computational differential equations*, Cambridge University Press, 1996.

[4] G. Forsythe,C.B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall I, 1967.

[5] C.W. Gear, R.D. Skeel, The development of ODE methods: A symbiosis between hardware and numerical analysis, *A history of scientific computing*, ed. by Stephen Nash, AC. Press 1990.

[6] A. Iserles, *A first course in the numerical analysis of differential equations*, Cambridge University Press, 1996.

[7] D. Knuth ,*The Art of Computer Programming*, Vol. I-V, Addison-Wesley, 2005.

[8] A. Quarteroni, R. Sacco y F. Saleri, *Numerical Mathematics*, Springer, 2000, ISBN 0-387-98959-5.

[9] A. Quarteroni y F. Saleri, *Cálculo Científico con MatLab y Octave*, Springer, 2006, ISBN 88-470-0503-5.

[10] J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis* Springer-Verlag, New York, 1980, ISBN 0-387-90420-4.

[11] R. Varga, *Matrix iterative analysis*, Springer Verlag, ISBN 3-540-66321, Springer 2000.

Índice alfabético

Este texto pretende iniciar a los estudiantes de grado de la UNED en el conocimiento de los métodos numéricos del Cálculo Científico.
Su estudio requiere conocimientos básicos de álgebra lineal, cálculo diferencial y ecuaciones diferenciales. Está especialmente diseñado para la enseñanza que no es presencial y contiene numerosos ejercicios resueltos.

Carlos Moreno se licenció y doctoró en Ciencias Matemáticas por la Universidad de Santiago de Compostela. Completó sus estudios de doctorado en el INRIA *(*Institut de Recherche en Informatique et en Automatique*)* en París, entre 1974 y 1976. Fue profesor en las Universidades de Santiago y Vigo. En 1979 obtuvo una plaza de profesor titular en la Universidad Autónoma de Madrid, y en 1987 de catedrático en ETSI de Caminos, Canales y Puertos de la Universidad Politécnica de Madrid. Desde 2002 es catedrático en la UNED. Ha realizado investigación en métodos numéricos en Ingeniería y Finanzas.

UNED Editorial

colección Grado

ISBN: 978-84-362-6257-5

9 788436 262575 02208

6102208GR01A01