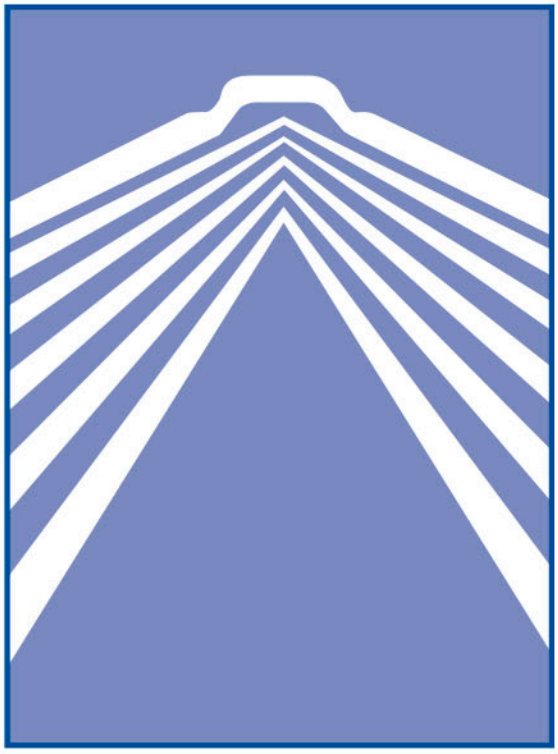
4.ª edición

MÉTODOS NUMÉRICOS

Teoría, problemas y prácticas con MATLAB



Juan Antonio Infante del Río
José María Rey Cabezas

PIRÁMIDE

4.ª edición

MÉTODOS NUMÉRICOS

Teoría, problemas y prácticas con MATLAB

JUAN ANTONIO INFANTE DEL RÍO
PROFESOR TITULAR DE UNIVERSIDAD
DEL DEPARTAMENTO DE MATEMÁTICA APLICADA
DE LA FACULTAD DE CIENCIAS MATEMÁTICAS
DE LA UNIVERSIDAD COMPLUTENSE DE MADRID

JOSÉ MARÍA REY CABEZAS
PROFESOR TITULAR DE UNIVERSIDAD
DE LA SECCIÓN DEPARTAMENTAL DE MATEMÁTICA
APLICADA DE LA FACULTAD DE CIENCIAS QUÍMICAS
DE LA UNIVERSIDAD COMPLUTENSE DE MADRID

4.ª edición

MÉTODOS NUMÉRICOS

Teoría, problemas y prácticas con MATLAB

EDICIONES PIRÁMIDE

COLECCIÓN «CIENCIA Y TÉCNICA»

Edición en versión digital

Para cualquier información pueden dirigirse a piramide_legal@anaya.es
Juan Ignacio Luca de Tena, 15. 28027 Madrid
Teléfono: 91 393 89 89
www.edicionespiramide.es
ISBN digital: 978-84-368-3336-2

A Julián Infante Lopesino,
Julián Infante del Río,
Carmen del Río Martín,
María Cabezas Pérez
y Julio Rey Vallejo.

In memoriam.

Índice

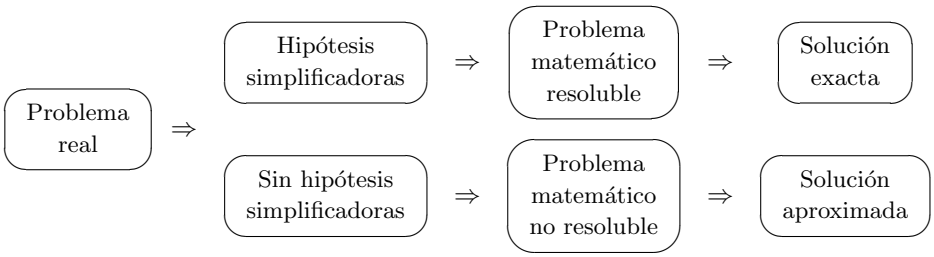
Prefacio

P. Henrici da una definición aproximada del Análisis Numérico como "*la teoría de los métodos constructivos en Análisis Matemático*", haciendo un especial énfasis en la palabra "constructivos". Durante mucho tiempo, las matemáticas fueron totalmente constructivas, pues su único objetivo era llegar a la solución de problemas concretos. No obstante, a medida que los problemas sujetos a la investigación matemática crecían en alcance y generalidad, los matemáticos fueron interesándose, cada vez más, por cuestiones como la *existencia*, *unicidad* y *propiedades cualitativas* de la solución, antes que por su construcción. Una de las causas que condujeron a esta situación fue la escasa capacidad de cálculo que hacía inútil el diseño de algoritmos constructivos de la solución de problemas complejos. No obstante, cuando parecía que las matemáticas habían olvidado cualquier matiz constructivo, surgieron los primeros *ordenadores*, los cuales devolvieron a los matemáticos la esperanza de poder construir las soluciones de los problemas. Fue entonces cuando nació lo que hoy denominamos *Análisis Numérico*. Aunque muchas de las ideas básicas en que se apoyan las técnicas numéricas actuales se conocen desde hace tiempo, ha sido la capacidad de cálculo aportada por los ordenadores (vertiginosamente acrecentada con el transcurso del tiempo) la que les ha dado mayor vigencia e importancia.

El Análisis Numérico es una herramienta fundamental en el campo de las ciencias aplicadas que trata de diseñar métodos que aproximen, de forma eficiente, las soluciones de problemas prácticos previamente formulados matemáticamente. En la mayoría de los casos, el problema matemático se deriva de un problema práctico en áreas experimentales como la Física, Química, Biología, Economía... Sobre él se aplican, típicamente, dos tipos de estrategias generales:

a) Se dan por supuestas algunas hipótesis de carácter simplificador que permiten llegar a una formulación matemática resoluble. (Así es como se procedió tradicionalmente, hasta que se contó con las técnicas numéricas.)

b) Se prescinde de alguna de estas hipótesis para llegar a una formulación matemática más complicada, que no se puede resolver explícitamente, pero cuya solución puede calcularse de forma aproximada.



Aunque de ninguna de las dos formas anteriores obtenemos la solución del problema original, a menudo resulta más apropiado utilizar la segunda. Para ello, se idea un *algoritmo*[1], es decir, una secuencia finita de operaciones algebraicas y lógicas, que se espera produzca una solución aproximada del problema matemático y, en consecuencia, del físico. La confirmación de esta esperanza es, precisamente, una de las principales tareas del Análisis Numérico. En otras palabras, el cometido de la disciplina que nos ocupa es el diseño de métodos que conduzcan (y no solamente de manera teórica) a la solución de los problemas planteados: estudiar los algoritmos en profundidad para que se pueda saber de antemano cuáles son sus ventajas e inconvenientes, qué dificultades presentan a la hora de llevar a cabo su programación efectiva y seleccionar, en cada caso, el algoritmo más *eficiente* en cuanto al almacenamiento de datos y al tiempo de cómputo. En definitiva, proporcionar métodos que permitan obtener realmente una aproximación de la solución buscada y conocer, en la medida de lo posible, el grado de aproximación entre la solución hallada y la real, es decir, dar una *estimación* del *error* cometido.

Este libro está escrito tras años de experiencia de los autores en la docencia de asignaturas relacionadas con el Análisis Numérico. En particular, de la asignatura *Métodos Numéricos* de la licenciatura en Ciencias Matemáticas de la Universidad Complutense de Madrid desde que se puso en marcha el nuevo plan de estudios. Esta asignatura tiene carácter troncal, por lo que se imparte, con similares contenidos, en todas las titulaciones de Matemáticas del Estado.

El objetivo de este manual es modesto, pero no por ello carente de importancia. Por una parte, pretendemos proporcionar al alumno una primera toma de contacto con las técnicas numéricas que le sirva para conocer un amplio catálogo de métodos que aproximan las soluciones de los problemas abordados (esencialmente, ecuaciones y sistemas lineales y no lineales, interpolación, derivación e integración). Por otra, se intenta cimentar una sólida base teórica que permita conocer los límites de validez y condiciones de aplicación de los métodos, así como el ulterior estudio en profundidad de otras técnicas más sofisticadas.

[1]La palabra algoritmo proviene del matemático persa *Abu Jafar Mohammed ibn Musa al–Khowarizmi*, que vivió en Bagdad hacia el año 840 d.C.

Cada capítulo tiene una sección de problemas de los que, alrededor de la mitad, se resuelven con todo detalle; pensamos que la escasez de textos que incluyan una buena cantidad de problemas resueltos en su totalidad, puede ser un valor añadido de esta obra. En algunos problemas se recogen resultados complementarios a los expuestos en la parte teórica, enunciados de forma que su resolución sea abordable.

Finalmente, destacar que en la última sección de cada capítulo se enuncian una serie de prácticas de ordenador, pensadas para ser implementadas en `MATLAB`. En muchas de ellas se pide realizar programas que se podrían sustituir por un único comando (de hecho, en general se indica que se compare con el comando en cuestión). Si hacemos esto es porque estamos convencidos (y la experiencia con nuestros alumnos así lo confirma) de que sólo cuando uno se enfrenta a la programación efectiva de los métodos es capaz de entenderlos en profundidad. En las direcciones de internet

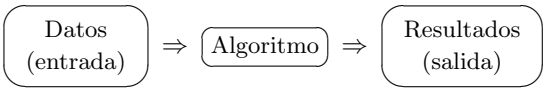
`http://www.mat.ucm.es/~infante` o `http://www.mat.ucm.es/~jrey`

pueden encontrarse programas que resuelven algunas de las prácticas y servirán de ayuda, eventualmente, para resolver otras.

1 Análisis de errores

1.1. Introducción

El Análisis Numérico es una disciplina que contempla el desarrollo y evaluación de métodos para calcular, a partir de ciertos datos numéricos, los resultados requeridos. Se puede pensar el problema que se quiere resolver como una aplicación f que transforma datos x en resultados y; los datos constituyen la *información de entrada*, los resultados requeridos son la *información de salida* y el método de cálculo se conoce como *algoritmo*. Los ingredientes esenciales en un problema de Análisis Numérico pueden resumirse, pues, en el siguiente diagrama:



A modo de ejemplo, podemos pensar en el problema de hallar $\sqrt{17}$. En este caso, $x = 17$ es el *dato*, f la función raíz cuadrada e $y = \sqrt{17}$ el *resultado* deseado. Como *algoritmo* de cálculo podemos construir, para $\lambda = 17$, la sucesión

$$\begin{cases} x_0 = \lambda \\ x_n = \dfrac{1}{2}\left(x_{n-1} + \dfrac{\lambda}{x_{n-1}}\right), \ n \in \mathbb{N}, \end{cases}$$

cuyo límite es $\sqrt{\lambda}$ (compruébese), parando en un valor de n suficientemente grande.

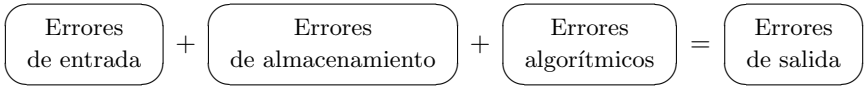
Con bastante frecuencia nos encontramos con distintos algoritmos para construir la información de salida que se requiere. Así, volviendo al ejemplo anterior, para aproximar el número $\sqrt{17}$ podemos utilizar también el algoritmo que se obtiene al hacer un desarrollo de Taylor de orden 2 de la función raíz cuadrada

$$\sqrt{x} \simeq \sqrt{a} + \frac{1}{2\sqrt{a}}(x-a) - \frac{1}{8\sqrt{a^3}}(x-a)^2$$

particularizando en $x = 17$ y $a = 16$, es decir,

$$\sqrt{17} \simeq 4 + \frac{1}{8} - \frac{1}{512} = \frac{2111}{512} = 4.123046875.$$

Por tanto, para escoger entre los diversos algoritmos disponibles deben estudiarse los aspectos teóricos que contribuirán a la elección del algoritmo más adecuado a cada caso concreto. En general, los criterios fundamentales para preferir un algoritmo frente a otro son la *rapidez* y la *precisión*; a igualdad de precisión el algoritmo más rápido tendrá, obviamente, la preferencia. El objetivo de este capítulo es, de hecho, el estudio de la precisión o, equivalentemente, del *error*. En muy pocas ocasiones la información de entrada que se suministra es exacta pues se obtiene, en general, mediante instrumentos de medida; como, por otra parte, tanto el almacenamiento de los datos como el propio algoritmo de cálculo introducen también errores, la información de salida contendrá errores que provendrán de las tres fuentes. Esquemáticamente:



Sobre el primer tipo de errores nada podemos decir: están relacionados con el diseño de los aparatos de medición o la precisión de la percepción a través de los órganos sensoriales. Los otros dos tipos de errores son los que analizaremos en este capítulo: en las dos primeras secciones se aborda el estudio de los errores de almacenamiento y en las dos últimas los algorítmicos.

Antes de comenzar este estudio recordemos cierta terminología estándar en el tratamiento de errores. El error cometido al calibrar cierta magnitud puede ser medido bien en términos *absolutos*, bien en términos *relativos* al tamaño de la magnitud que se aproxima. Así, si $\tilde{z}$ es una aproximación de una cierta cantidad $z \neq 0$ y $||\cdot||$ es una norma (véase la definición 2.19), entonces

$$||\tilde{z} - z|| \quad \text{es el } \textit{error absoluto}$$

y

$$\frac{||\tilde{z} - z||}{||z||} \quad \text{es el } \textit{error relativo}$$

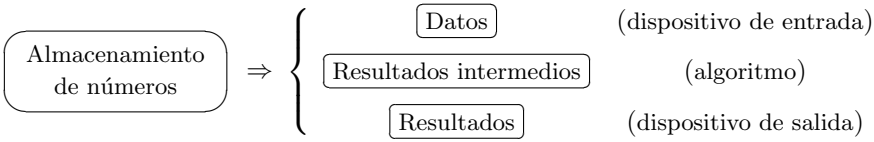
cometido en la aproximación anterior.[1]

[1] Si $z = 0$ se trabaja sólo con errores absolutos.

Desde el punto de vista de las aplicaciones el error relativo resulta más relevante, ya que conocer el error absoluto no es muy útil si no se conoce la magnitud de la cantidad que se está midiendo. Por ejemplo, un error de $0'25$ cm al determinar la estatura de una persona puede ser irrelevante pero sería inaceptable en microcirugía.

1.2. Números máquina

Los problemas que aborda el Análisis Numérico se resuelven, fundamentalmente, mediante la realización de cadenas de operaciones aritméticas (más o menos sofisticadas) en un ordenador. Deberán, por tanto, almacenarse en la máquina los datos de partida (que introduciremos mediante un *dispositivo de entrada*, como puede ser un teclado o la lectura de un fichero) así como los resultados intermedios. Finalmente, deberán comunicarse los resultados finales (mediante un *dispositivo de salida*, como es la pantalla o la escritura en un fichero).



En esta sección vamos a ocuparnos del almacenamiento de los números en la máquina. Como es sabido, la capacidad de almacenamiento (la *memoria*) de los ordenadores crece vertiginosamente gracias a los progresos de la electrónica pero, por mucho que crezca, siempre será una capacidad finita. Esto implica que ninguna máquina es capaz de guardar ni siquiera un solo número irracional "completo" (las infinitas cifras decimales de π, por ejemplo). De hecho, cada número se representa en el ordenador con una cantidad máxima de cifras decimales; consecuentemente, sólo se guardarán de forma exacta los números que no excedan de ese máximo.

Vamos a intentar aclarar estas afirmaciones, que hemos introducido de forma un tanto vaga. Los números pueden representarse en la denominada *notación decimal en coma flotante*, que contiene la información relevante: signo, fracción, signo para el exponente y exponente. Por ejemplo, el número $-623'45$ admite, entre otras, las siguientes representaciones en coma flotante

$$-623'45+0,\ -62'345+1,\ -6234'5-1, -6'2345+2,\ldots$$

que deben entenderse, respectivamente, como

$$-623'45\times 10^{0},\ -62'345\times 10^{1},\ -6234'5\times 10^{-1}, -6'2345\times 10^{2},\ldots$$

De las infinitas representaciones posibles nos quedaremos con la última (denominada *notación decimal* en *coma flotante normalizada*), que está caracterizada por que la fracción es un número comprendido entre 1 y 10. Es decir, la notación decimal en coma flotante normalizada de un número real no nulo es

$$\pm m \pm E \ \text{ con } \ 1 \leq m < 10 \ \text{ y } \ E \in \mathbb{N} \cup \{0\}$$

que se corresponde con $\pm m \times 10^{\pm E}$.

Observación 1.1. A partir de ahora emplearemos, como hacen los ordenadores, la notación anglosajona de punto decimal en lugar de coma decimal, aunque seguiremos hablando de notación en coma flotante. □

Los ordenadores almacenan la información en cantidades ingentes de *posiciones de memoria* (o *bits*).[2] Éstas son entes físicos que sólo pueden tomar los valores 0/1, encendido/apagado, positivo/negativo o cualquier otra dicotomía electrónicamente viable. Es por esto por lo que no se utiliza la representación decimal para los números, sino la *representación binaria*.

Observación 1.2. El *sistema binario* utiliza el 2 como base, de la misma forma que el *sistema decimal* utiliza el 10. Con el propósito de hacer una comparación, recuérdese primero cómo funciona el sistema de representación decimal que nos es más familiar. Cuando escribimos el número 547.154 de forma más explícita tenemos

$$547.154 = 5 \times 10^2 + 4 \times 10^1 + 7 \times 10^0 + 1 \times 10^{-1} + 5 \times 10^{-2} + 4 \times 10^{-3}.$$

La expresión del segundo miembro utiliza potencias de 10 junto con los dígitos

$$0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9.$$

En el sistema binario sólo se utilizan los dígitos 0 y 1. Por ejemplo,

$$101.101 = 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3}.$$

Este número real se representa como 5.625 en notación decimal.

En general, cualquier número natural $\alpha > 1$ puede utilizarse como *base* para un *sistema numérico*. Los números representados en base α incluirán símbolos $s_0,\ s_1,\ s_2, \ldots,\ s_{\alpha-1}$ (véase el problema 1.8). De esta forma, todo número real x admite una representación en base α de la forma

$$\begin{aligned} x &= \pm \sum_{k=-\infty}^{\infty} c_k \alpha^k \\ &= \pm \left(\cdots + c_{-2}\alpha^{-2} + c_{-1}\alpha^{-1} + c_0\alpha^0 + c_1\alpha^1 + c_2\alpha^2 + \cdots \right) \end{aligned} \tag{1.1}$$

[2] La palabra *bit* es la abreviatura de *Binary Digit*.

con $c_k \in \{s_0, s_1, s_2, \ldots, s_{\alpha-1}\}$ para $k \in \mathbb{Z}$. Si el contexto no aclara cuál es la base numérica que se utiliza para el número x, suele emplearse la notación $(x)_\alpha$. Así, por ejemplo,

$$(101.101)_2 = (5.625)_{10}.$$

Ya que un ordenador se comunica con el usuario utilizando el sistema decimal (aun cuando internamente esté utilizando el sistema binario), debe efectuar procesos de conversión que ocurren durante la entrada o salida de la información. Normalmente el usuario no tiene que preocuparse de dichas conversiones, pero debe tener conciencia de que éstas implican errores de *redondeo*, como se verá en la sección 1.3. Por ejemplo, algunos números decimales sencillos toman expresiones más complicadas en el sistema binario; así,

$$(0.1)_{10} = (0.0\,\overbrace{0011}^{}\,)_2$$

(véase el problema 1.2). □

Observación 1.3. La representación en base $\alpha > 1$ dada en (1.1) puede no ser única como, por ejemplo, ocurre con el número representado en base 10 por $0.\overbrace{9}^{}$ y 1. En efecto, al multiplicar $x = 0.\overbrace{9}^{}$ por 10 obtenemos $10x = 9.\overbrace{9}^{}$ de forma que al restar x a $10x$ queda $9x = 9$, de donde $x = 1$. □

Al igual que utilizábamos la notación decimal en coma flotante podemos también utilizar esta notación en el sistema binario. Un número binario representado en coma flotante normalizada tomará una expresión de la forma

$$(s)\,(S)\,E\,m$$

donde s es el signo del número (0 para los positivos y 1 para los negativos), S es el signo del exponente (también aquí 0 representará el signo + y 1 el signo −), el exponente E es un número entero en base 2 y m se denomina *mantisa* y es un número en base 2 verificando $1 \leq m < 2$ (es decir, $1 \leq m < (10)_2$). El número así representado será

$$(-1)^s \times m \times 2^{(-1)^S \times E_d}$$

donde E_d es la conversión decimal del número binario E (es decir, el número de lugares que hay que "correr la coma").

Ejemplo 1.1.

$$(1)\,(0)\;10\;1.01 \equiv -1.01 \times 2^2 \equiv -101\ (\equiv -(2^2 + 2^0) = -5 \text{ en decimal})$$

y

$$(0)\,(1)\;1\;1.1 \equiv 1.1 \times 2^{-1} \equiv 0.11\ (\equiv 2^{-1} + 2^{-2} = 0.75 \text{ en decimal}).$$ □

Cada número se almacena en la máquina en lo que se denomina una *palabra*. La mayoría de los ordenadores actuales tienen una longitud de palabra de 32 posiciones de memoria. Estos 32 bits se distribuyen, como veremos a continuación, de la siguiente forma: 1 se usa para el signo del número, 8 para el exponente con signo y 23 para la mantisa.

Observación 1.4. El hecho de que la mantisa m verifique $1 \leq m < 10$ en binario, hace que la primera cifra de m (la que está a la izquierda del punto decimal) sea siempre un 1. Este 1, común a todos los números representados, no se guarda (se sobreentiende) y, por tanto, los 23 bits que se destinan a la mantisa sirven para almacenar las 24 primeras cifras. □

Obviamente, utilizando palabras de 32 bits, tan sólo puede ser representada una cantidad finita de números distintos: del orden de 2^{32} números diferentes.[3] Estos números, representables en el ordenador de forma *exacta*, se denominan *números máquina*.

Es fácil encontrar números que no son números máquina; por ejemplo, 1.0×2^{128} (128 es, en binario, 10000000, que no cabe en la parte destinada al exponente con signo) y $1.1 \overset{25)}{\cdots} 1$ (no cabe en el espacio destinado a la mantisa). El primero es un ejemplo en el que se sobrepasa la capacidad del ordenador (*desbordamiento*) y el segundo muestra que hay números que no se pueden usar de forma exacta y habrá que aproximarlos (*redondeo*). Ambos fenómenos serán estudiados más adelante.

Continuando con la representación de los números máquina, podemos preguntarnos por la representación del número real 0. La primera respuesta que nos viene a la cabeza es: "una palabra en la que los 32 bits tengan el valor 0". Pero el convenio de que se sobreentiende el primer 1 de la mantisa hace que la palabra anterior represente el número binario 1.0 (1 en notación decimal); de hecho, el 0 no sería un número máquina si se usara esta representación.

Este problema, junto con el de la doble representación de los números con exponente $+0$ y -0, se resuelve mediante el *formato estándar de representación* (*IEEE*[4] *Storage Format*): para almacenar la información correspondiente al exponente y su signo se suma 127 al exponente y se utilizan 8 bits para almacenarlo; el exponente verdadero se obtendrá al restar 127 a la conversión decimal del exponente almacenado. Así pues, el número queda representado en la forma

Signo del número	Exponente +127	Mantisa
1 bit	8 bits	23 bits

[3] Como los exponentes $+0$ y -0 dan lugar a los mismos números, el número total es la diferencia entre las variaciones con repetición de 2 elementos tomados de 32 en 32 y las variaciones con repetición de 2 elementos tomados de 24 en 24, es decir, $2^{32} - 2^{24} = 4278190080$.

[4] *Institute of Electrical and Electronic Engineers.*

Como el rango de números que pueden representarse con 8 bits es de 0 a 255 (ya que $|E| \leq (11111111)_2 = (255)_{10}$), los exponentes verdaderos correspondientes están en el rango comprendido entre -127 y 128. Los valores extremos 0 y 255 se usarán para números "especiales". En concreto, si s es el signo del número almacenado, m son los 23 dígitos de mantisa y E_d es la conversión decimal del número binario guardado en los 8 bits dedicados al exponente; cuando éste toma valores entre 1 y 254 (ambos inclusive), el número anteriormente representado corresponde al número binario

$$\boxed{(-1)^s \times 1.m \times 2^{E_d - 127}}$$

A este tipo de números se les denomina *números normales*.

Ejemplo 1.2. El número binario $101.11 = 1.0111 \times 2^2$ se representa de la siguiente forma

$$\overbrace{0}^{\text{Signo del número}} \quad \overbrace{10000001}^{2+127=129} \overbrace{01110000000000000000000}^{\text{Mantisa}}.$$

A la inversa, el número almacenado como

$$1\ 01111011\ 10100000000000000000000$$

es el número $-1.101 \times 2^{123-127} = -1.101 \times 2^{-4} = -0.0001101$ en base 2. □

Cuando el exponente toma el valor 0, el convenio es ligeramente distinto: el valor representado es

$$\boxed{(-1)^s \times 0.m \times 2^{-126}}$$

(obsérvese que aquí no se sobreentiende el valor 1 para el primer dígito de la mantisa y se ha desplazado la coma adecuadamente). Esta otra interpretación permite trabajar con números máquina mucho más pequeños, en valor absoluto, que los que se obtendrían con el convenio habitual. Así, al exponente 0 le corresponderían números positivos comprendidos entre

$$1.0 \times 2^{-127} \ \text{ y } \ 1.1 \overset{23)}{\cdots} 1 \times 2^{-127}$$

con el convenio habitual, y entre

$$0.0 \overset{22)}{\cdots} 01 \times 2^{-126} = 1.0 \times 2^{-149} \ \text{ y } \ 0.1 \overset{23)}{\cdots} 1 \times 2^{-126}$$

con esta nueva interpretación. De esta forma, se ve ampliada la capacidad de manejo, por parte de la máquina, de números "pequeños". Los números máquina

así representados se denominan *números subnormales*. Nótese que el 0 es uno de ellos, estando representado por una palabra enteramente formada por ceros, como pedía nuestra intuición.

Los números con exponente 255 representan las *excepciones*: $\pm\infty$ y `NaN` (*Not a Number*). Los valores $\pm\infty$ se producen a causa de un desbordamiento por exceso, mientras que el valor `NaN` surge al realizar ciertas operaciones como las clásicas indeterminaciones ($\frac{0}{0}$, $\infty - \infty, \ldots$), como se verá en la sección 1.3.

En la tabla 1.1 se recogen la representación y el valor decimal de algunos números especialmente interesantes (véase la práctica 1.3).

TABLA 1.1:
Representación de diversos números ($2^{127} \simeq 10^{38}$)

Número	Representación	Valor decimal
$+0$	$0\ 00000000\ 0\overset{23)}{\cdots}0$	0.0
-0	$1\ 00000000\ 0\overset{23)}{\cdots}0$	-0.0
Máximo número normal	$0\ 11111110\ 1\overset{23)}{\cdots}1$	$3.40282347 \times 10^{38}$
Mínimo número normal (positivo)	$0\ 00000001\ 0\overset{23)}{\cdots}0$	$1.17549435 \times 10^{-38}$
Máximo número subnormal	$0\ 00000000\ 1\overset{23)}{\cdots}1$	$1.17549421 \times 10^{-38}$
Mínimo número subnormal (positivo)	$0\ 00000000\ 0\overset{22)}{\cdots}01$	$1.40129846 \times 10^{-45}$
$+\infty$	$0\ 11111111\ 0\overset{23)}{\cdots}0$	infinito
$-\infty$	$1\ 11111111\ 0\overset{23)}{\cdots}0$	–infinito
`NaN`	$*\ 11111111$ algún 1	no es un número

Observación 1.5.

1. Hasta ahora sólo hemos considerado la representación de números reales en lo que se denomina *precisión simple*. En general, las máquinas también permiten trabajar en *doble precisión*: cada número se almacena en dos palabras unidas y, de los 64 bits disponibles, 1 se destina al signo, 11 al exponente y 52 a la mantisa. Así, los números normales en doble precisión se representan

Signo del número	Exponente +1023	Mantisa
1 bit	11 bits	52 bits

   lo que corresponde al número binario

   $$(-1)^s \times 1.m \times 2^{E_d - 1023}$$

   y determinan un rango del orden de 10^{-308} a 10^{308} (véase la práctica 1.4).

2. En general, los ordenadores (y los lenguajes de programación) distinguen entre números reales (tipo `REAL` o `FLOAT`) y números enteros (tipo `INTEGER`). Una vez declarado un número como entero, su almacenamiento se hace en una palabra, usando el primer bit para el signo y los 31 restantes para el número en binario. Por ejemplo, el número $(-254)_{10} \equiv (-11111110)_2$ se almacena como $10\overset{23)}{\cdots}011111110$. Como el mayor número entero representable es

$$0\,1\overset{31)}{\cdots}1 = \sum_{k=0}^{30} 2^k = (2^{31}-1)_{10},$$

esto hace que el rango de los números máquina enteros vaya desde $-(2^{31}-1)$ a $2^{31}-1$. También se pueden representar *enteros largos* mediante dos palabras unidas, usando 1 bit para el signo y 63 para el número en binario. Existen otras formas más sofisticadas de representar los números enteros que aquí no trataremos. En especial, la conocida con el nombre de *complemento a* 2, que está íntimamente relacionada con el tipo de almacenamiento con traslación que se usa para los exponentes.

3. Los números complejos se representan como un par de números reales (dos palabras en simple precisión y cuatro en doble precisión).

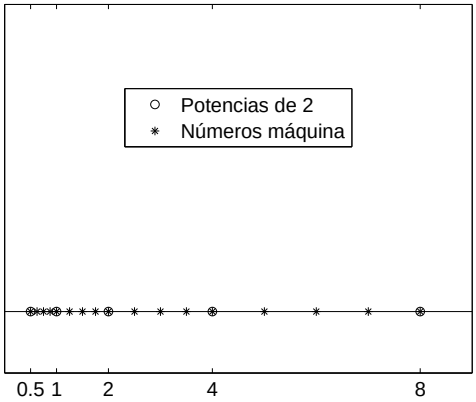


Figura 1.1: Distribución irregular de los números máquina.

4. En un ordenador binario, los números máquina están distribuidos de forma irregular, concentrándose más cuanto más cerca de cero se esté (véase la figura 1.1); téngase en cuenta que entre dos potencias consecutivas de 2 siempre existe la misma cantidad de números máquina: para un exponente fijo E y prefijado el signo, la cantidad total de números máquina normales

de la forma $\pm 1.m \times 2^E$ es 2^{23} (número de variaciones con repetición de 2 elementos tomados de 23 en 23). Dado que la distancia entre las potencias de 2 es más pequeña cerca del 0 y más grande lejos de él, se obtiene una distribución desigual de números máquina, siendo la densidad mayor cerca del origen. □

1.3. Desbordamiento y redondeo

Ya hemos visto la forma estándar de almacenar números máquina y hemos insistido en que éstos constituyen sólo una cantidad finita. ¿Qué podemos decir de los restantes números reales? El hecho de que un número real no sea un número máquina puede deberse a dos motivos:

a) Que su exponente (una vez normalizado el número) esté fuera del rango admitido (es decir, que el número sea demasiado grande o pequeño).

b) Que su mantisa (normalizada también) tenga más de 23 cifras (en cuyo caso el número tiene más cifras de las que se pueden almacenar).

Estas dos posibilidades se tratan de forma muy distinta:

a) Si el resultado de un cálculo o un número leído por la máquina tiene un exponente que se sale del rango admisible se produce el fenómeno conocido como *desbordamiento*, que puede ser de dos tipos: por *exceso* (*overflow*) o por *defecto* (*underflow*). La representación estándar en coma flotante trata el desbordamiento por exceso como una situación *excepcional*: toma el valor $+\infty$ (o $-\infty$ si el número es negativo)[5] y sigue trabajando, siempre que se pueda dar sentido a las operaciones siguientes (por ejemplo, $\frac{1}{\infty} = 0$). Para el desbordamiento por defecto se asigna el valor 0.

b) Cuando se necesita trabajar con números no máquina que no producen desbordamiento lo que se hace es aproximarlos por números máquina cercanos. Este proceso se denomina *redondeo*.

Veamos la forma en la que el redondeo se lleva a cabo: dado un número

$$x = \pm 1.a_1 a_2 \cdots a_{23} a_{24} a_{25} \cdots \times 2^E$$

con E dentro del rango admisible, sean x_i y x_d los números máquina más próximos a x por la izquierda y por la derecha respectivamente, esto es,

$$x_i = x_d = x \text{ si } x \text{ es un número máquina}$$

[5] En la tabla 1.1 se vio cómo se representaban estos valores en la máquina.

y, en otro caso,

$$\begin{cases} x_i = 1.a_1 \cdots a_{23} \times 2^E & \text{y} \quad x_d = (1.a_1 \cdots a_{23} + 2^{-23}) \times 2^E \quad \text{si} \quad x > 0 \\ x_i = -(1.a_1 \cdots a_{23} + 2^{-23}) \times 2^E & \text{y} \quad x_d = -1.a_1 \cdots a_{23} \times 2^E \quad \text{si} \quad x < 0. \end{cases}$$

Nótese que, en cualquier caso, para todo número $x \in \mathbb{R}$ se verifica que

$$x_i \leq x \leq x_d.$$

En coma flotante estándar, para cada número real x, están definidos cuatro tipos de redondeo $r(x)$:

a) *Redondeo a la derecha* (o por *exceso*): $r(x) = x_d$.

b) *Redondeo a la izquierda* (o por *defecto*): $r(x) = x_i$.

c) *Redondeo a cero*: $r(x) = x_i$ si $x > 0$ y $r(x) = x_d$ si $x < 0$ (es decir, se elige de entre x_i y x_d el que está entre 0 y x).

d) *Redondeo al más próximo*: se elige de entre x_i y x_d el que está más cerca de x; en el caso de que estén a igual distancia, se toma el que tenga $a_{23} = 0$.

Observación 1.6.

1. Nótese que, con cualquiera de los criterios anteriores, si x es un número máquina entonces coincide con su redondeo $r(x)$.

2. Aunque en principio puede utilizarse cualquiera de las cuatro posibilidades anteriores (incluso puede cambiarse de una a otra dentro de la misma máquina) nosotros solamente utilizaremos la última de ellas, por ser la más útil y usada (de hecho es la opción por defecto) y la denominaremos, simplemente, *redondeo*. □

La primera cuestión que se nos plantea es qué error se comete cuando en lugar de trabajar con un número real x se trabaja (como es obligado) con su redondeo. En términos absolutos, y manteniendo la notación, se tiene la siguiente *cota del error de redondeo absoluto*

$$|r(x) - x| \leq \frac{x_d - x_i}{2} = \frac{2^{-23} \times 2^E}{2} = 2^{-24} \times 2^E$$

mientras que, en términos relativos, una *cota del error de redondeo relativo* (teniendo en cuenta que $|x| \geq 2^E$) es

$$\left| \frac{r(x) - x}{x} \right| \leq \frac{2^{-24} \times 2^E}{2^E} = 2^{-24}.$$

El valor de 2^{-24} obtenido en la cota del error de redondeo relativo es, exactamente, la mitad de la distancia entre 1 y el siguiente número máquina $1.0 \overset{22)}{\cdots} 01$. Esta distancia ($2^{-23} \simeq 1.2 \times 10^{-7}$) se denomina *precisión* o *épsilon* de la máquina y se denota por *eps*. Así pues, trabajando en coma flotante estándar, como

$$r(x) = \left(\frac{x + r(x) - x}{x}\right) x = \left(1 + \frac{r(x) - x}{x}\right) x,$$

se verifica que

$$r(x) = (1 + \delta)x \ \text{ con } \ |\delta| \leq 2^{-24} = \frac{eps}{2}$$

para todo $x \in \mathbb{R}$. Esto, en términos de la representación decimal de x, quiere decir que el redondeo de x tendrá alrededor de 7 *cifras significativas* correctas (de forma más precisa, un mínimo de 6 y un máximo de 8, dependiendo de en cuántas cifras decimales se transformen las 23 cifras decimales en binario).

Observación 1.7. Todas las consideraciones que se han hecho acerca del desbordamiento y el redondeo con precisión simple pueden trasladarse, con las correspondientes modificaciones, al caso en que se utilice doble precisión. En concreto, el épsilon de la máquina en doble precisión es 2^{-52} y el redondeo de un número en doble precisión tiene un mínimo de 15 y un máximo de 16 cifras significativas exactas. □

1.4. Aritmética en coma flotante

Para llevar a cabo una operación aritmética $(+, -, \times, /)$ en un ordenador, lo primero que se debe tener en cuenta es que los operandos con los que se trabaja no son exactamente los de partida, sino sus redondeos. Incluso aunque los operandos sean números máquina, el asunto es más complicado de lo que puede parecer a primera vista; el resultado de una operación con números máquina no tiene por qué ser necesariamente un número máquina y habrá, por tanto, que redondearlo. De hecho, la aritmética estándar en coma flotante está diseñada mediante operaciones $\oplus$, $\ominus$, $\otimes$, $\oslash$ que verifican, para cada par de números máquina x e y:

$$x \oplus y = r(x + y),\ x \ominus y = r(x - y),\ x \otimes y = r(x \times y),\ x \oslash y = r(x/y).$$

La implementación efectiva de las operaciones es bastante sofisticada, pues debe enfrentarse con diversos problemas de los que veremos algunos ejemplos seguidamente.

En primer lugar, para llevar a cabo una suma (o resta) de dos números máquina se siguen los siguientes pasos:

1. Se igualan los exponentes al mayor de ambos.
2. Se realiza la operación de forma exacta.
3. Se normaliza el resultado (modificando, si es necesario, el exponente de forma que la mantisa esté entre 1 y 2, es decir, de la forma $\pm 1.m \times 2^E$).
4. Se redondea (en caso de que sea necesario).
5. Se normaliza si es necesario.[6]

Ejemplo 1.3. Para sumar $1 \equiv 1.0 \times 2^0$ con $2 \equiv 1.0 \times 2^1$ se escribe

$$\left\{\begin{array}{rcl} 0.10 \overset{22)}{\cdots} 0 & \times & 2^1 \\ 1.00 \overset{22)}{\cdots} 0 & \times & 2^1 \\ \hline 1.10 \overset{22)}{\cdots} 0 & \times & 2^1 \end{array}\right.$$

Como se observa, sólo han hecho falta los dos primeros pasos. Sin embargo, si restamos a $1 \equiv 1.0 \times 2^0$ el número máquina más cercano a él por la izquierda, es decir, $1.1 \overset{23)}{\cdots} 1 \times 2^{-1}$, tenemos:

$$\text{En primer lugar:} \quad \left\{\begin{array}{rcl} 1.0 \overset{23)}{\cdots} 00 & \times & 2^0 \\ 0.1 \overset{23)}{\cdots} 11 & \times & 2^0 \\ \hline 0.0 \overset{23)}{\cdots} 01 & \times & 2^0 \end{array}\right.$$

$$\text{En segundo lugar:} \quad 1.0 \overset{23)}{\cdots} 0 \times 2^{-24}$$

Nótese que para realizar la operación de forma exacta hemos tenido que utilizar una posición más de memoria (24 bits) para guardar el segundo número y el resultado. Esto nos hace ver que las operaciones no se pueden llevar a cabo en palabras, sino que hace falta almacenar provisionalmente los operandos y el resultado en pedazos "más largos" de memoria. De hecho, en la aritmética estándar en coma flotante se utilizan *registros* de 80 bits para realizar las operaciones. □

Para la multiplicación (o división) de dos números no es necesario igualar los exponentes; basta multiplicar las dos mantisas (lo que dará un número de, a lo más, 48 cifras significativas, que cabe en un registro) y se suman los exponentes, normalizando el resultado, redondeándolo si es preciso y, eventualmente, volviéndolo a normalizar.

[6] Por ejemplo, el redondeo de $1.1 \overset{24)}{\cdots} 1 \times 2^E$ es 10.0×2^E que debe normalizarse a $1.0 \times 2^{E+1}$.

En cualquier caso, este tipo de precauciones garantizan, únicamente, que la aritmética en coma flotante se realiza siempre del modo correcto, en el sentido de que el resultado almacenado tras operar dos números máquina coincida con el redondeo del resultado real de la operación. No obstante, esto no impide que se produzcan situaciones anómalas que son intrínsecas al hecho de estar trabajando con una precisión determinada, como por ejemplo:

a) Las operaciones en coma flotante no verifican, en general, la propiedad asociativa. Así, para la suma,

$$(1 \oplus 2^{-24}) \oplus 2^{-24} = r\left(r\left(1 + 2^{-24}\right) + 2^{-24}\right) = r\left(1 + 2^{-24}\right) = 1$$

mientras que

$$\begin{aligned} 1 \oplus (2^{-24} \oplus 2^{-24}) &= r\left(1 + r\left(2^{-24} + 2^{-24}\right)\right) = r\left(1 + 2^{-23}\right) \\ &= 1 + 2^{-23} = 1.0 \overset{22)}{\cdots} 01 \times 2^0. \end{aligned}$$

b) Si dos números máquina x e y verifican $x \oplus y = x$, esto no implica, en general, que $y = 0$. De hecho, basta tomar cualquier número máquina y de la forma

$$y = \theta x \ \text{ con } \ 0 < \theta < 2^{-24}$$

para que se dé la igualdad $x \oplus y = x$. En efecto,

$$x \oplus y = r(x + y) = r(x + \theta x) = r(x) = x,$$

ya que al efectuar la suma entre

$$x = 1.a_1 a_2 \cdots a_{23} \times 2^{E_x}$$

e

$$y = 1.b_1 b_2 \cdots b_{23} \times 2^{E_y} \ \text{ con } \ E_y < E_x - 24$$

queda

$$\left\{ \begin{array}{lcl} 1.a_1 a_2 \cdots a_{23} & \times & 2^{E_x} \\ 0.0 \overset{23)}{\cdots} 00 \overset{E_x - 24 - E_y)}{\cdots} 01 b_1 b_2 \cdots b_{23} & \times & 2^{E_x} \\ \hline 1.a_1 a_2 \cdots a_{23} 0 \overset{E_x - 24 - E_y)}{\cdots} 01 b_1 b_2 \cdots b_{23} & \times & 2^{E_x}. \end{array} \right.$$

El resultado de la suma es un número cuyo redondeo es

$$1.a_1 a_2 \cdots a_{23} \times 2^{E_x} = x.$$

Otro fenómeno a tener en cuenta es lo que se conoce como *cancelación* y aparece al restar dos números muy próximos entre sí, lo que puede producir grandes errores relativos.

Ejemplo 1.4. Al realizar la resta en el ordenador de los números

$$x = 1.1 \overset{23)}{\cdots} 101 \overset{15)}{\cdots} 1 \times 2^E \quad \text{e} \quad y = 1.1 \overset{15)}{\cdots} 1 \times 2^E$$

(teniendo en cuenta que y es ya un número máquina) se obtiene

$$\begin{aligned} x \ominus y = r(r(x) - y) &= r\left(1.1 \overset{23)}{\cdots} 1 \times 2^E - 1.1 \overset{15)}{\cdots} 1 \times 2^E\right) \\ &= r\left(0.0 \overset{15)}{\cdots} 01 \overset{8)}{\cdots} 1 \times 2^E\right) = 1.1 \overset{7)}{\cdots} 1 \times 2^{E-16}, \end{aligned}$$

mientras que el resultado exacto es

$$x - y = 0.0 \overset{15)}{\cdots} 01 \overset{8)}{\cdots} 101 \overset{15)}{\cdots} 1 \times 2^E = 1.1 \overset{7)}{\cdots} 101 \overset{15)}{\cdots} 1 \times 2^{E-16},$$

¡que es un número máquina! Como se observa, el resultado hallado coincide con el real en tan sólo las 9 primeras cifras significativas; esto se produce porque al normalizar $0.0 \overset{15)}{\cdots} 01 \overset{8)}{\cdots} 1 \times 2^E$ se "corre la coma" 16 lugares, añadiendo 16 ceros por la derecha que son independientes de los valores de partida. □

Veamos otro ejemplo patológico:

Ejemplo 1.5. Dada la función $f(x) = \operatorname{sen} x$, si se utiliza el cociente incremental

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

en $x = 1$ para aproximar

$$\cos 1 = f'(1) = \lim_{h \to 0} \frac{\operatorname{sen}(1+h) - \operatorname{sen} 1}{h}$$

se puede comprobar que, trabajando en doble precisión, se obtienen los resultados de la tabla 1.2 para diversos valores de h, donde se han comparado los resultados obtenidos con el valor de $\cos 1 \simeq 0.54030230586814$. ¿Qué es lo que está ocurriendo? Lo que sucede es que a medida que h va decreciendo, los números $\operatorname{sen}(1+h)$ y $\operatorname{sen} 1$ van teniendo, cada vez, más cifras significativas iguales y, por tanto, su diferencia va teniendo cada vez menor número de cifras exactas (de hecho, para valores de h menores o iguales a 10^{-16} la diferencia $\operatorname{sen}(1+h) - \operatorname{sen} 1$ da, en el ordenador, el valor 0). □

La consecuencia que podemos extraer de los ejemplos 1.4 y 1.5 es que, en la medida de lo posible, deben intentar evitarse sustracciones de números muy cercanos. A veces hay que prever estos problemas y cambiar el orden de las operaciones con vistas a obtener representaciones matemáticas equivalentes que eviten las cancelaciones.

TABLA 1.2:
Valores aproximados de $\cos 1$ *mediante el cociente* $\frac{\operatorname{sen}(1+h)-\operatorname{sen} 1}{h}$

h	Valor aproximado	Error absoluto	Error relativo (%)
10^{-5}	0.54029809850586	0.00000420736228	0.00077870522286
10^{-6}	0.54030188512133	0.00000042074681	0.00007787248080
10^{-7}	0.54030226404045	0.00000004182769	0.00000774153481
10^{-8}	0.54030230289825	0.00000000296989	0.00000054967103
10^{-9}	0.54030235840941	0.00000005254127	0.00000972442009
10^{-10}	0.54030224738710	0.00000005848104	0.00001082376215
10^{-11}	0.54030113716408	0.00000116870406	0.00021630558456
10^{-12}	0.54034554608506	0.00004324021692	0.00800296731187
10^{-13}	0.53956838996783	0.00073391590031	0.13583430837565
10^{-14}	0.54400928206633	0.00370697619819	0.68609298126735
10^{-15}	0.55511151231258	0.01480920644444	2.74091120537484
10^{-16}	0	0.54030230586814	100

Ejemplo 1.6.

1. Como es sabido, las raíces de la ecuación de segundo grado

$$ax^2 + bx + c = 0 \quad (a, b, c \in \mathbb{R},\ a \neq 0)$$

son

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{y} \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Ahora bien, cuando $b > 0$ y $0 < 4ac \ll b^2$, resulta conveniente utilizar la siguiente expresión equivalente para el cálculo de la primera raíz:

$$x_1 = \frac{\left(\sqrt{b^2 - 4ac} - b\right)\left(\sqrt{b^2 - 4ac} + b\right)}{2a\left(\sqrt{b^2 - 4ac} + b\right)} = -\frac{2c}{b + \sqrt{b^2 - 4ac}}.$$

2. En lugar de trabajar con la función

$$f(x) = \frac{x^4 - x^3 - x + 1}{(x-1)^2},\ x \in \mathbb{R} \tag{1.2}$$

cuando $x \simeq 1$, podemos escribir ésta en forma equivalente como

$$f(x) = \frac{(x-1)^2(x^2 + x + 1)}{(x-1)^2} = x^2 + x + 1. \tag{1.3}$$

Como se observa en la tabla 1.3, mediante la expresión (1.2) evaluamos f de forma errónea, debido al fenómeno de la cancelación. □

TABLA 1.3:
Valores de f en puntos próximos a $x = 1$

Abscisa x	Expresión (1.2)	Expresión (1.3)
1.00000762939453	3.00002288818359	3.00002288824180
1.00000381469727	3.00001525878906	3.00001144410635
1.00000190734863	3.00000000000000	3.00000572204954
1.00000095367432	3.00000000000000	3.00000286102386
1.00000047683716	3.00000000000000	3.00000143051170
1.00000023841858	3.00000000000000	3.00000071525579
1.00000011920929	3.00000000000000	3.00000035762788
1.00000005960464	3.00000000000000	3.00000017881394
1.00000002980232	3.00000000000000	3.00000008940697
1.00000001490116	3.00000000000000	3.00000004470348
1.00000000745058	4.00000000000000	3.00000002235174
1.00000000372529	0	3.00000001117587

1.5. Propagación del error

Según hemos visto, a la hora de resolver un problema lo primero que hay que tener en cuenta es que los datos de partida con los que va a trabajar el ordenador no tienen por qué ser exactamente los datos originales (debido al redondeo). La segunda cuestión relevante es que, aunque los datos se almacenaran de forma exacta en el ordenador, en el momento que empezamos a trabajar con ellos se comienzan a cometer errores, de forma que, en algunas ocasiones, el resultado obtenido dista mucho del deseado. Para clarificar lo anterior analicemos una simple operación aritmética como es la suma de dos números: dados dos números $x, y \in \mathbb{R}$, para calcular su suma $x + y$ el ordenador halla

$$\begin{aligned} x \oplus y &= r\left(r(x) + r(y)\right) = r\left((1+\delta_1)x + (1+\delta_2)y\right) \\ &= (1+\delta_3)\left[(1+\delta_1)x + (1+\delta_2)y\right] \\ &= (1+\delta_3)(x+y) + (1+\delta_3)\delta_1 x + (1+\delta_3)\delta_2 y \end{aligned} \tag{1.4}$$

con

$$|\delta_i| \leq 2^{-24} = \frac{eps}{2}$$

para $i = 1, 2, 3$. Vemos así que los errores de redondeo en los datos se propagan al resultado de la operación. Si ahora este resultado fuera un operando de otra

operación, este error se propagaría al consiguiente resultado, y así sucesivamente. De esta forma, si pensamos en un algoritmo que involucre una gran cantidad de operaciones elementales, las perspectivas pueden parecer no muy buenas.

En general, si denotamos por $*$ cualquier operación elemental $(+, -, \cdot, /)$, para cada par de números máquina x e y, se tendrá

$$r(x * y) = (1 + \delta)(x * y) \;\text{ con }\; |\delta| \leq \frac{eps}{2}.$$

Si x e y no son números máquina, se tiene

$$r(r(x) * r(y)) = r\left([(1 + \delta_1)x] * [(1 + \delta_2)]y\right) = (1 + \delta_3)\left((1 + \delta_1)x * (1 + \delta_2)y\right).$$

Para operaciones compuestas las cosas se complican. Supongamos que queremos calcular $x(y + z)$ donde x, y y z son números máquina. El resultado es

$$\begin{aligned} x \otimes (y \oplus z) &= r\left(xr(y + z)\right) = (1 + \delta_1)\left(xr(y + z)\right) \\ &= (1 + \delta_1)\left[(1 + \delta_2)x(y + z)\right] = (1 + \delta_1 + \delta_2 + \delta_1\delta_2)x(y + z) \\ &\simeq (1 + \delta_1 + \delta_2)x(y + z) = (1 + \delta_3)x(y + z) \end{aligned}$$

con $|\delta_3| \leq |\delta_1| + |\delta_2| \leq 2^{-24} + 2^{-24} = 2^{-23}$.

Parece claro, por tanto, que debemos estudiar cuánto influye la propagación del error en el resultado final del problema. Dos son los principales conceptos ligados a este estudio:

a) El *condicionamiento*. Mide la influencia que tendrían en el resultado eventuales errores en los datos, en el caso ideal de que se pudiese trabajar con aritmética exacta; está ligado, por tanto, al problema en sí y no depende del algoritmo que se utilice para resolverlo.

b) La *estabilidad*. Está relacionada con la influencia que tiene en los resultados finales la acumulación de los errores que se producen en las sucesivas operaciones elementales que se llevan a cabo para resolver el problema; es decir, depende del algoritmo que se utilice para obtener la solución.

Así pues, hablaremos de un problema bien o mal condicionado, mientras que, dado un algoritmo para resolver un problema, diremos si es o no numéricamente estable. Ambos conceptos son, en general, difíciles de analizar.

1.5.1. Condicionamiento

Diremos que un problema está *mal condicionado* cuando pequeños cambios en los datos dan lugar a grandes cambios en las respuestas. Las técnicas que se

emplean en el estudio del condicionamiento de un problema están fuertemente ligadas a la estructura del mismo. En general, a la hora de resolver un problema $y = f(x)$ se intenta definir un *número de condición* $\kappa = \kappa(x) \geq 0$ de forma que

$$\frac{||f(\widetilde{x}) - f(x)||}{||f(x)||} \simeq \kappa(x) \frac{||\widetilde{x} - x||}{||x||}. \tag{1.5}$$

Este número indicará, según sea cercano a 1 o alejado de éste, si el problema está bien o mal condicionado, respectivamente. Si el número de condición es menor que 1 o está próximo a 1, el error del dato no se amplificará mucho y el error del resultado será, a lo sumo, del mismo orden que el error en el dato; por el contrario, si este número de condición toma valores muy grandes, el error final será una gran amplificación del error en el dato.

Para casos concretos, podremos definir fácilmente un número de condición. Comencemos con un problema sencillo: la evaluación de una función diferenciable $f : \mathbb{R} \to \mathbb{R}$ en un punto x. Si en lugar de x tomamos una aproximación suya $\widetilde{x}$ con $|\widetilde{x} - x| \ll 1$ (por ejemplo su redondeo), el teorema del Valor Medio asegura que

$$f(\widetilde{x}) - f(x) = f'(\xi)(\widetilde{x} - x) \simeq f'(x)(\widetilde{x} - x).$$

De esta forma, si $f'(x)$ no es muy grande, el efecto de la perturbación sobre $f(x)$ es pequeño. Concretamente, el error relativo de la perturbación viene dado por

$$\overbrace{\left|\frac{f(\widetilde{x}) - f(x)}{f(x)}\right|}^{\text{Error relativo (resultados)}} \simeq \underbrace{\left|\frac{f'(x)}{f(x)} x\right|}_{\text{Número de condición}} \overbrace{\left|\frac{\widetilde{x} - x}{x}\right|}^{\text{Error relativo (datos)}}$$

Ejemplo 1.7. Consideremos la función

$$f(x) = 1 - \sqrt{1 - x^2},\ -1 \leq x \leq 1.$$

Como

$$f'(x) = \frac{x}{\sqrt{1 - x^2}},\ -1 < x < 1$$

entonces el número de condición de f es

$$\left|\frac{f'(x)}{f(x)}x\right| = \frac{x^2}{\sqrt{1-x^2}(1-\sqrt{1-x^2})}$$
$$= \frac{x^2(1+\sqrt{1-x^2})}{\sqrt{1-x^2}(1-\sqrt{1-x^2})(1+\sqrt{1-x^2})}$$
$$= \frac{1+\sqrt{1-x^2}}{\sqrt{1-x^2}}$$

que es muy grande para valores de x próximos a ± 1. □

Ejemplo 1.8. Supongamos que $\widetilde{x}_1$ y $\widetilde{x}_2$ son aproximaciones de x_1 y x_2 con errores ε_1 y ε_2 respectivamente, es decir, $\widetilde{x}_1 = x_1 + \varepsilon_1$ y $\widetilde{x}_2 = x_2 + \varepsilon_2$. Veamos que el error relativo del producto es, aproximadamente, igual a la suma de los errores relativos de los factores, esto es,

$$\frac{\widetilde{x}_1\widetilde{x}_2 - x_1x_2}{x_1x_2} \sim \frac{\varepsilon_1}{x_1} + \frac{\varepsilon_2}{x_2} \tag{1.6}$$

y, por tanto, el producto de dos números es siempre un problema bien condicionado. En efecto, como

$$\widetilde{x}_1\widetilde{x}_2 = (x_1+\varepsilon_1)(x_2+\varepsilon_2) = x_1x_2 + x_1\varepsilon_2 + x_2\varepsilon_1 + \varepsilon_1\varepsilon_2$$

entonces

$$\widetilde{x}_1\widetilde{x}_2 - x_1x_2 \sim x_1\varepsilon_2 + x_2\varepsilon_1,$$

de donde se sigue (1.6).

Sin embargo, la suma de dos números no va a ser siempre un problema bien condicionado. Teniendo en cuenta que

$$\widetilde{x}_1 + \widetilde{x}_2 = (x_1+\varepsilon_1) + (x_2+\varepsilon_2) = x_1 + x_2 + \varepsilon_1 + \varepsilon_2$$

se verifica que

$$(\widetilde{x}_1 + \widetilde{x}_2) - (x_1 + x_2) = \varepsilon_1 + \varepsilon_2,$$

de donde

$$\frac{(\widetilde{x}_1 + \widetilde{x}_2) - (x_1 + x_2)}{x_1 + x_2} = \frac{\varepsilon_1}{x_1 + x_2} + \frac{\varepsilon_2}{x_1 + x_2} = \frac{x_1}{x_1 + x_2}\frac{\varepsilon_1}{x_1} + \frac{x_2}{x_1 + x_2}\frac{\varepsilon_2}{x_2}.$$

Nótese que el problema para la suma de dos números está bien condicionado salvo cuando $x_1 \simeq -x_2$, en cuyo caso el condicionamiento puede ser muy malo. Esto explica el fenómeno ya estudiado de la cancelación. □

Otro tipo de problemas para los que se puede dar un número de condición de forma relativamente sencilla es la resolución de sistemas lineales $Ax = b$, como veremos en el capítulo 3. Anticipamos, a continuación, un ejemplo de sistema lineal mal condicionado.

Ejemplo 1.9 (R. S. Wilson). Consideremos el sistema lineal $Ax = b$ donde b es el vector

$$b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$$

y A es la aparentemente "inofensiva" matriz simétrica

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$$

que tiene por matriz inversa a

$$A^{-1} = \begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}$$

y cuyo determinante vale 1. La solución exacta de dicho sistema es

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Si consideramos las siguientes perturbaciones de los datos A y b

$$\widetilde{A} = \begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix} \text{ y } \widetilde{b} = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix},$$

las soluciones exactas de los sistemas lineales $\widetilde{A}y = b$ y $Az = \widetilde{b}$ vienen dadas, respectivamente, por

$$y = \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix} \text{ y } z = \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix}.$$

Como se aprecia, pequeños cambios en el dato A proporcionan un resultado muy alejado de la solución original, x. Análogamente, cuando se perturba ligeramente el dato b se obtiene también un resultado z también distante de x.

La justificación de estas propiedades sorprendentes, así como la forma precisa de medir el tamaño de las perturbaciones y de los errores, se verán en el capítulo 3 (véase, en particular, el ejemplo 3.1). □

1.5.2. Estabilidad

En términos generales, diremos que un algoritmo es *inestable* cuando los errores que se van produciendo en las diversas etapas van aumentando en etapas posteriores y afectan, en gran manera, a la exactitud del cálculo en su conjunto; un algoritmo es *estable* cuando no es *inestable*. Otra forma de aproximarse al concepto de estabilidad es pensar un *algoritmo estable* como aquel que consigue resultados cercanos a los exactos cuando se parte de datos próximos a los verdaderos.

Ejemplo 1.10. Para ilustrar el comentario anterior vamos a considerar dos algoritmos para el cálculo de $\left(\frac{1}{7}\right)^{100}$. Cada uno de ellos es una aplicación particular de dos algoritmos más generales que sirven para calcular las potencias sucesivas de un número positivo $\lambda > 0$, a saber

$$\begin{cases} x_0 = 1 \\ x_n = \lambda x_{n-1},\ n \in \mathbb{N} \end{cases} \quad \text{y} \quad \begin{cases} x_0 = 1,\ x_1 = \lambda \\ x_{n+1} = (3+\lambda)x_n - 3\lambda x_{n-1},\ n \in \mathbb{N}. \end{cases}$$

En ambos casos, el término general de la sucesión $\{x_n\}_{n=0}^{\infty}$ es

$$x_n = \lambda^n,\ n \in \mathbb{N} \cup \{0\}.$$

En efecto, en el primero de ellos se comprueba de forma inmediata y en el segundo basta observar que $x_0 = 1$, $x_1 = \lambda$ y para cada $n \in \mathbb{N}$ se verifica que

$$(3+\lambda)x_n - 3\lambda x_{n-1} = (3+\lambda)\lambda^n - 3\lambda\lambda^{n-1} = 3\lambda^n + \lambda^{n+1} - 3\lambda^n = \lambda^{n+1} = x_{n+1}.$$

En nuestro caso concreto queremos hallar el término x_{100} para la elección $\lambda = \frac{1}{7}$, por lo que basta calcular los 100 primeros términos de las sucesiones anteriores para dicho valor de λ, esto es,

$$(\varphi_1) \begin{cases} x_0 = 1 \\ x_n = \dfrac{x_{n-1}}{7},\ n = 1, 2, \ldots, 100 \end{cases}$$

y

$$(\varphi_2)\begin{cases} x_0 = 1,\ x_1 = \dfrac{1}{7} \\ x_{n+1} = \dfrac{22}{7}x_n - \dfrac{3}{7}x_{n-1},\ n = 1, 2, \ldots, 99. \end{cases}$$

Si se calculan los sucesivos valores de x_n mediante ambos algoritmos se obtienen los resultados expuestos en la tabla 1.4. Estos resultados muestran el buen comportamiento del algoritmo φ_1 frente al pésimo comportamiento de φ_2. □

TABLA 1.4:
Algoritmos φ_1 y φ_2 para el cálculo de 7^{-100} en doble precisión

n	Valor de x_n (φ_1)	Valor de x_n (φ_2)
0	1	1
1	0.1428571428571	0.1428571428571
2	$2.0408163265306 \times 10^{-2}$	$2.0408163265306 \times 10^{-2}$
3	$2.9154518950437 \times 10^{-3}$	$2.9154518950436 \times 10^{-3}$
4	$4.1649312786339 \times 10^{-4}$	$4.1649312786308 \times 10^{-4}$
11	$5.0573331066306 \times 10^{-10}$	$5.0505834009817 \times 10^{-10}$
12	$7.2247615809009 \times 10^{-11}$	$7.0222704114337 \times 10^{-11}$
13	$1.0321087972716 \times 10^{-11}$	$4.2463528886988 \times 10^{-12}$
14	$1.4744411389594 \times 10^{-12}$	$-1.6749764113091 \times 10^{-11}$
15	$2.1063444842277 \times 10^{-13}$	$-5.4461981307728 \times 10^{-11}$
16	$3.0090635488967 \times 10^{-14}$	$-1.6398775663296 \times 10^{-10}$
25	$7.4567399858374 \times 10^{-22}$	$-3.2283632877849 \times 10^{-6}$
30	$4.4366870862363 \times 10^{-26}$	$-7.8449227893174 \times 10^{-4}$
34	$1.8478496818977 \times 10^{-29}$	$-6.3543874593471 \times 10^{-2}$
35	$2.6397852598538 \times 10^{-30}$	-0.1906316237804
38	$7.6961669383493 \times 10^{-33}$	-5.1470538420712
40	$1.5706463139488 \times 10^{-34}$	$-4.6323484578640 \times 10$
45	$9.3451913723379 \times 10^{-39}$	$-1.1256606752610 \times 10^{4}$
50	$5.5602971216386 \times 10^{-43}$	$-2.7353554408841 \times 10^{6}$
60	$1.9684192301176 \times 10^{-51}$	$-1.6152000342877 \times 10^{11}$
70	$6.9684662181415 \times 10^{-60}$	$-9.5375946824653 \times 10^{15}$
80	$2.4669298435211 \times 10^{-68}$	$-5.6318542840489 \times 10^{20}$
90	$8.7332601785620 \times 10^{-77}$	$-3.3255536361880 \times 10^{25}$
100	$3.0916904080902 \times 10^{-85}$	$-1.9637061666327 \times 10^{30}$

Formalicemos a continuación el concepto de estabilidad. Para ello introducimos primero la siguiente notación debida a *Landau*.

Notación 1.1. Utilizaremos la expresión $\phi(r) = O(r)$ para denotar que la cantidad $\phi(r)$ es del orden de r para r pequeño, es decir, cuando existe una constante $C > 0$ tal que

$$|\phi(r)| \leq Cr,\ 0 < r \ll 1. \quad \square$$

Definición 1.1. Un algoritmo φ de resolución del problema $y = f(x)$ es *estable* si existe $\widetilde{x}$ con

$$\frac{||\widetilde{x} - x||}{||x||} = O(eps)$$

verificando

$$\frac{||\varphi(x) - f(\widetilde{x})||}{||f(\widetilde{x})||} = O(eps). \tag{1.7}$$

En caso contrario se dice que el algoritmo φ es *inestable*. $\square$

Observación 1.8. Nótese que la expresión (1.7) indica que un algoritmo estable φ proporciona un resultado relativamente cercano (*i.e.*, con error relativo del orden del épsilon de la máquina) al resultado correspondiente a un dato relativamente cercano (en el sentido anterior) al dato original. $\square$

Ejemplo 1.11. Estudiemos la estabilidad de los algoritmos φ_1 y φ_2 del ejemplo 1.10. En el algoritmo φ_2 es claro que $f(\widetilde{\lambda}) = \widetilde{\lambda}^{100}$ es muy pequeño para cualquier valor $\widetilde{\lambda}$ que esté cerca de $\frac{1}{7}$. De esta forma, con los datos de la tabla 1.4, se puede comprobar que

$$\frac{\left|\varphi_2\left(\frac{1}{7}\right) - f(\widetilde{\lambda})\right|}{\left|f(\widetilde{\lambda})\right|} \simeq 10^{114}$$

para cualquier $\widetilde{\lambda}$ próximo a $\frac{1}{7}$, lo que hace que φ_2 sea un algoritmo inestable.

En el estudio de la estabilidad del algoritmo φ_1 vamos a afinar un poco más;

los valores sucesivos que se van obteniendo con él son:

$$\begin{cases} x_0 = 1 \\ x_1 = r\left(\dfrac{1}{7}\right) \\ x_2 = \left(r\left(\dfrac{1}{7}\right)\right)^2 (1+\delta_1) \\ x_3 = \left(\left(r\left(\dfrac{1}{7}\right)\right)^2 (1+\delta_1) r\left(\dfrac{1}{7}\right)\right)(1+\delta_2) = \left(r\left(\dfrac{1}{7}\right)\right)^3 (1+\delta_1)(1+\delta_2) \\ \quad \cdots \\ x_{100} = \left(r\left(\dfrac{1}{7}\right)\right)^{100} (1+\delta_1)(1+\delta_2)\cdots(1+\delta_{99}) \end{cases}$$

donde $r\left(\dfrac{1}{7}\right)$ denota el redondeo de $\dfrac{1}{7}$ y

$$|\delta_i| \leq \frac{eps}{2}$$

para $i = 1, 2, \ldots, 99$. De esta forma, eligiendo

$$\widetilde{\lambda} = r\left(\frac{1}{7}\right) \sqrt[100]{(1+\delta_1)(1+\delta_2)\cdots(1+\delta_{99})}$$

se tiene que

$$\frac{\left|\widetilde{\lambda} - \dfrac{1}{7}\right|}{\left|\dfrac{1}{7}\right|} = O(eps) \;\text{ y }\; \varphi_1\left(\frac{1}{7}\right) = f(\widetilde{\lambda}) \tag{1.8}$$

con lo que

$$\frac{\left|\varphi_1\left(\dfrac{1}{7}\right) - f(\widetilde{\lambda})\right|}{\left|f(\widetilde{\lambda})\right|} = 0$$

y, por tanto, el algoritmo φ_1 es estable. □

Observación 1.9. El comportamiento asintótico hacia $-\infty$ de la sucesión definida en el algoritmo φ_2 puede justificarse de la siguiente forma. Dicho algoritmo es, a su vez, una particularización de la ley de recurrencia más general

$$\begin{cases} x_0 = \alpha + \beta,\ x_1 = \dfrac{\alpha}{7} + 3\beta \\ x_{n+1} = \dfrac{22}{7} x_n - \dfrac{3}{7} x_{n-1},\ n \in \mathbb{N} \end{cases} \qquad (\alpha, \beta \in \mathbb{R})$$

que tiene como término general

$$x_n = \alpha \left(\frac{1}{7}\right)^n + \beta 3^n,\ n \in \mathbb{N} \cup \{0\}.$$

Nuestro caso se corresponde con la elección $\alpha = 1$ y $\beta = 0$. No obstante, al tomar la máquina los datos iniciales redondeados

$$\widetilde{x}_0 = 1 \ \text{ y } \ \widetilde{x}_1 = r\left(\frac{1}{7}\right) = 0.14285714285714,$$

los valores de α y β correspondientes vienen dados por la solución del sistema lineal

$$\begin{cases} \alpha + \beta = 1 \\ \dfrac{\alpha}{7} + 3\beta = 0.14285714285714, \end{cases}$$

es decir,

$$\alpha = \frac{21 - 7r\left(\frac{1}{7}\right)}{20} = 1 + 10^{-15} \ \text{ y } \ \beta = \frac{7r\left(\frac{1}{7}\right) - 1}{20} = -10^{-15}.$$

El hecho de que $\beta \neq 0$ hace que para valores grandes de n la cantidad $\beta 3^n$ sea, en valor absoluto, muy grande y perturbe desmesuradamente los resultados finales. □

Una técnica estándar para estudiar la estabilidad de un algoritmo es lo que se denomina *análisis regresivo del error*: mediante las manipulaciones necesarias, se intenta escribir el resultado obtenido por el algoritmo como el verdadero valor que se obtendría si se aplicasen operaciones exactas a otros datos distintos; si la diferencia entre estos últimos y los reales no es grande, el algoritmo será *regresivamente estable*. Más concretamente,

Definición 1.2. Un algoritmo φ de resolución del problema $y = f(x)$ es *regresivamente estable* si existe $\widetilde{x}$ con

$$\frac{||\widetilde{x} - x||}{||x||} = O(eps)$$

verificando

$$\varphi(x) = f(\widetilde{x}). \quad \square \tag{1.9}$$

Observación 1.10. Claramente, un algoritmo regresivamente estable es estable puesto que (1.9) implica

$$\frac{||\varphi(x) - f(\widetilde{x})||}{||f(\widetilde{x})||} = 0$$

que, obviamente, es del orden del épsilon de la máquina. Un algoritmo regresivamente estable φ proporciona, por tanto, un resultado exacto correspondiente a un dato relativamente cercano (en el sentido anteriormente explicado) al dato verdadero. Es precisamente el hecho de que el error cometido se proyecte hacia atrás sobre los datos lo que justifica el calificativo de regresivo. □

Ejemplo 1.12. El algoritmo φ_1 considerado en el ejemplo 1.10 es regresivamente estable (véase (1.8)). □

Los conceptos de condicionamiento y estabilidad regresiva nos permiten estudiar, de manera sencilla, la *precisión* que un algoritmo proporciona a la hora de resolver un problema concreto.

Teorema 1.1. *Supongamos que queremos resolver un problema $y = f(x)$, cuyo número de condición es $\kappa(x) \geq 0$, mediante un algoritmo φ regresivamente estable. Entonces se verifica que*

$$\frac{||\varphi(x) - f(x)||}{||f(x)||} = O(\kappa(x)\ eps),$$

es decir, el error relativo cometido es del orden del número de condición multiplicado por el épsilon de la máquina.

Demostración. Por ser φ un algoritmo regresivamente estable se tiene que

$$\varphi(x) = f(\widetilde{x})$$

para un cierto $\widetilde{x}$ verificando

$$\frac{||\widetilde{x} - x||}{||x||} = O(eps)$$

(véase (1.9)). A partir de la relación (1.5), se concluye que

$$\frac{||\varphi(x) - f(x)||}{||f(x)||} = \frac{||f(\widetilde{x}) - f(x)||}{||f(x)||} \simeq \kappa(x)\frac{||\widetilde{x} - x||}{||x||} = O(\kappa(x)\ eps). \quad \square$$

Observación 1.11. El teorema 1.1 aporta información sobre la precisión del algoritmo en el siguiente sentido: determina el orden del error relativo que se comete cuando se toma el resultado que el algoritmo proporciona a partir del dato exacto, en lugar de tomar el resultado exacto del problema (que, como ya se ha indicado, es imposible de calcular en la mayoría de los casos). □

Ejemplo 1.13. Como el número de condición de la función

$$\psi(x) = x^{100},\ x > 0$$

es constante para todo $x > 0$ e igual a

$$\kappa = \left| x \frac{\psi'(x)}{\psi(x)} \right| = 100,$$

se tiene que el error relativo cometido al usar el algoritmo φ_1 considerado en el ejemplo 1.10 es del orden de 100 veces el épsilon de la máquina. □

1.6. Problemas

1.6.1. Problemas resueltos

1.1. Encontrar la expresión decimal de los números binarios 11.11 y $111.\overbrace{101}$.

SOLUCIÓN. Claramente se tiene que

$$(11.11)_2 = 1 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} = 2 + 1 + \frac{1}{2} + \frac{1}{4} = (3.75)_{10}$$

y

$$\begin{aligned}
\left(111.\overbrace{101}\right)_2 &= 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 + 1 \times 2^{-1} + 0 \times 2^{-2} \\
&\quad + 1 \times 2^{-3} + 1 \times 2^{-4} + 0 \times 2^{-5} + 1 \times 2^{-6} + \cdots \\
&= 4 + 2 + 1 + (4 + 0 + 1) \times 2^{-3} + (4 + 0 + 1) \times 2^{-6} \\
&\quad + (4 + 0 + 1) \times 2^{-9} + (4 + 0 + 1) \times 2^{-12} + \cdots \\
&= 7 + 5 \sum_{n=1}^{\infty} \frac{1}{2^{3n}} = 7 + 5 \sum_{n=1}^{\infty} \left(\frac{1}{8}\right)^n = 7 + 5 \frac{\frac{1}{8}}{1 - \frac{1}{8}} \\
&= 7 + \frac{5}{7} = \left(7.\overbrace{714285}\right)_{10}. \quad \square
\end{aligned}$$

1.2. Encontrar la expresión binaria de los números decimales 0.1 y $5.\overbrace{3}$.

SOLUCIÓN. En primer lugar, veamos el algoritmo general para convertir un número decimal x con $0 < x < 1$ a binario en la forma

$$x = (0.\alpha_1\alpha_2 \ldots \alpha_p \ldots)_2.$$

Como

$$x = \frac{\alpha_1}{2} + \frac{\alpha_2}{2^2} + \cdots + \frac{\alpha_p}{2^p} + \cdots$$

se tiene que

$$\alpha_1 = 2x - \left(\frac{\alpha_2}{2} + \frac{\alpha_3}{2^2} + \cdots + \frac{\alpha_{p+1}}{2^p} + \cdots\right) = 2x - (0.\alpha_2\alpha_3 \ldots \alpha_{p+1} \ldots)_2,$$

es decir, la primera cifra decimal α_1 de x en base 2 se halla tomando la parte entera de $2x$. Ahora bien, como el número $(0.\alpha_2\alpha_3 \ldots \alpha_{p+1} \ldots)_2$ está también entre 0 y 1, para hallar la siguiente cifra α_2 basta repetir el proceso y, así, sucesivamente. De esta forma, a partir de la siguiente tabla

$0.1 \times 2 = 0.2$	$\rightarrow$	0
$0.2 \times 2 = 0.4$	$\rightarrow$	0
$0.4 \times 2 = 0.8$	$\rightarrow$	0
$0.8 \times 2 = 1.6$	$\rightarrow$	1
$0.6 \times 2 = 1.2$	$\rightarrow$	1
$0.2 \times 2 = 0.4$	$\rightarrow$	0
$\cdots$	$\cdots$	$\cdots$

se obtiene que

$$(0.1)_{10} = (0.0\overbrace{0011})_2$$

y, como $5.\overbrace{3} = 5 + 0.\overbrace{3}$, a partir de

$0.\overbrace{3} \times 2 = 0.\overbrace{6}$	$\rightarrow$	0
$0.\overbrace{6} \times 2 = 1.\overbrace{3}$	$\rightarrow$	1
$0.\overbrace{3} \times 2 = 0.\overbrace{6}$	$\rightarrow$	0
$\cdots$	$\cdots$	$\cdots$

se concluye que

$$(5.\overbrace{3})_{10} = (5)_{10} + (0.\overbrace{3})_{10} = (101)_2 + (0.\overbrace{01})_2 = (101.\overbrace{01})_2. \quad \square$$

1.3. Determinar los números decimales que en simple precisión tienen la siguiente representación en coma flotante estándar:

a) 11100101110010010000000000000000.

b) 00000011110001111100000000000000.

c) 11111111100000000000000000000000.

d) 10000000010000000000000000000000.

Solución.

a) $-1.1001001 \times 2^{203-127} = -1.1001001 \times 2^{76} \simeq -1.186494578820998 \times 10^{23}$.

b) $1.10001111 \times 2^{7-127} = 1.10001111 \times 2^{-120} \simeq 1.172555614945232 \times 10^{-36}$.

c) $-\infty$.

d) $-0.1 \times 2^{-126} = -1.0 \times 2^{-127} \simeq -5.877471754111438 \times 10^{-39}$. □

1.4. Supongamos que tenemos un ordenador que almacena los números en base 10 con tan sólo dos dígitos de mantisa. Queremos calcular con esta máquina la menor raíz de la ecuación $x^2 - 20x + 1 = 0$.

a) ¿Qué valor se obtendría al calcularla como $10 - \sqrt{99}$?

b) Ídem calculándola como $\dfrac{1}{10 + \sqrt{99}}$.

Solución.

a) Las raíces de la ecuación $x^2 - 20x + 1 = 0$ son

$$10 - \sqrt{99} = 0.05012562893429393\ldots \text{ y } 10 + \sqrt{99} = 19.94987437106570\ldots .$$

Trabajando en una máquina con una mantisa de 2 dígitos decimales, 99 es un número máquina para ella (9.9×10^1). Como

$$\sqrt{99} = 9.949874371065706\ldots$$

se tiene que $r(\sqrt{99}) = 9.9 \times 10^0$, con lo que, al ser también 10 un número máquina $(10 = 1.0 \times 10^1)$,

$$r(10 - \sqrt{99}) = 10 - 9.9 = 0.1.$$

b) Podemos evitar la cancelación anterior escribiendo

$$10 - \sqrt{99} = \frac{1}{10 + \sqrt{99}}.$$

Puesto que $r\left(10 + \sqrt{99}\right) = r\left(10 + 9.9\right) = 20$, dado que 19.9 no es un número máquina y hay que tomar su redondeo, se verifica

$$r\left(\frac{1}{10 + \sqrt{99}}\right) = r\left(\frac{1}{r\left(10 + \sqrt{99}\right)}\right) = r\left(\frac{1}{20}\right) = 0.05.$$

Nótese que éste es el mejor resultado que puede obtenerse en esa máquina para el cálculo de $10 - \sqrt{99}$ puesto que coincide con el redondeo del resultado exacto. □

1.5. Se considera el problema de sumar tres números reales $x, y, z \in \mathbb{R}$ con el algoritmo

$$(x + y) + z,$$

es decir, se suman los dos primeros y al resultado obtenido se le suma el tercero. Demostrar que este algoritmo es regresivamente estable.

SOLUCIÓN. Basta observar que, a partir de la relación (1.4), se tiene que

$$\begin{aligned}(x \oplus y) \oplus z &= r\left((x \oplus y) + r(z)\right) = (1+\delta_5)\left((x \oplus y) + (1+\delta_4)z\right)\\ &= (1+\delta_5)\left((1+\delta_3)(x+y) + (1+\delta_3)\delta_1 x + (1+\delta_3)\delta_2 y + (1+\delta_4)z\right)\\ &= \widetilde{x} + \widetilde{y} + \widetilde{z}\end{aligned}$$

donde

$$|\delta_i| \leq \frac{eps}{2}$$

para $i = 1, 2, 3, 4, 5$, y los números

$$\begin{cases} \widetilde{x} = (1 + \delta_1 + \delta_3 + \delta_5 + \delta_1\delta_3 + \delta_1\delta_5 + \delta_3\delta_5 + \delta_1\delta_3\delta_5)\, x \\ \widetilde{y} = (1 + \delta_2 + \delta_3 + \delta_5 + \delta_2\delta_3 + \delta_2\delta_5 + \delta_3\delta_5 + \delta_2\delta_3\delta_5)\, y \\ \widetilde{z} = (1 + \delta_4 + \delta_5 + \delta_4\delta_5)\, z \end{cases}$$

verifican

$$\left|\frac{\widetilde{x} - x}{x}\right| = O(eps), \ \left|\frac{\widetilde{y} - y}{y}\right| = O(eps) \ \text{ y } \ \left|\frac{\widetilde{z} - z}{z}\right| = O(eps). \quad \square$$

1.6. Estudiar el condicionamiento de la suma de n números $x_1 + x_2 + \cdots + x_n$.

SOLUCIÓN. Consideremos $\widetilde{x}_i$, perturbaciones de x_i,

$$\widetilde{x}_i = x_i + \varepsilon_i$$

con $i = 1, 2, \ldots, n$. El error absoluto en la suma es

$$\left|\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \widetilde{x}_i\right| = \left|\sum_{i=1}^{n} (x_i - \widetilde{x}_i)\right| = \left|\sum_{i=1}^{n} \varepsilon_i\right|$$

y una cota del error relativo viene dada por

$$\frac{\left|\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \widetilde{x}_i\right|}{\left|\sum_{i=1}^{n} x_i\right|} = \frac{\left|\sum_{i=1}^{n} \varepsilon_i\right|}{\left|\sum_{i=1}^{n} x_i\right|} \leq \frac{\sum_{i=1}^{n} |\varepsilon_i|}{\left|\sum_{i=1}^{n} x_i\right|} = \sum_{i=1}^{n} \frac{|x_i|}{\left|\sum_{i=1}^{n} x_i\right|} \left|\frac{\varepsilon_i}{x_i}\right|$$

por lo que los números de condición son

$$\kappa(x_i) = \frac{|x_i|}{\left|\sum_{i=1}^{n} x_i\right|}$$

para $i = 1, 2, \ldots, n$. Si todos los números $\{x_1, x_2, \ldots, x_n\}$ tienen el mismo signo entonces el problema está bien condicionado. En otro caso, si

$$\sum_{i=1}^{n} x_i \simeq 0$$

los números de condición pueden ser muy grandes, lo que ocurre, en particular, cuando los números están próximos a cancelarse entre sí. □

1.6.2. Problemas propuestos

1.7. *Arquímedes* (278–212 a.C.) obtuvo las siguientes acotaciones del número π

$$\frac{223}{71} < \pi < \frac{22}{7}.$$

Determinar los errores absolutos y relativos cometidos en esas aproximaciones.

1.8. Algunos ordenadores utilizan, en lugar del sistema binario, el sistema hexadecimal, es decir, utilizan como base el 16 y los dígitos que se emplean son

$$0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ \mathrm{A},\ \mathrm{B},\ \mathrm{C},\ \mathrm{D},\ \mathrm{E},\ \mathrm{F}.$$

Encontrar la expresión decimal de los números hexadecimales E, 1A, A9B.A1 y A.$\widehat{\mathrm{A}}$, así como su expresión binaria. Comprobar lo cómodo que resulta expresar números hexadecimales en binario y viceversa.

1.9. Hallar la representación en coma flotante estándar en precisión simple de:

a) Los números máquina de los problemas 1.1 y 1.2.

b) El redondeo de los números de los problemas 1.1 y 1.2 que no son números máquina.

1.10. Hallar el número de condición para las siguientes funciones:

$$f(x) = x^{\alpha}\ (\alpha \in \mathbb{R}),\ g(x) = \operatorname{sen} x \ \text{ y } \ h(x) = e^{x}.$$

1.11. Se considera el problema de sumar n números reales $x_1, x_2, \ldots, x_n \in \mathbb{R}$ con el algoritmo

$$\begin{cases} s_1 = x_1 \\ s_i = s_{i-1} + x_i,\ i = 2, 3, \ldots, n. \end{cases}$$

Demostrar que el algoritmo anterior es regresivamente estable.

1.7. Prácticas

1.1. Teniendo en cuenta que `MATLAB` trabaja en doble precisión, calcular el número máquina inmediatamente anterior a 1 y comprobar que dista 2^{-53} de 1.

1.2. Calcular $1 \oplus 2^{-52}$, $1 \oplus 2^{-53}$, $1 \ominus 2^{-53}$ y $1 \ominus 2^{-54}$ y comprobar que los resultados coinciden con los que teóricamente deben obtenerse.

1.3. Determinar el mayor y menor número máquina normal positivo, así como el mayor y menor número subnormal positivo, cuando se trabaja en simple precisión. Comprobar que coinciden con los expuestos en la tabla 1.1.

1.4. Ídem para doble precisión. Comparar con los comandos `realmax` y `realmin` de `MATLAB`.

1.5. Hallar la suma y la resta de dos números con exponentes muy dispares y comprobar que el resultado obtenido concuerda con el previsto según la aritmética en coma flotante estándar.

1.6. Hallar las potencias sucesivas de 10 con exponente desde 1 hasta 309. Comprobar que el valor de 10^{309} es infinito y que si $x = 10^{309}$ entonces $\dfrac{1}{x} = 0$.

1.7. Determinar el épsilon de la máquina. Para ello, calcular $1 + x$ con $x = 2^{-i}$ para $i = 1, 2, \ldots$ mientras que $1 + x > 1$. Comparar con el comando `eps` de `MATLAB`.

1.8. Consideremos la *sucesión de Fibonacci*

$$\begin{cases} x_0 = 1,\ x_1 = 2 \\ x_{n+1} = x_n + x_{n-1},\ n \in \mathbb{N}. \end{cases}$$

a) A partir de la sucesión $\{x_n\}_{n=0}^{\infty}$ se define

$$y_n = \frac{x_n}{x_{n-1}},\ n \in \mathbb{N}.$$

Obtener la ley de recurrencia de la sucesión $\{y_n\}_{n=1}^{\infty}$ y demostrar que

$$\lim_{n\to+\infty} y_n = \frac{1+\sqrt{5}}{2}.$$

El número $\phi = \dfrac{1+\sqrt{5}}{2} \simeq 1.61803$ se denomina *razón áurea*.

b) Determinar los errores absoluto y relativo,

$$|y_n - \phi| \quad \text{y} \quad \left|\frac{y_n - \phi}{\phi}\right|,$$

para valores $n = 1, 2, \ldots, 100$.

1.9. Se considera la sucesión de números reales $\{\alpha_n\}_{n=0}^{\infty}$ siendo

$$\alpha_n = \int_0^1 \frac{x^n}{x+10}\, dx,\ n \in \mathbb{N} \cup \{0\}.$$

a) Demostrar que los números α_n verifican la ley de recurrencia

$$\alpha_n + 10\alpha_{n-1} = \frac{1}{n},\ n \in \mathbb{N}.$$

b) Comprobar que los valores de $\{\alpha_0, \alpha_1, \ldots, \alpha_{25}\}$ obtenidos a partir de las secuencias

$$\begin{cases} \alpha_0 = \log\dfrac{11}{10} \\ \alpha_n = \dfrac{1}{n} - 10\alpha_{n-1}, \quad n = 1, 2, \ldots, 25 \end{cases} \tag{1.10}$$

y

$$\begin{cases} \alpha_{25} = 3.509 \times 10^{-3} \\ \alpha_{n-1} = \dfrac{1}{10}\left(\dfrac{1}{n} - \alpha_n\right), \quad n = 25, 24, \ldots, 1 \end{cases} \tag{1.11}$$

son los que se obtienen en la tabla 1.5. Dar una explicación a los mismos.

1.10. Aproximar la derivada de la función $\operatorname{sen} x$ en $x = 1$ mediante la fórmula

$$\frac{\operatorname{sen}(1+h) - \operatorname{sen} 1}{h}$$

con $h = 10^{-i}$ para $i = 5, 6, \ldots, 16$ comparando los resultados obtenidos con el valor exacto como se hace en la tabla 1.2. Comprobar la pérdida de precisión por cancelación.

TABLA 1.5:
Valores obtenidos mediante la secuencia (1.10)

n	α_n	n	α_n
0	9.531018×10^{-2}	13	5.784969×10^{-3}
1	4.689820×10^{-2}	14	1.357888×10^{-2}
2	3.101798×10^{-2}	15	-6.912212×10^{-2}
3	2.315353×10^{-2}	16	7.537212×10^{-1}
4	1.846471×10^{-2}	17	-7.478389
5	1.535290×10^{-2}	18	7.483944×10
6	1.313766×10^{-2}	19	-7.483418×10^{2}
7	1.148056×10^{-2}	20	7.483468×10^{3}
8	1.019440×10^{-2}	21	-7.483463×10^{4}
9	9.167129×10^{-3}	22	7.483464×10^{5}
10	8.328714×10^{-3}	23	-7.483464×10^{6}
11	7.621952×10^{-3}	24	7.483464×10^{7}
12	7.113811×10^{-3}	25	-7.483464×10^{8}

Valores obtenidos mediante la secuencia (1.11)

n	α_n	n	α_n
25	3.509000×10^{-3}	12	7.038976×10^{-3}
24	3.649100×10^{-3}	11	7.629436×10^{-3}
23	3.801757×10^{-3}	10	8.327966×10^{-3}
22	3.967650×10^{-3}	9	9.167203×10^{-3}
21	4.148690×10^{-3}	8	1.019439×10^{-2}
20	4.347036×10^{-3}	7	1.148056×10^{-2}
19	4.565296×10^{-3}	6	1.313766×10^{-2}
18	4.806628×10^{-3}	5	1.535290×10^{-2}
17	5.074893×10^{-3}	4	1.846471×10^{-2}
16	5.374864×10^{-3}	3	2.315353×10^{-2}
15	5.712514×10^{-3}	2	3.101798×10^{-2}
14	6.095415×10^{-3}	1	4.689820×10^{-2}
13	6.533316×10^{-3}	0	9.531018×10^{-2}

1.11. Las raíces exactas de la ecuación de segundo grado

$$x^2 - (64 + 10^{-15})\, x + (64 \times 10^{-15}) = 0$$

son $x_1 = 64$ y $x_2 = 10^{-15}$. Calcular sus raíces comprobando que el resultado obtenido para la menor de ellas no coincide con el exacto en ninguna cifra significativa.

1.12. Comprobar los resultados del ejemplo 1.10 relativo al cálculo de los 100 primeros términos de la sucesión definida por

$$\begin{cases} x_0 = 1,\ x_1 = \dfrac{1}{7} \\ x_{n+1} = \dfrac{22}{7} x_n - \dfrac{3}{7} x_{n-1},\ n \in \mathbb{N}. \end{cases}$$

2 Complementos de álgebra matricial

2.1. Introducción

El primero de los problemas típicos de Análisis Numérico que se aborda en este libro es el de la resolución de sistemas lineales de ecuaciones. Dedicaremos los próximos capítulos al estudio de métodos de resolución de sistemas lineales de la forma $Au = b$ donde A es una matriz cuadrada conocida y b es un vector dado.

En el presente capítulo se recogen una serie de definiciones y resultados relativos a las matrices. Aunque buena parte de ellos puedan ser conocidos, el hecho de recopilarlos aquí servirá para, por una parte, fijar la notación utilizada y, por otra, refrescar la memoria y tener accesibles los resultados que se necesitarán en los capítulos siguientes.

2.2. Diversos tipos de matrices y propiedades

En todo lo que sigue, $\mathbb{K}$ denota el cuerpo $\mathbb{R}$ de los números reales o el cuerpo $\mathbb{C}$ de los números complejos y $\mathbf{V} = \mathbb{K}^n$, $n \in \mathbb{N}$.

Definición 2.1. Los elementos $v \in \mathbf{V}$ se denominan *vectores* y se representan como

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix},$$

denominándose $\{v_1, v_2, \ldots, v_n\}$ las *componentes* de v. El elemento

$$v^{\mathrm{T}} = (v_1, v_2, \ldots, v_n)$$

es el *vector traspuesto* de $v \in \mathbf{V}$ y

$$v^* = (\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_n)$$

es el *vector adjunto* de v donde, para cada $j \in \{1, 2, \ldots, n\}$,

$$\overline{v}_j = a_j - ib_j \text{ si } v_j = a_j + ib_j,\ a_j, b_j \in \mathbb{R}$$

siendo $i = \sqrt{-1}$ la *unidad imaginaria*. □

Definición 2.2. Una *matriz* A es una colección de elementos $a_{ij} \in \mathbb{K}$ dispuestos en la forma

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

A es una *matriz real* si $\mathbb{K} = \mathbb{R}$ y *compleja* cuando $\mathbb{K} = \mathbb{C}$. La matriz anterior tiene m filas y n columnas y se denomina de *tipo* (m, n). En particular:

a) Un *vector columna* (o *vector*) es una matriz de tipo $(m, 1)$.

b) Un *vector fila* (o *vector traspuesto*) es una matriz de tipo $(1, n)$.

Notación 2.1.

a) Una matriz A de tipo (m, n) de elementos $a_{ij} \in \mathbb{K}$ se denota $A = (a_{ij})_{i,j=1}^{m,n}$ (i fila, j columna). También denotaremos por $(A)_{ij}$ o $A(i, j)$ el elemento de A que ocupa la fila i y la columna j.

b) En algunas ocasiones, escribiremos la matriz A de tipo (m, n) en la forma

$$A = (a_1, a_2, \ldots, a_n)$$

donde $a_i \in \mathbb{K}^m$ representa la columna i–ésima de A para $i = 1, 2, \ldots, n$. □

Definición 2.3. Si $A = (a_{ij})_{i,j=1}^{m,n}$ y $B = (b_{ij})_{i,j=1}^{m,n}$ son matrices del tipo (m, n) se define la *suma* de matrices como la matriz

$$A + B = (c_{ij})_{i,j=1}^{m,n} \text{ con } c_{ij} = a_{ij} + b_{ij}$$

y el *producto* de una matriz $A = (a_{ij})_{i,j=1}^{m,n}$ por un *escalar* $\lambda \in \mathbb{K}$ como la matriz

$$\lambda A = (d_{ij})_{i,j=1}^{m,n} \text{ siendo } d_{ij} = \lambda a_{ij}. \quad \square$$

Observación 2.1. Es un sencillo ejercicio comprobar que el conjunto formado por todas las matrices del tipo (m, n) con elementos en $\mathbb{K}$ tiene, con las operaciones de suma y producto por escalares, estructura de espacio vectorial sobre $\mathbb{K}$ al que denotaremos por $\mathcal{M}_{m\times n} = \mathcal{M}_{m\times n}(\mathbb{K})$. □

Definición 2.4. Sea $A = (a_{ij})_{i,j=1}^{m,n} \in \mathcal{M}_{m\times n}$.

a) La matriz $A^* = (\overline{a_{ji}})_{j,i=1}^{n,m} \in \mathcal{M}_{n\times m}$ se denomina *matriz adjunta* de A.

b) La matriz $A^{\mathrm{T}} = (a_{ji})_{j,i=1}^{n,m} \in \mathcal{M}_{n\times m}$ se denomina *matriz traspuesta* de A. □

Observación 2.2. Claramente,

$$(A^*)^* = A \ \text{ y } \ (A^{\mathrm{T}})^{\mathrm{T}} = A.$$

Además, cuando A es real, $A^* = A^{\mathrm{T}}$. □

Definición 2.5. Si $A = (a_{ik})_{i,k=1}^{m,l} \in \mathcal{M}_{m\times l}$ y $B = (b_{kj})_{k,j=1}^{l,n} \in \mathcal{M}_{l\times n}$, entonces la *matriz producto* de A y B viene dada por $AB = (c_{ij})_{i,j=1}^{m,n} \in \mathcal{M}_{m\times n}$, siendo

$$c_{ij} = \sum_{k=1}^{l} a_{ik}b_{kj}. \quad \square$$

Observación 2.3. Fácilmente se comprueba que

$$(AB)^* = B^*A^* \ \text{ y } \ (AB)^{\mathrm{T}} = B^{\mathrm{T}}A^{\mathrm{T}}. \quad \square$$

Definición 2.6. Sea $A = (a_{ij})_{i,j=1}^{m,n} \in \mathcal{M}_{m\times n}$. Se llama *submatriz* de A a toda matriz de la forma

$$\begin{pmatrix} a_{i_1j_1} & a_{i_1j_2} & \cdots & a_{i_1j_q} \\ a_{i_2j_1} & a_{i_2j_2} & \cdots & a_{i_2j_q} \\ \cdots & \cdots & \cdots & \cdots \\ a_{i_pj_1} & a_{i_pj_2} & \cdots & a_{i_pj_q} \end{pmatrix}$$

donde $1 \le i_1 < i_2 < \cdots < i_p \le m$ y $1 \le j_1 < j_2 < \cdots < j_q \le n$. □

Definición 2.7. Una matriz $A \in \mathcal{M}_{n\times n}$ se denomina *matriz cuadrada* o *matriz de orden n*. Diremos que una matriz es *rectangular* si no es cuadrada. □

Notación 2.2. Denotaremos $\mathcal{M}_n = \mathcal{M}_n(\mathbb{K}) = \mathcal{M}_{n\times n} = \mathcal{M}_{n\times n}(\mathbb{K})$ al espacio vectorial de matrices cuadradas de orden n con elementos en $\mathbb{K}$. □

En todo lo que sigue, salvo que se mencione explícitamente lo contrario, trabajaremos con matrices $A \in \mathcal{M}_n$.

Observación 2.4. La *descomposición por bloques* o *en cajas* de una matriz cuadrada $A \in \mathcal{M}_n$ es de la forma

$$A = \left(\begin{array}{c|c|c|c} A_{11} & A_{12} & \cdots & A_{1p} \\ \hline A_{21} & A_{22} & \cdots & A_{2p} \\ \hline \cdots & \cdots & \ddots & \cdots \\ \hline A_{p1} & A_{p2} & \cdots & A_{pp} \end{array}\right)$$

donde cada submatriz $A_{ij} \in \mathcal{M}_{n_i \times n_j}$ con

$$A_{ij} = (a_{lk})_{l=n_1+n_2+\cdots+n_{i-1}+1,k=n_1+n_2+\cdots+n_{j-1}+1}^{n_1+n_2+\cdots+n_i,n_1+n_2+\cdots+n_j}$$

y

$$\sum_{i=1}^{p} n_i = n.$$

Nótese que los bloques diagonales A_{ii} para $i = 1, 2, \ldots, p$, son matrices cuadradas. Así, por ejemplo, si $A \in \mathcal{M}_4$, podemos hacer, entre otras, las siguientes descomposiciones en bloques de la matriz A :

$$\left(\begin{array}{c|ccc} a_{11} & a_{12} & a_{13} & a_{14} \\ \hline a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{array}\right), \left(\begin{array}{cc|cc} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{array}\right),$$

$$\left(\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ \hline a_{41} & a_{42} & a_{43} & a_{44} \end{array}\right), \left(\begin{array}{c|c|c|c} a_{11} & a_{12} & a_{13} & a_{14} \\ \hline a_{21} & a_{22} & a_{23} & a_{24} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} \\ \hline a_{41} & a_{42} & a_{43} & a_{44} \end{array}\right).$$

En el primer caso $p = 2$, $n_1 = 1$ y $n_2 = 3$; en el segundo, $p = 2$, $n_1 = n_2 = 2$; en el tercero $p = 2$, $n_1 = 3$ y $n_2 = 1$; en el último, $p = 3$, $n_1 = n_3 = 1$ y $n_2 = 2$. □

Definición 2.8. Sea $A = (a_{ij})_{i,j=1}^{n} \in \mathcal{M}_n$. Los elementos a_{ii} se denominan *elementos diagonales*. □

Definición 2.9. Con la notación habitual de la *delta de Kronecker*

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

se denomina *matriz identidad* a la matriz cuadrada $I = (\delta_{ij})_{i,j=1}^{n} \in \mathcal{M}_n$. □

Definición 2.10. Una matriz $A \in \mathcal{M}_n$ es *inversible* (o *regular* o *no singular*) si existe una matriz $B \in \mathcal{M}_n$ (única) de forma que

$$AB = BA = I.$$

En tal caso, la matriz B se denota A^{-1} y se le denomina *matriz inversa* de A. En el caso de que A no tenga la propiedad anterior se dice que A es una matriz *no inversible* (o *singular o no regular*). □

Observación 2.5. Si A y B son dos matrices inversibles se verifica:

$$(AB)^{-1} = B^{-1}A^{-1},\ (A^*)^{-1} = (A^{-1})^*\ \text{ y }\ (A^{\mathrm{T}})^{-1} = (A^{-1})^{\mathrm{T}}. \quad \square \tag{2.1}$$

Veamos seguidamente los diversos tipos de matrices con los que trabajaremos usualmente.

Definición 2.11. Una matriz $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ es:

a) *Hermítica* si $A = A^*$, es decir, $a_{ij} = \overline{a_{ji}}$ para $i, j = 1, 2, \ldots, n$.

b) *Simétrica* si A es real y $A = A^{\mathrm{T}}$, es decir, $a_{ij} = a_{ji}$ para $i, j = 1, 2, \ldots, n$.

c) *Unitaria* si $A^* = A^{-1}$, es decir, $AA^* = A^*A = I$.

d) *Ortogonal* si A es real y $A^{\mathrm{T}} = A^{-1}$, es decir, $AA^{\mathrm{T}} = A^{\mathrm{T}}A = I$.

e) *Normal* si $AA^* = A^*A$.

f) *Triangular superior* (resp. *inferior*) si $a_{ij} = 0$ para $i > j$ (resp. $i < j$).

g) *Diagonal* si $a_{ij} = 0$ cuando $i \neq j$. En este caso se denota

$$A = \operatorname{diag}(a_{ii}) = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn}).$$

h) *Banda* si existe $p \in \{1, 2, \ldots, n\}$ tal que $a_{ij} = 0$ para $|i - j| \geq p$. Al número p se le suele denominar *semi–ancho* de banda y a $2p-1$ *ancho* de banda. En el caso particular en que $p = 2$ la matriz A es *tridiagonal* y tiene la forma

$$A = \begin{pmatrix} a_{11} & a_{12} & & & & & \\ a_{21} & a_{22} & a_{23} & & & & \\ & a_{32} & a_{33} & a_{34} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ & & & & a_{n,n-1} & a_{nn} \end{pmatrix}.$$

i) De *diagonal estrictamente dominante* si

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

para $i = 1, 2, \ldots, n$. $\square$

Observación 2.6.

1. Toda matriz hermítica o unitaria es normal.
2. Si A es hermítica e inversible entonces A^{-1} es también hermítica (véase (2.1)).
3. Si A es normal e inversible entonces A^{-1} es también normal (véase el problema 2.24). □

Definición 2.12. Sea $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$. Se denomina *traza* de la matriz A al número

$$\operatorname{tr}(A) = \sum_{i=1}^n a_{ii}. \quad \square$$

La traza de una matriz es invariante por transformaciones unitarias como se muestra en el siguiente resultado:

Proposición 2.1. *Si $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ y U es una matriz unitaria, entonces*

$$\operatorname{tr}(A) = \operatorname{tr}(UAU^*).$$

Demostración. Escribiendo $U = (u_1, u_2, \ldots, u_n)$, como $U^*U = I$, se verifica que

$$u_j^* u_i = \delta_{ji}$$

para $i, j = 1, 2, \ldots, n$. De esta forma,

$$\begin{aligned}\operatorname{tr}(UAU^*) &= \sum_{k=1}^n (UAU^*)_{kk} = \sum_{k=1}^n \sum_{i=1}^n u_{ki}(AU^*)_{ik} = \sum_{k=1}^n \left(\sum_{i=1}^n \sum_{j=1}^n u_{ki} a_{ij} \overline{u_{kj}} \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n a_{ij} \left(\sum_{k=1}^n \overline{u_{kj}} u_{ki} \right) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} u_j^* u_i = \sum_{i=1}^n a_{ii} = \operatorname{tr}(A). \quad \square\end{aligned}$$

Definición 2.13. Si $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ y G_n es el grupo de las permutaciones de $\{1, 2, \ldots, n\}$, entonces

$$\det(A) = \sum_{\sigma \in G_n} \operatorname{sig}(\sigma)\, a_{\sigma(1)1} a_{\sigma(2)2} \ldots a_{\sigma(n)n}$$

es el *determinante* de la matriz A.[1] □

[1] $\operatorname{sig}(\sigma)$ denota la *signatura* de la permutación σ. El número de sumandos de la suma anterior es $n!$ (número de permutaciones de los índices $\{1, 2, \ldots, n\}$).

En la siguiente proposición recogemos algunos resultados conocidos relativos a la traza y el determinante de una matriz.

Proposición 2.2. *Sean* $A,\ B \in \mathcal{M}_n$.

a) $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

b) $\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B)$.

c) $\det(I) = 1$.

d) $\det(AB) = \det(BA) = \det(A)\det(B)$.

e) $\det(\lambda A) = \lambda^n \det(A),\ \lambda \in \mathbb{K}$.

f) $\det(A^*) = \overline{\det(A)}$.

g) *Si* $A = (a_{ij})_{i,j=1}^n$ *es una matriz triangular entonces*

$$\det(A) = \prod_{i=1}^{n} a_{ii}. \quad \square$$

Proposición 2.3. *Una matriz* $A \in \mathcal{M}_n$ *es inversible si y sólo si* $\det(A) \neq 0$.

Demostración.

$\boxed{\Rightarrow}$ Supongamos que $\det(A) = 0$. En ese caso, por el teorema de Rouché–Fröbenius, existe $v \in \mathbf{V}\backslash\{0\}$ tal que $Av = 0$. Por tanto, se obtiene la contradicción:

$$Av = 0 \ \Rightarrow\ A^{-1}Av = 0 \ \Rightarrow\ v = 0.$$

$\boxed{\Leftarrow}$ Como $\det(A) \neq 0$, aplicando nuevamente el teorema de Rouché–Fröbenius, sabemos que para cada $i \in \{1, 2, \ldots, n\}$ el sistema lineal

$$Av_i = \mathbf{e}_i,$$

donde $\mathbf{e}_i = (0, \ldots, 0, \overset{i)}{1}, 0, \ldots, 0)^{\mathrm{T}}$, admite una única solución. De esta forma la matriz $B = (v_1, v_2, \ldots, v_n) \in \mathcal{M}_n$ que tiene por columnas los vectores v_i verifica $AB = I$ (véase el problema 2.3). Por tanto, $\det(B) \neq 0$ y, aplicando el mismo razonamiento, existe $C \in \mathcal{M}_n$ tal que $BC = I$. Así pues,

$$AB = I = BC \ \Rightarrow\ A(AB) = (AB)C \ \Rightarrow\ A = C.$$

Luego $AB = BA = I$ y, por tanto, la matriz A es inversible . $\square$

Observación 2.7. De los apartados *c*) y *d*) de la proposición 2.2 se deduce que

$$\det(A^{-1}) = \frac{1}{\det(A)} \text{ si } A \text{ es inversible.} \quad \square \tag{2.2}$$

Definición 2.14. Sea $A \in \mathcal{M}_n$.

a) El polinomio de grado n

$$P_A(\lambda) = \det(A - \lambda I)$$

se denomina *polinomio característico* de A.

b) Las raíces del polinomio característico de A se denominan *autovalores* (o *valores propios*) de la matriz A, es decir, $\lambda_i(A)$ es un *autovalor* de A si $P_A(\lambda_i(A)) = 0$. Por el teorema Fundamental del Álgebra, la matriz A tiene n autovalores $\{\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A)\}$, ya sean reales o complejos, contados con su multiplicidad.

c) Llamaremos *espectro* de A al conjunto de todos los autovalores de la matriz A y lo representaremos por

$$\mathrm{sp}(A) = \{\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A)\}.$$

d) El número no negativo

$$\varrho(A) = \max_{1 \le i \le n} \{|\lambda_i(A)| : \lambda_i(A) \in \mathrm{sp}(A)\}$$

es el *radio espectral* de A. Como se observa, el radio espectral de una matriz es el radio del círculo más pequeño centrado en el origen que contiene a todos los autovalores de la matriz. $\square$

Observación 2.8.

a) A la vista de la proposición 2.3, para toda matriz $A \in \mathcal{M}_n$ se verifica:

$$A \text{ es inversible} \Leftrightarrow 0 \notin \mathrm{sp}(A).$$

b) Si $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ es una matriz triangular entonces

$$\mathrm{sp}(A) = \{a_{11}, a_{22}, \ldots, a_{nn}\}. \quad \square$$

Veamos una caracterización de los autovalores de una matriz.

Proposición 2.4. *Para toda matriz $A \in \mathcal{M}_n$ se verifica:*

$$\lambda \in \mathrm{sp}(A) \Leftrightarrow \textit{existe } v \in \mathbf{V}\backslash\{0\} \textit{ tal que } Av = \lambda v.$$

DEMOSTRACIÓN. Basta tener en cuenta las siguientes equivalencias:

$$\begin{aligned} Av = \lambda v,\ v \neq 0 &\Leftrightarrow (A - \lambda I)v = 0,\ v \neq 0 \\ &\Leftrightarrow \text{el sistema lineal } (A - \lambda I)x = 0 \text{ admite solución no trivial} \\ &\Leftrightarrow \det(A - \lambda I) = 0 \\ &\Leftrightarrow \lambda \in \mathrm{sp}(A). \quad \square \end{aligned}$$

Definición 2.15. Sea $A \in \mathcal{M}_n$. Un vector $v \in \mathbf{V}\backslash\{0\}$ es *autovector* (o *vector propio*) asociado al autovalor $\lambda = \lambda(A)$ de A si $Av = \lambda v$. El conjunto

$$\{v \in \mathbf{V} :\ Av = \lambda v\}$$

es el *subespacio propio* correspondiente al autovalor λ. $\square$

A continuación vamos a mostrar que todos los autovalores de una matriz hermítica son reales y que los autovalores de una matriz unitaria tienen módulo uno (lo que le confiere el calificativo de unitaria a la matriz).

Proposición 2.5. *Sea $A \in \mathcal{M}_n$.*

a) Si A es hermítica entonces $\mathrm{sp}(A) \subset \mathbb{R}$.

b) Si A es unitaria entonces $|\lambda| = 1$ para todo $\lambda \in \mathrm{sp}(A)$.

DEMOSTRACIÓN. Sea $\lambda \in \mathrm{sp}(A)$. Por la proposición 2.4 sabemos que existe un vector $v \in \mathbf{V}\backslash\{0\}$ tal que $Av = \lambda v$. Nótese que

$$(\lambda v)^* = (\overline{\lambda v_1}, \overline{\lambda v_2}, \ldots, \overline{\lambda v_n}) = (\overline{\lambda}\overline{v}_1, \overline{\lambda}\overline{v}_2, \ldots, \overline{\lambda}\overline{v}_n) = \overline{\lambda}(\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_n) = \overline{\lambda}v^*$$

y que, al ser $v \neq 0$, entonces

$$v^*v = \sum_{i=1}^{n} \overline{v}_i v_i = \sum_{i=1}^{n} |v_i|^2 > 0. \tag{2.3}$$

De esta forma:

a) Si A es hermítica entonces:

$$\lambda v^*v = v^*\lambda v = v^*Av = v^*A^*v = (Av)^*v = (\lambda v)^*v = \overline{\lambda}v^*v. \tag{2.4}$$

Así pues, de (2.3) y (2.4) se deduce $\lambda = \overline{\lambda}$ y, por tanto, $\lambda \in \mathbb{R}$.

b) Si A es unitaria, se verifica:

$$|\lambda|^2 v^* v = \overline{\lambda} v^* \lambda v = (\lambda v)^*(\lambda v) = (Av)^*(Av) = v^* A^* A v = v^* v.$$

Nuevamente, por (2.3), se sigue $|\lambda|^2 = 1$, de donde $|\lambda| = 1$. □

Como es sabido, toda aplicación lineal $\mathcal{A} : \mathbf{V} \to \mathbf{V}$ viene representada respecto a una base $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ por una matriz $A = (a_{ij})_{i,j=1}^n$. Si tomamos otra base $\widetilde{\mathcal{B}} = \{\widetilde{v}_1, \widetilde{v}_2, \ldots, \widetilde{v}_n\}$ de $\mathbf{V}$, la aplicación vendrá representada por otra matriz $\widetilde{A} = P^{-1}AP$, donde $P \in \mathcal{M}_n$ es la matriz cuya columna i–ésima está formada por las coordenadas del vector $\widetilde{v}_i$ respecto a la base $\mathcal{B}$. La matriz P se denomina *matriz de paso* de la base $\mathcal{B}$ a la base $\widetilde{\mathcal{B}}$. Estas matrices A y $\widetilde{A}$, que representan la misma aplicación lineal respecto a distintas bases, se denominan matrices semejantes. Así,

Definición 2.16. Dos matrices $A, B \in \mathcal{M}_n$ son *semejantes* si existe una matriz de paso P no singular tal que $B = P^{-1}AP$. □

Observación 2.9. Dos matrices semejantes tienen el mismo espectro puesto que sus polinomios característicos coinciden. En efecto,

$$\begin{aligned} P_B(\lambda) &= \det(B - \lambda I) = \det(P^{-1}AP - \lambda I) = \det(P^{-1}(A - \lambda I)P) \\ &= \det(P^{-1}) \det(A - \lambda I) \det(P) = \det(A - \lambda I) = P_A(\lambda) \end{aligned}$$

(véase (2.2)). □

Se plantea el problema de, a partir de una matriz dada A, encontrar una matriz semejante a A que sea lo más sencilla posible. El caso "más favorable" es cuando existe una matriz P de paso de forma que $P^{-1}AP$ es diagonal. Esto da pie a la siguiente definición:

Definición 2.17. Una matriz $A \in \mathcal{M}_n$ es *diagonalizable* si existe una matriz $P \in \mathcal{M}_n$ inversible tal que $P^{-1}AP$ es diagonal. □

Se tiene la siguiente caracterización:

Proposición 2.6. *Una matriz A es diagonalizable si y sólo si existe una base de $\mathbf{V}$ formada por autovectores de la matriz A.*

Demostración. Basta observar que

$$P^{-1}AP = D = \operatorname{diag}(\lambda_i) \Leftrightarrow AP = PD \Leftrightarrow Ap_i = \lambda_i p_i,\ i = 1, 2, \ldots, n$$

donde $\{p_1, p_2, \ldots, p_n\}$ son las columnas de la matriz P.

Así, si A es diagonalizable, las columnas de P (que son linealmente independientes por ser P inversible) constituyen una base del espacio y cada una de ellas es un autovector de la matriz A. Recíprocamente, si $\{p_1, p_2, \ldots, p_n\}$ es una base de autovectores de A, la matriz $P = (p_1, p_2, \ldots, p_n)$ diagonaliza la matriz A. □

El siguiente resultado muestra que toda matriz es semejante a una matriz triangular mediante una matriz de paso unitaria:

Teorema 2.1. *Sea $A \in \mathcal{M}_n$.*

a) Existe U unitaria tal que U^*AU es triangular.

b) A es normal si y sólo si existe U unitaria tal que U^*AU es diagonal.

Demostración.

a) En primer lugar vamos a demostrar un resultado más débil: que toda matriz es semejante a una triangular superior (sin imponer ninguna condición a la matriz de paso). Para ello, procedemos por inducción en el tamaño n de la matriz:

i) Para $n = 1$ el resultado se tiene trivialmente.

ii) Supongamos cierto el resultado para matrices de orden $n-1$. Consideremos un autovalor λ de A y un autovector u asociado a él. Obviamente, como $u \neq 0$, existen vectores $\{v_2, v_3, \ldots, v_n\}$ que hacen que la matriz

$$P = (u, v_2, \ldots, v_n)$$

sea inversible. De esta forma

$$AP = (Au, Av_2, \ldots, Av_n) = (\lambda u, Av_2, \ldots, Av_n)$$

y

$$P^{-1}AP = (\lambda P^{-1}u, P^{-1}Av_2, \ldots, P^{-1}Av_n).$$

Ahora bien, como

$$I = P^{-1}P = (P^{-1}u, P^{-1}v_2, \ldots, P^{-1}v_n),$$

entonces

$$P^{-1}u = \mathbf{e}_1$$

siendo $\mathbf{e}_1$ el primer vector de la base canónica. Así, podemos escribir

$$P^{-1}AP = (\lambda \mathbf{e}_1, P^{-1}Av_2, \ldots, P^{-1}Av_n) = \left(\begin{array}{c|c} \lambda & w^{\mathrm{T}} \\ \hline \mathbf{0} & A_{n-1} \end{array}\right) \qquad (2.5)$$

con $A_{n-1} \in \mathcal{M}_{n-1}$ y $w \in \mathbb{C}^{n-1}$. Por la hipótesis de inducción, existe $Q_{n-1} \in \mathcal{M}_{n-1}$ inversible tal que

$$(Q_{n-1})^{-1} A_{n-1} Q_{n-1} = T_{n-1} \tag{2.6}$$

siendo $T_{n-1} \in \mathcal{M}_{n-1}$ triangular superior. A partir de la matriz

$$Q = P \left(\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & Q_{n-1} \end{array}\right)$$

y de las relaciones (2.5) y (2.6) se tiene que la matriz

$$\begin{aligned} Q^{-1}AQ &= \left(\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & (Q_{n-1})^{-1} \end{array}\right) P^{-1}AP \left(\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & Q_{n-1} \end{array}\right) \\ &= \left(\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & (Q_{n-1})^{-1} \end{array}\right) \left(\begin{array}{c|c} \lambda & w^{\mathrm{T}} \\ \hline \mathbf{0} & A_{n-1} \end{array}\right) \left(\begin{array}{c|c} 1 & \mathbf{0} \\ \hline \mathbf{0} & Q_{n-1} \end{array}\right) \\ &= \left(\begin{array}{c|c} \lambda & w^{\mathrm{T}} Q_{n-1} \\ \hline \mathbf{0} & (Q_{n-1})^{-1} A_{n-1} Q_{n-1} \end{array}\right) = \left(\begin{array}{c|c} \lambda & w^{\mathrm{T}} Q_{n-1} \\ \hline \mathbf{0} & T_{n-1} \end{array}\right) \end{aligned}$$

es triangular superior.

Demostremos ahora el resultado enunciado. Dada la matriz A existe, según hemos visto, una matriz P inversible tal que $T = P^{-1}AP$ es triangular superior. Aplicando el resultado del problema 2.12 a la matriz P, se tiene que existen una matriz U unitaria y R triangular superior e inversible de forma que $P = UR$. Por tanto, $U = PR^{-1}$ y

$$U^*AU = U^{-1}AU = RP^{-1}APR^{-1} = RTR^{-1}$$

que es triangular superior por serlo T, R y su inversa.

b) Supongamos que A es normal. A partir del apartado *a*) se tiene que existe U unitaria tal que $T = U^*AU$ es triangular superior. Veamos que, además, es normal por serlo A. En efecto,

$$\begin{aligned} T^*T &= (U^*A^*U)(U^*AU) = U^*(A^*A)U \\ &= U^*(AA^*)U = (U^*AU)(U^*A^*U) = TT^*. \end{aligned}$$

Por tanto, al ser T triangular y normal, utilizando el problema 2.10, se llega a que T es diagonal. Recíprocamente, si $U^*AU = D = \operatorname{diag}(d_{ii})$, entonces

$$DD^* = D^*D = \operatorname{diag}\left(|d_{ii}|^2\right).$$

De esta forma, por ser U unitaria,

$$\begin{aligned} A^*A &= UD^*(U^*U)DU^* = UD^*DU^* \\ &= UDD^*U^* = (UDU^*)(UD^*U^*) = AA^*. \quad \square \end{aligned}$$

Observación 2.10.

1. Las matrices de paso que verifican las condiciones del enunciado no son únicas (basta considerar $A = I$).

2. Como consecuencia de la observación 2.6 y de la segunda afirmación del teorema 2.1 se deduce que toda matriz hermítica o unitaria es diagonalizable por una matriz de paso unitaria. Además, cuando la matriz A es simétrica puede demostrarse que la matriz de paso es también real, es decir, existe O ortogonal tal que $O^{\mathrm{T}}AO$ es diagonal.

3. La matriz triangular U^*AU se denomina *forma de Schur* de A. Tiene gran importancia en la práctica por la dificultad numérica de calcular otras formas canónicas como la *matriz de Jordan*.

4. Para toda matriz hermítica existe una base ortonormal formada por autovectores de A (véase el problema 2.23).

5. Nótese que los elementos diagonales de la matriz triangular U^*AU son los autovalores de la matriz A. De esta forma, la proposición 2.1 asegura que

$$\operatorname{tr}(A) = \operatorname{tr}(U^*AU) = \sum_{i=1}^{n} \lambda_i(A).$$

   Análogamente, se verifica que

$$\det(A) = \det(U^*AU) = \prod_{i=1}^{n} \lambda_i(A).$$

6. Del apartado 5 y de la observación 2.9 se deduce que la proposición 2.1 puede generalizarse a matrices de paso arbitrarias: dos matrices semejantes tienen la misma traza, esto es

$$\operatorname{tr}(A) = \operatorname{tr}(P^{-1}AP)$$

   para cualquier matriz $P \in \mathcal{M}_n$ inversible. □

Definición 2.18. Una matriz hermítica $A \in \mathcal{M}_n$ es *definida positiva* (resp. *semidefinida positiva*) si

$$v^*Av > 0,\ v \in \mathbf{V}\backslash\{0\} \quad (\text{resp. } v^*Av \geq 0,\ v \in \mathbf{V}). \quad \square$$

Observación 2.11. En el caso real, para que una matriz A simétrica sea definida positiva basta con que verifique

$$v^{\mathrm{T}}Av > 0,\ v \in \mathbb{R}^n\backslash\{0\} \tag{2.7}$$

puesto que la condición anterior implica que

$$v^*Av > 0,\ v \in \mathbb{C}^n\backslash\{0\}.$$

En efecto, para un vector arbitrario $v = a + bi \in \mathbb{C}^n\backslash\{0\}$, a partir de la relación (2.7) se tiene que

$$\begin{aligned} v^*Av &= (a+bi)^*A(a+bi) = \left(a^{\mathrm{T}} - b^{\mathrm{T}}i\right)A(a+bi) \\ &= a^{\mathrm{T}}Aa + a^{\mathrm{T}}Abi - b^{\mathrm{T}}Aai + b^{\mathrm{T}}Ab \\ &= a^{\mathrm{T}}Aa + b^{\mathrm{T}}Ab > 0, \end{aligned}$$

puesto que los vectores $a, b \in \mathbb{R}^n$ no son nulos simultánemente y la simetría de la matriz A hace que se tenga

$$a^{\mathrm{T}}Ab = (a^{\mathrm{T}}Ab)^{\mathrm{T}} = b^{\mathrm{T}}A^{\mathrm{T}}a = b^{\mathrm{T}}Aa. \quad \square$$

Veamos una caracterización de las matrices definidas y semidefinidas positivas a partir del signo de los autovalores de la matriz:

Proposición 2.7. *Si $A \in \mathcal{M}_n$ es una matriz hermítica, se verifica:*

a) A es definida positiva $\Leftrightarrow$ $\mathrm{sp}(A) \subset \mathbb{R}_+$.

b) A es semidefinida positiva $\Leftrightarrow$ $\mathrm{sp}(A) \subset \mathbb{R}_+ \cup \{0\}$.

Demostración. Mostramos únicamente la primera equivalencia, pues la segunda se prueba de forma análoga.

$\boxed{\Rightarrow}$ Si $\lambda \in \mathrm{sp}(A)$, por la proposición 2.4 existe $v \in \mathbf{V}\backslash\{0\}$ tal que $Av = \lambda v$. Por tanto,

$$\lambda v^*v = v^*Av > 0 \ \text{ y } \ v^*v = \sum_{i=1}^{n} |v_i|^2 > 0,$$

de donde

$$\lambda = \frac{v^*Av}{v^*v} > 0.$$

$\boxed{\Leftarrow}$ Por el teorema 2.1, puesto que toda matriz hermítica es normal, sabemos que existe una matriz U unitaria tal que $U^*AU = D$ siendo D una matriz diagonal. Más concretamente,

$$U^*AU = D = \mathrm{diag}\,(\lambda_i(A)).$$

Por tanto,

$$v^*Av = v^*(UDU^*)v = (U^*v)^*D(U^*v),\ v \in \mathbf{V}.$$

Dado $v \in \mathbf{V}\backslash\{0\}$, si denotamos $w = U^*v$, como U es una matriz inversible entonces $w \neq 0$ y, al ser $\lambda_i(A) > 0$ para todo $i = 1, 2, \ldots, n$, se verifica

$$v^*Av = (U^*v)^*D(U^*v) = \sum_{i=1}^{n} \lambda_i \, |w_i|^2 > 0, \; v \in \mathbf{V}\backslash\{0\}. \quad \square$$

Observación 2.12. Si $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ es una matriz hermítica y denotamos por

$$\delta_k = \det \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}$$

al *menor principal* de A de orden $k \in \{1, 2, \ldots, n\}$, se verifica:

$$A \text{ es definida positiva } \Leftrightarrow \; \delta_k > 0, \; k = 1, 2, \ldots, n$$

(véanse los problemas 2.16 y 4.9). $\square$

Veamos un resultado que nos va a permitir construir matrices semidefinidas positivas a partir de matrices cuadradas arbitrarias y matrices definidas positivas a partir de matrices inversibles.

Proposición 2.8. *Dada una matriz $A \in \mathcal{M}_n$ se verifica que A^*A es una matriz hermítica y semidefinida positiva. Además, cuando A es inversible la matriz A^*A es, de hecho, definida positiva.*

Demostración. Claramente $(A^*A)^* = A^*A$, luego la matriz A^*A es hermítica. Por otra parte, para todo vector $v \in \mathbf{V}\backslash\{0\}$ se verifica

$$v^*A^*Av = (Av)^*(Av) = \sum_{i=1}^{n} |w_i|^2 \geq 0$$

siendo $w = Av$, de donde se deduce que la matriz A^*A es semidefinida positiva. De hecho, cuando la matriz A es inversible se verifica que $w \neq 0$, por lo que

$$v^*A^*Av = \sum_{i=1}^{n} |w_i|^2 > 0,$$

obteniéndose así que A^*A es definida positiva. $\square$

Finalizamos esta sección con un resultado, que se utilizará con bastante frecuencia, relativo a la inversibilidad de las matrices de diagonal estrictamente dominante.

Teorema 2.2. *Toda matriz de diagonal estrictamente dominante es inversible.*

DEMOSTRACIÓN. Sea $A = (a_{ij})_{i,j=1}^{n} \in \mathcal{M}_n$ una matriz de diagonal estrictamente dominante, es decir,

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \tag{2.8}$$

para $i = 1, 2, \ldots, n$. Para demostrar que A es inversible veamos que la única solución del sistema lineal homogéneo $Az = 0$ es $z = 0$. Argumentamos por reducción al absurdo: supongamos que $z \neq 0$ y derivemos una contradicción. Para ello, sea $i_0 \in \{1, 2, \ldots, n\}$ tal que

$$|z_{i_0}| = \max_{1\leq i\leq n} |z_i|,$$

por lo que

$$|z_i| \leq |z_{i_0}|$$

para $i = 1, 2, \ldots, n$. El hecho de que $z \neq 0$ hace que $z_{i_0} \neq 0$ y, como $Az = 0$, entonces

$$\sum_{j=1}^{n} a_{ij} z_j = 0$$

para $i = 1, 2, \ldots, n$, de donde se deduce que

$$a_{ii} z_i = -\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} z_j$$

para todo índice $i = 1, 2, \ldots, n$. En particular, aplicando la propiedad (2.8) para $i = i_0$, se obtiene la contradicción

$$|a_{i_0 i_0}|\,|z_{i_0}| = \left| \sum_{\substack{j=1 \\ j\neq i_0}}^{n} a_{i_0 j} z_j \right| \leq \sum_{\substack{j=1 \\ j\neq i_0}}^{n} |a_{i_0 j}|\,|z_j| \leq \sum_{\substack{j=1 \\ j\neq i_0}}^{n} |a_{i_0 j}|\,|z_{i_0}| < |a_{i_0 i_0}|\,|z_{i_0}|\,. \quad \square$$

2.3. Normas matriciales

En esta sección se introduce la noción de *norma* de una matriz, herramienta básica en el estudio de la convergencia de los métodos iterativos que serán tratados en el capítulo 5. En primer lugar, recordamos el concepto de norma vectorial y algunas de sus propiedades.

Definición 2.19. Una aplicación $||\cdot|| : \mathbf{V} \to \mathbb{R}_+ \cup \{0\}$ es una *norma* en $\mathbf{V}$ si verifica:

a) $||v|| = 0 \Leftrightarrow v = 0$.

b) $||\lambda v|| = |\lambda| \, ||v||$, $v \in \mathbf{V}$, $\lambda \in \mathbb{K}$.

c) $||u + v|| \leq ||u|| + ||v||$, $u, v \in \mathbf{V}$ (*desigualdad triangular*).

Una norma en $\mathbf{V}$ se denomina también *norma vectorial*. □

Ejemplo 2.1. Son normas en $\mathbf{V}$ las aplicaciones:

$$||v||_1 = \sum_{i=1}^{n} |v_i|, \; ||v||_2 = \sqrt{\sum_{i=1}^{n} |v_i|^2} \;\text{ y }\; ||v||_\infty = \max_{1 \leq i \leq n} |v_i|$$

donde $v = (v_1, v_2, \ldots, v_n)^{\mathrm{T}}$. En general, también es una norma en $\mathbf{V}$

$$||v||_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}} \quad \text{para } \; 1 \leq p < +\infty.$$

Así, por ejemplo, para el vector $v = (1, -1, 1, -1)^{\mathrm{T}} \in \mathbb{R}^4$ se tiene que

$$||v||_p = \sqrt[p]{4}, \; 1 \leq p < +\infty \;\text{ y }\; ||v||_\infty = 1.$$

Además, como se verá en el problema 2.11,

$$\lim_{p \to +\infty} ||v||_p = ||v||_\infty, \; v \in \mathbf{V}. \quad \square \tag{2.9}$$

Observación 2.13. La desigualdad triangular de la norma determina, para todo par de vectores $u, v \in \mathbf{V}$ las desigualdades

$$\begin{cases} ||u|| = ||v + (u - v)|| \leq ||v|| + ||u - v|| \\ ||v|| = ||u + (v - u)|| \leq ||u|| + ||v - u|| \, . \end{cases}$$

Como $||u - v|| = ||v - u||$, se deduce la desigualdad

$$\big| \, ||u|| - ||v|| \, \big| \leq ||u - v|| \, , \; u, v \in \mathbf{V}. \quad \square \tag{2.10}$$

Observación 2.14. Si $v = (v_1, v_2, \ldots, v_n)^{\mathrm{T}} \in \mathbf{V}$ entonces

$$v^* v = (\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_n) \begin{pmatrix} v_1 \\ v_2 \\ \cdots \\ v_n \end{pmatrix} = \sum_{i=1}^{n} \overline{v}_i v_i = \sum_{i=1}^{n} |v_i|^2 = ||v||_2^2 \, .$$

Es decir, podemos expresar

$$||v||_2 = +\sqrt{v^* v},\ v \in \mathbf{V}. \quad \square \tag{2.11}$$

Observación 2.15. La bola unidad en $\mathbb{R}^2$ para las normas p e infinito viene dada por

$$\mathbf{B}_1^{(p)}(0,0) = \left\{(x,y) \in \mathbb{R}^2 :\ |x|^p + |y|^p < 1\right\},\ 1 \leq p < +\infty$$

y

$$\mathbf{B}_1^{(\infty)}(0,0) = \left\{(x,y) \in \mathbb{R}^2 :\ \max\{|x|,|y|\} < 1\right\}$$

(véase la figura 2.1). La relación (2.9) hace que se tenga

$$\lim_{p \to +\infty} \mathbf{B}_1^{(p)}(0,0) = \mathbf{B}_1^{(\infty)}(0,0). \quad \square$$

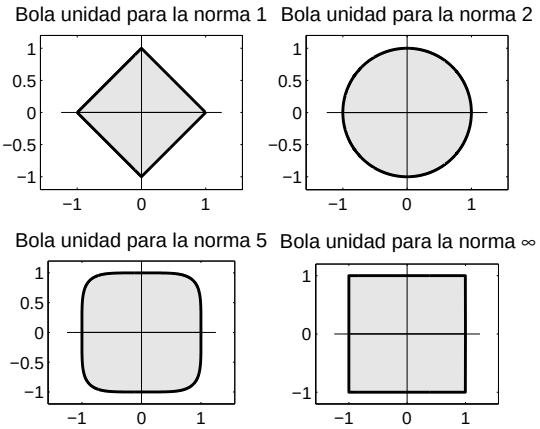


Figura 2.1: Forma de la bola unidad en $\mathbb{R}^2$ para diversas normas.

Definición 2.20. Dos normas $||\cdot||$ y $||\cdot||'$ en $\mathbf{V}$ son *equivalentes* si existen constantes c, $C > 0$ tales que

$$c\,||v||' \leq ||v|| \leq C\,||v||',\ v \in \mathbf{V}. \quad \square$$

Proposición 2.9. *En un espacio vectorial* $\mathbf{V}$ *de dimensión finita todas las normas vectoriales son equivalentes.* $\square$

La desigualdad (2.10) permite demostrar la continuidad de la norma.

Proposición 2.10. *Si* $||\cdot||$ *es una norma en* $\mathbf{V}$ *entonces la aplicación* $u \mapsto ||u||$ *es continua.*

DEMOSTRACIÓN. Dados $u \in \mathbf{V}$ y $\varepsilon > 0$ arbitrarios basta tomar $\delta = \varepsilon$ y $v \in \mathbf{V}$ con $||u - v|| < \delta$ para que, aplicando (2.10), se verifique que

$$\big|\, ||u|| - ||v|| \,\big| < \varepsilon. \quad \square$$

Introducimos, a continuación, la noción de norma de una matriz:

Definición 2.21. Una *norma matricial* es una aplicación $|\!|\!|\cdot|\!|\!| : \mathcal{M}_n \to \mathbb{R}_+ \cup \{0\}$ verificando las siguientes propiedades:

a) $|\!|\!|A|\!|\!| = 0 \Leftrightarrow A = 0$.

b) $|\!|\!|\lambda A|\!|\!| = |\lambda| \; |\!|\!|A|\!|\!|$, $A \in \mathcal{M}_n$, $\lambda \in \mathbb{K}$.

c) $|\!|\!|A + B|\!|\!| \leq |\!|\!|A|\!|\!| + |\!|\!|B|\!|\!|$, $A, B \in \mathcal{M}_n$.

d) $|\!|\!|AB|\!|\!| \leq |\!|\!|A|\!|\!| \; |\!|\!|B|\!|\!|$, $A, B \in \mathcal{M}_n$. $\square$

Observación 2.16. Las tres primeras propiedades aseguran que una norma matricial es una norma vectorial (cuando se considera una matriz $A \in \mathcal{M}_n$ como un vector de n^2 componentes) que verifica una propiedad extra. Esta cuarta condición proporciona la "compatibilidad" de la norma con el producto de matrices. $\square$

Proposición 2.11. *Sea* $||\cdot||$ *una norma en* $\mathbf{V}$. *La aplicación* $|\!|\!|\cdot|\!|\!| : \mathcal{M}_n \to \mathbb{R}_+ \cup \{0\}$ *dada por*

$$|\!|\!|A|\!|\!| = \sup_{v \neq 0} \frac{||Av||}{||v||} = \sup_{||v||=1} ||Av|| \tag{2.12}$$

es una norma matricial.

DEMOSTRACIÓN. La aplicación $|\!|\!|\cdot|\!|\!|$ está bien definida ya que $\sup\limits_{||v||=1} ||Av|| < +\infty$ debido a la continuidad de la aplicación $v \mapsto ||Av||$ (véase la proposición 2.10) sobre la esfera unidad que es compacta. Demostrar que las dos definiciones dadas son equivalentes y que la aplicación $|\!|\!|\cdot|\!|\!|$ cumple las propiedades de norma matricial se deja como ejercicio al lector. $\square$

Definición 2.22. La norma $|\!|\!|\cdot|\!|\!|$ dada en (2.12) se denomina *norma matricial subordinada* a la norma vectorial $||\cdot||$. $\square$

Usualmente utilizaremos las siguientes normas matriciales subordinadas:

$$|\!|\!|A|\!|\!|_1 = \sup_{v \neq 0} \frac{||Av||_1}{||v||_1}, \; |\!|\!|A|\!|\!|_2 = \sup_{v \neq 0} \frac{||Av||_2}{||v||_2} \;\; \text{y} \;\; |\!|\!|A|\!|\!|_\infty = \sup_{v \neq 0} \frac{||Av||_\infty}{||v||_\infty}. \tag{2.13}$$

Observación 2.17. Existen normas matriciales que no están subordinadas a ninguna norma vectorial (véase la proposicion 2.15). □

Proposición 2.12. *Sea* $|||\cdot|||$ *una norma matricial subordinada a una norma vectorial* $||\cdot||$. *Se verifica:*

a) $||Av|| \leq |||A||| \, ||v||$, $A \in \mathcal{M}_n$, $v \in \mathbf{V}$.

b) $|||A||| = \inf\{\lambda \geq 0 : \, ||Av|| \leq \lambda \, ||v||, \, v \in \mathbf{V}\}$.

c) Existe $u \in \mathbf{V} \backslash \{0\}$ *tal que* $||Au|| = |||A||| \, ||u||$.

d) $|||I||| = 1$.

Demostración. Los apartados *a*), *b*) y *d*) se obtienen directamente de (2.12). Para mostrar *c*) basta tener en cuenta, nuevamente, la continuidad de la aplicación $v \mapsto ||Av||$ sobre la esfera unidad (compacta) para concluir que el supremo de (2.12) se alcanza. Así, si $u \in \mathbf{V}$ con $||u|| = 1$ verifica $|||A||| = ||Au||$, entonces

$$||Au|| = |||A||| \, ||u||. \quad \square$$

Observación 2.18. A la vista del apartado *b*) de la proposición 2.12, si para una norma matricial $|||\cdot|||$ subordinada a una norma vectorial $||\cdot||$ se verifica que existe una constante $M \geq 0$ verificando:

a) $||Av|| \leq M \, ||v||$, $v \in \mathbf{V}$.

b) Existe $u \in \mathbf{V} \backslash \{0\}$ tal que $||Au|| = M \, ||u||$.

entonces $M = |||A|||$. □

Veamos cómo se pueden calcular, de forma directa, las normas matriciales subordinadas más usuales definidas en (2.13):

Teorema 2.3. *Si* $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ *se verifica:*

a) $|||A|||_1 = \max\limits_{1 \leq j \leq n} \sum\limits_{i=1}^n |a_{ij}|$, *es decir, la norma* $|||\cdot|||_1$ *viene dada por la mayor de todas las cantidades que se obtienen al sumar los módulos de los elementos de cada columna.*

b) $|||A|||_2 = +\sqrt{\varrho(A^*A)} = +\sqrt{\varrho(AA^*)} = |||A^*|||_2$.

c) $|||A|||_\infty = \max\limits_{1 \leq i \leq n} \sum\limits_{j=1}^n |a_{ij}|$, *es decir, la norma* $|||\cdot|||_\infty$ *viene dada por la mayor de todas las cantidades que se obtienen al sumar los módulos de los elementos de cada fila.*

Demostración.

a) Para todo $v \in \mathbf{V}$ se verifica que

$$\begin{aligned} ||Av||_1 &= \sum_{i=1}^{n} |(Av)_i| = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} v_j \right| \\ &\leq \sum_{i=1}^{n} \left(\sum_{j=1}^{n} |a_{ij}|\, |v_j| \right) = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} |a_{ij}|\, |v_j| \right) \\ &= \sum_{j=1}^{n} |v_j| \sum_{i=1}^{n} |a_{ij}| \leq \left(\max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \right) ||v||_1 . \end{aligned} \tag{2.14}$$

Consideremos el vector $u \in \mathbf{V}$ de componentes

$$u_i = \delta_{ij_0} = \begin{cases} 1 & \text{si} \quad i = j_0 \\ 0 & \text{si} \quad i \neq j_0 \end{cases}$$

donde j_0 es un índice que verifica

$$\max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| = \sum_{i=1}^{n} |a_{ij_0}|.$$

Como para este vector se verifica $||u||_1 = 1$ y

$$||Au||_1 = \sum_{i=1}^{n} |(Au)_i| = \sum_{i=1}^{n} |a_{ij_0} u_{j_0}| = \sum_{i=1}^{n} |a_{ij_0}| = \left(\max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \right) ||u||_1 ,$$

la observación 2.18 implica que

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

b) Como $\mathrm{sp}(A^*A) = \mathrm{sp}(AA^*)$ (véase el problema 2.13), entonces

$$\varrho(A^*A) = \varrho(AA^*).$$

Por otra parte, como vimos en la proposición 2.8, la matriz A^*A es hermítica, por lo que es diagonalizable por una matriz de paso unitaria (véase la observación 2.10), es decir,

$$U^*A^*AU = D = \mathrm{diag}\,(\lambda_i(A^*A)),$$

lo que hace que se tenga que

$$A^*A = UDU^*.$$

Por tanto, a partir de (2.11), para todo $v \in \mathbf{V}$ se verifica

$$\begin{aligned} ||Av||_2^2 &= (Av)^*Av = v^*A^*Av = v^*UDU^*v \\ &= (U^*v)^*D(U^*v) = \sum_{i=1}^{n} \lambda_i(A^*A)|w_i|^2 \end{aligned}$$

siendo $w = U^*v$. Consecuentemente, usando que los autovalores de A^*A son números reales no negativos (véanse las proposiciones 2.7 y 2.8)

$$\begin{aligned} ||Av||_2^2 &\leq \varrho(A^*A) \sum_{i=1}^{n} |w_i|^2 = \varrho(A^*A)(U^*v)^*U^*v \\ &= \varrho(A^*A)v^*UU^*v = \varrho(A^*A)v^*v = \varrho(A^*A)\,||v||_2^2\,, \end{aligned} \tag{2.15}$$

donde hemos utilizado nuevamente (2.11).

Elegimos ahora un autovalor de módulo máximo

$$\lambda_i(A^*A) = \max_{1\leq j\leq n} \lambda_j(A^*A) = \varrho(A^*A)$$

y $u \in \mathbf{V}\backslash\{0\}$ un autovector asociado a él, es decir,

$$A^*Au = \lambda_i(A^*A)u.$$

Como se verifica que

$$\begin{aligned} ||Au||_2^2 &= (Au)^*Au = u^*A^*Au = \lambda_i(A^*A)u^*u \\ &= \lambda_i(A^*A)\,||u||_2^2 = \varrho(A^*A)\,||u||_2^2 \end{aligned}$$

entonces, por la observación 2.18, obtenemos que

$$|\!|\!|A|\!|\!|_2 = +\sqrt{\varrho(A^*A)}.$$

c) Para todo $v \in \mathbf{V}$ se verifica

$$\begin{aligned} ||Av||_\infty &= \max_{1\leq i\leq n} |(Av)_i| = \max_{1\leq i\leq n} \left|\sum_{j=1}^{n} a_{ij}v_j\right| \\ &\leq \max_{1\leq i\leq n} \left(\sum_{j=1}^{n} |a_{ij}|\,|v_j|\right) \leq \left(\max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|\right) ||v||_\infty\,. \end{aligned} \tag{2.16}$$

Consideremos ahora el vector $u \in \mathbf{V}$ de componentes

$$u_j = \begin{cases} \dfrac{\overline{a_{i_0 j}}}{|a_{i_0 j}|} & \text{si} \quad a_{i_0 j} \neq 0 \\ 1 & \text{si} \quad a_{i_0 j} = 0 \end{cases}$$

siendo i_0 un índice que verifica

$$\max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{i_0 j}|.$$

Por un lado, se verifica la desigualdad (2.16) para u, es decir,

$$||Au||_\infty \leq \left(\max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| \right) ||u||_\infty . \tag{2.17}$$

Por otra parte, tenemos que

$$\begin{aligned} ||Au||_\infty \geq |(Au)_{i_0}| &= \left| \sum_{j=1}^{n} a_{i_0 j} u_j \right| = \left| \sum_{\substack{j=1 \\ a_{i_0 j} \neq 0}}^{n} a_{i_0 j} u_j \right| \\ &= \left| \sum_{\substack{j=1 \\ a_{i_0 j} \neq 0}}^{n} a_{i_0 j} \frac{\overline{a_{i_0 j}}}{|a_{i_0 j}|} \right| = \sum_{\substack{j=1 \\ a_{i_0 j} \neq 0}}^{n} \frac{|a_{i_0 j}|^2}{|a_{i_0 j}|} = \sum_{\substack{j=1 \\ a_{i_0 j} \neq 0}}^{n} |a_{i_0 j}| \\ &= \sum_{j=1}^{n} |a_{i_0 j}| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| = \left(\max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| \right) ||u||_\infty , \end{aligned} \tag{2.18}$$

donde se ha usado que $||u||_\infty = 1$, al ser $|u_j| = 1$, $j = 1, 2, \ldots, n$. De esta forma, las desigualdades (2.17) y (2.18) conducen a la igualdad

$$||Au||_\infty = \left(\max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| \right) ||u||_\infty .$$

Nuevamente, la observación 2.18 implica que

$$|\!|\!| A |\!|\!|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|. \quad \square$$

Observación 2.19. De los apartados *a*) y *c*) del teorema 2.3 se deduce que

$$\|A^*\|_1 = \|A\|_\infty . \quad \square$$

Como se ha visto en el teorema 2.3, las normas $\|\cdot\|_1$ y $\|\cdot\|_\infty$ son fácilmente calculables a partir de los elementos de la matriz. Por el contrario, la norma $\|\cdot\|_2$ no tiene una expresión explícita sencilla; sin embargo, esta norma tiene buenas propiedades desde el punto de vista teórico. Veamos algunas:

Proposición 2.13. *Sea* $A \in \mathcal{M}_n$.

a) *La norma* $\|\cdot\|_2$ *es invariante por transformaciones unitarias, es decir, si* $UU^* = I$ *entonces*

$$\|A\|_2 = \|AU\|_2 = \|UA\|_2 = \|U^*AU\|_2 .$$

b) *Si* A *es normal entonces*

$$\|A\|_2 = \varrho(A).$$

Demostración.

a) Según se ha visto en el apartado *b*) del teorema 2.3, se tiene que

$$\|A\|_2^2 = \varrho(A^*A) = \varrho(A^*U^*UA) = \varrho((UA)^*(UA)) = \|UA\|_2^2 ,$$

$$\|A\|_2^2 = \varrho(AA^*) = \varrho(AUU^*A^*) = \varrho((AU)(AU)^*) = \|AU\|_2^2 ,$$

y, por las dos propiedades anteriores,

$$\|U^*AU\|_2 = \|AU\|_2 = \|A\|_2 .$$

b) Si A es normal, por el teorema 2.1, existe U unitaria tal que

$$U^*AU = D = \operatorname{diag}(\lambda_i(A)).$$

Por tanto, el apartado anterior nos asegura que

$$\|A\|_2^2 = \|U^*AU\|_2^2 = \|D\|_2^2 = \varrho(D^*D). \tag{2.19}$$

Como

$$D^* = \operatorname{diag}(\overline{\lambda_i(A)})$$

entonces

$$D^*D = \operatorname{diag}(|\lambda_i(A)|^2)$$

y, consecuentemente,

$$\mathrm{sp}(D^*D) = \left\{ |\lambda_1(A)|^2, |\lambda_2(A)|^2, \ldots, |\lambda_n(A)|^2 \right\}.$$

De esta forma, regresando a la expresión (2.19), se concluye

$$\|\!|A|\!\|_2^2 = \varrho(D^*D) = \max_{1\le i\le n} |\lambda_i(A)|^2 = \left(\max_{1\le i\le n} |\lambda_i(A)| \right)^2 = (\varrho(A))^2. \quad \square$$

Observación 2.20. Sea $A \in \mathcal{M}_n$.

a) Si A es hermítica entonces $\|\!|A|\!\|_2 = \varrho(A)$.

b) Si A es unitaria entonces $\|\!|A|\!\|_2 = +\sqrt{\varrho(A^*A)} = +\sqrt{\varrho(I)} = 1$. $\square$

Como ya se ha dicho, existen normas matriciales que no están subordinadas a ninguna norma vectorial. Vamos a construir una de ellas (que, por otra parte, no es otra que la norma euclídea en $\mathcal{M}_n$ considerado como espacio vectorial de dimensión n^2) que servirá como complemento práctico a la norma $\|\!|\cdot|\!\|_2$ (véase la observación 2.22).

Observación 2.21 Si $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ entonces $A^* = (\overline{a_{ji}})_{j,i=1}^n$, por lo que

$$A^*A = (\alpha_{ij})_{i,j=1}^n$$

siendo

$$\alpha_{ij} = \sum_{k=1}^n \overline{a_{ki}} a_{kj}$$

para $i, j = 1, 2, \ldots, n$. En particular, los elementos diagonales son de la forma

$$\alpha_{ii} = \sum_{k=1}^n \overline{a_{ki}} a_{ki} = \sum_{k=1}^n |a_{ki}|^2$$

para $i = 1, 2, \ldots, n$; consecuentemente,

$$\mathrm{tr}(A^*A) = \sum_{i=1}^n \alpha_{ii} = \sum_{i,k=1}^n |a_{ki}|^2. \quad \square$$

Proposición 2.14. *La aplicación* $\|\!|\cdot|\!\|_{\mathrm{F}} : \mathcal{M}_n \to \mathbb{R}_+ \cup \{0\}$ *dada por*

$$\|\!|A|\!\|_{\mathrm{F}} = +\sqrt{\sum_{i,j=1}^n |a_{ij}|^2} = +\sqrt{\mathrm{tr}(A^*A)} \quad \left(= +\sqrt{\mathrm{tr}(AA^*)} \right) \tag{2.20}$$

es una norma matricial.

DEMOSTRACIÓN. La aplicación $\|\cdot\|_{\mathrm{F}}$ no es más que la norma euclídea en $\mathcal{M}_n$ considerado como espacio vectorial de dimensión n^2, por lo que:

a) $\|A\|_{\mathrm{F}} = 0 \Leftrightarrow A = 0$.

b) $\|\lambda A\|_{\mathrm{F}} = |\lambda| \, \|A\|_{\mathrm{F}}$, $A \in \mathcal{M}_n$, $\lambda \in \mathbb{K}$.

c) $\|A + B\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}} + \|B\|_{\mathrm{F}}$, $A, B \in \mathcal{M}_n$.

d) Para la cuarta propiedad aplicamos la desigualdad de Cauchy–Schwarz a los vectores

$$a_i = (a_{i1}, a_{i2}, \ldots, a_{in})^{\mathrm{T}} \ \text{ y } \ b_j = (b_{1j}, b_{2j}, \ldots, b_{nj})^{\mathrm{T}},$$

obteniendo

$$\begin{aligned} \|AB\|_{\mathrm{F}}^2 &= \sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 \leq \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} |a_{ik}|^2 \right) \left(\sum_{l=1}^{n} |b_{lj}|^2 \right) \\ &= \left(\sum_{i,k=1}^{n} |a_{ik}|^2 \right) \left(\sum_{j,l=1}^{n} |b_{lj}|^2 \right) = \|A\|_{\mathrm{F}}^2 \, \|B\|_{\mathrm{F}}^2 . \quad \square \end{aligned}$$

Definición 2.23. La norma matricial $\|\cdot\|_{\mathrm{F}}$ dada en (2.20) se denomina *norma Fröbenius*. $\square$

Entre las principales propiedades de la norma Fröbenius destacamos:

Proposición 2.15. *La norma Fröbenius $\|\cdot\|_{\mathrm{F}}$ es una norma matricial no subordinada si $n \geq 2$ e invariante por transformaciones unitarias. Además,*

$$\|A\|_2 \leq \|A\|_{\mathrm{F}} \leq \sqrt{n} \, \|A\|_2 , \ A \in \mathcal{M}_n. \tag{2.21}$$

DEMOSTRACIÓN. Como

$$\|I\|_{\mathrm{F}} = \sqrt{n} \neq 1 \ \text{ si } \ n \geq 2,$$

por la proposición 2.12 se obtiene que la norma $\|\cdot\|_{\mathrm{F}}$ no está subordinada si $n \geq 2$. Por otra parte, si U es una matriz unitaria, se verifica que

$$\|A\|_{\mathrm{F}}^2 = \mathrm{tr}(A^*A) = \mathrm{tr}(A^*U^*UA) = \mathrm{tr}((UA)^*(UA)) = \|UA\|_{\mathrm{F}}^2 ,$$

$$\|A\|_{\mathrm{F}}^2 = \mathrm{tr}(AA^*) = \mathrm{tr}(AUU^*A^*) = \mathrm{tr}((AU)(AU)^*) = \|AU\|_{\mathrm{F}}^2$$

y

$$\|U^*AU\|_{\mathrm{F}}^2 = \|AU\|_{\mathrm{F}}^2 = \|A\|_{\mathrm{F}}^2 .$$

Finalmente, como los autovalores de A^*A son números reales no negativos (véanse las proposiciones 2.7 y 2.8) entonces

$$\varrho(A^*A) \leq \sum_{i=1}^{n} \lambda_i(A^*A) \leq n\varrho(A^*A).$$

Así, por el teorema 2.3, se tiene que

$$\|A\|_2^2 = \varrho(A^*A) \leq \sum_{i=1}^{n} \lambda_i(A^*A) = \operatorname{tr}(A^*A) = \|A\|_{\mathrm{F}}^2 \leq n\varrho(A^*A) = n\|A\|_2^2,$$

obteniéndose (2.21). □

Observación 2.22. Ya se ha comentado que el teorema 2.3 proporciona la manera de calcular la norma $\|\cdot\|_1$ y la norma $\|\cdot\|_\infty$ de una matriz $A \in \mathcal{M}_n$ a partir de los elementos que la componen y que no ocurre así con la norma $\|\cdot\|_2$. El interés de la norma $\|\cdot\|_{\mathrm{F}}$ es que también se calcula directamente a partir de los elementos de la matriz y, mediante (2.21), puede usarse para obtener cotas de la norma $\|\cdot\|_2$. □

Las matrices normales verifican que su norma $\|\cdot\|_2$ coincide con su radio espectral. En el caso general (es decir, el caso de una matriz y una norma cualesquiera) el resultado se convierte en una desigualdad: el radio espectral es siempre menor o igual que la norma de la matriz. Por otra parte, siempre se puede encontrar una norma tan próxima al radio espectral como se quiera. Veámoslo:

Teorema 2.4. *Sea $A \in \mathcal{M}_n$.*

a) Para toda norma matricial $\|\cdot\|$ (subordinada o no) se verifica que

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

b) Para todo $\varepsilon > 0$ existe una norma matricial $\|\cdot\|_{A,\varepsilon}$ (que se puede tomar subordinada) tal que

$$\|A\|_{A,\varepsilon} \leq \varrho(A) + \varepsilon.$$

Demostración.

a) Sea $v \in \mathbf{V}\backslash\{0\}$ un autovector asociado al autovalor λ de A de módulo máximo, es decir,

$$Av = \lambda v \ \text{ con } \ |\lambda| = \varrho(A)$$

y $w \in \mathbf{V}$ de forma que la matriz $vw^{\mathrm{T}} \in \mathcal{M}_n$ es no nula. Entonces,

$$\varrho(A)\ \|\!|vw^{\mathrm{T}}|\!\| = |\lambda|\ \|\!|vw^{\mathrm{T}}|\!\| = \|\!|\lambda vw^{\mathrm{T}}|\!\| = \|\!|Avw^{\mathrm{T}}|\!\| \leq \|A\|\ \|\!|vw^{\mathrm{T}}|\!\|,$$

de donde se sigue el resultado al ser $\|\!|vw^{\mathrm{T}}|\!\| > 0$.

b) Como $A \in \mathcal{M}_n$, por el teorema 2.1, existe una matriz unitaria U tal que $U^{-1}AU$ es triangular. Supongamos que es de la forma

$$U^{-1}AU = \begin{pmatrix} \lambda_1 & t_{12} & t_{13} & \cdots & t_{1,n-1} & t_{1n} \\ & \lambda_2 & t_{23} & \cdots & t_{2,n-1} & t_{2n} \\ & & \lambda_3 & \cdots & t_{3,n-1} & t_{3n} \\ & & & \ddots & \cdots & \cdots \\ & & & & \lambda_{n-1} & t_{n-1,n} \\ & & & & & \lambda_n \end{pmatrix}$$

donde $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ son los autovalores de A. Si para cada $\delta > 0$ consideramos la matriz diagonal

$$D_\delta = \operatorname{diag}(1, \delta, \delta^2, \ldots, \delta^{n-1}),$$

entonces

$$(UD_\delta)^{-1}A(UD_\delta) = D_\delta^{-1}U^{-1}AUD_\delta$$
$$= \begin{pmatrix} \lambda_1 & \delta t_{12} & \delta^2 t_{13} & \cdots & \delta^{n-2}t_{1,n-1} & \delta^{n-1}t_{1n} \\ & \lambda_2 & \delta t_{23} & \cdots & \delta^{n-3}t_{2,n-1} & \delta^{n-2}t_{2n} \\ & & \lambda_3 & \cdots & \delta^{n-4}t_{3,n-1} & \delta^{n-3}t_{3n} \\ & & & \ddots & \cdots & \cdots \\ & & & & \lambda_{n-1} & \delta t_{n-1,n} \\ & & & & & \lambda_n \end{pmatrix}.$$

Dado $\varepsilon > 0$ tomamos $\delta > 0$ suficientemente pequeño para que

$$\sum_{j=i+1}^{n} \delta^{j-i}|t_{ij}| < \varepsilon$$

para $i = 1, 2, \ldots, n-1$, y consideramos la aplicación $\|\!|\cdot|\!\|_{A,\varepsilon} : \mathcal{M}_n \to \mathbb{R}_+ \cup \{0\}$ dada por

$$\|\!|B|\!\|_{A,\varepsilon} = \left\|\!\left|(UD_\delta)^{-1}B(UD_\delta)\right|\!\right\|_\infty, \; B \in \mathcal{M}_n.$$

Nótese que $\|\!|\cdot|\!\|_{A,\varepsilon}$ depende de la matriz A y de ε. Claramente, $\|\!|\cdot|\!\|_{A,\varepsilon}$ es una norma matricial subordinada a la norma vectorial

$$v \mapsto \left|\left|(UD_\delta)^{-1}v\right|\right|_\infty, \; v \in \mathbf{V}.$$

Además,

$$\|\!|A|\!\|_{A,\varepsilon} = \left\|\!\left|(UD_\delta)^{-1}A(UD_\delta)\right|\!\right\|_\infty = \max_{1\le i\le n}\left(\sum_{j=i+1}^{n} \delta^{j-i}|t_{ij}| + |\lambda_i|\right)$$

$$= \max_{1\le i\le n}\sum_{j=i+1}^{n} \delta^{j-i}|t_{ij}| + \max_{1\le i\le n}|\lambda_i| < \varepsilon + \varrho(A). \quad \square$$

Si A es una matriz inversible entonces es no nula y, por tanto, $\|A\| > 0$ para cualquier norma matricial $\|\cdot\|$ (subordinada o no). Como $I = AA^{-1}$, entonces

$$\|I\| = \left\|AA^{-1}\right\| \leq \|A\| \, \left\|A^{-1}\right\|,$$

obteniéndose una acotación inferior de la norma de la inversa de la matriz A

$$\left\|A^{-1}\right\| \geq \frac{\|I\|}{\|A\|}.$$

A continuación vemos cómo obtener una acotación superior de la norma de la inversa de A cuando la matriz A puede escribirse como $A = I + B$, donde B es una matriz con norma "pequeña". Para matrices de este tipo se tiene el siguiente resultado:

Teorema 2.5. *Sea $B \in \mathcal{M}_n$. Si existe $\|\cdot\|$ norma matricial (subordinada o no) tal que*

$$\|B\| < 1 \tag{2.22}$$

entonces $I + B$ es inversible y

$$\left\|(I+B)^{-1}\right\| \leq \frac{\|I\|}{1 - \|B\|}. \tag{2.23}$$

Demostración. El teorema 2.4 y la condición (2.22) hacen que se tenga

$$\varrho(B) \leq \|B\| < 1.$$

Consecuentemente, $\lambda = -1$ no es autovalor de B y, por tanto,

$$\det(I + B) = \det(B - (-1)I) \neq 0,$$

por lo que la matriz $I + B$ es inversible. Por otra parte, como

$$(I + B)(I + B)^{-1} = I$$

se tiene que

$$(I + B)^{-1} + B(I + B)^{-1} = I,$$

es decir,

$$(I + B)^{-1} = I - B(I + B)^{-1}.$$

Por tanto,

$$\left\|(I+B)^{-1}\right\| \leq \|I\| + \|B\| \, \left\|(I+B)^{-1}\right\|,$$

de donde se sigue (2.23). □

Observación 2.23.

1. El contrarrecíproco del teorema 2.5 afirma que si $I+B$ es una matriz singular entonces $\|B\| \geq 1$ para toda norma matricial $\|\cdot\|$.

2. En el caso particular de que $\|\cdot\|$ sea una norma matricial subordinada, entonces la expresión (2.23) queda en la forma

$$\left\|(I+B)^{-1}\right\| \leq \frac{1}{1-\|B\|}.$$

3. Aplicando el teorema 2.5 a la matriz $\dfrac{1}{r}A$, se obtiene el siguiente resultado más general: si existen $r > 0$ y una norma matricial $\|\cdot\|$ tales que $A = rI + B$ con $\|B\| < r$, entonces la matriz A es inversible y

$$\left\|A^{-1}\right\| \leq \frac{\|I\|}{r-\|B\|}$$

(véase el problema 2.28). □

2.4. Convergencia de las iteraciones de una matriz

Para el estudio de los métodos iterativos que se llevará a cabo en el capítulo 5, el paso fundamental será determinar cuándo las sucesivas potencias de una matriz tienden hacia cero. En esta sección abordamos esta cuestión. Para ello, comenzamos con la siguiente definición:

Definición 2.24. Sea $||\cdot||$ una norma vectorial. Una sucesión de vectores $\{v^k\}_{k=1}^{\infty}$ de $\mathbf{V}$ *converge* a un vector $v \in \mathbf{V}$ si

$$\lim_{k\to+\infty} \left|\left|v^k - v\right|\right| = 0.$$

En tal caso, se denota $v = \lim_{k\to+\infty} v^k$. □

Observación 2.24.

1. La equivalencia de las normas en $\mathbf{V}$ muestra que la convergencia de una sucesión de vectores es independiente de la norma elegida. De esta forma (tomando, por ejemplo, la norma $||\cdot||_\infty$) se demuestra fácilmente que:

$$\lim_{k\to+\infty} v^k = v \Leftrightarrow \lim_{k\to+\infty} v_i^k = v_i,\ i = 1, 2, \ldots, n$$

siendo $v^k = (v_1^k, v_2^k, \ldots, v_n^k)^{\mathrm{T}}$, $k \in \mathbb{N}$, y $v = (v_1, v_2, \ldots, v_n)^{\mathrm{T}}$.

2. En particular, la noción anterior incluye el caso de la convergencia de matrices, pues basta considerar $\mathcal{M}_n(\mathbb{K})$ como espacio vectorial de dimensión n^2. Concretamente, a partir de una norma matricial $|||\cdot|||$, una sucesión de matrices $\{A_k\}_{k=1}^{\infty} \subset \mathcal{M}_n$ *converge* a una matriz $A \in \mathcal{M}_n$, y lo denotaremos

$$A = \lim_{k \to +\infty} A_k,$$

si

$$\lim_{k \to +\infty} |||A_k - A||| = 0. \quad \square$$

Ejemplo 2.2.

1. La sucesión de vectores

$$v^k = \left(\frac{2}{k^3}, 1 - \frac{1}{k^2}, e^{\frac{1}{k}} \right)^{\mathrm{T}} \in \mathbb{R}^3$$

es convergente al vector

$$v = \lim_{k \to +\infty} v^k = (0, 1, 1)^{\mathrm{T}}.$$

2. La sucesión de matrices

$$A_k = \begin{pmatrix} 1 + \dfrac{k}{k^2 + 3} & \dfrac{4}{k} \\ \dfrac{1}{k} + \dfrac{2}{k^2} & 1 - e^{-\frac{3}{k^4}} \end{pmatrix} \in \mathcal{M}_2$$

converge a la matriz

$$A = \lim_{k \to +\infty} A_k = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}. \quad \square$$

El siguiente resultado caracteriza la convergencia a cero de las potencias sucesivas B^k de una matriz cuadrada B.

Teorema 2.6. *Sea $B \in \mathcal{M}_n$. Son equivalentes:*

a) $\lim_{k \to +\infty} B^k = 0.$

b) $\lim_{k \to +\infty} B^k v = 0,\ v \in \mathbf{V}.$

c) $\varrho(B) < 1.$

d) Existe $|||\cdot|||$ norma matricial (que se puede tomar subordinada) tal que

$$|||B||| < 1.$$

DEMOSTRACIÓN.

$\boxed{a)\Rightarrow b)}$ Sea $|||\cdot|||$ la norma matricial subordinada a una norma vectorial $||\cdot||$. Por definición,

$$\lim_{k\to+\infty} B^k = 0 \Leftrightarrow \lim_{k\to+\infty} |||B^k||| = 0.$$

Por tanto, como para todo $v \in \mathbf{V}$ se verifica que

$$||B^k v|| \leq |||B^k|||\ ||v||\,,\ k \in \mathbb{N}$$

entonces

$$\lim_{k\to+\infty} ||B^k v|| = 0$$

y, así,

$$\lim_{k\to+\infty} B^k v = 0.$$

$\boxed{b)\Rightarrow c)}$ Argumentamos por reducción al absurdo. Si $\varrho(B) \geq 1$ entonces existe un autovalor $\lambda = \lambda(B) \in \mathrm{sp}(B)$ con $|\lambda| \geq 1$; basta considerar un autovector $v \in \mathbf{V}\backslash\{0\}$ asociado a λ para llegar a una contradicción. En efecto, como $Bv = \lambda v$ entonces

$$B^k v = \lambda^k v,\ k \in \mathbb{N},$$

y, por tanto,

$$\lim_{k\to+\infty} ||B^k v|| = \lim_{k\to+\infty} |\lambda|^k\, ||v|| \neq 0.$$

$\boxed{c)\Rightarrow d)}$ Por el teorema 2.4, dado $\varepsilon > 0$ existe una norma matricial $|||\cdot|||_{B,\varepsilon}$ tal que $|||B|||_{B,\varepsilon} \leq \varrho(B) + \varepsilon$. Tomando

$$0 < \varepsilon < 1 - \varrho(B)$$

se obtiene que

$$|||B|||_{B,\varepsilon} < \varrho(B) + (1 - \varrho(B)) = 1.$$

$\boxed{d)\Rightarrow a)}$ Claramente,

$$|||B^k||| = |||B^{k-1}B||| \leq |||B^{k-1}|||\ |||B||| \leq \cdots \leq |||B|||^k\,,\ k \in \mathbb{N}. \tag{2.24}$$

Por tanto, la hipótesis $|||B||| < 1$ implica

$$\lim_{k\to+\infty} |||B^k||| = 0,$$

es decir,

$$\lim_{k\to+\infty} B^k = 0. \quad \square$$

Observación 2.25. En la práctica, el resultado anterior se utiliza del siguiente modo: si se quiere demostrar que las potencias sucesivas de una matriz B convergen a cero, bastará probar que todos sus autovalores tienen módulo menor que uno, o bien encontrar una norma matricial para la que $|||B||| < 1$. □

El siguiente resultado muestra que la norma de las sucesivas potencias de una matriz se comporta, en el límite, como las sucesivas potencias de su radio espectral:

Teorema 2.7. *Si $B \in \mathcal{M}_n$ y $|||\cdot|||$ es una norma matricial (subordinada o no) entonces*

$$\lim_{k\to+\infty} |||B^k|||^{\frac{1}{k}} = \varrho(B). \tag{2.25}$$

Demostración. Como

$$(\varrho(B))^k = \varrho(B^k),\ k \in \mathbb{N}, \tag{2.26}$$

(véase el problema 2.14), el teorema 2.4 determina, para todo $k \in \mathbb{N}$,

$$(\varrho(B))^k = \varrho(B^k) \leq |||B^k|||$$

y, por tanto,

$$\varrho(B) \leq |||B^k|||^{\frac{1}{k}},\ k \in \mathbb{N}. \tag{2.27}$$

Para obtener (2.25), debemos demostrar que para todo $\varepsilon > 0$ existe un número $k_0 = k_0(\varepsilon) \in \mathbb{N}$ tal que

$$\left| |||B^k|||^{\frac{1}{k}} - \varrho(B) \right| < \varepsilon,\ k \geq k_0,$$

lo que equivale, a la vista de (2.27), a mostrar que

$$|||B^k|||^{\frac{1}{k}} < \varrho(B) + \varepsilon,\ k \geq k_0.$$

Para ello, a partir de ε consideramos la matriz

$$B_\varepsilon = \frac{B}{\varrho(B) + \varepsilon}.$$

Como

$$\varrho(B_\varepsilon) = \frac{\varrho(B)}{\varrho(B) + \varepsilon} < 1,$$

aplicando el teorema 2.6 obtenemos

$$\lim_{k\to+\infty} B_\varepsilon^k = 0,$$

es decir,

$$\lim_{k\to+\infty} \left\|\left| B_\varepsilon^k \right|\right\| = 0.$$

Del hecho de que esta sucesión de números reales converja a cero deducimos que, dado $\widetilde{\varepsilon} = 1$, existe $k_0 \in \mathbb{N}$ tal que

$$\left\|\left| B_\varepsilon^k \right|\right\| < 1$$

para $k \geq k_0$. Consecuentemente, para todo $k \geq k_0$ se verifica

$$\frac{\left\|\left| B^k \right|\right\|}{(\varrho(B)+\varepsilon)^k} = \left\|\left| B_\varepsilon^k \right|\right\| < 1,$$

de donde

$$\left\|\left| B^k \right|\right\| < (\varrho(B)+\varepsilon)^k$$

y, por tanto,

$$\left\|\left| B^k \right|\right\|^{\frac{1}{k}} < \varrho(B)+\varepsilon$$

concluyendo, así, el resultado. □

2.5. Problemas

2.5.1. Problemas resueltos

2.1. Si $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \in \mathcal{M}_{n\times 1}$ y $B = \begin{pmatrix} b_1^{\mathrm{T}} \\ b_2^{\mathrm{T}} \\ \cdots \\ b_m^{\mathrm{T}} \end{pmatrix} \in \mathcal{M}_{m\times n}$ probar que

$$Bx = \begin{pmatrix} b_1^{\mathrm{T}} x \\ b_2^{\mathrm{T}} x \\ \cdots \\ b_m^{\mathrm{T}} x \end{pmatrix}.$$

Solución. Comprobemos que para cada $i = 1, 2, \ldots, m$ se tiene que $b_i^{\mathrm{T}} x$ es la componente i–ésima del vector $Bx \in \mathcal{M}_{m\times 1}$. En efecto,

$$b_i^{\mathrm{T}} x = (b_{i1}, b_{i2}, \ldots, b_{in}) \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix} = \sum_{j=1}^{n} b_{ij} x_j = (Bx)_i. \quad \square$$

2.2. Si $B = (b_1, b_2, \ldots, b_n) \in \mathcal{M}_{m\times n}$ y $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \in \mathcal{M}_{n\times 1}$ demostrar que

$$Bx = \sum_{i=1}^{n} x_i b_i = x_1 b_1 + x_2 b_2 + \cdots + x_n b_n.$$

SOLUCIÓN. Teniendo en cuenta que $Bx \in \mathcal{M}_{m\times 1}$ y

$$(Bx)_i = \sum_{j=1}^{n} b_{ij} x_j$$

para $i = 1, 2, \ldots, m$, se verifica que

$$Bx = \begin{pmatrix} \sum_{j=1}^{n} b_{1j} x_j \\ \sum_{j=1}^{n} b_{2j} x_j \\ \cdots \\ \sum_{j=1}^{n} b_{mj} x_j \end{pmatrix} = \sum_{j=1}^{n} x_j \begin{pmatrix} b_{1j} \\ b_{2j} \\ \cdots \\ b_{mj} \end{pmatrix} = \sum_{j=1}^{n} x_j b_j. \quad \square$$

2.3. Si $A \in \mathcal{M}_{m\times n}$ y $B = (b_1, b_2, \ldots, b_p) \in \mathcal{M}_{n\times p}$ probar que

$$AB = (Ab_1, Ab_2, \ldots, Ab_p).$$

SOLUCIÓN. Como $AB \in \mathcal{M}_{m\times p}$ y

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

para $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, p$, basta tener en cuenta que para cada índice $j \in \{1, 2, \ldots, p\}$ la columna j–ésima de la matriz AB es

$$(AB)_j = \begin{pmatrix} \sum_{k=1}^{n} a_{1k} b_{kj} \\ \sum_{k=1}^{n} a_{2k} b_{kj} \\ \cdots \\ \sum_{k=1}^{n} a_{mk} b_{kj} \end{pmatrix} = A \begin{pmatrix} b_{1j} \\ b_{2j} \\ \cdots \\ b_{nj} \end{pmatrix} = Ab_j. \quad \square$$

2.4. Demostrar que si $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{n\times p}$ y $x \in \mathcal{M}_{p\times 1}$ entonces

$$A(Bx) = (AB)x.$$

SOLUCIÓN. Por el problema 2.2 se tiene que si $B = (b_1, b_2, \ldots, b_p)$ entonces

$$A(Bx) = A\left(\sum_{j=1}^{p} x_j b_j\right) = \sum_{j=1}^{p} x_j \left(Ab_j\right).$$

Como, nuevamente, por el problema 2.2 se verifica que

$$\sum_{j=1}^{p} x_j \left(Ab_j\right) = \left(Ab_1, Ab_2, \ldots, Ab_p\right) x$$

podemos aplicar el problema 2.3 para concluir que

$$A(Bx) = \left(Ab_1, Ab_2, \ldots, Ab_p\right) x = (AB)x. \quad \square$$

2.5. Probar que si $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{n\times p}$ y $C \in \mathcal{M}_{p\times q}$ entonces

$$A(BC) = (AB)C.$$

SOLUCIÓN. Denotando por $C = (c_1, c_2, \ldots, c_q) \in \mathcal{M}_{p\times q}$ podemos escribir

$$\begin{aligned} A(BC) &= A\left(B\left(c_1, c_2, \ldots, c_q\right)\right) = A\left(Bc_1, Bc_2, \ldots, Bc_q\right) \\ &= \left(A\left(Bc_1\right), A\left(Bc_2\right), \ldots, A\left(Bc_q\right)\right) \end{aligned}$$

donde hemos utilizado dos veces el resultado del problema 2.3. Como por el problema 2.4 se verifica que

$$A\left(Bc_i\right) = \left(AB\right) c_i$$

para $i = 1, 2, \ldots, q$, el resultado se concluye aplicando nuevamente el problema 2.3, ya que

$$\left((AB)c_1, (AB)c_2, \ldots, (AB)c_q\right) = (AB)\left(c_1, c_2, \ldots, c_q\right) = (AB)C. \quad \square$$

2.6. Sean $A, D \in \mathcal{M}_n$ con $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$. Encontrar las expresiones de DA y AD.

SOLUCIÓN. Denotando por $A = (a_{ij})_{i,j=1}^{n}$ y $D = (d_{ij})_{i,j=1}^{n}$ con

$$d_{ij} = 0 \text{ si } i \neq j$$

para cada par de elementos $i, j \in \{1, 2, \ldots, n\}$ se verifica que

$$(AD)_{ij} = \sum_{k=1}^{n} a_{ik} d_{kj} = a_{ij} d_{jj} \text{ y } (DA)_{ij} = \sum_{k=1}^{n} d_{ik} a_{kj} = d_{ii} a_{ij}$$

por lo que las matrices AD y DA son de la forma

$$AD = \begin{pmatrix} a_{11}d_{11} & a_{12}d_{22} & \cdots & a_{1n}d_{nn} \\ a_{21}d_{11} & a_{22}d_{22} & \cdots & a_{2n}d_{nn} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1}d_{11} & a_{n2}d_{22} & \cdots & a_{nn}d_{nn} \end{pmatrix}$$

y

$$DA = \begin{pmatrix} d_{11}a_{11} & d_{11}a_{12} & \cdots & d_{11}a_{1n} \\ d_{22}a_{21} & d_{22}a_{22} & \cdots & d_{22}a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ d_{nn}a_{n1} & d_{nn}a_{n2} & \cdots & d_{nn}a_{nn} \end{pmatrix}.$$

Observación: nótese que si

$$\begin{cases} A = (a_1, a_2, \ldots, a_n) & \Rightarrow \quad AD = (d_{11}a_1, d_{22}a_2, \ldots, d_{nn}a_n) \\ A = \begin{pmatrix} a_1^{\mathrm{T}} \\ a_2^{\mathrm{T}} \\ \cdots \\ a_n^{\mathrm{T}} \end{pmatrix} & \Rightarrow \quad DA = \begin{pmatrix} d_{11}a_1^{\mathrm{T}} \\ d_{22}a_2^{\mathrm{T}} \\ \cdots \\ d_{nn}a_n^{\mathrm{T}} \end{pmatrix}. \end{cases}$$

Es decir, si se multiplica la matriz A por una matriz diagonal por la derecha las columnas de A quedan multiplicadas por los elementos correspondientes de la matriz diagonal, mientras que si se hace por la izquierda, son las filas las que quedan multiplicadas. □

2.7. Sean A y B matrices triangulares superiores (respectivamente inferiores) y λ un escalar. Demostrar que λA, $A + B$ y AB son matrices triangulares superiores (respectivamente inferiores) y determinar sus elementos diagonales.

Solución. Únicamente consideramos el caso en que A y B son triangulares superiores (el otro es totalmente análogo), es decir, $A = (a_{ij})_{i,j=1}^{n}$ y $B = (b_{ij})_{i,j=1}^{n}$ con

$$a_{ij} = b_{ij} = 0 \text{ si } i > j. \tag{2.28}$$

Claramente, $\lambda A = (\alpha_{ij})_{i,j=1}^{n}$, $A + B = (\beta_{ij})_{i,j=1}^{n}$ y $AB = (\gamma_{ij})_{i,j=1}^{n}$ siendo

$$\alpha_{ij} = \lambda a_{ij},\ \beta_{ij} = a_{ij} + b_{ij} \text{ y } \gamma_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

para $i, j = 1, 2, \ldots, n$. Además, la propiedad (2.28) hace que para $i > j$ se tenga

$$\alpha_{ij} = 0,\ \beta_{ij} = 0 \ \text{ y } \ \gamma_{ij} = \sum_{k=1}^{j} a_{ik} b_{kj} + \sum_{k=j+1}^{n} a_{ik} b_{kj} = 0 + 0 = 0.$$

Por otra parte, para cada $i \in \{1, 2, \ldots, n\}$ se verifica que

$$\gamma_{ii} = \sum_{k=1}^{n} a_{ik} b_{ki} = \sum_{k=1}^{i-1} a_{ik} b_{ki} + a_{ii} b_{ii} + \sum_{k=i+1}^{n} a_{ik} b_{ki} = 0 + a_{ii} b_{ii} + 0 = a_{ii} b_{ii}$$

por lo que los elementos diagonales de estas matrices son

$$\alpha_{ii} = \lambda a_{ii},\ \beta_{ii} = a_{ii} + b_{ii} \ \text{ y } \ \gamma_{ii} = a_{ii} b_{ii}$$

para $i = 1, 2, \ldots, n$. □

2.8. Demostrar que si D es una matriz diagonal e inversible, su inversa es también diagonal. Determinar D^{-1}.

SOLUCIÓN. Si la matriz $D = \operatorname{diag}(d_{11}, d_{22}, \ldots, d_{nn})$ es inversible, como

$$\det(D) = \prod_{i=1}^{n} d_{ii},$$

se verifica que

$$d_{ii} \neq 0$$

para $i = 1, 2, \ldots, n$. Por otra parte, si $D^{-1} = (a_1, a_2, \ldots, a_n) \in \mathcal{M}_n$ entonces

$$I = D^{-1} D = (d_{11} a_1, d_{22} a_2, \ldots, d_{nn} a_n)$$

(véase el problema 2.6) de donde se obtiene que

$$a_i = \frac{1}{d_{ii}} \mathbf{e}_i$$

para $i = 1, 2, \ldots, n$, siendo $\mathbf{e}_i = (0, \ldots, 0, \overset{i)}{1}, 0, \ldots, 0)^{\mathrm{T}}$ el i–ésimo vector de la base canónica. Es decir, que D^{-1} es también una matriz diagonal de la forma

$$D^{-1} = \operatorname{diag}\left(\frac{1}{d_{11}}, \frac{1}{d_{22}}, \ldots, \frac{1}{d_{nn}} \right). \quad \square$$

2.9. Si A es una matriz triangular superior (respectivamente inferior) e inversible probar que su inversa es también triangular superior (respectivamente inferior). Determinar los elementos diagonales de A^{-1}.

SOLUCIÓN. Si la matriz $A = (a_{ij})_{i,j=1}^{n}$ es triangular superior e inversible, como

$$\det(A) = \prod_{i=1}^{n} a_{ii},$$

entonces

$$a_{ii} \neq 0 \tag{2.29}$$

para $i = 1, 2, \ldots, n$. Por otra parte, si $A^{-1} = (b_1, b_2, \ldots, b_n) \in \mathcal{M}_n$, de la relación $AA^{-1} = I$ se obtiene que

$$(Ab_1, Ab_2, \ldots, Ab_n) = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$$

y, por tanto, para cada $i \in \{1, 2, \ldots, n\}$,

$$Ab_i = \mathbf{e}_i$$

siendo $\mathbf{e}_i = (0, \ldots, 0, \overset{i)}{1}, 0, \ldots, 0)^{\mathrm{T}}$ (véase el problema 2.3). De esta forma, para cada índice $i \in \{1, 2, \ldots, n\}$ se verifica que

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & \cdots & a_{1,n-1} & a_{1n} \\ & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & a_{i-1,i-1} & a_{i-1,i} & \cdots & a_{i-1,n-1} & a_{i-1,n} \\ & & & a_{ii} & \cdots & a_{i,n-1} & a_{in} \\ & & & & \ddots & \cdots & \cdots \\ & & & & & a_{n-1,n-1} & a_{n-1,n} \\ & & & & & & a_{nn} \end{pmatrix} \begin{pmatrix} b_{1i} \\ \cdots \\ b_{i-1,i} \\ b_{ii} \\ b_{i+1,i} \\ \cdots \\ b_{ni} \end{pmatrix} = \begin{pmatrix} 0 \\ \cdots \\ 0 \\ 1 \\ 0 \\ \cdots \\ 0 \end{pmatrix}.$$

Multiplicando por cajas se obtiene que

$$\begin{pmatrix} a_{i+1,i+1} & a_{i+1,i+2} & \cdots & a_{i+1,n} \\ & a_{i+2,i+2} & \cdots & a_{i+2,n} \\ & & \ddots & \cdots \\ & & & a_{nn} \end{pmatrix} \begin{pmatrix} b_{i+1,i} \\ b_{i+2,i} \\ \cdots \\ b_{ni} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix},$$

de donde, por ser esta submatriz inversible (véase (2.29)), se tiene que

$$b_{ji} = 0$$

para $j = i+1, i+2, \ldots, n$. Por tanto la matriz A^{-1} es también triangular superior. Además, los elementos diagonales de A^{-1} son

$$b_{ii} = \frac{1}{a_{ii}}$$

para $i = 1, 2, \ldots, n$. □

2.10. Demostrar que si $A \in \mathcal{M}_n$ es una matriz triangular y normal entonces A es diagonal.

SOLUCIÓN. Si la matriz $A = (a_{ij})_{i,j=1}^n$ es triangular superior entonces

$$a_{ij} = 0 \text{ si } i > j \tag{2.30}$$

y la matriz $A^* = (\overline{a_{ji}})_{j,i=1}^n$ es triangular inferior. Por otra parte, por ser A una matriz normal, se verifica

$$A^*A = AA^*,$$

por lo que, en particular, se cumple que

$$(A^*A)_{ii} = (AA^*)_{ii} \tag{2.31}$$

para $i = 1, 2, \ldots, n$. Vamos a probar por inducción en i que para todo índice $i \in \{1, 2, \ldots, n\}$ se verifica que

$$a_{ij} = 0 \text{ si } i < j. \tag{2.32}$$

i) Para $i = 1$, se tiene que

$$\begin{cases} (A^*A)_{11} = \displaystyle\sum_{k=1}^n \overline{a_{k1}} a_{k1} = |a_{11}|^2 \\ (AA^*)_{11} = \displaystyle\sum_{k=1}^n a_{1k}\, \overline{a_{1k}} = \sum_{k=1}^n |a_{1k}|^2 \end{cases}$$

(véase (2.30)), por lo que, gracias a (2.31), tendremos

$$a_{12} = a_{13} = \cdots = a_{1n} = 0.$$

ii) Supuesto cierto el resultado para las filas anteriores a la i–ésima, en particular

$$a_{1i} = a_{2i} = \cdots = a_{i-1,i} = 0,$$

por lo que, gracias a (2.30), se tiene que

$$\begin{cases} (A^*A)_{ii} = \displaystyle\sum_{k=1}^n \overline{a_{ki}} a_{ki} = |a_{1i}|^2 + |a_{2i}|^2 + \cdots + |a_{i-1,i}|^2 + |a_{ii}|^2 = |a_{ii}|^2 \\ (AA^*)_{ii} = \displaystyle\sum_{k=1}^n a_{ik}\, \overline{a_{ik}} = \sum_{k=1}^n |a_{ik}|^2 = \sum_{k=i}^n |a_{ik}|^2. \end{cases}$$

Nuevamente, la relación (2.31) conduce a

$$a_{i,i+1} = a_{i,i+2} = \cdots = a_{in} = 0.$$

Así, las relaciones (2.30) y (2.32) hacen que $A = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$. □

2.11. Demostrar que para todo $p \geq 1$ se verifica que

$$||v||_\infty \leq ||v||_p \leq \sqrt[p]{n}\ ||v||_\infty\,,\ v \in \mathbf{V}. \tag{2.33}$$

Concluir que

$$||v||_\infty = \lim_{p\to+\infty} ||v||_p\,,\ v \in \mathbf{V}. \tag{2.34}$$

SOLUCIÓN. Sea $p \geq 1$. Para todo $v = (v_1, v_2, \ldots, v_n)^{\mathrm{T}} \in \mathbf{V}$ se tiene que

$$||v||_\infty^p = \left(\max_{1\leq i\leq n} |v_i|\right)^p = \max_{1\leq i\leq n} |v_i|^p \leq \sum_{i=1}^n |v_i|^p = ||v||_p^p$$

y

$$||v||_p^p = \sum_{i=1}^n |v_i|^p \leq n \max_{1\leq i\leq n} |v_i|^p = n\,||v||_\infty^p$$

de donde se sigue (2.33). Finalmente, teniendo en cuenta que

$$\lim_{p\to+\infty} \sqrt[p]{n} = 1,$$

basta hacer tender $p \to +\infty$ en la expresión (2.33) para obtener (2.34). □

2.12. Factorización QR y método de Gram–Schmidt. Sea $A \in \mathcal{M}_n$ inversible. Demostrar que existen $Q \in \mathcal{M}_n$ unitaria y $R \in \mathcal{M}_n$ triangular superior e inversible de forma que $A = QR$.

SOLUCIÓN. La demostración de este resultado se basa en el *proceso de ortonormalización de Gram–Schmidt*. Si $A = (a_1, a_2, \ldots, a_n)$, se definen los vectores

$$q_j = \frac{1}{||b_j||_2} b_j$$

siendo

$$b_j = a_j - \sum_{k=1}^{j-1} (q_k^* a_j)\, q_k \tag{2.35}$$

para $j = 1, 2, \ldots, n$. Nótese que esta definición hace que cada columna a_j de A pueda escribirse como combinación lineal de los vectores $\{q_1, q_2, \ldots, q_j\}$, es decir,

$$a_j = r_{1j}q_1 + r_{2j}q_2 + \cdots + r_{jj}q_j$$

para $j = 1, 2, \ldots, n$. Por tanto, tomando

$$r_{ij} = 0 \ \text{ si } \ i > j,$$

$R = (r_{ij})_{i,j=1}^n$ y $Q = (q_1, q_2, \ldots, q_n)$, se verifica que $A = QR$. Obviamente, la matriz R es triangular superior; para finalizar la prueba basta demostrar que la matriz Q es unitaria y, para ello, es suficiente probar que

$$q_j^* q_i = \delta_{ij} \text{ si } i \leq j$$

para $j = 1, 2, \ldots, n$. Claramente, para todo índice j, se verifica que

$$q_j^* q_j = ||q_j||_2^2 = \frac{1}{||b_j||_2^2} ||b_j||_2^2 = 1$$

Mostremos entonces, por inducción en j, que

$$q_j^* q_i = 0 \text{ si } i < j$$

cuando $j = 2, 3, \ldots, n$:

i) Para $j = 2$

$$\begin{aligned} q_2^* q_1 &= \frac{1}{||b_2^*||_2} b_2^* q_1 = \frac{1}{||b_2^*||_2} \left(a_2^* - \overline{(q_1^* a_2)} q_1^*\right) q_1 \\ &= \frac{1}{||b_2^*||_2} \left(a_2^* q_1 - (a_2^* q_1)\, q_1^* q_1\right) = \frac{1}{||b_2^*||_2} \left(a_2^* q_1 - a_2^* q_1\right) = 0. \end{aligned}$$

ii) Supuesto cierto el resultado para los índices menores que j, es decir, suponiendo que

$$q_k^* q_i = 0 \text{ si } i < k < j,$$

se tiene, a partir de la relación (2.35),

$$\begin{aligned} q_j^* q_i &= \frac{1}{||b_j^*||_2} b_j^* q_i = \frac{1}{||b_j^*||_2} \left(a_j^* - \sum_{k=1}^{j-1} \overline{(q_k^* a_j)} q_k^*\right) q_i \\ &= \frac{1}{||b_j^*||_2} \left(a_j^* q_i - \left(a_j^* q_i\right) q_i^* q_i\right) = \frac{1}{||b_j^*||_2} \left(a_j^* q_i - a_j^* q_i\right) = 0. \end{aligned}$$

Finalmente, como A es inversible y $A = QR$, la matriz R es inversible. □

2.13. Dadas dos matrices $A, B \in \mathcal{M}_n$ demostrar que

$$\text{sp}(AB) = \text{sp}(BA).$$

SOLUCIÓN. Es suficiente probar que $\text{sp}(AB) \subset \text{sp}(BA)$, pues intercambiando los papeles de A y B se obtendría la otra inclusión. Para ello, dado $\lambda \in \text{sp}(AB)$ pueden presentarse dos casos:

a) Si $\lambda = 0$ entonces

$$0 = \det(AB) = \det(BA),$$

luego $\lambda = 0 \in \mathrm{sp}(BA)$.

b) Si $\lambda \neq 0$, sea $v \in \mathbf{V}\backslash\{0\}$ un autovector asociado al autovalor λ, es decir,

$$ABv = \lambda v.$$

De esta forma, multiplicando la anterior expresión por la matriz B se obtiene

$$BA(Bv) = \lambda Bv,$$

es decir,

$$BAw = \lambda w \tag{2.36}$$

siendo $w = Bv$. Veamos que $w \in \mathbf{V}\backslash\{0\}$ argumentando por reducción al absurdo: en el caso de que $Bv = w = 0$ se tendría

$$0 = ABv = \lambda v$$

lo cual no es posible por ser $\lambda \neq 0$ y $v \in \mathbf{V}\backslash\{0\}$. Por tanto, la relación (2.36) determina que λ es un autovalor de la matriz BA. □

2.14. Si $A \in \mathcal{M}_n$ probar que

$$\varrho(A^k) = (\varrho(A))^k,\ k \in \mathbb{N}.$$

Solución. Por el teorema 2.1, existe una matriz unitaria U tal que $T = U^*AU$ es triangular. Supongamos que

$$T = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ & t_{22} & \cdots & t_{2n} \\ & & \ddots & \cdots \\ & & & t_{nn} \end{pmatrix}$$

donde, como sabemos,

$$\mathrm{sp}(A) = \mathrm{sp}(U^*AU) = \mathrm{sp}(T) = \{t_{11}, t_{22}, \ldots, t_{nn}\}. \tag{2.37}$$

Sea $k \in \mathbb{N}$. Como el producto de matrices triangulares superiores es una matriz triangular superior (véase el problema 2.7), la matriz unitaria U que triangulariza A también triangulariza A^k ya que

$$A^k = A \overset{k)}{\cdots} A = (UTU^*) \overset{k)}{\cdots} (UTU^*) = UT^kU^*.$$

Como

$$T^k = \begin{pmatrix} (t_{11})^k & \nu_{12} & \cdots & \nu_{1n} \\ & (t_{22})^k & \cdots & \nu_{2n} \\ & & \ddots & \cdots \\ & & & (t_{nn})^k \end{pmatrix}$$

entonces

$$\mathrm{sp}(A^k) = \mathrm{sp}(UT^kU^*) = \mathrm{sp}(T^k) = \{(t_{11})^k, (t_{22})^k, \ldots, (t_{nn})^k\},$$

es decir, los autovalores de la matriz A^k son los autovalores de la matriz A elevados a la potencia k–ésima (véase (2.37)). Por tanto,

$$\varrho(A^k) = \max_{1\le i\le n} \left|\lambda_i(A^k)\right| = \left(\max_{1\le i\le n} |\lambda_i(A)|\right)^k = (\varrho(A))^k. \quad \square$$

2.15. Demostrar que si $A \in \mathcal{M}_n$ es triangular por bloques entonces

$$\mathrm{sp}(A) = \bigcup_{i=1}^{p} \mathrm{sp}(A_{ii})$$

siendo A_{ii}, $i = 1, 2, \ldots, p$ los bloques de la diagonal de A. Deducir que el determinante de una matriz triangular por bloques es el producto de los determinantes de los bloques de su diagonal.

Solución. Sea A una matriz triangular superior por bloques de la forma

$$A = \left(\begin{array}{c|c|c|c} A_{11} & A_{12} & \cdots & A_{1p} \\ \hline & A_{22} & \cdots & A_{2p} \\ \hline & & \ddots & \cdots \\ \hline & & & A_{pp} \end{array}\right)$$

donde $A_{ii} \in \mathcal{M}_{n_i}$ para $i = 1, 2, \ldots, p$, siendo $n_1 + n_2 + \cdots + n_p = n$. Para cada $i \in \{1, 2, \ldots, p\}$ dada la submatriz $A_{ii} \in \mathcal{M}_{n_i}$ existirá una matriz $U_i \in \mathcal{M}_{n_i}$ unitaria tal que $T_i = U_i^* A_{ii} U_i$ es una matriz triangular (véase el teorema 2.1). Supongamos que T_i es triangular superior de la forma

$$T_i = \begin{pmatrix} t^i_{11} & t^i_{12} & \cdots & t^i_{1n_i} \\ & t^i_{22} & \cdots & t^i_{2n_i} \\ & & \ddots & \cdots \\ & & & t^i_{n_i n_i} \end{pmatrix}$$

para $i = 1, 2, \ldots, p$. De esta forma

$$\mathrm{sp}(A_{ii}) = \mathrm{sp}\,(U_i^* A_{ii} U_i) = \mathrm{sp}(T_i)$$

y

$$\det(A_{ii}) = \det(T_i)$$

para $i = 1, 2, \ldots, p$. A partir de la matriz diagonal por bloques

$$U = \mathrm{diag}(U_1, U_2, \ldots, U_p) = \left(\begin{array}{c|c|c|c} U_1 & & & \\ \hline & U_2 & & \\ \hline & & \ddots & \\ \hline & & & U_p \end{array}\right)$$

se verifica que

$$\begin{aligned} U^* A U &= \left(\begin{array}{c|c|c|c} U_1^* & & & \\ \hline & U_2^* & & \\ \hline & & \ddots & \\ \hline & & & U_p^* \end{array}\right) \left(\begin{array}{c|c|c|c} A_{11} & A_{12} & \cdots & A_{1p} \\ \hline & A_{22} & \cdots & A_{2p} \\ \hline & & \ddots & \cdots \\ \hline & & & A_{pp} \end{array}\right) U \\ &= \left(\begin{array}{c|c|c|c} U_1^* A_{11} & U_1^* A_{12} & \cdots & U_1^* A_{1p} \\ \hline & U_2^* A_{22} & \cdots & U_2^* A_{2p} \\ \hline & & \ddots & \cdots \\ \hline & & & U_p^* A_{pp} \end{array}\right) \left(\begin{array}{c|c|c|c} U_1 & & & \\ \hline & U_2 & & \\ \hline & & \ddots & \\ \hline & & & U_p \end{array}\right) \\ &= \left(\begin{array}{c|c|c|c} U_1^* A_{11} U_1 & U_1^* A_{12} U_2 & \cdots & U_1^* A_{1p} U_p \\ \hline & U_2^* A_{22} U_2 & \cdots & U_2^* A_{2p} U_p \\ \hline & & \ddots & \cdots \\ \hline & & & U_p^* A_{pp} U_p \end{array}\right) \\ &= \left(\begin{array}{c|c|c|c} T_1 & U_1^* A_{12} U_2 & \cdots & U_1^* A_{1p} U_p \\ \hline & T_2 & \cdots & U_2^* A_{2p} U_p \\ \hline & & \ddots & \cdots \\ \hline & & & T_p \end{array}\right). \end{aligned}$$

Por tanto,

$$\mathrm{sp}(A) = \mathrm{sp}(U^* A U) = \bigcup_{i=1}^{p} \mathrm{sp}(T_i) = \bigcup_{i=1}^{p} \mathrm{sp}(A_{ii})$$

y

$$\begin{aligned} \det(A) = \det(U^* A U) &= \left(\prod_{i=1}^{n_1} t_{ii}^1\right)\left(\prod_{i=1}^{n_2} t_{ii}^2\right) \cdots \left(\prod_{i=1}^{n_p} t_{ii}^p\right) \\ &= \det(T_1)\det(T_2)\cdots\det(T_p) = \det(A_{11})\det(A_{22})\cdots\det(A_{pp}). \quad \square \end{aligned}$$

2.16. Demostrar que si $A \in \mathcal{M}_n$ es una matriz hermítica definida positiva y se descompone en bloques, los bloques diagonales son matrices hermíticas y definidas positivas. En particular, deducir que los elementos diagonales de A son números positivos, así como sus menores principales.

SOLUCIÓN. Supongamos que A está descompuesta en bloques de la forma

$$A = \left(\begin{array}{c|c|c|c} A_{11} & A_{12} & \cdots & A_{1p} \\ \hline A_{21} & A_{22} & \cdots & A_{2p} \\ \hline \cdots & \cdots & \cdots & \cdots \\ \hline A_{p1} & A_{p2} & \cdots & A_{pp} \end{array}\right)$$

donde $A_{ii} \in \mathcal{M}_{n_i}$ para $i = 1, 2, \ldots, p$, siendo $n_1 + n_2 + \cdots + n_p = n$. Como

$$A^* = \left(\begin{array}{c|c|c|c} A_{11}^* & A_{21}^* & \cdots & A_{p1}^* \\ \hline A_{12}^* & A_{22}^* & \cdots & A_{p2}^* \\ \hline \cdots & \cdots & \cdots & \cdots \\ \hline A_{1p}^* & A_{2p}^* & \cdots & A_{pp}^* \end{array}\right)$$

el hecho de que A sea hermítica hace que

$$A_{ii}^* = A_{ii}$$

para $i = 1, 2, \ldots, p$, por lo que las matrices A_{ii} también son hermíticas. Por tanto, para cada $i \in \{1, 2, \ldots, p\}$, la matriz A_{ii} es definida positiva si se verifica que

$$w_i^* A_{ii} w_i > 0,\ w_i \in \mathbf{V}_{n_i} \backslash \{0\}.$$

Ahora bien, a partir de un vector $w_i \in \mathbf{V}_{n_i} \backslash \{0\}$ basta considerar el vector "ampliado" $v = (0, \ldots, w_i, 0, \ldots, 0)^{\mathrm{T}}$ para el que se cumple

$$w_i^* A_{ii} w_i = v^* A v > 0$$

por ser $v \in \mathbf{V} \backslash \{0\}$ y A un matriz hermítica definida positiva.

Nótese que, en particular, si hacemos la siguiente descomposición de la matriz hermítica y definida positiva $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$

$$A = \left(\begin{array}{c|c|c|c} a_{11} & a_{12} & \cdots & a_{1n} \\ \hline a_{21} & a_{22} & \cdots & a_{2n} \\ \hline \cdots & \cdots & \cdots & \cdots \\ \hline a_{n1} & a_{n2} & \cdots & a_{nn} \end{array}\right)$$

se verifica que $a_{ii} > 0$ para $i = 1, 2, \ldots, n$. Por otra parte, por ser cada caja principal hermítica y definida positiva sus autovalores son positivos y, por tanto, su determinante también. □

2.17. Sea $A \in \mathcal{M}_n$ una matriz hermítica con espectro $\mathrm{sp}(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$.

a) Demostrar que para todo $\lambda \in \mathbb{R}$ y $v \in \mathbf{V}\backslash\{0\}$ se verifica que

$$\min_{1\leq j\leq n} |\lambda - \lambda_j| \leq \frac{||Av - \lambda v||_2}{||v||_2}.$$

b) Estudiar cómo se puede aplicar el resultado anterior para obtener aproximaciones de los autovalores de la matriz A.

SOLUCIÓN.

a) Por ser A una matriz hermítica existe una base ortonormal de autovectores $\mathcal{B} = \{u_1, u_2, \ldots, u_n\}$ del espacio $\mathbf{V}$ (véase el problema 2.23), por lo que podemos escribir el vector $v \in \mathbf{V}\backslash\{0\}$ en términos de la base $\mathcal{B}$ como

$$v = \sum_{j=1}^{n} \alpha_j u_j.$$

Por tanto,

$$Av = \sum_{j=1}^{n} \alpha_j A u_j = \sum_{j=1}^{n} \alpha_j \lambda_j u_j$$

y, por ser $\mathcal{B}$ una base ortonormal,

$$\begin{aligned} ||Av - \lambda v||_2^2 &= \left|\left| \sum_{j=1}^{n} \alpha_j \lambda_j u_j - \lambda \sum_{j=1}^{n} \alpha_j u_j \right|\right|_2^2 \\ &= \left|\left| \sum_{j=1}^{n} (\lambda_j - \lambda)\alpha_j u_j \right|\right|_2^2 = \sum_{j=1}^{n} |\alpha_j|^2 |\lambda_j - \lambda|^2 \, ||u_j||_2^2 \\ &\geq \min_{1\leq j\leq n} |\lambda - \lambda_j|^2 \sum_{j=1}^{n} |\alpha_j|^2 \, ||u_j||_2^2 \\ &= \min_{1\leq j\leq n} |\lambda - \lambda_j|^2 \, ||v||_2^2 \end{aligned}$$

de donde se sigue el resultado.

b) Dados $\lambda \in \mathbb{R}$ y $v \in \mathbf{V}\backslash\{0\}$, si denotamos por

$$r = \frac{||Av - \lambda v||_2}{||v||_2}$$

el resultado anterior muestra que existe un autovalor λ_j de A que dista de λ una cantidad inferior a r. De esta forma, en el caso de que $r \ll 1$ se puede tomar λ como una aproximación de λ_j. $\square$

2.18. Dados $a, b, c \in \mathbb{R}$ con $ac > 0$ se considera la matriz tridiagonal

$$A = \begin{pmatrix} b & c & & & \\ a & b & c & & \\ & \ddots & \ddots & \ddots & \\ & & a & b & c \\ & & & a & b \end{pmatrix} \in \mathcal{M}_n.$$

Comprobar que los autovalores de A son

$$\lambda_j = b + 2\,\mathrm{sign}(c)\sqrt{ac}\cos\left(\frac{j\pi}{n+1}\right)$$

para $j = 1, 2, \ldots, n$, con autovectores asociados $\{v^j\}_{j=1}^n$ de componentes

$$\left(v^j\right)_k = \left(\frac{a}{c}\right)^{\frac{k-1}{2}} \operatorname{sen}\left(\frac{j\pi k}{n+1}\right)$$

para $k = 1, 2, \ldots, n$.

Solución. Basta observar que para cada $j \in \{1, 2, \ldots, n\}$ se verifica que

$$\begin{aligned} \left(Av^j\right)_k &= a\left(v^j\right)_{k-1} + b\left(v^j\right)_k + c\left(v^j\right)_{k+1} \\ &= a\left(\frac{a}{c}\right)^{\frac{k-2}{2}} \operatorname{sen}\left(\frac{j\pi(k-1)}{n+1}\right) + b\left(\frac{a}{c}\right)^{\frac{k-1}{2}} \operatorname{sen}\left(\frac{j\pi k}{n+1}\right) \\ &\quad + c\left(\frac{a}{c}\right)^{\frac{k}{2}} \operatorname{sen}\left(\frac{j\pi(k+1)}{n+1}\right) \\ &= b\left(\frac{a}{c}\right)^{\frac{k-1}{2}} \operatorname{sen}\left(\frac{j\pi k}{n+1}\right) \\ &\quad + c\left(\frac{a}{c}\right)^{\frac{k}{2}} \left(\operatorname{sen}\left(\frac{j\pi(k-1)}{n+1}\right) + \operatorname{sen}\left(\frac{j\pi(k+1)}{n+1}\right)\right) \\ &= b\left(\frac{a}{c}\right)^{\frac{k-1}{2}} \operatorname{sen}\left(\frac{j\pi k}{n+1}\right) + 2c\left(\frac{a}{c}\right)^{\frac{k}{2}} \operatorname{sen}\left(\frac{j\pi k}{n+1}\right)\cos\left(\frac{j\pi}{n+1}\right). \end{aligned}$$

Por tanto,

$$\begin{aligned}\left(Av^j\right)_k &= b\left(\frac{a}{c}\right)^{\frac{k-1}{2}}\operatorname{sen}\left(\frac{j\pi k}{n+1}\right)\\ &\quad + 2\operatorname{sign}(c)\sqrt{ac}\left(\frac{a}{c}\right)^{\frac{k-1}{2}}\operatorname{sen}\left(\frac{j\pi k}{n+1}\right)\cos\left(\frac{j\pi}{n+1}\right)\\ &= \left(b + 2\operatorname{sign}(c)\sqrt{ac}\cos\left(\frac{j\pi}{n+1}\right)\right)\left(\frac{a}{c}\right)^{\frac{k-1}{2}}\operatorname{sen}\left(\frac{j\pi k}{n+1}\right)\\ &= \lambda_j\left(v^j\right)_k. \quad \square\end{aligned}$$

2.5.2. Problemas propuestos

2.19. Si $v, w \in \mathbf{V}$, determinar la forma de vw^*.

2.20. Producto de matrices por bloques.

a) Sean $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ y $N = \begin{pmatrix} E & F \\ G & H \end{pmatrix}$ particiones por bloques coherentes de las matrices M y N (es decir, $A, E \in \mathcal{M}_{n_1}$ y $D, H \in \mathcal{M}_{n_2}$ con $n_1 + n_2 = n$). Demostrar que

$$MN = \begin{pmatrix} AE + BG & AF + BH \\ CE + DG & CF + DH \end{pmatrix}.$$

b) Generalizar el resultado anterior para el producto de dos matrices cuadradas descompuestas en p bloques.

2.21. Sean D y E matrices diagonales y λ un escalar. Demostrar que λD, $D + E$ y DE son matrices diagonales y determinarlas.

2.22. Una matriz de la forma

$$P^{ij} = \begin{pmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & 0 & \cdots & \cdots & \cdots & 1 & & & \\ & & & \vdots & 1 & & & \vdots & & & \\ & & & \vdots & & \ddots & & \vdots & & & \\ & & & \vdots & & & 1 & \vdots & & & \\ & & & 1 & \cdots & \cdots & \cdots & 0 & & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ i) \\ \\ \\ \\ j) \\ \\ \\ \\ \end{matrix}$$

se denomina *matriz de permutación* de las líneas i y j. Si $B = (b_{kl})_{k,l=1}^{n}$ comprobar que

$$(P^{ij}B)_{kl} = \begin{cases} b_{kl} & \text{si} \quad k \neq i, j, \quad l = 1, 2, \ldots, n \\ b_{jl} & \text{si} \quad k = i, \quad\quad l = 1, 2, \ldots, n \\ b_{il} & \text{si} \quad k = j, \quad\quad l = 1, 2, \ldots, n \end{cases}$$

y

$$(BP^{ij})_{kl} = \begin{cases} b_{kl} & \text{si} \quad l \neq i, j, \quad k = 1, 2, \ldots, n \\ b_{kj} & \text{si} \quad l = i, \quad\quad k = 1, 2, \ldots, n \\ b_{ki} & \text{si} \quad l = j, \quad\quad k = 1, 2, \ldots, n \end{cases}$$

es decir, al multiplicar la matriz B a la izquierda (respectivamente, derecha) por P^{ij} se intercambian las filas (respectivamente, columnas) i y j de B. Además, probar que

$$\det(P^{ij}) = \begin{cases} 1 & \text{si} \quad i = j \\ -1 & \text{si} \quad i \neq j \end{cases} \quad \text{y} \quad (P^{ij})^{-1} = P^{ij}.$$

2.23. Demostrar que si $A \in \mathcal{M}_n$ es normal entonces existe una base ortonormal formada por autovectores de A, es decir, existe $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ base de $\mathbf{V}$ tal que

$$Av_i = \lambda_i v_i$$

para $i = 1, 2, \ldots, n$, siendo $\text{sp}(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ y

$$v_i^* v_j = \delta_{ij}$$

para $i, j = 1, 2, \ldots, n$.

2.24. Probar que si $A \in \mathcal{M}_n$ es normal e inversible entonces A^{-1} es también normal.

2.25. Si $A \in \mathcal{M}_n$ es inversible demostrar que

$$\lambda \in \text{sp}(A) \Leftrightarrow \frac{1}{\lambda} \in \text{sp}(A^{-1})$$

y

$$\varrho(A^{-1}) = \frac{1}{\min\limits_{1 \leq i \leq n} \{|\lambda_i(A)| : \lambda_i(A) \in \text{sp}(A)\}}.$$

2.26. Generalizar los resultados los problemas 2.8 y 2.9 para matrices diagonales y triangulares, superiores e inferiores, por bloques.

2.27. Sea $A = (a_{ij})_{i,j=1}^{n} \in \mathcal{M}_n$ una matriz simétrica y definida positiva. Demostrar las siguientes propiedades:

a) $|a_{ij}| \leq \dfrac{a_{ii} + a_{jj}}{2}$, $i, j = 1, 2, \ldots, n$.

b) $|a_{ij}| \leq \sqrt{a_{ii} a_{jj}}$, $i, j = 1, 2, \ldots, n$.

c) $\max\limits_{1 \leq i,j \leq n} |a_{ij}| = \max\limits_{1 \leq i \leq n} a_{ii}$.

2.28. Demostrar que si existen un número $r > 0$ y una norma matricial $\|\cdot\|$ tales que $A = rI + B$ con $\|B\| < r$ entonces A es inversible y

$$\left\|A^{-1}\right\| \leq \frac{\|I\|}{r - \|B\|}.$$

2.6. Prácticas

2.1. Comprobar el resultado del problema 2.20 apartado *a)* para descomposiciones coherentes por bloques de las matrices M y N.

2.2. Utilizar los comandos `det` y `eig` de `MATLAB` para comprobar, con elecciones arbitrarias de matrices $A, B \in \mathcal{M}_n$, las siguientes propiedades:

a) $\det(AB) = \det(BA) = \det(A)\det(B)$

b) $\det(\lambda A) = \lambda^n \det(A)$

c) $\det(A^*) = \overline{\det(A)}$

d) $\det(A) = \prod\limits_{i=1}^{n} \lambda_i(A)$ siendo $\mathrm{sp}(A) = \{\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A)\}$.

2.3. Escribir un programa que calcule las normas uno, infinito y Fröbenius de una matriz dada. Comparar los resultados con los obtenidos con el comando `norm` de `MATLAB`.

2.4. Escribir un programa específico para el producto de una matriz triangular superior (respectivamente, inferior) por un vector y el producto de dos matrices triangulares superiores (respectivamente, inferiores).

2.5. Escribir un programa que calcule las potencias sucesivas de una matriz A, verificando previamente si $\|A\|_1$, $\|A\|_\infty$ o $\|A\|_{\mathrm{F}}$ es menor que uno.

3 Condicionamiento de un sistema lineal

3.1. Introducción

Nos disponemos a estudiar, en los próximos capítulos, diversos métodos de resolución de sistemas lineales. Antes de ello, vamos a llevar a cabo un breve análisis del condicionamiento de este tipo de problemas. En el capítulo 1, el ejemplo 1.9 de R. S. Wilson nos mostraba la existencia de sistemas lineales muy mal condicionados, lo cual nos invita a hacer un estudio riguroso del tema. Como se argumentó en dicho capítulo, no es en general sencillo definir números de condición que nos indiquen si el problema que se pretende resolver en cada caso está bien o mal condicionado. Sin embargo, para los sistemas lineales, la teoría será fácil y el número de condición estará ligado a la matriz del sistema.

3.2. Condicionamiento de una matriz y de un sistema lineal

Veamos cómo definir el condicionamiento de un sistema lineal $Au = b$ siendo $A \in \mathcal{M}_n$ una matriz inversible y $b \in \mathbf{V}$ un vector no nulo. En el supuesto de que se tome como segundo miembro, en lugar del vector b, una perturbación de éste, $b + \delta b$, denotando por $u + \delta u$ la solución del problema perturbado, se verifica que

$$A(u + \delta u) = b + \delta b \;\Rightarrow\; A\delta u = \delta b \;\Rightarrow\; \delta u = A^{-1}\delta b,$$

luego, a partir de la norma matricial $|||\cdot|||$ subordinada a $||\cdot||$, se tiene

$$||\delta u|| \leq |||A^{-1}||| \; ||\delta b|| \,;$$

como, por otra parte,

$$Au = b \;\Rightarrow\; ||b|| \leq |||A||| \; ||u|| \;\Rightarrow\; \frac{1}{||u||} \leq \frac{|||A|||}{||b||},$$

se tiene

$$\overbrace{\frac{||\delta u||}{||u||}}^{\text{Error relativo (resultados)}} \leq |||A||| \; |||A^{-1}||| \; \overbrace{\frac{||\delta b||}{||b||}}^{\text{Error relativo (datos)}}$$

Parece claro, pues, que la cantidad $|||A||| \; |||A^{-1}|||$ servirá como número de condición para resolver un sistema lineal $Au = b$. De hecho, se tiene la siguiente definición:

Definición 3.1. Sea $|||\cdot|||$ una norma matricial subordinada y $A \in \mathcal{M}_n$ una matriz inversible. El número

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

se denomina *condicionamiento* (o *número de condición*) de la matriz A relativo a la norma $|||\cdot|||$. □

Notación 3.1. En general, siempre que escribamos $\text{cond}(A)$ nos estaremos refiriendo al condicionamiento de una matriz por una norma subordinada $|||\cdot|||$. En el caso particular en que tomemos la norma $|||\cdot|||_p$, $1 \leq p \leq +\infty$, escribiremos

$$\text{cond}_p(A) = |||A|||_p \; |||A^{-1}|||_p . \quad \square$$

Teorema 3.1. *Sea $|||\cdot|||$ la norma matricial subordinada a una norma vectorial $||\cdot||$, $A \in \mathcal{M}_n$ una matriz inversible y $b \in \mathbf{V}\backslash\{0\}$. Si u y $u + \delta u$ son las soluciones respectivas de los sistemas lineales*

$$Au = b \;\; y \;\; A(u + \delta u) = b + \delta b$$

entonces se verifica que

$$\frac{||\delta u||}{||u||} \leq \text{cond}(A) \frac{||\delta b||}{||b||}. \tag{3.1}$$

Además, $\text{cond}(A)$ *es el número más pequeño que verifica la desigualdad anterior, es decir, para cada matriz A inversible existen vectores b, $\delta b \in \mathbf{V}\backslash\{0\}$ tales que*

$$\frac{||\delta u||}{||u||} = \text{cond}(A) \frac{||\delta b||}{||b||}.$$

Demostración. La propiedad (3.1) ya se ha demostrado previamente. Para ver la optimalidad, por la proposición 2.12 existe $u \in \mathbf{V}\backslash\{0\}$ tal que

$$||Au|| = |||A||| \; ||u|| .$$

A partir de este vector u, definimos

$$b = Au.$$

Por otro lado, aplicando nuevamente la proposición 2.12, existe $\delta b \in \mathbf{V}$ tal que

$$\left|\left|A^{-1}\delta b\right|\right| = \left|\left|\left|A^{-1}\right|\right|\right| \, ||\delta b|| \, .$$

Así pues, considerando los sistemas lineales

$$Au = b \;\text{ y }\; A(u+\delta u) = b + \delta b,$$

tendremos, como antes, que

$$A\delta u = \delta b$$

y así

$$\delta u = A^{-1}\delta b,$$

con lo que

$$||\delta u|| = \left|\left|A^{-1}\delta b\right|\right| = \left|\left|\left|A^{-1}\right|\right|\right| \, ||\delta b|| \;\text{ y }\; ||b|| = ||Au|| = |||A||| \, ||u|| \, .$$

Por tanto,

$$\frac{||\delta u||}{||u||} = \left|\left|\left|A^{-1}\right|\right|\right| \, |||A||| \, \frac{||\delta b||}{||b||} = \mathrm{cond}(A)\frac{||\delta b||}{||b||}. \quad \square$$

Cuando se consideran perturbaciones de la matriz A en lugar de perturbaciones del vector b, el resultado que se obtiene no es tan nítido, pero el número $\mathrm{cond}(A)$ sigue siendo una buena herramienta para medir el condicionamiento del problema. En concreto, se tiene el siguiente resultado:

Teorema 3.2. *Sea $|||\cdot|||$ la norma matricial subordinada a una norma vectorial $||\cdot||$, $A \in \mathcal{M}_n$ una matriz inversible y $b \in \mathbf{V}\backslash\{0\}$. Si u y $u+\Delta u$ son las soluciones respectivas de los sistemas lineales*

$$Au = b \;\; y \;\; (A+\Delta A)(u+\Delta u) = b,$$

se verifica que

$$\frac{||\Delta u||}{||u+\Delta u||} \leq \mathrm{cond}(A)\frac{|||\Delta A|||}{|||A|||}$$

siendo $\mathrm{cond}(A)$ *el número más pequeño que verifica la desigualdad anterior, es decir, para toda matriz A inversible existen $b \in \mathbf{V}\backslash\{0\}$ y $\Delta A \in \mathcal{M}_n\backslash\{0\}$ tales que*

$$\frac{||\Delta u||}{||u+\Delta u||} = \mathrm{cond}(A)\frac{|||\Delta A|||}{|||A|||}.$$

Además,

$$\frac{||\Delta u||}{||u||} \leq \mathrm{cond}(A)\frac{|||\Delta A|||}{|||A|||}\left(1 + O(|||\Delta A|||)\right).$$

DEMOSTRACIÓN. Véase [**Ci**]. □

Recogemos ahora algunas propiedades de demostración inmediata que verifica el condicionamiento de una matriz.

Proposición 3.1. *Sea $|||\cdot|||$ una norma matricial subordinada y $A \in \mathcal{M}_n$ una matriz inversible. Se verifican las siguientes propiedades:*

a) $\mathrm{cond}(A) \geq 1$.

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

c) $\mathrm{cond}(\lambda A) = \mathrm{cond}(A)$ *para todo* $\lambda \in \mathbb{K}\backslash\{0\}$.

DEMOSTRACIÓN. Por ser $|||\cdot|||$ una norma matricial subordinada, se verifica

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

Por otra parte,

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

y para todo $\lambda \neq 0$ se tiene que

$$\mathrm{cond}(\lambda A) = |||\lambda A||| \ \left|\left|\left|(\lambda A)^{-1}\right|\right|\right| = |\lambda| \, |\lambda^{-1}| \ |||A||| \ \left|\left|\left|A^{-1}\right|\right|\right| = \mathrm{cond}(A). \quad \square$$

En el caso particular de que consideremos como norma matricial subordinada $|||\cdot|||_2$ se tiene el siguiente resultado:

Proposición 3.2. *Sea $A \in \mathcal{M}_n$ una matriz inversible. Se verifica que*

$$\mathrm{cond}_2(A) = +\sqrt{\frac{\lambda_n(A^*A)}{\lambda_1(A^*A)}}$$

donde $\lambda_1(A^*A)$ y $\lambda_n(A^*A)$ son, respectivamente, el menor y el mayor de los autovalores de la matriz A^*A.

DEMOSTRACIÓN. En primer lugar tengamos en cuenta que la matriz A^*A es hermítica y definida positiva por ser A una matriz inversible (véase la proposición 2.8), por lo que todos los autovalores de A^*A son positivos. Por otra parte, aplicando el teorema 2.3 se verifica que

$$|||A|||_2^2 = \varrho(A^*A) = \lambda_n(A^*A)$$

y

$$\left|\left|\left|A^{-1}\right|\right|\right|_2^2 = \varrho((A^{-1})^*A^{-1}) = \varrho(A^{-1}(A^{-1})^*) = \varrho((A^*A)^{-1}) = \frac{1}{\lambda_1(A^*A)}$$

(véase el problema 2.25). □

Observación 3.1 A partir del teorema 2.3 y de los problemas 2.24 y 2.25, se verifica que:

1. Si A es normal e inversible y $\mathrm{sp}(A) = \{\lambda_1(A), \lambda_2(A), \ldots, \lambda_n(A)\}$, entonces

$$\mathrm{cond}_2(A) = \|A\|_2 \, \left\|A^{-1}\right\|_2 = \varrho(A)\varrho(A^{-1}) = \frac{\varrho(A)}{\mu(A)} \tag{3.2}$$

   siendo

$$\mu(A) = \min_{1\leq i\leq n} |\lambda_i(A)|.$$

   Consecuentemente, para este tipo de matrices, se verifica que

$$\mathrm{cond}(A) = \|A\| \, \left\|A^{-1}\right\| \geq \varrho(A)\varrho(A^{-1}) = \mathrm{cond}_2(A) \tag{3.3}$$

   para cualquier norma matricial subordinada $\|\cdot\|$ (véanse el teorema 2.4 y (3.2)). Es decir, para matrices normales el condicionamiento cond_2 es el menor de todos.

2. En el caso particular de que A sea unitaria entonces $\mathrm{cond}_2(A) = 1$ (véase la proposición 2.5).

3. La invarianza por transformaciones unitarias de $\|\cdot\|_2$ hace que $\mathrm{cond}_2(A)$ sea invariante por transformaciones unitarias, es decir,

$$\mathrm{cond}_2(A) = \mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(U^*AU)$$

   si $UU^* = I$. □

Hagamos unas consideraciones finales respecto al problema que nos ocupa en este capítulo.

1. Como hemos visto en la proposición 3.1, se verifica que el condicionamiento de una matriz es siempre un número mayor o igual que 1. Por tanto, el sistema lineal $Au = b$ estará tanto mejor condicionado cuanto más próximo a 1 esté $\mathrm{cond}(A)$.

2. En el caso de que A sea una matriz unitaria, el sistema $Au = b$ siempre está bien condicionado para $\|\cdot\|_2$, ya que $\mathrm{cond}_2(A) = 1$; además, las transformaciones unitarias conservan el $\mathrm{cond}_2(A)$.

3. Cuando se necesita resolver un sistema lineal $Au = b$ siendo A una matriz con un número de condición elevado, se hace necesario utilizar un *precondicionador*. La idea básica es sencilla: tomar una matriz inversible M de forma que la matriz $\widetilde{A} = MA$ tenga un condicionamiento pequeño; después, bastará

resolver el sistema $\widetilde{A}u = \widetilde{b}$ siendo $\widetilde{b} = Mb$. Sin embargo, lo que no es sencillo es, precisamente, encontrar esta matriz M. Una posible elección, de fácil cálculo y a veces suficiente, es considerar $M = D^{-1}$ siendo $D = \mathrm{diag}\,(A)$. La idea aquí expuesta es la de un precondicionador por la izquierda. También se suelen tomar precondicionadores:

a) Por la derecha ($\widetilde{A} = AM$, $\widetilde{A}\widetilde{u} = b$, $u = M\widetilde{u}$).

b) Por ambos lados ($M = C^2$, $\widetilde{A} = CAC$, $\widetilde{b} = Cb$, $\widetilde{A}\widetilde{u} = \widetilde{b}$, $u = C\widetilde{u}$).

c) Simétricos ($M = CC^{\mathrm{T}}$, $\widetilde{A} = CAC^{\mathrm{T}}$, $\widetilde{b} = Cb$, $\widetilde{A}\widetilde{u} = \widetilde{b}$, $u = C^{\mathrm{T}}\widetilde{u}$).

lo que puede dar una idea de lo sofisticado de estas técnicas.

Analicemos ahora, con más detalle, el ejemplo 1.9.

Ejemplo 3.1 (R. S. Wilson). Consideremos

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \; b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \;\text{ y }\; \delta b = \begin{pmatrix} 0.1 \\ -0.1 \\ 0.1 \\ -0.1 \end{pmatrix}.$$

La solución exacta del sistema lineal $Au = b$ es

$$u = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

mientras que la solución del sistema $A(u + \delta u) = b + \delta b$ es

$$u + \delta u = \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix} \;\Rightarrow\; \delta u = \begin{pmatrix} 8.2 \\ -13.6 \\ 3.5 \\ -2.1 \end{pmatrix}.$$

El polinomio característico de A viene dado por

$$P(\lambda) = \det(A - \lambda I) = \lambda^4 - 35\lambda^3 + 146\lambda^2 - 100\lambda + 1$$

y tiene como raíces aproximadas los números

$$\lambda_1 \simeq 0.0101500484363,\; \lambda_2 \simeq 0.843107149904$$

$$\lambda_3 \simeq 3.85805745587 \;\text{ y }\; \lambda_4 \simeq 30.2886853457.$$

De esta forma, por ser A simétrica, la relación (3.2) determina

$$\mathrm{cond}_2(A) = \frac{\lambda_4}{\lambda_1} \simeq 2984.09269037.$$

Por tanto, no es de extrañar el mal comportamiento que, tras las perturbaciones en los datos, se observó anteriormente. □

3.3. Problemas

3.3.1. Problemas resueltos

3.1. Sea $A \in \mathcal{M}_n$ una matriz hermítica (respectivamente simétrica) definida positiva. Demostrar los siguientes resultados:

a) Existe $B \in \mathcal{M}_n$ hermítica (respectivamente simétrica) definida positiva tal que $A = B^2$. La matriz B se denomina la *raíz cuadrada* de la matriz A.

b) Si $\mathrm{cond}_2(A) > 1$ entonces $\mathrm{cond}_2(B) < \mathrm{cond}_2(A)$.

Solución.

a) Por ser A hermítica es diagonalizable (véase la observación 2.10), es decir, existe U unitaria tal que U^*AU es una matriz diagonal de la forma

$$U^*AU = D = \mathrm{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

donde

$$\mathrm{sp}(A) = \mathrm{sp}(D) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}. \tag{3.4}$$

Por otra parte, por ser A hermítica definida positiva entonces

$$\lambda_i > 0$$

para $i = 1, 2, \ldots, n$ (véase la proposición 2.7). Consecuentemente, a partir de la matriz diagonal

$$\Delta = \mathrm{diag}\,\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right)$$

se verifica que

$$U^*AU = D = \Delta^2,$$

de donde

$$A = U\Delta^2U^* = (U\Delta U^*)(U\Delta U^*) = B^2$$

considerando

$$B = U\Delta U^*.$$

Claramente la matriz B es hermítica y definida positiva, ya que

$$\mathrm{sp}(B) = \mathrm{sp}(\Delta) = \left\{\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right\} \subset \mathbb{R}_+. \tag{3.5}$$

b) Por la observación 3.1 sabemos que si $C \in \mathcal{M}_n$ es una matriz normal y

$$\{\lambda_1(C), \lambda_2(C), \ldots, \lambda_n(C)\}$$

son los autovalores de C, entonces

$$\mathrm{cond}_2(C) = \frac{\varrho(C)}{\mu(C)}$$

donde

$$\mu(C) = \min_{1\le i\le n} |\lambda_i(C)|.$$

Aplicando este resultado a las matrices A y B, que son hermíticas y, por tanto, normales, se verifica que

$$\mathrm{cond}_2(B) = \frac{\varrho(B)}{\mu(B)} = \sqrt{\frac{\varrho(A)}{\mu(A)}} = \sqrt{\mathrm{cond}_2(A)}$$

donde hemos tenido en cuenta las relaciones (3.4) y (3.5). Finalmente, como

$$\mathrm{cond}_2(A) > 1,$$

de la relación anterior se sigue que

$$\mathrm{cond}_2(B) = \sqrt{\mathrm{cond}_2(A)} < \mathrm{cond}_2(A). \quad \square$$

3.3.2. Problemas propuestos

3.2. Sea $A \in \mathcal{M}_n$ una matriz inversible. Demostrar las siguientes desigualdades:

a) $\dfrac{1}{n}\,\mathrm{cond}_2(A) \le \mathrm{cond}_1(A) \le n\,\mathrm{cond}_2(A)$.

b) $\dfrac{1}{n}\,\mathrm{cond}_\infty(A) \le \mathrm{cond}_2(A) \le n\,\mathrm{cond}_\infty(A)$.

c) $\dfrac{1}{n^2}\,\mathrm{cond}_1(A) \le \mathrm{cond}_\infty(A) \le n^2\,\mathrm{cond}_1(A)$.

3.4. Prácticas

3.1. Utilizar el comando `eig` de `MATLAB` para calcular $\mathrm{cond}_2(A)$ siendo A la matriz de Wilson del ejemplo 3.1. Calcular también, usando el comando `cond` de `MATLAB`, los condicionamientos de dicha matriz respecto a las normas $\|\cdot\|_1$, $\|\cdot\|_\infty$ y $\|\cdot\|_{\mathrm{F}}$; comprobar que los tres son mayores que $\mathrm{cond}_2(A)$.

3.2. La *matriz de Hilbert* de orden $n \in \mathbb{N}$ viene definida como

$$H_n = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{pmatrix},$$

es decir, $H_n = (h_{ij})_{i,j=1}^n$ siendo

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

para $i, j = 1, 2, \ldots, n$. Esta matriz viene implementada en `MATLAB` con el comando `hilb(n)`.

a) Utilizar el comando `cond` de `MATLAB` para determinar el $\mathrm{cond}_2(H_n)$ para valores de $n \in \{1, 2, \ldots, 14\}$ y observar cómo éste crece al aumentar n. Concretamente, comprobar que se obtienen los valores dados en la tabla 3.1.

b) Dado $n \in \{1, 2, \ldots, 14\}$ consideramos el sistema lineal

$$H_n u_n = b_n,$$

siendo $b_n \in \mathbf{V}$ el vector de coordenadas

$$(b_n)_i = \sum_{j=1}^{n} \frac{1}{i+j-1}$$

para $i = 1, 2, \ldots, n$, cuya solución exacta es

$$u_n = (1, 1, \ldots, 1)^{\mathrm{T}}.$$

Utilizar el comando `\` de `MATLAB` para comparar las soluciones que se van obteniendo, a medida que n aumenta, con las respectivas soluciones exactas.

TABLA 3.1:
Condicionamiento de la matriz de Hilbert de orden n

n	$\text{cond}_2(H_n)$	n	$\text{cond}_2(H_n)$
1	1	8	1.525758×10^{10}
2	1.928147×10^{1}	9	4.931532×10^{11}
3	5.240568×10^{2}	10	1.602534×10^{13}
4	1.551374×10^{4}	11	5.218389×10^{14}
5	4.766073×10^{5}	12	1.768065×10^{16}
6	1.495106×10^{7}	13	3.682278×10^{18}
7	4.753674×10^{8}	14	1.557018×10^{18}

3.3. Se considera la *matriz de Vandermonde*

$$V_n = \begin{pmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \dots & \dots & \dots & \dots \\ 1 & x_n & \dots & x_n^n \end{pmatrix}$$

donde $x_i \neq x_j$ si $i \neq j$. Calcular el $\text{cond}_2(V_n)$ para diversas elecciones de los puntos $\{x_0, x_1, \dots, x_n\}$.

4 Resolución de sistemas lineales: métodos directos

4.1. Introducción

Ya nos hemos referido previamente al énfasis especial que el Análisis Numérico hace en desarrollar técnicas pensadas para resolver problemas prácticos. Son muchas las aplicaciones (determinación de tensiones en los nudos de una red de corriente continua, cálculo de estructuras reticuladas definidas por vigas, modelos económicos del tipo *input–output*...) cuya resolución práctica pasa por plantear y resolver un sistema de ecuaciones lineales; más aún, la práctica totalidad de los problemas que surgen en las ciencias aplicadas se resuelven mediante algoritmos que involucran, en alguna de sus etapas, la resolución de un sistema lineal (desde el diseño asistido por ordenador, en el que se requiere conocer la intensidad de cada color básico en cada *pixel* de la pantalla, hasta la predicción meteorológica, donde se resuelven complicadas ecuaciones en derivadas parciales mediante los *métodos de elementos* y *diferencias finitas*).

Desde el punto de vista teórico, los sistemas lineales no plantean ninguna dificultad; el *teorema de Rouché–Fröbenius* y la *regla de Cramer* agotan el estudio del problema. Sin embargo, en la práctica, los sistemas lineales involucrados en la resolución de los problemas están asociados a matrices de órdenes grandes (piénsese en sistemas de n ecuaciones e incógnitas con n tomando valores como 10, 1000 o 100000). Para estos tamaños, la *regla de Cramer* es inviable por el elevado número de operaciones que exige. Ésta es la razón por la que se han diseñado otros métodos más eficaces para la resolución de sistemas lineales.

En este capítulo abordaremos el estudio de los *métodos directos* y trataremos los *métodos iterativos* en el capítulo siguiente. En los primeros se diseñan algoritmos que dan la solución exacta del problema (salvo, claro está, los errores debidos al redondeo) en contraposición a los métodos iterativos, que dan la solución como límite de una sucesión de vectores.

El problema que queremos resolver se plantea de la siguiente forma: "dada una matriz inversible $A \in \mathcal{M}_n$ y $b \in \mathbf{V}$, encontrar $u \in \mathbf{V}$ tal que $Au = b$". Para hacerlo, primero veremos cómo abordar el problema cuando la matriz A es particularmente "sencilla" (diagonal o triangular) y, después, expondremos el conocido *método de Gauss* y un par de variantes suyas (factorización *LU* y el *método de Cholesky*), menos generales pero más eficaces. Con estos métodos se transforma el sistema original en uno o varios sistemas triangulares y, por tanto, sencillos de resolver.

Observación 4.1. La solución del sistema planteado es $u = A^{-1}b$. No obstante, *jamás* se obtendrá de esta forma, calculando A^{-1} de la "forma habitual" y multiplicando A^{-1} por b (esto es menos efectivo, aún, que la *regla de Cramer*). De hecho, la estrategia es justamente la contraria: para calcular A^{-1} lo que haremos será resolver, con alguno de los métodos que expondremos, los n sistemas de ecuaciones

$$Au_i = \mathbf{e}_i$$

para $i = 1, 2, \ldots, n$, donde $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ es la base canónica de $\mathbf{V}$. La matriz $B = (u_1, u_2, \ldots, u_n)$ es la inversa de A, dado que, como se vio en el problema 2.3,

$$AB = (Au_1, Au_2, \ldots, Au_n) = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n) = I. \quad \square$$

4.2. Sistemas diagonales y triangulares

Si la matriz $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ es triangular (o, en particular, diagonal) e inversible entonces

$$\det(A) = \prod_{i=1}^{n} a_{ii} = a_{11}a_{22}\cdots a_{nn} \neq 0$$

(véase la proposición 2.3), lo que implica que

$$a_{ii} \neq 0$$

para $i = 1, 2, \ldots, n$ y, por tanto, podemos dividir por los elementos diagonales de la matriz A.

La solución del sistema $Au = b$ donde A es una matriz diagonal e inversible es

$$u_i = \frac{b_i}{a_{ii}}$$

para $i = 1, 2, \ldots, n$.

Para resolver $Au = b$ cuando A es una matriz triangular superior e inversible, es decir, el sistema

$$\left\{\begin{array}{ccccccccccc} a_{11}u_1 & + & a_{12}u_2 & + & \cdots & + & a_{1,n-1}u_{n-1} & + & a_{1n}u_n & = & b_1 \\ & & a_{22}u_2 & + & \cdots & + & a_{2,n-1}u_{n-1} & + & a_{2n}u_n & = & b_2 \\ & & & & \ddots & & \cdots & & \cdots & = & \cdots \\ & & & & & & a_{n-1,n-1}u_{n-1} & + & a_{n-1,n}u_n & = & b_{n-1} \\ & & & & & & & & a_{nn}u_n & = & b_n \end{array}\right.$$

basta considerar

$$\left\{\begin{array}{l} u_n = \dfrac{b_n}{a_{nn}} \\ \\ u_i = \dfrac{1}{a_{ii}}\left(b_i - \displaystyle\sum_{j=i+1}^{n} a_{ij}u_j\right), \quad i = n-1, n-2, \ldots, 1. \end{array}\right.$$

Esta técnica se conoce con el nombre de *método de remonte*: se despeja la última incógnita en la última ecuación y se sustituye en la anterior; se despeja en esta ecuación la penúltima incógnita y se sustituyen las dos incógnitas halladas en la ecuación anterior...

En el caso de que la matriz A sea triangular inferior e inversible, aplicando un *remonte* "hacia abajo" se obtiene que la solución del sistema

$$\left\{\begin{array}{ccccccccccc} a_{11}u_1 & & & & & & & & & = & b_1 \\ a_{21}u_1 & + & a_{22}u_2 & & & & & & & = & b_2 \\ \cdots & & \cdots & & \ddots & & & & & = & \cdots \\ a_{n-1,1}u_1 & + & a_{n-1,2}u_2 & + & \cdots & + & a_{n-1,n-1}u_{n-1} & & & = & b_{n-1} \\ a_{n1}u_1 & + & a_{n2}u_2 & + & \cdots & + & a_{n,n-1}u_{n-1} & + & a_{nn}u_n & = & b_n \end{array}\right.$$

es

$$\left\{\begin{array}{l} u_1 = \dfrac{b_1}{a_{11}} \\ \\ u_i = \dfrac{1}{a_{ii}}\left(b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}u_j\right), \quad i = 2, 3, \ldots, n. \end{array}\right.$$

Observación 4.2. Para matrices tridiagonales existen métodos específicos de resolución (véase el problema 4.2). □

4.3. Eliminación gaussiana

Suponemos al lector familiarizado con el método de Gauss de resolución de sistemas lineales. En cualquier caso, vamos a describirlo detalladamente con vistas a su formalización. Antes de comenzar con la descripción teórica, veamos un ejemplo.

4.3.1. Ejemplo para la formalización del método de Gauss

Vamos a aplicar el método de Gauss a la resolución del sistema lineal

$$Au = \begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 2 & 1 & 3 \\ 1 & 1 & -1 & 1 \\ 0 & 1 & 8 & 12 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 5 \\ 2 \end{pmatrix} = b.$$

Como el elemento $a_{11} = 0$, debemos intercambiar la primera fila con, por ejemplo, la segunda, obteniendo

$$\mathcal{A}_1 u = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 1 & 1 & -1 & 1 \\ 0 & 1 & 8 & 12 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 5 \\ 2 \end{pmatrix} = \beta_1,$$

donde $\mathcal{A}_1 = P_1 A$ y $\beta_1 = P_1 b$ siendo

$$P_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

El siguiente paso consiste en "hacer ceros" por debajo de la diagonal en la primera columna. Para ello, restamos a la tercera fila la primera, obteniendo

$$A_2 u = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & -1 & -2 & -2 \\ 0 & 1 & 8 & 12 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 5 \\ 2 \end{pmatrix} = b_2,$$

donde $A_2 = E_1 \mathcal{A}_1$ y $b_2 = E_1 \beta_1$, siendo

$$E_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Fijemos nuestra atención en el elemento $(A_2)_{22}$. Como es no nulo, no necesitamos permutar filas. Sin hacer ningún cambio escribimos

$$\mathcal{A}_2 u = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & -1 & -2 & -2 \\ 0 & 1 & 8 & 12 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 5 \\ 2 \end{pmatrix} = \beta_2,$$

donde $\mathcal{A}_2 = P_2 A_2$ y $\beta_2 = P_2 b_2$, siendo

$$P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Para "hacer ceros" por debajo de la diagonal de la segunda columna, a la tercera fila le sumamos la segunda y a la cuarta se la restamos. Así,

$$A_3 u = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 6 & 11 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 6 \\ 1 \end{pmatrix} = b_3,$$

donde $A_3 = E_2 \mathcal{A}_2$ y $b_3 = E_2 \beta_2$, siendo

$$E_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}.$$

Ahora, necesitamos permutar entre sí la tercera y cuarta filas. Haciéndolo, se obtiene

$$\mathcal{A}_3 u = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 6 & 11 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 6 \end{pmatrix} = \beta_3,$$

donde $\mathcal{A}_3 = P_3 A_3$ y $\beta_3 = P_3 b_3$ con

$$P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Aunque ya hemos llegado a un sistema triangular, en el caso general habría que hacer cero el elemento $(\mathcal{A}_3)_{43}$. En nuestro caso no hay que hacer nada respecto a la etapa anterior, quedando

$$A_4 u = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 6 & 11 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 6 \end{pmatrix} = b_4,$$

donde $A_4 = E_3 \mathcal{A}_3$ y $b_4 = E_3 \beta_3$ siendo

$$E_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Así pues, se ha conseguido el siguiente sistema triangular equivalente

$$(E_3 P_3 E_2 P_2 E_1 P_1 A) u = E_3 P_3 E_2 P_2 E_1 P_1 b,$$

cuya resolución puede llevarse a cabo mediante el método de remonte, obteniendo

$$(u_1, u_2, u_3, u_4)^{\mathrm{T}} = \left(\frac{75}{2}, -\frac{46}{3}, \frac{67}{6}, -6 \right)^{\mathrm{T}}.$$

Observación 4.3.

1. Las matrices $\{P_1, P_2, P_3\}$ son casos particulares de matrices de permutación. Una matriz de la forma

$$P^{ij} = \begin{pmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & 0 & \cdots & \cdots & \cdots & 1 & & \\ & & & \vdots & 1 & & & \vdots & & \\ & & & \vdots & & \ddots & & \vdots & & \\ & & & \vdots & & & 1 & \vdots & & \\ & & & 1 & \cdots & \cdots & \cdots & 0 & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ i) \\ \\ \\ \\ j) \\ \\ \\ \end{matrix}$$

es una matriz de permutación de las líneas i y j. Como ya se adelantó en el problema 2.22, dada una matriz B, la matriz $P^{ij}B$ es una matriz idéntica a B salvo que las filas i y j están permutadas entre sí. Además,

$$\det(P^{ij}) = \begin{cases} 1 & \text{si} \quad i = j \\ -1 & \text{si} \quad i \neq j \end{cases} \quad \text{y} \quad (P^{ij})^{-1} = P^{ij}. \tag{4.1}$$

2. Las matrices $\{E_1, E_2, E_3\}$ son casos particulares de matrices del tipo

$$E_k = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & \ell_{k+1,k} & 1 & & \\ & & & \vdots & & \ddots & \\ & & & \ell_{nk} & & & 1 \end{pmatrix}.$$

Estas matrices son inversibles (pues $\det(E_k) = 1$) y pueden escribirse como

$$E_k = I + \ell_k \mathbf{e}_k^{\mathrm{T}}$$

donde $\mathbf{e}_k$ es el k–ésimo vector de la base canónica y

$$\ell_k = (0, \overset{k)}{\cdots}, 0, \ell_{k+1,k}, \ell_{k+2,k}, \ldots, \ell_{nk})^{\mathrm{T}}.$$

Nótese que, en el caso particular de las matrices E_1, E_2 y E_3 del ejemplo anterior, el elemento ℓ_{ik}, $i = k+1, k+2, \ldots, n$, de cada E_k no es otra cosa que el número por el que se debe multiplicar, en el paso k, la fila k–ésima para que, al sumarla a la i–ésima, el elemento que ocupa la fila i y la columna k tome el valor cero. Estos números $\{\ell_{k+1,k}, \ell_{k+2,k}, \ldots, \ell_{nk}\}$ se denominan *multiplicadores*. □

4.3.2. Estudio general del método de Gauss

El método de Gauss para la resolución de un sistema lineal $Au = b$ donde $A \in \mathcal{M}_n$ y $b \in \mathbf{V}$ se basa en la transformación del sistema original en otro de la forma $MAu = Mb$ de manera que la matriz MA sea triangular superior. Una vez hecho esto se resuelve el sistema triangular $MAu = Mb$ por el método de remonte. Obviamente, la matriz $M \in \mathcal{M}_n$ debe ser inversible para que el sistema resuelto sea *equivalente* (en el sentido de que tenga la misma solución) al primero.

Observación 4.4. En la práctica, la matriz M *no se calcula* explícitamente, sino que directamente se obtienen la matriz MA y el vector Mb. □

El método de Gauss, en el caso general, se desarrolla como sigue: partimos de una matriz A cuadrada e inversible de la forma

$$A = (a_{ij})_{i,j=1}^{n} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

- Primera etapa de eliminación. Puesto que la matriz A es inversible, existe un índice $i \in \{1, 2, \ldots, n\}$ tal que $a_{i1} \neq 0$ (ya veremos, en la práctica, cómo se elige). Este elemento no nulo a_{i1} se denomina primer *pivote* de eliminación. Seguidamente permutamos la fila del pivote a_{i1} con la primera fila, lo que en escritura matricial equivale a multiplicar la matriz A, por la izquierda, por una matriz $P_1 = P^{1i}$. Llamando $\mathcal{A}_1 = P_1 A = (\alpha_{ij}^1)$ entonces $\alpha_{11}^1 = a_{i1} \neq 0$. Mediante combinaciones lineales apropiadas de la primera fila con las otras filas de la matriz $\mathcal{A}_1$ eliminamos todos los elementos de la primera columna de la matriz $\mathcal{A}_1$ salvo el primero, lo que en escritura matricial equivale a multiplicar la matriz $\mathcal{A}_1$ a la izquierda por la matriz

$$E_1 = \begin{pmatrix} 1 & & & \\ -\dfrac{\alpha_{21}^1}{\alpha_{11}^1} & 1 & & \\ \vdots & & \ddots & \\ -\dfrac{\alpha_{n1}^1}{\alpha_{11}^1} & & & 1 \end{pmatrix}.$$

  Sea

$$A_2 = E_1(P_1 A) = \begin{pmatrix} \alpha_{11}^1 & \alpha_{12}^1 & \cdots & \alpha_{1n}^1 \\ 0 & a_{22}^2 & \cdots & a_{2n}^2 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & a_{n2}^2 & \cdots & a_{nn}^2 \end{pmatrix}$$

  (obsérvese que la primera fila de A_2 coincide con la de $\mathcal{A}_1$). Como $\det(E_1) = 1$ y $\det(P_1) = \pm 1$, entonces

$$\alpha_{11}^1 \det \begin{pmatrix} a_{22}^2 & \cdots & a_{2n}^2 \\ \cdots & \cdots & \cdots \\ a_{n2}^2 & \cdots & a_{nn}^2 \end{pmatrix} = \det(A_2) = \pm \det(A) \neq 0.$$

  Por tanto, la submatriz

$$\begin{pmatrix} a_{22}^2 & \cdots & a_{2n}^2 \\ \cdots & \cdots & \cdots \\ a_{n2}^2 & \cdots & a_{nn}^2 \end{pmatrix}$$

  es inversible.

- k–ésima etapa de eliminación. Supongamos que hemos llevado a cabo $k-1$ pasos de la eliminación y veamos que se puede efectuar el siguiente. La hipótesis, por tanto, es que tenemos una matriz

$$A_k = E_{k-1}P_{k-1}\cdots E_2P_2E_1P_1A$$

de la forma

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

donde la submatriz

$$\begin{pmatrix} a_{kk}^k & \cdots & a_{kn}^k \\ \cdots & \cdots & \cdots \\ a_{nk}^k & \cdots & a_{nn}^k \end{pmatrix}$$

es inversible. Por ello, existe $i \in \{k, k+1, \ldots, n\}$ tal que $a_{ik}^k \neq 0$. Seguidamente permutamos la fila del pivote a_{ik}^k con la k–ésima fila de la matriz A_k, lo que equivale a multiplicar la matriz A_k, por la izquierda, por una matriz de permutación $P_k = P^{ik}$.

Llamando

$$\mathcal{A}_k = P_kA_k = \begin{pmatrix} \alpha_{11}^1 & \alpha_{12}^1 & \cdots & \alpha_{1,k-1}^1 & \alpha_{1k}^1 & \cdots & \alpha_{1n}^1 \\ & \alpha_{22}^2 & \cdots & \alpha_{2,k-1}^2 & \alpha_{2k}^2 & \cdots & \alpha_{2n}^2 \\ & & \ddots & \cdots & \cdots & \cdots & \cdots \\ & & & \alpha_{k-1,k-1}^{k-1} & \alpha_{k-1,k}^{k-1} & \cdots & \alpha_{k-1,n}^{k-1} \\ & & & & \alpha_{kk}^k & \cdots & \alpha_{kn}^k \\ & & & & \cdots & \ddots & \cdots \\ & & & & \alpha_{nk}^k & \cdots & \alpha_{nn}^k \end{pmatrix}$$

se verifica que $\alpha_{kk}^k = a_{ik}^k \neq 0$.

La eliminación propiamente dicha equivale a la multiplicación de la matriz $\mathcal{A}_k$ por la izquierda por la matriz

$$E_k = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & -\dfrac{\alpha_{k+1,k}^k}{\alpha_{kk}^k} & 1 & & \\ & & & \vdots & & \ddots & \\ & & & -\dfrac{\alpha_{nk}^k}{\alpha_{kk}^k} & & & 1 \end{pmatrix}.$$

Descomponiendo las matrices E_k y $\mathcal{A}_k$ en bloques de la forma

$$E_k = \left(\begin{array}{ccc|ccc} 1 & & & & & \\ & \ddots & & & \mathbf{0} & \\ & & 1 & & & \\ \hline & & & 1 & & \\ & \mathbf{0} & & -\frac{\alpha_{k+1,k}^k}{\alpha_{kk}^k} & 1 & \\ & & & \cdots & & \ddots \\ & & & -\frac{\alpha_{nk}^k}{\alpha_{kk}^k} & & 1 \end{array}\right)$$

y

$$\mathcal{A}_k = \left(\begin{array}{ccc|ccc} \alpha_{11}^1 & \cdots & \alpha_{1,k-1}^1 & \alpha_{1k}^1 & \alpha_{1,k+1}^1 & \cdots & \alpha_{1n}^1 \\ & \ddots & \vdots & \vdots & \vdots & & \vdots \\ & & \alpha_{k-1,k-1}^{k-1} & \alpha_{k-1,k}^{k-1} & \alpha_{k-1,k+1}^{k-1} & \cdots & \alpha_{k-1,n}^{k-1} \\ \hline & & & \alpha_{kk}^k & \alpha_{k,k+1}^k & \cdots & \alpha_{kn}^k \\ & \mathbf{0} & & \alpha_{k+1,k}^k & \alpha_{k+1,k+1}^k & \cdots & \alpha_{k+1,n}^k \\ & & & & & & \\ & & & \cdots & \cdots & \ddots & \cdots \\ & & & \alpha_{nk}^k & \alpha_{n,k+1}^k & \cdots & \alpha_{nn}^k \end{array}\right)$$

al multiplicarlas se tiene que la matriz $A_{k+1} = E_k \mathcal{A}_k$ es de la forma

$$A_{k+1} = \left(\begin{array}{ccc|ccc} \alpha_{11}^1 & \cdots & \alpha_{1,k-1}^1 & \alpha_{1k}^1 & \alpha_{1,k+1}^1 & \cdots & \alpha_{1n}^1 \\ & \ddots & \vdots & \vdots & \vdots & & \vdots \\ & & \alpha_{k-1,k-1}^{k-1} & \alpha_{k-1,k}^{k-1} & \alpha_{k-1,k+1}^{k-1} & \cdots & \alpha_{k-1,n}^{k-1} \\ \hline & & & \alpha_{kk}^k & \alpha_{k,k+1}^k & \cdots & \alpha_{kn}^k \\ & \mathbf{0} & & 0 & a_{k+1,k+1}^{k+1} & \cdots & a_{k+1,n}^{k+1} \\ & & & \cdots & \cdots & \ddots & \cdots \\ & & & 0 & a_{n,k+1}^{k+1} & \cdots & a_{nn}^{k+1} \end{array}\right),$$

operación que no modifica las k primeras filas de la matriz $\mathcal{A}_k$ y que consigue ceros por debajo de la diagonal en la columna k–ésima. Además, como

$$\alpha_{11}^1 \alpha_{22}^2 \cdots \alpha_{k-1,k-1}^{k-1} \alpha_{kk}^k \det \begin{pmatrix} a_{k+1,k+1}^{k+1} & \cdots & a_{k+1,n}^{k+1} \\ \cdots\cdots\cdots\cdots & & \cdots\cdots \\ a_{n,k+1}^{k+1} & \cdots & a_{nn}^{k+1} \end{pmatrix} = \det(A_{k+1})$$
$$= \pm \det(A) \neq 0,$$

la submatriz

$$\begin{pmatrix} a_{k+1,k+1}^{k+1} & \cdots & a_{k+1,n}^{k+1} \\ \cdots\cdots\cdots\cdots & & \cdots\cdots \\ a_{n,k+1}^{k+1} & \cdots & a_{nn}^{k+1} \end{pmatrix}$$

es inversible.

Si de este proceso se llevan a cabo $n-1$ pasos, se obtiene que

$$A_n = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1 A$$

es una matriz triangular superior. La matriz

$$M = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1$$

es inversible puesto que

$$\det(M) = \begin{cases} +1 & \text{si } \Lambda \text{ es par} \\ -1 & \text{si } \Lambda \text{ es impar,} \end{cases}$$

donde Λ es el número de matrices de permutación distintas de la identidad. Consecuentemente, la inversibilidad de la matriz M determina que

$$Au = b \Leftrightarrow MAu = Mb.$$

Observación 4.5. Por el proceso de eliminación se obtiene un procedimiento rápido para calcular $\det(A)$. Como

$$A_n = E_{n-1}P_{n-1}\cdots E_2P_2E_1P_1A = MA \text{ con } \det(M) = \pm 1$$

entonces

$$\det(A) = \frac{\det(A_n)}{\det(M)} = \pm\alpha_{11}^1\alpha_{22}^2\cdots\alpha_{n-1,n-1}^{n-1}a_{nn}^n. \quad \square$$

Veamos que el método de Gauss es aplicable a matrices arbitrarias (independientemente de la inversibilidad de éstas).

Teorema 4.1. *Sea $A \in \mathcal{M}_n$ (inversible o no). Se verifica que existe una matriz $M \in \mathcal{M}_n$ inversible de forma que la matriz MA es triangular superior.*

Demostración. El resultado está ya demostrado si la matriz A es inversible. Para lo único que se ha empleado la hipótesis de que la matriz A sea inversible es para asegurar, en la k–ésima etapa de eliminación, la existencia de algún $a_{ik}^k \neq 0$ con $i \in \{k, k+1, \ldots, n\}$. En el caso de que todos estos elementos fueran nulos, la matriz A_k ya tendría la forma A_{k+1}; bastaría, por tanto, tomar $P_k = E_k = I$ y continuar el proceso. $\square$

Observación 4.6. Contemos el número de operaciones elementales efectuadas en el método de Gauss:

1. Eliminación. Para construir la matriz A_{k+1} a partir de A_k se efectúan $n-k$ divisiones, $(n-k)(n-k) = (n-k)^2$ sumas y $(n-k)(n-k) = (n-k)^2$ multiplicaciones. Por otra parte, para pasar del vector $E_{k-1}P_{k-1}\cdots E_1P_1b$ al vector $E_kP_k\cdots E_1P_1b$ se efectúan $n-k$ sumas y $n-k$ multiplicaciones. Así, en total hay

$$\begin{cases} \displaystyle\sum_{k=1}^{n-1}(n-k)^2 + \sum_{k=1}^{n-1}(n-k) = \sum_{m=1}^{n-1}m^2 + \sum_{m=1}^{n-1}m = \frac{n^3-n}{3} & \text{sumas} \\ \displaystyle\sum_{k=1}^{n-1}(n-k)^2 + \sum_{k=1}^{n-1}(n-k) = \sum_{m=1}^{n-1}m^2 + \sum_{m=1}^{n-1}m = \frac{n^3-n}{3} & \text{productos} \\ \displaystyle\sum_{k=1}^{n-1}(n-k) = \sum_{m=1}^{n-1}m = \frac{n(n-1)}{2} & \text{divisiones} \end{cases}$$

2. Método de remonte. La resolución de un sistema triangular requiere

$$\begin{cases} \displaystyle\sum_{m=1}^{n-1} m = \frac{n(n-1)}{2} & \text{sumas} \\ \displaystyle\sum_{m=1}^{n-1} m = \frac{n(n-1)}{2} & \text{productos} \\ n & \text{divisiones} \end{cases}$$

Es decir, el método de Gauss necesita

$$\frac{2}{3}(n^3-n)+\frac{3}{2}n(n-1)+n=\frac{n(4n^2+9n-7)}{6}\sim\frac{2n^3}{3}$$

operaciones. Comparemos el número de operaciones elementales del método de Gauss con el número de operaciones elementales necesarias para la aplicación de las *fórmulas de Cramer*

$$u_i=\frac{\det(\beta_i)}{\det(A)}$$

para $i=1,2,\ldots,n$, donde

$$\beta_i=\begin{pmatrix} a_{11} & \cdots & a_{1,i-1} & b_1 & a_{1,i+1} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2,i-1} & b_2 & a_{2,i+1} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & \cdots & a_{n,i-1} & b_n & a_{n,i+1} & \cdots & a_{nn} \end{pmatrix}$$

en las que se deben evaluar $n+1$ determinantes y efectuar n divisiones. El cálculo de un determinante exige

$$\begin{cases} n!-1 & \text{sumas} \\ (n-1)n! & \text{productos} \end{cases}$$

Es decir, las fórmulas de Cramer necesitan

$$(n+1)(n!-1+(n-1)n!)+n=n(n+1)!-1$$

operaciones. Comparemos ambos métodos para algunos valores concretos del orden n de la matriz A:

n	operaciones (Gauss)	operaciones (Cramer)
10	805	399167999
100	681550	$\sim 10^{162}$
1000	668165500	$\sim 10^{2573}$

Hasta ahora hemos considerado como pivotes elementos cualesquiera siempre y cuando sean no nulos. Hagamos algunas observaciones en cuanto a la elección del pivote en cada etapa de eliminación.

Ejemplo 4.1. Consideremos el sistema lineal $Au = b$ donde

$$A = \begin{pmatrix} 2^{-26} & 1 \\ 1 & 1 \end{pmatrix},\ b = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \ \text{y} \ u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

La solución exacta del sistema viene dada por

$$\begin{cases} u_1 = 2 - u_2 \simeq 1.00000001490116 \\ u_2 = \dfrac{1 - 2^{-25}}{1 - 2^{-26}} \simeq 0.99999998509884. \end{cases}$$

Trabajando en precisión simple, el valor de estos números es

$$u_1 = u_2 = 1.$$

a) Tomamos 2^{-26} como primer pivote.

$$A = A_1 = \begin{pmatrix} 2^{-26} & 1 \\ 1 & 1 \end{pmatrix},\ P_1 = I,\ E_1 = \begin{pmatrix} 1 & 0 \\ -2^{26} & 1 \end{pmatrix}$$

$$A_2 = E_1 P_1 A = E_1 A = \begin{pmatrix} 1 & 0 \\ -2^{26} & 1 \end{pmatrix} \begin{pmatrix} 2^{-26} & 1 \\ 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 2^{-26} & 1 \\ 0 & 1 - 2^{26} \end{pmatrix} \simeq \begin{pmatrix} 2^{-26} & 1 \\ 0 & -2^{26} \end{pmatrix}$$

$$Mb = E_1 P_1 b = \begin{pmatrix} 1 & 0 \\ -2^{26} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 - 2^{26} \end{pmatrix} \simeq \begin{pmatrix} 1 \\ -2^{26} \end{pmatrix}.$$

De esta forma,

$$Au = b \ \Rightarrow\ MAu = Mb \ \Rightarrow\ \begin{pmatrix} 2^{-26} & 1 \\ 0 & -2^{26} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -2^{26} \end{pmatrix}.$$

Por tanto,

$$\begin{cases} -2^{26} u_2 = -2^{26} & \Rightarrow \quad u_2 = 1 \\ 2^{-26} u_1 + u_2 = 1 & \Rightarrow \quad u_1 = 0. \end{cases}$$

b) Tomamos 1 como primer pivote.

$$A = A_1 = \begin{pmatrix} 2^{-26} & 1 \\ 1 & 1 \end{pmatrix},\ P_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ E_1 = \begin{pmatrix} 1 & 0 \\ -2^{-26} & 1 \end{pmatrix}$$

$$A_2 = E_1 P_1 A = \begin{pmatrix} 1 & 0 \\ -2^{-26} & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 2^{-26} & 1 \\ 1 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} 1 & 1 \\ 0 & 1 - 2^{-26} \end{pmatrix} \simeq \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$Mb = E_1 P_1 b = \begin{pmatrix} 1 & 0 \\ -2^{-26} & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 - 2^{-25} \end{pmatrix} \simeq \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

Así, en este caso,

$$Au = b \;\Rightarrow\; MAu = Mb \;\Rightarrow\; \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

Por tanto,

$$\begin{cases} u_2 = 1 & \Rightarrow \quad u_2 = 1 \\ u_1 + u_2 = 2 & \Rightarrow \quad u_1 = 1. \end{cases} \quad \square$$

Este ejemplo pone de manifiesto que los errores de redondeo con efecto desastroso provienen de la división por pivotes "muy pequeños". En la práctica se utiliza una de las dos estrategias siguientes al comienzo de la k–ésima etapa de eliminación, $1 \leq k \leq n-1$:

a) Estrategia del *pivote parcial*: se toma como pivote el elemento de mayor módulo de entre los $n-k+1$ últimos elementos de la columna k–ésima; es decir, se elige a_{ik}^k, $k \leq i \leq n$, de forma que

$$|a_{ik}^k| = \max_{k \leq p \leq n} |a_{pk}^k|.$$

b) Estrategia del *pivote total*: se toma como pivote el elemento de mayor módulo de la submatriz correspondiente de la matriz A_k; es decir, se elige el elemento a_{ij}^k, $k \leq i, j \leq n$, de modo que

$$|a_{ij}^k| = \max_{k \leq p,q \leq n} |a_{pq}^k|.$$

Si el pivote elegido por esta estrategia no está situado en la k–ésima columna hay que efectuar un cambio de columnas, lo que equivale a multiplicar a la derecha la matriz A_k por una matriz P^{kj} de permutación. Esto no es otra cosa que intercambiar el orden de las incógnitas, intercambio que habrá que tener en cuenta a la hora de escribir en el orden adecuado el resultado. Esta dificultad añadida hace que, en general, en el método de Gauss se utilice habitualmente la estrategia del pivote parcial. $\square$

4.3.3. Factorización PA=LU

Veamos cómo, reinterpretando el método de eliminación de Gauss, se consigue transformar el sistema original en un par de sistemas triangulares con cuya resolución obtendremos la solución del problema de partida. Además, y esto es lo más importante, esta reinterpretación nos servirá para diseñar una implementación efectiva de la eliminación gaussiana. Para ello, establezcamos un resultado técnico previo.

Lema 4.1. *Sea*

$$E_k = I + \ell_k \mathbf{e}_k^{\mathrm{T}}$$

para $k = 1, 2, \ldots, n-1$, *donde* $\mathbf{e}_k$ *es el* k*–ésimo vector de la base canónica y*

$$\ell_k = (0, \overset{k)}{\cdots}, 0, \ell_{k+1,k}, \ell_{k+2,k}, \ldots, \ell_{nk})^{\mathrm{T}}.$$

a) Las matrices E_k *son inversibles y*

$$(E_k)^{-1} = I - \ell_k \mathbf{e}_k^{\mathrm{T}} \tag{4.2}$$

para $k = 1, 2, \ldots, n-1$.

b) Si $P = P^{ij}$ *es una matriz de permutación de las líneas* i *y* j *con* $j \geq i > k$ *entonces*

$$PE_kP = E'_k$$

para $k = 1, 2, \ldots, n-1$, *donde*

$$E'_k = I + \ell'_k \mathbf{e}_k^{\mathrm{T}},$$

siendo ℓ'_k *el vector* ℓ_k *al que se le han intercambiado las coordenadas* i *y* j, *es decir,*

$$\ell'_k = (0, \overset{k)}{\cdots}, 0, \ell_{k+1,k}, \ldots, \overset{i)}{\ell_{jk}}, \ldots, \overset{j)}{\ell_{ik}}, \ldots, \ell_{nk})^{\mathrm{T}}.$$

c) Para todo $k \in \{2, 3, \ldots, n-1\}$ *se verifica que*

$$E_1E_2 \cdots E_k = I + \sum_{i=1}^{k} \ell_i \mathbf{e}_i^{\mathrm{T}} = I + \ell_1 \mathbf{e}_1^{\mathrm{T}} + \ell_2 \mathbf{e}_2^{\mathrm{T}} + \cdots + \ell_k \mathbf{e}_k^{\mathrm{T}}.$$

En particular,

$$\begin{aligned} E_1E_2 \cdots E_{n-1} &= I + \ell_1 \mathbf{e}_1^{\mathrm{T}} + \ell_2 \mathbf{e}_2^{\mathrm{T}} + \cdots + \ell_{n-1} \mathbf{e}_{n-1}^{\mathrm{T}} \\ &= \begin{pmatrix} 1 & & & & \\ \ell_{21} & 1 & & & \\ \ell_{31} & \ell_{32} & 1 & & \\ \cdots & \cdots & \cdots & \ddots & \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{n,n-1} & 1 \end{pmatrix}. \end{aligned}$$

Demostración.

a) Como $\det(E_k) = 1$ entonces la matriz E_k es inversible. Además, para cada $k \in \{1, 2, \ldots, n-1\}$, se verifica que

$$\begin{aligned}\left(I + \ell_k \mathbf{e}_k^{\mathrm{T}}\right)\left(I - \ell_k \mathbf{e}_k^{\mathrm{T}}\right) &= I + \ell_k \mathbf{e}_k^{\mathrm{T}} - \ell_k \mathbf{e}_k^{\mathrm{T}} - \ell_k \mathbf{e}_k^{\mathrm{T}} \ell_k \mathbf{e}_k^{\mathrm{T}} \\ &= I - \ell_k \left(\mathbf{e}_k^{\mathrm{T}} \ell_k\right) \mathbf{e}_k^{\mathrm{T}} = I,\end{aligned}$$

ya que $\mathbf{e}_k^{\mathrm{T}} \ell_k = 0$, de donde se sigue (4.2).

b) Como $P^{-1} = P$ (véase (4.1)), se verifica que

$$PE_kP = P\left(I + \ell_k \mathbf{e}_k^{\mathrm{T}}\right) P = PIP + (P\ell_k)\left(\mathbf{e}_k^{\mathrm{T}} P\right) = I + \ell'_k \mathbf{e}_k^{\mathrm{T}},$$

dado que $P\ell_k = \ell'_k$ y $\mathbf{e}_k^{\mathrm{T}} P$ es la k–ésima fila de la matriz P, es decir, $\mathbf{e}_k^{\mathrm{T}}$.

c) Procedemos por inducción:

i) al multiplicar E_1 por E_2 se tiene que

$$\begin{aligned}E_1E_2 &= \left(I + \ell_1 \mathbf{e}_1^{\mathrm{T}}\right)\left(I + \ell_2 \mathbf{e}_2^{\mathrm{T}}\right) = I + \ell_1 \mathbf{e}_1^{\mathrm{T}} + \ell_2 \mathbf{e}_2^{\mathrm{T}} + \ell_1 \mathbf{e}_1^{\mathrm{T}} \ell_2 \mathbf{e}_2^{\mathrm{T}} \\ &= I + \ell_1 \mathbf{e}_1^{\mathrm{T}} + \ell_2 \mathbf{e}_2^{\mathrm{T}} + \ell_1 \left(\mathbf{e}_1^{\mathrm{T}} \ell_2\right) \mathbf{e}_2^{\mathrm{T}} = I + \ell_1 \mathbf{e}_1^{\mathrm{T}} + \ell_2 \mathbf{e}_2^{\mathrm{T}},\end{aligned}$$

ya que $\mathbf{e}_1^{\mathrm{T}} \ell_2 = 0$.

ii) suponiendo que

$$E_1E_2 \cdots E_k = I + \ell_1 \mathbf{e}_1^{\mathrm{T}} + \ell_2 \mathbf{e}_2^{\mathrm{T}} + \cdots + \ell_k \mathbf{e}_k^{\mathrm{T}},$$

se verifica que

$$\begin{aligned}E_1E_2 \cdots E_kE_{k+1} &= \left(I + \ell_1 \mathbf{e}_1^{\mathrm{T}} + \cdots + \ell_k \mathbf{e}_k^{\mathrm{T}}\right)\left(I + \ell_{k+1} \mathbf{e}_{k+1}^{\mathrm{T}}\right) \\ &= I + \ell_1 \mathbf{e}_1^{\mathrm{T}} + \cdots + \ell_k \mathbf{e}_k^{\mathrm{T}} + \ell_{k+1} \mathbf{e}_{k+1}^{\mathrm{T}} \\ &\quad + \sum_{i=1}^{k} \ell_i \left(\mathbf{e}_i^{\mathrm{T}} \ell_{k+1}\right) \mathbf{e}_{k+1}^{\mathrm{T}} \\ &= I + \ell_1 \mathbf{e}_1^{\mathrm{T}} + \cdots + \ell_k \mathbf{e}_k^{\mathrm{T}} + \ell_{k+1} \mathbf{e}_{k+1}^{\mathrm{T}},\end{aligned}$$

pues $\mathbf{e}_i^{\mathrm{T}} \ell_{k+1} = 0$, $i = 1, 2, \ldots, k$. □

Gracias al lema 4.1 podemos demostrar el resultado central de esta sección que asegura que, previa una permutación de filas, toda matriz puede escribirse como producto de una matriz L triangular inferior por una matriz U triangular superior.[1]

[1] Las notaciones L y U son anglosajonas: L por *lower* (inferior) y U por *upper* (superior).

Teorema 4.2. *Para toda matriz $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ existen una matriz P que es producto de $n-1$ matrices de permutación de líneas, una matriz triangular inferior $L = (l_{ij})_{i,j=1}^n$ con $l_{ii} = 1$, $i = 1, 2, \ldots, n$, y una matriz triangular superior U tales que $PA = LU$.*

DEMOSTRACIÓN. Si se aplica a la matriz A el método de Gauss se obtiene una matriz triangular superior

$$U = A_n = E_{n-1}P_{n-1}E_{n-2}P_{n-2}\cdots E_1P_1A \tag{4.3}$$

donde $E_k = I + \ell_k \mathbf{e}_k^{\mathrm{T}}$ y $P_kP_k = I$ para todo $k = 1, 2, \ldots, n-1$. Llamando

$$P = P_{n-1}P_{n-2}\cdots P_1,$$

se verifica que la matriz P es inversible y

$$P^{-1} = (P_{n-1}P_{n-2}\cdots P_1)^{-1} = (P_1)^{-1}(P_2)^{-1}\cdots(P_{n-1})^{-1} = P_1P_2\cdots P_{n-1}.$$

Veamos que la matriz U puede expresarse en la forma

$$U = \mathcal{E}_{n-1}\mathcal{E}_{n-2}\cdots\mathcal{E}_1PA$$

siendo

$$\mathcal{E}_k = P_{n-1}P_{n-2}\ldots P_{k+1}E_kP_{k+1}P_{k+2}\ldots P_{n-1}$$

para $k = 1, 2, \ldots, n-1$. En efecto, basta intercalar adecuadamente productos de la forma $P_kP_k = I$ en (4.3). Así, el producto de todas las filas de la tabla siguiente (en el orden en que aparecen) no es otra cosa que la matriz U, siendo el producto de matrices de la fila k–ésima la matriz $\mathcal{E}_{n-k}$ para $k = 1, 2, \ldots, n-1$, excepto la última fila para $k = n$, cuyo producto es PA.

$$
\begin{array}{cccccccccccc}
E_{n-1} & & & & & & & & & & & \\
P_{n-1} & E_{n-2} & \mathbf{P_{n-1}} & & & & & & & & & \\
\mathbf{P_{n-1}} & P_{n-2} & E_{n-3} & \mathbf{P_{n-2}} & \mathbf{P_{n-1}} & & & & & & & \\
\mathbf{P_{n-1}} & \mathbf{P_{n-2}} & P_{n-3} & E_{n-4} & \mathbf{P_{n-3}} & \mathbf{P_{n-2}} & \mathbf{P_{n-1}} & & & & & \\
 & & & & & & & & & & & \\
\ldots & & & \ddots & \ddots & & & & & & & \\
\ldots & & & & \ddots & \ddots & & & & & & \\
\ldots & & & & & \ddots & \ddots & & & & & \\
 & & & & & & & & & & & \\
\mathbf{P_{n-1}} & \mathbf{P_{n-2}} & \mathbf{P_{n-3}} & \mathbf{P_{n-4}} & \mathbf{P_{n-5}} & \ldots & P_3 & E_2 & \mathbf{P_3} & \mathbf{P_4} & \ldots & \mathbf{P_{n-1}} \\
\mathbf{P_{n-1}} & \mathbf{P_{n-2}} & \mathbf{P_{n-3}} & \mathbf{P_{n-4}} & \mathbf{P_{n-5}} & \ldots & \mathbf{P_3} & P_2 & E_1 & \mathbf{P_2} & \ldots & \mathbf{P_{n-1}} \\
\mathbf{P_{n-1}} & \mathbf{P_{n-2}} & \mathbf{P_{n-3}} & \mathbf{P_{n-4}} & \mathbf{P_{n-5}} & \ldots & \mathbf{P_3} & \mathbf{P_2} & P_1 & A & &
\end{array}
$$

Veamos ahora cómo determinar la matriz L. Aplicando reiteradas veces el segundo apartado del lema 4.1 se obtiene que

$$\mathcal{E}_k = I + \widehat{\ell}_k \mathbf{e}_k^{\mathrm{T}}$$

para $k = 1, 2, \ldots, n-1$, donde cada vector $\widehat{\ell}_k$ es el correspondiente vector ℓ_k al que se le han efectuado las permutaciones $\{P_i\}_{i=k+1}^{n-1}$. De esta forma, como las matrices $\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_{n-1}\}$ son inversibles, entonces

$$PA = (\mathcal{E}_{n-1}\mathcal{E}_{n-2}\cdots\mathcal{E}_1)^{-1}\, U = (\mathcal{E}_1)^{-1}(\mathcal{E}_2)^{-1}\cdots(\mathcal{E}_{n-1})^{-1}U.$$

Por el primer y el tercer apartados del lema 4.1 sabemos que

$$(\mathcal{E}_k)^{-1} = I - \widehat{\ell}_k \mathbf{e}_k^{\mathrm{T}}$$

y

$$(\mathcal{E}_1)^{-1}(\mathcal{E}_2)^{-1}\cdots(\mathcal{E}_{n-1})^{-1} = I - \sum_{i=1}^{n-1} \widehat{\ell}_i \mathbf{e}_i^{\mathrm{T}} = \begin{pmatrix} 1 & & & & \\ -\widehat{\ell}_{21} & 1 & & & \\ -\widehat{\ell}_{31} & -\widehat{\ell}_{32} & 1 & & \\ \cdots & \cdots & \cdots & \ddots & \\ -\widehat{\ell}_{n1} & -\widehat{\ell}_{n2} & \cdots & -\widehat{\ell}_{n,n-1} & 1 \end{pmatrix}.$$

El resultado se sigue al considerar

$$L = (\mathcal{E}_1)^{-1}(\mathcal{E}_2)^{-1}\cdots(\mathcal{E}_{n-1})^{-1}. \quad \square$$

Observación 4.7.

1. Los elementos $\widehat{\ell}_{jk}$ de la matriz L son, en realidad, los multiplicadores $-\frac{\alpha_{jk}^k}{\alpha_{kk}^k}$ que aparecen al aplicar el método de eliminación de Gauss, sobre los que se han efectuado las correspondientes permutaciones.

2. Si se quiere resolver el sistema $Au = b$ y se conoce la factorización $PA = LU$, como P es inversible, se tiene que

$$Au = b \;\Leftrightarrow\; PAu = Pb \;\Leftrightarrow\; LUu = Pb \;\Leftrightarrow\; \begin{cases} Lw = Pb \\ Uu = w. \end{cases}$$

   Obtenemos así dos sistemas triangulares que se resuelven mediante el método de remonte. Ésta va a ser la forma en que implementaremos la eliminación gaussiana como se verá en la subsección 4.3.4.

3. Obsérvese que, una vez calculadas las matrices P, L y U, si se quieren resolver varios sistemas con la misma matriz A, basta (para cada uno de ellos) calcular el "nuevo" segundo miembro Pb y resolver por el método de remonte dos sistemas triangulares. Así se hará, por ejemplo, para calcular A^{-1}, caso en el que se resolverán los n sistemas lineales $Az_i = \mathbf{e}_i$ para $i = 1, 2, \ldots, n$, donde $\mathbf{e}_i$ es el i–ésimo vector de la base canónica. □

4.3.4. Implementación del método de eliminación gaussiana

A continuación, describimos la forma en que vamos a implementar la eliminación gaussiana utilizando, como se ha indicado anteriormente, la estrategia del pivote parcial. Destacamos que todo el proceso de eliminación se llevará a cabo sobre la propia matriz A de forma que, al acabar, tendremos la matriz U en la parte triangular superior de A y la matriz L en la parte triangular inferior de A (obviamente, los unos de la diagonal de la matriz L no es necesario almacenarlos).

Para guardar la información con vistas a resolver otro sistema con la misma matriz, en la etapa k–ésima de eliminación debemos almacenar, entre otras cosas, las matrices E_k (o, mejor dicho, la parte relevante de ellas). Es decir, debemos guardar los valores de los multiplicadores $-\frac{\alpha_{jk}^k}{\alpha_{kk}^k}$, $j = k+1, k+2, \ldots, n$; en realidad, guardaremos sus opuestos, con vistas a obtener las matrices $\mathcal{E}_k$ que aparecen en la demostración del teorema 4.2. Los lugares más adecuados para hacerlo son los "huecos" que quedan en A tras "hacer ceros" en la columna k–ésima: los elementos $A(j,k)$, $j = k+1, k+2, \ldots, n$. Almacenando así estos valores, cuando en los pasos sucesivos hagamos permutaciones de filas en A estaremos cambiando el orden, también, de los multiplicadores. En otras palabras, estaremos realizando las transformaciones del tipo del lema 4.1 (apartado segundo) que se llevan a cabo en la demostración del teorema 4.2. Por otra parte, no se realizará el producto por la matriz E_k, sino que, simplemente, a cada trozo de fila j–ésima, $A(j, k+1:n)$, con $j \in \{k+1, k+2, \ldots, n\}$, se le restará el trozo correspondiente de la fila k–ésima, $A(k, k+1:n)$, multiplicada por el "multiplicador" almacenado $\frac{\alpha_{jk}^k}{\alpha_{kk}^k}$.

Actuando de esta forma, a partir de la matriz A, si denotamos $U = (u_{ij})_{i,j=1}^n$ y $L = (l_{ij})_{i,j=1}^n$, obtendremos una matriz del tipo

$$\begin{pmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1,n-1} & u_{1n} \\ l_{21} & u_{22} & u_{23} & \cdots & u_{2,n-1} & u_{2n} \\ l_{31} & l_{32} & u_{33} & \cdots & u_{3,n-1} & u_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ l_{n-1,1} & l_{n-1,2} & l_{n-1,3} & \cdots & u_{n-1,n-1} & u_{n-1,n} \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{n,n-1} & u_{nn} \end{pmatrix}.$$

Ejemplo 4.2. Veamos cómo se desarrolla el proceso anterior cuando lo aplicamos al ejemplo tratado en la subsección 4.3.1.

$$\begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 2 & 1 & 3 \\ 1 & 1 & -1 & 1 \\ 0 & 1 & 8 & 12 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 1 & 1 & -1 & 1 \\ 0 & 1 & 8 & 12 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 1 & 3 \\ \mathbf{0} & 1 & 2 & 1 \\ \mathbf{1} & -1 & -2 & -2 \\ \mathbf{0} & 1 & 8 & 12 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 2 & 1 & 3 \\ \mathbf{0} & 1 & 2 & 1 \\ \mathbf{1} & -1 & -2 & -2 \\ \mathbf{0} & 1 & 8 & 12 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 1 & 3 \\ \mathbf{0} & 1 & 2 & 1 \\ \mathbf{1} & \mathbf{-1} & 0 & -1 \\ \mathbf{0} & \mathbf{1} & 6 & 11 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 & 1 & 3 \\ \mathbf{0} & 1 & 2 & 1 \\ \mathbf{0} & \mathbf{1} & 6 & 11 \\ \mathbf{1} & \mathbf{-1} & 0 & -1 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 2 & 1 & 3 \\ \mathbf{0} & 1 & 2 & 1 \\ \mathbf{0} & \mathbf{1} & 6 & 11 \\ \mathbf{1} & \mathbf{-1} & \mathbf{0} & -1 \end{pmatrix}.$$

Nótese que

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 6 & 11 \\ 0 & 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 8 & 12 \\ 1 & 1 & -1 & 1 \end{pmatrix}.$$

Como

$$P = P_3P_2P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} I \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

entonces

$$PA = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 2 & 1 & 3 \\ 1 & 1 & -1 & 1 \\ 0 & 1 & 8 & 12 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 8 & 12 \\ 1 & 1 & -1 & 1 \end{pmatrix} = LU. \quad \square$$

Finalmente, destacamos que las permutaciones de filas no es necesario realizarlas "físicamente"; es suficiente conocer, en cada etapa, qué par de filas hay que intercambiar. Para ello, basta utilizar un *puntero* `punt` con el que intercambiamos los valores de `punt`(i) y `punt`(j) cuando haya que intercambiar las filas i y j. Así,

en el ejemplo 4.2 obtendríamos, después de todo el proceso,

$$\begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 2 & 1 & 3 \\ 1 & 1 & -1 & 1 \\ 0 & 1 & 8 & 12 \end{pmatrix} \to \begin{pmatrix} \mathbf{0} & 1 & 2 & 1 \\ 1 & 2 & 1 & 3 \\ \mathbf{1} & \mathbf{-1} & \mathbf{0} & -1 \\ \mathbf{0} & \mathbf{1} & 6 & 11 \end{pmatrix} \quad \text{y} \quad \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \overset{\texttt{punt}}{\longrightarrow} \begin{pmatrix} 2 \\ 1 \\ 4 \\ 3 \end{pmatrix}.$$

Para resolver ahora sistemas $Au = b$ con diversos segundos miembros b, bastará utilizar el método de remonte, con el nuevo orden de ecuaciones dado por `punt`, en los sistemas $Lw = Pb$ y $Uu = w$. Para ello, basta tomar

$$\begin{cases} w(1) = b(\texttt{punt}(1)) \\ w(i) = b(\texttt{punt}(i)) - \displaystyle\sum_{j=1}^{i-1} A(\texttt{punt}(i), j)w(j),\ i = 2, 3, \ldots, n \end{cases}$$

y

$$\begin{cases} u(n) = \dfrac{w(n)}{A(\texttt{punt}(n), n)} \\ u(i) = \dfrac{1}{A(\texttt{punt}(i), i)} \left(w(i) - \displaystyle\sum_{j=i+1}^{n} A(\texttt{punt}(i), j)u(j) \right),\ i = n-1, n-2, \ldots, 1, \end{cases}$$

donde, llegados a este punto, en la matriz A estarán almacenadas las matrices L y U como ya se ha indicado. Así, para el ejemplo 4.2 con

$$b = \begin{pmatrix} 1 \\ 0 \\ 5 \\ 2 \end{pmatrix},$$

se obtiene:

$$\begin{cases} w(1) = b(2) = 0, \\ w(2) = b(1) - A(1,1)w(1) = 1 - 0 \cdot 0 = 1, \\ w(3) = b(4) - A(4,1)w(1) - A(4,2)w(2) = 2 - 0 \cdot 0 - 1 \cdot 1 = 1, \\ w(4) = b(3) - A(3,1)w(1) - A(3,2)w(2) - A(3,3)w(3) \\ \qquad = 5 - 1 \cdot 0 + 1 \cdot 1 - 0 \cdot 1 = 6 \end{cases}$$

y

$$\begin{cases} u(4) = \dfrac{w(4)}{A(3,4)} = \dfrac{6}{-1} = -6, \\ u(3) = \dfrac{w(3) - A(4,4)u(4)}{A(4,3)} = \dfrac{1 + 11 \cdot 6}{6} = \dfrac{67}{6}, \\ u(2) = \dfrac{w(2) - A(1,4)u(4) - A(1,3)u(3)}{A(1,2)} = 1\left(1 + 1 \cdot 6 - 2\dfrac{67}{6}\right) = -\dfrac{46}{3} \\ u(1) = \dfrac{w(1) - A(2,4)u(4) - A(2,3)u(3) - A(2,2)u(2)}{A(2,1)} \\ \qquad = 1\left(0 + 3 \cdot 6 - 1\dfrac{67}{6} + 2\dfrac{46}{3}\right) = \dfrac{75}{2}. \end{cases}$$

Así, las soluciones de los sistemas $Lw = Pb$ y $Uu = w$ son, respectivamente,

$$w = (0, 1, 1, 6)^{\mathrm{T}} \text{ y } u = \left(\frac{75}{2}, -\frac{46}{3}, \frac{67}{6}, -6\right)^{\mathrm{T}}.$$

4.4. Factorización LU de una matriz

Consideremos nuevamente el sistema $Au = b$ siendo A una matriz inversible y $b \in \mathbf{V}\backslash\{0\}$. Supongamos que en el proceso de eliminación gaussiana no es necesario realizar permutaciones de filas porque en cada etapa de eliminación el elemento a_{kk}^k es no nulo. Entonces,

$$P_1 = P_2 = \cdots = P_{n-1} = I. \tag{4.4}$$

En este caso, por el teorema 4.2, la matriz $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ se podrá factorizar en la forma $A = LU$, donde $L = (l_{ij})_{i,j=1}^n$ es una matriz triangular inferior con $l_{ii} = 1$, $i = 1, 2, \ldots, n$, y U es una matriz triangular superior. En concreto, se tendrá

$$L = (E_{n-1}E_{n-2}\cdots E_2E_1)^{-1} = (E_1)^{-1}(E_2)^{-1}\cdots(E_{n-1})^{-1}$$

y

$$U = (E_{n-1}E_{n-2}\cdots E_2E_1)A$$

puesto que, en este caso, $\mathcal{E}_k = E_k$, $k = 1, 2, \ldots, n-1$. Cuando se dé esta circunstancia, se tendrá una ventaja adicional: las matrices L y U podrán calcularse, directamente, a partir de los elementos de A.

Vamos a dar una condición suficiente para que se pueda llevar a cabo el proceso de eliminación de Gauss sin permutar filas.

Teorema 4.3. *Sea $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ de forma que los n menores principales de la matriz A*

$$\delta_k = \det \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix} \tag{4.5}$$

para $k = 1, 2, \ldots, n$ son no nulos. Entonces existe una única matriz triangular inferior $L = (l_{ij})_{i,j=1}^n$ con $l_{ii} = 1$, $i = 1, 2, \ldots, n$, y existe una única matriz triangular superior U tal que $A = LU$.

Demostración.

a) Existencia. Basta con demostrar que no es preciso permutar filas al aplicar el método de Gauss a la matriz A. Para ello procedemos por inducción:

i) Para $k = 1$ el resultado es obvio, pues al ser $a_{11} = \delta_1 \neq 0$ podemos tomar $P_1 = I$.

ii) Suponiendo cierto el resultado para $k-1$, lo probamos para k. Es decir, suponemos que se han podido elegir $P_1 = P_2 = \cdots = P_{k-1} = I$; si esto es así, se tiene que

$$A_k = E_{k-1}E_{k-2}\cdots E_1 A = M_{k-1}A$$

donde M_{k-1} es una matriz triangular inferior con

$$\operatorname{diag}(M_{k-1}) = (1, 1, \ldots, 1)$$

(por ser producto de matrices triangulares inferiores con unos en la diagonal principal). De esta forma, a partir de las siguientes descomposiciones en bloques de las matrices

$$A_k = \left(\begin{array}{ccc|ccc} a_{11}^1 & \cdots & a_{1k}^1 & a_{1,k+1}^1 & \cdots & a_{1n}^1 \\ & \ddots & \cdots & \cdots & & \cdots \\ & & a_{kk}^k & a_{k,k+1}^k & \cdots & a_{kn}^k \\ \hline & & a_{k+1,k}^k & a_{k+1,k+1}^k & \cdots & a_{k+1,n}^k \\ & & \cdots & \cdots & & \cdots \\ & & a_{nk}^k & a_{n,k+1}^k & \cdots & a_{nn}^k \end{array}\right),$$

$$M_{k-1} = \left(\begin{array}{cccc|ccc} 1 & & & & & & \\ m_{21} & 1 & & & & & \\ \cdots & \cdots & \ddots & & & & \\ m_{k1} & m_{k2} & \cdots & 1 & & & \\ \hline m_{k+1,1} & m_{k+1,2} & \cdots & m_{k+1,k} & 1 & & \\ m_{k+2,1} & m_{k+2,2} & \cdots & m_{k+2,k} & m_{k+2,k+1} & & \\ \cdots & \cdots & & \cdots & \cdots & \ddots & \\ m_{n1} & m_{n2} & \cdots & m_{nk} & m_{n,k+1} & \cdots & 1 \end{array}\right)$$

y

$$A = \left(\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1k} & a_{1,k+1} & \cdots & a_{1n} \\ \cdots & & \cdots & \cdots & & \cdots \\ a_{k1} & \cdots & a_{kk} & a_{k,k+1} & \cdots & a_{kn} \\ \hline a_{k+1,1} & \cdots & a_{k+1,k} & a_{k+1,k+1} & \cdots & a_{k+1,n} \\ \cdots & & \cdots & \cdots & & \cdots \\ a_{n1} & \cdots & a_{nk} & a_{n,k+1} & \cdots & a_{nn} \end{array}\right),$$

utilizando las reglas de multiplicación de matrices por bloques en la igualdad $A_k = M_{k-1}A$, en particular se tiene que

$$\begin{pmatrix} a_{11}^1 & a_{12}^1 & \cdots & a_{1k}^1 \\ & a_{22}^2 & \cdots & a_{2k}^2 \\ & & \ddots & \cdots \\ & & & a_{kk}^k \end{pmatrix} = \begin{pmatrix} 1 & & & \\ m_{21} & 1 & & \\ \cdots & \cdots & \ddots & \\ m_{k1} & m_{k2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ a_{21} & \cdots & a_{2k} \\ \cdots & \cdots & \cdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix},$$

de donde, tomando determinantes,

$$\prod_{i=1}^{k} a_{ii}^i = 1 \cdot \delta_k = \delta_k \neq 0,$$

es decir,

$$a_{ii}^i \neq 0$$

para $i = 1, 2, \ldots, k$. En particular, $a_{kk}^k \neq 0$, por lo que podemos elegir a_{kk}^k como pivote y, consecuentemente, $P_k = I$.

b) Unicidad. Supongamos que existieran dos factorizaciones

$$L_1 U_1 = A = L_2 U_2 \tag{4.6}$$

donde U_k son matrices triangulares superiores y L_k son matrices triangulares inferiores con la propiedad

$$(L_1)_{ii} = (L_2)_{ii} = 1$$

para $i = 1, 2, \ldots, n$ y $k = 1, 2$. Como

$$\det(L_k) = 1 \ \text{ y } \ \det(U_k) = \det(A) = \delta_n \neq 0,$$

las matrices L_k y U_k son inversibles para $k = 1, 2$. A partir de (4.6) se tiene entonces que

$$L_2^{-1} L_1 = U_2 U_1^{-1}. \tag{4.7}$$

Como L_2 es una matriz triangular inferior con unos en la diagonal principal, la matriz L_2^{-1} es también una matriz triangular inferior con unos en la diagonal principal. Como L_1 es también una matriz triangular inferior con unos en la diagonal principal, entonces

$$L_2^{-1} L_1 = \begin{pmatrix} 1 & & & \\ \mu_{21} & 1 & & \\ \cdots & \cdots & \ddots & \\ \mu_{n1} & \cdots & \mu_{n,n-1} & 1 \end{pmatrix}.$$

Por otra parte, como las matrices U_2 y U_1^{-1} son triangulares superiores, se verifica

$$U_2 U_1^{-1} = \begin{pmatrix} \nu_{11} & \nu_{12} & \cdots & \nu_{1n} \\ & \nu_{22} & \cdots & \nu_{2n} \\ & & \ddots & \cdots \\ & & & \nu_{nn} \end{pmatrix}.$$

Por tanto, la relación (4.7) determina que tanto $L_2^{-1} L_1$ como $U_2 U_1^{-1}$ son iguales a la matriz identidad I. Consecuentemente,

$$L_1 = L_2 \ \text{ y } \ U_1 = U_2. \quad \square$$

Observación 4.8. De hecho, la condición (4.5) es también necesaria. Es decir, si una matriz inversible $A \in \mathcal{M}_n$ puede escribirse como producto de una matriz triangular inferior por una triangular superior, entonces todos sus menores principales δ_k (incluyendo, obviamente, el de orden n) son no nulos (véase el problema 4.15). $\square$

Observación 4.9.

1. Toda matriz $A \in \mathcal{M}_n$ hermítica y definida positiva cumple que sus n submatrices principales son inversibles (véase el problema 2.15), por lo que estamos en las condiciones de aplicarle el teorema 4.3. Sin embargo, para este tipo de matrices está especialmente indicado el *método de Cholesky*, que estudiaremos en la sección 4.5.

2. Análogamente a como ocurría con la factorización $PA = LU$, si se tienen que resolver varios sistemas lineales con la misma matriz A (por ejemplo, para calcular A^{-1}) es suficiente conservar la expresión de las matrices L y U y resolver los dos sistemas lineales de matrices triangulares

$$\begin{cases} Lw = b \\ Uu = w. \end{cases} \quad \square$$

Observación 4.10. Como comentamos al comienzo de esta sección, los elementos de las matrices L y U pueden calcularse, directamente, a partir de los elementos de A. Para ello, basta considerar la igualdad $A = LU$, elemento a elemento, teniendo en cuenta que

$$l_{ii} = 1 \tag{4.8}$$

para $i = 1, 2, \ldots, n$ y, si la matriz A es inversible,

$$\prod_{i=1}^{n} u_{ii} = \det(A) \neq 0,$$

lo que implica que

$$u_{ii} \neq 0$$

para $i = 1, 2, \ldots, n$. Obviamente

$$a_{ij} = \sum_{k=1}^{n} l_{ik} u_{kj}$$

para $i, j = 1, 2, \ldots, n$ pero, como las matrices L y U son triangular inferior y superior respectivamente, se tiene que

$$l_{ik} = 0 \text{ si } k > i$$

y

$$u_{kj} = 0 \text{ si } k > j$$

por lo que la expresión anterior se puede escribir como

$$a_{ij} = \sum_{k=1}^{\min\{i,j\}} l_{ik} u_{kj} \tag{4.9}$$

para $i, j = 1, 2, \ldots, n$. Distinguimos dos casos:

a) $\boxed{i \leq j}$ En esta situación la expresión (4.9) toma la forma

$$a_{ij} = \sum_{k=1}^{i} l_{ik}u_{kj} = \sum_{k=1}^{i-1} l_{ik}u_{kj} + u_{ij}$$

para $1 \leq i \leq j \leq n$, por lo que

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}$$

para $j = i, i+1, \ldots, n$. En la expresión anterior estamos siguiendo el convenio habitual

$$\sum_{k=p}^{q} \psi(k) = 0 \text{ si } q < p \tag{4.10}$$

donde $\psi(k)$ es cualquier función en la variable k. Así, por ejemplo, en el caso anterior

$$u_{1j} = a_{1j}$$

para $j = 1, 2, \ldots, n$.

b) $\boxed{i > j}$ Ahora (4.9) se convierte en

$$a_{ij} = \sum_{k=1}^{j} l_{ik}u_{kj} = \sum_{k=1}^{j-1} l_{ik}u_{kj} + l_{ij}u_{jj}$$

para $1 \leq j < i \leq n$, de donde se obtiene que

$$l_{ij} = \frac{1}{u_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj} \right)$$

para $i = j+1, j+2, \ldots, n$. Cambiando i por j en la expresión anterior y utilizando el convenio antes descrito se llega a que

$$l_{ji} = \frac{1}{u_{ii}} \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk}u_{ki} \right)$$

para $j = i+1, i+2, \ldots, n$.

De esta forma, la manera de obtener las matrices L y U es la siguiente: para cada $i \in \{1, 2, \ldots, n\}$ se calculan

$$\boxed{\begin{array}{ll} u_{ij} = a_{ij} - \displaystyle\sum_{k=1}^{i-1} l_{ik} u_{kj}, & j = i, i+1, \ldots, n \\ l_{ji} = \dfrac{1}{u_{ii}} \left(a_{ji} - \displaystyle\sum_{k=1}^{i-1} l_{jk} u_{ki} \right), & j = i+1, i+2, \ldots, n \end{array}}$$

Nótese que las fórmulas recursivas anteriores están bien definidas en el sentido de que para cada valor de $i \in \{1, 2, \ldots, n\}$ se calcula la fila i–ésima de U y, a continuación, la columna i–ésima de L, utilizando tan sólo elementos de U y L pertenecientes a las $i-1$ primeras filas y columnas, respectivamente. En el caso de que la matriz A no admita factorización LU se obtendrá, para algún índice $k \in \{1, 2, \ldots, n\}$, que $u_{kk} = 0$. □

Ejemplo 4.3. Efectuemos la factorización LU para resolver el sistema lineal

$$\begin{pmatrix} 1 & 2 & 1 & 3 \\ 1 & 1 & 1 & 4 \\ 2 & 1 & 4 & 10 \\ -1 & -3 & 7 & 5 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 45 \\ 48 \\ 101 \\ -4 \end{pmatrix}.$$

Aplicando las fórmulas anteriores, obtenemos:

$$\left\{ \begin{array}{l} \left\{ \begin{array}{lll} u_{1j} = a_{1j},\ j = 1,2,3,4 & \Rightarrow & u_{11} = 1,\ u_{12} = 2,\ u_{13} = 1,\ u_{14} = 3 \\ l_{i1} = \dfrac{a_{i1}}{u_{11}} = a_{i1},\ i = 1,2,3,4 & \Rightarrow & l_{11} = 1,\ l_{21} = 1,\ l_{31} = 2,\ l_{41} = -1 \end{array} \right. \\ \left\{ \begin{array}{lll} u_{2j} = a_{2j} - l_{21} u_{1j},\ j = 2,3,4 & \Rightarrow & u_{22} = -1,\ u_{23} = 0,\ u_{24} = 1 \\ l_{i2} = \dfrac{a_{i2} - l_{i1} u_{12}}{u_{22}},\ i = 2,3,4 & \Rightarrow & l_{22} = 1,\ l_{32} = 3,\ l_{42} = 1 \end{array} \right. \\ \left\{ \begin{array}{lll} u_{3j} = a_{3j} - l_{31} u_{1j} - l_{32} u_{2j},\ j = 3,4 & \Rightarrow & u_{33} = 2,\ u_{34} = 1 \\ l_{i3} = \dfrac{a_{i3} - l_{i1} u_{13} - l_{i2} u_{23}}{u_{33}},\ i = 3,4 & \Rightarrow & l_{33} = 1,\ l_{43} = 4 \end{array} \right. \\ u_{44} = a_{44} - l_{41} u_{14} - l_{42} u_{24} - l_{43} u_{34} \quad \Rightarrow \quad u_{44} = 3 \end{array} \right.$$

Por tanto,

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 3 & 1 & 0 \\ -1 & 1 & 4 & 1 \end{pmatrix} \text{ y } U = \begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

De esta forma, resolviendo por el método de remonte los sistemas triangulares

$$Lw = b \text{ y } Uu = w,$$

obtenemos

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 3 & 1 & 0 \\ -1 & 1 & 4 & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix} = \begin{pmatrix} 45 \\ 48 \\ 101 \\ -4 \end{pmatrix} \Rightarrow w = \begin{pmatrix} 45 \\ 3 \\ 2 \\ 30 \end{pmatrix}$$

y

$$\begin{pmatrix} 1 & 2 & 1 & 3 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 45 \\ 3 \\ 2 \\ 30 \end{pmatrix} \Rightarrow u = \begin{pmatrix} 5 \\ 7 \\ -4 \\ 10 \end{pmatrix}. \quad \square$$

Observación 4.11. La factorización LU preserva la estructura de matrices banda, es decir, si $a_{ij} = 0$ para $|i-j| \geq p$ entonces $l_{ij} = 0$ para $i - j \geq p$ y $u_{ij} = 0$ para $j - i \geq p$ (véase el problema 4.8). $\square$

4.5. Método de Cholesky

Consideremos ahora el sistema lineal $Au = b$ donde $A \in \mathcal{M}_n$ es una matriz simétrica y definida positiva. Como se ha comentado en la observación 4.9, este tipo de matrices admiten factorización LU. Vamos a ver que, de hecho, se puede obtener una factorización $A = BC$ siendo $B \in \mathcal{M}_n$ una matriz triangular inferior y $C = B^{\mathrm{T}}$.

Teorema 4.4. *Si $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ es simétrica y definida positiva, entonces existe una matriz real triangular inferior $B = (b_{ij})_{i,j=1}^n \in \mathcal{M}_n$ tal que $A = BB^{\mathrm{T}}$. Además, si se elige que*

$$b_{ii} > 0$$

para $i = 1, 2, \ldots, n$, entonces la factorización $A = BB^{\mathrm{T}}$ es única.

Demostración. Como las n cajas principales de la matriz A son simétricas y definidas positivas (véase el problema 2.16), todos los menores principales de A, δ_k, son positivos y, en particular, no nulos. Así, por el teorema 4.3, existe una

única matriz triangular inferior $L = (l_{ij})_{i,j=1}^{n}$ con $l_{ii} = 1$, $i = 1, 2, \ldots, n$, y una única matriz triangular superior $U = (u_{ij})_{i,j=1}^{n}$ tales que $A = LU$, es decir,

$$A = LU = \begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ \cdots & \cdots & \ddots & \\ l_{n1} & \cdots & l_{n,n-1} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ & & \ddots & \cdots \\ & & & u_{nn} \end{pmatrix}.$$

Por tanto, descomponiendo en cajas de tamaño $n_1 = k$ y $n_2 = n - k$, se tiene que

$$\prod_{i=1}^{k} u_{ii} = \delta_k > 0$$

para todo índice $k = 1, 2, \ldots, n$, de donde, comenzando con $k = 1$ y continuando con $k = 2, 3, \ldots, n$, se sigue que

$$u_{ii} > 0$$

para $i = 1, 2, \ldots, n$. Si intercalamos la matriz real diagonal e inversible

$$D = \text{diag } (\sqrt{u_{11}}, \sqrt{u_{22}}, \ldots, \sqrt{u_{nn}}) \tag{4.11}$$

en la factorización LU, como

$$D^{-1} = \text{diag } \left(\frac{1}{\sqrt{u_{11}}}, \frac{1}{\sqrt{u_{22}}}, \ldots, \frac{1}{\sqrt{u_{nn}}} \right),$$

se obtiene que

$$\begin{aligned} A = LU &= (LD)(D^{-1}U) \\ &= \begin{pmatrix} \sqrt{u_{11}} & & & \\ \mu_{21} & \sqrt{u_{22}} & & \\ \cdots & \cdots & \ddots & \\ \mu_{n1} & \cdots & \mu_{n,n-1} & \sqrt{u_{nn}} \end{pmatrix} \begin{pmatrix} \sqrt{u_{11}} & \nu_{12} & \cdots & \nu_{1n} \\ & \sqrt{u_{22}} & \cdots & \nu_{2n} \\ & & \ddots & \cdots \\ & & & \sqrt{u_{nn}} \end{pmatrix}. \end{aligned}$$

Denotando

$$B = LD \text{ y } C = D^{-1}U,$$

se verifica que $A = LU = BC$. Por ser la matriz A simétrica, $A = A^{\mathrm{T}}$; por tanto,

$$BC = A = A^{\mathrm{T}} = (BC)^{\mathrm{T}} = C^{\mathrm{T}}B^{\mathrm{T}}. \tag{4.12}$$

Como

$$\det(B) = \det\left(B^{\mathrm{T}}\right) = \prod_{i=1}^{n} \sqrt{u_{ii}} > 0$$

las matrices B y B^{T} son inversibles. Además, de la relación (4.12) se tiene que

$$C(B^{\mathrm{T}})^{-1} = B^{-1}C^{\mathrm{T}}. \tag{4.13}$$

Ahora bien, como

$$C(B^{\mathrm{T}})^{-1} = \begin{pmatrix} 1 & k_{12} & \cdots & k_{1n} \\ & 1 & \cdots & k_{2n} \\ & & \ddots & \cdots \\ & & & 1 \end{pmatrix} \text{ y } B^{-1}C^{\mathrm{T}} = \begin{pmatrix} 1 & & & \\ m_{21} & 1 & & \\ \cdots & \cdots & \ddots & \\ m_{n1} & \cdots & m_{n,n-1} & 1 \end{pmatrix},$$

la relación (4.13) hace que

$$C(B^{\mathrm{T}})^{-1} = B^{-1}C^{\mathrm{T}} = I$$

por lo que $C = B^{\mathrm{T}}$ y, por tanto, $A = BC = BB^{\mathrm{T}}$. Nótese que, en este caso, los elementos diagonales de la matriz B son

$$b_{ii} = \sqrt{u_{ii}} > 0$$

para $i = 1, 2, \ldots, n$.

Para demostrar la unicidad de la factorización $A = BB^{\mathrm{T}}$ cuando los elementos diagonales de la matriz triangular inferior B son positivos, supongamos que hubiera dos descomposiciones de la forma

$$B_1B_1^{\mathrm{T}} = A = B_2B_2^{\mathrm{T}} \tag{4.14}$$

donde

$$(B_1)_{ii} > 0 \text{ y } (B_2)_{ii} > 0 \tag{4.15}$$

para $i = 1, 2, \ldots, n$. La propiedad (4.15) hace que se tenga

$$\det(B_k) = \det\left((B_k)^{\mathrm{T}}\right) = \prod_{i=1}^{n} (B_k)_{ii} > 0,$$

lo que determina que las matrices B_k y $(B_k)^{\mathrm{T}}$ sean inversibles para $k = 1, 2$. Por tanto, a partir de (4.14), se verifica que

$$(B_1)^{\mathrm{T}}\left((B_2)^{\mathrm{T}}\right)^{-1} = (B_1)^{-1}B_2. \tag{4.16}$$

Como B_1 y B_2 son matrices triangulares inferiores, las matrices $(B_1)^{\mathrm{T}}$ y $\left((B_2)^{\mathrm{T}}\right)^{-1}$ son triangulares superiores y, por tanto,

$$(B_1)^{\mathrm{T}}\left((B_2)^{\mathrm{T}}\right)^{-1} = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ & \mu_{22} & \cdots & \mu_{2n} \\ & & \ddots & \cdots \\ & & & \mu_{nn} \end{pmatrix}.$$

Por otra parte las matrices $(B_1)^{-1}$ y B_2 son triangulares inferiores, por lo que

$$(B_1)^{-1}B_2 = \begin{pmatrix} \nu_{11} & & & \\ \nu_{21} & \nu_{22} & & \\ \cdots & \cdots & \ddots & \\ \nu_{n1} & \cdots & \nu_{n,n-1} & \nu_{nn} \end{pmatrix}.$$

De esta forma, la relación (4.16) determina que las matrices $(B_1)^{\mathrm{T}} \left((B_2)^{\mathrm{T}}\right)^{-1}$ y $(B_1)^{-1}B_2$ son diagonales, es decir,

$$(B_1)^{\mathrm{T}} \left((B_2)^{\mathrm{T}}\right)^{-1} = (B_1)^{-1}B_2 = D \tag{4.17}$$

siendo D una matriz diagonal, es decir,

$$D = \operatorname{diag}\left(\frac{(B_1)_{ii}}{(B_2)_{ii}}\right) = \operatorname{diag}\left(\frac{(B_2)_{ii}}{(B_1)_{ii}}\right).$$

Por tanto,

$$((B_1)_{ii})^2 = ((B_2)_{ii})^2$$

para $i = 1, 2, \ldots, n$ y, consecuentemente, a partir de (4.15), se tiene que

$$(B_1)_{ii} = (B_2)_{ii}$$

para $i = 1, 2, \ldots, n$. De esta forma,

$$D = \operatorname{diag}(1, 1, \ldots, 1) = I,$$

lo que, junto a (4.17), hace que se tenga $B_1 = B_2$. □

Observación 4.12.

1. Si se desean resolver varios sistemas lineales con la misma matriz $A = BB^{\mathrm{T}}$, de nuevo la estrategia es, una vez calculada la matriz B, resolver los sistemas triangulares
$$\begin{cases} Bw = b \\ B^{\mathrm{T}}u = w \end{cases}$$
con cada uno de los segundos miembros b requeridos.

2. Basta prefijar el signo de cada uno de los elementos de la diagonal de la matriz B para tener unicidad de la factorización de Cholesky.

3. Si $A = BB^{\mathrm{T}}$ es la factorización de Cholesky de una matriz $A \in \mathcal{M}_n$ simétrica y definida positiva, entonces
$$\det(A) = b_{11}^2 b_{22}^2 \cdots b_{nn}^2.$$

4. El teorema 4.4 puede generalizarse a matrices complejas. Concretamente, puede demostrarse que si $A = (a_{ij})_{i,j=1}^{n} \in \mathcal{M}_n$ es hermítica y definida positiva entonces existe una matriz triangular inferior $B = (b_{ij})_{i,j=1}^{n} \in \mathcal{M}_n$ tal que $A = BB^*$. Además, si se elige que

$$b_{ii} > 0$$

para $i = 1, 2, \ldots, n$, entonces la factorización $A = BB^*$ es única. □

Observación 4.13. Consideremos la factorización de Cholesky $A = BB^{\mathrm{T}}$ donde B es una matriz real triangular inferior con elementos diagonales

$$b_{ii} > 0 \tag{4.18}$$

para $i = 1, 2, \ldots, n$. De la relación $A = BB^{\mathrm{T}}$ se obtiene que

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

para $i, j = 1, 2, \ldots, n$, ya que

$$b_{pq} = 0 \text{ si } 1 \leq p < q \leq n.$$

Fijando nuestra atención en el caso en que $i \leq j$ se tiene que

$$a_{ij} = \sum_{k=1}^{i} b_{ik} b_{jk} = \sum_{k=1}^{i-1} b_{ik} b_{jk} + b_{ii} b_{ji} \tag{4.19}$$

para $1 \leq i \leq j \leq n$. Por tanto, para cada $i \in \{1, 2, \ldots, n\}$ podemos escribir

$$a_{ii} = \sum_{k=1}^{i-1} b_{ik} b_{ik} + b_{ii} b_{ii} = \sum_{k=1}^{i-1} b_{ik}^2 + b_{ii}^2,$$

igualdad que conduce a

$$b_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} b_{ik}^2}$$

donde nuevamente seguimos el criterio establecido en (4.10); de esta forma

$$b_{11} = \sqrt{a_{11}}.$$

Por otra parte, la relación (4.19) determina

$$b_{ji} = \frac{1}{b_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} b_{ik} b_{jk} \right)$$

para $j = i+1, i+2, \ldots, n$. Así pues, la forma de obtener la matriz B es la siguiente: para cada $i \in \{1, 2, \ldots, n\}$ se calculan

$$\boxed{\begin{aligned} b_{ii} &= \sqrt{a_{ii} - \sum_{k=1}^{i-1} b_{ik}^2} \\ b_{ji} &= \frac{1}{b_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} b_{ik} b_{jk} \right), \quad j = i+1, i+2, \ldots, n \end{aligned}}$$

1. Las fórmulas recursivas anteriores están bien definidas, pues para cada valor de $i \in \{1, 2, \ldots, n\}$ se calcula la columna i–ésima de B utilizando tan sólo elementos de la matriz B situados en las $i-1$ primeras columnas de B.

2. En el caso de que la matriz A sea simétrica pero no admita factorización de Cholesky se obtendrá, para algún índice $i \in \{1, 2, \ldots, n\}$, que

$$a_{ii} - \sum_{k=1}^{i-1} b_{ik}^2 \leq 0.$$

3. Cuando la matriz $A \in \mathcal{M}_n$ es hermítica y definida positiva entonces A admite factorización de Cholesky de la forma $A = BB^*$, donde $B \in \mathcal{M}_n$ es una matriz triangular inferior (véase la observación 4.12). En este caso, los elementos de la matriz B se obtienen recursivamente a partir de las relaciones

$$\boxed{\begin{aligned} b_{ii} &= \sqrt{a_{ii} - \sum_{k=1}^{i-1} |b_{ik}|^2} \\ b_{ji} &= \frac{1}{b_{ii}} \left(\overline{a_{ij}} - \sum_{k=1}^{i-1} \overline{b_{ik}} b_{jk} \right), \quad j = i+1, i+2, \ldots, n \end{aligned}} \quad \square$$

Observación 4.14. Al igual que ocurre con la factorización LU, la factorización de Cholesky preserva la estructura de matrices banda (véase el problema 4.8). $\square$

Observación 4.15. Contemos el número de operaciones elementales efectuadas en el método de Cholesky:

1. Factorización. El cálculo de la matriz B mediante las fórmulas dadas necesita:

$$\begin{cases} n & \text{raíces} \\ \displaystyle\sum_{k=1}^{n-1}(n-k) = \frac{n(n-1)}{2} & \text{divisiones} \\ \displaystyle\sum_{k=1}^{n-1}k(n-k) = \frac{n^3-n}{6} & \text{sumas} \\ \displaystyle\sum_{k=1}^{n-1}k(n-k) = \frac{n^3-n}{6} & \text{productos} \end{cases}$$

2. Resolución de los sistemas lineales $Bw = b$ y $B^{\mathrm{T}}u = w$. Como se ha visto, estas resoluciones necesitan:

$$\begin{cases} 2\dfrac{n(n-1)}{2} = n(n-1) & \text{sumas} \\ 2\dfrac{n(n-1)}{2} = n(n-1) & \text{productos} \\ 2n & \text{divisiones} \end{cases}$$

Es decir, el método de Cholesky necesita

$$\frac{n^3-n}{3} + \frac{5}{2}n(n-1) + 3n = \frac{(2n^2+15n+1)n}{6} \sim \frac{n^3}{3}$$

operaciones, lo que constituye del orden de la mitad de operaciones que el método de Gauss.

Ejemplo 4.4. Vamos a aplicar el método de Cholesky para resolver el sistema lineal

$$\begin{pmatrix} 1 & -1 & 1 & 0 \\ -1 & 2 & -1 & 2 \\ 1 & -1 & 5 & 2 \\ 0 & 2 & 2 & 6 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 4 \\ -3 \\ 16 \\ 8 \end{pmatrix}.$$

Puesto que A es una matriz simétrica y definida positiva, aplicando las fórmulas

anteriores para el cálculo de la matriz B obtenemos:

$$\left\{\begin{array}{l}
\left\{\begin{array}{l} b_{11}=\sqrt{a_{11}}=1 \\ b_{j1}=\dfrac{a_{1j}}{b_{11}}=a_{1j},\ j=2,3,4 \quad \Rightarrow \quad b_{21}=-1,\ b_{31}=1,\ b_{41}=0 \end{array}\right. \\
\left\{\begin{array}{l} b_{22}=\sqrt{a_{22}-b_{21}^2}=\sqrt{2-1}=1 \\ b_{32}=\dfrac{a_{23}-b_{21}b_{31}}{b_{22}}=\dfrac{-1+1\cdot 1}{1}=0, \\ b_{42}=\dfrac{a_{24}-b_{21}b_{41}}{b_{22}}=\dfrac{2+1\cdot 0}{1}=2 \end{array}\right. \\
\left\{\begin{array}{l} b_{33}=\sqrt{a_{33}-b_{31}^2-b_{32}^2}=\sqrt{5-1-0}=2 \\ b_{43}=\dfrac{a_{34}-b_{31}b_{41}-b_{32}b_{42}}{b_{33}}=\dfrac{2-1\cdot 0-0\cdot 2}{2}=1 \end{array}\right. \\
b_{44}=\sqrt{a_{44}-b_{41}^2-b_{42}^2-b_{43}^2}=\sqrt{6-0-4-1}=1
\end{array}\right.$$

Por tanto,

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

Resolviendo por el método de remonte los sistemas triangulares

$$Bw=b \ \text{ y } \ B^{\mathrm{T}}u=w$$

se obtiene que

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 0 & 2 & 1 & 1 \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{pmatrix}=\begin{pmatrix} 4 \\ -3 \\ 16 \\ 8 \end{pmatrix} \Rightarrow w=\begin{pmatrix} 4 \\ 1 \\ 6 \\ 0 \end{pmatrix}$$

y

$$\begin{pmatrix} 1 & -1 & 1 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix}=\begin{pmatrix} 4 \\ 1 \\ 6 \\ 0 \end{pmatrix} \Rightarrow u=\begin{pmatrix} 2 \\ 1 \\ 3 \\ 0 \end{pmatrix}.$$

Nótese que, a partir de B, podemos calcular el determinante de A; en este caso

$$\det(A)=4. \quad \square$$

4.6. Problemas

4.6.1. Problemas resueltos

4.1. Método de Gauss–Jordan para el cálculo de la inversa de una matriz. Consideremos una matriz $A \in \mathcal{M}_n$ inversible.

a) Adaptar la demostración del método de Gauss para probar que existe una matriz $\widetilde{M}$ de manera que $\widetilde{M}A$ sea la identidad.

b) Explicar cómo se puede utilizar el resultado del apartado anterior (*método de Gauss–Jordan*) para el cálculo de la inversa de la matriz A.

SOLUCIÓN.

a) Siguiendo el método de elimininación de Gauss, el resultado de la etapa $k-1$ es una matriz de la forma

$$\begin{aligned}\widetilde{A}_k &= \widetilde{E}_{k-1}\widetilde{P}_{k-1}\widetilde{E}_{k-2}\widetilde{P}_{k-2}\cdots\widetilde{E}_1\widetilde{P}_1 A \\ &= \begin{pmatrix} 1 & & & & \alpha_{1k}^1 & \cdots & \alpha_{1n}^1 \\ & 1 & & & \alpha_{2k}^2 & \cdots & \alpha_{2n}^2 \\ & & \ddots & & \cdots & \cdots & \cdots \\ & & & 1 & \alpha_{k-1,k}^{k-1} & \cdots & \alpha_{k-1,n}^{k-1} \\ & & & & a_{kk}^k & \cdots & a_{kn}^k \\ & & & & \cdots & \cdots & \cdots \\ & & & & a_{nk}^k & \cdots & a_{nn}^k \end{pmatrix}.\end{aligned}$$

Al igual que en el método de Gauss, la matriz $\widetilde{A}_k$ es inversible, por lo que existe un índice $i \in \{k, k+1, \ldots, n\}$ para el cual $a_{ik}^k \neq 0$, elemento que tomamos como pivote. A continuación multiplicamos a la izquierda la matriz $\widetilde{A}_k$ por una matriz $\widetilde{P}_k$ de permutación. Si

$$\widetilde{P}_k\widetilde{A}_k = \begin{pmatrix} 1 & & & & \alpha_{1k}^1 & \cdots & \alpha_{1n}^1 \\ & 1 & & & \alpha_{2k}^2 & \cdots & \alpha_{2n}^2 \\ & & \ddots & & \cdots & \cdots & \cdots \\ & & & 1 & \alpha_{k-1,k}^{k-1} & \cdots & \alpha_{k-1,n}^{k-1} \\ & & & & \alpha_{kk}^k & \cdots & \alpha_{kn}^k \\ & & & & \cdots & \cdots & \cdots \\ & & & & \alpha_{nk}^k & \cdots & \alpha_{nn}^k \end{pmatrix}$$

entonces $\alpha_{kk}^k = a_{ik}^k \neq 0$. La matriz $\widetilde{A}_{k+1}$ se obtiene multiplicando la matriz anterior a la izquierda por

$$\widetilde{E}_k = \begin{pmatrix} 1 & & & -\dfrac{\alpha_{1k}^1}{\alpha_{kk}^k} & & \\ & 1 & & -\dfrac{\alpha_{2k}^2}{\alpha_{kk}^k} & & \\ & & \ddots & \cdots & & \\ & & & \dfrac{1}{\alpha_{kk}^k} & & \\ & & & -\dfrac{\alpha_{k+1,k}^k}{\alpha_{kk}^k} & 1 & \\ & & & \cdots & & \ddots \\ & & & -\dfrac{\alpha_{nk}^k}{\alpha_{kk}^k} & & 1 \end{pmatrix}.$$

Después de la n–ésima etapa, la matriz $\widetilde{A}_{n+1} = \widetilde{M}A$ es la identidad, donde

$$\widetilde{M} = \widetilde{E}_n\widetilde{P}_n\widetilde{E}_{n-1}\widetilde{P}_{n-1}\cdots\widetilde{E}_2\widetilde{P}_2\widetilde{E}_1\widetilde{P}_1.$$

b) Basta ampliar la matriz A mediante la identidad I para construir la matriz

$$B = (A|I) \in \mathcal{M}_{n\times 2n}.$$

Aplicando a esta matriz B las transformaciones recogidas en $\widetilde{M}$ se tiene que

$$\widetilde{M}B = (\widetilde{M}A|\widetilde{M}) = (I|\widetilde{M})$$

donde, obviamente, $\widetilde{M} = A^{-1}$. □

4.2. Algoritmo para la resolución de sistemas tridiagonales. Se considera el sistema lineal $Ax = d$ para una matriz tridiagonal

$$A = \begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_n & b_n \end{pmatrix}. \tag{4.20}$$

a) Llamando $\delta_0 = 1$ y δ_k al menor principal de orden k de A, $k = 1, 2, \ldots, n$, probar, desarrollando el determinante por la última columna, que

$$\delta_k = b_k\delta_{k-1} - a_kc_{k-1}\delta_{k-2}$$

para $k = 2, 3, \ldots, n$.

b) Se definen las sucesiones

$$\begin{cases} m_1 = b_1 \\ m_k = b_k - \dfrac{c_{k-1}}{m_{k-1}} a_k, \ k = 2, 3, \ldots, n \end{cases}$$

y

$$\begin{cases} g_1 = \dfrac{d_1}{m_1} \\ g_k = \dfrac{d_k - g_{k-1} a_k}{m_k}, \ k = 2, 3, \ldots, n. \end{cases}$$

Probar que

$$m_k = \frac{\delta_k}{\delta_{k-1}}$$

y deducir que si

$$\delta_k \neq 0$$

para $k = 1, 2, \ldots, n$, entonces

$$m_k \neq 0$$

y, por tanto, los números m_k y g_k están bien definidos.

c) Demostrar que la solución del sistema $Ax = d$ viene dada por

$$x_n = g_n \ \text{ y } \ x_k = g_k - \frac{c_k}{m_k} x_{k+1} \tag{4.21}$$

para $k = n-1, n-2, \ldots, 1$.

d) Comprobar que

$$\begin{pmatrix} 1 & & & & \\ \frac{a_2}{m_1} & 1 & & & \\ & \ddots & \ddots & & \\ & & \frac{a_{n-1}}{m_{n-2}} & 1 & \\ & & & \frac{a_n}{m_{n-1}} & 1 \end{pmatrix} \begin{pmatrix} m_1 & c_1 & & & \\ & m_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & m_{n-1} & c_{n-1} \\ & & & & m_n \end{pmatrix}$$

es la factorización LU de la matriz A.

e) Determinar el número de operaciones que requiere este método y compararlo con el número de operaciones necesarias en el método de eliminación gaussiana (téngase en cuenta que los cocientes $\dfrac{c_k}{m_k}$ pueden calcularse una sola vez).

SOLUCIÓN.

a) Vamos a proceder desarrollando el determinante δ_k por los elementos de la última columna:

$$\begin{aligned}
\delta_k &= \det \begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{k-1} & b_{k-1} & c_{k-1} \\ & & & a_k & b_k \end{pmatrix} \\
&= b_k \det \begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{k-2} & b_{k-2} & c_{k-2} \\ & & & a_{k-1} & b_{k-1} \end{pmatrix} \\
&\quad - c_{k-1} \det \begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{k-2} & b_{k-2} & c_{k-2} \\ & & 0 & 0 & a_k \end{pmatrix} \\
&= b_k \delta_{k-1} - c_{k-1} a_k \det \begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{k-3} & b_{k-3} & c_{k-3} \\ & & & a_{k-2} & b_{k-2} \end{pmatrix} \\
&= b_k \delta_{k-1} - c_{k-1} a_k \delta_{k-2}.
\end{aligned}$$

b) Procedemos por inducción:

i) Para $k = 1$ el resultado es obvio.

ii) Suponemos cierto el resultado para $k-1$ y, bajo esta hipótesis, demostramos que también es cierto para k. En efecto, por el apartado *a*) se tiene que

$$\frac{\delta_k}{\delta_{k-1}} = \frac{b_k \delta_{k-1} - a_k c_{k-1} \delta_{k-2}}{\delta_{k-1}} = b_k - a_k c_{k-1} \frac{\delta_{k-2}}{\delta_{k-1}}.$$

Ahora bien, por la hipótesis de inducción

$$\frac{\delta_{k-1}}{\delta_{k-2}} = m_{k-1},$$

por tanto,

$$\frac{\delta_k}{\delta_{k-1}} = b_k - a_k c_{k-1}\frac{1}{m_{k-1}} = m_k.$$

c) En primer lugar, con el convenio $x_0 = x_{n+1} = 0$, podemos escribir el sistema anterior en la forma

$$a_k x_{k-1} + b_k x_k + c_k x_{k+1} = d_k$$

para $k = 1, 2, \ldots, n$. Mostramos, nuevamente por inducción, que (4.21) determina la solución del sistema lineal $Ax = d$:

i) Despejando x_1 en la primera ecuación del sistema se obtiene que

$$x_1 = \frac{d_1 - c_1 x_2}{b_1} = \frac{m_1 g_1 - c_1 x_2}{m_1} = g_1 - \frac{c_1}{m_1} x_2.$$

ii) Dado $k \in \{2, 3, \ldots, n\}$, supuesto cierto el resultado para $k-1$, usando la k–ésima ecuación podemos escribir

$$\begin{aligned} b_k x_k &= d_k - a_k x_{k-1} - c_k x_{k+1} \\ &= d_k - a_k \left(g_{k-1} - \frac{c_{k-1}}{m_{k-1}} x_k \right) - c_k x_{k+1}, \end{aligned}$$

gracias a la hipótesis de inducción. Por tanto,

$$\begin{aligned} b_k x_k &= d_k - g_{k-1} a_k + \frac{c_{k-1}}{m_{k-1}} a_k x_k - c_k x_{k+1} \\ &= m_k g_k + \frac{c_{k-1}}{m_{k-1}} a_k x_k - c_k x_{k+1}. \end{aligned}$$

De esta forma,

$$m_k x_k = \left(b_k - \frac{c_{k-1}}{m_{k-1}} a_k \right) x_k = m_k g_k - c_k x_{k+1}$$

de donde

$$x_k = g_k - \frac{c_k}{m_k} x_{k+1}.$$

La manera de proceder es la siguiente: se calculan recursivamente los valores

$$\{m_1, g_1, m_2, g_2, \ldots, m_n, g_n\}$$

y con ellos se obtienen

$$\{x_n, x_{n-1}, \ldots, x_1\}.$$

Obsérvese que la idea subyacente es despejar cada incógnita en la ecuación en la que ésta aparece en el valor central y sustituir en la ecuación siguiente, obteniendo así una fórmula regresiva para el cálculo de las incógnitas.

d) A partir de la definición de los elementos m_k es inmediato comprobar que $A = LU$, siendo

$$L = \begin{pmatrix} 1 & & & & \\ \frac{a_2}{m_1} & 1 & & & \\ & \ddots & \ddots & & \\ & & \frac{a_{n-1}}{m_{n-2}} & 1 & \\ & & & \frac{a_n}{m_{n-1}} & 1 \end{pmatrix} \text{ y } U = \begin{pmatrix} m_1 & c_1 & & & \\ & m_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & m_{n-1} & c_{n-1} \\ & & & & m_n \end{pmatrix},$$

por lo que ésta es la factorización LU de A.

Observación: en realidad, las secuencias g_k y x_k no son más que las soluciones de dos sistemas triangulares correspondientes a una factorización LU de A, con la propiedad de que la matriz U tiene unos en su diagonal. En efecto, como $m_k \neq 0$ para $k = 1, 2, \ldots, n$, podemos considerar la matriz diagonal

$$D = \operatorname{diag}\left(\frac{1}{m_1}, \frac{1}{m_2}, \ldots, \frac{1}{m_n}\right)$$

que es inversible y cuya inversa es

$$D^{-1} = \operatorname{diag}(m_1, m_2, \ldots, m_n).$$

Mediante esta matriz D escribimos

$$A = LU = BC$$

siendo

$$B = LD^{-1} = \begin{pmatrix} m_1 & & & & \\ a_2 & m_2 & & & \\ & \ddots & \ddots & & \\ & & a_{n-1} & m_{n-1} & \\ & & & a_n & m_n \end{pmatrix}$$

y

$$C = DU = \begin{pmatrix} 1 & \frac{c_1}{m_1} & & & \\ & 1 & \frac{c_2}{m_2} & & \\ & & \ddots & \ddots & \\ & & & 1 & \frac{c_{n-1}}{m_{n-1}} \\ & & & & 1 \end{pmatrix}.$$

Así pues,

$$Ax = d \Leftrightarrow BCx = d \Leftrightarrow \begin{cases} Bg = d \\ Cx = g. \end{cases}$$

Resolviendo por el método de remonte los dos sistemas triangulares anteriores se obtiene que

$$\begin{cases} g_1 = \dfrac{d_1}{m_1} \\ g_k = \dfrac{d_k - g_{k-1}a_k}{m_k},\ k = 2, 3, \ldots, n \end{cases}$$

y

$$\begin{cases} x_n = g_n \\ x_k = g_k - \dfrac{c_k}{m_k} x_{k+1},\ k = n-1, n-2, \ldots, 1. \end{cases}$$

Por tanto, $\{g_1, g_2, \ldots, g_n\}$ y $\{x_n, x_{n-1}, \ldots, x_1\}$ son las soluciones que se obtienen resolviendo por el método de remonte los sistemas triangulares $Bg = d$ y $Cx = g$, respectivamente.

e) Contabilicemos el número de operaciones necesarias en este método:

$$\begin{cases} \dfrac{c_k}{m_k} & \longrightarrow & n-1 \text{ cocientes} \\ m_k & \longrightarrow & n-1 \text{ multiplicaciones y } n-1 \text{ restas} \\ g_k & \longrightarrow & n \text{ divisiones, } n-1 \text{ multiplicaciones y } n-1 \text{ restas} \\ x_k & \longrightarrow & n-1 \text{ multiplicaciones y } n-1 \text{ restas.} \end{cases}$$

Luego el número de operaciones que requiere este método es $8n - 7$ frente a las

$$\frac{n(4n^2 + 9n - 7)}{6}$$

que se necesitan en el método de eliminación gaussiana. □

4.3. Se considera el sistema lineal $Ax = d$ donde A es una matriz tridiagonal de la forma (4.20) de diagonal estrictamente dominante. Demostrar que

$$||x||_\infty \leq c(A)\, ||d||_\infty$$

siendo

$$c(A) = \max_{1 \leq i \leq n} \left\{ \frac{1}{|b_i| - |a_i| - |c_i|} \right\} \qquad (a_1 = c_n = 0).$$

SOLUCIÓN. En primer lugar, por ser A de diagonal estrictamente dominante, se verifica que

$$|b_i| > |a_i| + |c_i|$$

para $i = 1, 2, \ldots, n$, lo que hace que la constante $c(A)$ esté bien definida.

Por otra parte, la relación $Ax = d$ determina que

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i$$

para $i = 1, 2, \ldots, n$, siendo $x_0 = x_{n+1} = 0$. Por tanto, para todo $i \in \{1, 2, \ldots, n\}$ se verifica que

$$|b_i||x_i| \leq |a_i||x_{i-1}| + |c_i||x_{i+1}| + |d_i| \leq (|a_i| + |c_i|)\,||x||_\infty + ||d||_\infty . \qquad (4.22)$$

Consideramos un índice $i_0 \in \{1, 2, \ldots, n\}$ verificando

$$||x||_\infty = \max_{1 \leq i \leq n} |x_i| = |x_{i_0}|.$$

Sustituyendo $i = i_0$ en la expresión (4.22) se obtiene que

$$|b_{i_0}|\,||x||_\infty \leq (|a_{i_0}| + |c_{i_0}|)\,||x||_\infty + ||d||_\infty$$

de donde se concluye que

$$||x||_\infty \leq \frac{||d||_\infty}{|b_{i_0}| - |a_{i_0}| - |c_{i_0}|} \leq \max_{1 \leq i \leq n} \left\{ \frac{1}{|b_i| - |a_i| - |c_i|} \right\} ||d||_\infty . \quad \square$$

4.4. Sea $A \in \mathcal{M}_n$ una matriz cuyos menores principales son todos no nulos.

a) Probar que A se puede factorizar en la forma $A = LDR$ siendo $L = (l_{ij})_{i,j=1}^n$ una matriz triangular inferior, $R = (r_{ij})_{i,j=1}^n$ triangular superior con

$$l_{ii} = r_{ii} = 1$$

para $i = 1, 2, \ldots, n$, y D diagonal. Encontrar fórmulas para los elementos de L, D y R.

b) Demostrar que si A es simétrica entonces $R = L^{\mathrm{T}}$. Adaptar las fórmulas anteriores para el caso de que A sea simétrica.

c) ¿Cómo se usaría la factorización dada en *a*) —o *b*), en el caso de que A sea simétrica— para resolver el sistema $Au = b$?

Solución.

a) Por ser todos los menores principales de A no nulos, la matriz A admite una única factorización LU (véase el teorema 4.3). Más concretamente, existe una única matriz triangular inferior $L = (l_{ij})_{i,j=1}^n$ con $l_{ii} = 1$, $i = 1, 2, \ldots, n$, y existe una única matriz triangular superior U tal que $A = LU$. Basta considerar las matrices

$$D = \text{diag } (u_{11}, u_{22}, \ldots, u_{nn}) \;\text{ y }\; R = D^{-1}U$$

(nótese que D es inversible por serlo A y, por tanto, U). De esta forma, se tiene que

$$A = LU = LD(D^{-1}U) = LDR.$$

Además, como

$$D^{-1} = \text{diag}\left(\frac{1}{u_{11}}, \frac{1}{u_{22}}, \ldots, \frac{1}{u_{nn}}\right)$$

entonces

$$r_{ii} = 1$$

para $i = 1, 2, \ldots, n$. Obtengamos los elementos de las matrices L, D y R a partir de la factorización

$$A = LDR.$$

Claramente,

$$a_{ij} = \sum_{k=1}^{n} l_{ik} d_{kk} r_{kj} = \sum_{k=1}^{\min\{i,j\}} l_{ik} d_{kk} r_{kj}$$

para $i, j = 1, 2, \ldots, n$, puesto que

$$l_{ik} = 0 \text{ si } k > i$$

y

$$r_{kj} = 0 \text{ si } k > j.$$

Distinguimos dos casos:

i) $\boxed{i \leq j}$ En esta situación:

$$a_{ij} = \sum_{k=1}^{i} l_{ik} d_{kk} r_{kj} = \sum_{k=1}^{i-1} l_{ik} d_{kk} r_{kj} + d_{ii} r_{ij}$$

para $j = i, i+1, \ldots, n$, ya que $l_{ii} = 1$. Por tanto,

$$\begin{cases} a_{ii} = \displaystyle\sum_{k=1}^{i-1} l_{ik} d_{kk} r_{ki} + d_{ii} r_{ii} \\ a_{ij} = \displaystyle\sum_{k=1}^{i-1} l_{ik} d_{kk} r_{kj} + d_{ii} r_{ij},\ j = i+1, i+2, \ldots, n, \end{cases}$$

de donde se obtiene que

$$\begin{cases} d_{ii} = a_{ii} - \displaystyle\sum_{k=1}^{i-1} l_{ik} d_{kk} r_{ki} \\ r_{ij} = \dfrac{1}{d_{ii}} \left(a_{ij} - \displaystyle\sum_{k=1}^{i-1} l_{ik} d_{kk} r_{kj} \right),\ j = i+1, i+2, \ldots, n. \end{cases}$$

ii) $\boxed{i > j}$ Ahora

$$a_{ij} = \sum_{k=1}^{j} l_{ik} d_{kk} r_{kj} = \sum_{k=1}^{j-1} l_{ik} d_{kk} r_{kj} + l_{ij} d_{jj}$$

para $i = j+1, j+2, \ldots, n$, ya que $r_{jj} = 1$. De esta forma,

$$l_{ij} = \frac{1}{d_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} d_{kk} r_{kj} \right)$$

para $i = j+1, j+2, \ldots, n$. Cambiando el índice i por el índice j en la expresión anterior se tiene que

$$l_{ji} = \frac{1}{d_{ii}} \left(a_{ji} - \sum_{k=1}^{i-1} l_{jk} d_{kk} r_{ki} \right)$$

para $j = i+1, i+2, \ldots, n$.

Es decir, para cada $i \in \{1, 2, \ldots, n\}$ se calculan

$$\boxed{\begin{array}{l} d_{ii} = a_{ii} - \displaystyle\sum_{k=1}^{i-1} l_{ik} d_{kk} r_{ki} \\ \\ r_{ij} = \dfrac{1}{d_{ii}} \left(a_{ij} - \displaystyle\sum_{k=1}^{i-1} l_{ik} d_{kk} r_{kj} \right), \; j = i+1, i+2, \ldots, n \\ \\ l_{ji} = \dfrac{1}{d_{ii}} \left(a_{ji} - \displaystyle\sum_{k=1}^{i-1} l_{jk}\, d_{kk} r_{ki} \right), \; j = i+1, i+2, \ldots, n \end{array}} \tag{4.23}$$

b) En el caso de que la matriz A sea simétrica se verifica que

$$LDR = A = A^{\mathrm{T}} = R^{\mathrm{T}} D L^{\mathrm{T}}$$

y, por tanto,

$$DR \left(L^{\mathrm{T}} \right)^{-1} = L^{-1} R^{\mathrm{T}} D \tag{4.24}$$

siendo $DR \left(L^{\mathrm{T}} \right)^{-1}$ una matriz triangular superior con

$$\left(DR \left(L^{\mathrm{T}} \right)^{-1} \right)_{ii} = d_{ii}$$

para $i = 1, 2, \ldots, n$, y $L^{-1}R^{\mathrm{T}}D$ una matriz triangular inferior con

$$\left(L^{-1}R^{\mathrm{T}}D\right)_{ii} = d_{ii}$$

para $i = 1, 2, \ldots, n$. Por tanto, de la relación (4.24) se deduce que

$$DR\left(L^{\mathrm{T}}\right)^{-1} = L^{-1}R^{\mathrm{T}}D = D.$$

Como D es una matriz inversible entonces

$$R\left(L^{\mathrm{T}}\right)^{-1} = D^{-1}D = I$$

de donde $R = L^{\mathrm{T}}$.

Para obtener los elementos de D y L, puesto que $L = R^{\mathrm{T}}$ y A es simétrica, bastará sustituir en la fórmula (4.23) r_{ki} por l_{ik}, obteniéndose así

$$\boxed{\begin{array}{l} d_{ii} = a_{ii} - \displaystyle\sum_{k=1}^{i-1} l_{ik}^2 d_{kk} \\[2ex] l_{ji} = \dfrac{1}{d_{ii}}\left(a_{ij} - \displaystyle\sum_{k=1}^{i-1} l_{jk}\, d_{kk} l_{ik}\right),\ j = i+1, i+2, \ldots, n \end{array}}$$

c) Para resolver el sistema lineal $Au = b$ cuando $A = LDR$ calcularíamos, a partir de A y por este orden, los elementos

$$\begin{array}{l} \left\{d_{11}, \{r_{1j}\}_{j=2}^{n}, \{l_{j1}\}_{j=2}^{n}, \ldots, d_{ii}, \{r_{ij}\}_{j=i+1}^{n}, \right. \\ \left. \{l_{ji}\}_{j=i+1}^{n}, \ldots, d_{n-1,n-1}, r_{n-1,n}, l_{n,n-1}, d_{nn}\right\} \end{array}$$

para obtener de esta forma las matrices D, R y L. Una vez calculadas, resolveríamos los sistemas lineales

$$\begin{cases} Lv = b \\ Dw = v \\ Ru = w \end{cases}$$

en ese orden, teniendo en cuenta que el primero y el último son sistemas triangulares (se utiliza remonte sin división puesto que $l_{ii} = u_{ii} = 1$ para $i = 1, 2, \ldots, n$) y el segundo es diagonal (despejando cada incógnita directamente).

Cuando la matriz A sea simétrica, una vez obtenidas D y L, se resolverían los sistemas

$$\begin{cases} Lv = b \\ Dw = v \\ L^{\mathrm{T}}u = w. \end{cases} \quad \square$$

4.5. Sea $M \in \mathcal{M}_n$ una matriz tridiagonal por bloques,

$$M = \begin{pmatrix} A_1 & C_1 & & & \\ B_2 & A_2 & C_2 & & \\ & \ddots & \ddots & \ddots & \\ & & B_{p-1} & A_{p-1} & C_{p-1} \\ & & & B_p & A_p \end{pmatrix}$$

de forma que las p submatrices

$$\Delta_k = \begin{pmatrix} A_1 & C_1 & & & \\ B_2 & A_2 & C_2 & & \\ & \ddots & \ddots & \ddots & \\ & & B_{k-1} & A_{k-1} & C_{k-1} \\ & & & B_k & A_k \end{pmatrix}$$

para $k = 1, 2, \ldots, p$ son inversibles.

a) Probar que existen matrices

$$L = \begin{pmatrix} I_1 & & & & \\ L_2 & I_2 & & & \\ & \ddots & \ddots & & \\ & & L_{p-1} & I_{p-1} & \\ & & & L_p & I_p \end{pmatrix} \tag{4.25}$$

y

$$U = \begin{pmatrix} U_1 & D_1 & & & \\ & U_2 & D_2 & & \\ & & \ddots & \ddots & \\ & & & U_{p-1} & D_{p-1} \\ & & & & U_p \end{pmatrix} \tag{4.26}$$

tales que $M = LU$.

b) Hallar la expresión de L_i, U_i y D_i en función de A_i, B_i y C_i.

SOLUCIÓN.

a) Vamos a demostrar que en el proceso de eliminación gaussiana se llega a una matriz triangular superior y tridiagonal (por tanto, bidiagonal) por bloques y no es necesario realizar permutaciones de filas, puesto que en cada etapa de eliminación, la matriz M_{kk}^k va a ser inversible. Lo probamos por inducción:

i) Como $\Delta_1 = A_1$ es una matriz inversible, considerando la matriz

$$E_1 = \begin{pmatrix} I_1 & & & & & \\ -B_2(A_1)^{-1} & I_2 & & & & \\ & & \ddots & & & \\ & & & I_{p-1} & & \\ & & & & I_p \end{pmatrix}$$

se verifica que

$$M_2 = E_1 M = \begin{pmatrix} A_1 & C_1 & & & & \\ & A_2^2 & C_2^2 & & & \\ & B_3^2 & A_3^2 & C_3^2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & B_{p-1}^2 & A_{p-1}^2 & C_{p-1}^2 \\ & & & & B_p^2 & A_p^2 \end{pmatrix}$$

(nótese que M_2 sigue siendo una matriz tridiagonal por bloques).

ii) Supuesto que se han efectuado $k-1$ pasos veamos que se puede efectuar el siguiente. La hipótesis de inducción es que tenemos una matriz triangular superior y tridiagonal por bloques de la forma

$$M_k = E_{k-1}E_{k-2}\cdots E_2E_1M = \left(\begin{array}{ccccc|ccccc} A_1 & C_1 & & & & & & & & \\ & A_2^2 & C_2^2 & & & & & & & \\ & & \ddots & \ddots & & & & & & \\ & & & A_k^k & & C_k^k & & & & \\ \hline & & & B_{k+1}^k & & A_{k+1}^k & C_{k+1}^k & & & \\ & & & & & \ddots & \ddots & \ddots & & \\ & & & & & & B_{p-1}^k & A_{p-1}^k & C_{p-1}^k & \\ & & & & & & & B_p^k & A_p^k \end{array}\right).$$

Como $\det E_j = 1$, multiplicando por cajas, de la expresión anterior se tiene que

$$\det(A_1)\det(A_2^2)\cdots\det(A_{k-1}^{k-1})\det(A_k^k) = \det(\Delta_k) \neq 0,$$

lo que hace que $\det(A_k^k) \neq 0$ y, por tanto, la matriz A_k^k es inversible. De esta forma, a partir de la matriz

$$E_k = \begin{pmatrix} I_1 & & & & & & & \\ & \ddots & & & & & & \\ & & I_k & & & & & \\ & & -B_{k+1}^k (A_k^k)^{-1} & I_{k+1} & & & & \\ & & & & I_{k+2} & & & \\ & & & & & \ddots & & \\ & & & & & & I_p \end{pmatrix}$$

se verifica que la matriz $M_{k+1} = E_k M_k$ es de la forma

$$M_{k+1} = \begin{pmatrix} A_1 & C_1 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & A_k^k & C_k^k & & & & \\ & & & A_{k+1}^{k+1} & C_{k+1}^{k+1} & & & \\ & & & B_{k+2}^{k+1} & A_{k+2}^{k+1} & C_{k+2}^{k+1} & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & B_{p-1}^{k+1} & A_{p-1}^{k+1} & C_{p-1}^{k+1} & \\ & & & & & B_p^{k+1} & A_p^{k+1} \end{pmatrix}.$$

Consecuentemente, en la etapa $n-1$ la matriz

$$M_n = E_{n-1} E_{n-2} \cdots E_2 E_1 M$$

es triangular superior y tridiagonal por bloques. Así pues, la matriz M se puede factorizar en la forma $M = LU$ tomando

$$U = M_n \ \text{ y } \ L = (E_{n-1} E_{n-2} \cdots E_2 E_1)^{-1} = (E_1)^{-1} (E_2)^{-1} \cdots (E_{n-1})^{-1},$$

siendo U una matriz triangular superior y tridiagonal por bloques de la forma (4.26) y L triangular inferior y tridiagonal por bloques del tipo (4.25).

b) Haciendo el producto por bloques de las matrices L y U se obtienen las relaciones:

$$\begin{cases} C_i = D_i, & i = 1, 2, \ldots, p-1 \\ B_i = L_i U_{i-1}, & i = 2, 3, \ldots, p \\ A_i = L_i D_{i-1} + U_i, & i = 2, 3, \ldots, p \ \text{ con } \ A_1 = U_1. \end{cases}$$

De esta forma,

$$D_i = C_i$$

para $i = 1, 2, \ldots, p-1$, y a partir de $U_1 = A_1$ se van calculando, para cada índice $i \in \{2, 3, \ldots, p\}$, las matrices

$$\begin{cases} L_i = B_i (U_{i-1})^{-1} \\ U_i = A_i - L_i C_{i-1}. \end{cases} \quad \square$$

4.6. Sea $A \in \mathcal{M}_n$ una matriz simétrica inversible (aunque no necesariamente definida positiva) que admite factorización LU. Demostrar que se puede escribir $A = B\widetilde{B}^{\mathrm{T}}$ donde B es una matriz real triangular inferior y las columnas de $\widetilde{B}$ son las de B salvo, quizá, el signo. Es decir, si $B = (b_1, b_2, \ldots, b_n)$ entonces

$$\widetilde{B} = (\sigma_1 b_1, \sigma_2 b_2, \ldots, \sigma_n b_n)$$

con $\sigma_i = \pm 1$ para $i = 1, 2, \ldots, n$.

Solución. A partir de la factorización $A = LU$ se verifica que

$$\det(A) = \det(L)\det(U) = \prod_{i=1}^{n} u_{ii} \tag{4.27}$$

ya que $l_{ii} = 1$ para $i = 1, 2, \ldots, n$. Por otra parte, al ser A una matriz inversible se tiene que $\det(A) \neq 0$. Por tanto, de la relación (4.27) se obtiene que $u_{ii} \neq 0$ para $i = 1, 2, \ldots, n$. De esta forma, la matriz diagonal

$$D = \operatorname{diag}\left(\sqrt{|u_{11}|}, \sqrt{|u_{22}|}, \ldots, \sqrt{|u_{nn}|}\right)$$

es inversible y

$$D^{-1} = \operatorname{diag}\left(\frac{1}{\sqrt{|u_{11}|}}, \frac{1}{\sqrt{|u_{22}|}}, \ldots, \frac{1}{\sqrt{|u_{nn}|}}\right).$$

Por tanto, podemos escribir A en la forma

$$A = LU = B\widetilde{B}^{\mathrm{T}}$$

siendo

$$B = LD = \begin{pmatrix} \sqrt{|u_{11}|} & & & \\ \mu_{12} & \sqrt{|u_{22}|} & & \\ \cdots & & \ddots & \\ \mu_{1n} & \cdots & \mu_{n-1,n} & \sqrt{|u_{nn}|} \end{pmatrix}$$

y

$$\widetilde{B}^{\mathrm{T}} = D^{-1}U = \begin{pmatrix} \sigma_1\sqrt{|u_{11}|} & \nu_{12} & \cdots & \nu_{1n} \\ & \sigma_2\sqrt{|u_{22}|} & \cdots & \nu_{2n} \\ & & \ddots & \cdots \\ & & & \sigma_n\sqrt{|u_{nn}|} \end{pmatrix}$$

donde

$$\sigma_i = \operatorname{sign} u_{ii}$$

para $i = 1, 2, \ldots, n$. Como la matriz A es simétrica se verifica

$$B\widetilde{B}^{\mathrm{T}} = A = A^{\mathrm{T}} = \widetilde{B}B^{\mathrm{T}};$$

por otro lado, como

$$\det(B) = \det\left(B^{\mathrm{T}}\right) = \prod_{i=1}^{n} \sqrt{|u_{ii}|} > 0$$

se verifica que las matrices B y B^{T} son inversibles. Por tanto,

$$\widetilde{B}^{\mathrm{T}}\left(B^{\mathrm{T}}\right)^{-1} = B^{-1}\widetilde{B}. \tag{4.28}$$

Ahora bien, como

$$\widetilde{B}^{\mathrm{T}}\left(B^{\mathrm{T}}\right)^{-1} = \begin{pmatrix} \sigma_1 & k_{12} & \cdots & k_{1n} \\ & \sigma_2 & \cdots & k_{2n} \\ & & \ddots & \cdots \\ & & & \sigma_n \end{pmatrix}$$

y

$$B^{-1}\widetilde{B} = \begin{pmatrix} \sigma_1 & & & \\ m_{21} & \sigma_2 & & \\ \cdots & & \ddots & \\ m_{n1} & \cdots & m_{n,n-1} & \sigma_n \end{pmatrix},$$

la relación (4.28) determina que

$$\widetilde{B}^{\mathrm{T}}\left(B^{\mathrm{T}}\right)^{-1} = B^{-1}\widetilde{B} = \Delta$$

siendo

$$\Delta = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n),$$

por lo que $\widetilde{B} = B\Delta$. Así, si

$$B = (b_1, b_2, \ldots, b_n)$$

entonces

$$\widetilde{B} = B\Delta = (\sigma_1 b_1, \sigma_2 b_2, \ldots, \sigma_n b_n)$$

(véase el problema 2.6). □

4.7. ¿Cuántas factorizaciones de Cholesky distintas (es decir, sin suponer que los elementos diagonales de B son positivos) admite una matriz simétrica definida positiva?

SOLUCIÓN. Consideremos dos factorizaciones de Cholesky de la matriz A

$$B_1 B_1^{\mathrm{T}} = A = B_2 B_2^{\mathrm{T}}.$$

Entonces, a partir de la relación

$$(B_1)^{\mathrm{T}} \left((B_2)^{\mathrm{T}}\right)^{-1} = (B_1)^{-1} B_2$$

se verifica, como vimos en la demostración del teorema 4.4, que

$$(B_1)^{\mathrm{T}} \left((B_2)^{\mathrm{T}}\right)^{-1} = (B_1)^{-1} B_2 = D \tag{4.29}$$

siendo D una matriz diagonal y

$$((B_1)_{ii})^2 = ((B_2)_{ii})^2$$

para $i = 1, 2, \ldots, n$. Por tanto,

$$(B_1)_{ii} = \pm (B_2)_{ii}$$

para $i = 1, 2, \ldots, n$, de donde

$$D = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n) \ \text{ con } \ \sigma_i = \pm 1. \tag{4.30}$$

Nótese que la matriz D es inversible y $D^{-1} = D$. Reemplazando (4.30) en (4.29) se obtiene que

$$B_2 = B_1 D.$$

De esta forma, si

$$B_1 = (b_1, b_2, \ldots, b_n)$$

entonces

$$B_2 = (\sigma_1 b_1, \sigma_2 b_2, \ldots, \sigma_n b_n).$$

Es decir, las matrices B_1 y B_2 tienen las columnas iguales salvo, eventualmente, el signo. Como hay n columnas y en cada una de ellas el signo puede ser positivo o negativo en total tenemos 2^n factorizaciones de Cholesky distintas. □

4.8. Demostrar que la factorización LU preserva la estructura de matrices banda, es decir, si

$$a_{ij} = 0 \text{ para } |i-j| \geq p$$

entonces

$$l_{ij} = 0 \text{ si } i-j \geq p \text{ y } u_{ij} = 0 \text{ si } j-i \geq p.$$

Probar el mismo resultado para la factorización de Cholesky.

SOLUCIÓN. Como $A = LU$ entonces para $i, j = 1, 2, \ldots, n$ se verifica que

$$a_{ij} = \sum_{k=1}^{n} l_{ik} u_{kj}.$$

Ahora bien, como las matrices L y U son triangular inferior y superior respectivamente,

$$l_{ik} = 0 \text{ si } k > i$$

y

$$u_{kj} = 0 \text{ si } k > j,$$

por lo que la expresión anterior se puede escribir como

$$a_{ij} = \sum_{k=1}^{\min\{i,j\}} l_{ik} u_{kj} \tag{4.31}$$

para $i, j = 1, 2, \ldots, n$. Así pues, si $i \leq j$ la expresión (4.31) toma la forma

$$a_{ij} = \sum_{k=1}^{i} l_{ik} u_{kj} = \sum_{k=1}^{i-1} l_{ik} u_{kj} + u_{ij} \tag{4.32}$$

para $j = i, i+1, \ldots, n$, ya que $l_{ii} = 1$ y, por ser A una matriz banda, se verifica que

$$a_{ij} = 0 \text{ si } i+p \leq j \leq n,\ i = 1, 2, \ldots, n-p. \tag{4.33}$$

Vamos a demostrar por inducción en i que

$$u_{ij} = 0 \text{ si } i+p \leq j \leq n,\ i = 1, 2, \ldots, n-p.$$

i) Para $i = 1$ la relación (4.32) determina

$$a_{1j} = u_{1j}$$

de donde, aplicando (4.33) para $i = 1$, se deduce

$$u_{1j} = 0 \text{ si } 1+p \leq j \leq n.$$

ii) Supuesto cierto el resultado para $k < i$ lo demostramos para i. A partir de (4.32) se tiene que

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}. \tag{4.34}$$

Ahora bien, como $i + p \leq j \leq n$ entonces $k + p < j \leq n$ si $k < i$. Así pues, por la hipótesis de inducción, se verifica que

$$u_{kj} = 0 \text{ si } k = 1, 2, \ldots, i-1.$$

De esta forma, sustituyendo estos valores en la expresión (4.34) y teniendo en cuenta (4.33), se obtiene que

$$u_{ij} = 0 \text{ si } i + p \leq j \leq n.$$

Para la matriz L el razonamiento es análogo considerando $j < i$ en (4.31) y demostrando por inducción en j que

$$l_{ij} = 0 \text{ si } j + p \leq i \leq n,\ j = 1, 2, \ldots, n - p.$$

A continuación demostramos el resultado para la factorización de Cholesky. Puesto que $A = BB^{\mathrm{T}}$, considerando $i \leq j$, se tiene que

$$a_{ij} = \sum_{k=1}^{\min\{i,j\}} b_{ik} b_{jk} = \sum_{k=1}^{i} b_{ik} b_{jk} = \sum_{k=1}^{i-1} b_{ik} b_{jk} + b_{ii} b_{ji} \tag{4.35}$$

para $j = i, i+1, \ldots, n$ y, por ser A una matriz banda, verifica nuevamente (4.33). Vamos a demostrar por inducción en i que

$$b_{ji} = 0 \text{ si } i + p \leq j \leq n,\ i = 1, 2, \ldots, n - p.$$

i) Para $i = 1$ la relación (4.35) implica que

$$a_{1j} = b_{11} b_{j1}$$

por lo que, al aplicar (4.33) para $i = 1$, se obtiene

$$b_{j1} = 0 \text{ si } 1 + p \leq j \leq n.$$

ii) Supuesto que el resultado es cierto para $k < i$ lo vamos a demostrar para i. A partir de (4.35) se tiene que

$$b_{ji} = \frac{1}{b_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} b_{ik} b_{jk} \right). \tag{4.36}$$

Como $i+p \leq j \leq n$ entonces $k+p < j \leq n$ si $k < i$, por lo que la hipótesis de inducción hace que se tenga

$$b_{jk} = 0 \;\text{ si }\; k = 1, 2, \ldots, i-1.$$

Así pues, al sustituir estos valores en (4.36) y teniendo en cuenta (4.33), se obtiene que

$$b_{ji} = 0 \;\text{ si }\; j+p \leq i \leq n. \quad \square$$

4.9. Factorización de Cholesky: demostración alternativa.

a) Se considera una matriz $A \in \mathcal{M}_n$ simétrica cuyos menores principales son todos positivos. Si A se escribe en la forma

$$A = \left(\begin{array}{c|c} A_{n-1} & a \\ \hline a^{\mathrm{T}} & \alpha \end{array}\right)$$

siendo $A_{n-1} \in \mathcal{M}_{n-1}$, $a \in \mathbb{R}^{n-1}$ y $\alpha \in \mathbb{R}$, demostrar que

$$\alpha - a^{\mathrm{T}}(A_{n-1})^{-1}a > 0.$$

b) En el supuesto de que A_{n-1} admita factorización de Cholesky de la forma

$$A_{n-1} = B_{n-1}(B_{n-1})^{\mathrm{T}}$$

¿cómo deben elegirse $x \in \mathbb{R}^{n-1}$ y $\beta \in \mathbb{R}$ para que

$$A = \left(\begin{array}{c|c} B_{n-1} & \mathbf{0} \\ \hline x^{\mathrm{T}} & \beta \end{array}\right)\left(\begin{array}{c|c} (B_{n-1})^{\mathrm{T}} & x \\ \hline \mathbf{0} & \beta \end{array}\right)?$$

Probar que tal elección de x y β es posible.

c) Demostrar, por inducción sobre la dimensión de la matriz, la existencia de factorización de Cholesky para una matriz simétrica cuyos menores principales son todos positivos.

d) Deducir, del apartado *c*), que toda matriz simétrica con menores principales positivos es definida positiva.

Solución.

a) Como la matriz A es simétrica se verifica que

$$\left(\begin{array}{c|c} A_{n-1} & a \\ \hline a^{\mathrm{T}} & \alpha \end{array}\right) = A = A^{\mathrm{T}} = \left(\begin{array}{c|c} (A_{n-1})^{\mathrm{T}} & a \\ \hline a^{\mathrm{T}} & \alpha \end{array}\right).$$

Por tanto $(A_{n-1})^{\mathrm{T}} = A_{n-1}$, es decir, la matriz A_{n-1} es también simétrica. Al ser los menores principales δ_k de A positivos, lo son los de A_{n-1}. En particular,

$$\det(A_{n-1}) = \delta_{n-1} > 0$$

y, consecuentemente, la matriz A_{n-1} es inversible. Multiplicando por bloques la matriz A a la izquierda por la matriz

$$E = \left(\begin{array}{c|c} I_{n-1} & \mathbf{0} \\ \hline -\,a^{\mathrm{T}}(A_{n-1})^{-1} & 1 \end{array}\right)$$

se obtiene que

$$EA = \left(\begin{array}{c|c} A_{n-1} & a \\ \hline \mathbf{0} & \alpha - a^{\mathrm{T}}(A_{n-1})^{-1}a \end{array}\right).$$

De esta forma, como $\det(E) = 1$, se verifica que

$$\begin{aligned}\left(\alpha - a^{\mathrm{T}}(A_{n-1})^{-1}a\right)\det(A_{n-1}) &= \det(EA) = \det(E)\det(A) \\ &= \det(A) = \delta_n > 0\end{aligned}$$

y, por tanto,

$$\alpha - a^{\mathrm{T}}(A_{n-1})^{-1}a > 0.$$

b) Como la matriz A_{n-1} admite factorización de Cholesky de la forma

$$A_{n-1} = B_{n-1}(B_{n-1})^{\mathrm{T}} \tag{4.37}$$

podemos considerar la matriz

$$B = \left(\begin{array}{c|c} B_{n-1} & \mathbf{0} \\ \hline x^{\mathrm{T}} & \beta \end{array}\right).$$

Vamos a determinar $x \in \mathbb{R}^{n-1}$ y $\beta \in \mathbb{R}$ para que se verifique la igualdad

$$A = BB^{\mathrm{T}},$$

es decir,

$$\begin{aligned}\left(\begin{array}{c|c} A_{n-1} & a \\ \hline a^{\mathrm{T}} & \alpha \end{array}\right) &= \left(\begin{array}{c|c} B_{n-1} & \mathbf{0} \\ \hline x^{\mathrm{T}} & \beta \end{array}\right)\left(\begin{array}{c|c} (B_{n-1})^{\mathrm{T}} & x \\ \hline \mathbf{0} & \beta \end{array}\right) \\ &= \left(\begin{array}{c|c} A_{n-1} & B_{n-1}x \\ \hline x^{\mathrm{T}}(B_{n-1})^{\mathrm{T}} & \beta^2 + x^{\mathrm{T}}x \end{array}\right).\end{aligned}$$

Por tanto, debe verificarse

$$B_{n-1}x = a \;\text{ y }\; \alpha = \beta^2 + x^{\mathrm{T}}x.$$

Puesto que B_{n-1} es inversible, la elección del vector

$$x = (B_{n-1})^{-1}a$$

está bien definida. En cuanto a β, se tiene que

$$\begin{aligned}\alpha &= \beta^2 + \left((B_{n-1})^{-1}a\right)^{\mathrm{T}}\left((B_{n-1})^{-1}a\right)\\ &= \beta^2 + a^{\mathrm{T}}\left((B_{n-1})^{-1}\right)^{\mathrm{T}}(B_{n-1})^{-1}a\\ &= \beta^2 + a^{\mathrm{T}}\left((B_{n-1})^{\mathrm{T}}\right)^{-1}(B_{n-1})^{-1}a\\ &= \beta^2 + a^{\mathrm{T}}\left(B_{n-1}(B_{n-1})^{\mathrm{T}}\right)^{-1}a\\ &= \beta^2 + a^{\mathrm{T}}(A_{n-1})^{-1}a\end{aligned}$$

(véase (4.37)). Como por el apartado a) se sabe que

$$\beta^2 = \alpha - a^{\mathrm{T}}(A_{n-1})^{-1}a > 0,$$

la elección

$$\beta = \sqrt{\alpha - a^{\mathrm{T}}(A_{n-1})^{-1}a}$$

está bien justificada.

c) Procedemos por inducción:

i) Para $n = 1$ el resultado es inmediato, pues $A = a_{11} > 0$ y, por tanto, $A = \sqrt{a_{11}}\sqrt{a_{11}}$.

ii) Supongamos cierto el resultado para matrices de orden $n-1$ y demostrémoslo para matrices de orden n. Para ello consideramos una matriz $A \in \mathcal{M}_n$ simétrica con menores principales positivos y la escribimos como

$$A = \left(\begin{array}{c|c} A_{n-1} & a \\ \hline a^{\mathrm{T}} & \alpha \end{array}\right).$$

Como la matriz $A_{n-1} \in \mathcal{M}_{n-1}$ es simétrica y con menores principales positivos, por la hipótesis de inducción, A_{n-1} admite una factorización de Cholesky de la forma

$$A_{n-1} = B_{n-1}(B_{n-1})^{\mathrm{T}}$$

para una matriz real $B_{n-1} \in \mathcal{M}_{n-1}$ triangular inferior. Tomando

$$\begin{cases} x = (B_{n-1})^{-1}a \\ \beta = \sqrt{\alpha - a^{\mathrm{T}}(A_{n-1})^{-1}a} \end{cases}$$

se sabe, por el apartado b), que la matriz real y triangular inferior

$$B = \left(\begin{array}{c|c} B_{n-1} & \mathbf{0} \\ \hline x^{\mathrm{T}} & \beta \end{array}\right)$$

verifica

$$A = BB^{\mathrm{T}}.$$

d) Como $A = BB^{\mathrm{T}}$ entonces

$$\det(B)\det\left(B^{\mathrm{T}}\right) = \det(A) = \delta_n > 0$$

por lo que las matrices B y B^{T} son inversibles. De esta forma, para todo $v \in \mathbf{V}\backslash\{0\}$ se verifica que

$$v^{\mathrm{T}}Av = v^{\mathrm{T}}BB^{\mathrm{T}}v = \left(B^{\mathrm{T}}v\right)^{\mathrm{T}}\left(B^{\mathrm{T}}v\right) = ||B^{\mathrm{T}}v||_2^2 > 0$$

por ser B^{T} inversible y $v \neq 0$. □

4.10. Cálculo recursivo de la inversa de una matriz.

a) Fórmula de Sherman–Morrison. Sea $B \in \mathcal{M}_n$ real e inversible y sean $u, v \in \mathbb{R}^n$ tales que la matriz $B + uv^{\mathrm{T}}$ es inversible. Comprobar que

$$(B + uv^{\mathrm{T}})^{-1} = B^{-1} - \frac{B^{-1}uv^{\mathrm{T}}B^{-1}}{1 + v^{\mathrm{T}}B^{-1}u}.$$

b) Sea $A \in \mathcal{M}_n$ una matriz real escrita en la forma

$$A = D + \sum_{i=1}^{m} u_i v_i^{\mathrm{T}}$$

donde D es una matriz diagonal e inversible y los vectores $u_i, v_i \in \mathbb{R}^n$ son tales que las m matrices

$$M_k = D + \sum_{i=1}^{k} u_i v_i^{\mathrm{T}}$$

para $k = 1, 2, \ldots, m$ son inversibles. Si $C_k = (M_k)^{-1}$, encontrar una fórmula recurrente para C_{k+1}.

c) Sea $A \in \mathcal{M}_n$ una matriz simétrica definida positiva. Demostrar que A se puede escribir en la forma

$$A = D + \sum_{i=1}^{n} u_i \mathbf{e}_i^{\mathrm{T}}$$

siendo $\mathbf{e}_i$ el i–ésimo vector de la base canónica, verificándose las hipótesis requeridas en *b*).

d) Razonar cómo pueden usarse los resultados anteriores para calcular la inversa de una matriz simétrica definida positiva.

SOLUCIÓN.

a) Denotando por

$$C = B^{-1} - \frac{B^{-1}uv^{\mathrm{T}}B^{-1}}{1 + v^{\mathrm{T}}B^{-1}u},$$

de la relación

$$\begin{aligned} C\left(B + uv^{\mathrm{T}}\right) &= \left(B^{-1} - \frac{B^{-1}uv^{\mathrm{T}}B^{-1}}{1 + v^{\mathrm{T}}B^{-1}u}\right)\left(B + uv^{\mathrm{T}}\right) \\ &= I + B^{-1}uv^{\mathrm{T}} - \frac{B^{-1}uv^{\mathrm{T}}}{1 + v^{\mathrm{T}}B^{-1}u} - \frac{B^{-1}u(v^{\mathrm{T}}B^{-1}u)v^{\mathrm{T}}}{1 + v^{\mathrm{T}}B^{-1}u} \\ &= I + B^{-1}uv^{\mathrm{T}} - \frac{B^{-1}uv^{\mathrm{T}}}{1 + v^{\mathrm{T}}B^{-1}u} - v^{\mathrm{T}}B^{-1}u\frac{B^{-1}uv^{\mathrm{T}}}{1 + v^{\mathrm{T}}B^{-1}u} \\ &= I + B^{-1}uv^{\mathrm{T}} - \left(1 + v^{\mathrm{T}}B^{-1}u\right)\frac{B^{-1}uv^{\mathrm{T}}}{1 + v^{\mathrm{T}}B^{-1}u} = I \end{aligned}$$

se deduce que $C = \left(B + uv^{\mathrm{T}}\right)^{-1}$.

b) Como para cada $k \in \{1, 2, \ldots, m-1\}$

$$M_{k+1} = D + \sum_{i=1}^{k+1} u_i v_i^{\mathrm{T}} = M_k + u_{k+1}v_{k+1}^{\mathrm{T}},$$

por el apartado *a*) se tiene que

$$\begin{aligned} C_{k+1} = (M_{k+1})^{-1} &= \left(M_k + u_{k+1}v_{k+1}^{\mathrm{T}}\right)^{-1} \\ &= (M_k)^{-1} - \frac{(M_k)^{-1}u_{k+1}v_{k+1}^{\mathrm{T}}(M_k)^{-1}}{1 + v_{k+1}^{\mathrm{T}}(M_k)^{-1}u_{k+1}} \\ &= C_k - \frac{C_k u_{k+1}v_{k+1}^{\mathrm{T}}C_k}{1 + v_{k+1}^{\mathrm{T}}C_k u_{k+1}}. \end{aligned}$$

c) Vamos a considerar la matriz diagonal

$$D = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$$

y los vectores

$$u_i = (a_{1i}, \ldots, a_{i-1,i}, 0, a_{i+1,i}, \ldots, a_{ni})^{\mathrm{T}}$$

para $i = 1, 2, \ldots, n$; es decir, cada vector u_i es la columna i–ésima de la matriz A salvo la i–ésima componente que es nula. Como para cada índice $i \in \{1, 2, \ldots, n\}$ se verifica que

$$u_i \mathbf{e}_i^{\mathrm{T}} = \begin{pmatrix} a_{1i} \\ \cdots \\ a_{i-1,i} \\ 0 \\ a_{i+1,i} \\ \cdots \\ a_{ni} \end{pmatrix} (0, \ldots, 0, \overset{i)}{1}, 0, \ldots, 0) = \begin{pmatrix} 0 & \cdots & 0 & a_{1i} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & a_{i-1,i} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & a_{i+1,i} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & a_{ni} & 0 & \cdots & 0 \end{pmatrix},$$

entonces, claramente, la matriz A puede escribirse como

$$A = D + \sum_{i=1}^{n} u_i \mathbf{e}_i^{\mathrm{T}}.$$

Veamos que con estas elecciones se verifican las hipótesis requeridas en el apartado *b*). En efecto:

i) Como A es simétrica y definida positiva, la matriz D es también simétrica y definida positiva y, por tanto, inversible.

ii) Las matrices M_k son de la forma

$$M_k = D + \sum_{i=1}^{k} u_i \mathbf{e}_i^{\mathrm{T}} = \left(\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1k} & & & \\ \cdots & & \cdots & & \mathbf{0} & \\ a_{k1} & \cdots & a_{kk} & & & \\ \hline a_{k+1,1} & \cdots & a_{k+1,k} & a_{k+1,k+1} & & \\ \cdots & & \cdots & & \ddots & \\ a_{n1} & \cdots & a_{nk} & & & a_{nn} \end{array}\right)$$

para $k = 1, 2, \ldots, n$. Consecuentemente, por ser cada M_k una matriz triangular inferior por bloques, se verifica que

$$\det(M_k) = \det \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \cdots & \cdots & \cdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix} \prod_{j=k+1}^{n} a_{jj} = \delta_k \prod_{j=k+1}^{n} a_{jj} > 0,$$

ya que al ser A simétrica y definida positiva sus menores principales δ_k son positivos y sus elementos diagonales también (véase el problema 2.16). Con esto se tiene que todas las matrices M_k son inversibles.

d) Los resultados anteriores sirven para calcular la inversa de la matriz A de forma recursiva; dado que $A = M_n$ entonces

$$A^{-1} = (M_n)^{-1} = C_n,$$

por lo que bastará calcular $C_0 = D^{-1}$ (lo cual es trivial) y, a partir de ella, utilizar la recurrencia del apartado *b*) hasta llegar a C_n. Nótese que estos cálculos no requieren invertir ninguna matriz. □

4.6.2. Problemas propuestos

4.11. Demostrar que la factorización LU sigue siendo posible si A es singular, siempre que los $n-1$ menores principales δ_k, $k = 1, 2, \ldots, n-1$, sean no nulos. Probar que, en ese caso, $u_{nn} = 0$.

4.12. Sea $A \in \mathcal{M}_n$ una matriz con todos sus menores principales no nulos. Demostrar que existen matrices $B \in \mathcal{M}_n$ triangular inferior y $C = (c_{ij})_{i,j=1}^n \in \mathcal{M}_n$ triangular superior con

$$c_{ii} = 1$$

para $i = 1, 2, \ldots, n$ tales que $A = BC$. ¿Es única la factorización anterior?

4.13. Sea $A \in \mathcal{M}_n$.

a) Adaptar el método de eliminación de Gauss para probar que existe $M \in \mathcal{M}_n$ inversible tal que MA es triangular inferior.

b) Encontrar condiciones suficientes para que la matriz A pueda factorizarse en la forma $A = UL$ con U triangular superior y L triangular inferior. ¿Es única tal factorización?

4.14. Se considera una matriz simétrica A cuyos menores principales son todos no nulos. Demostrar que, en la factorización LU de A, cada columna de L es múltiplo de la correspondiente fila de U. Explicar cómo facilita esto el cálculo de la factorización LU de una matriz simétrica.

4.15. Sea $A \in \mathcal{M}_n$ una matriz que admite factorización LU y δ_k el menor principal de orden k de la matriz A, $k = 1, 2, \ldots, n$.

a) Si A es inversible demostrar que

$$\delta_k \neq 0 \tag{4.38}$$

para $k = 1, 2, \ldots, n$.

b) Si A no es inversible y existe un índice k_0 de forma que $\delta_{k_0} = 0$, probar que

$$\delta_k = 0$$

para $k = k_0, k_0 + 1, \ldots, n$.

4.16. Sea $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ una matriz de diagonal estrictamente dominante.

a) Demostrar que A admite factorización LU.

b) Si además A es simétrica y verifica que

$$a_{ii} > 0$$

para $i = 1, 2, \ldots, n$, probar que A admite factorización de Cholesky.

4.17. Sea $A \in \mathcal{M}_n$ una matriz simétrica definida positiva.

a) Probar que la matriz A puede factorizarse en la forma $A = CDC^{\mathrm{T}}$ donde $C = (c_{ij})_{i,j=1}^n$ es triangular inferior con

$$c_{ii} = 1,$$

para $i = 1, 2, \ldots, n$, y D es diagonal. ¿Es única tal factorización?

b) Obtener las matrices C y D y demostrar que

$$d_{ii} = (b_{ii})^2$$

para $i = 1, 2, \ldots, n$, siendo $A = BB^{\mathrm{T}}$ una factorización de Cholesky de A.

c) Deducir un método directo de resolución del sistema $Au = b$, con A simétrica definida positiva, basado en los apartados *a*) y *b*).

4.18. Se dice que una matriz $A \in \mathcal{M}_n$ es *reducible* si existen dos conjuntos no vacíos S y T tales que $S \cup T = \{1, 2, \ldots, n\}$, $S \cap T = \emptyset$ y $a_{ij} = 0$ cuando $i \in S$ y $j \in T$. En caso contrario, se dice que la matriz A es *irreducible*.

a) Probar que si A es reducible existe una matriz de permutaciones P tal que

$$P^{\mathrm{T}} A P = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline \mathbf{0} & A_{22} \end{array} \right)$$

siendo $A_{11} \in \mathcal{M}_t$, $A_{22} \in \mathcal{M}_s$, con $s = \mathrm{card}(S)$ y $t = \mathrm{card}(T)$.

b) Demostrar que el sistema $Ax = b$ con A reducible y regular puede descomponerse en N sistemas, $2 \leq \mathrm{N} \leq n$, de la forma

$$\sum_{k=j}^{\mathrm{N}} A_{jk} x_k = b_j$$

con $A_{jk} \in \mathcal{M}_{m_j m_k}$, $x_k \in \mathbb{R}^{m_k}$ y $b \in \mathbb{R}^{m_j}$, siendo $\displaystyle\sum_{j=1}^{\mathrm{N}} m_j = n$ y A_{jj} irreducibles y regulares.

c) Basándose en *b*), proponer un método de resolución de sistemas con matrices reducibles y no singulares.

4.19. Probar que una matriz tridiagonal es irreducible si y sólo si sus elementos no diagonales son no nulos.

4.20. Consideremos una matriz simétrica definida positiva escrita en la forma

$$\left(\begin{array}{c|c} A & B \\ \hline B^{\mathrm{T}} & C \end{array}\right).$$

Elegir adecuadamente el vector $w \in \mathbf{V}$ verificando

$$\begin{pmatrix} w^{\mathrm{T}} & v^{\mathrm{T}} \end{pmatrix} \left(\begin{array}{c|c} A & B \\ \hline B^{\mathrm{T}} & C \end{array}\right) \begin{pmatrix} w \\ v \end{pmatrix} = v^{\mathrm{T}} M v$$

para demostrar que la matriz $M = C - B^{\mathrm{T}} A^{-1} B$ es también simétrica definida positiva.

4.7. Prácticas

4.1. Construir un sistema lineal de comportamiento análogo al del ejemplo 4.1 cuando se trabaja con doble precisión. Realizar las operaciones con `MATLAB` para comprobar los resultados.

4.2. Escribir funciones que proporcionen la solución de un sistema triangular inferior con unos en la diagonal, un sistema triangular inferior arbitrario y un sistema triangular superior, respectivamente, que tengan como argumentos de entrada la matriz del sistema y el vector segundo miembro.

4.3. Programar el método de eliminación gaussiana de forma que sirva para resolver sucesivos sistemas con la misma matriz, implementándolo siguiendo las indicaciones dadas en la subsección 4.3.4. Comparar con el comando \ de `MATLAB`.

4.4. Escribir un programa que calcule la inversa de una matriz mediante el método de Gauss–Jordan introducido en el problema 4.1. Comparar con el comando `inv` de `MATLAB`.

4.5. Programar el método de la factorización LU de forma que se puedan resolver varios sistemas con la misma matriz. Comparar con el comando `lu` de `MATLAB`.

4.6. Hacer una versión del programa anterior pensada para matrices banda.

4.7. Programar el método de la factorización de Cholesky de forma que se puedan resolver varios sistemas con la misma matriz. Comparar con el comando `chol` de `MATLAB`.

4.8. Hacer una versión del programa anterior pensada para matrices banda.

4.9. Escribir una función que implemente el método del problema 4.2 para matrices tridiagonales.

4.10. Programar el método de cálculo de la inversa de una matriz simétrica definida positiva estudiado en el problema 4.10.

5 Resolución de sistemas lineales: métodos iterativos

5.1. Introducción

En este capítulo abordaremos el estudio de los métodos iterativos para la resolución de sistemas lineales. Estos métodos proporcionan la solución de un sistema lineal $Au = b$ como límite de una sucesión de vectores; por tanto, en general no se obtiene la solución de forma exacta sino sólo de una manera aproximada. No obstante, esto no impide que su uso esté muy extendido debido, especialmente, a dos limitaciones que presentan los métodos directos:

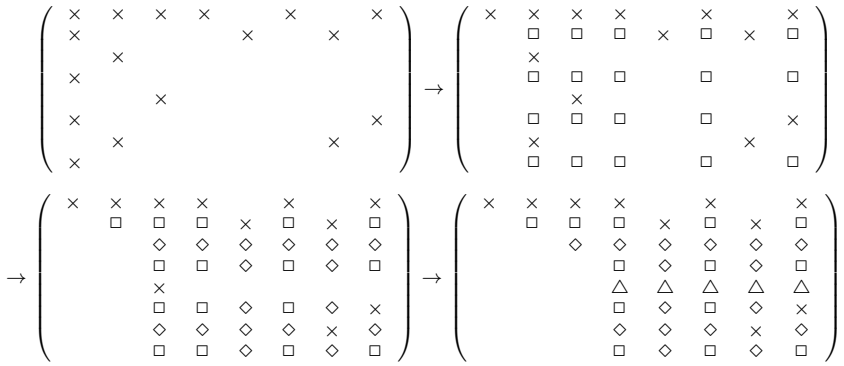
a) Cuando el tamaño de la matriz A del sistema es muy grande ($n \gg 100$) la propagación del error de redondeo es también grande y los resultados obtenidos (teóricamente exactos) pueden diferir bastante de los reales.

b) Muchas de las matrices de los sistemas lineales que aparecen en las aplicaciones son de gran tamaño ($n \simeq 100000$), pero tienen la particularidad de que la mayoría de sus elementos son nulos. A este tipo de matrices suele calificárselas como *vacías* o *huecas*; en ellas el número de elementos no nulos es de orden n:

   i) Si los elementos no nulos están distribuidos alrededor de la diagonal principal, todavía son de aplicación los métodos directos (recuérdese que la factorización LU y la de Cholesky preservan la estructura de matriz banda).

   ii) Cuando lo anterior no ocurre se dice que la matriz es *dispersa* y, al aplicar métodos directos a este tipo de matrices, se produce el fenómeno conocido como *rellenado* (*fill–in* en inglés): al aplicar el proceso de eliminación gaussiana se consiguen anular los elementos de una columna por debajo de la diagonal principal, pero se convierten en elementos no nulos algunos que previamente lo eran. Esto hace que la matriz, que en

principio era "vacía", vaya "llenándose", con lo que en cada paso crece el número de elementos que hay que anular. En el ejemplo siguiente se muestra cómo se va produciendo este fenómeno en las tres primeras etapas de eliminación para una matriz dispersa



donde los elementos denotados por $\square$ son elementos eventualmente no nulos que se introducen en la primera etapa de eliminación, los denotados por $\diamond$ se introducen en la segunda y los denotados por $\triangle$, en la tercera. Por tanto, si no se realiza una adaptación de los métodos directos al caso particular de matrices dispersas, los resultados no van a ser, en general, buenos. Existen técnicas sofisticadas que adaptan los métodos directos a este tipo de matrices, pero su estudio desborda los objetivos de este curso. Puede encontrarse una descripción suficientemente detallada en el capítulo 5 de [**La–Th**].

Los métodos iterativos obvian estos problemas puesto que se basan en la resolución (reiteradas veces) de sistemas diagonales o triangulares (ya sean por puntos o por bloques). La premisa fundamental en los métodos iterativos es que el coste de cada iteración no sea mayor que el de la multiplicación de una matriz por un vector, objetivo que se consigue si se está resolviendo, en cada paso, un sistema diagonal o triangular.

5.2. Estudio general

Para ver en qué consisten los métodos iterativos, supongamos que, dado un sistema lineal $Au = b$ con $A \in \mathcal{M}_n$ inversible, encontramos una matriz $B \in \mathcal{M}_n$ y un vector $c \in \mathbf{V}$ de forma que la matriz $I - B$ es inversible y la única solución del sistema lineal $u = Bu + c$ es la solución de $Au = b$.

La forma del sistema $u = Bu + c$ sugiere abordar la resolución del sistema lineal $Au = b$ mediante un *método iterativo* asociado a la matriz B del siguiente modo: dado un vector inicial $u^0 \in \mathbf{V}$ arbitrario se construye la sucesión de vectores $\{u^k\}_{k=0}^{\infty}$ dada por

$$u^{k+1} = Bu^k + c \tag{5.1}$$

para $k \in \mathbb{N} \cup \{0\}$, con la esperanza de que converja a la solución del sistema lineal.

Definición 5.1. El método iterativo (5.1) es *convergente* si existe $u \in \mathbf{V}$ tal que

$$\lim_{k \to +\infty} u^k = u$$

para cualquier vector inicial $u^0 \in \mathbf{V}$. Nótese que, en tal caso, este vector u verifica $u = Bu + c$. □

Si para cada $k \in \mathbb{N} \cup \{0\}$ denotamos el error cometido en cada iteración por

$$e^k = u^k - u \tag{5.2}$$

se verifica que

$$e^k = u^k - u = (Bu^{k-1} + c) - (Bu + c) = B(u^{k-1} - u) = Be^{k-1}$$

y, por tanto,

$$e^k = Be^{k-1} = B^2 e^{k-2} = \cdots = B^k e^0. \tag{5.3}$$

Así pues, el error en las iteraciones depende, en esencia, de las potencias sucesivas de la matriz B. Obsérvese que el resultado siguiente, que da el criterio fundamental de convergencia de los métodos iterativos, sólo involucra la matriz B del método iterativo considerado.

Teorema 5.1. *Sea $B \in \mathcal{M}_n$. Son equivalentes:*

a) El método iterativo asociado a la matriz B es convergente.

b) $\varrho(B) < 1$.

c) Existe una norma matricial $\|\cdot\|$ (que se puede tomar subordinada) tal que

$$\|B\| < 1.$$

DEMOSTRACIÓN. A partir del teorema 2.6 y de la relación (5.3), se tienen las equivalencias:

$$\begin{aligned}\text{El método es convergente } &\Leftrightarrow \lim_{k\to+\infty} e^k = 0 \text{ para todo } e^0 \in \mathbf{V}\\ &\Leftrightarrow \lim_{k\to+\infty} B^k e^0 = 0 \text{ para todo } e^0 \in \mathbf{V}\\ &\Leftrightarrow \varrho(B) < 1\\ &\Leftrightarrow |\!|\!| B |\!|\!| < 1 \text{ para una norma matricial } |\!|\!|\cdot|\!|\!|\,.\end{aligned}$$ □

Se plantea la cuestión de cómo elegir entre diversos métodos iterativos convergentes para la resolución de un mismo sistema lineal $Au = b$. En esta línea, se tiene el siguiente resultado:

Proposición 5.1. *Sea* $||\cdot||$ *una norma vectorial y consideremos* $u \in \mathbf{V}$ *tal que*

$$u = Bu + c.$$

Para el método iterativo

$$\begin{cases} u^0 \in \mathbf{V} \\ u^{k+1} = Bu^k + c,\ k \in \mathbb{N} \cup \{0\}, \end{cases}$$

se verifica que

$$\lim_{k\to+\infty} \left(\sup_{||e^0||=1} \left|\left| e^k \right|\right|^{\frac{1}{k}} \right) = \varrho(B)$$

donde $e^k = u^k - u$.

DEMOSTRACIÓN. A partir de (5.3), si $|\!|\!|\cdot|\!|\!|$ es la norma matricial subordinada a la norma vectorial $||\cdot||$, para todo $k \in \mathbb{N} \cup \{0\}$ se verifica que

$$|\!|\!| B^k |\!|\!| = \sup_{||e^0||=1} \left|\left| B^k e^0 \right|\right| = \sup_{||e^0||=1} \left|\left| e^k \right|\right|,$$

y el resultado se deduce directamente del teorema 2.7. □

Observación 5.1. La proposición 5.1 afirma que

$$\sup_{||u^0-u||=1} \left|\left| u^k - u \right|\right| \sim (\varrho(B))^k \text{ si } k \to +\infty.$$

Por tanto, en el caso de que el método iterativo sea convergente, la convergencia a u de la sucesión $\{u^k\}_{k=0}^{\infty}$ será igual de rápida que la convergencia a cero de la sucesión de números reales $\left\{(\varrho(B))^k\right\}_{k=0}^{\infty}$ y, consecuentemente, tanto más rápida cuanto menor sea el radio espectral de la matriz B que define dicho método. □

A la hora de resolver un sistema lineal mediante un método iterativo deberemos, en primer lugar, asegurar su convergencia (por ejemplo, encontrando alguna norma para la cual $\|\!|B|\!\| < 1$ o viendo que $\varrho(B) < 1$). A continuación, y en caso de disponer de varios a nuestro alcance, elegir aquel cuya matriz asociada tenga una norma o un radio espectral menor.

5.3. Métodos de Jacobi, Gauss–Seidel y relajación

En esta sección vamos a introducir tres de los métodos iterativos más clásicos para la resolución de un sistema lineal $Au = b$ con A inversible. Todos ellos comparten, como idea básica en su construcción, el descomponer o partir (*split* en inglés) la matriz del sistema en suma de dos matrices. Más concretamente, la estrategia que se va a utilizar es descomponer la matriz A en la forma

$$\boxed{A = M - N}$$

donde M va a ser una matriz inversible fácil de invertir (en el sentido de que sea fácil resolver un sistema asociado a dicha matriz como ocurre, por ejemplo, cuando M es una matriz diagonal o triangular ya sea por puntos o por bloques). Con esta descomposición de A se verifica:

$$Au = b \Leftrightarrow (M - N)u = b \Leftrightarrow Mu = Nu + b \Leftrightarrow u = Bu + c$$

donde

$$\boxed{B = M^{-1}N} \text{ y } \boxed{c = M^{-1}b} \tag{5.4}$$

De esta forma, podemos considerar el método iterativo

$$\begin{cases} u^0 \in \mathbf{V} \text{ arbitrario} \\ u^{k+1} = Bu^k + c,\ k \in \mathbb{N} \cup \{0\}. \end{cases} \tag{5.5}$$

Como $N = M - A$, entonces $B = M^{-1}N = M^{-1}(M - A) = I - M^{-1}A$. Así,

$$I - B = M^{-1}A$$

es una matriz inversible, por lo que el sistema $(I - B)u = c$ tiene solución única. En la práctica, para calcular u^{k+1}, se resolverá el sistema

$$Mu^{k+1} = Nu^k + b$$

en vez de trabajar, directamente, con (5.5). Es por esto por lo que requerimos que M sea una matriz fácil de invertir.

Observación 5.2. Como ya se ha comentado, todos los métodos iterativos que vamos a estudiar responden a una descomposición $M - N$ de la matriz A. Intuitivamente, cuanto más de A haya en M, tanto más se parecerá cada iteración al cálculo de la solución exacta (de hecho, en el caso límite $M = A$ la solución se obtiene en la primera iteración). No obstante, esto va en contra de la idea inicial de que el coste de cada iteración sea bajo. Un buen método iterativo será aquel que mantenga un equilibrio entre estas dos estrategias enfrentadas. □

A continuación describimos los diversos métodos iterativos que vamos a estudiar. Para ello, introducimos la siguiente notación:

Notación 5.1. Dada una matriz $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ inversible con

$$a_{ii} \neq 0 \tag{5.6}$$

para $i = 1, 2, \ldots, n$, consideramos la siguiente descomposición de la matriz

$$A = \begin{pmatrix} & & -F \\ & D & \\ -E & & \end{pmatrix}$$

que podemos escribir en la forma

$$\boxed{A = D - E - F}$$

donde

$$D = \text{diag}\,(a_{11}, a_{22}, \ldots, a_{nn}),\ E = (e_{ij})_{i,j=1}^n \ \text{ y } \ F = (f_{ij})_{i,j=1}^n$$

siendo

$$e_{ij} = \begin{cases} -a_{ij} & \text{si} \quad i > j \\ 0 & \text{si} \quad i \leq j \end{cases} \quad \text{y} \quad f_{ij} = \begin{cases} -a_{ij} & \text{si} \quad i < j \\ 0 & \text{si} \quad i \geq j. \end{cases}$$

A esta descomposición de A la denominaremos *descomposición $D - E - F$ por puntos* de la matriz A. □

5.3.1. Método de Jacobi

Consiste en tomar

$$\boxed{M = D} \ \text{ y } \ \boxed{N = E + F}$$

Así pues,

$$Au = b \ \Leftrightarrow \ Du = (E + F)u + b \ \Leftrightarrow \ u = D^{-1}(E + F)u + D^{-1}b$$

que conduce al *método iterativo de Jacobi* (por puntos)

$$\begin{cases} u^0 \in \mathbf{V} \text{ arbitrario} \\ u^{k+1} = D^{-1}(E+F)u^k + D^{-1}b,\ k \in \mathbb{N} \cup \{0\} \end{cases}$$

o, equivalentemente,

$$\begin{cases} u^0 \in \mathbf{V} \text{ arbitrario} \\ Du^{k+1} = (E+F)u^k + b,\ k \in \mathbb{N} \cup \{0\}. \end{cases} \tag{5.7}$$

Nótese que la hipótesis (5.6) determina que la matriz $M = D$ es inversible. La matriz de este método es

$$\boxed{\mathcal{J} = D^{-1}(E+F) = I - D^{-1}A}$$

que se denomina *matriz de Jacobi* (por puntos). La iteración definida en (5.7) puede escribirse, coordenada a coordenada, como

$$\begin{aligned} a_{ii}u_i^{k+1} &= b_i - a_{i1}u_1^k - \cdots - a_{i,i-1}u_{i-1}^k - a_{i,i+1}u_{i+1}^k - \cdots - a_{in}u_n^k \\ &= b_i - \sum_{j=1}^{i-1} a_{ij}u_j^k - \sum_{j=i+1}^{n} a_{ij}u_j^k \end{aligned} \tag{5.8}$$

para $i = 1, 2, \ldots, n$, donde

$$u^k = \left(u_1^k, u_2^k, \ldots, u_n^k\right)^{\mathrm{T}} \quad \text{y} \quad u^{k+1} = \left(u_1^{k+1}, u_2^{k+1}, \ldots, u_n^{k+1}\right)^{\mathrm{T}}.$$

Como se observa, las n componentes del vector u^{k+1} pueden calcularse de forma simultánea a partir de las n componentes del vector u^k; de hecho, el método de Jacobi también se conoce como método de las *iteraciones simultáneas*. La implementación efectiva de este método se estudiará en la sección 5.5.

5.3.2. Método de Gauss–Seidel

En el cálculo de la componente u_i^{k+1}, parece claro que una estrategia adecuada para mejorar la convergencia sería emplear las componentes ya calculadas

$$\{u_1^{k+1}, u_2^{k+1}, \ldots, u_{i-1}^{k+1}\}$$

en vez de utilizar las "antiguas"

$$\{u_1^k, u_2^k, \ldots, u_{i-1}^k\}.$$

Esta consideración nos lleva a reemplazar el sistema (5.8) por

$$\begin{aligned} a_{ii}u_i^{k+1} &= b_i - a_{i1}u_1^{k+1} - \cdots - a_{i,i-1}u_{i-1}^{k+1} - a_{i,i+1}u_{i+1}^{k} - \cdots - a_{in}u_n^{k} \\ &= b_i - \sum_{j=1}^{i-1} a_{ij}u_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}u_j^{k} \end{aligned}$$

para $i = 1, 2, \ldots, n$. Matricialmente, estas ecuaciones se escriben

$$Du^{k+1} = Eu^{k+1} + Fu^k + b,$$

es decir,

$$(D - E)u^{k+1} = Fu^k + b.$$

Tenemos así definido un nuevo método iterativo tomando

$$\boxed{M = D - E} \text{ y } \boxed{N = F}$$

De esta forma,

$$Au = b \Leftrightarrow (D - E)u = Fu + b \Leftrightarrow u = (D - E)^{-1}Fu + (D - E)^{-1}b$$

que conduce al *método iterativo de Gauss–Seidel* (por puntos)

$$\begin{cases} u^0 \in \mathbf{V} \text{ arbitrario} \\ u^{k+1} = (D - E)^{-1}Fu^k + (D - E)^{-1}b, \; k \in \mathbb{N} \cup \{0\} \end{cases}$$

o, en forma equivalente,

$$\begin{cases} u^0 \in \mathbf{V} \text{ arbitrario} \\ (D - E)u^{k+1} = Fu^k + b, \; k \in \mathbb{N} \cup \{0\}. \end{cases}$$

Nótese que, por (5.6), la matriz $M = D - E$ es inversible. La matriz de este método es

$$\boxed{\mathcal{L}_1 = (D - E)^{-1}F = I - (D - E)^{-1}A}$$

que se denomina *matriz de Gauss–Seidel* (por puntos).

Contrariamente a lo que sucedía en el método de Jacobi, las n componentes del vector u^{k+1} deben obtenerse de manera sucesiva a partir de las componentes ya calculadas de u^{k+1} y las restantes del vector u^k; por ello, a este método a veces se le denomina método de las *iteraciones sucesivas*. Además, según lo dicho anteriormente, el método de Gauss–Seidel será, en principio, más "rápido" pues la matriz M contiene más elementos de A. Detallaremos algunos aspectos sobre la implementación efectiva de este método en la sección 5.5.

5.3.3. Método de relajación

La idea que subyace en los métodos de relajación es tomar como valor siguiente, en cada paso de un método iterativo, no el que resultaría de aplicar directamente el método, sino una media ponderada entre éste y el valor anteriormente hallado, es decir,

$$\boxed{\text{Valor anterior } u^k} \Longrightarrow \boxed{\text{Método } u^{k+1}} \Longrightarrow \boxed{\text{Valor siguiente } wu^{k+1} + (1-w)u^k}$$

para un factor de peso $w \neq 0$. Así, por ejemplo, aplicando esta estrategia al método de Jacobi se obtiene

$$u^{k+1} = wu_{\mathcal{J}}^{k+1} + (1-w)u^k,\ w \neq 0$$

donde $u_{\mathcal{J}}^{k+1}$ es el valor obtenido al realizar una iteración en el método de Jacobi a partir de u^k. En términos de coordenadas, tendríamos:

$$u_i^{k+1} = \frac{w}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}u_j^k - \sum_{j=i+1}^{n} a_{ij}u_j^k\right) + (1-w)u_i^k \tag{5.9}$$

para $i = 1, 2, \ldots, n$, lo que matricialmente se escribe como

$$\begin{aligned} u^{k+1} &= wD^{-1}\left(b + (E+F)u^k\right) + (1-w)u^k \\ &= wD^{-1}\left(\frac{1-w}{w}D + E + F\right)u^k + wD^{-1}b. \end{aligned} \tag{5.10}$$

Este método, conocido como *método de relajación–Jacobi*, no se utiliza apenas debido a que no constituye una mejora sustancial del método de Jacobi. No obstante, nos servirá para deducir un método (conocido simplemente como *método de relajación*) con muy buenas propiedades. Para obtenerlo, razonaremos de forma análoga a como lo hicimos en la deducción del método de Gauss–Seidel a partir del método de Jacobi. A la vista de las ecuaciones dadas en (5.9) es razonable pensar que los resultados se mejorarían si usáramos cada coordenada de u^{k+1} desde el primer momento en que se haya calculado. Esto conduciría a las ecuaciones

$$u_i^{k+1} = \frac{w}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}u_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}u_j^k\right) + (1-w)u_i^k$$

para $i = 1, 2, \ldots, n$, lo que, en términos matriciales, es

$$u^{k+1} = wD^{-1}\left(b + Eu^{k+1} + Fu^k\right) + (1-w)u^k.$$

Agrupando, se tiene que

$$(D - wE)u^{k+1} = ((1-w)D + wF)\,u^k + wb$$

o, equivalentemente,

$$\left(\frac{D}{w} - E\right)u^{k+1} = \left(\frac{1-w}{w}D + F\right)u^k + b.$$

La matriz A puede ser escrita como $A = M - N$ siendo

$$\boxed{M = \frac{D}{w} - E} \text{ y } \boxed{N = \frac{1-w}{w}D + F}$$

Por tanto,

$$\begin{aligned} Au = b \Leftrightarrow & \left(\frac{D}{w} - E\right)u = \left(\frac{1-w}{w}D + F\right)u + b \\ \Leftrightarrow & \; u = \left(\frac{D}{w} - E\right)^{-1}\left(\frac{1-w}{w}D + F\right)u + \left(\frac{D}{w} - E\right)^{-1} b, \end{aligned}$$

lo que conduce al *método iterativo de relajación* (por puntos)

$$\begin{cases} u^0 \in \mathbf{V} \text{ arbitrario} \\ u^{k+1} = \left(\frac{D}{w} - E\right)^{-1}\left(\frac{1-w}{w}D + F\right)u^k + \left(\frac{D}{w} - E\right)^{-1} b, \;\; k \in \mathbb{N} \cup \{0\} \end{cases}$$

o, equivalentemente,

$$\begin{cases} u^0 \in \mathbf{V} \text{ arbitrario} \\ \left(\frac{D}{w} - E\right)u^{k+1} = \left(\frac{1-w}{w}D + F\right)u^k + b, \;\; k \in \mathbb{N} \cup \{0\}. \end{cases}$$

La hipótesis (5.6) hace que la matriz $M = \dfrac{D}{w} - E$ con $w \neq 0$ sea inversible. La matriz de este método es

$$\boxed{\mathcal{L}_w = \left(\frac{D}{w} - E\right)^{-1}\left(\frac{1-w}{w}D + F\right) = (D - wE)^{-1}\left((1-w)D + wF\right)}$$

denominada *matriz de relajación* (por puntos). Este método también se denomina en muchas ocasiones *método SOR* (de *sucessive overrelaxation*, sobrerrelajación sucesiva, en inglés). Algunos autores distinguen y denominan *sobrerrelajación* cuando $w > 1$ y *subrelajación* si $w < 1$. Nótese que para $w = 1$ se tiene el método de Gauss–Seidel, lo que hace coherente la notación $\mathcal{L}_1$ para la matriz asociada al mismo.

Observación 5.3. Un estudio detallado del método de relajación (como puede verse en [**Ci**] o [**La–Th**]) consistiría en hallar un intervalo $I \subset \mathbb{R}$ tal que

$$0 \notin I \ \text{ y } \ \varrho(\mathcal{L}_w) < 1,\ w \in I$$

y en determinar el *parámetro óptimo* $w_0 \in I$ dado por

$$\varrho(\mathcal{L}_{w_0}) = \inf_{w \in I} \varrho(\mathcal{L}_w).$$

En particular, cuando la matriz A es hermítica definida positiva y tridiagonal, el parámetro óptimo w_0 es

$$w_0 = \frac{2}{1 + \sqrt{1 - \varrho(\mathcal{L}_1)}} = \frac{2}{1 + \sqrt{1 - (\varrho(\mathcal{J}))^2}}. \quad \square$$

Veamos a continuación algunos ejemplos que ponen de manifiesto que, en general, la conveniencia de utilizar un método u otro está ligada al problema, lo que no permite asegurar que un método iterativo sea siempre mejor que otro.

Ejemplo 5.1. Consideremos la matriz

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

donde $\alpha \in \mathbb{R}$. Claramente, $A = D - E - F$ siendo

$$D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix},\ E = \begin{pmatrix} 0 & 0 & 0 \\ -2 & 0 & 0 \\ -\alpha & 0 & 0 \end{pmatrix} \ \text{ y } \ F = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

A partir de la definición,

$$\mathcal{J} = D^{-1}(E+F) = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 2 & 0 \\ -2 & 0 & 1 \\ -\alpha & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{2}{3} & 0 & \frac{1}{3} \\ -\frac{\alpha}{2} & 0 & 0 \end{pmatrix}$$

y

$$\mathcal{L}_1 = (D-E)^{-1}F = \begin{pmatrix} 2 & 0 & 0 \\ 2 & 3 & 0 \\ \alpha & 0 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ -\frac{1}{3} & \frac{1}{3} & 0 \\ -\frac{\alpha}{4} & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -\frac{2}{3} & \frac{1}{3} \\ 0 & -\frac{\alpha}{2} & 0 \end{pmatrix}.$$

En la tabla 5.1 se dan los radios espectrales de las matrices de los métodos de Jacobi y Gauss–Seidel para algunos valores concretos del parámetro α.

TABLA 5.1:
Radios espectrales de las matrices $\mathcal{J}$ y $\mathcal{L}_1$

α	$\varrho(\mathcal{J})$	$\varrho(\mathcal{L}_1)$
-1	0.84865653915700	0.86037961002806
-3	0.97263258335935	1.11506929330390
-5	1.08264845639125	1.30515864914088

De esta forma, a partir de los resultados del teorema 5.1 se concluye que para $\alpha = -1$ ambos métodos son convergentes, para $\alpha = -3$ el método de Jacobi converge y el de Gauss–Seidel diverge, mientras que para $\alpha = -5$ los dos métodos son divergentes. □

Ejemplo 5.2. Consideremos la matriz

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

donde $\alpha \in \mathbb{R}$. Podemos escribir, $A = D - E - F$ siendo

$$D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix},\ E = \begin{pmatrix} 0 & 0 & 0 \\ -2 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \text{ y } F = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & -\alpha \\ 0 & 0 & 0 \end{pmatrix}.$$

Así,

$$\mathcal{J} = D^{-1}(E+F) = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 2 & 0 \\ -2 & 0 & -\alpha \\ -1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{2}{3} & 0 & -\frac{\alpha}{3} \\ -\frac{1}{2} & 0 & 0 \end{pmatrix}$$

y

$$\mathcal{L}_1 = (D-E)^{-1}F = \begin{pmatrix} 2 & 0 & 0 \\ 2 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & -\alpha \\ 0 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ -\frac{1}{3} & \frac{1}{3} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & -\alpha \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -\frac{2}{3} & -\frac{\alpha}{3} \\ 0 & -\frac{1}{2} & 0 \end{pmatrix}.$$

En la tabla 5.2 se muestran los radios espectrales de las matrices de los métodos de Jacobi y Gauss–Seidel para algunos valores del parámetro α.

TABLA 5.2:
Radios espectrales de las matrices $\mathcal{J}$ y $\mathcal{L}_1$

α	$\varrho(\mathcal{J})$	$\varrho(\mathcal{L}_1)$
-1	0.84865653915700	0.40824829046386
-4	1.03018084965341	0.81649658092773
-7	1.17502381317383	1.08012344973464

Las conclusiones que se obtienen ahora, a partir del teorema 5.1, es que para $\alpha = -1$ ambos métodos son convergentes, para $\alpha = -4$ el método de Jacobi diverge y el de Gauss–Seidel converge mientras que para $\alpha = -7$ los dos métodos son divergentes. □

5.3.4. Métodos por bloques

Hasta ahora sólo hemos considerado la descomposición $D - E - F$ por puntos de A y, a partir de ella, hemos definido los diversos métodos iterativos. Un proceso análogo puede seguirse para obtener métodos iterativos por bloques: supongamos

la matriz A del sistema lineal descompuesta por bloques en la forma

$$\boxed{A = D - E - F}$$

donde

$$D = \begin{pmatrix} A_{11} & & & & \\ & A_{22} & & & \\ & & A_{33} & & \\ & & & \ddots & \\ & & & & A_{pp} \end{pmatrix},$$

$$E = \begin{pmatrix} & & & & \\ -A_{21} & & & & \\ -A_{31} & -A_{32} & & & \\ \cdots & \cdots & \ddots & & \\ -A_{p1} & -A_{p2} & \cdots & -A_{p,p-1} & \end{pmatrix}$$

y

$$F = \begin{pmatrix} & -A_{12} & -A_{13} & \cdots & -A_{1p} \\ & & -A_{23} & \cdots & -A_{2p} \\ & & & \ddots & \cdots \\ & & & & -A_{p-1,p} \\ & & & & \end{pmatrix}.$$

Si la matriz D es inversible (o, equivalentemente, si las matrices diagonales A_{ii} son inversibles para todo $i = 1, 2, \ldots, p$) se definen los métodos iterativos y las matrices de Jacobi, Gauss–Seidel y de relajación por bloques (asociadas a la descomposición por bloques $A = D - E - F$) de modo análogo, donde las letras D, E y F representan, ahora, las "nuevas" matrices que aparecen en las descomposición por bloques considerada.

Así, por ejemplo, una iteración del método de Jacobi por bloques toma la forma

$$\begin{aligned} A_{ii}u_i^{k+1} &= b_i - A_{11}u_1^k - \cdots - A_{i,i-1}u_{i-1}^k - A_{i,i+1}u_{i+1}^k - \cdots - A_{ip}u_p^k \\ &= b_i - \sum_{j=1}^{i-1} A_{ij}u_j^k - \sum_{j=i+1}^{p} A_{ij}u_j^k \end{aligned}$$

para $i = 1, 2, \ldots, p$ (nótese que, ahora, u_i^k, u_i^{k+1} y b_i son vectores) donde habrá que resolver los p sistemas asociados a los bloques A_{ii}, $i = 1, 2, \ldots, p$.

Según hemos visto en la observación 5.2, los métodos por bloques parece que serán mejores (es decir, más rápidos) que los métodos por puntos correspondientes,

ya que la matriz M tiene más elementos de A; pero hay que tener en cuenta también que para hallar las n_i–tuplas de coordenadas de u^{k+1} en cada iteración deben resolverse p sistemas lineales cuyas matrices son las submatrices A_{ii}, $i = 1, 2, \ldots, p$. Cada uno de estos p sistemas se resolverá mediante un método directo; como la matriz A_{ii} permanece de una iteración a otra y sólo cambia el segundo miembro, tendremos que resolver para cada $i \in \{1, 2, \ldots, p\}$ varios sistemas (uno para cada iteración) con la misma matriz. De ahí el énfasis que hemos puesto en el capítulo anterior en implementar los métodos directos de forma que se factorice de una vez para siempre la matriz del sistema y, para cada segundo miembro, baste resolver dos sistemas triangulares.

Nuevamente, habrá que atender al compromiso entre las dos estrategias contrapuestas, de forma que se utilizará un método por bloques, preferentemente al método por puntos correspondiente, sólo cuando el aumento en la duración de una iteración (debido a la resolución de los p sistemas lineales) esté suficientemente compensado por la aceleración de la convergencia.

Veamos un ejemplo en el que se muestra que los métodos iterativos por bloques son, en general, mejores que los métodos iterativos por puntos (aunque no siempre es así, como puede verse en el problema 5.2).

Ejemplo 5.3. En el ejemplo 5.1 se mostró que para el valor $\alpha = -5$ los métodos de Jacobi y Gauss–Seidel asociados a la matriz

$$A = \begin{pmatrix} 2 & -2 & 0 \\ 2 & 3 & -1 \\ -5 & 0 & 2 \end{pmatrix}$$

eran ambos divergentes. Veamos qué ocurre cuando se hacen las siguientes descomposiciones en bloques de la matriz A:

- Cuando $n_1 = 1$ y $n_2 = 2$ escribimos la matriz A como

$$A = \left(\begin{array}{c|cc} 2 & -2 & 0 \\ \hline 2 & 3 & -1 \\ -5 & 0 & 2 \end{array}\right).$$

  En este caso,

$$D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & -1 \\ 0 & 0 & 2 \end{pmatrix}, \; E = \begin{pmatrix} 0 & 0 & 0 \\ -2 & 0 & 0 \\ 5 & 0 & 0 \end{pmatrix} \text{ y } F = \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Así, las matrices de los métodos de Jacobi y Gauss–Seidel vienen dadas por

$$\mathcal{J} = D^{-1}(E+F) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & -1 \\ 0 & 0 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 2 & 0 \\ -2 & 0 & 0 \\ 5 & 0 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{6} \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 2 & 0 \\ -2 & 0 & 0 \\ 5 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{6} & 0 & 0 \\ \frac{5}{2} & 0 & 0 \end{pmatrix}$$

y

$$\mathcal{L}_1 = (D-E)^{-1}F = \begin{pmatrix} 2 & 0 & 0 \\ 2 & 3 & -1 \\ -5 & 0 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ \frac{1}{12} & \frac{1}{3} & \frac{1}{6} \\ \frac{5}{4} & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{6} & 0 \\ 0 & \frac{5}{2} & 0 \end{pmatrix}.$$

Los polinomios característicos de $\mathcal{J}$ y $\mathcal{L}_1$ son (compruébese como ejercicio)

$$P_{\mathcal{J}}(\lambda) = \frac{\lambda}{6}\left(1-6\lambda^2\right) \quad \text{y} \quad P_{\mathcal{L}_1}(\lambda) = \frac{\lambda^2}{6}(1-6\lambda).$$

Por tanto,

$$\operatorname{sp}(\mathcal{J}) = \left\{0, \pm\sqrt{\frac{1}{6}}\right\} \quad \text{y} \quad \operatorname{sp}(\mathcal{L}_1) = \left\{0, \frac{1}{6}\right\}$$

y se verifica que

$$\varrho(\mathcal{J}) = \sqrt{\frac{1}{6}} \simeq 0.408248 \quad \text{y} \quad \varrho(\mathcal{L}_1) = \frac{1}{6} = 0.1\,\widehat{6},$$

por lo que, aplicando el teorema 5.1, se concluye que los métodos de Jacobi y Gauss–Seidel son convergentes.

- Si $n_1 = 2$ y $n_2 = 1$ la matriz A toma la forma

$$A = \left(\begin{array}{cc|c} 2 & -2 & 0 \\ 2 & 3 & -1 \\ \hline -5 & 0 & 2 \end{array}\right),$$

es decir, en este caso,

$$D = \begin{pmatrix} 2 & -2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \; E = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 5 & 0 & 0 \end{pmatrix} \text{ y } F = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Ahora las matrices de los métodos de Jacobi y Gauss–Seidel toman la forma

$$\mathcal{J} = D^{-1}(E+F) = \begin{pmatrix} 2 & -2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 5 & 0 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} \frac{3}{10} & \frac{1}{5} & 0 \\ -\frac{1}{5} & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 5 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{5} \\ \frac{5}{2} & 0 & 0 \end{pmatrix}$$

y

$$\mathcal{L}_1 = (D-E)^{-1}F = \begin{pmatrix} 2 & -2 & 0 \\ 2 & 3 & 0 \\ -5 & 0 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} \frac{3}{10} & \frac{1}{5} & 0 \\ -\frac{1}{5} & \frac{1}{5} & 0 \\ \frac{3}{4} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{5} \\ 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

Sus respectivos polinomios característicos son (compruébese)

$$P_{\mathcal{J}}(\lambda) = \frac{\lambda}{2}\left(1-2\lambda^2\right) \text{ y } P_{\mathcal{L}_1}(\lambda) = \frac{\lambda^2}{2}(1-2\lambda).$$

Como

$$\mathrm{sp}(\mathcal{J}) = \left\{0, \pm\sqrt{\frac{1}{2}}\right\} \text{ y } \mathrm{sp}(\mathcal{L}_1) = \left\{0, \frac{1}{2}\right\},$$

entonces

$$\varrho(\mathcal{J}) = \sqrt{\frac{1}{2}} \simeq 0.707107 \text{ y } \varrho(\mathcal{L}_1) = \frac{1}{2} = 0.5,$$

por lo que, aplicando nuevamente el teorema 5.1, se concluye que ambos métodos vuelven a ser convergentes. □

5.4. Resultados de convergencia

El estudio de la convergencia de los métodos iterativos puede ser demasiado prolijo puesto que no existen teoremas que aseguren la convergencia para una clase general de matrices. No obstante, pueden darse resultados parciales para determinados tipos de matrices; aquí presentamos un resultado de carácter general y sendas condiciones de convergencia para el método de relajación y el de Jacobi recogiendo otros resultados en la sección de problemas.

Teorema 5.2. *Sea $A \in \mathcal{M}_n$ una matriz hermítica definida positiva escrita como $A = M - N$ siendo $M \in \mathcal{M}_n$ una matriz inversible. Si la matriz hermítica $M^* + N$ es definida positiva entonces*

$$\varrho(M^{-1}N) < 1.$$

Consecuentemente, en la situación anterior, el método iterativo definido por la matriz $B = M^{-1}N$ es convergente.

Demostración. En primer lugar, por ser A hermítica,

$$\begin{aligned}(M^* + N)^* &= M + N^* = (A + N) + N^* \\ &= (A^* + N) + N^* = (A + N)^* + N = M^* + N\end{aligned}$$

por lo que la matriz $M^* + N$ es hermítica. Por otra parte, sea $\lambda \in \mathrm{sp}(M^{-1}N)$ y $v \in \mathbf{V}\backslash\{0\}$ un autovector asociado al autovalor λ, es decir,

$$M^{-1}Nv = \lambda v. \tag{5.11}$$

A partir de v consideremos el vector

$$w = M^{-1}Nv. \tag{5.12}$$

En primer lugar, nótese que $w \neq v$. En efecto, en caso contrario se obtendría, a partir de (5.12),

$$v = M^{-1}Nv \;\Rightarrow\; Mv = Nv \;\Rightarrow\; Av = (M - N)v = 0,$$

lo que contradice que A sea inversible. Por otra parte, teniendo en cuenta que

$$Mw = Nv,$$

se verifica que

$$\begin{aligned}(v-w)^*(M^*+N)(v-w) &= (v-w)^*M^*(v-w)+(v-w)^*N(v-w)\\ &= (Mv-Mw)^*(v-w)+(v-w)^*(Nv-Nw)\\ &= (Mv-Nv)^*(v-w)+(v-w)^*(Mw-Nw)\\ &= v^*(M-N)^*(v-w)+(v-w)^*(M-N)w\\ &= v^*A^*(v-w)+(v-w)^*Aw\\ &= v^*Av-v^*Aw+v^*Aw-w^*Aw\\ &= v^*Av-w^*Aw\end{aligned}$$

por ser $A = M - N$ una matriz hermítica. Por tanto,

$$v^*Av - w^*Aw = (v-w)^*(M^*+N)(v-w) > 0 \tag{5.13}$$

ya que $w \neq v$ y $M^* + N$ es definida positiva. Ahora bien, a partir de (5.11), (5.12) y (5.13) se obtiene que

$$\begin{aligned}0 < v^*Av - w^*Aw &= v^*Av - (\lambda v)^*A(\lambda v)\\ &= v^*Av - v^*\overline{\lambda}A\lambda v = (1-|\lambda|^2)v^*Av.\end{aligned}$$

Como $v^*Av > 0$ por ser A definida positiva y $v \neq 0$, entonces

$$1 - |\lambda|^2 > 0,$$

de donde

$$|\lambda| < 1,$$

obteniéndose así el resultado. □

A continuación vamos a dar una condición suficiente para la convergencia del método de relajación:

Teorema 5.3 (Ostrowski–Reich). *Si $A \in \mathcal{M}_n$ es una matriz hermítica definida positiva y $0 < w < 2$, entonces el método de relajación (por puntos o bloques) es convergente. En particular, el método de Gauss–Seidel es convergente.*

Demostración. La descomposición $A = M - N$ asociada al método de relajación es

$$A = \left(\frac{D}{w} - E\right) - \left(\frac{1-w}{w}D + F\right),\ w \neq 0.$$

Como la matriz A es hermítica se tiene que

$$D - E - F = A = A^* = D^* - E^* - F^*.$$

Identificando en la igualdad anterior los elementos diagonales y los que quedan en la parte triangular inferior y superior de A, se verifica que $D^* = D$ y $E^* = F$. Por tanto,

$$M^* + N = \frac{D}{w} - E^* + \frac{1-w}{w}D + F = \frac{2-w}{w}D. \tag{5.14}$$

Como la matriz D es definida positiva (véase el problema 2.16), entonces, a partir de la relación (5.14), la matriz $M^* + N$ es definida positiva para valores del parámetro $0 < w < 2$ ya que

$$v^*(M^* + N)v = \frac{2-w}{w}v^*Dv > 0,\ v \in \mathbf{V}\backslash\{0\}$$

al ser

$$\frac{2-w}{w} > 0 \ \Leftrightarrow\ 0 < w < 2.$$

Aplicando el teorema 5.2 concluimos el resultado. □

Veamos ahora que la condición $0 < w < 2$ es necesaria para la convergencia del método de relajación.

Teorema 5.4 (Kahan). *El radio espectral de la matriz de relajación* (*por puntos o bloques*) *siempre verifica*

$$\varrho(\mathcal{L}_w) \geq |1-w|,\ w \neq 0.$$

Consecuentemente, el método de relajación (*por puntos o bloques*) *sólo puede ser convergente cuando* $0 < w < 2$.

Demostración. Por definición,

$$\det(\mathcal{L}_w) = \det\left(\left(\frac{D}{w} - E\right)^{-1}\left(\frac{1-w}{w}D + F\right)\right) = \frac{\det\left(\dfrac{1-w}{w}D + F\right)}{\det\left(\dfrac{D}{w} - E\right)}.$$

Como

$$\det\left(\frac{1-w}{w}D + F\right) = \det\left(\frac{1-w}{w}D\right) \quad \text{y} \quad \det\left(\frac{D}{w} - E\right) = \det\left(\frac{D}{w}\right)$$

(véase el problema 2.15), entonces

$$\det(\mathcal{L}_w) = \frac{\det\left(\dfrac{1-w}{w}D\right)}{\det\left(\dfrac{D}{w}\right)} = \frac{\dfrac{(1-w)^n}{w^n}\det(D)}{\dfrac{1}{w^n}\det(D)} = (1-w)^n. \tag{5.15}$$

Por otra parte, si

$$\mathrm{sp}\,(\mathcal{L}_w) = \{\lambda_i\,(\mathcal{L}_w) :\ i = 1, 2, \ldots, n\}$$

entonces

$$\det(\mathcal{L}_w) = \prod_{i=1}^{n} \lambda_i\,(\mathcal{L}_w) \tag{5.16}$$

(véase la observación 2.10). Así, usando (5.15) y (5.16) se obtiene que

$$\prod_{i=1}^{n} |\lambda_i\,(\mathcal{L}_w)| = |1 - w|^n,$$

lo que permite concluir que

$$\varrho(\mathcal{L}_w) \geq \left(\prod_{i=1}^{n} |\lambda_i\,(\mathcal{L}_w)| \right)^{\frac{1}{n}} = |1 - w|. \quad \square$$

En las aplicaciones aparecen, con mucha frecuencia, matrices $A = (a_{ij})_{i,j=1}^{n}$ de diagonal estrictamente dominante (véase la definición 2.11). Como se vio en el teorema 2.2, estas matrices son inversibles y, además, por la propia definición, verifican

$$a_{ii} \neq 0$$

para $i = 1, 2, \ldots, n$. Para este tipo de matrices se tiene el siguiente resultado de convergencia del método de Jacobi por puntos:

Teorema 5.5. *Si $A \in \mathcal{M}_n$ es una matriz de diagonal estrictamente dominante, el método iterativo de Jacobi por puntos es convergente.*

Demostración. La matriz del método de Jacobi $\mathcal{J} = D^{-1}(E + F)$ verifica que

$$(\mathcal{J})_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}} & \text{si} \quad i \neq j \\ 0 & \text{si} \quad i = j. \end{cases}$$

Por tanto, a partir del teorema 2.3, se tiene que

$$\|\mathcal{J}\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |(\mathcal{J})_{ij}| = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} = \max_{1 \leq i \leq n} \left(\frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \right) < 1.$$

De esta forma, aplicando el teorema 5.1 se concluye el resultado. $\square$

Veamos a continuación otra demostración del teorema 5.5. Su interés radica en que puede ser adaptada para demostrar otros resultados de convergencia en contextos más generales, como se verá en los problemas resueltos:

DEMOSTRACIÓN ALTERNATIVA. Supongamos que el método de Jacobi para A no es convergente; en tal caso, existe un autovalor $\lambda \in \mathrm{sp}(\mathcal{J})$ con $|\lambda| \geq 1$. De esta forma,

$$\det\left(D^{-1}(E+F) - \lambda I\right) = \det\left(\mathcal{J} - \lambda I\right) = P_{\mathcal{J}}(\lambda) = 0$$

y, por tanto,

$$\det(E + F - \lambda D) = \det(D)\det\left(D^{-1}(E+F) - \lambda I\right) = 0,$$

por lo que la matriz $E + F - \lambda D$ no es inversible y, en concreto, no es de diagonal estrictamente dominante (véase el teorema 2.2). Consecuentemente, existe un índice $i_0 \in \{1, 2, \ldots, n\}$ tal que

$$|-\lambda|\,|a_{i_0 i_0}| \leq \sum_{j=1}^{i_0-1} |-a_{i_0 j}| + \sum_{j=i_0+1}^{n} |-a_{i_0 j}| = \sum_{\substack{j=1 \\ j\neq i_0}}^{n} |a_{i_0 j}|.$$

Como $\lambda \neq 0$ podemos dividir la expresión anterior por $|\lambda|$ obteniendo

$$|a_{i_0 i_0}| \leq \frac{1}{|\lambda|} \sum_{\substack{j=1 \\ j\neq i_0}}^{n} |a_{i_0 j}| \leq \sum_{\substack{j=1 \\ j\neq i_0}}^{n} |a_{i_0 j}|$$

puesto que $|\lambda| \geq 1$. Se llega así a una contradicción con el hecho de que la matriz A sea de diagonal estrictamente dominante. □

5.5. Test de parada de las iteraciones

Como ya se ha dicho, cuando un método iterativo es convergente, la solución del sistema lineal $Au = b$ se obtiene como límite de la sucesión $\{u^k\}_{k=0}^{\infty}$ de iteraciones. Ante la imposibilidad de calcular todas las iteraciones, se plantea el problema de determinar $k \in \mathbb{N}$ para el cual podemos tomar u^k como una "buena" aproximación de u. Es decir, si se desea que el error relativo sea inferior a una cantidad prefijada $\varepsilon > 0$, el valor $k \in \mathbb{N}$ debe cumplir

$$\left|\left|u^k - u\right|\right| < \varepsilon\,||u||$$

para alguna norma vectorial $||\cdot||$. Por supuesto, al ser el vector u desconocido, no se puede trabajar directamente con esas cantidades.

El test más sencillo que podemos emplear es detener el proceso cuando la diferencia entre dos iteraciones consecutivas sea, en términos relativos, menor que la *tolerancia* admisible ε, es decir,

$$\left|\left|u^{k+1}-u^k\right|\right| < \varepsilon \left|\left|u^{k+1}\right|\right|. \tag{5.17}$$

Sin embargo, este test tiene el inconveniente de que puede cumplirse la relación (5.17) sin que el vector u^{k+1} esté próximo a u.

Una condición de parada de las iteraciones más adecuada viene dada a partir del *vector residuo*

$$r^k = b - Au^k = A(u-u^k),\ k \in \mathbb{N} \cup \{0\}.$$

Es razonable pensar que si $u^k \simeq u$ entonces $Au^k \simeq b$ y recíprocamente. Por tanto, pararemos las iteraciones cuando

$$\frac{\left|\left|r^k\right|\right|}{||b||} = \frac{\left|\left|Au^k - Au\right|\right|}{||Au||} < \varepsilon,$$

es decir, para valores de $k \in \mathbb{N}$ verificando

$$\left|\left|r^k\right|\right| < \varepsilon\, ||b||.$$

Obviamente debe procurarse que la comprobación de los tests de parada no incremente en exceso el número de operaciones necesarias para realizar una iteración. Veamos cómo organizando los cálculos de forma adecuada esto puede conseguirse tanto en el método de Jacobi como en el de relajación:

a) En el método de Jacobi podemos reescribir la iteración como

$$Du^{k+1} = b + (E+F)u^k = b - Au^k + Du^k = r^k + Du^k,$$

es decir,

$$D(u^{k+1}-u^k) = r^k.$$

De esta forma, calculando en primer lugar el vector r^k, resolviendo a continuación el sistema $Dd^k = r^k$ y tomando

$$u^{k+1} = u^k + d^k$$

obtenemos la información necesaria para los tests de parada así como la iteración siguiente u^{k+1} sin haber incrementado sustancialmente el número

de operaciones. En el caso particular del método de Jacobi por puntos, para cada $i \in \{1, 2, \ldots, n\}$, se calculan

$$\boxed{\begin{aligned} r_i^k &= b_i - \sum_{j=1}^{n} a_{ij} u_j^k \\ d_i^k &= \frac{r_i^k}{a_{ii}} \\ u_i^{k+1} &= u_i^k + d_i^k \end{aligned}}$$

b) En el método de relajación podemos reescribir la iteración como

$$\left(\frac{D}{w} - E\right) u^{k+1} = \left(\frac{1-w}{w} D + F\right) u^k + b,$$

es decir,

$$\frac{D}{w} u^{k+1} = E u^{k+1} - D u^k + F u^k + \frac{D}{w} u^k + b = \widetilde{r}^k + \frac{D}{w} u^k,$$

siendo

$$\widetilde{r}^k = b - \left((D - F) u^k - E u^{k+1}\right)$$

y, de esta forma,

$$D(u^{k+1} - u^k) = w \widetilde{r}^k.$$

Aunque el vector $\widetilde{r}^k$ no coincide con el vector residuo

$$r^k = b - A u^k = b - (D - E - F) u^k$$

puede demostrarse (véase [**La–Th**]) que el test de parada

$$\left|\left|\widetilde{r}^k\right|\right| < \varepsilon \, ||b||$$

sigue siendo válido. En el caso particular del método de relajación por puntos se tiene que

$$\widetilde{r}_i^k = b_i - \left(A u^{k,i}\right)_i$$

para $i = 1, 2, \ldots, n$, donde

$$u^{k,i} = \left(u_1^{k+1}, u_2^{k+1}, \ldots, u_{i-1}^{k+1}, u_i^k, u_{i+1}^k, \ldots, u_n^k\right)^{\mathrm{T}}.$$

Es decir, para cada $i \in \{1, 2, \ldots, n\}$, se calculan

$$\boxed{\begin{aligned} \widetilde{r}_i^k &= b_i - \sum_{j=1}^{i-1} a_{ij} u_j^{k+1} - \sum_{j=i}^{n} a_{ij} u_j^k \\ d_i^k &= w \frac{\widetilde{r}_i^k}{a_{ii}} \\ u_i^{k+1} &= u_i^k + d_i^k \end{aligned}}$$

Para acabar, simplemente reseñar que las normas vectoriales que suelen emplearse con mayor frecuencia en este tipo de tests son $||\cdot||_2$ y $||\cdot||_\infty$.

5.6. Problemas

5.6.1. Problemas resueltos

5.1. Estudiar la convergencia de los métodos de Jacobi y Gauss–Seidel por puntos para las matrices

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} \text{ y } \widetilde{A} = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}.$$

SOLUCIÓN.

a) Para la matriz A se tiene que

$$\mathcal{J} = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix} \text{ y } \mathcal{L}_1 = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{pmatrix},$$

de donde se obtiene que

$$P_{\mathcal{J}}(\lambda) = -\lambda^3 \text{ y } P_{\mathcal{L}_1}(\lambda) = -\lambda(2-\lambda)^2.$$

De esta forma, como $\mathrm{sp}(\mathcal{J}) = \{0\}$, entonces $\varrho(\mathcal{J}) = 0 < 1$, por lo que el método de Jacobi para la matriz A es convergente, mientras que al ser $\mathrm{sp}(\mathcal{L}_1) = \{0, 2\}$ se tiene que $\varrho(\mathcal{L}_1) = 2 > 1$ y, consecuentemente, el método de Gauss–Seidel para A no es convergente (véase el teorema 5.1).

b) En el caso de la matriz $\widetilde{A}$ se verifica que

$$\mathcal{J} = \begin{pmatrix} 0 & 0.5 & -0.5 \\ -1 & 0 & -1 \\ 0.5 & 0.5 & 0 \end{pmatrix} \text{ y } \mathcal{L}_1 = \begin{pmatrix} 0 & 0.5 & -0.5 \\ 0 & -0.5 & -0.5 \\ 0 & 0 & 0.5 \end{pmatrix}.$$

De esta forma,

$$P_{\mathcal{J}}(\lambda) = -\lambda\left(\lambda^2 + \frac{5}{4}\right) \quad \text{y} \quad P_{\mathcal{L}_1}(\lambda) = -\lambda\left(\frac{1}{2} + \lambda\right)^2.$$

Así, como $\varrho(\mathcal{J}) = \dfrac{\sqrt{5}}{2} > 1$, el método de Jacobi para la matriz $\widetilde{A}$ no es convergente, mientras al ser $\varrho(\mathcal{L}_1) = \dfrac{1}{2} < 1$ el método de Gauss–Seidel para $\widetilde{A}$ es convergente. □

5.2. Estudiar la convergencia de los métodos de Jacobi y Gauss–Seidel cuando se consideran descomposiciones en bloques de la matriz A del problema 5.1 de tamaños $n_1 = 1$, $n_2 = 2$ y $n_1 = 2$, $n_2 = 1$, respectivamente.

SOLUCIÓN.

a) Cuando $n_1 = 1$ y $n_2 = 2$ se está considerando la matriz

$$A = \left(\begin{array}{c|cc} 1 & 2 & -2 \\ \hline 1 & 1 & 1 \\ 2 & 2 & 1 \end{array}\right).$$

En este caso,

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 2 & 1 \end{pmatrix}, \; E = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ -2 & 0 & 0 \end{pmatrix} \quad \text{y} \quad F = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Así se obtienen las matrices

$$\mathcal{J} = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{y} \quad \mathcal{L}_1 = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -2 \\ 0 & 0 & 0 \end{pmatrix}$$

y, a partir de ellas, sus polinomios característicos

$$P_{\mathcal{J}}(\lambda) = \lambda(2 - \lambda^2) \quad \text{y} \quad P_{\mathcal{L}_1}(\lambda) = \lambda^2(2 - \lambda).$$

Consecuentemente,

$$\begin{cases} \operatorname{sp}(\mathcal{J}) = \{0, \pm\sqrt{2}\} & \Rightarrow \quad \varrho(\mathcal{J}) = \sqrt{2} > 1 \\ \operatorname{sp}(\mathcal{L}_1) = \{0, 2\} & \Rightarrow \quad \varrho(\mathcal{L}_1) = 2 > 1 \end{cases} \tag{5.18}$$

por lo que, a raíz del teorema 5.1, ninguno de los dos métodos es convergente.

b) Cuando $n_1 = 2$ y $n_2 = 1$ se tiene que

$$A_1 = \left(\begin{array}{cc|c} 1 & 2 & -2 \\ 1 & 1 & 1 \\ \hline 2 & 2 & 1 \end{array}\right).$$

Ahora

$$D = \begin{pmatrix} 1 & 2 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ E = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -2 & -2 & 0 \end{pmatrix} \text{ y } F = \begin{pmatrix} 0 & 0 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix},$$

por lo que

$$\mathcal{J} = \begin{pmatrix} 0 & 0 & -4 \\ 0 & 0 & 3 \\ -2 & -2 & 0 \end{pmatrix} \text{ y } \mathcal{L}_1 = \begin{pmatrix} 0 & 0 & -4 \\ 0 & 0 & 3 \\ 0 & 0 & 2 \end{pmatrix}.$$

Se comprueba que

$$P_{\mathcal{J}}(\lambda) = \lambda(2 - \lambda^2) \text{ y } P_{\mathcal{L}_1}(\lambda) = \lambda^2(2 - \lambda),$$

con lo que se vuelve a obtener (5.18) y, en consecuencia, ambos métodos son divergentes. □

5.3. Se considera una matriz

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \in \mathcal{M}_2$$

verificando

$$a_{11} \neq 0 \text{ y } a_{22} \neq 0.$$

a) Demostrar que los métodos de Jacobi y Gauss–Seidel para A convergen o divergen simultáneamente.

b) Encontrar una condición necesaria y suficiente sobre los elementos de la matriz A para que ambos converjan.

c) En el caso de que ambos métodos converjan, ¿cuál lo hace más rápidamente? Justificar la respuesta.

Solución. Consideremos la descomposición $D - E - F$ de la matriz A siendo

$$D = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix},\ E = \begin{pmatrix} 0 & 0 \\ -a_{21} & 0 \end{pmatrix} \text{ y } F = \begin{pmatrix} 0 & -a_{12} \\ 0 & 0 \end{pmatrix}.$$

Las matrices de los métodos de Jacobi y Gauss–Seidel son, respectivamente,

$$\mathcal{J} = D^{-1}(E+F) = \begin{pmatrix} \dfrac{1}{a_{11}} & 0 \\ 0 & \dfrac{1}{a_{22}} \end{pmatrix} \begin{pmatrix} 0 & -a_{12} \\ -a_{21} & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\dfrac{a_{12}}{a_{11}} \\ -\dfrac{a_{21}}{a_{22}} & 0 \end{pmatrix}$$

y

$$\mathcal{L}_1 = (D-E)^{-1}F = \begin{pmatrix} \dfrac{1}{a_{11}} & 0 \\ -\dfrac{a_{21}}{a_{11}a_{22}} & \dfrac{1}{a_{22}} \end{pmatrix} \begin{pmatrix} 0 & -a_{12} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\dfrac{a_{12}}{a_{11}} \\ 0 & \dfrac{a_{12}a_{21}}{a_{11}a_{22}} \end{pmatrix}.$$

Por tanto,

$$\det(\mathcal{J}-\lambda I) = \lambda^2 - \frac{a_{12}a_{21}}{a_{11}a_{22}} \;\text{ y }\; \det(\mathcal{L}_1-\lambda I) = -\lambda\left(\frac{a_{12}a_{21}}{a_{11}a_{22}} - \lambda\right).$$

Consecuentemente, se tiene que

$$\det(\mathcal{J}-\lambda I) = 0 \;\Leftrightarrow\; \lambda = \pm\sqrt{\frac{a_{12}a_{21}}{a_{11}a_{22}}}$$

y

$$\det(\mathcal{L}_1-\lambda I) = 0 \;\Leftrightarrow\; \begin{cases} \lambda_1 = 0 \\ \lambda_2 = \dfrac{a_{12}a_{21}}{a_{11}a_{22}} \end{cases}$$

por lo que los radios espectrales de las matrices $\mathcal{J}$ y $\mathcal{L}_1$ vienen dados por

$$\varrho(\mathcal{J}) = +\sqrt{\left|\frac{a_{12}a_{21}}{a_{11}a_{22}}\right|} \;\text{ y }\; \varrho(\mathcal{L}_1) = \left|\frac{a_{12}a_{21}}{a_{11}a_{22}}\right| \tag{5.19}$$

a) Como se observa,

$$\varrho(\mathcal{L}_1) = (\varrho(\mathcal{J}))^2,$$

por lo que se verifica que

$$\varrho(\mathcal{L}_1) < 1 \;\Leftrightarrow\; \varrho(\mathcal{J}) < 1.$$

De esta forma, ambos métodos convergen o divergen simultáneamente.

b) A la vista de (5.19) la condición necesaria y suficiente para que ambos métodos converjan es que se verifique la desigualdad

$$|a_{12}a_{21}| < |a_{11}a_{22}|. \tag{5.20}$$

c) En el caso de que se satisfaga la condición (5.20) se tendrá que

$$\varrho(\mathcal{L}_1) = (\varrho(\mathcal{J}))^2 < \varrho(\mathcal{J}),$$

por lo que el método de Gauss–Seidel convergerá más rápidamente que el método de Jacobi. □

5.4. Se considera una matriz $A \in \mathcal{M}_n$ descompuesta en bloques de la forma

$$A = \left(\begin{array}{c|c} A_1 & A_2 \\ \hline A_3 & A_4 \end{array}\right) \tag{5.21}$$

donde las matrices A_1 y A_4 son inversibles.

a) Demostrar que

$$\det(A) = \det(A_1)\det\left(A_4 - A_3A_1^{-1}A_2\right).$$

b) Utilizar el apartado *a*) para probar que para todo $\lambda \in \mathbb{C}\backslash\{0\}$ se verifica que

$$\det\left(\begin{array}{c|c} \lambda^2 A_1 & A_2 \\ \hline \lambda^2 A_3 & \lambda^2 A_4 \end{array}\right) = \det\left(\begin{array}{c|c} \lambda^2 A_1 & \lambda A_2 \\ \hline \lambda A_3 & \lambda^2 A_4 \end{array}\right).$$

c) Se considera la descomposición $D - E - F$ por bloques de la matriz A asociada a la descomposición (5.21) y los métodos de Jacobi y Gauss–Seidel por bloques correspondientes. Probar que para todo $\lambda \in \mathbb{C}\backslash\{0\}$ se verifica que

$$\lambda^n \det(-D) P_{\mathcal{J}}(\lambda) = \det(E - D) P_{\mathcal{L}_1}(\lambda^2)$$

donde $P_{\mathcal{J}}$ y $P_{\mathcal{L}_1}$ son, respectivamente, los polinomios característicos de las matrices de los métodos de Jacobi y Gauss–Seidel.

d) Encontrar la relación existente entre $\varrho(\mathcal{J})$ y $\varrho(\mathcal{L}_1)$. Deducir que ambos métodos convergen o divergen simultáneamente.

e) Si la matriz A es, además, hermítica y definida positiva, demostrar que los métodos de Jacobi y Gauss–Seidel asociados a esta descomposición por bloques son convergentes. ¿Cuál lo hace más rápidamente?

Solución.

a) Vamos a utilizar el método de eliminación gaussiana por bloques para anular el bloque ocupado por A_3. Para ello, consideramos la matriz

$$E_1 = \left(\begin{array}{c|c} I_1 & \mathbf{0} \\ \hline -A_3A_1^{-1} & I_2 \end{array}\right)$$

de forma que la matriz

$$E_1 A = \left(\begin{array}{c|c} A_1 & A_2 \\ \hline \mathbf{0} & A_4 - A_3 A_1^{-1} A_2 \end{array}\right)$$

es triangular superior por bloques. Así pues, como $\det(E_1) = 1$, se tiene que

$$\det(A) = \det(E_1 A) = \det(A_1) \det\left(A_4 - A_3 A_1^{-1} A_2\right).$$

b) Sea $\lambda \in \mathbb{C}^n \backslash \{0\}$. Aplicando dos veces el apartado *a*) se tiene que

$$\begin{aligned} \det\left(\begin{array}{c|c} \lambda^2 A_1 & A_2 \\ \hline \lambda^2 A_3 & \lambda^2 A_4 \end{array}\right) &= \det\left(\lambda^2 A_1\right) \det\left(\lambda^2 A_4 - \lambda^2 A_3 \left(\lambda^2 A_1\right)^{-1} A_2\right) \\ &= \det\left(\lambda^2 A_1\right) \det\left(\lambda^2 A_4 - (\lambda A_3) \left(\lambda^2 A_1\right)^{-1} (\lambda A_2)\right) \\ &= \det\left(\begin{array}{c|c} \lambda^2 A_1 & \lambda A_2 \\ \hline \lambda A_3 & \lambda^2 A_4 \end{array}\right). \end{aligned}$$

c) Sea $\lambda \in \mathbb{C}^n \backslash \{0\}$. Por una parte se tiene que

$$\begin{aligned} \lambda^n \det(-D) P_{\mathcal{J}}(\lambda) &= \lambda^n \det(-D) \det(D^{-1}(E+F) - \lambda I) \\ &= \lambda^n \det(\lambda D - E - F) = \lambda^n \det\left(\begin{array}{c|c} \lambda A_1 & A_2 \\ \hline A_3 & \lambda A_4 \end{array}\right) \\ &= \det\left(\begin{array}{c|c} \lambda^2 A_1 & \lambda A_2 \\ \hline \lambda A_3 & \lambda^2 A_4 \end{array}\right) \end{aligned}$$

y, por otra,

$$\begin{aligned} \det(E-D) P_{\mathcal{L}_1}(\lambda^2) &= \det(E-D) \det((D-E)^{-1} F - \lambda^2 I) \\ &= \det(\lambda^2 (D-E) - F) = \det\left(\begin{array}{c|c} \lambda^2 A_1 & A_2 \\ \hline \lambda^2 A_3 & \lambda^2 A_4 \end{array}\right), \end{aligned}$$

por lo que el apartado *b*) concluye el resultado.

d) Vamos a demostrar que se verifica que

$$\varrho(\mathcal{L}_1) = \left(\varrho(\mathcal{J})\right)^2. \tag{5.22}$$

Para ello, probaremos que para todo $\lambda \in \mathbb{C} \backslash \{0\}$

$$\lambda \in \mathrm{sp}(\mathcal{J}) \;\Leftrightarrow\; \lambda^2 \in \mathrm{sp}(\mathcal{L}_1).$$

En efecto:

$\boxed{\Rightarrow}$ Si $\lambda \in \mathrm{sp}(\mathcal{J})$ entonces $P_{\mathcal{J}}(\lambda) = 0$ y, por el apartado c), se tiene que $P_{\mathcal{L}_1}(\lambda^2) = 0$; es decir, $\lambda^2 \in \mathrm{sp}(\mathcal{L}_1)$.

$\boxed{\Leftarrow}$ Si $\lambda^2 \in \mathrm{sp}(\mathcal{L}_1)$ entonces $P_{\mathcal{L}_1}(\lambda^2) = 0$ y, nuevamente por el apartado c), $P_{\mathcal{J}}(\lambda) = 0$; esto es, $\lambda \in \mathrm{sp}(\mathcal{J})$.

Para acabar, basta tener en cuenta, gracias a la equivalencia que hemos demostrado, que no puede ocurrir que una de las dos matrices tenga $\lambda = 0$ como único autovalor y la otra no. Por tanto, se verifica (5.22) y, consecuentemente,

$$\varrho(\mathcal{J}) < 1 \;\Leftrightarrow\; \varrho(\mathcal{L}_1) < 1,$$

por lo que ambos métodos convergen o divergen simultáneamente.

e) Como se demostró en el teorema 5.3, el método de Gauss–Seidel por bloques es convergente. Aplicando el apartado *d*) se concluye que el método de Jacobi por bloques es también convergente. Además, como

$$\varrho(\mathcal{J}) < 1 \;\text{ y }\; \varrho(\mathcal{L}_1) = \left(\varrho(\mathcal{J})\right)^2$$

se deduce que

$$\varrho(\mathcal{L}_1) < \varrho(\mathcal{J}) < 1,$$

por lo que el método de Gauss–Seidel converge más rápidamente que el de Jacobi. □

5.5. A partir de un vector $u^0 \in \mathbf{V}$ dado, se considera el método iterativo

$$u^{k+1} = Bu^k + c.$$

Estudiar el comportamiento de la sucesión $\{u^k\}_{k=0}^{\infty}$ cuando $\varrho(B) = 0$.

Solución. Como $\varrho(B) = 0$ entonces $\mathrm{sp}(B) = \{0\}$, es decir, $\lambda = 0$ es el único autovalor de B con multiplicidad n. Por el teorema 2.1 sabemos que existe $U \in \mathcal{M}_n$ unitaria tal que $T = U^*BU$ es una matriz triangular superior con ceros en la diagonal. De esta forma, $B = UTU^*$ y, por tanto,

$$B^k = UT^kU^*,\ k \in \mathbb{N}.$$

Ahora bien, como

$$T^2 = \begin{pmatrix} 0 & 0 & \times & \times & \cdots & \times & \times \\ & 0 & 0 & \times & \ddots & & \times \\ & & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & & 0 & 0 & \times & \times \\ & & & & 0 & 0 & \times \\ & & & & & 0 & 0 \\ & & & & & & 0 \end{pmatrix}, \; T^3 = \begin{pmatrix} 0 & 0 & 0 & \times & \cdots & \times & \times \\ & 0 & 0 & 0 & \ddots & & \times \\ & & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & & 0 & 0 & 0 & \times \\ & & & & 0 & 0 & 0 \\ & & & & & 0 & 0 \\ & & & & & & 0 \end{pmatrix},$$

y, así sucesivamente, se tiene que

$$T^k = 0 \text{ si } k \geq n.$$

Consecuentemente,

$$B^k = UT^kU^* = 0 \text{ si } k \geq n.$$

Así pues, a partir del vector

$$e^k = u^k - u,\ k \in \mathbb{N} \cup \{0\},$$

donde $u \in \mathbf{V}$ es la solución del sistema lineal $u = Bu + c$, se verifica que

$$e^k = B^k e^0 = 0 \text{ si } k \geq n$$

(véase (5.3)) o, lo que es lo mismo,

$$u^k = u \text{ si } k \geq n.$$

Es decir, el método iterativo asociado a B converge, a lo sumo, en n pasos. □

5.6. Sea $A \in \mathcal{M}_n$ una matriz triangular superior por bloques. Estudiar la convergencia de los métodos de Jacobi, Gauss–Seidel y relajación asociados a la descomposición por bloques de A. Ídem si A es triangular inferior.

Solución.

a) Si la matriz A es triangular superior por bloques entonces $E = 0$, por lo que

$$\mathcal{J} = \mathcal{L}_1 = D^{-1}F$$

es la matriz de los métodos de Jacobi y de Gauss–Seidel y

$$\begin{aligned}\mathcal{L}_w &= \left(\frac{D}{w}\right)^{-1}\left(\frac{1-w}{w}D + F\right) = wD^{-1}\left(\frac{1-w}{w}D + F\right)\\ &= (1-w)I + wD^{-1}F\end{aligned}$$

es la matriz del método de relajación de parámetro w. Como $\mathcal{J} = \mathcal{L}_1$ es una matriz triangular superior con elementos nulos en la diagonal, entonces

$$\operatorname{sp}(\mathcal{J}) = \operatorname{sp}(\mathcal{L}_1) = \{0\},$$

por lo que, aplicando el problema 5.5, se tiene que los métodos de Jacobi y Gauss–Seidel convergen, a lo sumo, en n pasos. Por otra parte, como $\mathcal{L}_w$ es una matriz triangular superior que tiene a $1-w$ por elementos en la diagonal, se verifica que

$$\varrho(\mathcal{L}_w) = |1 - w|.$$

Por tanto, el método de relajación de parámetro w es convergente si y sólo si $0 < w < 2$.

b) Si A es una matriz triangular inferior por bloques entonces $F = 0$, por lo que

$$\mathcal{J} = D^{-1}E$$

es la matriz del método de Jacobi,

$$\mathcal{L}_1 = 0$$

es la matriz del método de Gauss–Seidel y

$$\begin{aligned}\mathcal{L}_w &= \left(\frac{D}{w} - E\right)^{-1} \frac{1-w}{w} D = (1-w)\left(wD^{-1}\left(\frac{D}{w} - E\right)\right)^{-1} \\ &= (1-w)\left(I - wD^{-1}E\right)^{-1}\end{aligned}$$

es la matriz del método de relajación de parámetro w. Como $\mathcal{J}$ es una matriz triangular inferior con elementos nulos en la diagonal, entonces

$$\operatorname{sp}(\mathcal{J}) = \{0\},$$

por lo que, aplicando nuevamente el problema 5.5, se tiene que el método de Jacobi converge, a lo sumo, en n pasos. Por otro lado, como $\mathcal{L}_1 = 0$ entonces el método de Gauss–Seidel es directo, en el sentido de que la solución

$$u = (D - E)^{-1}b$$

se obtiene en la primera iteración. Finalmente, como $\mathcal{L}_w$ es nuevamente una matriz triangular inferior con elementos $1 - w$ en la diagonal, se verifica que

$$\varrho(\mathcal{L}_w) = |1 - w|.$$

Por tanto, el método de relajación es convergente si y sólo si $0 < w < 2$. □

5.7. Demostrar que si $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ verifica

$$|a_{jj}| > \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}| \tag{5.23}$$

para $j = 1, \ldots, n$, entonces el método de Jacobi por puntos para A es convergente.

Solución. Como la matriz A verifica (5.23) entonces la matriz A^{T} es de diagonal estrictamente dominante y, por tanto, por el teorema 5.5, el método de Jacobi para A^{T} es convergente, es decir,

$$\varrho\left(\mathcal{J}_{A^{\mathrm{T}}}\right) < 1. \tag{5.24}$$

Por otra parte, si $A = D - E - F$ y denotamos por $A^{\mathrm{T}} = \widetilde{D} - \widetilde{E} - \widetilde{F}$ la descomposición $D - E - F$ de la matriz A^{T}, se tiene que

$$\widetilde{D} = D,\ \widetilde{E} = F^{\mathrm{T}} \ \text{ y } \ \widetilde{F} = E^{\mathrm{T}}.$$

De esta forma, se verifica que

$$\mathcal{J}_{A^{\mathrm{T}}} = \widetilde{D}^{-1}\left(\widetilde{E} + \widetilde{F}\right) = D^{-1}\left(F^{\mathrm{T}} + E^{\mathrm{T}}\right) = D^{-1}\left(E + F\right)^{\mathrm{T}} = \left(\left(E + F\right) D^{-1}\right)^{\mathrm{T}}$$

y, por tanto,

$$\mathcal{J}_{A^{\mathrm{T}}}^{\mathrm{T}} = \left(E + F\right) D^{-1},$$

de donde

$$D^{-1}\mathcal{J}_{A^{\mathrm{T}}}^{\mathrm{T}} D = D^{-1}\left(E + F\right) = \mathcal{J}.$$

Es decir, $\mathcal{J}$ y $\mathcal{J}_{A^{\mathrm{T}}}^{\mathrm{T}}$ (y, consecuentemente, $\mathcal{J}_{A^{\mathrm{T}}}$) son matrices semejantes, por lo que tienen el mismo espectro (véase la observación 2.9). En particular,

$$\varrho\left(\mathcal{J}\right) = \varrho\left(\mathcal{J}_{A^{\mathrm{T}}}\right) < 1$$

(véase (5.24)). Por tanto, el método de Jacobi para A es convergente. □

5.8. Sea $A \in \mathcal{M}_n$.

a) Probar que si $0 < w \leq 1$ y $\lambda \in \mathbb{C}$ con $|\lambda| \geq 1$ entonces

$$\left|\frac{1 - w - \lambda}{\lambda w}\right| \geq 1.$$

b) Demostrar que si A es una matriz de diagonal estrictamente dominante, el método de relajación por puntos para A es convergente si $0 < w \leq 1$.

Solución.

a) Aplicando la desigualdad

$$|x - y| \geq |\,|x| - |y|\,|\,,\ x, y \in \mathbb{C}$$

(véase (2.10)) a $x = 1 - w \geq 0$ e $y = \lambda \in \mathbb{C}$ se obtiene que

$$|(1 - w) - \lambda| \geq |\,|1 - w| - |\lambda|\,| = |\,(1 - w) - |\lambda|\,|\,. \tag{5.25}$$

Como $|\lambda| \geq 1$ entonces

$$0 \leq 1 - w \leq |\lambda|(1 - w) = |\lambda| - w|\lambda|$$

y, por tanto,

$$1 - w - |\lambda| \leq -w|\lambda| \leq 0;$$

consecuentemente,

$$|\,1 - w - |\lambda|\,| \geq w|\lambda|.$$

De esta forma, regresando a la expresión (5.25) se obtiene que

$$|(1-w) - \lambda| \geq w|\lambda|,$$

de donde se deduce la desigualdad buscada.

b) Si el método de relajación con parámetro $w \in (0,1]$ para A no converge existe un autovalor $\lambda \in \operatorname{sp}(\mathcal{L}_w)$ con $|\lambda| \geq 1$. De esta forma,

$$\det\left(\left(\frac{D}{w} - E\right)^{-1}\left(\frac{1-w}{w}D + F\right) - \lambda I\right) = \det\left(\mathcal{L}_w - \lambda I\right) = P_{\mathcal{L}_w}(\lambda) = 0$$

y, por tanto,

$$\begin{aligned} 0 &= \det\left(\frac{D}{w} - E\right)\det\left(\left(\frac{D}{w} - E\right)^{-1}\left(\frac{1-w}{w}D + F\right) - \lambda I\right) \\ &= \det\left(\left(\frac{1-w}{w}D + F\right) - \lambda\left(\frac{D}{w} - E\right)\right) \\ &= \det\left(\frac{1-w-\lambda}{w}D + F + \lambda E\right), \end{aligned}$$

por lo que la matriz

$$\frac{1-w-\lambda}{w}D + F + \lambda E$$

no es inversible y, en consecuencia, no es de diagonal estrictamente dominante. Por tanto, existe un índice $i_0 \in \{1, 2, \ldots, n\}$ tal que

$$\left|\frac{1-w-\lambda}{w}\right| |a_{i_0 i_0}| \leq |\lambda| \sum_{j=1}^{i_0-1} |a_{i_0 j}| + \sum_{j=i_0+1}^{n} |a_{i_0 j}|.$$

Como $\lambda \neq 0$ entonces

$$\left|\frac{1-w-\lambda}{w\lambda}\right| |a_{i_0 i_0}| \leq \sum_{j=1}^{i_0-1} |a_{i_0 j}| + \frac{1}{|\lambda|}\sum_{j=i_0+1}^{n} |a_{i_0 j}| \leq \sum_{\substack{j=1 \\ j\neq i_0}}^{n} |a_{i_0 j}|,$$

ya que $|\lambda| \geq 1$. De esta forma, por el apartado a) se obtiene que

$$|a_{i_0 i_0}| \leq \left| \frac{1-w-\lambda}{w\lambda} \right| |a_{i_0 i_0}| \leq \sum_{\substack{j=1 \\ j \neq i_0}}^{n} |a_{i_0 j}|,$$

lo que contradice el hecho de que la matriz A es de diagonal estrictamente dominante. □

5.9. Sea $A \in \mathcal{M}_n$ una matriz de diagonal estrictamente dominante. Demostrar que el método de relajación por bloques para A es convergente si $0 < w \leq 1$.

SOLUCIÓN. Consideramos la descomposición $D - E - F$ por bloques de $A \in \mathcal{M}_n$ donde los bloques diagonales $A_{ii} \in \mathcal{M}_{n_i}$, $i = 1, 2, \ldots, p$ con $n_1 + n_2 + \cdots + n_p = n$. Argumentamos por reducción al absurdo: si el método de relajación por bloques con parámetro $w \in (0, 1]$ para A no converge existe un autovalor $\lambda \in \operatorname{sp}(\mathcal{L}_w)$ con $|\lambda| \geq 1$. De esta forma,

$$\det\left(\left(\frac{D}{w} - E \right)^{-1} \left(\frac{1-w}{w} D + F \right) - \lambda I \right) = \det(\mathcal{L}_w - \lambda I) = P_{\mathcal{L}_w}(\lambda) = 0$$

y

$$\begin{aligned} 0 &= \det\left(\frac{D}{w} - E \right) \det\left(\left(\frac{D}{w} - E \right)^{-1} \left(\frac{1-w}{w} D + F \right) - \lambda I \right) \\ &= \det\left(\left(\frac{1-w}{w} D + F \right) - \lambda \left(\frac{D}{w} - E \right) \right) \\ &= \det\left(\frac{1-w-\lambda}{w} D + F + \lambda E \right), \end{aligned}$$

por lo que la matriz

$$\frac{1-w-\lambda}{w} D + F + \lambda E$$

no es inversible y, por tanto, no es de diagonal estrictamente dominante. Entonces, existe un índice $i_0 \in \{1, 2, \ldots, n\}$ con

$$n_1 + n_2 + \cdots + n_k < i_0 \leq n_1 + n_2 + \cdots + n_{k+1}$$

para algún $k \in \{0, 1, \ldots, p-1\}$, para el que se verifica

$$\left| \frac{1-w-\lambda}{w} \right| |a_{i_0 i_0}| \leq |\lambda| \sum_{j=1}^{n_1+n_2+\cdots+n_k} |a_{i_0 j}| + \left| \frac{1-w-\lambda}{w} \right| \sum_{\substack{j=n_1+n_2+\cdots+n_k+1 \\ j \neq i_0}}^{n_1+n_2+\cdots+n_{k+1}} |a_{i_0 j}|$$

$$+ \sum_{j=n_1+n_2+\cdots+n_{k+1}+1}^{n} |a_{i_0 j}|.$$

Dividiendo la expresión anterior por $\left|\dfrac{1-w-\lambda}{w}\right|$ se obtiene

$$\begin{aligned}|a_{i_0i_0}| &\leq \left|\frac{w\lambda}{1-w-\lambda}\right| \sum_{j=1}^{n_1+n_2+\cdots+n_k} |a_{i_0j}| + \sum_{\substack{j=n_1+n_2+\cdots+n_k+1 \\ j\neq i_0}}^{n_1+n_2+\cdots+n_{k+1}} |a_{i_0j}| \\ &\quad + \left|\frac{w}{1-w-\lambda}\right| \sum_{j=n_1+n_2+\cdots+n_{k+1}+1}^{n} |a_{i_0j}| \\ &\leq \sum_{j=1}^{n_1+n_2+\cdots+n_k} |a_{i_0j}| + \sum_{\substack{j=n_1+n_2+\cdots+n_k+1 \\ j\neq i_0}}^{n_1+n_2+\cdots+n_{k+1}} |a_{i_0j}| \\ &\quad + \sum_{j=n_1+n_2+\cdots+n_{k+1}+1}^{n} |a_{i_0j}| = \sum_{\substack{j=1 \\ j\neq i_0}}^{n} |a_{i_0j}|\end{aligned}$$

ya que, al ser $|\lambda| \geq 1$, se verifica

$$\left|\frac{w}{1-w-\lambda}\right| \leq \left|\frac{w\lambda}{1-w-\lambda}\right| \leq 1$$

(véase el apartado *a*) del problema 5.8). Por tanto, hemos llegado así a una contradicción con el hecho de que la matriz A sea de diagonal estrictamente dominante. □

5.10. Sea $A \in \mathcal{M}_n$ una matriz hermítica e inversible descompuesta en la forma $A = M - N$ con M inversible.

a) Se considera la sucesión

$$v^{n+1} = M^{-1}Nv^n$$

con $v^0 \in \mathbf{V}\backslash\{0\}$ arbitrario. Probar que si la matriz $M^* + N$ es definida positiva entonces la sucesión $\left\{(v^n)^* Av^n\right\}_{n=0}^{\infty}$ es monótona no creciente.

b) Demostrar que si $M^* + N$ es definida positiva y $\varrho(M^{-1}N) < 1$ entonces A es definida positiva.

Solución.

a) Debe probarse que

$$\left(v^{n+1}\right)^* Av^{n+1} \leq (v^n)^* Av^n$$

para $n \in \mathbb{N} \cup \{0\}$. Basta aplicar la igualdad (5.13) a los vectores $v = v^n$ y $w = v^{n+1}$, pues entonces

$$(v^n)^* Av^n - \left(v^{n+1}\right)^* Av^{n+1} = (v^n - v^{n+1})^*(M^* + N)(v^n - v^{n+1}) \geq 0$$

por ser $M^* + N$ definida positiva.

b) Argumentamos por contradicción. Si A no es definida positiva existe un vector $v^0 \in \mathbf{V}\backslash\{0\}$ tal que

$$\left(v^0\right)^* Av^0 \leq 0.$$

A partir del vector $v^0 \neq 0$ se considera la sucesión $\{v^n\}_{n=0}^{\infty}$ como en el apartado *a*). Nótese que $v^1 \neq v^0$: en efecto, en caso contrario se obtendría la contradicción

$$v^0 = M^{-1}Nv^0 \;\Rightarrow\; Mv^0 = Nv^0 \;\Rightarrow\; Av^0 = (M - N)v^0 = 0 \;\Rightarrow\; v^0 = 0$$

por ser A inversible. Razonando como en el apartado anterior se llega, en este caso, a que

$$\left(v^1\right)^* Av^1 < \left(v^0\right)^* Av^0.$$

Así, utilizando nuevamente *a*), se verifica

$$(v^n)^* Av^n \leq \left(v^{n-1}\right)^* Av^{n-1} \leq \cdots \leq \left(v^1\right)^* Av^1 < \left(v^0\right)^* Av^0 \leq 0$$

para todo $n \in \mathbb{N}$, por lo que

$$\lim_{n\to+\infty} (v^n)^* Av^n \neq 0. \tag{5.26}$$

Ahora bien, como para cada $n \in \mathbb{N}$

$$v^n = M^{-1}Nv^{n-1} = \left(M^{-1}N\right)^2 v^{n-2} = \cdots = \left(M^{-1}N\right)^n v^0$$

y $\varrho(M^{-1}N) < 1$, entonces

$$\lim_{n\to+\infty} v^n = \lim_{n\to+\infty} \left(M^{-1}N\right)^n v^0 = 0$$

(véase el teorema 2.6). Así pues, la continuidad de la función $v \mapsto v^*Av$ hace que se tenga

$$\lim_{n\to+\infty} (v^n)^* Av^n = 0^* A0 = 0,$$

hecho que contradice (5.26). □

5.11. Se considera el sistema lineal $Ax = d$ donde A es una matriz tridiagonal de la forma (4.20).

a) Demostrar que si

$$|b_i| > |a_i| \tag{5.27}$$

para $i = 1, 2, \ldots, n$ (con $a_1 = 0$) entonces

$$\varrho(\mathcal{L}_1) \leq \max_{1 \leq i \leq n} \left\{ \frac{|c_i|}{|b_i| - |a_i|} \right\}.$$

b) Deducir la convergencia del método de Gauss–Seidel por puntos cuando A es una matriz de diagonal estrictamente dominante.

Solución.

a) Sea $\lambda \in \mathrm{sp}(\mathcal{L}_1)$ un autovalor arbitrario de $\mathcal{L}_1$. Por definición se verifica que

$$\det((D-E)^{-1}F - \lambda I) = \det(\mathcal{L}_1 - \lambda I) = 0$$

y, por tanto,

$$\det(F - \lambda(D-E)) = 0.$$

Esto hace que la matriz $F - \lambda(D-E)$ no sea inversible ni, por tanto, de diagonal estrictamente dominante. Así pues, existe un índice $i_0 \in \{1, 2, \ldots, n\}$ tal que

$$|\lambda||b_{i_0}| \leq |c_{i_0}| + |\lambda||a_{i_0}|,$$

con lo que

$$(|b_{i_0}| - |a_{i_0}|)|\lambda| \leq |c_{i_0}|.$$

Gracias a la hipótesis (5.27) podemos dividir la expresión anterior sin invertir la desigualdad, obteniendo

$$|\lambda| \leq \frac{|c_{i_0}|}{|b_{i_0}| - |a_{i_0}|} \leq \max_{1 \leq i \leq n} \left\{ \frac{|c_i|}{|b_i| - |a_i|} \right\}.$$

b) Si la matriz A es de diagonal estrictamente dominante se verifica que

$$|b_i| > |a_i| + |c_i| \geq |a_i|$$

para $i = 1, 2, \ldots, n$, por lo que se cumple la condición (5.27) y, además,

$$\frac{|c_i|}{|b_i| - |a_i|} < 1$$

para $i = 1, 2, \ldots, n$. De esta forma, aplicando el apartado *a*), se obtiene que

$$\varrho(\mathcal{L}_1) \leq \max_{1 \leq i \leq n} \left\{ \frac{|c_i|}{|b_i| - |a_i|} \right\} < 1,$$

por lo que el método de Gauss–Seidel por puntos asociado a la matriz A es convergente. □

5.12. Teorema de los círculos de Gershgorin. Sea $A \in \mathcal{M}_n$. Demostrar que

$$\mathrm{sp}(A) \subset \bigcup_{i=1}^{n} \overline{\mathbf{B}}_{r_i}(a_{ii}) \ \text{ siendo } \ r_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

donde $\overline{\mathbf{B}}_{r_i}(a_{ii})$ denota la bola cerrada de centro a_{ii} y radio r_i, esto es,

$$\overline{\mathbf{B}}_{r_i}(a_{ii}) = \{z \in \mathbb{C} : \ |z - a_{ii}| \leq r_i\}.$$

Es decir, los autovalores de la matriz A se encuentran en la unión de las bolas cerradas de centro los elementos diagonales a_{ii} de A y radios respectivos la suma de los módulos de los elementos de la fila i–ésima excluyendo el elemento diagonal.

Aplicar el teorema anterior para obtener una acotación de los autovalores de la matriz

$$A = \begin{pmatrix} 1+i & 0 & 2 \\ 1 & 5-i & 3 \\ 0 & 1 & -4 \end{pmatrix}. \tag{5.28}$$

Solución. Si $\lambda \in \mathrm{sp}(A)$ la matriz $A - \lambda I$ no es inversible ni, por tanto, de diagonal estrictamente dominante, luego existe un índice $i_0 \in \{1, 2, \ldots, n\}$ tal que

$$|\lambda - a_{i_0 i_0}| \leq \sum_{\substack{j=1 \\ j\neq i_0}}^{n} |a_{i_0 j}| = r_{i_0}.$$

De esta forma,

$$\lambda \in \overline{\mathbf{B}}_{r_{i_0}}(a_{i_0 i_0}) \subset \bigcup_{i=1}^{n} \overline{\mathbf{B}}_{r_i}(a_{ii}).$$

Aplicando el teorema de Gershgorin a la matriz A dada en (5.28) obtenemos la siguiente acotación:

$$\mathrm{sp}(A) \subset \overline{\mathbf{B}}_2(1+i) \cup \overline{\mathbf{B}}_4(5-i) \cup \overline{\mathbf{B}}_1(-4)$$

(véase la figura 5.1). De hecho, los autovalores de la matriz A son

$$\lambda_1 \simeq 5.3561 - 0.9457i, \ \lambda_2 \simeq 0.9241 + 0.9825i \ \text{ y } \ \lambda_3 \simeq -4.2802 - 0.0368i. \quad \square$$

5.13. Sea $A = (a_{ij})_{i,j=1}^{n} \in \mathcal{M}_n$ una matriz de diagonal estrictamente dominante con elementos diagonales

$$a_{ii} > 0$$

para $i = 1, \ldots, n$.

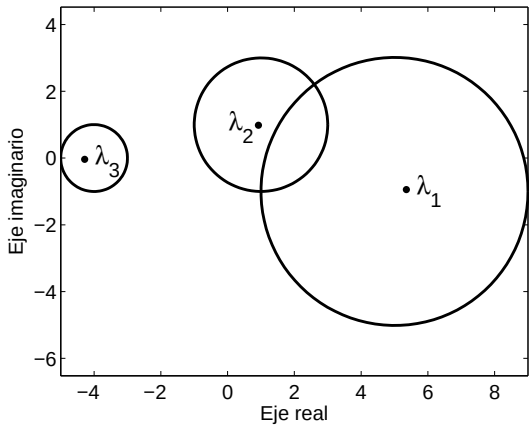


Figura 5.1: Acotación de Gershgorin para los autovalores de A.

a) Probar que si

$$\mathrm{sp}(A) \subset \mathbb{R}$$

entonces los autovalores de A son estrictamente positivos.

b) Demostrar que si, además, A es hermítica, el método de relajación por bloques para A es convergente si y sólo si $0 < w < 2$.

Solución.

a) Sea $\lambda \in \mathrm{sp}(A) \subset \mathbb{R}$. Supongamos que $\lambda \leq 0$ y derivemos una contradicción. Por el problema 5.12, existe $i_0 \in \{1, 2, \ldots, n\}$ tal que

$$|\lambda - a_{i_0 i_0}| \leq \sum_{\substack{j=1 \\ j \neq i_0}}^{n} |a_{i_0 j}| \,. \tag{5.29}$$

Ahora bien, como $\lambda \leq 0$ y $a_{i_0 i_0} > 0$ entonces

$$|\lambda - a_{i_0 i_0}| = a_{i_0 i_0} - \lambda \geq a_{i_0 i_0} = |a_{i_0 i_0}| \,,$$

por lo que reemplazando esta desigualdad en (5.29) se obtiene que

$$|a_{i_0 i_0}| \leq \sum_{\substack{j=1 \\ j \neq i_0}}^{n} |a_{i_0 j}| \,,$$

lo que contradice el hecho de que la matriz A sea de diagonal estrictamente dominante.

b) Como A es una matriz hermítica entonces

$$\mathrm{sp}(A) \subset \mathbb{R}$$

(véase la proposición 2.5) y, por el apartado *a*), de hecho

$$\mathrm{sp}(A) \subset \mathbb{R}_+.$$

Por tanto, la matriz hermítica A es definida positiva (véase la proposición 2.7). De esta forma, los teoremas 5.3 y 5.4 concluyen el resultado. □

5.14. Método iterativo para el cálculo de la inversa de una matriz. Se consideran las sucesiones de matrices

$$A_n = A_{n-1}\left(I + E_n + E_n^2\right) \quad \text{y} \quad E_n = I - AA_{n-1}$$

siendo $A \in \mathcal{M}_n$ inversible y $A_0 \in \mathcal{M}_n$ una matriz arbitraria.

a) Demostrar que $E_n = (E_1)^{3^{n-1}}$.

b) Probar que si $\varrho(E_1) < 1$ entonces $\lim_{n\to+\infty} A_n = A^{-1}$.

c) Mostrar que si se toma $A_0 = \dfrac{A^*}{\mathrm{tr}(AA^*)}$ entonces $\lim_{n\to+\infty} A_n = A^{-1}$.

Solución.

a) Lo probamos por inducción en n:

i) Si $n = 1$ el resultado es obvio, pues $E_1 = (E_1)^{3^0}$.

ii) Supuesto cierto el resultado para $n-1$ lo probamos para n. Por definición,

$$E_n = I - AA_{n-1} = I - AA_{n-2}\left(I + E_{n-1} + E_{n-1}^2\right).$$

Puesto que, por definición, $E_{n-1} = I - AA_{n-2}$, reemplazando el valor de $AA_{n-2} = I - E_{n-1}$ en la expresión anterior se obtiene que

$$\begin{aligned} E_n &= I + (E_{n-1} - I)\left(I + E_{n-1} + E_{n-1}^2\right) \\ &= I + E_{n-1} + E_{n-1}^2 + E_{n-1}^3 - I - E_{n-1} - E_{n-1}^2 = E_{n-1}^3. \end{aligned}$$

Aplicando la hipótesis de inducción se concluye que

$$E_n = E_{n-1}^3 = \left((E_1)^{3^{n-2}}\right)^3 = (E_1)^{3^{n-1}}$$

como queríamos demostrar.

b) Como $E_{n+1} = I - AA_n$ y la matriz A es inversible, entonces

$$A^{-1}E_{n+1} = A^{-1} - A_n,\ n \in \mathbb{N}$$

por lo que, por el apartado *a*), para cualquier norma matricial $|||\cdot|||$ se verifica

$$|||A^{-1} - A_n||| = |||A^{-1}E_{n+1}||| \leq |||A^{-1}|||\ |||E_{n+1}||| = |||A^{-1}|||\ |||(E_1)^{3^n}|||$$

para $n \in \mathbb{N}$. El resultado se concluye haciendo tender $n \to \infty$ en la desigualdad anterior, teniendo en cuenta que

$$\lim_{n\to+\infty} |||(E_1)^{3^n}||| = 0$$

puesto que $\varrho(E_1) < 1$ (véase el teorema 2.6).

c) A la vista del apartado *b*) basta demostrar que para la elección

$$A_0 = \frac{A^*}{\mathrm{tr}(AA^*)}$$

se verifica que $\varrho(E_1) < 1$. Podemos escribir E_1 en la forma

$$E_1 = I - AA_0 = I - B$$

siendo

$$B = \frac{AA^*}{\mathrm{tr}(AA^*)}.$$

Puesto que la matriz A es inversible se verifica que AA^* es una matriz hermítica definida positiva (véase la proposición 2.8). Por tanto, por la proposición 2.7,

$$\mathrm{sp}(AA^*) \subset \mathbb{R}_+,$$

es decir,

$$\lambda_i(AA^*) > 0$$

para $i = 1, 2, \ldots, n$. En consecuencia,

$$0 < \lambda_i(B) = \frac{\lambda_i(AA^*)}{\mathrm{tr}(AA^*)} = \frac{\lambda_i(AA^*)}{\displaystyle\sum_{i=1}^{n}\lambda_i(AA^*)} < 1$$

para $i = 1, 2, \ldots, n$. De esta forma, la igualdad

$$\det(E_1 - \lambda I) = \det\left((1-\lambda)I - B\right)$$

determina

$$\lambda \in \mathrm{sp}(E_1) \ \Rightarrow\ (1-\lambda) \in \mathrm{sp}(B) \ \Rightarrow\ 0 < 1 - \lambda < 1 \ \Rightarrow\ 0 < \lambda < 1,$$

de donde se deduce que $\varrho(E_1) < 1$. □

5.15. Convergencia de los métodos asociados a matrices no negativas.

a) Sea $B \in \mathcal{M}_n$ tal que $\varrho(B) < 1$. Demostrar que $I-B$ es inversible y comprobar que se verifica

$$(I-B)^{-1} = \sum_{k=0}^{\infty} B^k$$

(nótese la analogía existente con la suma de los términos de la serie geométrica

$$\sum_{n=0}^{\infty} z^n = \frac{1}{1-z} = (1-z)^{-1}$$

cuando $|z| < 1$).

b) Una matriz $B = (b_{ij})_{i,j=1}^n$ es *no negativa* (y se representa $B \geq 0$) si

$$b_{ij} \geq 0,\ i,j = 1,2,\ldots,n.$$

Si B es una matriz no negativa, demostrar la equivalencia:

$$I - B \text{ es inversible e } (I-B)^{-1} \geq 0 \Leftrightarrow \varrho(B) < 1.$$

c) Se considera la descomposición $D - E - F$ por puntos de una matriz real A inversible que verifica

$$a_{ij} \leq 0 \text{ si } i \neq j \text{ y } A^{-1} \geq 0.$$

i) Probar que $a_{ii} > 0$. Deducir que las matrices de los métodos de Jacobi y relajación por puntos asociados a A están bien definidas.

ii) Demostrar que el método de Jacobi por puntos para A es convergente.

iii) Utilizar el apartado *b*) para demostrar que:

α) $\left(\dfrac{D}{w} - E\right)^{-1} \geq 0$ si $w > 0$.

β) El método de relajación por puntos para A es convergente para valores del parámetro $0 < w \leq 1$.

Solución.

a) Como $\varrho(B) < 1$ entonces $\lambda = 1 \notin \operatorname{sp}(B)$. Por tanto, $\det(B-I) \neq 0$ y la matriz $I - B$ es inversible. Por otra parte, para cada $m \in \mathbb{N}$

$$\begin{aligned}(I-B)\sum_{k=0}^{m} B^k &= \sum_{k=0}^{m} B^k - \sum_{k=0}^{m} B^{k+1} \\ &= \left(I + B + \cdots + B^{m-1} + B^m\right) \\ &\quad - \left(B + B^2 + \cdots + B^m + B^{m+1}\right) \\ &= I - B^{m+1}.\end{aligned}$$

Así pues, haciendo tender $m \to \infty$ se tiene que

$$(I-B)\sum_{k=0}^{\infty} B^k = \lim_{m\to+\infty}\left((I-B)\sum_{k=0}^{m} B^k\right) = \lim_{m\to+\infty}\left(I-B^{m+1}\right) = I$$

(véase el teorema 2.6). De la igualdad anterior se deduce el resultado buscado.

b) Mostremos la equivalencia:

$\boxed{\Leftarrow}$ Por el apartado *a*) sabemos que la matriz $I-B$ es inversible, y el hecho de que $B \geq 0$ hace que se tenga

$$B^k \geq 0,\ k \in \mathbb{N},$$

por lo que, utilizando nuevamente el apartado *a*), se obtiene que

$$(I-B)^{-1} = \sum_{k=0}^{\infty} B^k \geq 0.$$

$\boxed{\Rightarrow}$ Sea $\lambda \in \mathrm{sp}(B)$ y $v \in \mathbf{V}\backslash\{0\}$ tal que $Bv = \lambda v$. De esta forma

$$|Bv| = |\lambda||v|$$

donde

$$|w| = \begin{pmatrix} |w_1| \\ |w_2| \\ \cdots \\ |w_n| \end{pmatrix},\ w \in \mathbf{V}.$$

Como $B \geq 0$, entonces

$$B|v| \geq |Bv| = |\lambda||v|$$

y, por tanto,

$$(I-B)|v| = |v| - B|v| \leq |v| - |\lambda||v| = (1-|\lambda|)\,|v|.$$

Al ser $(I-B)^{-1} \geq 0$, al multiplicar la desigualdad anterior por $(I-B)^{-1}$ no se invierte su sentido, obteniéndose

$$|v| \leq (1-|\lambda|)\,(I-B)^{-1}|v|. \tag{5.30}$$

Como v es un vector no nulo, existe $i \in \{1,2,\ldots,n\}$ tal que $v_i \neq 0$. De esta forma, a partir de (5.30), se tiene que

$$0 < |v_i| \leq (1-|\lambda|)\left((I-B)^{-1}|v|\right)_i.$$

Como, por hipótesis,

$$\left((I-B)^{-1}|v|\right)_i \geq 0,$$

de la desigualdad anterior se deduce que $1-|\lambda|>0$ o, lo que es lo mismo, $|\lambda|<1$, como queríamos demostrar.

c) Consideremos cada uno de los apartados:

i) Supongamos que existe $i \in \{1,2,\ldots,n\}$ tal que $a_{ii} \leq 0$ y derivemos una contradicción. En tal caso se tendría, por hipótesis, que

$$A\mathbf{e}_i = \begin{pmatrix} a_{1i} \\ a_{2i} \\ \cdots \\ a_{ni} \end{pmatrix} \leq 0$$

donde $\mathbf{e}_i$ es el i–ésimo vector de la base canónica. Puesto que $A^{-1} \geq 0$, se llegaría así a la contradicción

$$\mathbf{e}_i = A^{-1}A\mathbf{e}_i \leq 0.$$

Por tanto,

$$a_{ii} > 0$$

para $i = 1,2,\ldots,n$, lo que hace que tengan sentido las matrices de los métodos de Jacobi y de relajación, al ser el único requisito necesario que

$$a_{ii} \neq 0$$

para $i = 1,2,\ldots,n$.

ii) Por hipótesis

$$\begin{cases} a_{ii} > 0,\ i = 1,2,\ldots,n & \Rightarrow \quad D \geq 0 \ \Rightarrow \ D^{-1} \geq 0 \\ a_{ij} \leq 0,\ i \neq j & \Rightarrow \quad E \geq 0 \ \text{ y } \ F \geq 0. \end{cases} \tag{5.31}$$

De esta forma $\mathcal{J} = D^{-1}(E+F) \geq 0$. Por otra parte la matriz

$$I - \mathcal{J} = I - D^{-1}(E+F) = D^{-1}(D-E-F) = D^{-1}A$$

es inversible, por serlo A. Finalmente,

$$(I-\mathcal{J})^{-1} = \left(D^{-1}A\right)^{-1} = A^{-1}D \geq 0.$$

Así pues, estamos en condiciones de aplicar el apartado *b*) a la matriz $\mathcal{J}$ obteniendo que $\varrho(\mathcal{J}) < 1$ o, equivalentemente, que el método de Jacobi para A es convergente.

iii) α) Claramente,

$$\left(\frac{D}{w}-E\right)^{-1}=w(D-wE)^{-1}=w\left(I-wD^{-1}E\right)^{-1}D^{-1} \quad (5.32)$$

ya que

$$D-wE=D(I-wD^{-1}E).$$

Por hipótesis $D\geq 0$ y, por tanto, $D^{-1}\geq 0$. Por otra parte, la matriz no negativa

$$B=wD^{-1}E$$

es triangular inferior con elementos nulos en la diagonal; consecuentemente, $\varrho(B)=0<1$. Aplicando el apartado b) a la matriz B se obtiene que $I-B$ es inversible y

$$\left(I-wD^{-1}E\right)^{-1}=(I-B)^{-1}\geq 0.$$

De esta forma, a partir de (5.32), se deduce que

$$\left(\frac{D}{w}-E\right)^{-1}=w\left(I-wD^{-1}E\right)^{-1}D^{-1}\geq 0$$

siempre que $w>0$.

β) Por (5.31) se verifica que $D\geq 0$, $E\geq 0$ y $F\geq 0$. Luego por el apartado α) se tiene que

$$\mathcal{L}_w=\left(\frac{D}{w}-E\right)^{-1}\left(\frac{1-w}{w}D+F\right)\geq 0$$

si $0<w\leq 1$. Por otra parte, la matriz

$$\begin{aligned}I-\mathcal{L}_w&=I-\left(\frac{D}{w}-E\right)^{-1}\left(\frac{1-w}{w}D+F\right)\\&=\left(\frac{D}{w}-E\right)^{-1}\left(\frac{D}{w}-E-\frac{1-w}{w}D-F\right)\\&=\left(\frac{D}{w}-E\right)^{-1}(D-E-F)=\left(\frac{D}{w}-E\right)^{-1}A\end{aligned}$$

es inversible por serlo A e

$$\begin{aligned}(I-\mathcal{L}_w)^{-1} &= \left(\left(\frac{D}{w}-E\right)^{-1}A\right)^{-1} = A^{-1}\left(\frac{D}{w}-E\right)\\ &= A^{-1}\left(\frac{D}{w}+A-D+F\right) = A^{-1}\left(A+\frac{1-w}{w}D+F\right)\\ &= I+A^{-1}\left(\frac{1-w}{w}D+F\right) \geq 0\end{aligned}$$

donde hemos utilizado que $-E = A - D + F$. De esta forma, aplicando el apartado b) a la matriz $\mathcal{L}_w$ se obtiene que $\varrho(\mathcal{L}_w) < 1$ o, lo que es lo mismo, que el método de relajación de parámetro $w \in (0,1]$ para A es convergente. □

5.6.2. Problemas propuestos

5.16. Construir matrices para las cuales el método de Jacobi asociado sea convergente y el método de Gauss–Seidel diverja y recíprocamente.

5.17. Si $\{u^k\}_{k=0}^{\infty}$ es la sucesión de vectores de un método iterativo convergente definido por una matriz B y $u = \lim_{k\to+\infty} u^k$, demostrar la acotación

$$\left|\left|u^k - u\right|\right| \leq \frac{|||B|||^k}{1-|||B|||}\left|\left|u^1-u^0\right|\right|,\ k \in \mathbb{N}.$$

5.18. Sea $A \in \mathcal{M}_n$ una matriz de diagonal estrictamente dominante.

a) Demostrar que si A se descompone en la forma $A = M - N$, siendo

$$m_{ii} = a_{ii} \ \text{ y } \ m_{ij}n_{ij} = 0$$

para $i, j = 1, \ldots, n$, entonces el método iterativo asociado a tal descomposición de A está bien definido y es convergente.

b) Deducir, a partir de *a*), resultados de convergencia para los métodos de Jacobi y Gauss–Seidel.

5.19. Sea A una matriz de diagonal estrictamente dominante y $A = D-E-F$ su descomposición $D-E-F$ por puntos. Se consideran $\alpha, \beta \in [0,1]$, $M = D-\alpha E-\beta F$ y $N = M - A$.

a) Demostrar que el método asociado a la descomposición $Mu = Nu + b$ está bien definido.

b) Probar que dicho método es convergente y deducir la convergencia de los métodos de Gauss–Seidel y Jacobi por puntos para matrices de diagonal estrictamente dominante.

c) Demostrar que el método también es convergente si se considera la descomposición $D - E - F$ por bloques de A.

5.20. Demostrar que si $A \in \mathcal{M}_n$ es una matriz de diagonal estrictamente dominante y $0 < w \leq 1$, el método de relajación–Jacobi es convergente.

5.21. Sea $A \in \mathcal{M}_n$ una matriz de diagonal estrictamente dominante. Demostrar que el método de Jacobi por bloques para A es convergente.

5.22. Se considera la matriz tridiagonal

$$A = \begin{pmatrix} 2+\lambda_1 & 1 & & & \\ 1 & 2+\lambda_2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 2+\lambda_{n-1} & 1 \\ & & & 1 & 2+\lambda_n \end{pmatrix}$$

donde $\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \subset \mathbb{R}$.

a) Probar que si

$$\lambda_i > 0$$

para $i = 1, 2, \ldots, n$, el método de Jacobi por puntos para A es convergente.

b) Bajo el supuesto de que

$$\lambda_i \geq 0$$

para $i = 1, 2, \ldots, n$:

i) mostrar que

$$\delta_k > \delta_{k-1} > \cdots > \delta_0$$

para $k = 1, 2, \ldots, n$, donde δ_k denota el menor principal de orden k de la matriz A.

ii) utilizar el apartado *i*) para demostrar que el método de relajación por puntos para A es convergente.

5.23. Sea $\|\cdot\|$ una norma matricial subordinada y $A \in \mathcal{M}_n$ una matriz inversible escrita en la forma $A = M - N$ con M inversible. Demostrar que si

$$\|N\| \, \left\|A^{-1}\right\| < \frac{1}{2}$$

el método asociado a esta descomposición de A es convergente. Deducir que si A verifica las condiciones

$$\mathrm{cond}_\infty(A) = 1$$

y

$$\sum_{i \geq j} |a_{ij}| > \sum_{i<j} |a_{ij}|$$

para $i = 1, \ldots, n$, entonces el método de Gauss–Seidel por puntos para A es convergente.

5.24. Convergencia de los métodos de Jacobi y relajación para matrices irreducibles de diagonal fuertemente dominante.

a) Se considera una matriz $A = (a_{ij})_{i,j=1}^n \in \mathcal{M}_n$ irreducible (véase el problema 4.18) y n^2 números no nulos α_{ij}, $i, j = 1, \ldots, n$. Demostrar que la matriz

$$(\alpha_{ij} a_{ij})_{i,j=1}^n$$

es también irreducible.

b) Sea A una *matriz de diagonal fuertemente dominante*, es decir,

$$\begin{cases} |a_{ii}| \geq \displaystyle\sum_{j \neq i} |a_{ij}|, & i = 1, 2, \ldots, n \\ |a_{i_0 i_0}| > \displaystyle\sum_{j \neq i_0} |a_{i_0 j}| & \text{para algún } i_0. \end{cases}$$

Demostrar que si A es irreducible entonces A es inversible. Para ello, suponer que $x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \in \mathbf{V} \backslash \{0\}$ es tal que $Ax = 0$; entonces A es reducible con $S = \{k : |x_k| = ||x||_\infty\}$ y $T = \{1, 2, \ldots, n\} \backslash S$.

c) Demostrar que si una matriz A es de diagonal fuertemente dominante e irreducible entonces el método de Jacobi por puntos para A es convergente.

d) Ídem para el método de relajación por puntos con $0 < w \leq 1$.

e) Demostrar *c*) y *d*) para los métodos por bloques.

5.25. Sea $A \in \mathcal{M}_n$ escrita en la forma $A = M - N$ siendo $M \in \mathcal{M}_n$ una matriz inversible y sea $B = M^{-1}N$. Dado $\alpha \in \mathbb{R} \backslash \{-1\}$ se definen las matrices

$$M_\alpha = (1 + \alpha)M, \; N_\alpha = M_\alpha - A \;\text{ y }\; B_\alpha = M_\alpha^{-1} N_\alpha.$$

a) Demostrar que

$$B_\alpha = \frac{1}{1+\alpha}(B + \alpha I).$$

b) Probar la equivalencia

$$\lambda \in \mathrm{sp}(B) \;\Leftrightarrow\; \frac{\lambda+\alpha}{1+\alpha} \in \mathrm{sp}(B_\alpha).$$

c) Suponiendo que los autovalores de B verifican la relación

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n < 1,$$

demostrar que el método asociado a B_α converge para $\alpha > -\dfrac{1+\lambda_1}{2}$.

d) ¿Qué ocurre si $\lambda_n \geq \cdots \geq \lambda_1 > 1$?

e) Comprobar que el método asociado a B_α es, de hecho, un método de relajación de parámetro $w = \dfrac{1}{1+\alpha}$ aplicado al método asociado a B, en el sentido introducido en la subsección 5.3.3.

5.26. Sea A una matriz hermítica definida positiva con autovalores

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

y u la solución del sistema $Au = b$. Se consideran la sucesiones

$$u^{k+1} = u^k + \alpha_k r^k,$$

donde

$$r^k = b - Au^k \;\text{ y }\; \alpha_k = \frac{(r^k)^* r^k}{(r^k)^* A r^k},$$

y

$$E_k = (u^k - u)^* A (u^k - u)$$

para $k \in \mathbb{N}$.

a) Demostrar que

$$E_{k+1} = E_k - \alpha_k \, ||r_k||_2^2 \, .$$

b) Probar que $E_k = (r^k)^* A^{-1} r^k$ y deducir que

$$E_{k+1} = E_k(1 - \alpha_k \beta_k)$$

siendo

$$\beta_k = \frac{(r^k)^* r^k}{(r^k)^* A^{-1} r^k}.$$

c) Probar que $E_{k+1} \leq E_k \left(1 - \dfrac{\lambda_1}{\lambda_n}\right)$. Deducir que

$$\lim_{k \to +\infty} E_k = 0$$

y, por tanto, la sucesión $\{u^k\}_{k=1}^{\infty}$ converge a la solución del sistema $Au = b$.

d) ¿Cómo se utilizaría este método para la resolución de un sistema lineal con matriz hermítica definida positiva?

5.27. Se considera el sistema lineal $Au = b$ donde $A \in \mathcal{M}_n$ es una matriz hermítica y definida positiva. Dado el método iterativo

$$u^{k+1} = (I - \theta A)u^k + \theta b \ \text{ con } \ \theta > 0$$

determinar para qué valores de θ el método anterior es convergente a la solución del sistema $Au = b$.

5.28. Se considera el sistema de $2n$ ecuaciones con $2n$ incógnitas

$$\begin{cases} x + Sy = b_1 \\ S^{\mathrm{T}}x + y = b_2 \end{cases} \qquad (b_1, b_2 \in \mathbb{R}^n \text{ dados})$$

con $x, y \in \mathbb{R}^n$ y $S \in \mathcal{M}_n$.

a) Escribir el anterior sistema en la forma

$$A \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

para una matriz $A \in \mathcal{M}_{2n}$.

b) Dar condiciones sobre la matriz S para que el método de Jacobi por bloques asociado a la descomposición en bloques de A correspondiente a *a*), (es decir, $2n = n + n$), sea convergente. Para ello, denotando para cada $k \in \mathbb{N}$

$$\begin{pmatrix} e_k \\ g_k \end{pmatrix} = \begin{pmatrix} x_k - x \\ y_k - y \end{pmatrix},$$

encontrar una fórmula recursiva para e_k que no involucre a g_k, y viceversa; deducir, a partir de estas dos fórmulas, bajo qué hipótesis sobre S las sucesiones $\{e_k\}_{k=1}^{\infty}$ y $\{g_k\}_{k=1}^{\infty}$ tienden a cero.

c) ¿Qué se puede decir para el método de Gauss–Seidel asociado a la misma descomposición por bloques?

5.29. Método de Gauss–Seidel simétrico. Sea $A \in \mathcal{M}_n$ una matriz simétrica definida positiva descompuesta por puntos $A = D - E - F$. Dado $u^0 \in \mathbf{V}$ arbitrario se define la sucesión $\{u^k\}_{k=0}^{\infty}$ mediante

$$\begin{cases} (D-E)u^{k+\frac{1}{2}} = Fu^k + b \\ (D-F)u^{k+1} = Eu^{k+\frac{1}{2}} + b \end{cases}$$

siendo $b \in \mathbf{V}$ un vector fijo.

a) Escribir $u^{k+1} = Bu^k + c$ determinando explícitamente B y c.

b) Demostrar la equivalencia

$$Bv = \lambda v \;\Leftrightarrow\; \lambda Av + (\lambda - 1)ED^{-1}Fv = 0.$$

c) Probar que la sucesión $\{u^k\}_{k=0}^{\infty}$ converge a la solución de $Au = b$.

5.30. Método de direcciones alternadas. Sea H una matriz hermítica definida positiva (respectivamente, semidefinida positiva).

a) Probar que para cada $r > 0$ la matriz $rI+H$ es inversible, $(rI-H)(rI+H)^{-1}$ es hermítica y

$$\left\|\!\left|(rI - H)(rI + H)^{-1}\right|\!\right|_2 < 1 \text{ (resp. } \leq 1).$$

b) Sean H_1 y H_2 hermíticas definida y semidefinida positiva, respectivamente. Dado $u^0 \in \mathbf{V}$ arbitrario se define la sucesión $\{u^k\}_{k=0}^{\infty}$ por

$$\begin{cases} (H_1 + rI)u^{k+\frac{1}{2}} = (rI - H_2)u^k + b \\ (H_2 + rI)u^{k+1} = (rI - H_1)u^{k+\frac{1}{2}} + b \end{cases}$$

siendo $r > 0$ y $b \in \mathbf{V}$ fijos. Expresar u^{k+1} en la forma

$$u^{k+1} = Bu^k + c$$

dando explícitamente la matriz B y el vector c. Probar que $\varrho(B) < 1$.

c) Demostrar que la sucesión $\{u^k\}_{k=0}^{\infty}$ es convergente y su límite es solución de un sistema lineal independiente de r. ¿Cuál es dicho sistema?

5.7. Prácticas

5.1. Escribir un programa en `MATLAB` que resuelva un sistema lineal mediante el método de Jacobi por puntos, pidiendo por pantalla, además de la matriz y el segundo miembro, el número máximo de iteraciones y la precisión para el test de parada.

5.2. Escribir un programa en `MATLAB` que resuelva un sistema lineal mediante el método de relajación por puntos, pidiendo por pantalla, además de la matriz y el segundo miembro, el parámetro de relajación, el número máximo de iteraciones y la precisión para el test de parada.

5.3. Dado $n \in \mathbb{N}$ se considera la matriz $A = (a_{ij})_{i,j=1}^{n}$ donde

$$a_{ij} = \begin{cases} 20 + i & \text{si} \quad i = j \\ \dfrac{(-1)^{i+j}}{i+j} & \text{si} \quad i \neq j \end{cases}$$

y el vector $b = (b_i)_{i=1}^{n}$ con

$$b_i = \frac{1}{i}$$

para $i = 1, 2, \ldots, n$. Resolver, para valores grandes de n, el sistema lineal $Au = b$ mediante el método de relajación por puntos, tomando como valores del parámetro $w = 0.1 : 0.1 : 1.9$ y con una precisión en el test de parada de 10^{-10}. Determinar el mejor valor de w entre todos los anteriores.

5.4. Programar en `MATLAB` el método de Jacobi por bloques para matrices de diagonal estrictamente dominante de forma que, en cada iteración, los sistemas asociados a los bloques diagonales se resuelvan mediante factorización LU.

5.5. Programar en `MATLAB` el método de relajación por bloques para matrices simétricas definidas positivas de forma que, en cada iteración, los sistemas asociados a los bloques diagonales se resuelvan mediante factorización de Cholesky.

5.6. Dada la matriz A descompuesta por bloques en la forma $A = (A_{ij})_{i,j=1}^{20}$ donde

$$A_{ij} = \frac{(-1)^{i+j}}{i+j} I_{20} \quad \text{si} \quad i \neq j,$$

siendo I_{20} la matriz identidad de orden 20, y

$$A_{ii} = \begin{pmatrix} 8 & -1 & -1 & & & & \\ -1 & 8 & -1 & -1 & & & \\ -1 & -1 & 8 & -1 & -1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & -1 & 8 & -1 & -1 \\ & & & -1 & -1 & 8 & -1 \\ & & & & -1 & -1 & 8 \end{pmatrix}$$

aplicar los programas de las prácticas 5.4 y 5.5 para resolver el sistema lineal $Au = b$ siendo $b = (1, 1, \ldots, 1)^{\mathrm{T}}$.

6 Interpolación numérica

6.1. Introducción

Son muchas y muy distintas las situaciones en las que, en el quehacer científico, aparecen series de datos o resultados de mediciones experimentales de los cuales sólo se conoce cómo se comportan en una cierta cantidad finita de ítems (por ejemplo, la población de un país durante los años de una década) y para los cuales se necesita encontrar una "ley general" que sirva para su tratamiento (como, continuando con el ejemplo anterior, conocer la población a mediados de uno de los años o la tasa de crecimiento de ésta). Esa "ley general" a la que se adaptan esos datos no es otra cosa que una función que tome los valores predeterminados.

Éste es, precisamente, el cometido de la *interpolación*: dada una tabla de datos (o, lo que es lo mismo, una función de la que se desconoce su expresión general y sólo se conocen los valores que toma en unos cuantos puntos) se trata de encontrar una función que tome los valores requeridos en los puntos dados. La función interpoladora servirá para sustituir a la función desconocida, tanto para evaluarla en puntos en los que no se conoce su valor (*interpolación*, en sentido estricto), como para conocer su tasa de variación (*diferenciación numérica*) o su distribución acumulativa (*integración numérica*). De hecho, y en ello reside su mayor importancia, la interpolación sirve para fundamentar una amplia gama de métodos para diferenciación e integración numéricas, así como para el tratamiento numérico de ecuaciones diferenciales.

La función interpoladora debe ser, por tanto, fácil de evaluar, derivar e integrar. Dependiendo del tipo de funciones que se consideren, se distinguen varios tipos de *interpolación*, entre los que podemos citar:

- Polinómica (Lagrange, Hermite).
- Trigonométrica.
- Funciones *spline* (interpolación polinomial a trozos).
- Funciones racionales.
- Exponencial.

En este capítulo centraremos nuestro estudio en la *interpolación de Lagrange* y la interpolación mediante funciones *spline* cúbicas.

6.2. Interpolación de Lagrange

El problema de la *interpolación de Lagrange* consiste en lo siguiente: "Dada una función $f : [a, b] \to \mathbb{R}$ y $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$ con $x_i \neq x_j$ si $i \neq j$ (donde no se supone que estos puntos sean equidistantes ni tampoco que estén enunciados en su orden natural) encontrar un polinomio P verificando

$$P(x_i) = f(x_i)$$

para $i = 0, 1, \ldots, n$".

Notación 6.1. En todo lo que sigue denotaremos por

$$\mathbb{R}[x] = \left\{P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,\ a_i \in \mathbb{R},\ n \in \mathbb{N} \cup \{0\}\right\}$$

al *conjunto de polinomios* reales en la variable x, por

$$\mathcal{P}_n = \{P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{R}[x]\},\ n \in \mathbb{N} \cup \{0\}$$

al conjunto de polinomios reales de grado menor o igual que n y por ∂P al *grado* del polinomio P. □

Para ilustrar el problema anterior supongamos que queremos encontrar un polinomio P que tome valores $\{y_0, y_1, y_2, y_3\}$ en un conjunto de puntos $\{x_0, x_1, x_2, x_3\}$. Si el polinomio buscado es de la forma

$$P(x) = a_3 x^3 + a_2 x^2 + a_1 x + a_0 \in \mathcal{P}_3$$

las condiciones anteriores determinan que

$$y_i = P(x_i) = a_3 x_i^3 + a_2 x_i^2 + a_1 x_i + a_0$$

para $i = 0, 1, 2, 3$. Luego los coeficientes buscados $\{a_0, a_1, a_2, a_3\}$ vienen dados como la solución del siguiente sistema lineal de cuatro ecuaciones con cuatro incógnitas:

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + a_3 x_0^3 = y_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + a_3 x_1^3 = y_1 \\ a_0 + a_1 x_2 + a_2 x_2^2 + a_3 x_2^3 = y_2 \\ a_0 + a_1 x_3 + a_2 x_3^2 + a_3 x_3^3 = y_3 \end{cases}$$

o, equivalentemente, en forma matricial, $Ax = b$ donde

$$A = \begin{pmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \end{pmatrix}, \; x = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} \text{ y } b = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{pmatrix}.$$

El determinante del sistema anterior es de Vandermonde:

$$\begin{aligned} \det(A) &= \begin{vmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \end{vmatrix} = \prod_{i>j} (x_i - x_j) \\ &= (x_3 - x_2)(x_3 - x_1)(x_3 - x_0)(x_2 - x_1)(x_2 - x_0)(x_1 - x_0). \end{aligned}$$

Así, si los puntos $\{x_0, x_1, x_2, x_3\}$ son distintos, entonces $\det(A) \neq 0$, por lo que existe una única solución $\{a_0, a_1, a_2, a_3\}$ para cualquier valor de $\{y_0, y_1, y_2, y_3\}$. Sin embargo, el hecho de que las matrices de Vandermonde estén mal condicionadas, unido a argumentos de otro tipo que se comentarán más adelante, hacen que, en la práctica, el polinomio de interpolación no se calcule mediante la resolución del sistema lineal así asociado.

El resultado fundamental de esta sección es el que muestra la existencia y unicidad del polinomio de interpolación así como la forma concreta que éste va a tener. Más precisamente,

Teorema 6.1 (Fórmula de interpolación de Lagrange). *Sea $f : [a, b] \to \mathbb{R}$ y $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$ con $x_i \neq x_j$ si $i \neq j$. Existe un único polinomio $P_n \in \mathcal{P}_n$ que verifica*

$$P_n(x_i) = f(x_i)$$

para $i = 0, 1, \ldots, n$. Además, este polinomio viene dado por

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x) \tag{6.1}$$

donde, para cada $i \in \{0, 1, \ldots, n\}$,

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}. \tag{6.2}$$

DEMOSTRACIÓN.

a) Existencia: teniendo en cuenta que para cada $i \in \{0, 1, \ldots, n\}$

$$L_i(x) = \frac{x - x_0}{x_i - x_0} \frac{x - x_1}{x_i - x_1} \cdots \frac{x - x_{i-1}}{x_i - x_{i-1}} \frac{x - x_{i+1}}{x_i - x_{i+1}} \cdots \frac{x - x_n}{x_i - x_n}$$

es inmediato comprobar que

$$\begin{cases} L_i \in \mathcal{P}_n,\ i = 0, 1, \ldots, n & \Rightarrow \quad P_n \in \mathcal{P}_n \\ L_i(x_j) = \delta_{ij} & \Rightarrow \quad P_n(x_j) = f(x_j),\ j = 0, 1, \ldots, n. \end{cases}$$

b) Unicidad: supongamos que existen dos polinomios $P_n, Q_n \in \mathcal{P}_n$ tales que

$$P_n(x_i) = f(x_i) = Q_n(x_i)$$

para $i = 0, 1, \ldots, n$. De esta forma, el polinomio $D_n = P_n - Q_n$ verifica $D_n \in \mathcal{P}_n$ y, para cada $i \in \{0, 1, \ldots, n\}$,

$$D_n(x_i) = P_n(x_i) - Q_n(x_i) = 0.$$

Es decir, D_n es un polinomio de grado menor o igual que n con $n+1$ raíces; consecuentemente, por el teorema Fundamental del Álgebra, $D_n \equiv 0$ de donde se concluye que $P_n \equiv Q_n$. □

Definición 6.1. El polinomio P_n dado en (6.1) se denomina *polinomio de interpolación de Lagrange* relativo a la función f (o, simplemente, *polinomio de interpolación* de f) en los puntos $\{x_0, x_1, \ldots, x_n\}$ y los polinomios $L_i(x)$ dados en (6.2) son los *polinomios básicos de interpolación de Lagrange.* □

Ejemplo 6.1. Encontrar el polinomio de interpolación para la siguiente tabla

x_i	2	3	-1	4
$f(x_i)$	1	2	3	4

(6.3)

En este caso,

$$\begin{cases} L_0(x) = \dfrac{(x-3)(x+1)(x-4)}{(-1)\cdot 3 \cdot (-2)} = \dfrac{(x-3)(x+1)(x-4)}{6} \\ L_1(x) = \dfrac{(x-2)(x+1)(x-4)}{1 \cdot 4 \cdot (-1)} = -\dfrac{(x-2)(x+1)(x-4)}{4} \\ L_2(x) = \dfrac{(x-2)(x-3)(x-4)}{(-3)\cdot(-4)\cdot(-5)} = -\dfrac{(x-2)(x-3)(x-4)}{60} \\ L_3(x) = \dfrac{(x-2)(x-3)(x+1)}{2 \cdot 1 \cdot 5} = \dfrac{(x-2)(x-3)(x+1)}{10}. \end{cases}$$

De esta forma, el polinomio de interpolación buscado es

$$\begin{aligned} P_3(x) = \sum_{i=0}^{3} f(x_i)L_i(x) &= \frac{1}{6}(x-3)(x+1)(x-4) - \frac{1}{2}(x-2)(x+1)(x-4) \\ &\quad - \frac{1}{20}(x-2)(x-3)(x-4) + \frac{2}{5}(x-2)(x-3)(x+1) \\ &= \frac{1}{60}(x^3 + 21x^2 - 64x + 96) \end{aligned}$$

y viene representado en la figura 6.1. □

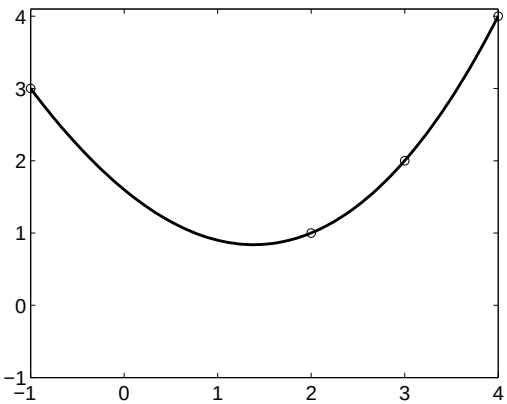


Figura 6.1: Polinomio de interpolación de la tabla (6.3).

Observación 6.1.

1. El cálculo del polinomio de interpolación a partir de la fórmula (6.1) requiere muchas operaciones. Además, una vez determinado el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$, si añadimos un nuevo punto x_{n+1} distinto a los anteriores y queremos hallar el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_n, x_{n+1}\}$, debemos repetir todo el proceso, dado que cambian todos los polinomios básicos $\{L_0(x), L_1(x), \ldots, L_n(x)\}$. Por esta razón, tampoco será éste el algoritmo que se use para calcular el polinomio de interpolación de Lagrange.

2. Al interpolar una función f en $n+1$ puntos distintos puede ocurrir que $\partial P_n \neq n$; de hecho, el grado de P_n puede ser pequeño aunque n sea grande. Así, por ejemplo, debido a la unicidad, $P_3(x) = 2x - 3$ es el polinomio de interpolación de la función $f(x) = x^4 - 2x^3 - x^2 + 4x - 3$ en los puntos $\{-1, 0, 1, 2\}$ (véase la figura 6.2). □

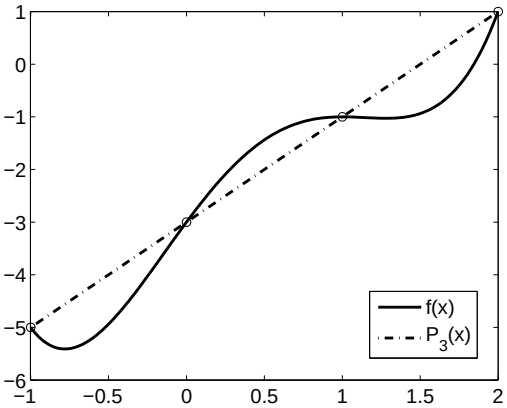


Figura 6.2: Polinomio de interpolación $P_3(x) = 2x - 3$.

En lo que sigue, utilizaremos con bastante frecuencia la siguiente:

Notación 6.2. Dados $n+1$ puntos distintos $\{x_0, x_1, \ldots, x_n\}$ denotaremos

$$\Pi_n(x) = \prod_{i=0}^{n}(x - x_i) = (x - x_0)(x - x_1)\cdots(x - x_n). \tag{6.4}$$

6.2.1. El error de interpolación

Como deseamos usar el polinomio de interpolación para sustituir el valor de la función f en puntos que no pertenecen al conjunto de puntos de interpolación $\{x_0, x_1, \ldots, x_n\}$, estamos interesados en estimar el *error*

$$E_n(x) = f(x) - P_n(x),\ x \in [a, b].$$

Sin hipótesis adicionales, nada podemos decir acerca de esa cantidad pues podemos cambiar la función f en puntos que no sean los de interpolación sin que cambie el polinomio de interpolación (véase la figura 6.3). Además, si la función f está tabulada y no se conoce su expresión analítica, entonces, estrictamente hablando, es imposible estimar el error que comete el polinomio de interpolación.

No obstante, cuando la función f es suficientemente regular, podemos precisar el error que se comete en cada punto de interpolación en términos de las derivadas de f. Concretamente,

Teorema 6.2. *Sea $f \in \mathcal{C}^{n+1}([a, b])$, $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$ con $x_i \neq x_j$ si $i \neq j$ y $P_n \in \mathcal{P}_n$ el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_n\}$. Para*

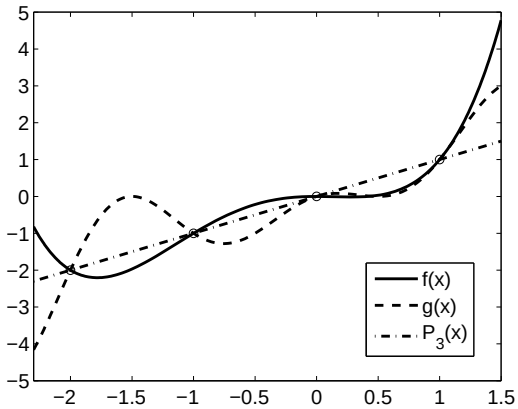


Figura 6.3: $f(x) = 0.5x^4 + x^3 - 0.5x^2$, $g(x) = x(1 - \operatorname{sen} \pi x)$ y $P_3(x) = x$.

cada $x \in [a, b]$ *existe* $\xi_x \in I_x$, *siendo* I_x *el mínimo intervalo cerrado que contiene a los puntos* $\{x_0, x_1, \ldots, x_n, x\}$, *tal que*

$$E_n(x) = f(x) - P_n(x) = \frac{f^{n+1)}(\xi_x)}{(n+1)!}\Pi_n(x). \tag{6.5}$$

DEMOSTRACIÓN. Dado $x \in [a, b]$ pueden presentarse dos casos:

a) $x = x_i$ para algún $i \in \{0, 1, \ldots, n\}$. En este caso el resultado es evidente, pues $f(x_i) = P_n(x_i)$ y $\Pi_n(x_i) = 0$.

b) $x \neq x_i$ para todo $i \in \{0, 1, \ldots, n\}$. Consideramos la función $F : [a, b] \to \mathbb{R}$ dada por

$$F(y) = [f(y) - P_n(y)]\Pi_n(x) - [f(x) - P_n(x)]\Pi_n(y),\ y \in [a, b].$$

Claramente $F \in \mathcal{C}^{n+1}([a, b])$,

$$F(x_i) = [f(x_i) - P_n(x_i)]\Pi_n(x) - [f(x) - P_n(x)]\Pi_n(x_i) = 0$$

para $i = 0, 1, \ldots, n$ y

$$F(x) = [f(x) - P_n(x)]\Pi_n(x) - [f(x) - P_n(x)]\Pi_n(x) = 0.$$

Es decir, la función F tiene, al menos, $n + 2$ raíces distintas en el intervalo $[a, b]$. Consecuentemente, por el teorema de Rolle, la función F' tiene al menos $n + 1$ raíces en I_x, el menor de los intervalos cerrados que contiene a x y a los puntos $\{x_0, x_1, \ldots, x_n\}$, la función F'' tiene al menos n raíces en

dicho intervalo y, reiterando el argumento, la función $F^{n+1)}$ tiene al menos una raíz $\xi_x \in I_x$. Como $\partial P_n \leq n$, entonces $P_n^{n+1)}(y) \equiv 0$; análogamente,

$$\Pi_n^{n+1)}(y) = (n+1)!$$

por ser Π_n un polinomio de grado $n+1$ cuyo coeficiente del término de mayor grado es 1. De esta forma,

$$\begin{aligned} F^{n+1)}(y) &= \left[f^{n+1)}(y) - P_n^{n+1)}(y)\right] \Pi_n(x) - [f(x) - P_n(x)]\, \Pi_n^{n+1)}(y) \\ &= f^{n+1)}(y)\Pi_n(x) - [f(x) - P_n(x)](n+1)! \end{aligned}$$

para cada $y \in [a,b]$. En particular,

$$0 = F^{n+1)}(\xi_x) = f^{n+1)}(\xi_x)\Pi_n(x) - [f(x) - P_n(x)](n+1)!$$

de donde se concluye el resultado. □

Observación 6.2.

1. La función $E_n(x)$ dada en (6.5) no puede usarse para calcular el valor exacto del error $E_n = f - P_n$, pues ξ_x como función de x es, en general, desconocida (salvo cuando $f^{n+1)} \equiv$ cte, es decir, $f \in \mathcal{P}_{n+1}$). No obstante, podemos utilizar la fórmula anterior para obtener una *cota del error de interpolación*, pues

$$|f(x) - P_n(x)| \leq \frac{|\Pi_n(x)|}{(n+1)!} \left|\left| f^{n+1)} \right|\right|_{\mathrm{L}^\infty(a,b)}, \; x \in [a,b] \tag{6.6}$$

siendo

$$||g||_{\mathrm{L}^\infty(a,b)} = \max_{a \leq x \leq b} |g(x)| \tag{6.7}$$

la *norma del máximo* de una función continua $g : [a,b] \to \mathbb{R}$.

2. La estimación del error precedente es *óptima* en el sentido de que existe una función para la que se da la igualdad. En efecto, basta considerar la función

$$f(x) = \Pi_n(x) = \prod_{i=0}^{n} (x - x_i)$$

para la que se verifica

$$P_n(x) = 0 \;\text{ y }\; f^{n+1)}(x) = (n+1)!$$

lo que determina la igualdad $|\Pi_n(x)| = |\Pi_n(x)|$ en (6.6). □

Corolario 6.1. *En las condiciones del teorema* 6.2*, la función* $x \mapsto f^{n+1)}(\xi_x)$ *puede extenderse de forma continua a todo el intervalo* $[a, b]$.

DEMOSTRACIÓN. Consideremos la función $g : [a, b] \to \mathbb{R}$ definida como

$$g(x) = \begin{cases} (n+1)! \, \dfrac{f(x) - P_n(x)}{\Pi_n(x)}, & x \notin \{x_0, x_1, \ldots, x_n\} \\ \\ (n+1)! \, \dfrac{f'(x_i) - P_n'(x_i)}{\Pi_n'(x_i)}, & x = x_i, \; i = 0, 1, \ldots, n. \end{cases} \tag{6.8}$$

En primer lugar, por ser los puntos $\{x_0, x_1, \ldots, x_n\}$ distintos, se verifica que $\Pi_n'(x_i) \neq 0$, $i = 0, 1, \ldots, n$ (véase el problema 6.5) por lo que la función g está bien definida. Claramente, la función g es continua en los puntos que no son de interpolación y, por la regla de L'Hôpital, también en los puntos $\{x_0, x_1, \ldots, x_n\}$; así pues $g \in \mathcal{C}([a, b])$. Por otra parte, por el teorema 6.2 sabemos que

$$f^{n+1)}(\xi_x) = (n+1)! \, \frac{f(x) - P_n(x)}{\Pi_n(x)} = g(x)$$

siempre que $x \notin \{x_0, x_1, \ldots, x_n\}$, con lo que se tiene el resultado. □

Ejemplo 6.2. Si consideramos la función

$$f(x) = \operatorname{sen} x, \; x \in \left[0, \frac{\pi}{2}\right]$$

se verifica que

$$P_1(x) = \frac{\left(\dfrac{\pi}{2} - x\right) f(0) + (x - 0) f\left(\dfrac{\pi}{2}\right)}{\dfrac{\pi}{2} - 0} = \frac{2}{\pi} x$$

es el polinomio de interpolación de f en los puntos $\left\{x_0 = 0, x_1 = \dfrac{\pi}{2}\right\}$. En la figura 6.4 se representan la función f y su polinomio de interpolación. Como, en este caso,

$$|f''(x)| = |\operatorname{sen} x| \leq 1, \; x \in \left[0, \frac{\pi}{2}\right]$$

y el máximo de la función

$$|\Pi_1(x)| = \left| x \left(x - \frac{\pi}{2}\right) \right| = x \left(\frac{\pi}{2} - x\right), \; x \in \left[0, \frac{\pi}{2}\right]$$

se presenta en $\widetilde{x} = \dfrac{\pi}{4}$ y toma el valor $|\Pi_1(\widetilde{x})| = \dfrac{\pi^2}{16}$ entonces

$$|E_1(x)| = |f(x) - P_1(x)| \leq \frac{\pi^2}{32} \simeq 0.308425, \; x \in \left[0, \frac{\pi}{2}\right]. \quad \square$$

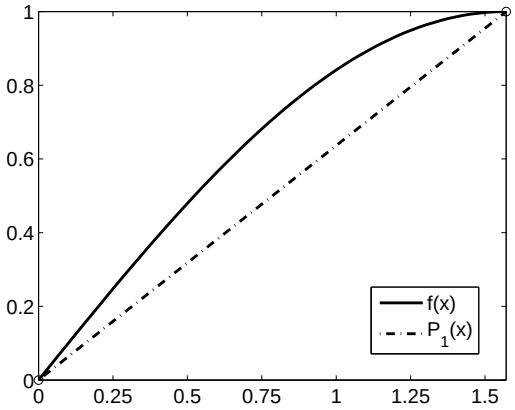


Figura 6.4: Función $f(x) = \operatorname{sen} x$ y polinomio $P_1(x) = \dfrac{2}{\pi}x$.

Sea $f : [a, b] \to \mathbb{R}$ y $P_n \in \mathcal{P}_n$ el polinomio de interpolación de f en puntos distintos $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$. Es razonable cuestionarse si será

$$\lim_{n \to +\infty} P_n(x) = f(x)$$

para todo $x \in [a, b]$. La respuesta es, en general, negativa como se muestra en el ejemplo siguiente:

Ejemplo 6.3. Consideremos la función $f : [-1, 1] \to \mathbb{R}$ dada por

$$f(x) = |x|$$

y los puntos de interpolación

$$x_k = -1 + \frac{2k}{n}$$

para $k = 0, 1, \ldots, n$, siendo $n \in \mathbb{N}$. En particular, para $n = 5$, $n = 10$, $n = 15$ y $n = 20$ los polinomios de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_n\}$ están representados en la figura 6.5. Puede apreciarse que, a medida que va aumentando n, se producen fuertes oscilaciones en los extremos del intervalo $[-1, 1]$ conocidas como *efectos de borde*. □

Observación 6.3. El ejemplo anterior muestra que una función continua no puede aproximarse, en general, de manera arbitrariamente precisa mediante polinomios de Lagrange relativos a puntos de interpolación equidistantes. Esto no excluye que pueda aproximarse por una sucesión adecuada, por ejemplo, mediante los *polinomios de Bernstein* que se emplean en el teorema de Weierstrass. □

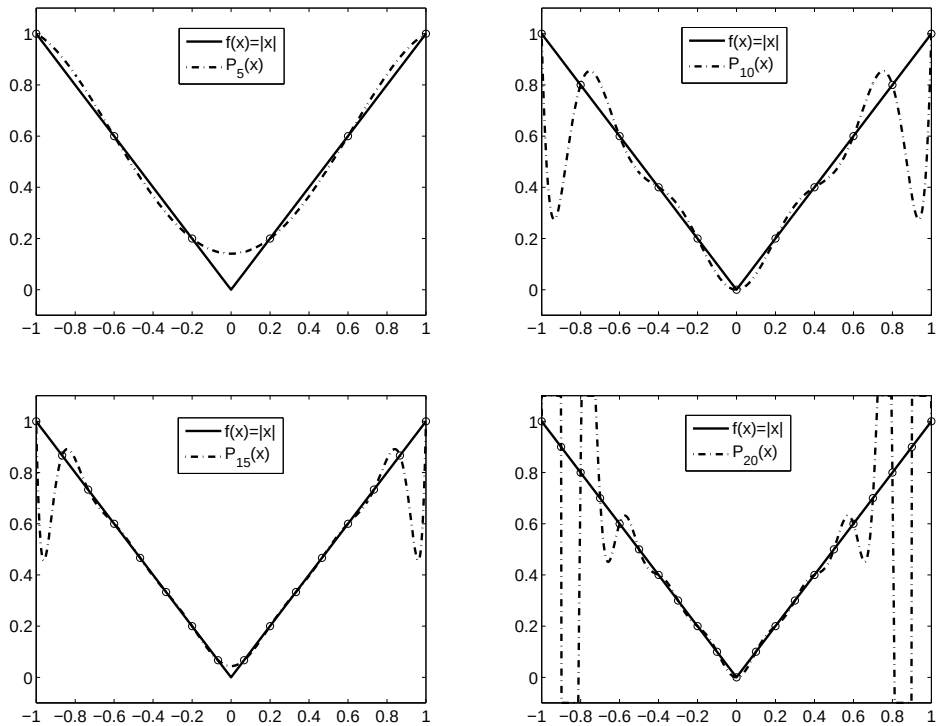


Figura 6.5: Polinomios de interpolación $P_5(x), P_{10}(x), P_{15}(x)$ y $P_{20}(x)$.

6.2.2. Fórmula de interpolación de Newton

Hasta ahora hemos visto cómo construir el polinomio de interpolación de Lagrange y conocemos una cota del error que cometemos al sustituir una función por su polinomio de interpolación.

En esta sección vamos a estudiar un método más eficiente de construcción del polinomio de interpolación. Supongamos que, una vez calculado dicho polinomio, necesitamos incluir otro punto de interpolación. Las razones pueden ser diversas: se han obtenido nuevos valores experimentales para la tabla que estamos interpolando, queremos obtener paulatinamente los polinomios de distintos grados para saber cuál nos interesa más, etc. En esta situación la fórmula de Lagrange no nos sirve pues, con sólo añadir un punto, todos los polinomios básicos $L_i(\cdot)$ cambian. Necesitamos una fórmula que permita encontrar el nuevo polinomio de interpolación a partir del que ya habíamos hallado, de forma que el trabajo realizado pueda aprovecharse.

La fórmula buscada es la *fórmula de interpolación de Newton* para el polinomio de interpolación de Lagrange, que nos va a permitir una representación del polinomio de interpolación en términos de "diferencias" (ya sean divididas o finitas) de valores de la función en los puntos de interpolación. Comencemos con la definición de estas "diferencias".

Definición 6.2. Sean $f : [a, b] \to \mathbb{R}$ y $\{x_0, x_1, x_2, \ldots\} \subset [a, b]$ tales que $x_i \neq x_j$ si $i \neq j$. Para cada $i \in \mathbb{N} \cup \{0\}$ sean

$$\begin{cases} f[x_i] = f(x_i) \\ f[x_i, x_{i+1}, \ldots, x_{i+m}] = \dfrac{f[x_i, x_{i+1}, \ldots, x_{i+m-1}] - f[x_{i+1}, x_{i+2}, \ldots, x_{i+m}]}{x_i - x_{i+m}}. \end{cases}$$

$f[x_i, x_{i+1}, \ldots, x_{i+m}]$ se denomina *diferencia dividida* de orden $m \in \mathbb{N} \cup \{0\}$ de f en el punto x_i. Análogamente, sean

$$\begin{cases} \Delta^0 f(x_i) = f(x_i) \\ \Delta^m f(x_i) = \Delta^{m-1} f(x_{i+1}) - \Delta^{m-1} f(x_i). \end{cases}$$

$\Delta^m f(x_i)$ es la *diferencia finita* de orden $m \in \mathbb{N} \cup \{0\}$ de f en el punto x_i. □

Observación 6.4. Los operadores Δ^m son *lineales*, es decir,

$$\Delta^m(\alpha f + \beta g)(x_i) = \alpha \Delta^m f(x_i) + \beta \Delta^m g(x_i),\ \alpha, \beta \in \mathbb{R},\ m \in \mathbb{N} \cup \{0\}. \quad \square$$

Ejemplo 6.4. Vamos a calcular las diferencias finitas y divididas de orden dos.

$$\begin{aligned} \Delta^2 f(x_i) &= \Delta f(x_{i+1}) - \Delta f(x_i) \\ &= (f(x_{i+2}) - f(x_{i+1})) - (f(x_{i+1}) - f(x_i)) \\ &= f(x_{i+2}) - 2f(x_{i+1}) + f(x_i) \end{aligned}$$

y

$$\begin{aligned} f[x_i, x_{i+1}, x_{i+2}] &= \frac{f[x_i, x_{i+1}] - f[x_{i+1}, x_{i+2}]}{x_i - x_{i+2}} \\ &= \frac{1}{x_i - x_{i+2}} \left(\frac{f(x_i) - f(x_{i+1})}{x_i - x_{i+1}} - \frac{f(x_{i+1}) - f(x_{i+2})}{x_{i+1} - x_{i+2}} \right). \end{aligned}$$

Si los puntos están *equiespaciados*, es decir,

$$x_i = x_0 + ih,\ i = 0, 1, \ldots, n \quad (h > 0) \tag{6.9}$$

entonces

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{1}{2h}\left(\frac{f(x_i) - f(x_{i+1})}{h} - \frac{f(x_{i+1}) - f(x_{i+2})}{h}\right)$$
$$= \frac{f(x_{i+2}) - 2f(x_{i+1}) + f(x_i)}{2h^2} = \frac{\Delta^2 f(x_i)}{2h^2}. \quad \square$$

En general, este último resultado puede generalizarse, obteniéndose la siguiente relación entre las diferencias divididas y las finitas:

Teorema 6.3. *Si $f : [a, b] \to \mathbb{R}$ y $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$ es una red de puntos de paso $h > 0$, es decir, $x_i = x_0 + ih$, $i = 0, 1, \ldots, n$, entonces*

$$f[x_i, x_{i+1}, \ldots, x_{i+m}] = \frac{\Delta^m f(x_i)}{m! h^m}.$$

DEMOSTRACIÓN. Lo mostramos por inducción sobre el orden de las diferencias:

i) Para las diferencias de orden $m = 1$, como $x_{i+1} = x_i + h$ entonces

$$f[x_i, x_{i+1}] = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{f(x_{i+1}) - f(x_i)}{h} = \frac{\Delta f(x_i)}{h}.$$

ii) Supongamos cierto el resultado para las diferencias de orden $m - 1$ y lo probamos para las de orden m. Por definición se tiene que

$$f[x_i, x_{i+1}, \ldots, x_{i+m}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+m}] - f[x_i, x_{i+1}, \ldots, x_{i+m-1}]}{x_{i+m} - x_i}.$$

Como $x_{i+m} - x_i = mh$, aplicando la hipótesis de inducción, podemos escribir

$$f[x_i, x_{i+1}, \ldots, x_{i+m}] = \frac{1}{mh}\left(\frac{\Delta^{m-1} f(x_{i+1})}{(m-1)! h^{m-1}} - \frac{\Delta^{m-1} f(x_i)}{(m-1)! h^{m-1}}\right)$$
$$= \frac{\Delta^{m-1} f(x_{i+1}) - \Delta^{m-1} f(x_i)}{m! h^m} = \frac{\Delta^m f(x_i)}{m! h^m}. \quad \square$$

Veamos a continuación que la i–ésima diferencia dividida $f[x_0, x_1, \ldots, x_i]$ es la misma independientemente del orden en que se tomen los puntos $\{x_0, x_1, \ldots, x_i\}$. Para ello, basta demostrar el siguiente resultado:

Teorema 6.4. *Si $f : [a, b] \to \mathbb{R}$ y $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$ con $x_i \neq x_j$ para $i \neq j$, entonces para cada $i \in \{0, 1, \ldots, n\}$ se verifica que*

$$f[x_0, x_1, \ldots, x_i] = \sum_{j=0}^{i} \frac{f(x_j)}{\prod\limits_{\substack{k=0 \\ k\neq j}}^{i} (x_j - x_k)}.$$

Demostración. Lo probamos por inducción sobre el número de puntos. Para dos puntos el resultado es obvio, pues

$$f[x_0, x_1] = \frac{f(x_0) - f(x_1)}{x_0 - x_1} = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}.$$

Supongamos cierto el resultado para i puntos y lo probamos para $i + 1$ puntos. Por definición, se tiene

$$f[x_0, x_1, \ldots, x_i] = \frac{f[x_0, x_1, \ldots, x_{i-1}] - f[x_1, x_2, \ldots, x_i]}{x_0 - x_i},$$

por lo que la hipótesis de inducción determina que

$$\begin{aligned} f[x_0, x_1, \ldots, x_i] &= \frac{1}{x_0 - x_i} \left(\sum_{j=0}^{i-1} \frac{f(x_j)}{\prod\limits_{\substack{k=0 \\ k\neq j}}^{i-1} (x_j - x_k)} - \sum_{j=1}^{i} \frac{f(x_j)}{\prod\limits_{\substack{k=1 \\ k\neq j}}^{i} (x_j - x_k)} \right) \\ &= \frac{1}{\prod\limits_{k=1}^{i-1} (x_0 - x_k)} \frac{f(x_0)}{x_0 - x_i} - \frac{1}{\prod\limits_{k=1}^{i-1} (x_i - x_k)} \frac{f(x_i)}{x_0 - x_i} \\ &\quad + \sum_{j=1}^{i-1} \left(\frac{1}{\prod\limits_{\substack{k=0 \\ k\neq j}}^{i-1} (x_j - x_k)} - \frac{1}{\prod\limits_{\substack{k=1 \\ k\neq j}}^{i} (x_j - x_k)} \right) \frac{f(x_j)}{x_0 - x_i}. \end{aligned}$$

De esta forma,

$$\begin{aligned} f[x_0,\ldots,x_i] &= \frac{f(x_0)}{\prod\limits_{\substack{k=0\\k\neq 0}}^{i}(x_0-x_k)} + \sum_{j=1}^{i-1}\frac{x_j-x_i-(x_j-x_0)}{\prod\limits_{\substack{k=0\\k\neq j}}^{i}(x_j-x_k)}\frac{f(x_j)}{x_0-x_i} + \frac{f(x_i)}{\prod\limits_{\substack{k=0\\k\neq i}}^{i}(x_i-x_k)} \\ &= \frac{f(x_0)}{\prod\limits_{\substack{k=0\\k\neq 0}}^{i}(x_0-x_k)} + \sum_{j=1}^{i-1}\frac{f(x_j)}{\prod\limits_{\substack{k=0\\k\neq j}}^{i}(x_j-x_k)} + \frac{f(x_i)}{\prod\limits_{\substack{k=0\\k\neq i}}^{i}(x_i-x_k)} \\ &= \sum_{j=0}^{i}\frac{f(x_j)}{\prod\limits_{\substack{k=0\\k\neq j}}^{i}(x_j-x_k)}. \quad \square \end{aligned}$$

Se obtiene, como consecuencia inmediata, la propiedad anteriormente comentada sobre la invarianza de las diferencias divididas:

Corolario 6.2. *Si $f:[a,b]\to\mathbb{R}$, $\{x_0,x_1,\ldots,x_n\}\subset[a,b]$ con $x_i\neq x_j$ para $i\neq j$ y σ es una permutación de $\{0,1,\ldots,i\}$, entonces*

$$f[x_0,x_1,\ldots,x_i] = f[x_{\sigma(0)},x_{\sigma(1)},\ldots,x_{\sigma(i)}]$$

para $i=0,1,\ldots,n$. $\square$

Estamos ya en condiciones de dar el resultado fundamental de esta subsección.

Teorema 6.5 (Fórmula de interpolación de Newton). *Sea $f:[a,b]\to\mathbb{R}$ y $\{x_0,x_1,\ldots,x_n\}\subset[a,b]$ con $x_i\neq x_j$ si $i\neq j$. El polinomio de interpolación de f en los puntos $\{x_0,x_1,\ldots,x_n\}$ viene dado por*

$$\begin{aligned} P_n(x) &= f(x_0)+\sum_{i=1}^{n}\Pi_{i-1}(x)f[x_0,x_1,\ldots,x_i] \\ &= f(x_0)+(x-x_0)f[x_0,x_1]+(x-x_0)(x-x_1)f[x_0,x_1,x_2] \\ &\quad +\cdots+(x-x_0)(x-x_1)\cdots(x-x_{n-1})f[x_0,x_1,\ldots,x_n]. \end{aligned}$$

Además, si $x\notin\{x_0,x_1,\ldots,x_n\}$, entonces

$$E_n(x) = f(x)-P_n(x) = \Pi_n(x)f[x_0,x_1,\ldots,x_n,x]. \tag{6.10}$$

DEMOSTRACIÓN. Procedemos por inducción sobre el grado del polinomio:

i) Para $n = 0$, $P_0(x) = f(x_0)$ es el polinomio de interpolación de f en x_0 y, para todo punto $x \neq x_0$, se verifica que

$$f[x_0, x] = \frac{f(x) - f(x_0)}{x - x_0},$$

por lo que

$$f(x) = f(x_0) + (x - x_0)f[x_0, x] = P_0(x) + \Pi_0(x)f[x_0, x].$$

ii) Suponemos cierto el resultado para $n - 1$, es decir, que

$$\begin{aligned} P_{n-1}(x) = f(x_0) &+ (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \\ &+ \cdots + (x - x_0)(x - x_1)\cdots(x - x_{n-2})f[x_0, x_1, \ldots, x_{n-1}] \end{aligned}$$

es el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_{n-1}\}$ y

$$f(x) - P_{n-1}(x) = \Pi_{n-1}(x)f[x_0, x_1, \ldots, x_{n-1}, x] \tag{6.11}$$

para $x \notin \{x_0, x_1, \ldots, x_{n-1}\}$. Consideremos el polinomio

$$\begin{aligned} Q(x) = f(x_0) &+ (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \\ &+ \cdots + (x - x_0)(x - x_1)\cdots(x - x_{n-2})f[x_0, x_1, \ldots, x_{n-1}] \\ &+ (x - x_0)(x - x_1)\cdots(x - x_{n-1})f[x_0, x_1, \ldots, x_n] \end{aligned}$$

que, por la hipótesis de inducción, podemos expresarlo como

$$Q(x) = P_{n-1}(x) + \Pi_{n-1}(x)f[x_0, x_1, \ldots, x_n].$$

Obviamente $Q \in \mathcal{P}_n$,

$$Q(x_i) = P_{n-1}(x_i) + \Pi_{n-1}(x_i)f[x_0, x_1, \ldots, x_n] = P_{n-1}(x_i) = f(x_i)$$

para $i = 0, 1, \ldots, n - 1$ y

$$Q(x_n) = P_{n-1}(x_n) + \Pi_{n-1}(x_n)f[x_0, x_1, \ldots, x_n] = f(x_n)$$

aplicando (6.11) en el punto $x = x_n$. Por tanto, por unicidad del polinomio de interpolación, Q es el polinomio de interpolación P_n de f en los puntos $\{x_0, x_1, \ldots, x_n\}$. Por otra parte, para todo punto $x \notin \{x_0, x_1, \ldots, x_n\}$ se verifica, teniendo en cuenta el corolario 6.2, que

$$\begin{aligned} f[x_0, x_1, \ldots, x_n, x] &= f[x, x_0, \ldots, x_{n-1}, x_n] \\ &= \frac{f[x, x_0, \ldots, x_{n-2}, x_{n-1}] - f[x_0, x_1, \ldots, x_{n-1}, x_n]}{x - x_n} \\ &= \frac{f[x_0, x_1, \ldots, x_{n-1}, x] - f[x_0, x_1, \ldots, x_{n-1}, x_n]}{x - x_n}, \end{aligned}$$

de donde

$$f[x_0, x_1, \ldots, x_{n-1}, x] = f[x_0, x_1, \ldots, x_{n-1}, x_n] + (x - x_n) f[x_0, x_1, \ldots, x_n, x].$$

Sustituyendo este valor en (6.11) se obtiene que

$$f(x) - P_{n-1}(x) = \Pi_{n-1}(x) \left(f[x_0, \ldots, x_{n-1}, x_n] + (x - x_n) f[x_0, \ldots, x_n, x] \right)$$

para $x \notin \{x_0, x_1, \ldots, x_n\}$, es decir,

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

de donde se sigue (6.10). □

Observación 6.5.

1. De (6.5) y (6.10) se deduce que, cuando la función $f \in \mathcal{C}^{n+1}([a, b])$, para cada punto $x \in [a, b] \backslash \{x_0, x_1, \ldots, x_n\}$ existe $\xi_x \in I_x$ verificando

$$f[x_0, x_1, \ldots, x_n, x] = \frac{f^{n+1)}(\xi_x)}{(n+1)!}. \tag{6.12}$$

   Nótese que, aunque de esta última expresión pudiera parecer que se determina el error de forma exacta, esto no es del todo cierto, pues para calcular $f[x_0, x_1, \ldots, x_n, x]$ se necesita el valor de $f(x)$ que, en general, es desconocido (obviamente, si conocemos el valor que toma la función f en el punto x entonces el error que se comete $f(x) - P_n(x)$ se determina explícitamente).

2. En el caso particular de que los puntos $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$ constituyan una red de paso $h > 0$, entonces el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_n\}$ viene dado por:

$$\begin{aligned} P_n(x) &= f(x_0) + \sum_{i=1}^{n} \Pi_{i-1}(x) \frac{\Delta^i f(x_0)}{i! h^i} \\ &= f(x_0) + (x - x_0) \frac{\Delta f(x_0)}{h} + (x - x_0)(x - x_1) \frac{\Delta^2 f(x_0)}{2h^2} \\ &\quad + \cdots + (x - x_0)(x - x_1) \cdots (x - x_{n-1}) \frac{\Delta^n f(x_0)}{n! h^n}. \end{aligned}$$

3. En la tabla 6.1 se muestran los algoritmos para construir el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_n\}$ a partir de las diferencias

TABLA 6.1:
Algoritmo de Newton (*diferencias divididas*)

$f(x_0)$	$f[x_0, x_1]$	$\cdots$	$f[x_0, x_1, \ldots, x_{n-1}]$	$f[x_0, x_1, \ldots, x_n]$
$f(x_1)$	$f[x_1, x_2]$	$\cdots$	$f[x_1, x_2, \ldots, x_n]$	
$f(x_2)$	$f[x_2, x_3]$	$\cdots$		
$\cdots$	$\cdots$	$\cdots$		
$f(x_{n-2})$	$f[x_{n-2}, x_{n-1}]$			
$f(x_{n-1})$	$f[x_{n-1}, x_n]$			
$f(x_n)$				

Algoritmo de Newton (*diferencias finitas*)

$f(x_0)$	$\Delta f(x_0)$	$\Delta^2 f(x_0)$	$\cdots$	$\Delta^{n-1} f(x_0)$	$\Delta^n f(x_0)$
$f(x_1)$	$\Delta f(x_1)$	$\Delta^2 f(x_1)$	$\cdots$	$\Delta^{n-1} f(x_1)$	
$f(x_2)$	$\Delta f(x_2)$	$\Delta^2 f(x_2)$	$\cdots$		
$\cdots$	$\cdots$	$\cdots$	$\cdots$		
$f(x_{n-2})$	$\Delta f(x_{n-2})$	$\Delta^2 f(x_{n-2})$			
$f(x_{n-1})$	$\Delta f(x_{n-1})$				
$f(x_n)$					

divididas y finitas, respectivamente. Como se observa, en la fórmula de interpolación de Newton sólo intervienen los elementos de la primera fila de las tablas triangulares citadas, aunque es necesario considerar todos los demás elementos para obtener éstos. □

La propiedad más importante de la fórmula de interpolación de Newton es que permite obtener el polinomio de interpolación de f en ciertos puntos a partir de polinomios de interpolación en subconjuntos de ellos. En particular,

Corolario 6.3. *Sea* $f : [a, b] \to \mathbb{R}$ *y* $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$ *con* $x_i \neq x_j$ *si* $i \neq j$. *Si* P_n *es el polinomio de interpolación de* f *en los puntos* $\{x_0, x_1, \ldots, x_n\}$ *y* x_{n+1} *es otro punto de* $[a, b]$ *tal que* $x_{n+1} \notin \{x_0, x_1, \ldots, x_n\}$, *entonces*

$$P_{n+1}(x) = P_n(x) + \Pi_n(x) f[x_0, x_1, \ldots, x_{n+1}]$$

es el polinomio de interpolación de f *en los puntos* $\{x_0, x_1, \ldots, x_n, x_{n+1}\}$. □

Observación 6.6. La relación (6.12) muestra que si $\{x_0, x_1, \ldots, x_n\}$ son puntos distintos de $[a, b]$ y $f \in \mathcal{C}^n([a, b])$ entonces existe un valor ξ intermedio entre los

puntos $\{x_0, x_1, \ldots, x_n\}$ tal que

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{n)}(\xi)}{n!}. \tag{6.13}$$

Nótese que para $n = 1$ se trata del clásico teorema del Valor Medio

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} = f[x_0, x_1] = f'(\xi)$$

con ξ entre x_0 y x_1.

Por otra parte, para $f \in \mathcal{C}^{n+1}([a, b])$, a partir de la fórmula de interpolación de Newton podemos escribir

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

y, utilizando la relación (6.13),

$$\begin{aligned} f(x) = f(x_0) &+ f'(\xi_1)(x - x_0) + \frac{f''(\xi_2)}{2!}(x - x_0)(x - x_1) \\ &+ \cdots + \frac{f^{n)}(\xi_n)}{n!}(x - x_0)(x - x_1) \ldots (x - x_{n-1}) \\ &+ \frac{f^{n+1)}(\xi_{n+1})}{(n+1)!}(x - x_0)(x - x_1) \ldots (x - x_n) \end{aligned}$$

donde cada ξ_k se encuentra entre los puntos $\{x_0, x_1, \ldots, x_k\}$ para $k = 1, 2, \ldots, n$ y ξ_{n+1} entre $\{x_0, x_1, \ldots, x_n, x\}$. Si hacemos que todos los puntos de interpolación $\{x_0, x_1, \ldots, x_n\}$ converjan a x_0 entonces los valores $\{\xi_0, \xi_1, \ldots, \xi_n\}$ también convergerán a x_0, convirtiéndose la expresión anterior en

$$\begin{aligned} f(x) = f(x_0) &+ f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 \\ &+ \cdots + \frac{f^{n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}, \end{aligned}$$

siendo ξ un valor entre x_0 y x. Hemos obtenido así la *fórmula de Taylor* como caso límite de un proceso de interpolación: en lugar de considerar $n+1$ puntos distintos de interpolación tenemos un único punto de interpolación de multiplicidad $n + 1$. En este caso, el polinomio de Taylor

$$P_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{n)}(x_0)}{n!}(x - x_0)^n$$

interpola a f, y las derivadas de P_n a las derivadas de f hasta el orden n, en el punto x_0. □

6.2.3. Minimización del error

Según se ha visto, si $f \in \mathcal{C}^{n+1}([a,b])$ y P_n es el polinomio de interpolación de f en $n+1$ puntos distintos $\{x_0, x_1, \ldots, x_n\} \subset [a,b]$, para cada $x \in [a,b]$ existe $\xi_x \in [a,b]$ tal que

$$E_n(x) = f(x) - P_n(x) = \frac{f^{n+1)}(\xi_x)}{(n+1)!}\Pi_n(x).$$

De esta forma, llamando

$$M_{n+1} = \left|\left| f^{n+1)} \right|\right|_{\mathrm{L}^\infty(a,b)}$$

se tiene que

$$||f - P_n||_{\mathrm{L}^\infty(a,b)} \leq \frac{M_{n+1}}{(n+1)!}\,||\Pi_n||_{\mathrm{L}^\infty(a,b)}$$

donde la norma del máximo está definida en (6.7). El problema que nos planteamos ahora es el de encontrar, de entre todos los polinomios de interpolación de Lagrange de grado menor o igual que n, el que minimice esta acotación óptima (véase la Observación 6.2) del error. Los únicos parámetros con los que se puede jugar son las abscisas de interpolación, puesto que la función y el grado del polinomio están fijados. Por tanto, debemos elegir los puntos de interpolación $\{x_0, x_1, \ldots, x_n\}$ que hagan mínimo el valor de $||\Pi_n||_{\mathrm{L}^\infty(a,b)}$. El primero en resolver este problema fue el matemático ruso *P. L. Tchebychev* (1821–1894) y su solución conduce a una clase de polinomios que también sirven para tratar otro tipo de problemas.

Observación 6.7. En todo lo que sigue, con vistas a simplificar la exposición, supondremos que estamos trabajando en el intervalo $[-1,1]$. Esto no supone ninguna pérdida de generalidad, como se prueba en el problema 6.15. □

Definición 6.3. Consideremos la sucesión de polinomios $\{T_n\}_{n=0}^{\infty}$ dada por

$$\begin{cases} T_0(x) = 1,\ T_1(x) = x \\ T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),\ n \in \mathbb{N}. \end{cases} \tag{6.14}$$

T_n se denomina *polinomio de Tchebychev* de orden $n \in \mathbb{N} \cup \{0\}$. □

Ejemplo 6.5. Los polinomios de Tchebychev que siguen a $T_0(x) = 1$ y $T_1(x) = x$ son:

$$T_2(x) = 2x^2 - 1,\ T_3(x) = 4x^3 - 3x,\ T_4(x) = 8x^4 - 8x^2 + 1,$$
$$T_5(x) = 16x^5 - 20x^3 + 5x,\ T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1, \ldots$$

En la figura 6.6 se representan gráficamente algunos de ellos. □

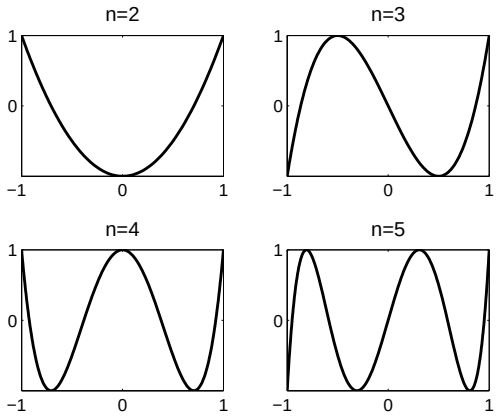


Figura 6.6: Polinomios de Tchebychev $T_2(x)$, $T_3(x)$, $T_4(x)$ y $T_5(x)$.

Destacamos, a continuación, las principales propiedades que tienen estos polinomios.

Proposición 6.1.

a) T_n *es un polinomio de grado* n *con coeficiente director* 2^{n-1}, $n \in \mathbb{N}$.

b) *Para cada* $n \in \mathbb{N} \cup \{0\}$ *se verifica que*

$$T_n(x) = \cos(n \arccos x),\ x \in [-1, 1].$$

c) *Para cada* $n \in \mathbb{N}$ *las* n *raíces de* T_n *se localizan en el intervalo* $[-1, 1]$ *en los puntos*

$$x_k = \cos \frac{(2k+1)\pi}{2n}$$

para $k = 0, 1, \ldots, n-1$.

d) *Para cada* $n \in \mathbb{N}$ *los valores extremos de* T_n *en el intervalo* $[-1, 1]$ *son* 1 *y* -1, *alcanzados alternativamente en los* $n+1$ *puntos*

$$z_k = \cos \frac{k\pi}{n}$$

para $k = 0, 1, \ldots, n$. *Consecuentemente, para cada* $n \in \mathbb{N} \cup \{0\}$,

$$||T_n||_{\mathrm{L}^\infty(-1,1)} = 1.$$

DEMOSTRACIÓN.

a) Véase el problema 6.13.

b) Para cada $n \in \mathbb{N} \cup \{0\}$ denotemos

$$\tau_n(x) = \cos(n \arccos x),\ x \in [-1, 1].$$

Si $\phi(x) = \arccos x$ entonces

$$\tau_n(x) = \cos n\phi(x),\ x \in [-1, 1],$$

lo que nos permite escribir

$$\begin{cases} \tau_{n+1}(x) = \cos\left((n+1)\phi(x)\right) = \cos n\phi(x)\cos\phi(x) - \operatorname{sen} n\phi(x)\operatorname{sen}\phi(x) \\ \tau_{n-1}(x) = \cos\left((n-1)\phi(x)\right) = \cos n\phi(x)\cos\phi(x) + \operatorname{sen} n\phi(x)\operatorname{sen}\phi(x); \end{cases}$$

sumando ambas expresiones se obtiene

$$\tau_{n+1}(x) + \tau_{n-1}(x) = 2\cos\phi(x)\cos n\phi(x) = 2x\tau_n(x).$$

Puesto que

$$\tau_0(x) = \cos(0 \arccos x) = 1 \ \text{ y } \ \tau_1(x) = \cos(\arccos x) = x,$$

las funciones τ_n verifican también la relación de recurrencia (6.14). Consecuentemente,

$$\tau_n(x) = T_n(x),\ x \in [-1, 1],\ n \in \mathbb{N} \cup \{0\}.$$

c) Busquemos las raíces de T_n en el intervalo $[-1, 1]$. Por la propiedad anterior se verifica que

$$\begin{aligned} T_n(x) = 0 \Leftrightarrow \cos n\phi(x) = 0 \ &\Leftrightarrow\ n\phi(x) = \frac{(2k+1)\pi}{2},\ k \in \mathbb{Z} \\ &\Leftrightarrow\ x = \cos\phi(x) = \cos\frac{(2k+1)\pi}{2n},\ k \in \mathbb{Z}. \end{aligned}$$

Es decir, las raíces de T_n en el intervalo $[-1, 1]$ se encuentran entre los números

$$x_k = \cos\frac{(2k+1)\pi}{2n},\ k \in \mathbb{Z}.$$

Ahora bien, los valores de $k \in \{0, 1, \ldots, n-1\}$ determinan n raíces distintas $\{x_0, x_1, \ldots, x_{n-1}\}$ situadas en el intervalo abierto $(-1, 1)$; como $\partial T_n = n$, entonces $\{x_0, x_1, \ldots, x_{n-1}\}$ son las n raíces de T_n (los valores de x_k que se obtienen para otros valores de k son repetición de éstos).

d) Como $T_n'(x) = -n \operatorname{sen} n\phi(x)\phi'(x)$ y

$$\phi'(x) = -\frac{1}{\sqrt{1-x^2}} < 0,\ x \in (-1,1)$$

entonces, para $x \in (-1,1)$ se verifica que

$$T_n'(x) = 0 \Leftrightarrow \operatorname{sen} n\phi(x) = 0 \Leftrightarrow n\phi(x) = k\pi,\ k \in \mathbb{Z}$$
$$\Leftrightarrow x = \cos\phi(x) = \cos\frac{k\pi}{n},\ k \in \mathbb{Z}.$$

Como ocurría antes, los valores de $k \in \{1, 2, \ldots, n-1\}$ determinan $n-1$ raíces distintas $\{z_1, z_2, \ldots, z_{n-1}\}$ de T_n' situadas en el intervalo abierto $(-1,1)$; como $\partial T_n' = n-1$ entonces $\{z_1, z_2, \ldots, z_{n-1}\}$ son las $n-1$ raíces de T_n'. Como en estos puntos

$$T_n(z_k) = \cos\left(n \arccos\left(\cos\frac{k\pi}{n}\right)\right) = \cos k\pi = (-1)^k$$

para $k = 1, 2, \ldots, n-1$, entonces T_n va alcanzando, alternativamente, los valores -1 y 1. Por otra parte, como en los extremos del intervalo se verifica

$$T_n(-1) = \cos n\pi = (-1)^n \ \text{ y } \ T_n(1) = \cos 0 = 1,$$

es en los puntos $\{z_0, z_1, \ldots, z_n\}$ (con $z_0 = 1$ y $z_n = -1$) donde T_n alcanza el máximo y el mínimo absolutos en el intervalo $[-1,1]$. □

Observación 6.8.

1. La ley de recurrencia (6.14) define una sucesión de polinomios definidos, por tanto, en todo $\mathbb{R}$. Cada uno de estos polinomios $T_n(x)$ coincide con la correspondiente función $\cos(n \arccos x)$ solamente en los puntos del intervalo $[-1,1]$, que son los únicos valores en los que la función arccos está definida.
2. A partir de la proposición 6.1, la expresión del polinomio T_n es:

$$T_n(x) = 2^{n-1} \prod_{k=0}^{n-1} \left(x - \cos\frac{(2k+1)\pi}{2n}\right). \quad \square$$

Observación 6.9. Otras propiedades de estos polinomios son (ver [**Ap**]):

1. La expresión explícita de T_n es

$$T_n(x) = \sum_{k=0}^{\left[\frac{n}{2}\right]} \binom{n}{2k} x^{n-2k}(x^2-1)^k.$$

2. T_{2n} es una función par con término independiente $(-1)^n$, mientras que T_{2n-1} es una función impar y, por tanto, sin término independiente (véase el problema 6.22).

3. Entre dos raíces consecutivas de la ecuación $T_{n+1}(x) = 0$ existe una única raíz de $T_n(x) = 0$.

4. Los polinomios de Tchebychev constituyen una familia ortogonal de polinomios (véase la definición 7.1) en el intervalo $[-1, 1]$ respecto a la función peso $w(x) = \dfrac{1}{\sqrt{1-x^2}}$. □

Regresemos al problema de encontrar un polinomio de grado prefijado para el que la norma del máximo sea lo más pequeña posible. Para ello introduzcamos la siguiente:

Notación 6.3. Vamos a denotar por

$$\mathcal{P}_n^{\mathrm{m}} = \{P(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathcal{P}_n\},\ n \in \mathbb{N} \cup \{0\}$$

al conjunto de polinomios *mónicos* de grado n. □

Teorema 6.6. *El polinomio mónico*

$$T_n^{\mathrm{m}}(x) = \frac{1}{2^{n-1}} T_n(x)$$

minimiza la norma del máximo en el intervalo $[-1, 1]$ *entre los polinomios mónicos de grado* n*, es decir,*

$$||P||_{\mathrm{L}^\infty(-1,1)} \geq ||T_n^{\mathrm{m}}||_{\mathrm{L}^\infty(-1,1)}$$

para todo $P \in \mathcal{P}_n^{\mathrm{m}}$.

Demostración. Por la proposición 6.1 sabemos que T_n^{m} es un polinomio mónico que toma, en el intervalo $[-1, 1]$, los valores extremos $\pm\dfrac{1}{2^{n-1}}$ alternativamente en los $n + 1$ puntos distintos

$$z_k = \cos\frac{k\pi}{n}$$

para $k = 0, 1, \ldots, n$; por tanto,

$$||T_n^{\mathrm{m}}||_{\mathrm{L}^\infty(-1,1)} = \frac{1}{2^{n-1}}.$$

Supongamos que existiera un polinomio $P \in \mathcal{P}_n^{\mathrm{m}}$ tal que

$$||P||_{\mathrm{L}^\infty(-1,1)} < ||T_n^{\mathrm{m}}||_{\mathrm{L}^\infty(-1,1)} = \frac{1}{2^{n-1}} \tag{6.15}$$

y derivemos una contradicción. Consideremos el polinomio

$$Q(x) = T_n^{\mathrm{m}}(x) - P(x).$$

Como T_n^{m} y P son polinomios mónicos de grado n, entonces $\partial Q \leq n-1$. Por otra parte,

$$Q(z_k) = T_n^{\mathrm{m}}(z_k) - P(z_k) = \frac{(-1)^k}{2^{n-1}} - P(z_k) = (-1)^k \left(\frac{1}{2^{n-1}} - (-1)^k P(z_k) \right)$$

para $k = 0, 1, \ldots, n$. El supuesto (6.15) conduce a

$$\frac{1}{2^{n-1}} - (-1)^k P(z_k) > 0$$

para $k = 0, 1, \ldots, n$, lo que implica que Q cambia alternativamente de signo en los $n+1$ puntos distintos $\{z_0, z_1, \ldots, z_n\}$. Así, por el teorema de Bolzano, sabemos que Q se anula al menos una vez entre dos cambios de signo consecutivos, teniendo, al menos, n raíces distintas; como, por otra parte, $\partial Q \leq n-1$, entonces $Q \equiv 0$, es decir, $P \equiv T_n^{\mathrm{m}}$ lo que es una contradicción. □

Observación 6.10. De hecho, puede demostrarse que si $P \in \mathcal{P}_n^{\mathrm{m}}$ entonces

$$||P||_{\mathrm{L}^\infty(-1,1)} = ||T_n^{\mathrm{m}}||_{\mathrm{L}^\infty(-1,1)} \Leftrightarrow P(x) = T_n^{\mathrm{m}}(x),\ x \in [-1,1]. \quad \square$$

Estamos ya en condiciones de aplicar las propiedades que tienen los polinomios de Tchebychev a la acotación del error.

Teorema 6.7. *Dada una función $f \in \mathcal{C}^{n+1}([-1,1])$ la menor cota del error en la norma del máximo en la interpolación de f se obtiene cuando se toman como abscisas de interpolación $\{x_0, x_1, \ldots, x_n\} \subset [-1,1]$ las $n+1$ raíces del polinomio de Tchebychev T_{n+1}, es decir,*

$$x_k = \cos \frac{(2k+1)\pi}{2(n+1)} \tag{6.16}$$

para $k = 0, 1, \ldots, n$. En este caso, se tiene que

$$||E_n||_{\mathrm{L}^\infty(-1,1)} = ||f - P_n||_{\mathrm{L}^\infty(-1,1)} \leq \frac{1}{2^n (n+1)!} \left|\left| f^{n+1)} \right|\right|_{\mathrm{L}^\infty(-1,1)}.$$

Demostración. Como se comentó al comienzo de esta subsección, fijada la función f y el grado del polinomio, es $\Pi_n(x)$ el que permite modificar el valor de la cota de $||E_n||_{\mathrm{L}^\infty(-1,1)}$. Como $\Pi_n(x)$ es un polinomio mónico de grado $n+1$, el menor valor de dicha cota se alcanzará cuando $\Pi_n(x)$ sea el polinomio mónico

con menor norma del máximo en el intervalo $[-1,1]$. Esto ocurre, en virtud del teorema 6.6, cuando

$$\Pi_n(x) = T_{n+1}^{\mathrm{m}}(x) = \frac{T_{n+1}(x)}{2^n}$$

o, lo que es lo mismo, cuando $\{x_0, x_1, \ldots, x_n\}$ son las raíces del polinomio $T_{n+1}(x)$. Para estos valores de $\{x_0, x_1, \ldots, x_n\}$ se tiene que

$$\begin{aligned} |E_n(x)| = |f(x) - P_n(x)| &= \frac{|f^{n+1)}(\xi_x)|}{(n+1)!}|\Pi_n(x)| \\ &= \frac{|f^{n+1)}(\xi_x)|}{2^n(n+1)!}|T_{n+1}(x)| \leq \frac{1}{2^n(n+1)!}\left|\left|f^{n+1)}\right|\right|_{\mathrm{L}^\infty(-1,1)}, \ x \in [-1,1]. \quad \square \end{aligned}$$

Observación 6.11. Nótese que los puntos $\{x_0, x_1, \ldots, x_n\}$ definidos en (6.16) no son equiespaciados, sino que están agrupados cerca de los extremos del intervalo. $\square$

Observación 6.12. Aunque el resultado anterior sólo hace referencia al intervalo $[-1,1]$, puede extenderse a un intervalo arbitrario $[a,b]$. Si $f : [a,b] \to \mathbb{R}$, los puntos de interpolación

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{(2k+1)\pi}{2(n+1)} \tag{6.17}$$

para $k = 0, 1, \ldots, n$ minimizan la cota del error en la interpolación de f. Concretamente, si $f \in \mathcal{C}^{n+1}([a,b])$ se tiene que

$$||E_n||_{\mathrm{L}^\infty(a,b)} = ||f - P_n||_{\mathrm{L}^\infty(a,b)} \leq \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}\left|\left|f^{n+1)}\right|\right|_{\mathrm{L}^\infty(a,b)}$$

(véase el problema 6.15). $\square$

Observación 6.13. Las abscisas (6.17) sirven para minimizar la *cota* genérica del error. No obstante, puede ocurrir que para una función concreta f, existan $n+1$ puntos distintos de forma que el error de interpolación con estos puntos sea inferior al error que se comete considerando dichas abscisas. Por ejemplo, si se considera $f(x) = 2x^3$ y $P \in \mathcal{P}_1$ el polinomio de interpolación asociado a las abscisas de Tchebychev en $[-1,1]$, $x_0 = -\dfrac{\sqrt{2}}{2}$ y $x_1 = \dfrac{\sqrt{2}}{2}$, se tiene que

$$||f - P||_{\mathrm{L}^\infty(-1,1)} = 1;$$

sin embargo, si consideramos $Q \in \mathcal{P}_1$ el polinomio de interpolación de f en los puntos $x_0 = -1$ y $x_1 = 1$, se tiene que

$$||f - Q||_{\mathrm{L}^\infty(-1,1)} = \frac{4}{3\sqrt{3}} \simeq 0.7698 < 1. \quad \square$$

Ejemplo 6.6. Consideremos nuevamente la función del ejemplo 6.3, es decir,

$$f(x) = |x|,\ [-1, 1]$$

tomando ahora como puntos de interpolación los dados por las abscisas de Tchebychev

$$x_k = \cos \frac{(2k+1)\pi}{2(n+1)}$$

para $k = 0, 1, \ldots, n$, donde $n \in \mathbb{N}$. En particular, para los valores $n = 5$, $n = 10$, $n = 15$ y $n = 20$ los polinomios de interpolación de f en $\{x_0, x_1, \ldots, x_n\}$ vienen representados en la figura 6.7. Obsérvese que, a diferencia de lo que ocurría en el ejemplo 6.3, ahora los *efectos de borde* desaparecen. □

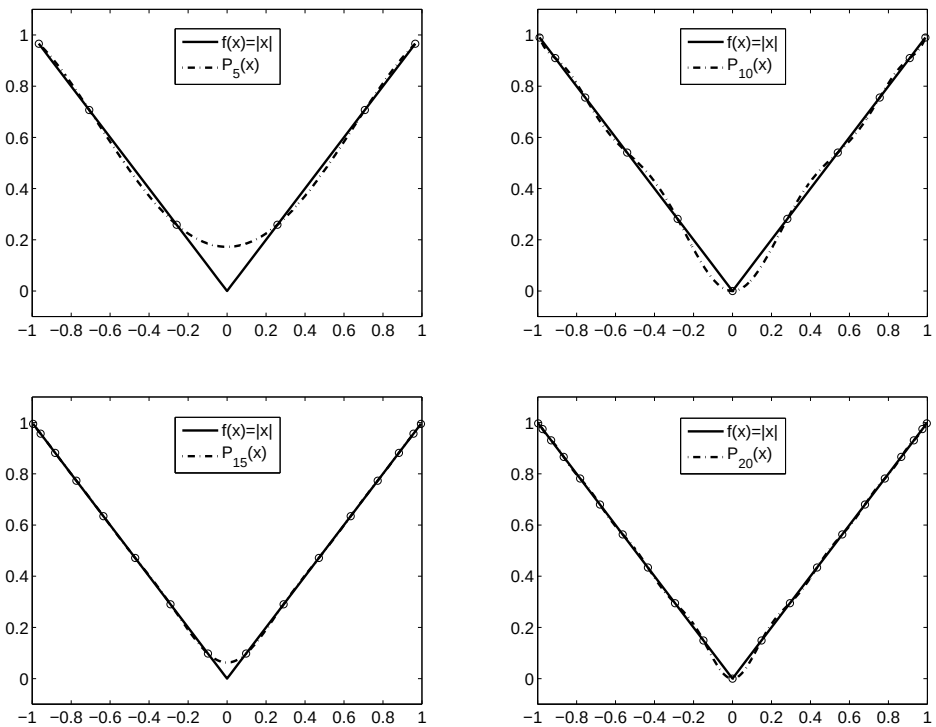


Figura 6.7: Polinomios de interpolación $P_5(x), P_{10}(x), P_{15}(x)$ y $P_{20}(x)$.

6.3. Interpolación mediante funciones spline

La palabra inglesa *spline* denota un instrumento flexible usado en dibujo técnico que sirve para trazar curvas suaves (de hecho, algunas veces se traduce como "trazador"); se trata de una regla que puede ser adaptada, flexionándola, a la forma que tome la curva que se desee dibujar. Precisamente, la propiedad de adaptarse bien a formas dadas que tienen las funciones *spline* es lo que hace que se les dé tal nombre.

Desde el punto de vista matemático, una función *spline* en un intervalo $[a, b]$ está formada por polinomios definidos en subintervalos de $[a, b]$ obedeciendo a ciertas condiciones de regularidad. Concretamente,

Definición 6.4. Sea $\Delta = \{a = x_0 < x_1 < \cdots < x_n = b\}$ una *partición* del intervalo $[a, b]$. $S_\Delta : [a, b] \to \mathbb{R}$ es una *función spline* de *orden* $k \in \mathbb{N}$ asociada a Δ si $S_\Delta \in \mathcal{C}^{k-1}([a, b])$ y S_Δ coincide en cada intervalo $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, n-1$, con un polinomio de grado $\leq k$. □

Observación 6.14.

1. Cuando se toma el valor $k = 1$ se obtiene una función lineal a trozos, es decir, una *poligonal*.

2. En lo que sigue nos restringiremos al caso en que $k = 3$ pues, en la práctica, las *funciones spline cúbicas* son las que se utilizan con mayor frecuencia. □

Notación 6.4. Dado $y = (y_0, y_1, \ldots, y_n) \in \mathbb{R}^{n+1}$ denotaremos por $S_\Delta(y, \cdot)$ a una función *spline* cúbica de interpolación que verifica

$$S_\Delta(y, x_i) = y_i$$

para $i = 0, 1, \ldots, n$. □

Observación 6.15. Estas condiciones de interpolación no determinan de forma única una función *spline* cúbica. Por ejemplo, si tomamos $y = (y_0, y_1, y_2, y_3) \in \mathbb{R}^4$ entonces

$$S_\Delta(y, x) = \begin{cases} a_3x^3 + a_2x^2 + a_1x + a_0, & x \in [x_0, x_1] \\ b_3x^3 + b_2x^2 + b_1x + b_0, & x \in [x_1, x_2] \\ c_3x^3 + c_2x^2 + c_1x + c_0, & x \in [x_2, x_3] \end{cases}$$

con $S_\Delta \in \mathcal{C}^2([x_0, x_3])$ y $S_\Delta(y, x_i) = y_i,\ i = 0, 1, 2, 3$. Por tanto,

$$\begin{cases} a_3x_0^3 + a_2x_0^2 + a_1x_0 + a_0 = y_0 \\ a_3x_1^3 + a_2x_1^2 + a_1x_1 + a_0 = y_1 \\ b_3x_1^3 + b_2x_1^2 + b_1x_1 + b_0 = y_1 \\ b_3x_2^3 + b_2x_2^2 + b_1x_2 + b_0 = y_2 \\ c_3x_2^3 + c_2x_2^2 + c_1x_2 + c_0 = y_2 \\ c_3x_3^3 + c_2x_3^2 + c_1x_3 + c_0 = y_3 \\ 3a_3x_1^2 + 2a_2x_1 + a_1 = 3b_3x_1^2 + 2b_2x_1 + b_1 \\ 3b_3x_2^2 + 2b_2x_2 + b_1 = 3c_3x_2^2 + 2c_2x_2 + c_1 \\ 6a_3x_1 + 2a_2 = 6b_3x_1 + 2b_2 \\ 6b_3x_2 + 2b_2 = 6c_3x_1 + 2c_2. \end{cases}$$

Puede demostrarse que las 10 ecuaciones anteriores son independientes entre sí; como tenemos 12 incógnitas, el sistema anterior tiene dos *grados de libertad*. En el caso general en que $y = (y_0, y_1, \ldots, y_n) \in \mathbb{R}^{n+1}$, se obtiene un sistema de $4n - 2$ ecuaciones ($2n$ debidas a los valores que interpola S_Δ, $n - 1$ a la continuidad de S'_Δ y $n - 1$ a la continuidad de S''_Δ) con $4n$ incógnitas. □

Normalmente, para determinar unívocamente las funciones *spline* se impone uno de los siguientes tipos de condiciones:

a) Tipo I: $S''_\Delta(y, a) = S''_\Delta(y, b) = 0$. Se suelen denominar condiciones *naturales*.

b) Tipo II: $S'_\Delta(y, a) = y'_0,\ S'_\Delta(y, b) = y'_n$ siendo $y'_0,\ y'_n \in \mathbb{R}$ valores prefijados.

c) Tipo III: $S^{k)}_\Delta(y, a) = S^{k)}_\Delta(y, b),\ k = 0, 1, 2$.

Observación 6.16. Las condiciones de tipo III se utilizan cuando se precisa interpolar una función periódica de periodo $b - a$. Obviamente se tendrá $y_0 = y_n$. Por ello, estas condiciones se suelen denominar *periódicas*. □

6.3.1. Método de cálculo de las funciones spline cúbicas

Vamos a caracterizar la función $S_\Delta(y, \cdot)$ a través de lo que denominaremos sus *momentos*

$$M_j = S''_\Delta(y, x_j)$$

para $j = 0, 1, \ldots, n$. De hecho, vamos a probar que, conocidos los momentos de una función *spline* cúbica, ésta viene determinada unívocamente a partir de ellos. Posteriormente, veremos que los momentos están también unívocamente determinados por los datos del problema. De esta forma, habremos probado la existencia

y unicidad de la función *spline* cúbica interpoladora con condiciones de alguno de los tres tipos anteriores. Además, el proceso será constructivo y proporcionará un algoritmo de cálculo de las funciones *spline* cúbicas.

Para cada $j \in \{0, 1, \ldots, n-1\}$ vamos a denotar

$$h_{j+1} = x_{j+1} - x_j.$$

Como $S_\Delta(y, \cdot)$ coincide en cada intervalo $[x_j, x_{j+1}]$ con un polinomio de grado ≤ 3, la función $S''_\Delta(y, \cdot)$ coincide en cada intervalo $[x_j, x_{j+1}]$ con una función lineal que puede ser descrita en términos de los momentos de $S_\Delta(y, \cdot)$ en los valores extremos del intervalo:

$$S''_\Delta(y, x) = M_j \frac{x_{j+1} - x}{h_{j+1}} + M_{j+1} \frac{x - x_j}{h_{j+1}},\ x \in [x_j, x_{j+1}]. \tag{6.18}$$

Integrando esta igualdad, para cada $j = 0, 1, \ldots, n-1$, se obtiene

$$S'_\Delta(y, x) = -M_j \frac{(x_{j+1} - x)^2}{2h_{j+1}} + M_{j+1} \frac{(x - x_j)^2}{2h_{j+1}} + A_j,\ x \in [x_j, x_{j+1}]. \tag{6.19}$$

Integrando nuevamente obtenemos

$$S_\Delta(y, x) = M_j \frac{(x_{j+1} - x)^3}{6h_{j+1}} + M_{j+1} \frac{(x - x_j)^3}{6h_{j+1}} + A_j(x - x_j) + B_j,\ x \in [x_j, x_{j+1}].$$

Determinemos las constantes A_j y B_j para $j = 0, 1, \ldots, n-1$. Como

$$S_\Delta(y, x_j) = y_j$$

entonces

$$\begin{cases} y_j = M_j \dfrac{(x_{j+1} - x_j)^3}{6h_{j+1}} + B_j = M_j \dfrac{h_{j+1}^2}{6} + B_j \\ y_{j+1} = M_{j+1} \dfrac{(x_{j+1} - x_j)^3}{6h_{j+1}} + A_j(x_{j+1} - x_j) + B_j = M_{j+1} \dfrac{h_{j+1}^2}{6} + A_j h_{j+1} + B_j, \end{cases}$$

de donde

$$B_j = y_j - M_j \frac{h_{j+1}^2}{6}$$

y

$$\begin{aligned} A_j &= \frac{1}{h_{j+1}} \left(y_{j+1} - B_j - M_{j+1} \frac{h_{j+1}^2}{6} \right) \\ &= \frac{1}{h_{j+1}} \left(y_{j+1} - y_j + M_j \frac{h_{j+1}^2}{6} - M_{j+1} \frac{h_{j+1}^2}{6} \right) \\ &= \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6}(M_{j+1} - M_j). \end{aligned}$$

Así pues, para cada $j \in \{0, 1, \ldots, n-1\}$, podemos escribir

$$S_\Delta(y, x) = \alpha_j + \beta_j(x - x_j) + \gamma_j(x - x_j)^2 + \delta_j(x - x_j)^3,\ x \in [x_j, x_{j+1}]$$

donde, utilizando (6.18) y (6.19), estos coeficientes vienen dados por

$$\begin{cases} \alpha_j = S_\Delta(y, x_j) = y_j \\ \beta_j = S'_\Delta(y, x_j) = -M_j \dfrac{h_{j+1}}{2} + A_j = \dfrac{y_{j+1} - y_j}{h_{j+1}} - \dfrac{2M_j + M_{j+1}}{6} h_{j+1} \\ \gamma_j = \dfrac{S''_\Delta(y, x_j)}{2!} = \dfrac{M_j}{2} \\ \delta_j = \dfrac{S'''_\Delta(y, x_j^+)}{3!} = \dfrac{M_{j+1} - M_j}{6h_{j+1}}. \end{cases}$$

Es decir, para cada $j \in \{0, 1, \ldots, n-1\}$, se tiene que

$$\begin{aligned} S_\Delta(y, x) = y_j + \left(\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{2M_j + M_{j+1}}{6} h_{j+1} \right)(x - x_j) \\ + \frac{M_j}{2}(x - x_j)^2 + \frac{M_{j+1} - M_j}{6h_{j+1}}(x - x_j)^3 \end{aligned}$$

para todo $x \in [x_j, x_{j+1}]$.

La fórmula anterior determina unívocamente la función $S_\Delta(y, \cdot)$ siempre y cuando se conozcan sus momentos. Veamos a continuación cómo efectuar el cálculo de los mismos.

Como la función $S_\Delta(y, \cdot) \in \mathcal{C}^2([a, b])$ entonces, en particular, verifica

$$S'_\Delta(y, x_j^-) = S'_\Delta(y, x_j^+) \tag{6.20}$$

para $j = 1, 2, \ldots, n-1$. Usando la relación (6.19) con el valor de A_j ya calculado, para cada índice $j \in \{1, 2, \ldots, n-1\}$, se verifica que

$$\begin{aligned} S'_\Delta(y, x_j^-) &= M_j \frac{h_j}{2} + \frac{y_j - y_{j-1}}{h_j} - \frac{h_j}{6}(M_j - M_{j-1}) \\ &= \frac{y_j - y_{j-1}}{h_j} + \frac{h_j}{3} M_j + \frac{h_j}{6} M_{j-1} \end{aligned}$$

y

$$\begin{aligned} S'_\Delta(y, x_j^+) &= -M_j \frac{h_{j+1}}{2} + \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6}(M_{j+1} - M_j) \\ &= \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{3} M_j - \frac{h_{j+1}}{6} M_{j+1}. \end{aligned}$$

De esta forma, para $j = 1, 2, \ldots, n-1$, la relación (6.20) determina que

$$\boxed{\frac{h_j}{6}M_{j-1} + \frac{h_j + h_{j+1}}{3}M_j + \frac{h_{j+1}}{6}M_{j+1} = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j}} \tag{6.21}$$

Introduciendo la notación

$$\lambda_j = \frac{h_{j+1}}{h_j + h_{j+1}},\ \mu_j = 1 - \lambda_j = \frac{h_j}{h_j + h_{j+1}}$$

y

$$d_j = \frac{6}{h_j + h_{j+1}}\left(\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j}\right),$$

para $j = 1, 2, \ldots, n-1$, y multiplicando por $\dfrac{6}{h_j + h_{j+1}}$, la igualdad (6.21) se transforma en

$$\mu_j M_{j-1} + 2M_j + \lambda_j M_{j+1} = d_j \tag{6.22}$$

para $j = 1, 2, \ldots, n-1$. Para ver lo que ocurre en los extremos (correspondientes a $j = 0$ y $j = n$) y, de esta forma, obtener otras dos ecuaciones, tengamos en cuenta los diversos tipos de condiciones:

a) Tipo I: $S''_\Delta(y, a) = S''_\Delta(y, b) = 0$. En este caso:

$$M_0 = S''_\Delta(y, a) = 0 = S''_\Delta(y, b) = M_n \ \Rightarrow\ \boxed{M_0 = M_n = 0}$$

Considerando

$$\lambda_0 = d_0 = \mu_n = d_n = 0$$

se verifica, obviamente,

$$\begin{cases} 2M_0 + \lambda_0 M_1 = d_0 \\ \mu_n M_{n-1} + 2M_n = d_n \end{cases} \tag{6.23}$$

b) Tipo II: $S'_\Delta(y, a) = y'_0$, $S'_\Delta(y, b) = y'_n$. En este caso,

$$\begin{aligned} y'_0 = S'_\Delta(y, a) &= -M_0\frac{h_1}{2} + \frac{y_1 - y_0}{h_1} - \frac{h_1}{6}(M_1 - M_0) \\ &= -M_0\frac{h_1}{3} + \frac{y_1 - y_0}{h_1} - M_1\frac{h_1}{6} \end{aligned}$$

e

$$\begin{aligned} y'_n = S'_\Delta(y, b) &= M_n\frac{h_n}{2} + \frac{y_n - y_{n-1}}{h_n} - \frac{h_n}{6}(M_n - M_{n-1}) \\ &= M_n\frac{h_n}{3} + \frac{y_n - y_{n-1}}{h_n} + M_{n-1}\frac{h_n}{6}, \end{aligned}$$

es decir,

$$\boxed{\begin{array}{c}\dfrac{h_1}{3}M_0+\dfrac{h_1}{6}M_1=\dfrac{y_1-y_0}{h_1}-y_0'\\ \dfrac{h_n}{6}M_{n-1}+\dfrac{h_n}{3}M_n=y_n'-\dfrac{y_n-y_{n-1}}{h_n}\end{array}}$$

Multiplicando la primera expresión por $\dfrac{6}{h_1}$ y la segunda por $\dfrac{6}{h_n}$ y llamando

$$\lambda_0=\mu_n=1,\ d_0=\frac{6}{h_1}\left(\frac{y_1-y_0}{h_1}-y_0'\right)\ \text{ y }\ d_n=\frac{6}{h_n}\left(y_n'-\frac{y_n-y_{n-1}}{h_n}\right)$$

volvemos a obtener la relación expresada en (6.23).

c) Tipo III: $S_\Delta^{k)}(y,a)=S_\Delta^{k)}(y,b),\ k=0,1,2$. Las condiciones de tipo periódico hacen que $y_0=y_n$. De la propia definición se verifica que

$$\boxed{M_0=M_n}$$

y, de la relación $S'_\Delta(y,a)=S'_\Delta(y,b)$, se obtiene

$$-M_0\frac{h_1}{2}+\frac{y_1-y_0}{h_1}-\frac{h_1}{6}(M_1-M_0)=M_n\frac{h_n}{2}+\frac{y_n-y_{n-1}}{h_n}-\frac{h_n}{6}(M_n-M_{n-1});$$

Como $y_0=y_n$ y $M_0=M_n$ entonces

$$-M_n\frac{h_1}{2}+\frac{y_1-y_n}{h_1}-\frac{h_1}{6}(M_1-M_n)=M_n\frac{h_n}{2}+\frac{y_n-y_{n-1}}{h_n}-\frac{h_n}{6}(M_n-M_{n-1}),$$

es decir,

$$\boxed{\frac{h_n}{6}M_{n-1}+\frac{h_1+h_n}{3}M_n+\frac{h_1}{6}M_1=\frac{y_1-y_n}{h_1}-\frac{y_n-y_{n-1}}{h_n}}$$

Si multiplicamos la igualdad anterior por $\dfrac{6}{h_1+h_n}$ y denotamos

$$\lambda_n=\frac{h_1}{h_1+h_n},\ \mu_n=1-\lambda_n=\frac{h_n}{h_1+h_n}$$

y

$$d_n=\frac{6}{h_1+h_n}\left(\frac{y_1-y_n}{h_1}-\frac{y_n-y_{n-1}}{h_n}\right),$$

obtenemos

$$\mu_nM_{n-1}+2M_n+\lambda_nM_1=d_n.$$

Por otra parte, el hecho de que $M_0 = M_n$ hace que podamos escribir la ecuación

$$\mu_1 M_0 + 2M_1 + \lambda_1 M_2 = d_1$$

como

$$2M_1 + \lambda_1 M_2 + \mu_1 M_n = d_1.$$

De esta forma, con las notaciones anteriores, para las funciones *spline* de tipo I y tipo II se obtiene el sistema lineal $AM = d$ donde

$$A = \begin{pmatrix} 2 & \lambda_0 & & & & \\ \mu_1 & 2 & \lambda_1 & & & \\ & \mu_2 & 2 & \lambda_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mu_{n-1} & 2 & \lambda_{n-1} \\ & & & & \mu_n & 2 \end{pmatrix}, \; M = \begin{pmatrix} M_0 \\ M_1 \\ \cdots \\ M_n \end{pmatrix} \text{ y } d = \begin{pmatrix} d_0 \\ d_1 \\ \cdots \\ d_n \end{pmatrix}.$$

Para las funciones de tipo III (caso periódico) se llega al sistema $A_p M_p = d_p$ siendo

$$A_p = \begin{pmatrix} 2 & \lambda_1 & & & & \mu_1 \\ \mu_2 & 2 & \lambda_2 & & & \\ & \mu_3 & 2 & \lambda_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mu_{n-1} & 2 & \lambda_{n-1} \\ \lambda_n & & & & \mu_n & 2 \end{pmatrix}, \; M_p = \begin{pmatrix} M_1 \\ M_2 \\ \cdots \\ M_n \end{pmatrix} \text{ y } d_p = \begin{pmatrix} d_1 \\ d_2 \\ \cdots \\ d_n \end{pmatrix}.$$

Observación 6.17.

1. Nótese que los valores λ_i y μ_i verifican que $\lambda_i \geq 0$, $\mu_i \geq 0$, $\lambda_i + \mu_i = 1$ y son independientes del valor $y = (y_0, y_1, \ldots, y_n) \in \mathbb{R}^{n+1}$ (sólo dependen de la partición Δ). En consecuencia, si necesitamos interpolar otra función en la misma partición, basta que calculemos el nuevo vector d y resolvamos otro sistema lineal con la misma matriz ya calculada.

2. Cuando los puntos de la partición Δ son equiespaciados entonces

$$\lambda_i = \mu_i = \frac{1}{2} \text{ y } d_i = 3\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}$$

para todo $i \in \{1, 2, \ldots, n-1\}$. □

Teorema 6.8. *Los sistemas $AM = d$ y $A_p M_p = d_p$ admiten una única solución para cada partición Δ de $[a, b]$.*

Demostración. Basta tener en cuenta que las matrices A y A_p son de diagonal estrictamente dominante, por lo que son inversibles. □

La existencia y unicidad de la función *spline* cúbica interpoladora se deducen del teorema 6.8 y del hecho de que tal función está unívocamente determinada a partir de sus momentos.

Corolario 6.4 (Existencia y unicidad). *Sea* Δ *una partición del intervalo* $[a,b]$ *de la forma* $\Delta = \{a = x_0 < x_1 < \cdots < x_n = b\}$, $y = (y_0, y_1, \ldots, y_n) \in \mathbb{R}^{n+1}$ *y* $f : [a,b] \to \mathbb{R}$ *tal que*

$$f(x_i) = y_i$$

para $i = 0, 1, \ldots, n$. *Toda función* spline *cúbica de interpolación* $S_\Delta(y,\cdot)$ *que verifique una de las tres condiciones siguientes:*

a) $S''_\Delta(y,a) = S''_\Delta(y,b) = 0$.

b) $S'_\Delta(y,a) = f'(a)$, $S'_\Delta(y,b) = f'(b)$.

c) $S^{k)}_\Delta(y,a) = S^{k)}_\Delta(y,b)$, $k = 0,1,2$.

está unívocamente determinada. □

Ejemplo 6.7. Vamos a determinar la función *spline* cúbica de tipo I que interpola la siguiente tabla

x_i	0	1	3
$f(x_i)$	1	2	0

(6.24)

En este caso

$$\lambda_0 = d_0 = \mu_2 = d_2 = 0,\ \lambda_1 = \frac{h_2}{h_1 + h_2} = \frac{2}{3},\ \mu_1 = 1 - \lambda_1 = \frac{1}{3}$$

y

$$d_1 = \frac{6}{h_1 + h_2}\left(\frac{y_2 - y_1}{h_2} - \frac{y_1 - y_0}{h_1}\right) = \frac{6}{3}\left(\frac{-2}{2} - \frac{1}{1}\right) = -4.$$

Por tanto, el sistema que obtenemos es

$$\begin{pmatrix} 2 & 0 & 0 \\ \frac{1}{3} & 2 & \frac{2}{3} \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} M_0 \\ M_1 \\ M_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -4 \\ 0 \end{pmatrix},$$

que tiene como solución $M_0 = M_2 = 0$ y $M_1 = -2$. La caracterización de $S_\Delta(y,\cdot)$ por sus momentos

$$S_\Delta(y,x) = y_j + \left(\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{2M_j + M_{j+1}}{6}h_{j+1}\right)(x - x_j) + \frac{M_j}{2}(x - x_j)^2 + \frac{M_{j+1} - M_j}{6h_{j+1}}(x - x_j)^3,\ x \in [x_j, x_{j+1}],\ j = 0,1$$

donde $\{x_0, x_1, x_2\} = \{0, 1, 3\}$, hace que (teniendo en cuenta que $M_0 = M_2 = 0$)

$$\begin{aligned} S_\Delta(y, x) &= y_0 + \left(\frac{y_1 - y_0}{h_1} - \frac{M_1}{6}h_1\right)(x - x_0) + \frac{M_1}{6h_1}(x - x_0)^3 \\ &= 1 + \left(\frac{1}{1} - \frac{-2}{6}\right)x + \frac{-2}{6}x^3 = 1 + \frac{4}{3}x - \frac{1}{3}x^3,\ x \in [0, 1] \end{aligned}$$

y

$$\begin{aligned} S_\Delta(y, x) &= y_1 + \left(\frac{y_2 - y_1}{h_2} - \frac{2M_1}{6}h_2\right)(x - x_1) + \frac{M_1}{2}(x - x_1)^2 - \frac{M_1}{6h_2}(x - x_1)^3 \\ &= 2 + \left(\frac{-2}{2} - \frac{-4}{6}2\right)(x - 1) + \frac{-2}{2}(x - 1)^2 - \frac{-2}{12}(x - 1)^3 \\ &= 2 + \frac{1}{3}(x - 1) - (x - 1)^2 + \frac{1}{6}(x - 1)^3,\ x \in [1, 3]. \end{aligned}$$

Es decir, la función *spline* de interpolación buscada es

$$S_\Delta(y, x) = \begin{cases} 1 + \dfrac{4}{3}x - \dfrac{1}{3}x^3, & x \in [0, 1] \\ 2 + \dfrac{1}{3}(x - 1) - (x - 1)^2 + \dfrac{1}{6}(x - 1)^3, & x \in [1, 3] \end{cases}$$

y viene representada en la figura 6.8. □

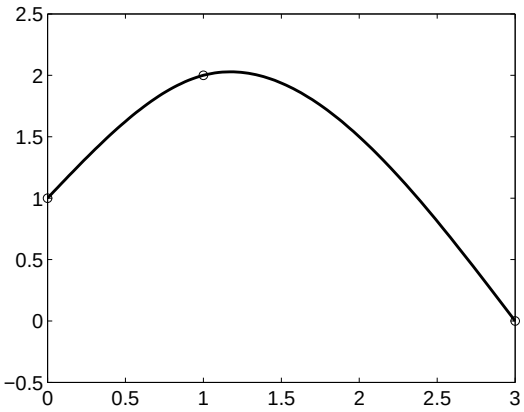


Figura 6.8: Función *spline* de tipo I que interpola la tabla (6.24).

Ejemplo 6.8. Vamos a hallar la función *spline* cúbica de tipo II que interpola la siguiente tabla

x_i	-1	0	1	3
$f(x_i)$	-4	-1	0	20
$f'(x_i)$	6	—	—	22

(6.25)

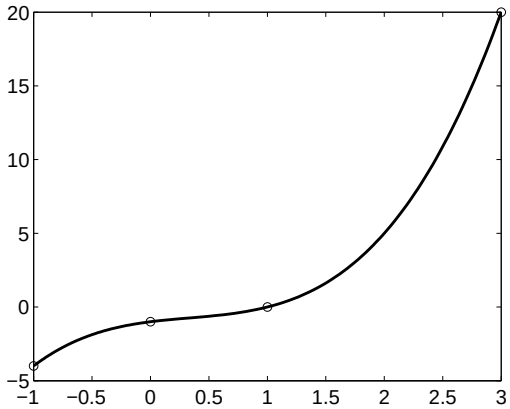


Figura 6.9: Función *spline* de tipo II que interpola la tabla (6.25).

Compruébese que, en este caso, los momentos son

$$M_0 = -8,\ M_1 = -2,\ M_2 = 4 \text{ y } M_3 = 16$$

y que determinan la función

$$S_\Delta(y, x) = x^3 - x^2 + x - 1,\ x \in [-1, 3].$$

La gráfica de $S_\Delta(y, \cdot)$ viene dada en la figura 6.9. Nótese que la función *spline* obtenida es el mismo polinomio de grado 3 en todo el intervalo $[-1, 3]$. □

6.3.2. Convergencia en la interpolación por funciones spline

Hemos visto que en la interpolación de Lagrange los polinomios de interpolación no convergen a la función al tomar particiones arbitrariamente finas. En cambio, las funciones *spline* convergen uniformemente a la función interpolada bajo hipótesis bastante generales.

Definición 6.5. Sea $\Delta = \{a = x_0 < x_1 < \cdots < x_n = b\}$ una partición del intervalo $[a, b]$. Se denomina *diámetro* de Δ a la cantidad

$$\varrho(\Delta) = \max_{0 \le j \le n-1} |x_{j+1} - x_j|. \quad \square$$

Observación 6.18. Si los puntos de la partición Δ son equiespaciados se verifica que $\varrho(\Delta) = h$. □

Teorema 6.9. *Sean* $\Delta = \{a = x_0 < x_1 < \cdots < x_n = b\}$ *una partición de* $[a, b]$ *de diámetro* $\varrho(\Delta)$ *con la propiedad*

$$\frac{\varrho(\Delta)}{|x_{j+1} - x_j|} \leq \mathrm{K}$$

para $j = 0, 1, \ldots, n-1$, $f \in \mathcal{C}^4([a, b])$ *verificando*

$$|f^{iv)}(x)| \leq \mathrm{L},\ x \in [a, b],$$

y S_Δ *la función spline cúbica que interpola a* f *en los puntos* $\{x_0, x_1, \ldots, x_n\}$ *con la propiedad*

$$S'_\Delta(y, a) = f'(a) \ \ y \ \ S'_\Delta(y, b) = f'(b).$$

Existen constantes $0 < c_i \leq 2$, $i = 0, 1, 2, 3$ (*independientes de* Δ) *tales que para todo* $x \in [a, b]$ *se verifica que*

$$\left| f^{i)}(x) - S^{i)}_\Delta(x) \right| \leq c_i \mathrm{LK}(\varrho(\Delta))^{4-i}$$

para $i = 0, 1, 2, 3$.

Demostración. Véase [**St–Bu**]. □

Observación 6.19.

1. K mide la *uniformidad* de la partición.
2. En particular, si los puntos de la partición están equiespaciados, se puede tomar $\mathrm{K} = 1$, obteniéndose

   $$\left| f^{i)}(x) - S^{i)}_\Delta(x) \right| \leq c_i \mathrm{L} h^{4-i}$$

   para todo $x \in [a, b]$ e $i = 0, 1, 2, 3$.
3. Si $\Delta_k = \{a = x_0^{(k)} < x_1^{(k)} < \cdots < x_n^{(k)} = b\}$, $k \in \mathbb{N}$, son particiones del intervalo $[a, b]$ verificando

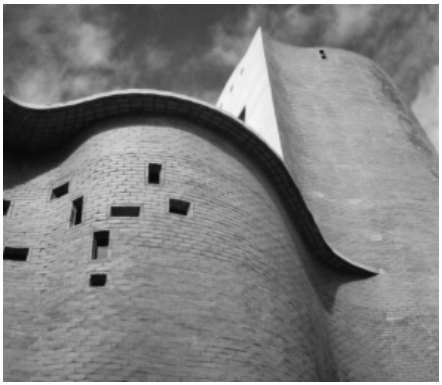
   $$\lim_{k \to +\infty} \varrho(\Delta_k) = 0 \ \text{ y } \ \sup_k \left\{ \max_j \frac{\varrho(\Delta_k)}{|x_{j+1}^{(k)} - x_j^{(k)}|} \right\} \leq \mathrm{K},$$

   entonces las correspondientes funciones *spline* S_{Δ_k} que interpolan a f en los puntos de Δ_k y tales que

   $$S'_{\Delta_k}(a) = f'(a) \ \text{ y } \ S'_{\Delta_k}(b) = f'(b)$$

   convergen uniformemente a f en $[a, b]$. También sus tres primeras derivadas convergen uniformemente a las derivadas respectivas de f. □

A modo de ejemplo, a partir de la fotografía de la iglesia del arquitecto uruguayo Eladio Dieste, utilizando tres funciones *spline* de tipo II que interpolan la tabla 6.2, se obtiene la aproximación de la cubierta dada en la figura 6.10.



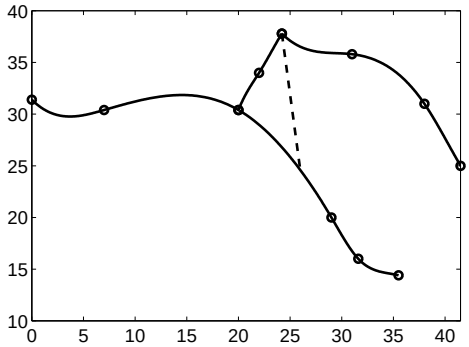


Figura 6.10: Iglesia de Eladio Dieste y cubierta aproximada.

TABLA 6.2:
Puntos de interpolación para la cubierta de la figura 6.10

x_i	y_i		x_i	y_i		x_i	y_i
0.0	31.4		20.0	30.4		24.2	37.8
7.0	30.4		22.0	34.0		31.0	35.8
20.0	30.4		24.2	37.8		38.0	31.0
29.0	20.0					41.5	25.0
31.6	16.0						
35.5	14.4						
$f'(0.0)$	-1.00		$f'(20.0)$	2.50		$f'(24.2)$	-1.00
$f'(35.5)$	-0.25		$f'(24.20)$	1.66		$f'(41.5)$	-1.71

6.4. Problemas

6.4.1. Problemas resueltos

6.1. Dada la función

$$f(x) = \cos \pi x,$$

hallar el polinomio P que interpola a f en los puntos $\{0, 0.5, 1, 1.5\}$.

Solución. Tenemos que interpolar la tabla

x_i	0	0.5	1	1.5
$f(x_i)$	1	0	-1	0

(6.26)

Por la fórmula de interpolación de Newton para diferencias finitas con paso $h = 0.5$ (véase la observación 6.5), el polinomio de interpolación de la función f en los puntos $\{x_0, x_1, x_2, x_3\}$ viene dado por

$$P(x) = f(x_0) + (x - x_0)\frac{\Delta f(x_0)}{h} + (x - x_0)(x - x_1)\frac{\Delta^2 f(x_0)}{2!h^2} + (x - x_0)(x - x_1)(x - x_2)\frac{\Delta^3 f(x_0)}{3!h^3}.$$

La tabla de diferencias finitas para f es

$f(x_0) = 1$	$\Delta f(x_0) = -1$	$\Delta^2 f(x_0) = 0$	$\Delta^3 f(x_0) = 2$
$f(x_1) = 0$	$\Delta f(x_1) = -1$	$\Delta^2 f(x_1) = 2$	
$f(x_2) = -1$	$\Delta f(x_2) = 1$		
$f(x_3) = 0$			

por lo que el polinomio de interpolación buscado es

$$P(x) = 1 - 2x + \frac{8}{3}x\left(x - \frac{1}{2}\right)(x - 1) = \frac{8}{3}x^3 - 4x^2 - \frac{2}{3}x + 1$$

y viene representado en la figura 6.11. □

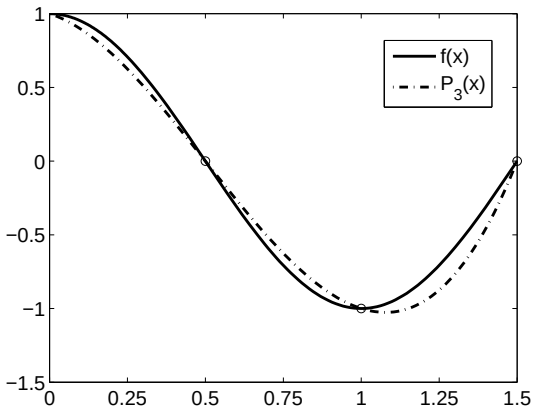


Figura 6.11: Función $f(x) = \cos \pi x$ y polinomio de interpolación de la tabla (6.26).

6.2. Sean $f, g : [a, b] \to \mathbb{R}$ y $\{x_0, x_1, \ldots, x_n\}$ puntos distintos del intervalo $[a, b]$. Si P y Q son, respectivamente, los polinomios de interpolación de f y g en los puntos $\{x_0, x_1, \ldots, x_n\}$:

a) ¿Es $\alpha P + \beta Q$ ($\alpha, \beta \in \mathbb{R}$) el polinomio de interpolación de $\alpha f + \beta g$ en los puntos $\{x_0, x_1, \ldots, x_n\}$?

b) ¿Es PQ el polinomio de interpolación de fg en los puntos $\{x_0, x_1, \ldots, x_n\}$?

SOLUCIÓN.

a) Claramente $\alpha P + \beta Q \in \mathcal{P}_n$ y para cada $i = 0, 1, \ldots, n$ se verifica

$$\begin{aligned}(\alpha P + \beta Q)(x_i) &= \alpha P(x_i) + \beta Q(x_i) \\ &= \alpha f(x_i) + \beta g(x_i) = (\alpha f + \beta g)(x_i),\end{aligned}$$

por lo que, por la unicidad del polinomio de interpolación, $\alpha P + \beta Q$ es el polinomio de interpolación de $\alpha f + \beta g$ en los puntos $\{x_0, x_1, \ldots, x_n\}$.

b) La respuesta es, en general, negativa puesto que, ahora, $PQ \in \mathcal{P}_{2n}$. Así, por ejemplo, si consideramos las funciones

$$f(x) = g(x) = x,\ x \in [0, 1]$$

se verifica que $P(x) = Q(x) = x$ es el polinomio de interpolación de las funciones f y g en los puntos $\{x_0 = 0, x_1 = 1\}$; en cambio $(PQ)(x) = x^2$ no es el polinomio de interpolación de la función $(fg)(x) = x^2$ en los puntos $\{x_0 = 0, x_1 = 1\}$ dado que éste es $P_1(x) = x$. □

6.3. Sean $\{x_0, x_1, \ldots, x_n\} \subset \mathbb{R}$ con $x_i \neq x_j$ si $i \neq j$ y $f(x) = x^{n+1}$. Calcular el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_n\}$ utilizando la fórmula del error y determinar el término independiente de dicho polinomio.

SOLUCIÓN. Aplicando la fórmula del error en la interpolación de Lagrange se verifica que

$$x^{n+1} - P_n(x) = \frac{f^{n+1)}(\xi)}{(n+1)!}\Pi_n(x) = \Pi_n(x)$$

(véase (6.5)), ya que, en este caso

$$f^{n+1)}(x) = (n+1)!$$

De la expresión anterior se deduce que

$$P_n(x) = x^{n+1} - \Pi_n(x)$$

es el polinomio de interpolación de la función f en los puntos $\{x_0, x_1, \ldots, x_n\}$. Nótese que $P_n \in \mathcal{P}_n$ (ya que Π_n es un polinomio mónico de grado $n+1$) y el término independiente de P_n es

$$P_n(0) = -\Pi_n(0) = -\prod_{i=0}^{n}(-x_i) = (-1)^n x_0 x_1 \cdots x_n. \quad \square$$

6.4. Sean $\{x_0, x_1, \ldots, x_n\} \subset \mathbb{R}$ puntos distintos y $c_i = L_i(0)$, $i = 0, 1, \ldots, n$ siendo $\{L_0(x), L_1(x), \ldots, L_n(x)\}$ los polinomios básicos de Lagrange. Demostrar que

$$\sum_{i=0}^{n} c_i x_i^j = \begin{cases} 1 & \text{si} \quad j = 0 \\ 0 & \text{si} \quad j = 1, 2, \ldots, n \\ (-1)^n x_0 x_1 \cdots x_n & \text{si} \quad j = n+1 \end{cases}$$

y

$$\sum_{i=0}^{n} L_i(x) = 1, \; x \in \mathbb{R}. \tag{6.27}$$

Concluir que si P_n es el polinomio de interpolación de una función $f : [a, b] \to \mathbb{R}$ en los puntos $\{x_0, x_1, \ldots, x_n\}$ entonces se verifica que

$$E_n(x) = f(x) - P_n(x) = \sum_{i=0}^{n} \left(f(x) - f(x_i)\right) L_i(x), \; x \in [a, b]. \tag{6.28}$$

Solución. Por la fórmula de interpolación de Lagrange (véase el teorema 6.1) se tiene que

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x) \tag{6.29}$$

es el polinomio de interpolación de la función f en los puntos $\{x_0, x_1, \ldots, x_n\}$. En particular, si hacemos $x = 0$ en la expresión anterior obtenemos que

$$\sum_{i=0}^{n} c_i f(x_i) = P_n(0).$$

Observamos que el polinomio de interpolación de un polinomio de grado menor o igual que n en $n+1$ puntos distintos es, por unicidad, el propio polinomio. De esta forma, para:

a) $j = 0$, tomando $f(x) \equiv 1$ se tiene que $P_n(x) \equiv 1$ y, en particular, $P_n(0) = 1$, es decir,

$$\sum_{i=0}^{n} c_i = 1.$$

b) $1 \leq j \leq n$, tomando $f(x) = x^j$ se tiene que $P_n(x) = x^j$ y, en particular, $P_n(0) = 0$, por lo que

$$\sum_{i=0}^{n} c_i x_i^j = 0.$$

c) $j = n+1$, tomando ahora $f(x) = x^{n+1}$ y aplicando el problema 6.3, se obtiene que

$$\sum_{i=0}^{n} c_i x_i^{n+1} = (-1)^n x_0 x_1 \cdots x_n.$$

Por otra parte, si en (6.29) tomamos $f(x) \equiv 1$, como $P_n(x) \equiv 1$, entonces

$$\sum_{i=0}^{n} L_i(x) = 1.$$

La propiedad (6.28) se sigue de forma inmediata a partir de (6.27) y (6.29). □

6.5. Demostrar que los polinomios básicos de interpolación de Lagrange pueden ser expresados en la forma

$$L_i(x) = \frac{\Pi_n(x)}{(x - x_i)\Pi_n'(x_i)}$$

para $i = 0, 1, \ldots, n$.

SOLUCIÓN. Como

$$\Pi_n'(x) = \sum_{i=0}^{n} \left(\prod_{\substack{j=0 \\ j \neq i}}^{n} (x - x_j) \right)$$

entonces, para cada $i \in \{0, 1, \ldots, n\}$, se tiene que

$$\Pi_n'(x_i) = \prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j) \tag{6.30}$$

y, de esta forma,

$$\frac{\Pi_n(x)}{(x - x_i)\Pi_n'(x_i)} = \frac{\displaystyle\prod_{j=0}^{n} (x - x_j)}{(x - x_i) \displaystyle\prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j)} = \frac{\displaystyle\prod_{\substack{j=0 \\ j \neq i}}^{n} (x - x_j)}{\displaystyle\prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j)} = L_i(x). \quad \square$$

6.6. Sean $a > 0$, $\{-x_n, -x_{n-1}, \ldots, -x_1, 0, x_1, \ldots, x_n\} \subset [-a, a]$ y P_{2n} el polinomio de interpolación de una función $f : [-a, a] \to \mathbb{R}$ en los puntos anteriores. Demostrar los siguientes resultados:

a) Si f es una función par (respectivamente, impar) entonces P_{2n} es par (respectivamente, impar).

b) Si f es una función par existe $Q_n \in \mathcal{P}_n$ tal que $P_{2n}(x) = Q_n(x^2)$. ¿Quién es Q_n? ¿Qué utilidad tiene esto?

Solución.

a) Trataremos únicamente el caso en que f es una función par (el caso impar se aborda de forma análoga). Consideremos el polinomio

$$P(x) = P_{2n}(-x).$$

Claramente $P \in \mathcal{P}_{2n}$ y verifica, por ser f par,

$$P(x_i) = P_{2n}(-x_i) = f(-x_i) = f(x_i),\ i = 1, 2, \ldots, n$$

$$P(0) = P_{2n}(0) = f(0)$$

y

$$P(-x_i) = P_{2n}(x_i) = f(x_i) = f(-x_i),\ i = 1, 2, \ldots, n.$$

Consecuentemente, $P_{2n}(-x)$ es el polinomio de interpolación de f en los puntos

$$\{-x_n, -x_{n-1}, \ldots, -x_1, 0, x_1, \ldots, x_{n-1}, x_n\}$$

y, por unicidad,

$$P_{2n}(x) = P_{2n}(-x)$$

por lo que se verifica que P_{2n} es una función par.

b) Al ser f una función par sabemos, por el apartado anterior, que P_{2n} es también una función par que puede factorizarse, por tanto, en la forma

$$P_{2n}(x) = a_n \prod_{i=1}^{n} (x^2 - \alpha_i^2)$$

(téngase en cuenta que si un número α es raíz de un polinomio par también es raíz el opuesto $-\alpha$). De esta forma

$$P_{2n}(x) = Q_n(x^2)$$

siendo

$$Q_n(x) = a_n \prod_{i=1}^{n} (x - \alpha_i^2) \in \mathcal{P}_n.$$

Además, como

$$Q_n(0) = P_{2n}(0) = f(0)$$

y

$$Q_n(x_i^2) = P_{2n}(x_i) = f(x_i)$$

para $i = 1, 2, \ldots, n$, se tiene que Q_n es el polinomio de interpolación de Lagrange de la tabla

0	x_1^2	$\cdots$	x_n^2
$f(0)$	$f(x_1)$	$\cdots$	$f(x_n)$

La utilidad de lo anterior es que, para calcular el polinomio P_{2n}, podemos calcular Q_n (que es un polinomio de grado la mitad que el anterior) y luego cambiar la variable x por x^2, con el consiguiente ahorro de operaciones. □

6.7. Demostrar que si $f \in \mathcal{C}([a,b])$ y existe $f'(x_i)$ para algún $x_i \in [a,b]$ entonces la función

$$g(x) = \begin{cases} f[x, x_i] & \text{si} \quad x \neq x_i \\ f'(x_i) & \text{si} \quad x = x_i \end{cases}$$

es continua en el intervalo $[a,b]$ (lo que permite definir la diferencia dividida $f[x,x_i]$ en todo punto del intervalo $[a,b]$). ¿Tiene $f[x,x_i]$ alguna propiedad similar a la *fórmula de Leibniz* $(\varphi\psi)' = \varphi'\psi + \varphi\psi'$?

SOLUCIÓN. En primer lugar, como g es continua en $[a,b]\backslash\{x_i\}$ (por serlo f) y

$$\lim_{x\to x_i} g(x) = \lim_{x\to x_i} f[x,x_i] = \lim_{x\to x_i} \frac{f(x)-f(x_i)}{x-x_i} = f'(x_i) = g(x_i)$$

se tiene que la función g es continua en todo el intervalo $[a,b]$.

Por otra parte, denotando $h(x) = f(x)g(x)$, se verifica que

$$\begin{aligned}(fg)[x,x_i] = h[x,x_i] &= \frac{h(x)-h(x_i)}{x-x_i} = \frac{f(x)g(x)-f(x_i)g(x_i)}{x-x_i} \\ &= \frac{f(x)g(x)-f(x)g(x_i)+f(x)g(x_i)-f(x_i)g(x_i)}{x-x_i} \\ &= f(x)\frac{g(x)-g(x_i)}{x-x_i} + \frac{f(x)-f(x_i)}{x-x_i}g(x_i) \\ &= f(x)g[x,x_i] + f[x,x_i]g(x_i).\end{aligned}$$

Análogamente se cumple que

$$(fg)[x,x_i] = f[x,x_i]g(x) + f(x_i)g[x,x_i]. \quad \square$$

6.8. Sean $\{x_0, x_1, x_2, \ldots\} \subset \mathbb{R}$ tales que $x_i \neq x_j$ si $i \neq j$. Demostrar que para cada $i \in \mathbb{N} \cup \{0\}$ y para cada $m \in \mathbb{N}$ se verifica que

$$\Delta^m f(x_i) = \sum_{k=0}^{m} (-1)^k \binom{m}{k} f(x_{m+i-k}).$$

SOLUCIÓN. Fijado el índice i procedemos por inducción sobre m:

i) Para $m = 1$ el resultado es obvio, pues $\Delta f(x_i) = f(x_{i+1}) - f(x_i)$.

ii) Supuesto cierto el resultado para $m-1$ lo probamos para m. Por definición,

$$\Delta^m f(x_i) = \Delta^{m-1} f(x_{i+1}) - \Delta^{m-1} f(x_i)$$

por lo que, aplicando la hipótesis de inducción, se tiene que

$$\begin{aligned}
\Delta^m f(x_i) &= \sum_{k=0}^{m-1} (-1)^k \binom{m-1}{k} f(x_{m+i-k}) - \sum_{k=0}^{m-1} (-1)^k \binom{m-1}{k} f(x_{m-1+i-k}) \\
&= \sum_{k=0}^{m-1} (-1)^k \binom{m-1}{k} f(x_{m+i-k}) - \sum_{k=1}^{m} (-1)^{k-1} \binom{m-1}{k-1} f(x_{m+i-k}) \\
&= f(x_{m+i}) + \sum_{k=1}^{m-1} (-1)^k \left(\binom{m-1}{k} + \binom{m-1}{k-1} \right) f(x_{m+i-k}) + (-1)^m f(x_i) \\
&= \sum_{k=0}^{m} (-1)^k \binom{m}{k} f(x_{m+i-k})
\end{aligned}$$

donde se ha utilizado la siguiente igualdad entre números combinatorios:

$$\begin{aligned}
\binom{m-1}{k} + \binom{m-1}{k-1} &= \frac{(m-1)!}{k!(m-1-k)!} + \frac{(m-1)!}{(k-1)!(m-k)!} \\
&= \frac{(m-k)(m-1)! + k(m-1)!}{k!(m-k)!} \\
&= \frac{m(m-1)!}{k!(m-k)!} = \frac{m!}{k!(m-k)!} = \binom{m}{k}. \quad \square
\end{aligned}$$

6.9. Fórmula baricéntrica para el polinomio de interpolación. Sean $f : [a, b] \to \mathbb{R}$ y $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$ con $x_i \neq x_j$ si $i \neq j$. Denotando por

$$w_i = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{1}{x_i - x_j}$$

para $i = 0, 1, \ldots, n$, demostrar que el polinomio de interpolación de la función f en los puntos $\{x_0, x_1, \ldots, x_n\}$ puede expresarse en la forma

$$P_n(x) = \frac{\displaystyle\sum_{i=0}^{n} \frac{w_i}{x - x_i} f(x_i)}{\displaystyle\sum_{i=0}^{n} \frac{w_i}{x - x_i}}. \tag{6.31}$$

SOLUCIÓN. Esta forma de escribir el polinomio de interpolación puede resultar de utilidad si se necesita evaluar el polinomio en una gran cantidad de puntos. Nótese que, una vez calculados los $n + 1$ números w_i, la evaluación del polinomio P_n en un punto x exige del orden de $5n$ operaciones.

Con la notación aquí introducida, podemos escribir

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} = w_i \prod_{\substack{j=0 \\ j \neq i}}^{n} (x - x_j) = \frac{w_i}{x - x_i} \Pi_n(x)$$

para $i = 0, 1, \ldots, n$. Así, usando (6.27), se tiene que

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x) = \frac{\displaystyle\sum_{i=0}^{n} f(x_i) L_i(x)}{\displaystyle\sum_{i=0}^{n} L_i(x)} = \frac{\displaystyle\sum_{i=0}^{n} f(x_i) \frac{w_i}{x - x_i} \Pi_n(x)}{\displaystyle\sum_{i=0}^{n} \frac{w_i}{x - x_i} \Pi_n(x)}$$

de donde se deduce (6.31) sin más que dividir numerador y denominador por $\Pi_n(x)$. □

6.10. Lema de Aitken. Sean $\Omega = \{x_0, x_1, \ldots, x_n\}$ y S, T subconjuntos de Ω tales que $S \cap T = T - \{x_j\} = S - \{x_i\}$ con $i \neq j$, es decir, S y T tienen todos sus puntos en común salvo $x_i \in S$ y $x_j \in T$. Demostrar que

$$P^{S \cup T}(x) = \frac{(x_i - x) P^T(x) - (x_j - x) P^S(x)}{x_i - x_j},$$

donde P^Λ denota el polinomio de interpolación de la función f en los puntos del conjunto Λ.

SOLUCIÓN. Por hipótesis, S y T tienen la misma cantidad de puntos. Denotemos por

$$P(x) = \frac{(x_i - x) P^T(x) - (x_j - x) P^S(x)}{x_i - x_j}.$$

Por un lado, si $P^S, P^T \in \mathcal{P}_m$ entonces $P \in \mathcal{P}_{m+1}$. Además,

$$P(x_i) = P^S(x_i) = f(x_i),$$

$$P(x_j) = P^T(x_j) = f(x_j)$$

y, como $P^S(x_k) = P^T(x_k) = f(x_k)$ para $k \notin \{i, j\}$, se verifica que

$$P(x_k) = \frac{(x_i - x_k)f(x_k) - (x_j - x_k)f(x_k)}{x_i - x_j} = f(x_k).$$

Consecuentemente, como $P \in \mathcal{P}_{m+1}$ e interpola a f en todos los puntos de $S \cup T$ entonces $P = P^{S\cup T}$. □

6.11. Algoritmos de Aitken y de Neville. En este problema se presentan dos algoritmos clásicos para evaluar el polinomio de interpolación en un punto dado (obviamente, sin construir previamente dicho polinomio):

a) Algoritmo de Aitken. Se consideran los polinomios $P^{i,k}$ definidos como

$$P^{i,0}(x) = f(x_i),\ i = 0, 1, \ldots, n$$

y, para $k = 1, 2, \ldots, n$

$$P^{i,k}(x) = \frac{(x_i - x)P^{k-1,k-1}(x) - (x_{k-1} - x)P^{i,k-1}(x)}{x_i - x_{k-1}},\ i = k, k+1, \ldots, n.$$

Probar que $P^{i,k}$ es el polinomio de interpolación de f en el conjunto de puntos $\{x_0, x_1, \ldots, x_{k-1}, x_i\}$ para $k = 0, 1, \ldots, n$, $i = k, k+1, \ldots, n$. En particular, $P^{n,n}$ es el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_{n-1}, x_n\}$. ¿Cómo se calcularía el valor del polinomio de interpolación en un punto α?

b) Algoritmo de Neville. Se consideran los polinomios $P^{i,k}$ dados por

$$P^{i,0}(x) = f(x_i),\ i = 0, 1, \ldots, n$$

y, para $k = 1, 2, \ldots, n$

$$P^{i,k}(x) = \frac{(x_i - x)P^{i-1,k-1}(x) - (x_{i-k} - x)P^{i,k-1}(x)}{x_i - x_{i-k}},\ i = k, k+1, \ldots, n.$$

Demostrar que $P^{i,k}$ es el polinomio de interpolación de f en los puntos $\{x_{i-k}, x_{i-k+1}, \ldots, x_i\}$ para $k = 0, 1, \ldots, n$, $i = k, k+1, \ldots, n$. En particular, $P^{n,n}$ es el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_{n-1}, x_n\}$. ¿Cómo se evaluaría el polinomio de interpolación en un punto α?

SOLUCIÓN.

a) Demostramos el resultado por inducción sobre k:

i) $k = 0$. Evidente, pues $P^{i,0}(x) = f(x_i),\ i = 0, 1, \ldots, n$.

ii) Suponemos cierto el resultado para k y lo probamos para $k+1$. Por el lema de Aitken (véase el problema 6.10), el polinomio $P^{i,k+1}$ interpola a f en los puntos

$$\overbrace{\{x_0, x_1, \ldots, x_{k-1}, x_k\}}^{T} \cup \overbrace{\{x_0, x_1, \ldots, x_{k-1}, x_i\}}^{S} = \{x_0, x_1, \ldots, x_k, x_i\}.$$

Para evaluar el polinomio de interpolación en un punto $x = \alpha$ basta construir la siguiente tabla triangular

$$\begin{array}{|ccccc|}
\hline
y^{00} & & & & \\
y^{10} & y^{11} & & & \\
y^{20} & y^{21} & y^{22} & & \\
\cdots & \cdots & \cdots & & \\
y^{n0} & y^{n1} & y^{n2} & \cdots & y^{nn} \\
\hline
\end{array}$$

donde

$$y^{i0} = f(x_i),\ i = 0, 1, \ldots, n$$

y, para $k = 1, 2, \ldots, n$

$$y^{ik} = \frac{(x_i - \alpha)y^{k-1,k-1} - (x_{k-1} - \alpha)y^{i,k-1}}{x_i - x_{k-1}},\ i = k, k+1, \ldots, n.$$

El valor buscado es y^{nn}.

b) Procedemos nuevamente por inducción sobre k:

i) $k = 0$. Evidente, pues $P^{i,0}(x) = f(x_i),\ i = 0, 1, \ldots, n$.

ii) Suponemos cierto el resultado para k y lo probamos para $k+1$. El lema de Aitken concluye que el polinomio $P^{i,k+1}$ interpola a f en

$$\overbrace{\{x_{i-1-k}, x_{i-k}, \ldots, x_{i-1}\}}^{T} \cup \overbrace{\{x_{i-k}, x_{i-k+1}, \ldots, x_i\}}^{S}$$

es decir, en los puntos $\{x_{i-k-1}, x_{i-k}, \ldots, x_i\}$.

Para evaluar el polinomio de interpolación en un punto $x = \alpha$ basta construir la misma tabla triangular del apartado *a*) donde, ahora, los valores de y^{ik} vienen dados por

$$y^{i0} = f(x_i),\ i = 0, 1, \ldots, n$$

y, para $k = 1, 2, \ldots, n$

$$y^{ik} = \frac{(x_i - \alpha)y^{i-1,k-1} - (x_{i-k} - \alpha)y^{i,k-1}}{x_i - x_{i-k}},\ i = k, k+1, \ldots, n.$$

Nuevamente, el valor buscado es y^{nn}. □

6.12. Interpolación de Lagrange en dimensión 2. Consideremos un dominio rectangular $\Omega = [a, b] \times [c, d]$, una función $f : \Omega \to \mathbb{R}$, $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$ con $x_i \neq x_j$ si $i \neq j$ e $\{y_0, y_1, \ldots, y_m\} \subset [c, d]$ con $y_i \neq y_j$ si $i \neq j$ y el mallado

$$\Delta = \{(x_i, y_j) \in \mathbb{R}^2 :\ i = 0, 1, \ldots, n,\ j = 0, 1, \ldots, m\}.$$

a) Demostrar que el polinomio

$$P_{nm}(x, y) = \sum_{i=0}^{n} \sum_{j=0}^{m} f(x_i, y_j) L_{ni}(x) L_{mj}(y)$$

donde L_{ni} y L_{mj} son los polinomios de interpolación básicos de Lagrange, es decir,

$$L_{ni}(x) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k} \quad \text{y} \quad L_{mj}(y) = \prod_{\substack{k=0 \\ k \neq j}}^{m} \frac{y - y_k}{y_j - y_k}$$

verifica las $(n + 1)(m + 1)$ condiciones de interpolación

$$P_{nm}(x_i, y_j) = f(x_i, y_j) \tag{6.32}$$

para $i = 0, 1, \ldots, n$ y $j = 0, 1, \ldots, m$.

b) Probar que si el polinomio

$$Q(x, y) = \sum_{l=0}^{n} \sum_{k=0}^{m} a_{lk} x^l y^k \tag{6.33}$$

se anula en todos los puntos (x_i, y_j), $i = 0, 1, \ldots, n$, $j = 0, 1, \ldots, m$, entonces $Q(x, y)$ es idénticamente nulo.

c) Deducir un resultado de existencia y unicidad del polinomio de interpolación de Lagrange para funciones de dos variables.

d) Como aplicación, hallar el polinomio que interpola la siguiente tabla

	$x_0 = -1$	$x_1 = 2$	$x_2 = 4$
$y_0 = 1$	$f(x_0, y_0) = 1$	$f(x_1, y_0) = -1$	$f(x_2, y_0) = 1$
$y_1 = 2$	$f(x_0, y_1) = -1$	$f(x_1, y_1) = 0$	$f(x_2, y_1) = 3$

SOLUCIÓN.

a) La relación (6.32) se obtiene directamente a partir de la propiedad

$$L_{ni}(x_k) = \delta_{ik} \quad \text{y} \quad L_{mj}(y_k) = \delta_{jk}$$

que tienen los polinomios $L_{ni}(x)$ y $L_{mj}(y)$.

b) Para cada $j \in \{0, 1, \ldots, m\}$ consideramos el polinomio

$$q_j(x) = Q(x, y_j) = \sum_{l=0}^{n} b_{lj} x^l$$

siendo

$$b_{lj} = \sum_{k=0}^{m} a_{lk} y_j^k. \tag{6.34}$$

Como polinomio de una sola variable x que es, $q_j \in \mathcal{P}_n$ y tiene $n+1$ raíces distintas pues, por hipótesis,

$$q_j(x_i) = Q(x_i, y_j) = 0$$

para $i = 0, 1, \ldots, n$. Consecuentemente, por el teorema Fundamental del Álgebra se verifica que $q_j \equiv 0$. Por tanto, se tiene que

$$b_{lj} = 0$$

para $l = 0, 1, \ldots, n$ y $j = 0, 1, \ldots, m$. De esta forma, la expresión (6.34) nos indica que los polinomios

$$c_l(y) = \sum_{k=0}^{m} a_{lk} y^k$$

para $l = 0, 1, \ldots, n$, tienen, todos ellos, las $m+1$ raíces $\{y_0, y_1, \ldots, y_m\}$; consecuentemente, son todos nulos. Luego se verifica que

$$a_{lk} = 0$$

para $l = 0, 1, \ldots, n$ y $k = 0, 1, \ldots, m$, es decir, $Q \equiv 0$.

c) $P_{nm}(x,y)$ es, gracias al apartado *a*), un polinomio de grado $\leq n$ en x y grado $\leq m$ en y que interpola la función f en los puntos del mallado Δ. Veamos que es el único, esto es, que si $P_1(x,y)$ y $P_2(x,y)$ son dos polinomios de grado $\leq n$ en x y grado $\leq m$ en y verificando

$$P_1(x_i,y_j) = f(x_i,y_j) = P_2(x_i,y_j)$$

para $i = 0,1,\ldots,n$ y $j = 0,1,\ldots,m$ entonces P_1 coincide con P_2. Basta considerar el polinomio

$$Q(x,y) = P_1(x,y) - P_2(x,y)$$

que es de la forma (6.33) y verifica que

$$Q(x_i,y_j) = P_1(x_i,y_j) - P_2(x_i,y_j) = 0$$

para $i = 0,1,\ldots,n$ y $j = 0,1,\ldots,m$, por lo que el apartado *b*) asegura que $Q \equiv 0$ y, por tanto, $P_1 \equiv P_2$.

d) De la definición se obtienen los polinomios básicos de interpolación de Lagrange que, en este caso, vienen dados por

$$\begin{cases} L_{20}(x) = \dfrac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \dfrac{1}{15}(x-2)(x-4) \\ L_{21}(x) = \dfrac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = -\dfrac{1}{6}(x+1)(x-4) \\ L_{22}(x) = \dfrac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \dfrac{1}{10}(x+1)(x-2) \\ L_{10}(y) = \dfrac{y-y_1}{y_0-y_1} = 2-y \\ L_{11}(y) = \dfrac{y-y_0}{y_1-y_0} = y-1. \end{cases}$$

El polinomio de interpolación buscado

$$\begin{aligned} P_{21}(x,y) &= \sum_{i=0}^{2} \left(f(x_i,y_0)L_{10}(y) + f(x_i,y_1)L_{11}(y)\right) L_{2i}(x) \\ &= \left(f(x_0,y_0)L_{10}(y) + f(x_0,y_1)L_{11}(y)\right) L_{20}(x) \\ &\quad + \left(f(x_1,y_0)L_{10}(y) + f(x_1,y_1)L_{11}(y)\right) L_{21}(x) \\ &\quad + \left(f(x_2,y_0)L_{10}(y) + f(x_2,y_1)L_{11}(y)\right) L_{22}(x) \end{aligned}$$

es, en nuestro caso concreto,

$$\begin{aligned} P_{21}(x,y) &= (L_{10}(y) - L_{11}(y))\, L_{20}(x) - L_{10}(y) L_{21}(x) \\ &\quad + (L_{10}(y) + 3L_{11}(y))\, L_{22}(x) = \frac{1}{15}(x-2)(x-4)(3-2y) \\ &\quad + \frac{1}{6}(x+1)(x-4)(2-y) + \frac{1}{10}(x+1)(x-2)(2y-1) \\ &= -\frac{1}{30}\left((3y-13)x^2 + 3(21-11y)x + 2(12y-7)\right). \end{aligned}$$

La figura 6.12 muestra la gráfica del polinomio de interpolación anterior. □

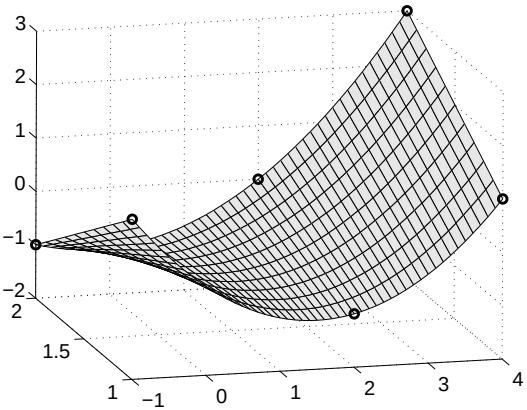


Figura 6.12: Interpolación de Lagrange en dimensión 2.

6.13. Demostrar que el polinomio de Tchebychev $T_n(x)$ tiene grado n y coeficiente director 2^{n-1} para cada $n \in \mathbb{N}$.

SOLUCIÓN. Recordemos que los polinomios $\{T_n(x)\}_{n=0}^{\infty}$ vienen dados, de forma recursiva, por

$$\begin{cases} T_0(x) = 1,\ T_1(x) = x \\ T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),\ n \in \mathbb{N}. \end{cases}$$

Probamos el resultado por inducción:

i) Para $n = 1$ el resultado es obvio, pues el polinomio $T_1(x)$ tiene grado 1 y coeficiente director $2^0 = 1$.

ii) Supuesto cierto el resultado hasta un $n \in \mathbb{N}$ arbitrario, es decir,

$$T_k(x) = 2^{k-1}x^k + t_{k-1}(x)$$

para $k = 1, 2, \ldots, n$, con $t_{k-1} \in \mathcal{P}_{k-1}$, el polinomio $T_{n+1}(x)$ se obtiene como

$$\begin{aligned} T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x) \\ &= 2x(2^{n-1}x^n + t_{n-1}(x)) - (2^{n-2}x^{n-1} + t_{n-2}(x)) \\ &= 2^n x^{n+1} + t_n(x) \end{aligned}$$

con $t_n \in \mathcal{P}_n$, que es el resultado buscado. □

6.14. Demostrar que si $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathcal{P}_n$ con $a_n \neq 0$ entonces

$$||P||_{\mathrm{L}^\infty(a,b)} \geq |a_n| \frac{(b-a)^n}{2^{2n-1}}$$

y, además, para todo $x \in [a, b]$, se verifica que

$$||P||_{\mathrm{L}^\infty(a,b)} = |a_n| \frac{(b-a)^n}{2^{2n-1}} \Leftrightarrow P(x) = a_n \frac{(b-a)^n}{2^{2n-1}} T_n\left(\frac{2x-a-b}{b-a}\right).$$

Solución. Mediante el cambio de variable

$$y = \frac{2x-a-b}{b-a}$$

transformamos el intervalo $[a, b]$ en $[-1, 1]$. Como

$$x = \frac{a+b}{2} + \frac{b-a}{2} y,$$

entonces

$$\begin{aligned} x^n &= \frac{1}{2^n}\left((a+b) + (b-a)y\right)^n = \frac{(b-a)^n}{2^n}\left(\frac{a+b}{b-a} + y\right)^n \\ &= \left(\frac{b-a}{2}\right)^n y^n + q_{n-1}(y) \end{aligned}$$

con $q_{n-1} \in \mathcal{P}_{n-1}$. De esta forma, podemos escribir P como

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = a_n \left(\frac{b-a}{2}\right)^n Q(y)$$

siendo $Q \in \mathcal{P}_n^{\mathrm{m}}$. Aplicando el teorema 6.6 obtenemos

$$||Q||_{\mathrm{L}^\infty(-1,1)} \geq ||T_n^{\mathrm{m}}||_{\mathrm{L}^\infty(-1,1)} = \frac{1}{2^{n-1}},$$

lo que implica que

$$||P||_{\mathrm{L}^\infty(a,b)} = |a_n|\frac{(b-a)^n}{2^n}\,||Q||_{\mathrm{L}^\infty(-1,1)} \geq |a_n|\frac{(b-a)^n}{2^n}\frac{1}{2^{n-1}} = |a_n|\frac{(b-a)^n}{2^{2n-1}}.$$

Además, por la observación 6.10, se verifica que

$$\begin{aligned}
||Q||_{\mathrm{L}^\infty(-1,1)} = ||T_n^{\mathrm{m}}||_{\mathrm{L}^\infty(-1,1)} &\Leftrightarrow\ Q(x) \equiv T_n^{\mathrm{m}}(x)\\
&\Leftrightarrow\ P(x) = a_n\left(\frac{b-a}{2}\right)^n Q(y)\\
&\qquad = a_n\left(\frac{b-a}{2}\right)^n T_n^{\mathrm{m}}(y)\\
&\qquad = a_n\left(\frac{b-a}{2}\right)^n \frac{T_n(y)}{2^{n-1}}\\
&\qquad = a_n\frac{(b-a)^n}{2^{2n-1}}T_n\left(\frac{2x-a-b}{b-a}\right). \quad \square
\end{aligned}$$

6.15. Sea $f \in \mathcal{C}^{n+1}([a,b])$. Demostrar que los puntos de interpolación

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{(2k+1)\pi}{2(n+1)}$$

para $k = 0, 1, \ldots, n$ minimizan la cota del error en la interpolación de la función f en el intervalo $[a,b]$ y que para cada $x \in [a,b]$ se tiene que

$$|E_n(x)| = |f(x) - P_n(x)| \leq \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}\left|\left|f^{n+1)}\right|\right|_{\mathrm{L}^\infty(a,b)}.$$

Solución. Basta considerar el cambio de variable

$$y = \frac{2x-a-b}{b-a} \ \Leftrightarrow\ x = \frac{a+b}{2} + \frac{b-a}{2}y$$

que lleva el intervalo $[a,b]$ al $[-1,1]$ (y viceversa) y la función $\widetilde{f} : [-1,1] \to \mathbb{R}$ dada por

$$\widetilde{f}(y) = f(x) = f\left(\frac{a+b}{2} + \frac{b-a}{2}y\right),$$

que verifica $\widetilde{f} \in \mathcal{C}^{n+1}([-1,1])$ y

$$\widetilde{f}^{k)}(y) = \left(\frac{b-a}{2}\right)^k f^{k)}\left(\frac{a+b}{2} + \frac{b-a}{2}y\right) = \left(\frac{b-a}{2}\right)^k f^{k)}(x)$$

para $k = 0, 1, \ldots, n+1$. Por el teorema 6.7, las abscisas de Tchebychev

$$y_k = \cos \frac{(2k+1)\pi}{2(n+1)}$$

determinan la cota mínima del error en la interpolación para la función $\widetilde{f}$ en el intervalo $[-1, 1]$. Llamando $P_n(x)$ al polinomio de interpolación de la función f en los puntos

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} y_k = \frac{a+b}{2} + \frac{b-a}{2} \cos \frac{(2k+1)\pi}{2(n+1)}$$

para $k = 0, 1, \ldots, n$, es fácil comprobar que

$$P_n(x) = \widetilde{P}_n(y)$$

donde $\widetilde{P}_n(y)$ es el polinomio de interpolación de $\widetilde{f}$ en los puntos $\{y_0, y_1, \ldots, y_n\}$. Por tanto,

$$E_n(x) = f(x) - P_n(x) = \widetilde{f}(y) - \widetilde{P}_n(y)$$

y los puntos $\{x_0, x_1, \ldots, x_n\}$ son los que minimizan la cota del error en la interpolación de f en el intervalo $[a, b]$. Además, para cada $x \in [a, b]$, se tiene que

$$\begin{aligned} |E_n(x)| = |\widetilde{f}(y) - \widetilde{P}_n(y)| &\leq \frac{1}{2^n (n+1)!} \left|\left| \widetilde{f}^{n+1)} \right|\right|_{\mathrm{L}^\infty(-1,1)} \\ &= \frac{1}{2^n (n+1)!} \left(\frac{b-a}{2} \right)^{n+1} \left|\left| f^{n+1)} \right|\right|_{\mathrm{L}^\infty(a,b)} \\ &= \frac{(b-a)^{n+1}}{2^{2n+1} (n+1)!} \left|\left| f^{n+1)} \right|\right|_{\mathrm{L}^\infty(a,b)}. \quad \square \end{aligned}$$

6.16. Sean λ, $\mu \in \mathbb{R}$.

a) Determinar los valores de λ y μ para que

$$S(x) = \begin{cases} \lambda x(x^2+1), & 0 \leq x \leq 1 \\ -\lambda x^3 + \mu x^2 - 5\lambda x + 1, & 1 \leq x \leq 2 \end{cases}$$

sea una función *spline* cúbica.

b) Con los valores de λ y μ obtenidos en el apartado *a*), ¿puede ser S una función *spline* cúbica de interpolación de la función

$$f(x) = x^2,\ 0 \leq x \leq 2$$

respecto de la partición $\Delta = \{0, 1, 2\}$?

SOLUCIÓN.

a) Para que S sea una función *spline* cúbica debe cumplir:

$$\begin{cases} S \in \mathcal{C}^2([0,2]) \\ S \text{ es un polinomio de grado} \leq 3 \text{ en los intervalos } [0,1] \text{ y } [1,2]. \end{cases}$$

Como la segunda condición se cumple independientemente de los valores de λ y μ, usaremos la primera:

i) Continuidad de S.

$$S(1^-) = S(1^+) \Leftrightarrow 2\lambda = -6\lambda + \mu + 1 \Leftrightarrow \mu = 8\lambda - 1. \tag{6.35}$$

ii) Continuidad de S'. Como

$$S'(x) = \begin{cases} \lambda(3x^2+1), & 0 < x < 1 \\ -3\lambda x^2 + 2\mu x - 5\lambda, & 1 < x < 2 \end{cases}$$

entonces

$$S'(1^-) = S'(1^+) \Leftrightarrow 4\lambda = -8\lambda + 2\mu \Leftrightarrow \mu = 6\lambda. \tag{6.36}$$

De las relaciones (6.35) y (6.36) se obtiene que

$$\begin{cases} \mu = 8\lambda - 1 \\ \mu = 6\lambda \end{cases} \Leftrightarrow \lambda = \frac{1}{2} \text{ y } \mu = 3,$$

por lo que la función S viene dada por

$$S(x) = \begin{cases} \frac{1}{2}x(x^2+1), & 0 \leq x \leq 1 \\ -\frac{1}{2}x^3 + 3x^2 - \frac{5}{2}x + 1, & 1 \leq x \leq 2 \end{cases}$$

(véase la figura 6.13). Nótese que para estos valores de λ y μ se tiene que

$$S''(x) = \begin{cases} 3x, & 0 < x < 1 \\ -3x + 6, & 1 < x < 2 \end{cases}$$

por lo que $S''(1^-) = 3 = S''(1^+)$ y, por tanto, $S \in \mathcal{C}^2([0,2])$.

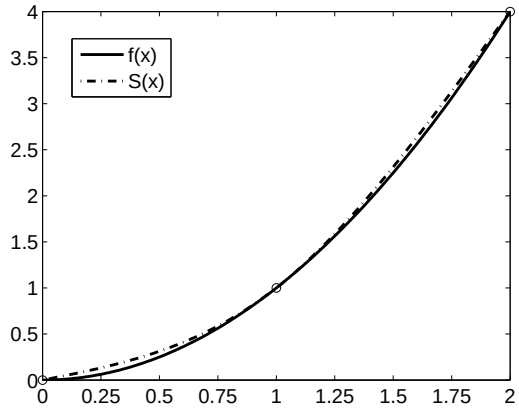


Figura 6.13: Función $f(x) = x^2$ y función *spline* de tipo I.

b) Como se verifica que

$$f(0) = 0 = S(0),\ f(1) = 1 = S(1)\ \text{ y }\ f(2) = 4 = S(2)$$

entonces S es una función *spline* cúbica que interpola la función $f(x) = x^2$ en los puntos de la partición $\Delta = \{0, 1, 2\}$. Como, además,

$$S''(0) = 0 = S''(2)$$

entonces S es una función *spline* cúbica interpoladora de tipo I. □

6.17. Sea $\Delta = \{a = x_0 < x_1 < \cdots < x_n = b\}$ una partición del intervalo $[a, b]$. Demostrar los siguientes resultados:

a) Si $n < 4$ toda función *spline* cúbica que verifique

$$S_\Delta^{k)}(a) = S_\Delta^{k)}(b) = 0 \tag{6.37}$$

para $k = 0, 1, 2$, es idénticamente nula.

b) Si $n = 4$ la anterior función *spline* cúbica está unívocamente determinada por el valor que tome en x_2.

Solución.

a) Distinguimos los posibles casos que pueden presentarse:

i) $\underline{n=1}$. En este caso, sólo son necesarias cuatro de las seis condiciones dadas. De hecho, vamos a probar que

$$\begin{cases} S''_\Delta(a) = S''_\Delta(b) = 0 \\ S_\Delta(a) = S'_\Delta(a) = 0 \end{cases} \Rightarrow S_\Delta \equiv 0 \text{ en } [a,b], \tag{6.38}$$

lo que nos será de utilidad en los restantes casos. Puesto que S''_Δ es una recta en $[a,b]$ y

$$S''_\Delta(a) = S''_\Delta(b) = 0,$$

la función S''_Δ es nula en $[a,b]$. Por tanto, S'_Δ será una constante, necesariamente nula, por ser $S'_\Delta(a) = 0$. Así, la función S_Δ es una constante en $[a,b]$; como $S_\Delta(a) = 0$, dicha constante es nula.

ii) $\underline{n=2}$. Puesto que

$$S_\Delta(a) = S_\Delta(b) = 0,$$

el teorema de Rolle asegura la existencia de un punto $\xi_1 \in (a,b)$ tal que $S'_\Delta(\xi_1) = 0$. Como $S'_\Delta(a) = 0$, nuevamente el teorema de Rolle implica que existe $\xi_2 \in (a,\xi_1)$ con $S''_\Delta(\xi_2) = 0$. Por tanto, la función S''_Δ tiene tres raíces distintas (a, ξ_2 y b) en $[a,b]$; consecuentemente, en uno de los intervalos $[a,x_1]$ o $[x_1,b]$ tendrá dos raíces y, por ser una recta, será idénticamente nula. Supongamos que esto ocurre en $[a,x_1]$ (en el otro intervalo, la argumentación sería análoga). Puesto que $S''_\Delta \equiv 0$ en $[a,x_1]$, en particular,

$$S''_\Delta(a) = S''_\Delta(x_1) = 0;$$

como, además,

$$S_\Delta(a) = S'_\Delta(a) = 0,$$

gracias a (6.38) se tiene que $S_\Delta \equiv 0$ en $[a,x_1]$. El hecho de que la función $S_\Delta \in \mathcal{C}^2([a,b])$ hace que

$$S^{k)}_\Delta(x_1) = 0 \tag{6.39}$$

para $k = 0,1,2$ y, por el caso $n=1$, se concluye que $S_\Delta \equiv 0$ en $[a,b]$.

iii) $\underline{n=3}$. A partir de las condiciones (6.37) se llega, análogamente, a que la función S''_Δ tiene dos raíces distintas en el intervalo abierto (a,b). Como además

$$S''_\Delta(a) = S''_\Delta(b) = 0,$$

la función S''_Δ tendrá dos raíces distintas en uno de los intervalos $[a,x_1]$, $[x_1,x_2]$ o $[x_2,b]$, es decir, $S''_\Delta \equiv 0$ en alguno de estos tres intervalos. Si esto ocurre en el intervalo:

α) $[a, x_1]$, como en el caso $n = 2$, se tiene que

$$\begin{cases} S''_\Delta(a) = S''_\Delta(x_1) = 0 \\ S_\Delta(a) = S'_\Delta(a) = 0 \end{cases}$$

por lo que (6.38) asegura que $S_\Delta \equiv 0$ en $[a, x_1]$. Nuevamente, por ser $S_\Delta \in \mathcal{C}^2([a, b])$, se tiene la propiedad (6.39), por lo que el caso $n = 2$ determina $S_\Delta \equiv 0$ en $[a, b]$.

β) $[x_1, x_2]$, como por un lado se tiene que $S^{k)}_\Delta(a) = 0$, $k = 0, 1, 2$ y $S''_\Delta(x_1) = 0$, la propiedad (6.38) implica que

$$S_\Delta \equiv 0 \text{ en } [a, x_1]. \tag{6.40}$$

Por otra parte, como $S^{k)}_\Delta(b) = 0$, $k = 0, 1, 2$ y $S''_\Delta(x_2) = 0$, el resultado simétrico a (6.38) (que se prueba de forma análoga) asegura que

$$S_\Delta \equiv 0 \text{ en } [x_2, b]. \tag{6.41}$$

A partir de (6.40) y (6.41) y puesto que $S_\Delta \in \mathcal{C}^2([a, b])$, aplicando el resultado para $n = 1$ se concluye que $S_\Delta \equiv 0$ en $[a, b]$.

γ) $[x_2, b]$, se argumenta de forma similar al caso en que las raíces se hallan en el intervalo $[a, x_1]$.

b) Sean S^1_Δ y S^2_Δ dos funciones *spline* cúbicas verificando (6.37) y

$$S^1_\Delta(x_2) = S^2_\Delta(x_2).$$

Para probar que $S^1_\Delta \equiv S^2_\Delta$ consideramos la función

$$S_\Delta(x) = S^1_\Delta(x) - S^2_\Delta(x)$$

y veamos que es la función idénticamente nula. Como

$$S_\Delta(a) = S_\Delta(x_2) = S_\Delta(b) = 0$$

la función S_Δ tiene tres raíces distintas en el intervalo $[a, b]$, por lo que, aplicando el teorema de Rolle, la función S'_Δ tiene, al menos, dos raíces distintas en (a, b); como además

$$S'_\Delta(a) = S'_\Delta(b) = 0,$$

la función S'_Δ tiene cuatro raíces distintas en el intervalo $[a, b]$. De esta forma, aplicando nuevamente el teorema de Rolle, la función S''_Δ tiene, al menos, tres raíces distintas en (a, b); como además

$$S''_\Delta(a) = S''_\Delta(b) = 0,$$

la función S''_Δ tiene, al menos, cinco raíces distintas en el intervalo $[a, b]$. Por tanto, existe $i \in \{0, 1, 2, 3\}$ tal que la función S''_Δ tiene dos raíces distintas en el intervalo $[x_i, x_{i+1}]$. Ahora bien, como S''_Δ es una recta en el intervalo $[x_i, x_{i+1}]$ entonces $S''_\Delta \equiv 0$ en $[x_i, x_{i+1}]$. Razonando como en el apartado a), distinguiendo los cuatro casos posibles, se deduce el resultado. □

6.4.2. Problemas propuestos

6.18. Consideramos la función $f(x) = e^x$. Utilizar las fórmulas de Lagrange y de Newton para obtener el polinomio de interpolación de f en los puntos

$$\{-2, -1, 0, 1, 2\} \quad \text{y} \quad \{-2, -1, 0, 1, 2, 3\}.$$

6.19. Si $\{x_0, x_1, \ldots, x_n\} \subset \mathbb{R}$ son $n+1$ puntos distintos, hallar el valor que toma la diferencia dividida $P[x_0, x_1, \ldots, x_n]$ cuando $P \in \mathcal{P}_n$.

6.20. Considerando la función $f(x) = \dfrac{1}{x}$, demostrar que

$$f[x_0, x_1, \ldots, x_n] = (-1)^n \prod_{i=0}^{n} \frac{1}{x_i}.$$

6.21. Si $T_k(x)$ denota el polinomio de Tchebychev de orden k, probar que para todo $n, m \in \mathbb{N} \cup \{0\}$ se verifica que

$$T_n\left(T_m(x)\right) = T_{nm}(x), \; x \in \mathbb{R}.$$

6.22. Demostrar que las raíces del n–ésimo polinomio de Tchebychev T_n son simétricas respecto al punto $x = 0$. Deducir que T_n se puede escribir como $P_k(x^2)$ o $xP_k(x^2)$, según que $n = 2k$ o $n = 2k + 1$, donde $P_k \in \mathcal{P}_k$.

6.23. Decidir si es verdadera o falsa la siguiente afirmación relativa a los polinomios de Tchebychev: "si n es un divisor de m entonces cada raíz de T_n es también raíz de T_m".

6.24. Sea $\Delta = \{x_0 = a < x_1 < \cdots < x_n = b\}$ una partición del intervalo $[a, b]$ y $S_\Delta(y, \cdot)$ la función *spline* cúbica de tipo I que interpola las componentes del vector $y = (y_0, y_1, \ldots, y_n)^{\mathrm{T}}$. ¿Qué deben verificar los valores $\{y_0, y_1, \ldots, y_n\}$ para que $S_\Delta(y, \cdot)$ coincida en todo el intervalo $[a, b]$ con un polinomio $P \in \mathcal{P}_3$?

6.25. Demostrar, a partir del teorema de Rolle, la unicidad de la función *spline* cúbica de interpolación de los tipos I y II.

6.26. A partir de una partición $\Delta = \{x_0 = a < x_1 < \cdots < x_n = b\}$ del intervalo $[a, b]$ se considera el espacio vectorial $\mathcal{S}$ de las funciones *spline* cúbicas de tipo I asociadas a dicha partición y las funciones $\{S_0, S_1, \ldots, S_n\}$ de $\mathcal{S}$ definidas por

$$S_i(x_j) = \delta_{ij}$$

para $i, j = 0, 1, \ldots, n$. Demostrar que $\mathcal{B} = \{S_0, S_1, \ldots, S_n\}$ constituye una base de $\mathcal{S}$ y deducir que si $S_\Delta(y, \cdot)$ es una función *spline* cúbica interpoladora de tipo I entonces

$$S_\Delta(y, x) = \sum_{i=0}^{n} y_i S_i(x).$$

6.5. Prácticas

6.1. Escribir un programa en `MATLAB` que calcule el polinomio de interpolación de Lagrange de una función en unos puntos dados mediante la fórmula de Newton, que permita añadir nuevos puntos de interpolación (de uno en uno) y que dibuje la función y el polinomio de interpolación obtenido.

6.2. Hacer una versión del programa anterior que sirva para interpolar los valores de una tabla dibujando el polinomo de interpolación y los valores interpolados.

6.3. Programar el cálculo de una función *spline* cúbica interpoladora de una función dada, dibujando ambas funciones.

6.4. Hacer una versión del programa anterior que interpole los valores de una tabla, dibujando la función *spline* cúbica obtenida y los valores interpolados.

6.5. Calcular los polinomios de interpolación de Lagrange de grados 5, 10, 15 y 20 de la función

$$f(x) = \exp\left(-\frac{1}{x^2}\right),\ x \in \left[-\frac{3}{2}, \frac{3}{2}\right]$$

tomando, por una parte, puntos equiespaciados y, por otra, las abscisas de Tchebychev. Comparar gráficamente los resultados.

7 Diferenciación e integración numéricas

7.1. Introducción

En el capítulo anterior se vio cómo el polinomio de interpolación de Lagrange de una función f se manifiesta como una útil herramienta cuando se quiere aproximar el valor de la función en puntos donde dicho valor es desconocido. En este capítulo haremos uso de esta herramienta para resolver dos problemas clásicos del Análisis Numérico como son la diferenciación y la integración aproximadas. En ambos casos, la idea básica será la misma: aproximar la derivada de la función en un punto y la integral de la función en un intervalo, mediante la derivada en dicho punto de su polinomio de interpolación y la integral del polinomio de interpolación en dicho intervalo, respectivamente.

Trataremos en una primera sección el problema de la diferenciación numérica y en la siguiente el de la integración numérica.

7.2. Diferenciación numérica

La aproximación del valor de la derivada de una función en un punto no es, hoy en día, objeto del que se deba ocupar la diferenciación numérica; los paquetes de cálculo simbólico (por ejemplo, el comando `diff` de `MATLAB`) proporcionan la expresión de la derivada y, por tanto, basta con evaluar dicha expresión en el punto dado. Muy diferente es el caso en que tan sólo se conozcan los valores de la función en algunos puntos, pues aquí las fórmulas de derivación aproximada sí serán de gran utilidad. El estudio teórico de la diferenciación numérica y del error en la derivación desempeña un papel de gran relevancia en el tratamiento de los métodos de resolución de ecuaciones diferenciales y del método de las *diferencias finitas* (véase el problema 7.3).

Veamos, mediante un ejemplo, cómo pueden obtenerse, de forma sencilla, fórmulas de derivación aproximada e, incluso, el error cometido en esta aproximación.

Ejemplo 7.1. Vamos a aproximar la derivada primera de una función f, que suponemos regular, en un punto x utilizando los valores de la función en dos puntos x y $x+h$. Para ello hacemos un desarrollo de Taylor de segundo orden

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(\xi),$$

que determina

$$\frac{f(x+h)-f(x)}{h} = f'(x) + \frac{h}{2}f''(\xi).$$

Por tanto, la fórmula

$$f'(x) \simeq \frac{f(x+h)-f(x)}{h}$$

aproxima la derivada primera de f en x con un *error*

$$E = \left|\frac{h}{2}f''(\xi)\right| \leq \frac{M_2}{2}h$$

donde denotamos

$$M_k = \max_{a\leq x\leq b}\left|f^{k)}(x)\right| = \left|\left|f^{k)}\right|\right|_{\mathrm{L}^\infty(a,b)}$$

para $k \in \mathbb{N}$. No obstante, podemos encontrar otra fórmula más exacta que la anterior si conocemos el valor que toma la función f en los puntos $x-h$, x y $x+h$. En efecto, como

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(\xi)$$

y

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(\eta),$$

restando ambas expresiones y dividiendo por $2h$ obtenemos

$$\frac{f(x+h)-f(x-h)}{2h} = f'(x) + \frac{h^2}{12}(f'''(\xi)+f'''(\eta)) = f'(x) + \frac{h^2}{6}f'''(\theta)$$

donde hemos aplicado el teorema de los Valores Intermedios a la función f'''. Así, la fórmula

$$f'(x) \simeq \frac{f(x+h)-f(x-h)}{2h} \tag{7.1}$$

aproxima la primera derivada de f en el punto x con un error

$$E = \left|\frac{h^2}{6}f'''(\theta)\right| \leq \frac{M_3}{6}h^2.$$

Análogamente se pueden hallar fórmulas de aproximación para las derivadas sucesivas. Por ejemplo, a partir de los desarrollos

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{iv)}(\xi)$$

y

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{iv)}(\eta)$$

obtenemos

$$\begin{aligned}\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} &= f''(x) + \frac{h^2}{24}(f^{iv)}(\xi) + f^{iv)}(\eta)) \\ &= f''(x) + \frac{h^2}{12}\frac{f^{iv)}(\xi) + f^{iv)}(\eta)}{2} \\ &= f''(x) + \frac{h^2}{12}f^{iv)}(\theta)\end{aligned}$$

a partir del teorema de los Valores Intermedios aplicado a la función $f^{iv)}$. De esta manera, la fórmula

$$f''(x) \simeq \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} \tag{7.2}$$

aproxima la segunda derivada de f en el punto x con un error

$$E = \left|\frac{h^2}{12}f^{iv)}(\theta)\right| \leq \frac{M_4}{12}h^2. \quad \square$$

Las fórmulas (7.1) y (7.2) son las más utilizadas para aproximar las derivadas primera y segunda, respectivamente, de una función en un punto. No obstante, se pueden deducir otras fórmulas de derivación aproximada sin más que explotar la idea ya expuesta de derivar el polinomio de interpolación de f en abscisas adecuadas; así se hará en la subsección 7.2.2.

7.2.1. El error en la diferenciación numérica

Sean $f : [a,b] \to \mathbb{R}$ y $\{x_0, x_1, \ldots, x_n\} \subset [a,b]$ con $x_i \neq x_j$ si $i \neq j$. Si $P_n \in \mathcal{P}_n$ es el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_n\}$, aproximamos la derivada de f en x por

$$\boxed{f'(x) \simeq P_n'(x) = \sum_{i=0}^{n} f(x_i)L_i'(x)}$$

y deseamos obtener una fórmula para el error que se comete en esta aproximación. El siguiente resultado ofrece una respuesta:

Teorema 7.1. *Sea $f \in \mathcal{C}^{n+2}([a,b])$, $\{x_0, x_1, \ldots, x_n\} \subset [a,b]$ con $x_i \neq x_j$ si $i \neq j$ y $P_n \in \mathcal{P}_n$ el polinomio de interpolación de f en los puntos $\{x_0, x_1, \ldots, x_n\}$. Para todo $x \in [a,b] \backslash \{x_0, x_1, \ldots, x_n\}$ existen $\xi_x \in (a,b)$ y $\eta_x \in (a,b)$ tales que*

$$E_n'(x) = f'(x) - P_n'(x) = \frac{f^{n+1)}(\xi_x)}{(n+1)!}\Pi_n'(x) + \frac{f^{n+2)}(\eta_x)}{(n+2)!}\Pi_n(x).$$

DEMOSTRACIÓN. Sea $x \in [a,b]$ con $x \neq x_i$ para $i = 0, 1, \ldots, n$. Puesto que

$$E_n(x) = f(x) - P_n(x) = \Pi_n(x) f[x_0, x_1, \ldots, x_n, x]$$

(véase (6.10)), formalmente

$$E_n'(x) = \Pi_n'(x) f[x_0, x_1, \ldots, x_n, x] + \Pi_n(x) \frac{d}{dx}\left(f[x_0, x_1, \ldots, x_n, x]\right). \tag{7.3}$$

Por una parte, gracias a (6.12), existe $\xi_x \in (a,b)$ tal que

$$f[x_0, x_1, \ldots, x_n, x] = \frac{f^{n+1)}(\xi_x)}{(n+1)!}.$$

Justifiquemos, a continuación, que la función

$$h(x) = f[x_0, x_1, \ldots, x_n, x]$$

es derivable en el punto x. Teniendo en cuenta la invarianza de las diferencias divididas respecto al orden en que aparecen los puntos, se tiene que

$$\begin{aligned}
\lim_{\varepsilon \to 0} \frac{h(x+\varepsilon) - h(x)}{\varepsilon} &= \lim_{\varepsilon \to 0} \frac{f[x_0, x_1, \ldots, x_n, x+\varepsilon] - f[x_0, x_1, \ldots, x_n, x]}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{f[x+\varepsilon, x_0, x_1, \ldots, x_n] - f[x_0, x_1, \ldots, x_n, x]}{(x+\varepsilon) - x} \\
&= \lim_{\varepsilon \to 0} f[x+\varepsilon, x_0, x_1, \ldots, x_n, x] \\
&= \lim_{\varepsilon \to 0} f[x_0, x_1, \ldots, x_n, x, x+\varepsilon].
\end{aligned}$$

De esta forma, a partir nuevamente de la relación (6.12), se verifica que

$$\lim_{\varepsilon \to 0} \frac{h(x+\varepsilon) - h(x)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{f^{n+2)}(\eta_{x,\varepsilon})}{(n+2)!} = \frac{f^{n+2)}(\eta_x)}{(n+2)!}$$

donde se usa la continuidad de la función $f^{n+2)}(\eta_x)$ demostrada en el corolario 6.1. Por tanto, la función h es derivable en x y

$$h'(x) = \frac{f^{n+2)}(\eta_x)}{(n+2)!}.$$

Sustituyendo este valor en (7.3) concluimos el resultado. □

Haciendo tender x a los puntos de interpolación y teniendo en cuenta la relación (6.30), se obtiene:

Corolario 7.1. *Sea* $f \in \mathcal{C}^{n+2}([a,b])$, $\{x_0, x_1, \ldots, x_n\} \subset [a,b]$ *con* $x_i \neq x_j$ *si* $i \neq j$ *y* $P_n \in \mathcal{P}_n$ *el polinomio de interpolación de* f *en los puntos* $\{x_0, x_1, \ldots, x_n\}$. *Entonces, para cada* $i \in \{0, 1, \ldots, n\}$, *se verifica que*

$$E_n'(x_i) = f'(x_i) - P_n'(x_i) = \frac{f^{n+1)}(\xi_{x_i})}{(n+1)!} \prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j). \quad \square \tag{7.4}$$

Observación 7.1. Una cota superior del error $E_n'(x_i)$ viene dada por

$$|E_n'(x_i)| \leq \frac{M_{n+1}}{(n+1)!} \prod_{\substack{j=0 \\ j \neq i}}^{n} |x_i - x_j|. \quad \square$$

Observación 7.2. En el caso particular de puntos equiespaciados, es decir, si $x_i = x_0 + ih$, $i = 0, 1, \ldots, n$ entonces, para cada $i \in \{0, 1, \ldots, n\}$, se verifica que

$$E_n'(x_i) = f'(x_i) - P_n'(x_i) = \frac{h^n f^{n+1)}(\xi_{x_i})}{(n+1)!} \prod_{\substack{j=0 \\ j \neq i}}^{n} (i - j).$$

Como veremos a continuación, las fórmulas de derivación numérica más utilizadas aproximan el valor de la derivada en puntos que son abscisas de interpolación. Por tanto, este último será el resultado de estimación del error que más se utilice. $\square$

7.2.2. Ejemplos de fórmulas de derivación

Para construir fórmulas de derivación aproximada vamos a considerar puntos equiespaciados que, por conveniencia de notación, vamos a suponer centrados alrededor de uno dado x_0; así, tendremos una red $\{\ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots\}$. Considerando el polinomio de interpolación de Lagrange de f en dos, tres o cinco puntos, derivando, y teniendo en cuenta la observación 7.2, se obtienen las siguientes fórmulas, algunas de las cuales ya se dedujeron en el ejemplo 7.1:

- $\boxed{n = 1}$ (dos puntos de interpolación, $\{x_0, x_1\}$):

$$\begin{cases} f'(x_0) = \dfrac{f(x_1) - f(x_0)}{h} - \dfrac{h}{2} f''(\xi_0) \\ f'(x_1) = \dfrac{f(x_0) - f(x_1)}{h} + \dfrac{h}{2} f''(\xi_1). \end{cases}$$

- $\boxed{n=2}$ (tres puntos de interpolación, $\{x_{-1}, x_0, x_1\}$):

$$\begin{cases} f'(x_{-1}) = \dfrac{-3f(x_{-1}) + 4f(x_0) - f(x_1)}{2h} + \dfrac{h^2}{3} f'''(\xi_{-1}) \\ f'(x_0) = \dfrac{f(x_1) - f(x_{-1})}{2h} - \dfrac{h^2}{6} f'''(\xi_0) \\ f'(x_1) = \dfrac{f(x_{-1}) - 4f(x_0) + 3f(x_1)}{2h} + \dfrac{h^2}{3} f'''(\xi_1). \end{cases}$$

- $\boxed{n=4}$ (cinco puntos de interpolación, $\{x_{-2}, x_{-1}, x_0, x_1, x_2\}$):

$$\begin{cases} f'(x_{-2}) = \dfrac{-25f(x_{-2}) + 48f(x_{-1}) - 36f(x_0) + 16f(x_1) - 3f(x_2)}{12h} + \dfrac{h^4}{5} f^{v)}(\xi_{-2}) \\ f'(x_{-1}) = \dfrac{-3f(x_{-2}) - 10f(x_{-1}) + 18f(x_0) - 6f(x_1) + f(x_2)}{12h} - \dfrac{h^4}{20} f^{v)}(\xi_{-1}) \\ f'(x_0) = \dfrac{f(x_{-2}) - 8f(x_{-1}) + 8f(x_1) - f(x_2)}{12h} + \dfrac{h^4}{30} f^{v)}(\xi_0) \\ f'(x_1) = \dfrac{-f(x_{-2}) + 6f(x_{-1}) - 18f(x_0) + 10f(x_1) + 3f(x_2)}{12h} - \dfrac{h^4}{20} f^{v)}(\xi_1) \\ f'(x_2) = \dfrac{3f(x_{-2}) - 16f(x_{-1}) + 36f(x_0) - 48f(x_1) + 25f(x_2)}{12h} + \dfrac{h^4}{5} f^{v)}(\xi_2). \end{cases}$$

Observación 7.3. Un somero examen de las fórmulas anteriores muestra que si el número de puntos es impar y se toma la derivada en el punto central, la fórmula correspondiente para diferenciación numérica tiene una expresión más sencilla y de mayor exactitud (además, el punto central no aparece en la fórmula y se requiere una evaluación menos de la función). □

Por último, destacar que la misma estrategia conduce a fórmulas para aproximar derivadas de orden superior.

7.3. Integración numérica

Uno de los problemas matemáticos más antiguos es el del cálculo del área que encierra una curva. Quizá el ejemplo más significativo haya sido el intento de conseguir la cuadratura del círculo, que condujo a la aparición y estudio del número π.

Las fórmulas de integración numérica (que, de hecho, se conocen también por el nombre de *fórmulas de cuadratura*) tienen como objetivo aproximar el valor de la integral de una función en un intervalo: de la que sólo se conocen los valores en algunos puntos, o cuya primitiva es difícil de calcular, o cuya primitiva no se puede expresar en términos de funciones elementales.

Nuevamente, el cálculo simbólico puede ayudar (en particular, el comando `int` de `MATLAB`) cuando se conoce la fórmula explícita de la función integrando y ésta admite una primitiva. En los restantes casos habrá que acudir a las fórmulas de integración aproximada. Además, no debe olvidarse que el estudio teórico de dichas fórmulas tiene especial relevancia en el diseño de los métodos de resolución de ecuaciones diferenciales (*métodos multipaso*).

Para conseguir nuestro objetivo, utilizaremos fórmulas del tipo

$$\boxed{\int_a^b f(x)\,dx \simeq \sum_{i=0}^{n} c_i f(x_i)}$$

que provienen de aproximar la integral de la función f en el intervalo $[a,b]$ por la integral en $[a,b]$ de su polinomio de interpolación en ciertos puntos $\{x_0, x_1, \ldots, x_n\}$. Se obtienen así las fórmulas de *tipo interpolatorio*. Más concretamente,

$$\int_a^b f(x)\,dx \simeq \int_a^b P_n(x)\,dx$$

donde

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x),\ x \in [a,b] \tag{7.5}$$

es el polinomio de interpolación de f en los $n+1$ puntos distintos $\{x_0, x_1, \ldots, x_n\}$ del intervalo $[a,b]$ (véase (6.1)). Integrando en $[a,b]$ la expresión (7.5) obtenemos

$$\int_a^b P_n(x)\,dx = \sum_{i=0}^{n} c_i f(x_i)$$

siendo

$$c_i = \int_a^b L_i(x)\,dx$$

para $i = 0, 1, \ldots, n$. Nótese que los coeficientes $\{c_0, c_1, \ldots, c_n\}$ son independientes de f y, por tanto, una vez calculados proporcionan una fórmula que se puede aplicar a cualquier función $f : [a,b] \to \mathbb{R}$. Por supuesto, será necesario estudiar el *error* que se comete en este tipo de fórmulas, es decir, el valor de

$$R_{(a,b)}(f) = \int_a^b f(x)\,dx - \int_a^b P_n(x)\,dx = \int_a^b (f(x) - P_n(x))\,dx.$$

Obviamente, si $f \in \mathcal{P}_n$ entonces f coincidirá con su polinomio de interpolación; consecuentemente, las fórmulas de tipo interpolatorio de $n+1$ puntos distintos son *exactas* para polinomios de grado menor o igual que n en el sentido de que

$$R_{(a,b)}(f) = 0.$$

A la hora de encontrar este tipo de fórmulas se distinguen dos tipos de estrategias, en función del contexto en el que se ha de trabajar:

a) Si los puntos $\{x_0, x_1, \ldots, x_n\}$ están fijados a priori, se trata de determinar los coeficientes $\{c_0, c_1, \ldots, c_n\}$. En el caso particular de que los puntos estén equiespaciados se obtienen las denominadas *fórmulas de Newton–Côtes*.

b) En otro caso se determinan los puntos $\{x_0, x_1, \ldots, x_n\}$ y los coeficientes $\{c_0, c_1, \ldots, c_n\}$. Se llega así a las *fórmulas de cuadratura de Gauss*.

7.3.1. Fórmulas de Newton–Côtes

Como acabamos de decir, las fórmulas de *Newton–Côtes* se obtienen como fórmulas de tipo interpolatorio cuando las abscisas de interpolación se toman equiespaciadas.

Antes de comenzar con el estudio general de dichas fórmulas, vamos a demostrar un resultado de gran utilidad que servirá para estimar el error cometido con ellas:

Teorema 7.2 (Valor Medio Integral Generalizado). *Sean $f, g \in \mathcal{C}([a,b])$ de forma que g no cambia de signo en el intervalo $[a,b]$. Existe $\theta \in (a,b)$ tal que*

$$\int_a^b f(x)g(x)\,dx = f(\theta)\int_a^b g(x)\,dx.$$

En particular, si $g \equiv 1$, entonces

$$\int_a^b f(x)\,dx = f(\theta)(b-a)$$

con $\theta \in (a,b)$.

Demostración. Sin pérdida de generalidad, supongamos que

$$g(x) \geq 0,\ x \in [a,b] \quad (g \not\equiv 0) \tag{7.6}$$

(el otro caso se aborda de forma análoga). Como $f \in \mathcal{C}([a,b])$, existen

$$m_0 = \min_{a\leq x\leq b} f(x) \ \text{ y } \ M_0 = \max_{a\leq x\leq b} f(x).$$

La no negatividad de g hace que se verifique

$$m_0 g(x) \leq f(x)g(x) \leq M_0 g(x),\ x \in [a,b],$$

con lo que

$$m_0 \int_a^b g(x)\,dx \leq \int_a^b f(x)g(x)\,dx \leq M_0 \int_a^b g(x)\,dx.$$

y, por tanto, como $\displaystyle\int_a^b g(x)\,dx > 0$ (véase (7.6)), entonces

$$m_0 \leq \frac{\displaystyle\int_a^b f(x)g(x)\,dx}{\displaystyle\int_a^b g(x)\,dx} \leq M_0.$$

Nuevamente, al ser $f \in \mathcal{C}([a,b])$, el teorema de los Valores Intermedios implica la existencia de $\theta \in [a,b]$ tal que

$$f(\theta) = \frac{\displaystyle\int_a^b f(x)g(x)\,dx}{\displaystyle\int_a^b g(x)\,dx}.$$

Para concluir, nótese que si $\theta = a$ o $\theta = b$ entonces

$$\int_a^b (f(\theta) - f(x))g(x)\,dx = 0$$

y, por tanto, existiría $\eta \in (a,b)$ tal que $f(\eta) = f(\theta)$. □

Presentamos algunos casos particulares de fórmulas de Newton–Côtes:

a) Fórmula del trapecio. Consiste en sustituir la integral de f por el área del trapecio inscrito en la gráfica (véase la figura 7.1), es decir, recordando que el área de un trapecio es la semisuma de las bases multiplicada por la altura:

$$\int_a^b f(x)\,dx \simeq \frac{f(a)+f(b)}{2}(b-a).$$

Esta aproximación es la que se obtiene al integrar el polinomio de interpolación de la función f en los puntos extremos del intervalo $[a,b]$. En efecto, denotando por $P_1 \in \mathcal{P}_1$ a dicho polinomio, es decir,

$$P_1(x) = f(a) + (x-a)f[a,b] = f(a) + \frac{f(b)-f(a)}{b-a}(x-a),$$

se tiene que

$$\begin{aligned}\int_a^b f(x)\,dx \simeq \int_a^b P_1(x)\,dx &= f(a)x + \frac{f(b)-f(a)}{b-a}\frac{(x-a)^2}{2}\Bigg|_a^b \\ &= f(a)(b-a) + \frac{f(b)-f(a)}{b-a}\frac{(b-a)^2}{2} \\ &= f(a)(b-a) + \frac{f(b)-f(a)}{2}(b-a) \\ &= \frac{b-a}{2}(f(a)+f(b)).\end{aligned}$$

Para calcular el error cometido en esta aproximación, como por el teorema 6.2 sabemos que

$$E_1(x) = f(x) - P_1(x) = \frac{f''(\xi_x)}{2!}\Pi_1(x),\ x \in [a,b]$$

entonces

$$R_{(a,b)}(f) = \int_a^b (f(x) - P_1(x))\,dx = \frac{1}{2}\int_a^b f''(\xi_x)\Pi_1(x)\,dx.$$

Como la función

$$x \mapsto f''(\xi_x)$$

es continua (véase el corolario 6.1) y

$$\Pi_1(x) = (x-a)(x-b) \leq 0,\ x \in [a,b],$$

por el teorema 7.2 se tiene que existe $\theta \in (a,b)$ tal que

$$\begin{aligned}R_{(a,b)}(f) &= \frac{f''(\theta)}{2}\int_a^b (x-a)(x-b)\,dx \\ &= \frac{f''(\theta)}{2}\int_a^b (x^2-(a+b)x+ab)\,dx \\ &= \frac{f''(\theta)}{2}\left(\frac{x^3}{3} - (a+b)\frac{x^2}{2} + abx\right)\Bigg|_a^b \\ &= \frac{f''(\theta)}{2}\left(\frac{b^3-a^3}{3} - (a+b)\frac{b^2-a^2}{2} + ab(b-a)\right),\end{aligned}$$

es decir,

$$\begin{aligned} R_{(a,b)}(f) &= \frac{f''(\theta)}{12}\left(2(b^3-a^3)-3(a+b)(b^2-a^2)+6ab(b-a)\right) \\ &= \frac{f''(\theta)}{12}\left(a^3-3a^2b+3ab^2-b^3\right) \\ &= \frac{f''(\theta)}{12}(a-b)^3 = -\frac{(b-a)^3}{12}f''(\theta). \end{aligned}$$

De esta forma, llamando $x_0 = a$, $x_1 = b$ y $h = b - a$,

$$\boxed{\int_a^b f(x)\,dx \simeq \frac{h}{2}\left(f(x_0)+f(x_1)\right)}$$

donde el error cometido en esta fórmula viene dado por

$$R_{(a,b)}(f) = -\frac{h^3}{12}f''(\theta)$$

para algún $\theta \in (a, b)$.

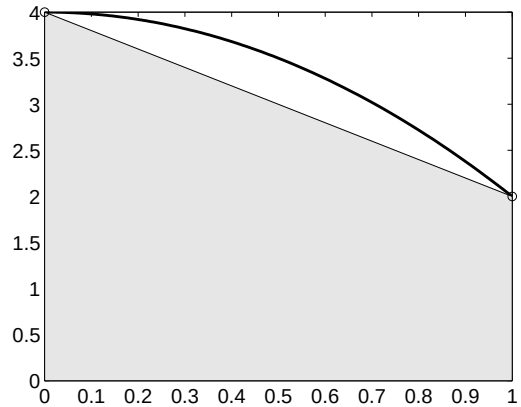


Figura 7.1: Fórmula del trapecio.

Ejemplo 7.2. Dada la función

$$f(x) = 4 - 2x^2,\ x \in [0, 1],$$

el valor aproximado de la integral de f en $[0, 1]$ que se obtiene mediante la fórmula del trapecio es

$$\int_0^1 (4-2x^2)\,dx \simeq \frac{f(0)+f(1)}{2} = 3$$

(véase la figura 7.1). En este caso concreto, el error cometido se puede hallar explícitamente y vale

$$R_{(0,1)}(f) = -\frac{h^3}{12} f''(\theta) = -\frac{1}{12}(-4) = \frac{1}{3}.$$

Se trata, por tanto, de una aproximación por *defecto*. □

b) Fórmula del trapecio abierta. Se divide el intervalo $[a, b]$ en tres partes iguales y se considera el polinomio de interpolación de f en los dos puntos intermedios $\{x_0 = a+h, x_1 = a+2h\}$ donde $h = \dfrac{b-a}{3}$. Aproximamos la integral de f por el área del trapecio correspondiente. Como puede demostrarse (véase el problema 7.6) se obtiene

$$\boxed{\int_a^b f(x)\,dx \simeq \frac{3h}{2}\left(f(x_0) + f(x_1)\right)}$$

siendo

$$R_{(a,b)}(f) = \frac{3h^3}{4} f''(\theta)$$

el error cometido en esta fórmula para algún $\theta \in (a, b)$.

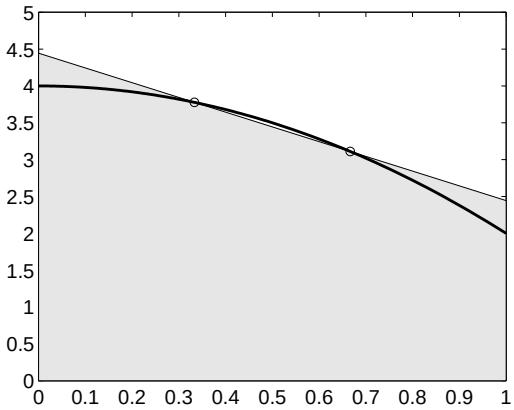


Figura 7.2: Fórmula del trapecio abierta.

Ejemplo 7.3. Para la función del ejemplo 7.2, mediante la fórmula del trapecio abierta se obtiene que

$$\int_0^1 (4 - 2x^2)\,dx \simeq \frac{f\left(\dfrac{1}{3}\right) + f\left(\dfrac{2}{3}\right)}{2} = \frac{1}{2}\left(\frac{34}{9} + \frac{28}{9}\right) = \frac{31}{9}$$

(véase la figura 7.2). El error cometido es

$$R_{(0,1)}(f) = \frac{3h^3}{4} f''(\theta) = \frac{3}{4}\left(\frac{1}{3}\right)^3 (-4) = -\frac{1}{9}.$$

Así pues, en este caso, se obtiene una aproximación por *exceso*. □

c) Fórmula de Simpson. Se considera el polinomio de interpolación de f en los puntos $\left\{x_0 = a, x_1 = \dfrac{a+b}{2}, x_2 = b\right\}$. Tomando como paso

$$h = \frac{b-a}{2},$$

podemos escribir

$$x_i = x_0 + ih,\ i = 0, 1, 2,$$

con lo que el polinomio de interpolación buscado es

$$\begin{aligned} P_2(x) &= f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \\ &= f(x_0) + (x - x_0)\frac{\Delta f(x_0)}{h} + (x - x_0)(x - x_1)\frac{\Delta^2 f(x_0)}{2h^2} \\ &= f(x_0) + (x - x_0)\frac{f(x_1) - f(x_0)}{h} \\ &\quad + (x - x_0)(x - x_1)\frac{f(x_0) - 2f(x_1) + f(x_2)}{2h^2}. \end{aligned}$$

Se deja como ejercicio al lector comprobar que

$$\int_a^b P_2(x)\, dx = \frac{h}{3}\left(f(x_0) + 4f(x_1) + f(x_2)\right).$$

Ejemplo 7.4. Considerando la función

$$f(x) = x^3 - 3x^2 + 2x + 3,\ x \in [0, 3],$$

el valor aproximado de la integral de f en $[0, 3]$ que se obtiene mediante la fórmula de Simpson es

$$\int_0^3 (x^3 - 3x^2 + 2x + 3)\, dx \simeq \frac{f(0) + 4f\left(\dfrac{3}{2}\right) + f(3)}{2} = \frac{1}{2}\left(3 + 4\frac{21}{8} + 9\right) = \frac{45}{4}$$

(véase la figura 7.3). Comparando con el valor exacto de la integral

$$\int_0^3 (x^3 - 3x^2 + 2x + 3)\, dx = \frac{1}{4}x^4 - x^3 + x^2 + 3x\Big|_0^3 = \frac{45}{4} \qquad (7.7)$$

se observa que el error cometido es nulo. □

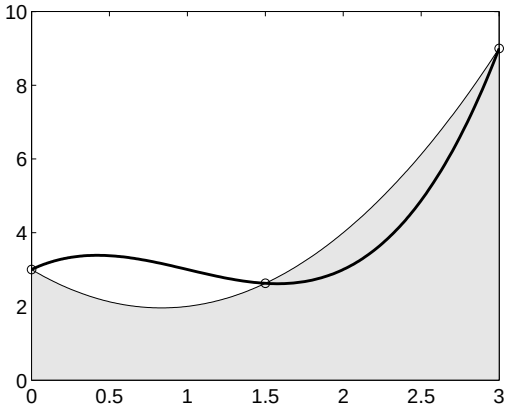


Figura 7.3: Fórmula de Simpson.

Para calcular el error en la fórmula de Simpson

$$\begin{aligned} R_{(a,b)}(f) &= \int_a^b (f(x) - P_2(x))\,dx \\ &= \int_a^b f(x)\,dx - \frac{h}{3}(f(x_0) + 4f(x_1) + f(x_2)) \end{aligned}$$

consideramos la función $\psi : [0, h] \to \mathbb{R}$ dada por

$$\psi(r) = \int_{x_1-r}^{x_1+r} f(x)\,dx - \frac{r}{3}(f(x_1 - r) + 4f(x_1) + f(x_1 + r)).$$

Claramente

$$R_{(a,b)}(f) = \psi(h) \tag{7.8}$$

y si $f \in \mathcal{C}^4([a, b])$ entonces $\psi \in \mathcal{C}^4([0, h])$. Bajo este supuesto, para todo $r \in [0, h]$ se verifica que

$$\psi'(r) = \frac{2}{3}[f(x_1 - r) - 2f(x_1) + f(x_1 + r)] + \frac{r}{3}[f'(x_1 - r) - f'(x_1 + r)],$$

$$\psi''(r) = \frac{1}{3}[f'(x_1 + r) - f'(x_1 - r)] - \frac{r}{3}[f''(x_1 - r) + f''(x_1 + r)]$$

y

$$\psi'''(r) = -\frac{r}{3}[f'''(x_1 + r) - f'''(x_1 - r)].$$

Además, como se observa,

$$\psi(0) = \psi'(0) = \psi''(0) = \psi'''(0) = 0. \tag{7.9}$$

A la vista de la forma que tiene la derivada tercera de ψ, por el teorema del Valor Medio, para todo $r \in [0, h]$ existe $\xi_r \in (x_1 - r, x_1 + r)$ tal que

$$\psi'''(r) = -\frac{r}{3}[f'''(x_1 + r) - f'''(x_1 - r)] = -\frac{2r^2}{3} f^{iv)}(\xi_r).$$

Integrando esta expresión obtenemos

$$\psi''(r) = -\frac{2}{3}\int_0^r f^{iv)}(\xi_s) s^2\, ds + \psi''(0) = -\frac{2}{3}\int_0^r f^{iv)}(\xi_s) s^2\, ds$$

(véase (7.9)). Como la función

$$s \mapsto f^{iv)}(\xi_s)$$

es continua (véase el corolario 6.1) y la función

$$s \mapsto s^2$$

no cambia de signo en el intervalo $[0, h]$, por el teorema 7.2 existe $\eta_r \in (a, b)$ tal que

$$\psi''(r) = -\frac{2}{3} f^{iv)}(\eta_r) \int_0^r s^2\, ds = -\frac{2}{9} f^{iv)}(\eta_r) r^3.$$

Reiterando este argumento, de forma análoga se obtiene que

$$\begin{aligned}\psi'(r) &= -\frac{2}{9}\int_0^r f^{iv)}(\eta_s) s^3\, ds + \psi'(0) = -\frac{2}{9}\int_0^r f^{iv)}(\eta_s) s^3\, ds \\ &= -\frac{2}{9} f^{iv)}(\nu_r) \int_0^r s^3\, ds = -\frac{1}{18} f^{iv)}(\nu_r) r^4\end{aligned}$$

para algún $\nu_r \in (a, b)$ y, finalmente,

$$\begin{aligned}\psi(r) &= -\frac{1}{18}\int_0^r f^{iv)}(\nu_s) s^4\, ds + \psi(0) = -\frac{1}{18}\int_0^r f^{iv)}(\nu_s) s^4\, ds \\ &= -\frac{1}{18} f^{iv)}(\theta_r) \int_0^r s^4\, ds = -\frac{1}{90} f^{iv)}(\theta_r) r^5\end{aligned}$$

para algún $\theta_r \in (a, b)$. En particular,

$$R_{(a,b)}(f) = \psi(h) = -\frac{1}{90} f^{iv)}(\theta_h) h^5 \tag{7.10}$$

(véase (7.8)) y, por tanto,

$$\boxed{\int_a^b f(x)\,dx \simeq \frac{h}{3}\left(f(x_0)+4f(x_1)+f(x_2)\right)}$$

donde el error cometido en la fórmula anterior viene expresado en (7.10).

La regla de Simpson es de especial interés porque su precisión es mayor de lo que podría esperarse a partir del conocimiento de la función f en tan sólo tres puntos (nótese que es *exacta* para polinomios de grado menor o igual que 3). Esta propiedad justifica que no se cometiera error al aplicarla a la función del ejemplo 7.4.

d) Fórmula de Simpson abierta. En este caso se subdivide el intervalo $[a,b]$ en cuatro subintervalos iguales de longitud

$$h = \frac{b-a}{4}$$

y se consideran los puntos $\{x_0 = a+h, x_1 = a+2h, x_2 = a+3h\}$. El polinomio de interpolación de f en los puntos $\{x_0, x_1, x_2\}$ viene dado por

$$\begin{aligned} P_2(x) = f(x_0) &+ (x-x_0)\frac{f(x_1)-f(x_0)}{h} \\ &+ (x-x_0)(x-x_1)\frac{f(x_0)-2f(x_1)+f(x_2)}{2h^2}. \end{aligned}$$

Se deja nuevamente al lector que compruebe que

$$\boxed{\int_a^b f(x)\,dx \simeq \frac{4h}{3}\left(2f(x_0)-f(x_1)+2f(x_2)\right)}$$

donde la expresión del error viene dada por

$$R_{(a,b)}(f) = \frac{14h^5}{45}f^{iv)}(\theta)$$

para algún $\theta \in (a,b)$.

Ejemplo 7.5. Para la función del ejemplo 7.4 el valor que se obtiene mediante la fórmula de Simpson abierta es

$$\begin{aligned} \int_0^3 (x^3-3x^2+2x+3)\,dx &\simeq 2f\left(\frac{3}{4}\right) - f\left(\frac{3}{2}\right) + 2f\left(\frac{9}{4}\right) \\ &= 2\frac{207}{64} - \frac{21}{8} + 2\frac{237}{64} = \frac{45}{4} \end{aligned}$$

(véase la figura 7.4) por lo que, nuevamente, el error cometido es nulo (véase (7.7)). □

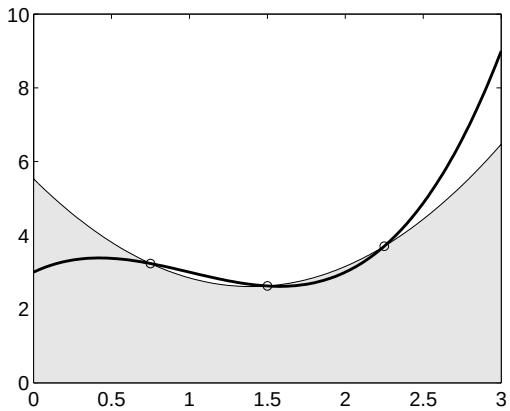


Figura 7.4: Fórmula de Simpson abierta.

Observación 7.4. Otra fórmula utilizada con bastante frecuencia en las aplicaciones es la denominada de los *tres octavos*, en la que se utilizan cuatro puntos de interpolación $x_i = a + ih$, $i = 0, 1, 2, 3$, con $h = \dfrac{b-a}{3}$:

$$\boxed{\int_a^b f(x)\,dx \simeq \frac{3h}{8}\left(f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right)}$$

donde el error cometido en la aproximación anterior viene dado por

$$R_{(a,b)}(f) = -\frac{3h^5}{80} f^{iv)}(\theta)$$

con $\theta \in (a, b)$. □

Como se observa en los ejemplos anteriores, las fórmulas de Newton–Côtes para $n + 1$ puntos de la forma

$$x_i = x_0 + ih$$

para $i = 0, 1, \ldots, n$, son de dos tipos:

a) *Fórmulas cerradas*: se toman los extremos del intervalo como puntos de interpolación, es decir, se consideran

$$h = \frac{b-a}{n} \quad \text{y} \quad x_0 = a.$$

b) *Fórmulas abiertas*: no se consideran los extremos del intervalo como puntos de interpolación. En este caso, se toman

$$h = \frac{b-a}{n+2} \quad \text{y} \quad x_0 = a + h.$$

Veamos a continuación un resultado de carácter general para el error cometido en las fórmulas de Newton–Côtes.

Teorema 7.3. *Sea* $n = 2k$ (*resp.* $n = 2k+1$) *y* $f \in \mathcal{C}^{2k+2}([a,b])$. *Entonces, el error en las fórmulas de Newton–Côtes* (*abierta y cerrada*) *para* $n+1$ *puntos viene dado por*

$$R_{(a,b)}(f) = \frac{C}{(2k+2)!} h^{2k+3} f^{2k+2)}(\theta) \tag{7.11}$$

para un cierto $\theta \in (a,b)$, *donde el valor de la constante* C *es independiente de* f *pero es distinto en cada fórmula.*

DEMOSTRACIÓN. Véase [**Is–Ke**]. □

Observación 7.5.

1. Teniendo en cuenta que si n es par la derivada que aparece en la fórmula del error es la de orden $n+2$, las fórmulas de Newton–Côtes de $n+1$ puntos con n par son exactas para polinomios de grado $n+1$ (superior en una unidad al grado esperado).

2. A simple vista puede parecer que las fórmulas abiertas y las cerradas de igual número de puntos tienen el mismo orden de error, dado que aparece la misma potencia de h; no obstante, no hay que olvidar que el valor de h difiere de las fórmulas abiertas a las cerradas. Así pues, debemos comparar una fórmula abierta con una cerrada para el mismo valor de h. Dado un paso $h = \dfrac{b-a}{p}$, éste corresponde a la fórmula cerrada de $p+1$ puntos y a la fórmula abierta de $p-1$ puntos; el exponente de h será en la expresión del error de la fórmula abierta inferior en dos unidades al exponente correspondiente a la fórmula cerrada. Consecuentemente, el orden de error de las fórmulas cerradas es superior en dos unidades al de las fórmulas abiertas. □

Veamos ahora cómo pueden hallarse los coeficientes de las fórmulas de Newton–Côtes. Fijaremos nuestra atención en las fórmulas cerradas (en las fórmulas abiertas se procede de forma análoga). Recuérdese que los coeficientes de estas fórmulas vienen dados por

$$c_i = \int_a^b L_i(x)\,dx$$

para $i = 0, 1, \ldots, n$. Como los puntos $\{x_0, x_1, \ldots, x_n\}$ son equidistantes, el cambio de variable

$$x = a + sh \;\text{ siendo }\; h = \frac{b-a}{n},$$

conduce a

$$c_i = \int_a^b \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}\, dx = \int_0^n \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{h(s-j)}{h(i-j)} h\, ds = h\alpha_i$$

para $i = 0, 1, \ldots, n$, donde

$$\alpha_i = \int_0^n \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{s-j}{i-j}\, ds.$$

De esta forma,

$$\boxed{\int_a^b P_n(x)\, dx = h \sum_{i=0}^{n} \alpha_i f(x_i)} \tag{7.12}$$

Nótese que los valores $\{\alpha_0, \alpha_1, \ldots, \alpha_n\}$ son independientes del intervalo (sólo dependen del número de puntos de interpolación). En consecuencia, estos valores pueden ser calculados de una vez por todas (de hecho, están tabulados) y, por tanto, los coeficientes de las fórmulas de Newton–Côtes son "casi" independientes también del intervalo de integración (salvo el factor $h = \dfrac{b-a}{n}$ que aparece en (7.12)).

Observación 7.6. Basta considerar la función $f(x) \equiv 1$, $x \in [a, b]$ para demostrar que

$$\sum_{i=0}^{n} \alpha_i = n.$$

En efecto, como $P_n \equiv 1$ entonces

$$b - a = \int_a^b P_n(x)\, dx = h \sum_{i=0}^{n} \alpha_i f(x_i) = h \sum_{i=0}^{n} \alpha_i. \quad \square$$

En el caso de que se desee obtener una fórmula general de tipo interpolatorio puede recurrirse al *método de los coeficientes indeterminados*, donde la idea básica es utilizar el hecho de que, si se utilizan $n + 1$ puntos distintos, la fórmula de integración es exacta para polinomios de grado menor o igual que n (ésta es una forma de hallar los coeficientes $\{c_0, c_1, \ldots, c_n\}$ cuando se carece de tablas o cuando las abscisas $\{x_0, x_1, \ldots, x_n\}$ no son equidistantes y, por tanto, no están tabulados).

Trabajando con la base natural de polinomios de grado menor o igual que n, es decir, $\{1, x, x^2, \ldots, x^n\}$, se obtienen las siguientes ecuaciones:

$$\sum_{i=0}^{n} c_i x_i^k = \int_a^b x^k \, dx = \frac{b^{k+1} - a^{k+1}}{k+1}$$

para $k = 0, 1, \ldots, n$. De esta forma, los coeficientes $\{c_0, c_1, \ldots, c_n\}$ vienen dados como la solución de un sistema lineal $Ac = d$ de $n+1$ ecuaciones con $n+1$ incógnitas, siendo

$$A = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_0 & x_1 & \ldots & x_n \\ x_0^2 & x_1^2 & \ldots & x_n^2 \\ \ldots & \ldots & \ldots & \ldots \\ x_0^n & x_1^n & \ldots & x_n^n \end{pmatrix}, \; c = \begin{pmatrix} c_0 \\ c_1 \\ \ldots \\ c_n \end{pmatrix} \quad \text{y} \quad d = \begin{pmatrix} b-a \\ \dfrac{b^2-a^2}{2} \\ \ldots \\ \dfrac{b^{n+1}-a^{n+1}}{n+1} \end{pmatrix}.$$

Como el determinante de A es de Vandermonde

$$\det(A) = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ x_0 & x_1 & \ldots & x_n \\ x_0^2 & x_1^2 & \ldots & x_n^2 \\ \ldots & \ldots & \ldots & \ldots \\ x_0^n & x_1^n & \ldots & x_n^n \end{vmatrix} = \prod_{i>j} (x_i - x_j) = \prod_{j=0}^{n-1} \left(\prod_{i=j+1}^{n} (x_i - x_j) \right) \neq 0$$

(ya que $x_i \neq x_j$ si $i \neq j$), se verifica que los coeficientes $\{c_0, c_1, \ldots, c_n\}$ están unívocamente determinados.

Ejemplo 7.6. Vamos a encontrar un fórmula que aproxime la integral de una función f en el intervalo $[1, 3]$, utilizando los valores de f en los puntos 0, 2 y 4 y que sea exacta para polinomios del mayor grado posible. Suponiendo que la fórmula buscada es del tipo

$$\int_1^3 f(x) \, dx \simeq Af(0) + Bf(2) + Cf(4)$$

el problema reside en determinar las constantes $A, B, C \in \mathbb{R}$. Para ello, vamos a imponer que la fórmula de integración sea exacta para polinomios de grado n, teniendo en cuenta que

$$\int_1^3 x^n \, dx = \frac{3^{n+1} - 1}{n+1}$$

para $n \in \mathbb{N} \cup \{0\}$. Dando valores crecientes a n se obtiene que

$$\begin{cases} n = 0 & \Rightarrow \quad 2 = A + B + C \\ n = 1 & \Rightarrow \quad 4 = 2B + 4C \\ n = 2 & \Rightarrow \quad \dfrac{26}{3} = 4B + 16C. \end{cases}$$

La solución de este sistema lineal de tres ecuaciones con tres incógnitas es (compruébese)

$$A = C = \frac{1}{12} \ \text{ y } \ B = \frac{11}{6},$$

por lo que la fórmula buscada es

$$\int_1^3 f(x)\,dx \simeq \frac{1}{12}\left(f(0) + 22f(2) + f(4)\right) \tag{7.13}$$

que, al menos, es exacta para polinomios de grado ≤ 2. Como

$$\int_1^3 x^3\,dx = 20 = \frac{1}{12}\left(0^3 + 22 \times 2^3 + 4^3\right)$$

pero

$$\int_1^3 x^4\,dx = \frac{242}{5} \neq \frac{152}{3} = \frac{1}{12}\left(0^4 + 22 \times 2^4 + 4^4\right)$$

se concluye que la fórmula (7.13) es exacta hasta polinomios de grado 3. □

7.3.2. Fórmulas de integración compuesta

Si se observa la expresión del error en las fórmulas de Newton–Côtes (véase (7.11)) es claro que, para que éste sea pequeño, hace falta tomar un gran número de puntos de interpolación, lo que implica a su vez una fuerte exigencia de regularidad sobre la función f. Por otra parte, los coeficientes de las fórmulas de Newton–Côtes de un elevado número de puntos son poco manejables. Esto lleva a plantearse una estrategia distinta en el empleo de estas fórmulas: en la práctica, para aproximar $\displaystyle\int_a^b f(x)\,dx$ se subdivide el intervalo $[a, b]$ en subintervalos pequeños y se aplica, en cada uno de ellos, una fórmula de Newton–Côtes con un número bajo de puntos.

Vamos a ilustrar esta idea con la *regla de los trapecios*. Según se ha visto, la fórmula del trapecio

$$\int_a^b f(x)\,dx \simeq \frac{f(a) + f(b)}{2}(b - a)$$

tiene como error

$$R_{(a,b)}(f) = -\frac{(b-a)^3}{12} f''(\theta)$$

que, para intervalos de gran longitud puede ser excesivo. Para solucionar este problema dividimos el intervalo $[a,b]$ en subintervalos de longitud $h = \dfrac{b-a}{m}$ con $m \in \mathbb{N}$ obteniendo los puntos

$$x_i = a + ih$$

para $i = 0, 1, \ldots, m$.

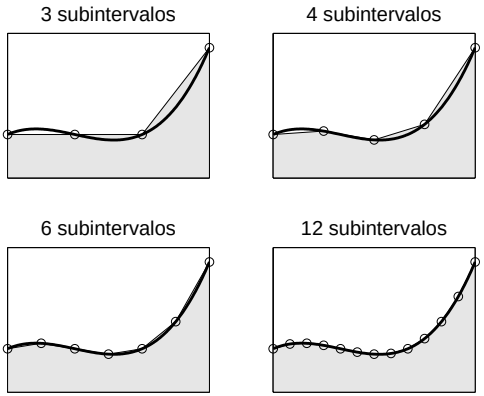


Figura 7.5: Regla de los trapecios.

Aplicando la fórmula del trapecio a cada uno de estos subintervalos se obtiene que

$$\begin{aligned}\int_{x_{i-1}}^{x_i} f(x)\,dx &= \frac{f(x_{i-1}) + f(x_i)}{2}(x_i - x_{i-1}) - \frac{h^3}{12} f''(\theta_i) \\ &= \frac{h}{2}(f(x_{i-1}) + f(x_i)) - \frac{h^3}{12} f''(\theta_i)\end{aligned}$$

para $i = 1, 2, \ldots, m$. Por tanto,

$$\begin{aligned}\int_a^b f(x)\,dx = \sum_{i=1}^{m} \int_{x_{i-1}}^{x_i} f(x)\,dx &= \frac{h}{2} \sum_{i=1}^{m} (f(x_{i-1}) + f(x_i)) - \frac{h^3}{12} \sum_{i=1}^{m} f''(\theta_i) \\ &= \frac{h}{2}\left(f(a) + 2\sum_{i=1}^{m-1} f(x_i) + f(b)\right) - \frac{h^3}{12} \sum_{i=1}^{m} f''(\theta_i) \\ &= \frac{h}{2}\left(f(a) + 2\sum_{i=1}^{m-1} f(x_i) + f(b)\right) - \frac{b-a}{m}\frac{h^2}{12} \sum_{i=1}^{m} f''(\theta_i).\end{aligned}$$

En el caso de que $f'' \in \mathcal{C}([a,b])$ podemos aplicar el teorema de los Valores Intermedios que determina la existencia de $\theta \in [a,b]$ tal que

$$f''(\theta) = \frac{1}{m}\sum_{i=1}^{m} f''(\theta_i). \tag{7.14}$$

De esta forma, hemos demostrado:

Teorema 7.4 (Regla de los trapecios). *Sean $f \in \mathcal{C}^2([a,b])$, $h = \dfrac{b-a}{m}$ con $m \in \mathbb{N}$ y $x_i = a+ih$, $i = 0,1,\ldots,m$. La expresión que toma la regla de los trapecios con m subintervalos es*

$$\int_a^b f(x)\,dx \simeq \frac{h}{2}\left(f(a) + 2\sum_{i=1}^{m-1} f(x_i) + f(b)\right)$$

donde el error cometido viene dado por

$$R_{(a,b)}(f) = -(b-a)\frac{h^2}{12}f''(\theta)$$

para algún $\theta \in [a,b]$. □

Si se conocen los valores de la función f en $2m+1$ puntos equiespaciados $\{a, a+h, \ldots, a+2mh\}$ y aplicamos la fórmula de Simpson en cada uno de los m subintervalos $[a, a+2h]$, $[a+2h, a+4h]$, $\ldots$, $[a+2(m-1)h, a+2mh]$ se obtiene:

Teorema 7.5 (Regla de Simpson compuesta). *Consideremos $f \in \mathcal{C}^4([a,b])$, $h = \dfrac{b-a}{2m}$ con $m \in \mathbb{N}$ y $x_i = a+ih$, $i = 0,1,\ldots,2m$. La regla de Simpson compuesta con m subintervalos se escribe como*

$$\int_a^b f(x)\,dx \simeq \frac{h}{3}\left(f(a) + 4\sum_{i=1}^{m} f(x_{2i-1}) + 2\sum_{i=1}^{m-1} f(x_{2i}) + f(b)\right)$$

donde el error cometido en esta aproximación viene dado por

$$R_{(a,b)}(f) = -(b-a)\frac{h^4}{180}f^{iv)}(\theta)$$

para algún $\theta \in [a,b]$.

DEMOSTRACIÓN. Como tenemos m intervalos de la forma

$$[x_{2(i-1)}, x_{2i}]$$

para $i = 1, 2, \ldots, m$, si aplicamos en cada intervalo $[x_{2(i-1)}, x_{2i}]$ la fórmula de Simpson y denotamos por

$$S = \frac{h}{3}\left(f(a) + 4\sum_{i=1}^{m} f(x_{2i-1}) + 2\sum_{i=1}^{m-1} f(x_{2i}) + f(b)\right),$$

obtenemos

$$\begin{aligned}\int_a^b f(x)\,dx &= \sum_{i=1}^{m}\int_{x_{2(i-1)}}^{x_{2i}} f(x)\,dx \\ &= \frac{h}{3}\sum_{i=1}^{m}\big(f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})\big) - \frac{h^5}{90}\sum_{i=1}^{m} f^{iv)}(\theta_i) \\ &= S - \frac{b-a}{m}\frac{h^4}{180}\sum_{i=1}^{m} f^{iv)}(\theta_i) \\ &= S - (b-a)\frac{h^4}{180}f^{iv)}(\theta)\end{aligned}$$

donde hemos aplicado el teorema de los Valores Intermedios a la función $f^{iv)}$. □

Análogamente, si se conoce el valor que toma la función f en $4m+1$ puntos igualmente espaciados $\{a, a+h, \ldots, a+4mh\}$ y aplicamos la fórmula de Simpson abierta en cada subintervalo $[a, a+4h]$, $[a+4h, a+8h]$, $\ldots$, $[a+4(m-1)h, a+4mh]$ se obtiene:

Teorema 7.6 (Regla de Simpson abierta compuesta). *Supongamos que* $f \in \mathcal{C}^4([a,b])$, $h = \dfrac{b-a}{4m}$ *con* $m \in \mathbb{N}$ *y* $x_i = a + ih$, $i = 0, 1, \ldots, 4m$. *Entonces, la regla de Simpson abierta compuesta con* m *subintervalos viene dada por*

$$\int_a^b f(x)\,dx \simeq \frac{4h}{3}\left(2\sum_{i=1}^{m} f(x_{4i-3}) - \sum_{i=1}^{m} f(x_{4i-2}) + 2\sum_{i=1}^{m} f(x_{4i-1})\right)$$

y el error cometido se expresa como

$$R_{(a,b)}(f) = (b-a)\frac{7h^4}{90}f^{iv)}(\theta)$$

para algún $\theta \in [a, b]$.

DEMOSTRACIÓN. Ahora tenemos m intervalos de la forma

$$[x_{4(i-1)}, x_{4i}]$$

para $i = 1, 2, \ldots, m$. Aplicando en cada intervalo $[x_{4(i-1)}, x_{4i}]$ la fórmula de Simpson abierta y denotando por

$$S = \frac{4h}{3}\left(2\sum_{i=1}^{m} f(x_{4i-3}) - \sum_{i=1}^{m} f(x_{4i-2}) + 2\sum_{i=1}^{m} f(x_{4i-1})\right),$$

se obtiene

$$\begin{aligned}\int_a^b f(x)\,dx &= \sum_{i=1}^{m} \int_{x_{4(i-1)}}^{x_{4i}} f(x)\,dx \\ &= \frac{4h}{3}\sum_{i=1}^{m}\big(2f(x_{4i-3})) - f(x_{4i-2}) + 2f(x_{4i-1})\big) + \frac{14h^5}{45}\sum_{i=1}^{m} f^{iv)}(\theta_i) \\ &= S + \frac{b-a}{m}\frac{7h^4}{90}\sum_{i=1}^{m} f^{iv)}(\theta_i) \\ &= S + (b-a)\frac{7h^4}{90} f^{iv)}(\theta)\end{aligned}$$

donde se ha vuelto a aplicar el teorema de los Valores Intermedios a la función $f^{iv)}$. □

7.3.3. Fórmulas de cuadratura de Gauss

Vamos a estudiar el problema de evaluar la integral definida de una función continua f en un intervalo fijo $[a, b]$, suponiendo que podemos evaluar f en puntos arbitrarios de $[a, b]$. Veremos que algunos de los inconvenientes de las fórmulas de Newton–Côtes se pueden evitar utilizando puntos de interpolación no equidistantes. *K. F. Gauss* (1777–1855) descubrió que, mediante una elección adecuada de los puntos de interpolación, se pueden construir fórmulas de integración que, usando $n+1$ puntos de interpolación, dan el valor exacto de la integral para polinomios de grado menor o igual que $2n + 1$.

Introduciendo una *función peso* $w \in \mathcal{C}([a, b])$ con $w(x) > 0$, $x \in (a, b)$, nos planteamos el problema de encontrar $n + 1$ coeficientes $\{c_0, c_1, \ldots, c_n\}$ tales que

$$\boxed{\int_a^b w(x)f(x)\,dx \simeq \sum_{i=0}^{n} c_i f(x_i)} \tag{7.15}$$

eligiendo los puntos $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$ que maximicen el orden del método. Como ahora están por fijar los coeficientes y las abscisas, al haber un mayor número de grados de libertad, podremos conseguir una mayor efectividad con estas fórmulas. Para llegar a determinar estos $2n + 2$ parámetros, necesitaremos previamente introducir el concepto de familia ortogonal de polinomios.

Definición 7.1. Un conjunto de polinomios $\{P_0, P_1, \ldots, P_m\}$ constituye una familia de *polinomios ortogonales* en el intervalo $[a, b]$ respecto al peso $w(x)$ si:

a) $\partial P_i = i$ para $i = 0, 1, \ldots, m$.

b) $\displaystyle\int_a^b w(x)P_i(x)P_j(x)\,dx = 0$ si $i \neq j$.

Observación 7.7.

1. Si $\{P_0, P_1, \ldots, P_m\}$ es una familia de polinomios ortogonales en el intervalo $[a, b]$ respecto al peso $w(x)$ entonces $\mathcal{B} = \{P_0, P_1, \ldots, P_m\}$ constituye una base del espacio $\mathcal{P}_m$, puesto que los elementos de $\mathcal{B}$ son linealmente independientes y todo polinomio $P \in \mathcal{P}_m$ puede expresarse en la forma

$$P(x) = \sum_{i=0}^{m} \alpha_i P_i(x). \tag{7.16}$$

Además, podemos determinar explícitamente la forma concreta de los coeficientes $\{\alpha_0, \alpha_1, \ldots, \alpha_m\}$. En efecto, si para cada $j \in \{0, 1, \ldots, m\}$ multiplicamos la expresión (7.16) por $w(x)P_j(x)$ e integramos en $[a, b]$, se obtiene que

$$\int_a^b w(x)P(x)P_j(x)\,dx = \sum_{i=0}^{m} \alpha_i \int_a^b w(x)P_i(x)P_j(x)\,dx = \alpha_j \int_a^b w(x)P_j^2(x)\,dx,$$

gracias a la propiedad de ortogonalidad de la familia. Consecuentemente,

$$\alpha_j = \frac{\displaystyle\int_a^b w(x)P(x)P_j(x)\,dx}{\displaystyle\int_a^b w(x)P_j^2(x)\,dx}$$

para $j = 0, 1, \ldots, m$. Los números $\{\alpha_0, \alpha_1, \ldots, \alpha_m\}$, en determinados contextos, se suelen denominar *coeficientes de Fourier* del polinomio $P(x)$ en el intervalo $[a, b]$ respecto al peso $w(x)$.

2. Los polinomios ortogonales en un intervalo $[a, b]$ respecto a una función peso $w(x)$ están unívocamente determinados, salvo constantes multiplicativas.

3. Puede demostrarse que las raíces de cada polinomio de una familia de polinomios ortogonales son, todas ellas, reales y distintas. □

Las familias de polinomios ortogonales que con mayor frecuencia se utilizan en las aplicaciones son las que se muestran en la tabla 7.1. En la tabla 7.2 se recogen los primeros polinomios de cada una de estas familias.

TABLA 7.1:
Polinomios ortogonales

Polinomios de	Intervalo	Peso $w(x)$	Polinomio genérico
Legendre	$[-1, 1]$	1	$P_k(x) = \dfrac{1}{2^k k!}\dfrac{d^k}{dx^k}(x^2-1)^k$
Tchebychev	$[-1, 1]$	$\dfrac{1}{\sqrt{1-x^2}}$	$T_k(x) = \cos(k \arccos x)$
Laguerre	$[0, +\infty)$	e^{-x}	$L_k(x) = e^x \dfrac{d^k}{dx^k}(e^{-x}x^k)$
Hermite	$(-\infty, +\infty)$	e^{-x^2}	$H_k(x) = (-1)^k e^{x^2} \dfrac{d^k}{dx^k}(e^{-x^2})$

TABLA 7.2:
Polinomios de Legendre, Tchebychev, Laguerre y Hermite

Polinomios	$k=0$	$k=1$	$k=2$	$k=3$
$P_k(x)$	1	x	$\dfrac{3x^2-1}{2}$	$\dfrac{x(5x^2-3)}{2}$
$T_k(x)$	1	x	$2x^2-1$	$x(4x^2-3)$
$L_k(x)$	1	$1-x$	$2-4x+x^2$	$-x^3+9x^2-18x+6$
$H_k(x)$	1	$2x$	$4x^2-2$	$4x(2x^2-3)$

El resultado fundamental de esta sección muestra que la fórmula de integración (7.15) es exacta para polinomios de grado $2n+1$ y, además, la condición necesaria y suficiente para ello es que si P_{n+1} es el $(n+1)$–ésimo polinomio de una familia de polinomios ortogonales en $[a, b]$ respecto al peso $w(x)$, las abscisas $\{x_0, x_1, \ldots, x_n\}$ se tomen como sus raíces y los coeficientes $\{c_0, c_1, \ldots, c_n\}$ sean

$$c_i = \frac{1}{P'_{n+1}(x_i)} \int_a^b w(x) \frac{P_{n+1}(x)}{x - x_i}\, dx \tag{7.17}$$

para $i = 0, 1, \ldots, n$.

Teorema 7.7.

a) Si $\{x_0, x_1, \ldots, x_n\}$ *son las raíces del polinomio* P_{n+1} *y* $\{c_0, c_1, \ldots, c_n\}$ *son los coeficientes dados en* (7.17)*, entonces se verifica que*

$$\int_a^b w(x)P(x)\,dx = \sum_{i=0}^{n} c_i P(x_i),\ P \in \mathcal{P}_{2n+1}. \tag{7.18}$$

b) Recíprocamente, si existen $n+1$ *puntos distintos* $\{x_0, x_1, \ldots, x_n\}$ *y coeficientes no nulos* $\{c_0, c_1, \ldots, c_n\}$ *verificando* (7.18) *entonces* $\{x_0, x_1, \ldots, x_n\}$ *son las raíces de* P_{n+1} *y* $\{c_0, c_1, \ldots, c_n\}$ *son los coeficientes dados en* (7.17)*.*

Demostración.

a) Sea $P \in \mathcal{P}_{2n+1}$. Si dividimos P por P_{n+1} obtenemos

$$P(x) = C(x)P_{n+1}(x) + R(x)$$

donde $C, R \in \mathcal{P}_n$. Si $\{P_0, P_1, \ldots, P_n\}$ son los primeros polinomios ortogonales de la familia entonces, por la observación 7.7, podemos escribir

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

siendo

$$a_i = \frac{\displaystyle\int_a^b w(x)C(x)P_i(x)\,dx}{\displaystyle\int_a^b w(x)P_i^2(x)\,dx}.$$

Por hipótesis,

$$P(x_i) = C(x_i)P_{n+1}(x_i) + R(x_i) = R(x_i) \tag{7.19}$$

para $i = 0, 1, \ldots, n$, y

$$\int_a^b w(x)C(x)P_{n+1}(x)\,dx = \sum_{i=0}^{n} a_i \int_a^b w(x)P_i(x)P_{n+1}(x)\,dx = 0.$$

Así pues, se verifica que

$$\int_a^b w(x)P(x)\,dx = \int_a^b w(x)R(x)\,dx. \tag{7.20}$$

Por otra parte, como R coincide con su polinomio de interpolación en los puntos $\{x_0, x_1, \ldots, x_n\}$, utilizando la fórmula de interpolación de Lagrange y la expresión dada en el problema 6.5 de los polinomios básicos de Lagrange, podemos expresar $R(x)$ como

$$R(x) = \sum_{i=0}^{n} R(x_i)L_i(x) = \sum_{i=0}^{n} R(x_i)\frac{\Pi_n(x)}{(x - x_i)\Pi_n'(x_i)}.$$

Como

$$\Pi_n(x) = \prod_{i=0}^{n}(x - x_i) = kP_{n+1}(x)$$

entonces

$$\frac{\Pi_n(x)}{\Pi_n'(x_i)} = \frac{P_{n+1}(x)}{P_{n+1}'(x_i)}$$

para $i = 0, 1, \ldots, n$, por lo que

$$R(x) = \sum_{i=0}^{n} R(x_i)\frac{P_{n+1}(x)}{(x - x_i)P_{n+1}'(x_i)}. \tag{7.21}$$

De esta forma, las relaciones (7.19) y (7.21) hacen que se tenga

$$R(x) = \sum_{i=0}^{n} P(x_i)\frac{P_{n+1}(x)}{(x - x_i)P_{n+1}'(x_i)}$$

que, junto con (7.20), determina la igualdad

$$\int_a^b w(x)P(x)\,dx = \sum_{i=0}^{n} P(x_i)\frac{1}{P_{n+1}'(x_i)}\int_a^b w(x)\frac{P_{n+1}(x)}{x - x_i}\,dx = \sum_{i=0}^{n} c_iP(x_i).$$

b) Sea $Q \in \mathcal{P}_n$. Si $\{P_0, P_1, \ldots, P_n\}$ son los primeros polinomios ortogonales en el intervalo $[a, b]$ respecto al peso $w(x)$, se verifica que

$$Q(x) = \sum_{i=0}^{n} b_iP_i(x)$$

donde, como antes,

$$b_i = \frac{\displaystyle\int_a^b w(x)Q(x)P_i(x)\,dx}{\displaystyle\int_a^b w(x)P_i^2(x)\,dx}$$

para $i = 0, 1, \ldots, n$. Por hipótesis, como $QP_{n+1} \in \mathcal{P}_{2n+1}$, entonces

$$\int_a^b w(x)Q(x)P_{n+1}(x)\,dx = \sum_{i=0}^{n} c_i Q(x_i)P_{n+1}(x_i).$$

Por otra parte, como por la propiedad de ortogonalidad se tiene que

$$\int_a^b w(x)Q(x)P_{n+1}(x)\,dx = \sum_{i=0}^{n} b_i \int_a^b w(x)P_i(x)P_{n+1}(x)\,dx = 0$$

entonces

$$\sum_{i=0}^{n} c_i Q(x_i)P_{n+1}(x_i) = 0.$$

El hecho de que la fórmula anterior sea válida independientemente de los valores $\{Q(x_0), Q(x_1), \ldots, Q(x_n)\}$ obliga a que

$$c_i P_{n+1}(x_i) = 0$$

para $i = 0, 1, \ldots, n$. Como los coeficientes $\{c_0, c_1, \ldots, c_n\}$ son no nulos, se concluye que los puntos $\{x_0, x_1, \ldots, x_n\}$ son las raíces del polinomio P_{n+1}.

Para hallar el valor de los coeficientes $\{c_0, c_1, \ldots, c_n\}$ consideremos ahora un polinomio $P \in \mathcal{P}_{2n+1}$. Dividiendo P por P_{n+1} obtenemos

$$P(x) = C(x)P_{n+1}(x) + R(x)$$

donde $C, R \in \mathcal{P}_n$. Como

$$P(x_i) = C(x_i)P_{n+1}(x_i) + R(x_i) = R(x_i) \tag{7.22}$$

para $i = 0, 1, \ldots, n$, y la fórmula (7.18) es exacta para P y R, se verifica que

$$\int_a^b w(x)P(x)\,dx = \sum_{i=0}^{n} c_i P(x_i) = \sum_{i=0}^{n} c_i R(x_i) = \int_a^b w(x)R(x)\,dx.$$

El hecho de que R coincida con su polinomio de interpolación en los puntos $\{x_0, x_1, \ldots, x_n\}$ permite expresar $R(x)$ en la forma dada en (7.21). Reemplazando este valor en la expresión anterior se obtiene que

$$\begin{aligned}\int_a^b w(x)P(x)\,dx &= \sum_{i=0}^{n} R(x_i)\frac{1}{P'_{n+1}(x_i)}\int_a^b w(x)\frac{P_{n+1}(x)}{x - x_i}\,dx \\ &= \sum_{i=0}^{n} P(x_i)\frac{1}{P'_{n+1}(x_i)}\int_a^b w(x)\frac{P_{n+1}(x)}{x - x_i}\,dx\end{aligned}$$

(véase (7.22)). Así, los coeficientes $\{c_0, c_1, \ldots, c_n\}$ obtenidos son los dados en (7.17). □

Observación 7.8. Las fórmulas de Gauss de $n+1$ puntos no son exactas para polinomios de grado $2n+2$, es decir, dados $n+1$ puntos $\{x_0, x_1, \ldots, x_n\}$ y $n+1$ coeficientes $\{c_0, c_1, \ldots, c_n\}$ cualesquiera, existe un polinomio P de grado $2n+2$ para el que se verifica que

$$\int_a^b w(x)P(x)\,dx \neq \sum_{i=0}^{n} c_i P(x_i).$$

En efecto, basta considerar el polinomio

$$P(x) = \prod_{i=0}^{n} (x - x_i)^2 \in \mathcal{P}_{2n+2}$$

que verifica

$$\sum_{i=0}^{n} c_i P(x_i) = 0 \ \text{ y } \ \int_a^b w(x)P(x)\,dx > 0. \quad \square$$

Observación 7.9.

1. Al igual que ocurría con las fórmulas de Newton–Côtes, las abscisas y coeficientes de las fórmulas de Gauss para las familias de polinomios ortogonales habituales, están tabulados y no es necesario efectuar su cálculo.

2. Puesto que las fórmulas de Gauss de $n+1$ puntos son exactas para polinomios de grado menor o igual que $2n+1$, en particular lo son para el polinomio $P \equiv 1$, de donde se deduce que

$$\sum_{i=0}^{n} c_i = \int_a^b w(x)\,dx.$$

3. Puesto que las familias de polinomios ortogonales están definidas respecto a intervalos y funciones peso particulares, para aplicar las fórmulas de Gauss a un intervalo arbitrario basta con realizar un cambio de variable (que será lineal) y aplicar la fórmula correspondiente a la función integrando dividida por la función peso. $\square$

El siguiente resultado da una estimación del error en la integración gaussiana.

Teorema 7.8. *Si $f \in \mathcal{C}^{2n+2)}([a,b])$ entonces existe $\theta \in (a,b)$ tal que*

$$\int_a^b w(x)f(x)\,dx = \sum_{i=0}^{n} c_i f(x_i) + \frac{f^{2n+2)}(\theta)}{(2n+2)!} \int_a^b w(x)P_{n+1}^2(x)\,dx \tag{7.23}$$

donde las abscisas $\{x_0, x_1, \ldots, x_n\}$ son las raíces del polinomio ortogonal P_{n+1} en el intervalo $[a,b]$ respecto al peso $w(x)$ y $\{c_0, c_1, \ldots, c_n\}$ son los coeficientes dados por (7.17).

DEMOSTRACIÓN. Véase [**Is**–**Ke**]. $\square$

Ejemplo 7.7. Para aproximar, mediante una fórmula de Gauss de tres puntos, la integral

$$\int_{-1}^{1} e^{-x^2}\,dx, \tag{7.24}$$

vamos a utilizar la fórmula de *Gauss–Legendre* y *Gauss–Tchebychev*:

a) Las raíces del polinomio de Legendre P_3 son

$$x_0 = -\sqrt{\frac{3}{5}},\ x_1 = 0 \ \text{ y } \ x_2 = \sqrt{\frac{3}{5}}$$

y los coeficientes valen

$$c_0 = c_2 = \frac{5}{9} \ \text{ y } \ c_1 = \frac{8}{9}.$$

Por tanto, la aproximación de (7.24) obtenida mediante esta fórmula es

$$\begin{aligned}\int_{-1}^{1} e^{-x^2}\,dx &\simeq c_0 e^{-x_0^2} + c_1 e^{-x_1^2} + c_2 e^{-x_2^2}\\ &\simeq \frac{1}{9}\left(5 \times 0.548812 + 8 + 5 \times 0.548812\right)\\ &\simeq 1.498679.\end{aligned} \tag{7.25}$$

b) Las raíces del polinomio de Tchebychev T_3 son

$$x_0 = -\frac{\sqrt{3}}{2},\ x_1 = 0 \ \text{ y } \ x_2 = \frac{\sqrt{3}}{2}$$

y los coeficientes valen ahora

$$c_0 = c_1 = c_2 = \frac{\pi}{3}.$$

La aproximación de (7.24) que se obtiene es

$$\begin{aligned}\int_{-1}^{1} e^{-x^2}\,dx &= \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}}\left(\sqrt{1-x^2}e^{-x^2}\right)dx\\ &\simeq c_0\sqrt{1-x_0^2}e^{-x_0^2} + c_1\sqrt{1-x_1^2}e^{-x_1^2} + c_2\sqrt{1-x_2^2}e^{-x_2^2}\\ &\simeq \frac{\pi}{3}\left(0.236183 + 1 + 0.236183\right)\\ &\simeq 1.541858.\end{aligned} \tag{7.26}$$

Para la fórmula de *Gauss–Legendre* se puede utilizar, de forma sencilla, la expresión (7.23) con $n = 2$ para el estudio del error. Puesto que la derivada sexta de la función $f(x) = e^{-x^2}$ es

$$f^{vi)}(x) = 8(-15 + 90x^2 - 60x^4 + 8x^6)e^{-x^2},\ x \in \mathbb{R},$$

puede comprobarse que

$$\max_{-1\le x\le 1} |f^{vi)}(x)| = -f^{vi)}(0) = 120.$$

Consecuentemente, una cota del error cometido en la aproximación (7.25) viene dada por

$$R(f) \le \frac{120}{6!}\int_{-1}^{1} \frac{x^2(5x^2-3)^2}{4}\,dx = \frac{1}{6}\frac{2}{7} = \frac{1}{21} \simeq 0.047619.$$

Nótese que el valor buscado es

$$\int_{-1}^{1} e^{-x^2}\,dx = 1.4936482\ldots \quad \square$$

7.4. Problemas

7.4.1. Problemas resueltos

7.1. Demostrar que si f es una función suficientemente regular entonces

$$f'(x) = \frac{-3f(x-h) - 10f(x) + 18f(x+h) - 6f(x+2h) + f(x+3h)}{12h} + O(h^4).$$

Solución. Mediante desarrollos de Taylor se tiene que

$$\begin{cases} f(x \pm h) = f(x) \pm hf'(x) + \dfrac{h^2}{2}f''(x) \pm \dfrac{h^3}{6}f'''(x) + \dfrac{h^4}{24}f^{iv)}(x) + O(h^5) \\ f(x+2h) = f(x) + 2hf'(x) + 2h^2f''(x) + \dfrac{4h^3}{3}f'''(x) + \dfrac{2h^4}{3}f^{iv)}(x) + O(h^5) \\ f(x+3h) = f(x) + 3hf'(x) + \dfrac{9h^2}{2}f''(x) + \dfrac{9h^3}{2}f'''(x) + \dfrac{27h^4}{8}f^{iv)}(x) + O(h^5) \end{cases}$$

por lo que

$$-3f(x-h) - 10f(x) + 18f(x+h) - 6f(x+2h) + f(x+3h) = 12hf'(x) + O(h^5)$$

expresión de la que se sigue el resultado buscado. $\square$

7.2. Sea $f \in \mathcal{C}^{n+1}([a,b])$ y P_n el polinomio de interpolación de la función f en los puntos $\{x_0, x_1, \ldots, x_n\} \subset [a,b]$. Probar que para cada $x \in [a,b]$ existe un punto $\eta_x \in [a,b]$ tal que

$$E_n'(x) = f'(x) - P_n'(x) = \frac{f^{n+1)}(\eta_x)}{n!} \prod_{i=0}^{n-1} (x - \xi_i),$$

donde los puntos $\xi_i \in (x_i, x_{i+1})$ son independientes de x.

SOLUCIÓN. Consideremos la función auxiliar

$$g(x) = f(x) - P_n(x),\ x \in [a,b].$$

Como $g \in \mathcal{C}^{n+1}([a,b])$ y

$$g(x_i) = f(x_i) - P_n(x_i) = 0$$

para $i = 0, 1, \ldots, n$, aplicando el teorema de Rolle se llega a que existen n puntos $\xi_i \in (x_i, x_{i+1})$ para $i = 0, 1, \ldots, n-1$, de forma que $g'(\xi_i) = 0$. Por tanto,

$$f'(\xi_i) = P_n'(\xi_i)$$

para $i = 0, 1, \ldots, n-1$. Como $P_n' \in \mathcal{P}_{n-1}$, esto quiere decir que P_n' es el polinomio de interpolación de la función f' en los n puntos distintos $\{\xi_0, \xi_1, \ldots, \xi_{n-1}\}$. Aplicando la fórmula del error en la interpolación de Lagrange se tiene que para cada $x \in [a,b]$ existe η_x perteneciente al mínimo intervalo cerrado que contiene a los puntos $\{\xi_0, \xi_1, \ldots, \xi_{n-1}, x\}$ tal que

$$f'(x) - P_n'(x) = \frac{f^{n+1)}(\eta_x)}{n!} \prod_{i=0}^{n-1} (x - \xi_i)$$

(nótese que $f^{n+1)}$ es la derivada n–ésima de la función f'). □

7.3. Método de diferencias finitas para problemas de contorno. A partir de una función dada $f : [0,1] \to \mathbb{R}$ se considera la función $u : [0,1] \to \mathbb{R}$ solución de la *ecuación diferencial*

$$-u''(x) + u(x) = f(x),\ x \in (0,1) \tag{7.27}$$

verificando además las *condiciones de contorno* $u(0) = u(1) = 0$. Fijado $n \in \mathbb{N}$, se consideran $h = \dfrac{1}{n+1}$ y los puntos $x_i = ih$, $i = 0, 1, \ldots, n+1$.

a) Para cada punto $x = x_i$ con $i \in \{1, 2, \ldots, n\}$ utilizar el valor aproximado

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

y reemplazarlo en la ecuación (7.27) para obtener un sistema lineal de n ecuaciones con n incógnitas (que aproximará la ecuación diferencial anterior) con matriz tridiagonal. Demostrar que el sistema lineal asociado tiene una única solución.

b) Aplicar el método anterior para obtener una aproximación de la solución del problema de contorno

$$(\mathcal{P}) \begin{cases} -u''(x) + u(x) = (1 + \pi^2) \operatorname{sen} \pi x, & 0 < x < 1 \\ u(0) = u(1) = 0 \end{cases}$$

con $n = 2, 4, 6, 10$ y comparar los resultados obtenidos con la solución exacta del problema $(\mathcal{P})$, que es $u(x) = \operatorname{sen} \pi x,\ x \in [0, 1]$.

SOLUCIÓN.

a) Para cada $i \in \{1, 2, \ldots, n\}$ se verifica que

$$\begin{aligned} f(x_i) = -u''(x_i) + u(x_i) &\simeq -\frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2} + u(x_i) \\ &= \frac{-u(x_{i-1}) + (2 + h^2)u(x_i) - u(x_{i+1})}{h^2}. \end{aligned}$$

Por tanto, en términos matriciales, podemos considerar el sistema lineal aproximado

$$Au = f$$

donde

$$A = \frac{1}{h^2} \begin{pmatrix} 2 + h^2 & -1 & & & & \\ -1 & 2 + h^2 & -1 & & & \\ & -1 & 2 + h^2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 + h^2 & -1 \\ & & & & -1 & 2 + h^2 \end{pmatrix},$$

$$u = (u_1, u_2, \ldots, u_n)^{\mathrm{T}} \ \text{ y } \ f = (f_1, f_2, \ldots, f_n)^{\mathrm{T}},$$

siendo

$$f_i = f(x_i)$$

para $i = 1, 2, \ldots, n$. De esta forma

$$u_j \simeq u(x_j)$$

para cada $j = 0, 1, \ldots, n+1$ (denotando $u_0 = u_{n+1} = 0$). Como la matriz A es de diagonal estrictamente dominante entonces A es inversible, por lo que el sistema $Au = f$ tiene una única solución.

b) La solución del sistema $Au = f$ es, en cada uno de los casos,

$$u^{(2)} \simeq (0.9414, 0.9414)^{\mathrm{T}},$$

$$u^{(4)} \simeq (0.6056, 0.9799, 0.9799, 0.6056)^{\mathrm{T}},$$

$$u^{(6)} \simeq (0.4406, 0.7938, 0.9899, 0.9899, 0.7938, 0.4406)^{\mathrm{T}}$$

y

$$u^{(10)} \simeq (0.2835, 0.5440, 0.7604, 0.9153, 0.9960, 0.9960, 0.9153, 0.7604, 0.5440, 0.2835)^{\mathrm{T}}.$$

La figura 7.6 muestra los resultados aproximados obtenidos con el método anterior y la solución exacta del problema $(\mathcal{P})$. □

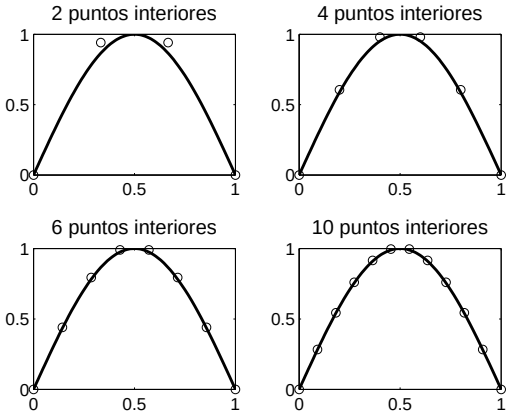


Figura 7.6: Soluciones exacta y aproximadas del problema $(\mathcal{P})$.

7.4. Se considera la fórmula de integración

$$\int_0^1 f(x)\,dx \simeq A(f(x_0) + f(x_1)).$$

Hallar el valor de A, x_0 y x_1 para que la fórmula sea exacta para polinomios del mayor grado posible. ¿Cuál es éste?

SOLUCIÓN. Impongamos que la fórmula de integración sea exacta para polinomios de grado n, esto es,

$$\frac{1}{n+1} = \int_0^1 x^n \, dx = A(x_0^n + x_1^n) \tag{7.28}$$

y vayamos dando valores, en forma creciente, a n :

$$\begin{cases} n = 0 \;\Rightarrow\; 2A = 1 \;\Rightarrow\; A = \dfrac{1}{2} \\ n = 1 \;\Rightarrow\; \dfrac{1}{2}(x_0 + x_1) = \dfrac{1}{2} \;\Rightarrow\; x_1 = 1 - x_0 \\ n = 2 \;\Rightarrow\; \dfrac{1}{2}(x_0^2 + x_1^2) = \dfrac{1}{3} \;\Rightarrow\; x_0^2 + x_1^2 = \dfrac{2}{3}. \end{cases} \tag{7.29}$$

De esta forma, a partir de (7.29), se obtiene la ecuación

$$2x_0^2 - 2x_0 + \frac{1}{3} = 0$$

que admite las soluciones

$$x_0 = \frac{1}{2} \pm \frac{\sqrt{3}}{6}.$$

Utilizando nuevamente (7.29) obtenemos que

$$x_1 = \frac{1}{2} \mp \frac{\sqrt{3}}{6}.$$

La simetría de la fórmula de integración con respecto a los coeficientes x_0 y x_1 nos permite tomar

$$x_0 = \frac{1}{2} + \frac{\sqrt{3}}{6} \;\text{ y }\; x_1 = \frac{1}{2} - \frac{\sqrt{3}}{6}.$$

Para comprobar si la fórmula de integración que hemos obtenido para estos valores de A, x_0 y x_1 es también exacta para polinomios de grado 3, verificamos si se cumple la relación (7.28) para el valor $n = 3$, obteniendo

$$A(x_0^3 + x_1^3) = \frac{1}{2}\left(\left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right)^3 + \left(\frac{1}{2} - \frac{\sqrt{3}}{6}\right)^3\right) = \frac{1}{2}\left(2\frac{1}{2^3} + 6\frac{1}{2}\frac{1}{12}\right) = \frac{1}{4}$$

puesto que

$$(a+b)^3 + (a-b)^3 = 2a^3 + 6ab^2, \; a, b \in \mathbb{R}.$$

Consecuentemente, se verifica la igualdad (7.28) para $n = 3$ con los valores de A, x_0 y x_1 obtenidos. No obstante, puede comprobarse que no se da la igualdad cuando $n = 4$ por lo que la fórmula de integración aproximada

$$\int_0^1 f(x)\,dx \simeq \frac{1}{2}\left(f\left(\frac{1}{2}+\frac{\sqrt{3}}{6}\right)+f\left(\frac{1}{2}-\frac{\sqrt{3}}{6}\right)\right)$$

sólo es exacta para polinomios de grado menor o igual que 3. □

7.5. Demostrar que la regla de los trapecios está bien condicionada, en términos absolutos, respecto a los errores de redondeo en las evaluaciones de la función.

SOLUCIÓN. Sean $\widetilde{f}(x_i)$ perturbaciones de los valores $f(x_i)$, es decir,

$$\widetilde{f}(x_i) = f(x_i) + \varepsilon_i$$

para $i = 0, 1, \ldots, m$, con $x_0 = a$ y $x_m = b$. El error absoluto cometido en la fórmula de los trapecios es

$$\begin{aligned}\Sigma &= \left|\frac{h}{2}\left(f(a)+2\sum_{i=1}^{m-1}f(x_i)+f(b)\right)-\frac{h}{2}\left(\widetilde{f}(a)+2\sum_{i=1}^{m-1}\widetilde{f}(x_i)+\widetilde{f}(b)\right)\right| \\ &= \frac{h}{2}\left|\varepsilon_0+2\sum_{i=1}^{m-1}\varepsilon_i+\varepsilon_m\right| \le \frac{h}{2}\left(|\varepsilon_0|+2\sum_{i=1}^{m-1}|\varepsilon_i|+|\varepsilon_m|\right) \\ &\le \frac{h}{2}\left(\varepsilon+2(m-1)\varepsilon+\varepsilon\right) = \frac{h}{2}2m\varepsilon = hm\varepsilon\end{aligned}$$

donde

$$\varepsilon = \max_{0\le i\le m}|\varepsilon_i|.$$

Ahora bien, como $h = \dfrac{b-a}{m}$, entonces se obtiene que

$$\Sigma \le hm\varepsilon = (b-a)\varepsilon$$

de donde se deduce que el error que se produce cuando se aplica la fórmula de los trapecios tras cometer errores en las evaluaciones de la función, es del orden de dichos errores. □

7.6. Regla de los trapecios abierta.

a) Deducir la expresión de la fórmula del trapecio abierta.

b) Determinar la expresión de la fórmula anterior compuesta (*regla de los trapecios abierta*).

c) ¿Qué valor se obtiene si se utiliza la fórmula de *b*) con 100 subintervalos para aproximar $\displaystyle\int_{-5}^{5} |x|\,dx$? ¿Qué ocurre si se toman 99 subintervalos?

SOLUCIÓN.

a) Consideramos

$$h = \frac{b-a}{3} \quad \text{y} \quad x_i = a + ih$$

para $i = 0, 1, 2, 3$. Por definición

$$P_1(x) = f(x_1) + (x - x_1) f[x_1, x_2] = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1)$$

es el polinomio de interpolación de la función f en los puntos $\{x_1, x_2\}$. De esta forma, teniendo en cuenta que $h = \dfrac{b-a}{3}$, se verifica que

$$\begin{aligned}
\int_a^b f(x)\,dx \simeq \int_a^b P_1(x)\,dx &= f(x_1)x + \frac{f(x_2) - f(x_1)}{x_2 - x_1}\frac{(x - x_1)^2}{2}\bigg|_a^b \\
&= f(x_1)(b-a) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}\frac{(b - x_1)^2 - (a - x_1)^2}{2} \\
&= 3hf(x_1) + \frac{f(x_2) - f(x_1)}{h}\frac{4h^2 - h^2}{2} \\
&= 3hf(x_1) + \frac{f(x_2) - f(x_1)}{h}\frac{3h^2}{2} \\
&= 3h\left(f(x_1) + \frac{f(x_2) - f(x_1)}{2}\right) = \frac{3h}{2}\left(f(x_1) + f(x_2)\right).
\end{aligned}$$

b) La fórmula abierta de Newton–Côtes de dos puntos compuesta con m subintervalos para

$$h = \frac{b-a}{3m} \quad \text{y} \quad x_i = a + ih$$

para $i = 0, 1, \ldots, 3m$, viene dada por

$$\begin{aligned}
\int_a^b f(x)\,dx = \sum_{i=1}^{m} \int_{x_{3(i-1)}}^{x_{3i}} f(x)\,dx &\simeq \sum_{i=1}^{m} \left(\frac{3h}{2}\left(f(x_{3i-2}) + f(x_{3i-1})\right)\right) \\
&= \frac{3h}{2}\left(\sum_{i=1}^{m} f(x_{3i-2}) + \sum_{i=1}^{m} f(x_{3i-1})\right).
\end{aligned}$$

c) Si se utilizan 100 subintervalos, los puntos a considerar son

$$x_i = -5 + \frac{i}{30}$$

para $i = 0, 1, \ldots, 300$. Por tanto, en cada subintervalo de la forma $[x_{3(i-1)}, x_{3i}]$, $i = 1, 2, \ldots, 100$, la función f es un polinomio de grado 1. Concretamente

$$f(x) = \begin{cases} -x, \ x \in [x_{3(i-1)}, x_{3i}] & \text{si} \quad 1 \leq i \leq 50 \\ x, \ x \in [x_{3(i-1)}, x_{3i}] & \text{si} \quad 51 \leq i \leq 100. \end{cases}$$

Como la fórmula es exacta para polinomios lineales, el error cometido es nulo. En cambio, con 99 subintervalos, los puntos son

$$x_i = -5 + \frac{10i}{297}$$

para $i = 0, 1, \ldots, 297$, por lo que ahora

$$f(x) = \begin{cases} -x, \ x \in [x_{3(i-1)}, x_{3i}] & \text{si} \quad 1 \leq i \leq 49 \\ |x|, \ x \in [x_{147}, x_{150}] & \\ x, \ x \in [x_{3(i-1)}, x_{3i}] & \text{si} \quad 51 \leq i \leq 99 \end{cases}$$

donde

$$x_{147} = -\frac{5}{99} \ \text{ y } \ x_{150} = \frac{5}{99}.$$

Consecuentemente el error es nulo en todos los subintervalos de la forma $[x_{3(i-1)}, x_{3i}]$, $i = 1, 2, \ldots, 99$, salvo en el central $[x_{147}, x_{150}]$ obteniéndose, por un lado,

$$\int_{-\frac{5}{99}}^{\frac{5}{99}} |x| dx = 2 \int_0^{\frac{5}{99}} x \, dx = \left(\frac{5}{99}\right)^2 = \frac{25}{9801} \simeq 0.002550$$

y, por otro,

$$\int_{-\frac{5}{99}}^{\frac{5}{99}} P_1(x) \, dx = \frac{5}{297} \frac{10}{99} = \frac{50}{29403} \simeq 0.001700,$$

puesto que

$$P_1(x) \equiv \frac{5}{297}$$

es el polinomio de interpolación de la función $f(x) = |x|$ en los puntos

$$\left\{ x_{148} = -\frac{5}{297}, x_{149} = \frac{5}{297} \right\}$$

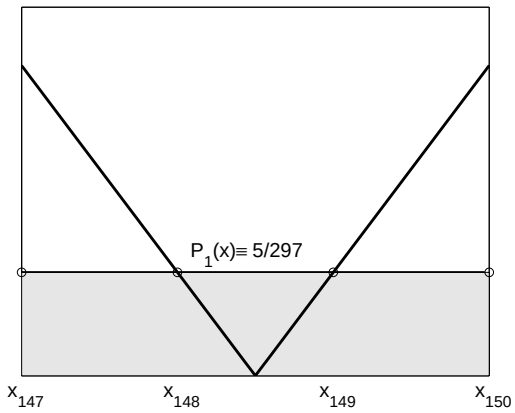


Figura 7.7: Aproximación en el intervalo central.

(véase la figura 7.7). Así, el error que se comete en la aproximación es

$$R_{(-5,5)}(f) = \frac{25}{9801} - \frac{50}{29403} = \frac{25}{29403} \simeq 8.502534 \times 10^{-4}. \quad \square$$

7.7. Aplicar la regla de Simpson compuesta a la integral

$$\int_1^x \frac{dt}{t}$$

para obtener una aproximación del logaritmo neperiano de 2, determinando el número m de subintervalos necesario para que el error cometido en esa aproximación sea inferior a 10^{-3}.

SOLUCIÓN. Por el teorema 7.5 sabemos que

$$\int_1^2 f(x)\,dx = \frac{h}{3}\left(f(1) + 4\sum_{i=1}^{m} f(x_{2i-1}) + 2\sum_{i=1}^{m-1} f(x_{2i}) + f(2)\right) - R_{(1,2)}(f)$$

siendo

$$R_{(1,2)}(f) = (2-1)\frac{h^4}{180} f^{iv)}(\theta)$$

para algún $\theta \in (1,2)$. Como

$$h = \frac{1}{2m} \;\text{ y }\; f^{iv)}(t) = \frac{24}{t^5},\ t \in (1,2)$$

entonces

$$R_{(1,2)}(f) = \frac{1}{2880m^4}\frac{24}{\theta^5} = \frac{1}{120m^4\theta^5} < \frac{1}{120m^4}$$

ya que $1 < \theta < 2$. De esta forma, como

$$\frac{1}{120m^4} \leq 10^{-3} \Leftrightarrow m \geq \sqrt[4]{\frac{25}{3}} \simeq 1.6990,$$

basta tomar $m = 2$ para que $R_{(1,2)}(f) < 10^{-3}$. Para este valor $m = 2$ las abscisas que intervienen en la fórmula de Simpson cerrada compuesta son

$$x_i = 1 + ih \ \text{ con } \ h = \frac{1}{4},$$

es decir,

$$\left\{x_0 = 1, x_1 = \frac{5}{4}, x_2 = \frac{3}{2}, x_3 = \frac{7}{4}, x_4 = 2\right\}$$

por lo que aplicando la fórmula de integración aproximada se obtiene que

$$\begin{aligned}
\ln 2 = \int_1^2 \frac{dt}{t} &\simeq \frac{h}{3}\left(f(1) + 4\left(f(x_1) + f(x_3)\right) + 2f(x_2) + f(2)\right) \\
&= \frac{1}{12}\left(f(1) + 4\left(f\left(\frac{5}{4}\right) + f\left(\frac{7}{4}\right)\right) + 2f\left(\frac{3}{2}\right) + f(2)\right) \\
&= \frac{1}{12}\left(1 + 4\left(\frac{4}{5} + \frac{4}{7}\right) + 2\frac{2}{3} + \frac{1}{2}\right) = \frac{1747}{2520} \\
&\simeq 0.693253968253.
\end{aligned}$$

Compárese este valor aproximado con $\ln 2 = 0.693147137704\ldots$ □

7.8. Hallar el valor de la aproximación que se obtiene al calcular

$$\int_{-4}^{4} |x-2|^3(1 - \operatorname{sen}\pi x)\, dx$$

mediante la regla de Simpson compuesta para cuatro subintervalos.

Solución. Consideremos los cuatro subintervalos

$$[-4,-2],\ [-2,0],\ [0,2],\ [2,4]$$

y los puntos

$$x_i = -4 + ih$$

para $i = 0, 1, \ldots, 8$ con $h = 1$. Como la función

$$f(x) = |x-2|^3(1 - \operatorname{sen} \pi x)$$

cumple que

$$f(k) = |k-2|^3,\ k = 0, \pm 1, \pm 2, \pm 3, \pm 4,$$

se tiene que f y el polinomio

$$P(x) = (2-x)^3$$

comparten ambos el mismo polinomio de interpolación en cada una de las tres ternas de puntos $\{x_0 = -4, x_1 = -3, x_2 = -2\}$, $\{x_2 = -2, x_3 = -1, x_4 = 0\}$ y $\{x_4 = 0, x_5 = 1, x_6 = 2\}$; de la misma forma, f y el polinomio

$$Q(x) = (x-2)^3$$

tienen el mismo polinomio de interpolación en $\{x_6 = 2, x_7 = 3, x_8 = 4\}$. Por lo tanto, la aproximación que da la fórmula de Simpson compuesta para cuatro subintervalos es (recuérdese que esta fórmula es exacta para polinomios de grado menor o igual que tres)

$$\begin{aligned}\int_{-4}^{4} f(x)\,dx &\simeq \int_{-4}^{2} (2-x)^3\,dx + \int_{2}^{4} (x-2)^3\,dx \\ &= -\left.\frac{(2-x)^4}{4}\right|_{-4}^{2} + \left.\frac{(x-2)^4}{4}\right|_{2}^{4} = \frac{6^4 + 2^4}{4} = 328. \quad \square\end{aligned}$$

7.9. Suponiendo que los puntos $\{x_0, x_1, \ldots, x_n\} \subset [-a, a]$ con n par están distribuidos simétricamente respecto al origen y que la fórmula

$$\int_{-a}^{a} f(x)\,dx \simeq \sum_{i=0}^{n} c_i f(x_i) \tag{7.30}$$

es exacta para polinomios de grado menor o igual que n, demostrar que también es exacta para polinomios de grado $n+1$.

SOLUCIÓN. Por ser la fórmula exacta para polinomios de grado menor o igual que n basta probar que

$$\int_{-a}^{a} x^{n+1}\,dx = \sum_{i=0}^{n} c_i x_i^{n+1}.$$

Sea $P_n \in \mathcal{P}_n$ el polinomio de interpolación de la función $f(x) = x^{n+1}$ en los puntos $\{x_0, x_1, \ldots, x_n\}$. Como la fórmula (7.30) es exacta para P_n, se tiene que

$$\int_{-a}^{a} P_n(x)\,dx = \sum_{i=0}^{n} c_i P_n(x_i) = \sum_{i=0}^{n} c_i x_i^{n+1}.$$

Ahora bien, como n es par entonces x^{n+1} es una función impar con respecto al origen y, por el problema 6.6, también lo es P_n. Así,

$$\int_{-a}^{a} x^{n+1}\,dx = 0 = \int_{-a}^{a} P_n(x)\,dx = \sum_{i=0}^{n} c_i x_i^{n+1}$$

como queríamos demostrar. □

7.10. Demostrar que los coeficientes $\{c_0, c_1, \ldots, c_n\}$ de las fórmulas de Gauss de $n+1$ puntos distintos $\{x_0, x_1, \ldots, x_n\}$ son positivos.

SOLUCIÓN. Para cada $j \in \{0, 1, \ldots, n\}$ consideramos el polinomio

$$q_j(x) = \prod_{\substack{k=0 \\ k\neq j}}^{n} (x - x_k)^2.$$

Como q_j es no negativo, $q_j \in \mathcal{P}_{2n}$ y las fórmulas de Gauss son exactas para polinomios de grado menor o igual que $2n+1$ se tiene que

$$\begin{aligned} 0 < \int_a^b w(x) q_j(x)\,dx &= \sum_{i=0}^{n} c_i q_j(x_i) \\ &= \sum_{i=0}^{n} c_i \prod_{\substack{k=0 \\ k\neq j}}^{n} (x_i - x_k)^2 = c_j \prod_{\substack{k=0 \\ k\neq j}}^{n} (x_j - x_k)^2, \end{aligned}$$

por lo que $c_j > 0$. □

7.4.2. Problemas propuestos

7.11. Determinar la fórmula abierta de Newton–Côtes de un punto (denominada *fórmula del punto medio*). Hallar la expresión de la *regla del punto medio compuesta*.

7.12. Sea $\Delta = \{x_0, x_1, \ldots, x_n\}$ una partición equiespaciada de $[a, b]$ con

$$h = \frac{b-a}{n} \quad \text{y} \quad x_i = a + ih$$

para $i = 0, 1, \ldots, n$. Para aproximar $\displaystyle\int_a^b f(x)\,dx$ se puede considerar la interpolación mediante funciones *spline* cúbicas. Encontrar una fórmula de integración aproximada en función de $f(x_i)$ y los momentos M_i de una función *spline* cúbica $S_\Delta(y, \cdot)$ siendo

$$y_i = f(x_i)$$

para $i = 0, 1, \ldots, n$.

7.13. Dado $n \in \mathbb{N}$ se consideran $h = \dfrac{1}{n}$, $x_i = ih$ para $i = 0, 1, \ldots, n$ y la función

$$f_n(x) = x^n \cos(2\pi n x).$$

Hallar el valor de la aproximación de $\displaystyle\int_0^1 f_n(x)\,dx$ que se obtiene utilizando la fórmula de Newton–Côtes cerrada de $n+1$ puntos.

7.14. Se considera la fórmula de integración aproximada

$$\int_{-1}^{1} f(x)\,dx = A\left(f(x_1) + f(0) + f(x_2)\right).$$

a) Determinar la constante $A \in \mathbb{R}$ y los puntos $x_1, x_2 \in [-1, 1]$ para que sea exacta para polinomios del mayor grado posible. ¿Cuál es éste?

b) ¿Son $\{x_1, 0, x_2\}$ las raíces del polinomio de Legendre de grado 3? Razonar la respuesta.

7.15. Determinar los valores de las constantes α, β y γ que hacen que la fórmula de integración

$$\int_0^3 f(x)\,dx \simeq \alpha f(0) + \beta f(1) + \gamma f(3)$$

sea exacta para polinomios del mayor grado posible. ¿Cuál es éste?

7.16. Hallar la aproximación que se obtiene de $\displaystyle\int_a^b f(x)\,dx$ cuando se considera la integral de la interpolación lineal a trozos de f en una partición equiespaciada.

7.17. Demostrar el resultado análogo al del problema 7.5 para la regla de Simpson compuesta.

7.5. Prácticas

7.1. Programar una función en `MATLAB` que implemente la regla de los trapecios. Calcular con este programa diversas integrales de valor conocido y comparar los resultados hallados con los exactos.

7.2. Programar en `MATLAB` la regla de Simpson compuesta como una función. Calcular con este programa diversas integrales de valor conocido y comparar los resultados hallados con los exactos.

7.3. Utilizando las fórmulas de Simpson cerrada y abierta aproximar

$$\int_0^1 \frac{\operatorname{sen} x}{x}\, dx$$

dando una cota del error cometido en cada caso.

7.4. Utilizar las fórmulas cerradas de Newton–Côtes de 2 y 3 puntos y las fórmulas abiertas de Newton–Côtes de $1, 2$ y 3 puntos, para aproximar

$$\int_0^{\frac{\pi}{4}} \operatorname{sen} x\, dx,$$

calculando el error cometido y comparando los resultados obtenidos con cada fórmula.

7.5. Utilizar los programas de las prácticas 7.1 y 7.2 para obtener

$$\int_0^1 e^{-x^2}\, dx$$

con un error inferior a 10^{-4} (estudiar previamente, en cada caso, cuál debe ser el número de subintervalos).

7.6. Escribir un programa en `MATLAB` para las fórmulas de Gauss–Legendre y Gauss–Tchebychev de tres puntos. Utilizarlos para aproximar

$$\int_{-1}^1 \frac{\operatorname{sen} x}{x}\, dx.$$

8 Resolución de ecuaciones no lineales

8.1. Introducción

En una gran variedad de situaciones surge, de manera natural, el problema de hallar las raíces de una ecuación. La experiencia muestra que ésta no es, en general, una tarea sencilla. Así, por ejemplo, para ecuaciones como

$$e^{-x} - 2x = 0 \ \text{ o } \ \tan x - x = 0,$$

podemos demostrar que tienen raíces reales (una única la primera y una cantidad infinita numerable la segunda) pero no existe un método para calcularlas de forma exacta. Otro aspecto relevante es que, en muchas ocasiones, los coeficientes de las ecuaciones sólo se conocen de forma aproximada, por lo que carecería de sentido su cálculo exacto (en el hipotético caso de que se pudieran hallar las raíces exactas). Este tipo de argumentos lleva a plantearse el estudio de métodos, necesariamente iterativos, para el cálculo de las raíces de una ecuación

$$F(x) = 0 \ \text{ en } \ [a, b]$$

que permitan aproximarlas con el grado de precisión que se desee.

Los comentarios anteriores no sólo son válidos para ecuaciones trascendentes, sino también para ecuaciones algebraicas. Así, no es un problema sencillo el de encontrar las raíces de la ecuación

$$x^{56} - 5x^{23} + 18x^{15} - 4x^8 + 32x^7 - x^5 + 4x^3 + 128 = 0.$$

Los métodos iterativos que veremos en este capítulo son también de aplicación a las ecuaciones algebraicas; no obstante, en el capítulo 10 veremos técnicas específicas, adaptadas al cálculo de las raíces de un polinomio.

A continuación establecemos el marco en el que vamos a trabajar:

Definición 8.1. Sea $F : D \subset \mathbb{R} \to \mathbb{R}$. Un punto $\xi \in D$ es una *raíz* de la ecuación

$$F(x) = 0 \ \text{ en } \ D \tag{8.1}$$

si $F(\xi) = 0$. También se dice que ξ es un *cero* de la función F. □

En todo lo que sigue supondremos que la ecuación (8.1) tiene las *raíces aisladas*, es decir, para cada raíz de la ecuación existe un entorno que no contiene más raíces de la ecuación. No todas las ecuaciones tienen las raíces aisladas como se muestra en el siguiente ejemplo:

Ejemplo 8.1. Consideremos la función $F : \mathbb{R} \to \mathbb{R}$ dada por

$$F(x) = \begin{cases} x \operatorname{sen} \dfrac{1}{x} & \text{si} \quad x \neq 0 \\ 0 & \text{si} \quad x = 0 \end{cases}$$

cuya gráfica viene dada en la figura 8.1.

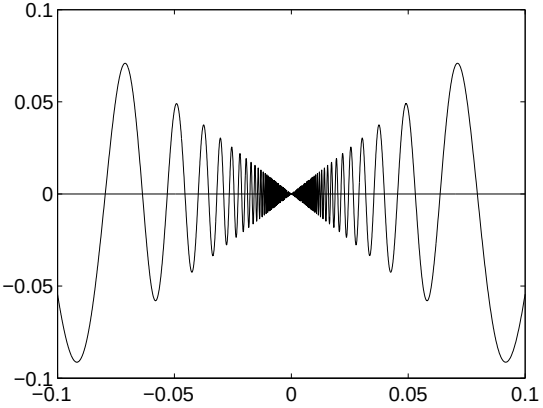


Figura 8.1: Función $F(x) = x \operatorname{sen} \dfrac{1}{x}$.

Como

$$\lim_{x \to 0} F(x) = \lim_{x \to 0} x \operatorname{sen} \frac{1}{x} = 0 = F(0)$$

entonces $F \in \mathcal{C}(\mathbb{R})$. Por la propia definición de F, $x = 0$ es una raíz de $F(x) = 0$. Por otra parte, para cada $\varepsilon > 0$ existe $\eta_\varepsilon \in (-\varepsilon, \varepsilon) \backslash \{0\}$ tal que $F(\eta_\varepsilon) = 0$ (es decir, cualquier entorno de $x = 0$ contiene otras raíces de la ecuación). En efecto, como

$$\lim_{n \to +\infty} \frac{1}{n\pi} = 0,$$

para todo $\varepsilon > 0$ existe $n_\varepsilon \in \mathbb{N}$ tal que

$$0 < \frac{1}{n_\varepsilon \pi} < \varepsilon.$$

Por tanto, tomando

$$\eta_\varepsilon = \frac{1}{n_\varepsilon \pi}$$

se verifica que $\eta_\varepsilon \in (0, \varepsilon)$ y

$$F(\eta_\varepsilon) = \frac{1}{n_\varepsilon \pi} \operatorname{sen} n_\varepsilon \pi = 0. \quad \square$$

Consideraremos dos etapas en el cálculo aproximado de las raíces reales aisladas de la ecuación

$$F(x) = 0 \ \text{ en } \ [a, b]. \tag{8.2}$$

a) *Separación de raíces*: se establecen subintervalos de $[a, b]$ que contengan una y sólo una raíz de la ecuación (8.2). Para ello las herramientas fundamentales van a ser el teorema de Bolzano (para asegurar la existencia de raíces) y el teorema de Rolle (para acotar el número de raíces que puede haber). Así, son formas típicas de argumentar las siguientes:

   i) Si $F \in \mathcal{C}([a, b])$, F es derivable en (a, b), $F(a)F(b) < 0$ y F' tiene un signo constante en (a, b) entonces F tiene una única raíz ξ en (a, b). En efecto, el teorema de Bolzano garantiza la existencia de $\xi \in (a, b)$ tal que $F(\xi) = 0$ y, por otra parte, el hecho de que F sea estrictamente creciente o decreciente en (a, b) determina, gracias al teorema de Rolle, que la raíz ξ sea única.

   ii) Si F' sólo se anula en n puntos la función F tendrá, a lo sumo, $n + 1$ raíces.

   iii) Si $F \in \mathcal{C}^2([a, b])$ es tal que F'' tiene un signo constante en $[a, b]$ entonces F tiene, a lo sumo, dos raíces reales en $[a, b]$.

b) En cada uno de estos intervalos se calcula la raíz ξ de la ecuación mediante un método iterativo como límite de una sucesión $\{x_n\}_{n=0}^{\infty}$ que converge a ξ. De esta forma, al igual que ocurría en los sistemas lineales estudiados en el capítulo 5, deberemos tomar como aproximación de la solución ξ un elemento x_n de la sucesión próximo a ella.

En las siguientes secciones se estudian los métodos más utilizados para resolver este tipo de problemas, en el supuesto de que se ha llevado a cabo la primera etapa y, por tanto, F tiene una única raíz ξ en $[a, b]$.

8.2. Método de la bisección

Sea $F \in \mathcal{C}([a,b])$ tal que

$$F(a)F(b) < 0.$$

Por el teorema de Bolzano existe, al menos, $\xi \in (a,b)$ tal que $F(\xi) = 0$. Suponiendo que hemos separado las raíces de la ecuación y que ξ es la única raíz de

$$F(x) = 0 \ \text{ en } \ [a,b], \tag{8.3}$$

dividimos el intervalo $[a,b]$ por la mitad. Así, pueden presentarse dos casos:

a) $F\left(\dfrac{a+b}{2}\right) = 0$. Entonces $\xi = \dfrac{a+b}{2}$.

b) $F\left(\dfrac{a+b}{2}\right) \neq 0$. En esta situación se elige el intervalo

$$\left[a, \frac{a+b}{2}\right] \ \text{ o } \ \left[\frac{a+b}{2}, b\right]$$

en cuyos extremos la función F toma signos opuestos para, de esta forma, poder seguir aplicando el teorema de Bolzano (obsérvese que, como ξ es la única raíz de F en $[a,b]$, sólo uno de estos dos intervalos tendrá la propiedad de que F cambie de signo en los extremos del mismo). Denotando a este intervalo por $[a_1, b_1]$, lo dividimos por la mitad y procedemos de igual forma.

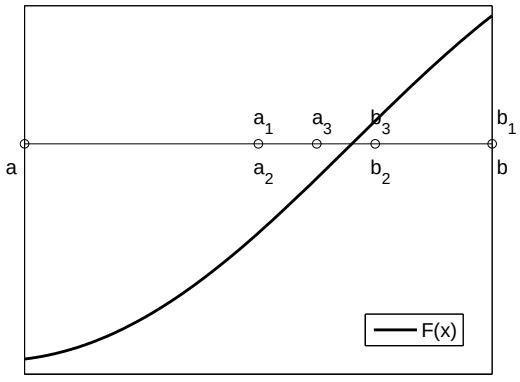


Figura 8.2: Método de la bisección.

Así, reiterando este proceso, o bien obtenemos la raíz exacta de (8.3) o bien una sucesión de intervalos cerrados $\{[a_n, b_n]\}_{n=1}^{\infty}$ *encajados*, es decir,

$$a \leq a_1 \leq \cdots \leq a_n \leq a_{n+1} \leq b_{n+1} \leq b_n \leq \cdots \leq b_1 \leq b,\ n \in \mathbb{N}, \tag{8.4}$$

en los cuales

$$F(a_n)F(b_n) < 0,\ n \in \mathbb{N} \tag{8.5}$$

y

$$b_n - a_n = \frac{b-a}{2^n},\ n \in \mathbb{N}. \tag{8.6}$$

La propiedad (8.4) determina que $\{a_n\}_{n=1}^{\infty}$ es una sucesión monótona creciente y acotada y que $\{b_n\}_{n=1}^{\infty}$ es una sucesión monótona decreciente y acotada. Por tanto, existen $\eta_i \in [a, b],\ i = 1, 2$, tales que

$$\eta_1 = \lim_{n \to +\infty} a_n \leq \lim_{n \to +\infty} b_n = \eta_2.$$

Ahora bien, de (8.6) se deduce que

$$\lim_{n \to +\infty} (b_n - a_n) = 0 \ \Rightarrow\ \eta_1 = \eta_2 = \eta$$

por lo que la continuidad de F junto con (8.5) hacen que

$$(F(\eta))^2 \leq 0,$$

de donde $F(\eta) = 0$. De esta forma, $\eta = \xi$ es la única raíz de la ecuación (8.3). Además,

$$\left| \xi - \frac{a_n + b_n}{2} \right| \leq \frac{b_n - a_n}{2} \leq \frac{b-a}{2^{n+1}},\ n \in \mathbb{N}. \tag{8.7}$$

Por tanto, para encontrar un valor que aproxime a ξ con un error inferior a $\varepsilon > 0$ basta tomar

$$\zeta = \frac{a_n + b_n}{2}$$

siendo $n \in \mathbb{N}$ tal que

$$\frac{b-a}{2^{n+1}} < \varepsilon,$$

es decir,

$$n > \frac{\ln(b-a) - \ln \varepsilon}{\ln 2} - 1.$$

Observación 8.1. El método de la bisección es útil para dar una idea rápida de la *localización* de las raíces, es decir, para determinar intervalos de longitud pequeña que contengan una única raíz; en cambio, los cálculos aumentan considerablemente si deseamos una buena aproximación de la raíz, al ser la convergencia muy lenta. Esto hace que este método se aplique, principalmente, como un paso previo a la utilización de otros métodos iterativos de convergencia más rápida. □

8.3. Métodos de punto fijo

Una forma alternativa de abordar el problema de hallar las raíces de una ecuación

$$F(x) = 0 \ \text{ en } \ [a, b]$$

es considerar otra ecuación

$$G(x) = 0 \ \text{ en } \ [a, b]$$

que sea *equivalente* a la anterior en el sentido de que ambas tengan las mismas raíces. De las múltiples formas en que esto puede llevarse a cabo vamos a considerar, en esta sección, el caso de que

$$G(x) = f(x) - x, \ x \in [a, b]$$

para alguna función $f : [a, b] \to \mathbb{R}$. De esta forma,

$$F(x) = 0 \ \text{ en } \ [a, b] \ \Leftrightarrow \ f(x) = x \ \text{ en } \ [a, b].$$

Esto nos lleva a dar la siguiente definición:

Definición 8.2. Sea $f : D \subset \mathbb{R} \to \mathbb{R}$ una aplicación. Un elemento $\xi \in D$ es un *punto fijo* de f si

$$f(\xi) = \xi. \quad \square$$

Observación 8.2. Obviamente, los puntos fijos de una función f son aquellos en los que su gráfica $y = f(x)$ corta a la recta $y = x$. En la figura 8.3 se representa la función

$$f(x) = 2x^4 - (2\sqrt{2} + 7)x^3 + 7(\sqrt{2} + 1)x^2 - (7\sqrt{2} + 1)x + 2\sqrt{2} \tag{8.8}$$

cuyos puntos fijos son

$$\xi_1 = \frac{1}{2}, \ \xi_2 = 1, \ \xi_3 = \sqrt{2} \ \text{ y } \ \xi_4 = 2. \quad \square$$

A continuación vamos a considerar una clase de funciones para las que, bajo ciertas hipótesis, va a existir un único punto fijo.

Definición 8.3. Una aplicación $f : [a, b] \to \mathbb{R}$ es *contractiva* (o una *contracción*) en $[a, b]$ si existe $0 \leq k < 1$ tal que

$$|f(x) - f(y)| \leq k|x - y|, \ x, y \in [a, b]. \tag{8.9}$$

k se denomina *constante de contractividad.* $\square$

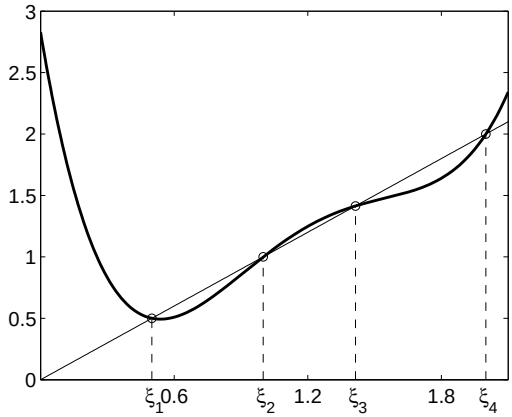


Figura 8.3: Puntos fijos de la función f dada en (8.8).

Ejemplo 8.2. La aplicación $f : [1, 2] \to \mathbb{R}$ dada por

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

es contractiva en $[1, 2]$ de constante $k = \dfrac{1}{2}$ pues para todo $x, y \in [1, 2]$ se verifica:

$$|f(x) - f(y)| = \left|\left(\frac{x}{2} + \frac{1}{x}\right) - \left(\frac{y}{2} + \frac{1}{y}\right)\right| = \left|\frac{x-y}{2} - \frac{x-y}{xy}\right|$$

$$= \left|\frac{1}{2} - \frac{1}{xy}\right| |x - y| \leq \frac{1}{2}|x - y|. \quad \square$$

Observación 8.3.

1. Si f es contractiva en $[a, b]$ entonces, a partir de (8.9), se tiene que

   $$|f(x) - f(y)| \leq k|x - y| < |x - y|,\ x, y \in [a, b].$$

   Por tanto, la distancia de $f(x)$ a $f(y)$ es menor que la distancia entre x e y. Es decir, al aplicar la función se "contrae" la distancia entre los puntos; de ahí el calificativo de contractiva.

2. Si f es contractiva en $[a, b]$ entonces es uniformemente continua (y, por tanto, continua) en $[a, b]$. El recíproco, en general, no es cierto. En efecto, basta considerar, por ejemplo, la función

   $$f(x) = \sqrt{|x|},\ x \in [-1, 1].$$

Como f es continua en el compacto $[-1,1]$ entonces f es uniformemente continua en $[-1,1]$. En cambio, para todo $n \in \mathbb{N}$ se verifica que

$$\left| f\left(\frac{1}{n^2}\right) - f(0) \right| = \frac{1}{n} = n \left| \frac{1}{n^2} - 0 \right|$$

por lo que la función f no es contractiva en el intervalo $[-1,1]$. De hecho, como se observa en la figura 8.4, la derivada de la función $f(x)$ tiende a infinito cuando $x \to 0$.

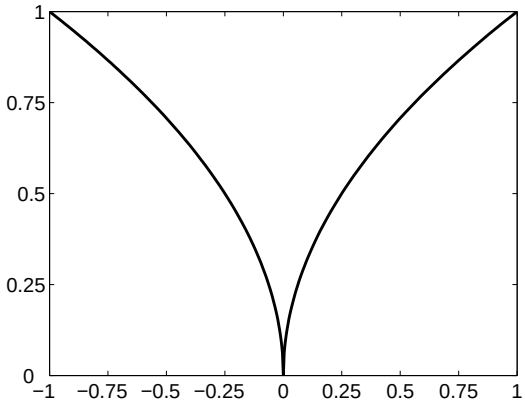


Figura 8.4: Función $f(x) = \sqrt{|x|}$.

3. Si $f \in \mathcal{C}([a,b])$ es derivable en (a,b) y existe $0 \leq k < 1$ tal que

$$|f'(x)| \leq k,\ x \in (a,b)$$

entonces la aplicación f es contractiva en $[a,b]$ de constante k pues, a partir del teorema del Valor Medio, se tiene que

$$|f(x) - f(y)| = |f'(\eta)|\,|x-y| \leq k|x-y|,\ x,y \in [a,b].$$

4. Si $f \in \mathcal{C}^1([a,b])$ y

$$|f'(x)| < 1,\ x \in [a,b] \tag{8.10}$$

entonces f es contractiva en $[a,b]$ de constante

$$k = \max_{a \leq x \leq b} |f'(x)| = ||f'||_{\mathrm{L}^\infty(a,b)}\,.$$

Para que este resultado sea cierto es esencial que el intervalo $[a,b]$ que aparece en (8.10) sea cerrado. En efecto, la función $f : [0,1] \to [0,1]$ dada por

$$f(x) = \frac{(1-x)^2}{2},\ x \in [0,1]$$

verifica

$$|f'(x)| = 1 - x < 1,\ x \in (0,1]$$

y no es contractiva en $[0,1]$, ya que

$$\left|f\left(\frac{1}{2n}\right) - f\left(\frac{1}{n}\right)\right| = \frac{1}{2}\left|\left(1 - \frac{1}{n} + \frac{1}{4n^2}\right) - \left(1 - \frac{2}{n} + \frac{1}{n^2}\right)\right|$$

$$= \left(1 - \frac{3}{4n}\right)\frac{1}{2n} = \frac{4n-3}{4n}\left|\frac{1}{2n} - \frac{1}{n}\right|$$

y

$$\lim_{n\to+\infty} \frac{4n-3}{4n} = 1. \quad \square$$

Teorema 8.1 (Punto Fijo de Banach). *Si $f : [a,b] \to \mathbb{R}$ es contractiva de constante $k \in [0,1)$ y verifica*

$$f([a,b]) \subset [a,b] \tag{8.11}$$

entonces f tiene un único punto fijo en $[a,b]$, es decir, existe un único $\xi \in [a,b]$ tal que $f(\xi) = \xi$. Además, ξ es el límite de la sucesión definida por

$$\begin{cases} x_0 \in [a,b] \ \ \textit{arbitrario} \\ x_n = f(x_{n-1}),\ n \in \mathbb{N} \end{cases} \tag{8.12}$$

y se tiene la siguiente estimación del error:

$$|x_n - \xi| \leq \frac{k^n}{1-k}|x_1 - x_0|,\ n \in \mathbb{N}. \tag{8.13}$$

Demostración.

a) Unicidad: supongamos que existieran $\xi_i \in [a,b]$, $i = 1,2$, tales que

$$f(\xi_i) = \xi_i$$

para $i = 1,2$, y veamos que, de hecho, $\xi_1 = \xi_2$. En efecto, como

$$|\xi_1 - \xi_2| = |f(\xi_1) - f(\xi_2)| \leq k\,|\xi_1 - \xi_2|$$

entonces

$$0 \leq (1-k)\,|\xi_1 - \xi_2| \leq 0$$

de donde, al ser $0 \leq k < 1$, se deduce que $\xi_1 = \xi_2$.

b) Existencia: la hipótesis (8.11) hace que la sucesión $\{x_n\}_{n=0}^{\infty}$ definida en (8.12) verifique

$$x_n \in [a,b],\ n \in \mathbb{N} \cup \{0\}.$$

Veamos que esta sucesión es de Cauchy en $[a,b]$. Para ello debemos demostrar que para todo $\varepsilon > 0$ existe $n_\varepsilon \in \mathbb{N}$ tal que

$$|x_m - x_n| < \varepsilon \ \text{ si } \ m,n \geq n_\varepsilon.$$

En primer lugar, observemos que para todo $q \in \mathbb{N} \cup \{0\}$ se cumple que

$$\begin{aligned} |x_{q+1} - x_q| &= |f(x_q) - f(x_{q-1})| \leq k|x_q - x_{q-1}| \\ &\leq k^2|x_{q-1} - x_{q-2}| \leq \cdots \leq k^q|x_1 - x_0|. \end{aligned}$$

Suponiendo que $m > n$ podemos escribir $m = n + p$ con $p \in \mathbb{N}$. De esta forma, aplicando la propiedad anterior, se verifica

$$\begin{aligned} |x_{n+p} - x_n| &= |(x_{n+p} - x_{n+p-1}) + (x_{n+p-1} - x_{n+p-2}) + \cdots + (x_{n+1} - x_n)| \\ &\leq |x_{n+p} - x_{n+p-1}| + |x_{n+p-1} - x_{n+p-2}| + \cdots + |x_{n+1} - x_n| \\ &\leq k^{n+p-1}|x_1 - x_0| + k^{n+p-2}|x_1 - x_0| + \cdots + k^n|x_1 - x_0| \\ &= \left(k^{n+p-1} + k^{n+p-2} + \cdots + k^n\right)|x_1 - x_0| \\ &= k^n\left(k^{p-1} + k^{p-2} + \cdots + k + 1\right)|x_1 - x_0| \\ &= k^n|x_1 - x_0|\sum_{i=0}^{p-1} k^i \leq k^n|x_1 - x_0|\sum_{i=0}^{\infty} k^i = \frac{k^n}{1-k}|x_1 - x_0|, \end{aligned}$$

por ser $0 \leq k < 1$. Por la misma razón,

$$\lim_{n\to+\infty} k^n = 0$$

y, en consecuencia, dado $\varepsilon > 0$ existe $n_\varepsilon \in \mathbb{N}$ tal que

$$\frac{k^n}{1-k}|x_1 - x_0| < \varepsilon \ \text{ si } \ n \geq n_\varepsilon.$$

Así, si $n \geq n_\varepsilon$ se tiene que

$$|x_{n+p} - x_n| < \varepsilon,$$

luego $\{x_n\}_{n=0}^{\infty}$ es una sucesión de Cauchy en $[a,b]$ y, por tanto, convergente en $[a,b]$, es decir, existe $\xi \in [a,b]$ tal que

$$\xi = \lim_{n\to+\infty} x_n.$$

Como $f \in \mathcal{C}([a,b])$ entonces, de la relación anterior, se obtiene que

$$f(\xi) = f\left(\lim_{n\to+\infty} x_n\right) = \lim_{n\to+\infty} f(x_n) = \lim_{n\to+\infty} x_{n+1} = \xi,$$

o sea, ξ es (el único) punto fijo de f. Finalmente, para cada $n \in \mathbb{N}$ fijo hemos obtenido la estimación

$$|x_n - x_{n+p}| \leq \frac{k^n}{1-k}|x_1 - x_0|,\ p \in \mathbb{N}.$$

Haciendo tender $p \to +\infty$, se obtiene la cota del error dada en (8.13). □

Observación 8.4.

1. Es de especial importancia el hecho de que el teorema del Punto Fijo no sólo asegura la existencia y unicidad del punto fijo de la ecuación sino que, al ser la demostración constructiva, diseña el método de cálculo del mismo, mediante lo que se conoce como el *método de las aproximaciones sucesivas* de *Picard* o *método del Punto Fijo*.

2. Aunque en las condiciones de aplicación del teorema del Punto Fijo la sucesión $\{x_n\}_{n=0}^{\infty}$ va a converger a ξ independientemente de x_0, conviene tomar, en la medida de lo posible, datos iniciales próximos a ξ.

3. A partir de (8.13), si queremos aproximar ξ de forma que el error cometido sea inferior a $\varepsilon > 0$, basta tomar $n \in \mathbb{N}$ verificando

$$\frac{k^n}{1-k}|x_1 - x_0| < \varepsilon \Leftrightarrow n > \frac{\ln((1-k)\varepsilon) - \ln(|x_1 - x_0|)}{\ln k}, \tag{8.14}$$

siendo x_n el valor aproximado buscado. □

Observación 8.5. Regresando a los métodos iterativos de resolución de un sistema lineal $Au = b$ estudiados en el capítulo 5, nótese que el método iterativo considerado en la definición 5.1 no es más que el de las aproximaciones sucesivas de Picard: hallar un punto fijo u de la aplicación $f : \mathbf{V} \to \mathbf{V}$ dada por

$$f(v) = Bv + c,\ v \in \mathbf{V}.$$

Cuando existe una norma matricial $|||\cdot|||$ (subordinada a una norma vectorial $||\cdot||$) con $|||B||| < 1$ se verifica que la aplicación f es contractiva, puesto que

$$||f(v) - f(w)|| = ||B(v-w)|| \leq |||B|||\ ||v-w||,\ v, w \in \mathbf{V},$$

por lo que la función f tiene un único punto fijo $u \in \mathbf{V}$ que es solución del sistema lineal $u = Bu + c$ y, por tanto, de $Au = b$. □

Observación 8.6 (Interpretación geométrica). En la figura 8.5 se muestran las iteraciones sucesivas del método de Punto Fijo para la ecuación

$$f(x) = x$$

donde hemos supuesto, para simplificar, que $f \in \mathcal{C}^1([a,b])$, f' tiene signo constante y, además,

$$|f'(x)| < 1,\ x \in [a,b].$$

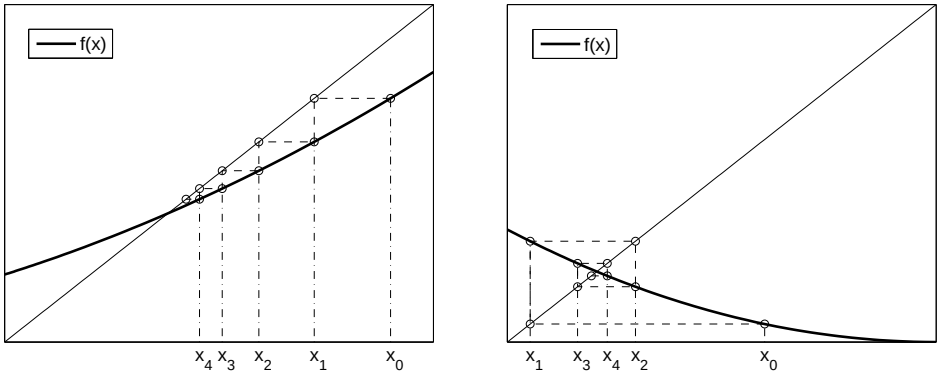


Figura 8.5: Iteraciones del Punto Fijo para $0 < f'(x) < 1$ y $-1 < f'(x) < 0$.

Ejemplo 8.3. Vamos a aproximar las raíces de la ecuación

$$e^{-x} - x = 0$$

con un orden de error inferior a 10^{-3}. Para ello, consideramos la función

$$F(x) = e^{-x} - x,\ x \in \mathbb{R}.$$

Como

$$F(0) = 1 \ \text{ y } \ F(1) = \frac{1}{e} - 1 = \frac{1-e}{e} < 0,$$

por el teorema de Bolzano existe $\xi \in (0,1)$ tal que $F(\xi) = 0$. Por otra parte, como

$$F'(x) = -e^{-x} - 1 < 0,\ x \in \mathbb{R}$$

la raíz ξ de F es única. A partir de la función

$$f(x) = e^{-x},\ x \in \mathbb{R}$$

se da la equivalencia

$$F(x) = 0 \ \Leftrightarrow \ f(x) = x.$$

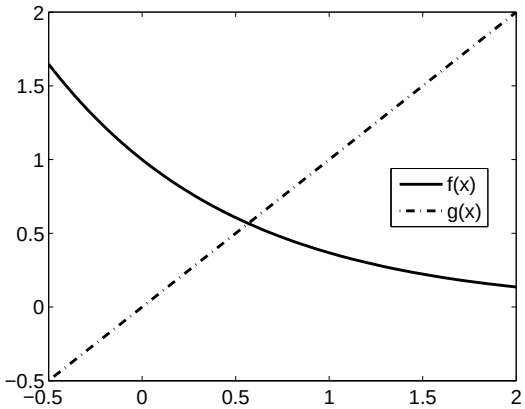


Figura 8.6: Funciones $f(x) = e^{-x}$ y $g(x) = x$.

Como se verifica que

$$f'(x) = -e^{-x} < 0,\ x \in \mathbb{R},$$

entonces la función f es estrictamente decreciente en $\mathbb{R}$ y, por tanto,

$$f([0,1]) = [f(1), f(0)] = \left[\frac{1}{e}, 1\right] \subset [0,1].$$

Aplicando ahora el teorema del Valor Medio obtenemos que

$$|e^{-x} - e^{-y}| = e^{-\eta}\,|x-y|,\ x, y \in [0,1].$$

Puesto que

$$\lim_{\eta \to 0}\ e^{-\eta} = 1$$

no podemos probar la contractividad de f en el intervalo $[0,1]$. Busquemos un intervalo más pequeño: como

$$f\left(\frac{1}{3}\right) = \frac{1}{e^{\frac{1}{3}}} < 1 \ \text{ y } \ f(1) = \frac{1}{e} > \frac{1}{3}$$

entonces, por ser f estrictamente decreciente, se verifica que

$$f\left(\left[\frac{1}{3}, 1\right]\right) \subset \left[\frac{1}{3}, 1\right].$$

Por otra parte, al aplicar el teorema del Valor Medio, se tiene que

$$|e^{-x} - e^{-y}| = e^{-\eta}\,|x-y| \leq e^{-\frac{1}{3}}|x-y|,\ x, y \in \left[\frac{1}{3}, 1\right]$$

ya que

$$\frac{1}{3} < \eta < 1 \Rightarrow e^{-\eta} < e^{-\frac{1}{3}}.$$

Por tanto, la función f es contractiva en $\left[\frac{1}{3}, 1\right]$ de constante

$$k = e^{-\frac{1}{3}} < 1.$$

Así, por el teorema 8.1, se tiene que existe un único $\xi \in \left[\frac{1}{3}, 1\right]$ tal que

$$\xi = e^{-\xi} \Leftrightarrow e^{-\xi} - \xi = 0$$

y, además, si consideramos la sucesión

$$\begin{cases} x_0 = \dfrac{1}{3} \\ x_n = e^{-x_{n-1}},\ n \in \mathbb{N} \end{cases}$$

se verifica que

$$\lim_{n \to +\infty} x_n = \xi.$$

Para aproximar ξ con el orden de error deseado, tomamos $n \in \mathbb{N}$ verificando (8.14), es decir,

$$\frac{e^{-\frac{n}{3}}}{1 - e^{-\frac{1}{3}}} \left| e^{-\frac{1}{3}} - \frac{1}{3} \right| < 10^{-3} \Leftrightarrow n > 21.6276.$$

Por tanto, tomando $n = 22$, tenemos asegurado que el valor

$$x_{22} = 0.56714233424982$$

aproxima a ξ con un error inferior a 10^{-3}. Puesto que la acotación del error es simplemente eso, una acotación, es probable que para valores más pequeños de n también se tengan aproximaciones de ξ con la precisión requerida. Así, en este caso, puesto que

$$\xi \simeq 0.56714329040978,$$

basta tomar términos de la sucesión a partir de $x_{13} = 0.56730080374378$. □

Para poder comparar los distintos métodos necesitamos definir las distintas velocidades de convergencia:

Definición 8.4. Sea $\{a_n\}_{n=0}^{\infty} \subset \mathbb{R}$ una sucesión convergente,

$$\ell = \lim_{n\to+\infty} a_n \ \text{ y } \ e_n = |a_n - \ell|,\ n \in \mathbb{N} \cup \{0\}.$$

La sucesión $\{a_n\}_{n=0}^{\infty}$ converge a ℓ:

a) Al menos *linealmente* si existe $M > 0$ tal que

$$\frac{e_n}{e_{n-1}} \leq M,\ n \in \mathbb{N}.$$

b) Al menos *cuadráticamente* si existe $M > 0$ tal que

$$\frac{e_n}{(e_{n-1})^2} \leq M,\ n \in \mathbb{N}. \quad \square$$

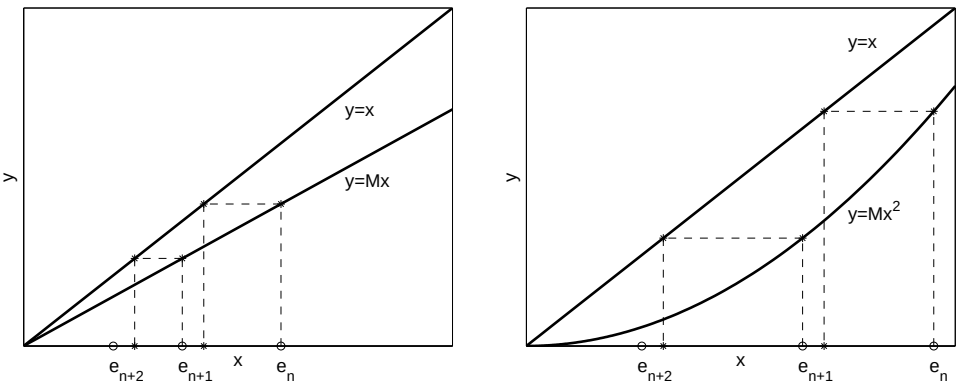


Figura 8.7: Convergencia lineal y convergencia cuadrática.

Observación 8.7.

1. Cuando se verifica que

$$\lim_{n\to+\infty} \frac{e_n}{e_{n-1}} = 0 \ \text{ o } \ \lim_{n\to+\infty} \frac{e_n}{(e_{n-1})^2} = 0$$

el orden de convergencia de la sucesión $\{a_n\}_{n=0}^{\infty}$ a ℓ se denomina, respectivamente, *superlineal* o *supercuadrático*.

2. Las nociones anteriores pueden generalizarse: la sucesión $\{a_n\}_{n=0}^{\infty}$ converge a ℓ al menos con un *orden de convergencia* $\alpha > 0$ si existe $M > 0$ tal que

$$\frac{e_n}{(e_{n-1})^{\alpha}} \leq M,\ n \in \mathbb{N}. \tag{8.15}$$

3. Si $\{a_n\}_{n=0}^{\infty}$ converge cuadráticamente a ℓ entonces el error en el paso n es del orden del cuadrado del error cometido en la etapa $n-1$. En particular, si $M < 1$ y $e_{n-1} < 10^{-m}$ entonces $e_n < 10^{-2m}$, es decir, si a_{n-1} aproxima a ℓ con m decimales exactos entonces a_n aproximará, al menos, con $2m$ decimales exactos el valor de ℓ; por tanto, cuando $M < 1$, el número de decimales exactos se va duplicando en cada iteración. En general, si $\{a_n\}_{n=0}^{\infty}$ converge a ℓ con un orden de convergencia $\alpha > 0$ entonces es inmediato demostrar por inducción, a partir de (8.15), la desigualdad

$$e_n \leq \Psi_n(M, \alpha, e_0),\ n \in \mathbb{N} \tag{8.16}$$

siendo

$$\Psi_n(M, \alpha, e_0) = M^{\alpha^{n-1}+\alpha^{n-2}+\cdots+\alpha+1} e_0^{\alpha^n} = \begin{cases} M^n e_0 & \text{si} \quad \alpha = 1 \\ M^{\frac{1-\alpha^n}{1-\alpha}} e_0^{\alpha^n} & \text{si} \quad \alpha \neq 1. \end{cases}$$

Para hacernos una idea de la estimación anterior, supongamos que

$$M = e_0 = \frac{1}{2},$$

en cuyo caso la función $\Psi_n(\alpha)$ es de la forma

$$\Psi_n(\alpha) = \begin{cases} \left(\dfrac{1}{2}\right)^{n+1} & \text{si} \quad \alpha = 1 \\ \left(\dfrac{1}{2}\right)^{\frac{1-\alpha^n}{1-\alpha}} \left(\dfrac{1}{2}\right)^{\alpha^n} = \left(\dfrac{1}{2}\right)^{\frac{1-\alpha^{n+1}}{1-\alpha}} & \text{si} \quad \alpha \neq 1 \end{cases}$$

y toma los valores de la tabla 8.1 para la convergencia lineal, cuadrática y cúbica.

TABLA 8.1:
Valores de la función $\Psi_n(\alpha)$ para $M = e_0 = \frac{1}{2}$

n	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$
1	2.500000×10^{-1}	1.250000×10^{-1}	6.250000×10^{-2}
2	1.250000×10^{-1}	7.812500×10^{-3}	1.220703×10^{-4}
3	6.250000×10^{-2}	3.051758×10^{-5}	9.094947×10^{-13}
4	3.125000×10^{-2}	4.656613×10^{-10}	3.761582×10^{-37}
5	1.562500×10^{-2}	1.084202×10^{-19}	$2.661225 \times 10^{-110}$

Nótese que cuanto mayor es el orden de convergencia α y más grande es el valor de n la función $\Psi_n(\alpha)$ va decayendo a cero más rápidamente, lo que hace que e_n también lo haga (véase (8.16)). □

Veamos que el orden de convergencia de la sucesión definida en el teorema 8.1 es, al menos, lineal.

Proposición 8.1. *Si $f : [a,b] \to [a,b]$ es contractiva de constante $k \in [0,1)$, entonces la sucesión $\{x_n\}_{n=0}^{\infty}$ dada en (8.12) converge, al menos linealmente, al único punto fijo ξ de f en $[a,b]$.*

Demostración. Denotando por

$$e_n = |x_n - \xi|,\ n \in \mathbb{N} \cup \{0\},$$

por ser ξ punto fijo de f, se tiene que

$$e_n = |x_n - \xi| = |f(x_{n-1}) - f(\xi)| \leq k|x_{n-1} - \xi| = ke_{n-1},$$

de donde se sigue el resultado. □

Observación 8.8. Cuanto menor sea el valor de k, más rápidamente convergerá la sucesión $\{x_n\}_{n=0}^{\infty}$ al punto fijo ξ. □

8.4. Método de Newton

Dada $F : [a,b] \to \mathbb{R}$ estamos interesados en encontrar un valor aproximado de las raíces de la ecuación

$$F(x) = 0 \ \text{ en } \ [a,b].$$

Para ello, vamos a considerar una ecuación de punto fijo equivalente a la anterior, es decir,

$$F(x) = 0 \ \text{ en } \ [a,b] \ \Leftrightarrow \ f(x) = x \ \text{ en } \ [a,b].$$

Obviamente, la equivalencia anterior se da para cualquier función f de la forma

$$f(x) = x - \phi(x)F(x),\ x \in [a,b] \tag{8.17}$$

donde $\phi : [a,b] \to \mathbb{R}$ es una función arbitraria verificando

$$\phi(x) \neq 0,\ x \in [a,b].$$

Supongamos que ξ es la única raíz de F en $[a,b]$ y, por tanto, el único punto fijo de f en $[a,b]$. Si la sucesión

$$\begin{cases} x_0 \in [a,b] \ \text{ arbitrario} \\ x_n = f(x_{n-1}),\ n \in \mathbb{N} \end{cases}$$

es convergente, tendrá por límite el punto ξ. Vamos a elegir la función ϕ de forma que la convergencia de la sucesión $\{x_n\}_{n=0}^{\infty}$ sea, al menos, cuadrática. Formalmente, haciendo un desarrollo de Taylor obtenemos

$$0 = F(\xi) = F(x_{n-1}) + F'(x_{n-1})(\xi - x_{n-1}) + \frac{F''(\eta_{n-1})}{2}(\xi - x_{n-1})^2$$

con η_{n-1} entre ξ y x_{n-1}. Despejando se obtiene que

$$\xi - x_{n-1} = -\frac{1}{F'(x_{n-1})}\left(F(x_{n-1}) + \frac{F''(\eta_{n-1})}{2}(\xi - x_{n-1})^2\right). \tag{8.18}$$

Por otra parte,

$$x_n = f(x_{n-1}) = x_{n-1} - \phi(x_{n-1})F(x_{n-1}),$$

es decir,

$$x_{n-1} - x_n = \phi(x_{n-1})F(x_{n-1}). \tag{8.19}$$

Consecuentemente, la suma de las relaciones (8.18) y (8.19) determina que

$$\xi - x_n = \left(\phi(x_{n-1}) - \frac{1}{F'(x_{n-1})}\right)F(x_{n-1}) - \frac{F''(\eta_{n-1})}{2F'(x_{n-1})}(\xi - x_{n-1})^2.$$

Así pues, si existen $m_1,\ M_2 > 0$ tales que

$$|F'(x)| \geq m_1 > 0,\ x \in [a,b] \tag{8.20}$$

y

$$|F''(x)| \leq M_2,\ x \in (a,b), \tag{8.21}$$

basta tomar

$$\phi(x) = \frac{1}{F'(x)},\ x \in [a,b] \tag{8.22}$$

para que

$$e_n \leq \frac{M_2}{2m_1}(e_{n-1})^2,\ n \in \mathbb{N} \tag{8.23}$$

y, por lo tanto, la sucesión $\{x_n\}_{n=0}^{\infty}$ tendrá, al menos, convergencia cuadrática. Es decir, tomamos

$$\boxed{f(x) = x - \frac{F(x)}{F'(x)},\ x \in [a,b]} \tag{8.24}$$

a partir de la cual se obtiene el *método de Newton*

$$\begin{cases} x_0 \in [a,b] \text{ dado} \\ x_n = x_{n-1} - \dfrac{F(x_{n-1})}{F'(x_{n-1})},\ n \in \mathbb{N}. \end{cases} \tag{8.25}$$

Observación 8.9 (Interpretación geométrica). Una iteración de este método consiste en tomar como x_{n+1} el punto de corte de la recta tangente a la gráfica de la función F en el punto x_n con el eje de abscisas (véase la figura 8.8). En efecto, la ecuación de la recta tangente a la curva $y = F(x)$ en el punto $(x_n, F(x_n))$ viene dada por

$$y = F(x_n) + F'(x_n)(x - x_n)$$

y la intersección de esta recta con el eje de abscisas es

$$x = x_n - \frac{F(x_n)}{F'(x_n)},$$

es decir, la iteración siguiente x_{n+1} en el método de Newton. □

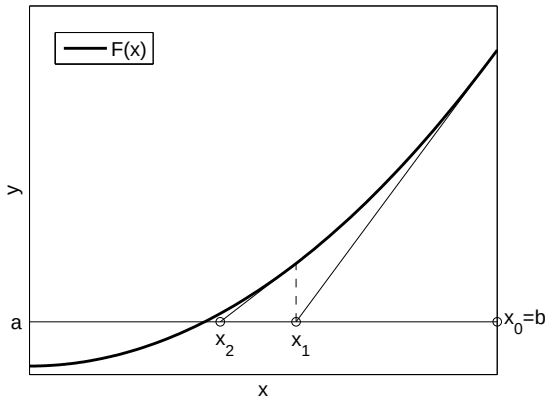


Figura 8.8: Método de Newton.

Para demostrar la convergencia del método de Newton vamos a requerir la siguiente hipótesis sobre la función $F : [a, b] \to \mathbb{R}$:

$$(\mathcal{H}) \begin{cases} F \in \mathcal{C}^2([a,b]) \\ F(a)F(b) < 0 \\ F' \text{ tiene signo constante en } [a,b] \\ F'' \text{ tiene signo constante en } [a,b]. \end{cases}$$

Estas condiciones serán también las que se utilicen para probar la convergencia de las diversas variantes del método de Newton que se describen en la sección 8.5.

Observación 8.10.

1. Toda función que verifique $(\mathcal{H})$ tiene una única raíz en $[a, b]$ que es simple.

2. Teniendo en cuenta las distintas posibilidades de los signos en $(\mathcal{H})$, pueden presentarse cuatro casos (véase la figura 8.9):

$$(\mathcal{H})_1 \begin{cases} F \in \mathcal{C}^2([a,b]) \\ F(a) < 0 < F(b) \\ F'(x) > 0,\ x \in [a,b] \\ F''(x) > 0,\ x \in [a,b] \end{cases} \qquad (\mathcal{H})_2 \begin{cases} F \in \mathcal{C}^2([a,b]) \\ F(a) < 0 < F(b) \\ F'(x) > 0,\ x \in [a,b] \\ F''(x) < 0,\ x \in [a,b] \end{cases}$$

$$(\mathcal{H})_3 \begin{cases} F \in \mathcal{C}^2([a,b]) \\ F(a) > 0 > F(b) \\ F'(x) < 0,\ x \in [a,b] \\ F''(x) < 0,\ x \in [a,b] \end{cases} \qquad (\mathcal{H})_4 \begin{cases} F \in \mathcal{C}^2([a,b]) \\ F(a) > 0 > F(b) \\ F'(x) < 0,\ x \in [a,b] \\ F''(x) > 0,\ x \in [a,b] \end{cases}$$

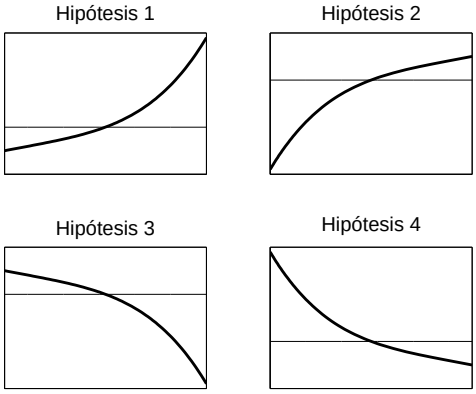


Figura 8.9: Tipos de funciones verificando la hipótesis $(\mathcal{H})$.

Para demostrar la convergencia del método de Newton cuando F verifica la hipótesis $(\mathcal{H})$ podemos suponer, sin pérdida de generalidad, que F verifica $(\mathcal{H})_1$. En efecto, la función

$$G(x) = \begin{cases} -F(-x),\ x \in [-b,-a] & \text{si } F \text{ verifica } (\mathcal{H})_2 \\ -F(x),\ x \in [a,b] & \text{si } F \text{ verifica } (\mathcal{H})_3 \\ F(-x),\ x \in [-b,-a] & \text{si } F \text{ verifica } (\mathcal{H})_4 \end{cases}$$

verifica $(\mathcal{H})_1$. Es un sencillo ejercicio comprobar que si la sucesión del método de Newton para G es convergente, también es convergente la sucesión del

método de Newton para F en cada uno de los tres casos; si F verifica $(\mathcal{H})_3$ ambas sucesiones coinciden mientras que si F verifica $(\mathcal{H})_2$ o $(\mathcal{H})_4$ una sucesión es la opuesta de la otra.

Teorema 8.2. *Si F verifica $(\mathcal{H})$ y $c \in [a,b]$ es el extremo de $[a,b]$ tal que*

$$\operatorname{sign} F(c) = \operatorname{sign} F''(c),$$

entonces el método de Newton para F con $x_0 = c$ converge, al menos cuadráticamente, a la única raíz ξ de F en $[a,b]$.

Demostración. Por la observación 8.10 sabemos que F tiene una única raíz ξ en $[a,b]$ y que, sin pérdida de generalidad, podemos suponer que F verifica la hipótesis $(\mathcal{H})_1$, lo que hace que $c = b$. Vamos a probar, por inducción, que

$$x_n \in (\xi, b],\ n \in \mathbb{N} \cup \{0\}.$$

i) Para $n = 0$ el resultado es obvio, pues $x_0 = b$.

ii) Suponiendo cierto el resultado para n, es decir,

$$\xi < x_n \leq b \tag{8.26}$$

lo demostramos para $n+1$. Desarrollando por Taylor la función F obtenemos

$$0 = F(\xi) = F(x_n) + F'(x_n)(\xi - x_n) + \frac{F''(\eta_n)}{2}(\xi - x_n)^2$$

para algún $\eta_n \in (\xi, x_n)$. Por la hipótesis de inducción y $(\mathcal{H})_1$

$$\frac{F''(\eta_n)}{2}(\xi - x_n)^2 > 0$$

y, por tanto,

$$F(x_n) + F'(x_n)(\xi - x_n) < 0.$$

Así,

$$\xi < x_n - \frac{F(x_n)}{F'(x_n)} = x_{n+1}$$

pues $F'(x_n) > 0$ y no se invierte el sentido de la desigualdad. Además, como $F'(x_n) > 0$ y por la hipótesis de inducción $F(x_n) > 0$, entonces

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)} < x_n \leq b \tag{8.27}$$

y, de esta forma,

$$\xi < x_{n+1} \leq b.$$

Por otro lado, de (8.27) se tiene que la sucesión $\{x_n\}_{n=0}^{\infty} \subset (\xi, b]$ es estrictamente decreciente y acotada. Consecuentemente, existe

$$\eta = \lim_{n \to +\infty} x_n.$$

Ahora bien, la continuidad de la función f definida en (8.24) determina que

$$f(\eta) = \lim_{n \to +\infty} f(x_n) = \lim_{n \to +\infty} x_{n+1} = \eta$$

o, equivalentemente,

$$F(\eta) = 0$$

por lo que la unicidad de raíces de F en $[a, b]$ hace que $\eta = \xi$. Además, la hipótesis $(\mathcal{H})_1$ hace que se verifiquen las condiciones (8.20) y (8.21) que implican la convergencia al menos cuadrática, como ya se detalló en la introducción de este método (véase (8.23)). □

Observación 8.11.

1. El teorema 8.2 sigue siendo válido si se toma como dato inicial cualquier valor $x_0 \in [a, b]$ que verifique

$$\operatorname{sign} F(x_0) = \operatorname{sign} F''(x_0).$$

2. Recuérdese la gran velocidad de la convergencia cuadrática. Esto hace que el método de Newton sea el que se emplee con una mayor frecuencia en las aplicaciones. □

Ejemplo 8.4. Aproximemos las raíces reales de la función

$$F(x) = x^5 + 5x + 8,\ x \in \mathbb{R}$$

mediante el método de Newton. En primer lugar, se verifica que $F \in \mathcal{C}^2(\mathbb{R})$,

$$F'(x) = 5(x^4 + 1) > 0,\ x \in \mathbb{R} \ \text{ y } \ F''(x) = 20x^3 \begin{cases} > 0,\ x > 0 \\ = 0,\ x = 0 \\ < 0,\ x < 0 \end{cases}$$

por lo que la función F es creciente en $\mathbb{R}$, cóncava en el intervalo $(-\infty, 0)$ y convexa en el intervalo $(0, +\infty)$, siendo $x = 0$ un *punto de inflexión* de F (véase la figura 8.10). Como, además,

$$F(-2) = -34 < 0 < 2 = F(-1),$$

por el teorema de Bolzano la función F tiene una única raíz $\xi \in (-2,-1)$. Por otra parte, como

$$F'(x) \neq 0,\ x \in \mathbb{R},$$

el teorema de Rolle garantiza que ξ es la única raíz de la función en todo $\mathbb{R}$. Puesto que

$$\begin{cases} F'(x) > 0,\ x \in [-2,-1] \\ F''(x) < 0,\ x \in [-2,-1] \end{cases}$$

si consideramos $x_0 = -2$ entonces el método de Newton es convergente y el límite de la sucesión

$$\begin{cases} x_0 = -2 \\ x_n = x_{n-1} - \dfrac{x_{n-1}^5 + 5x_{n-1} + 8}{5(x_{n-1}^4 + 1)} = \dfrac{4}{5}\dfrac{x_{n-1}^5 - 2}{x_{n-1}^4 + 1},\ n \in \mathbb{N} \end{cases}$$

es

$$\lim_{n \to +\infty} x_n = \xi.$$

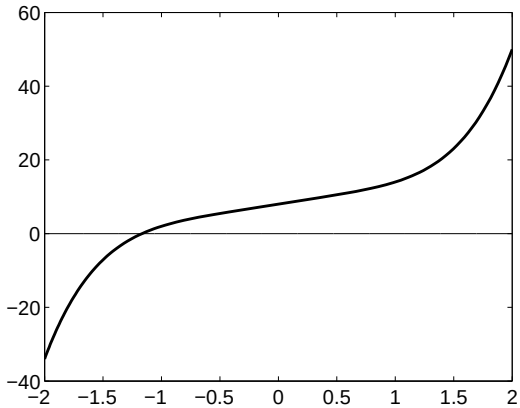


Figura 8.10: Función $F(x) = x^5 + 5x + 8$.

Los primeros términos de la sucesión anterior vienen dados en la siguiente tabla:

n	x_n
0	-2
1	-1.60000000000000
2	-1.32236390595213
3	-1.16769339203490
4	-1.16703666447529
5	-1.16703618370190
6	-1.16703618370164

Teniendo en cuenta que el valor de la raíz es

$$\xi = -1.16703618370164\ldots$$

y observando estos valores, se puede apreciar el efecto de la convergencia cuadrática, puesto que x_2 no tiene, a la derecha de la coma, ninguna cifra correcta, x_3 tiene tres, x_4 tiene seis, x_5 tiene doce y x_6 tiene, de hecho, veinticuatro. □

A continuación vamos a establecer una cota del error en términos de la diferencia entre dos términos consecutivos de la sucesión, resultado que, como se detallará más adelante, servirá como test de parada de las iteraciones.

Proposición 8.2. *Sea $F \in \mathcal{C}^2([a,b])$ y denotemos por*

$$m_1 = \min_{a\le x\le b} |F'(x)| \;\; y \;\; M_2 = \max_{a\le x\le b} |F''(x)|. \tag{8.28}$$

Si $m_1 > 0$ la sucesión del método de Newton verifica que

$$|x_n - \xi| \le \frac{M_2}{2m_1}|x_n - x_{n-1}|^2,\ n \in \mathbb{N}.$$

Demostración. Dado $n \in \mathbb{N}$, si hacemos un desarrollo de Taylor de la función F en torno al punto x_{n-1} y particularizamos en $x = x_n$, obtenemos que

$$\begin{aligned} F(x_n) &= F(x_{n-1}) + F'(x_{n-1})(x_n - x_{n-1}) + \frac{F''(\eta_{n-1})}{2}(x_n - x_{n-1})^2 \\ &= \frac{F''(\eta_{n-1})}{2}(x_n - x_{n-1})^2 \end{aligned}$$

con η_{n-1} entre x_{n-1} y x_n, ya que

$$x_n = x_{n-1} - \frac{F(x_{n-1})}{F'(x_{n-1})}.$$

Por tanto,

$$|F(x_n)| \le \frac{M_2}{2}|x_n - x_{n-1}|^2. \tag{8.29}$$

Como, por otra parte,

$$|F(x_n)| = |F(x_n) - F(\xi)| = |F'(\nu_n)|\,|x_n - \xi| \ge m_1|x_n - \xi|,$$

entonces, a partir de (8.29), se verifica que

$$|x_n - \xi| \le \frac{|F(x_n)|}{m_1} \le \frac{M_2}{2m_1}|x_n - x_{n-1}|^2. \quad \square$$

8.5. Variantes del método de Newton

Seguimos interesados en resolver la ecuación

$$F(x) = 0 \text{ en } [a, b].$$

Como hemos visto, en las condiciones de aplicación del método de Newton, éste tiene convergencia al menos cuadrática. No obstante, el inconveniente que puede presentarse es que en las sucesivas iteraciones hay que evaluar la derivada de F y ésta puede ser difícil de calcular y evaluar, haciendo que el "gasto" en cada iteración sea grande.

En esta sección vamos a considerar algunas variantes de este método que "sustituyen" la derivada de F por algo menos costoso de calcular. De esta forma las iteraciones serán más sencillas aunque, en contrapartida, la convergencia sea más lenta.

8.5.1. Método de Whittaker

En este método se toma, en lugar de $F'(x)$, un valor constante $\lambda \neq 0$. Nótese que, a partir de (8.22), esto es equivalente a tomar

$$\phi(x) \equiv \frac{1}{\lambda},\ x \in [a, b]$$

en (8.17). En consecuencia,

$$\boxed{f(x) = x - \frac{F(x)}{\lambda},\ x \in [a, b]}$$

a partir de lo cual se obtiene el *método de Whittaker*

$$\begin{cases} x_0 \in [a, b] \text{ dado} \\ x_n = x_{n-1} - \dfrac{F(x_{n-1})}{\lambda},\ n \in \mathbb{N}. \end{cases} \tag{8.30}$$

Observación 8.12 (Interpretación geométrica). La ecuación de la recta de pendiente λ que pasa por el punto $(x_n, F(x_n))$ viene dada por

$$y = F(x_n) + \lambda(x - x_n)$$

y la intersección de esta recta con el eje de abscisas es

$$x = x_n - \frac{F(x_n)}{\lambda},$$

esto es, la iteración siguiente x_{n+1} en el método de Whittaker. Es decir, este método consiste en intersecar el eje de abscisas con rectas paralelas que pasan por la imagen por F de la iteración anterior x_n y tienen pendiente constante λ (véase la figura 8.11). □

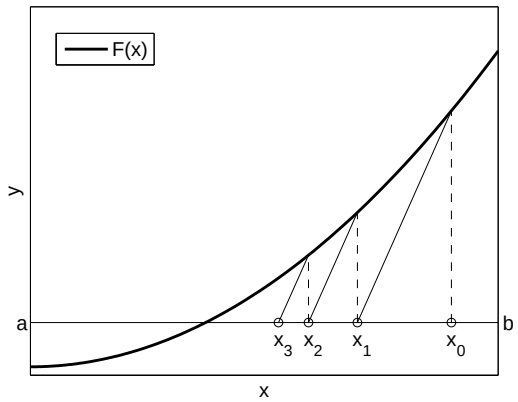


Figura 8.11: Método de Whittaker.

Veamos un resultado que asegura la convergencia de este método bajo las mismas hipótesis que en el método de Newton.

Teorema 8.3. *Si F verifica $(\mathcal{H})$, $c \in [a,b]$ es el extremo de $[a,b]$ tal que*

$$\operatorname{sign} F(c) = \operatorname{sign} F''(c)$$

y $\lambda \in \mathbb{R}\backslash\{0\}$ verifica que

$$\operatorname{sign} \lambda = \operatorname{sign} F' \;\; y \;\; |\lambda| \geq |F'(c)|,$$

entonces el método de Whittaker para F con $x_0 \in [a,b]$ arbitrario converge, al menos linealmente, a la única raíz ξ de F en $[a,b]$.

Demostración. Por la observación 8.10 sabemos que F tiene una única raíz ξ en $[a,b]$ y que, sin pérdida de generalidad, podemos suponer que F verifica la hipótesis $(\mathcal{H})_1$ lo que hace que $c = b$. Por lo tanto, el rango de valores de λ es

$$\lambda \geq F'(b) > 0.$$

En las condiciones anteriores vamos a demostrar que la función

$$f(x) = x - \frac{F(x)}{\lambda},\ x \in [a,b]$$

verifica las hipótesis del teorema del Punto Fijo, utilizando que $f \in \mathcal{C}^2([a,b])$ por serlo F:

a) $f([a,b]) \subset [a,b]$. Como $F'' > 0$ en $[a,b]$, la función F' es estrictamente creciente en $[a,b]$. Por tanto,

$$0 < F'(x) \leq F'(b) \leq \lambda,\ x \in [a,b],$$

lo que implica que

$$f'(x) = 1 - \frac{F'(x)}{\lambda} = \frac{\lambda - F'(x)}{\lambda} \geq 0,\ x \in [a,b]. \tag{8.31}$$

De esta forma, la función f es creciente en $[a,b]$. En concreto,

$$f(a) \leq f(x) \leq f(b),\ x \in [a,b]$$

y, con ello,

$$a < a - \frac{F(a)}{\lambda} = f(a) \leq f(x) \leq f(b) = b - \frac{F(b)}{\lambda} < b,\ x \in [a,b]$$

ya que $F(a) < 0 < F(b)$ y $\lambda > 0$. Así,

$$f([a,b]) \subset (a,b) \subset [a,b].$$

b) f es contractiva en $[a,b]$. Para ello vamos a probar que existe $0 \leq k < 1$ tal que

$$|f'(x)| \leq k,\ x \in [a,b].$$

Como

$$f''(x) = -\frac{F''(x)}{\lambda} < 0,\ x \in [a,b]$$

entonces la función f' es estrictamente decreciente en $[a,b]$; utilizando (8.31), llegamos a que

$$k = \max_{a \leq x \leq b} |f'(x)| = \max_{a \leq x \leq b} f'(x) = f'(a) = 1 - \frac{F'(a)}{\lambda} < 1.$$

Por tanto, aplicando el teorema 8.1 se tiene que ξ (que es la única raíz de F en $[a,b]$ y el único punto fijo de f en $[a,b]$) se obtiene como límite de la sucesión dada por

$$\begin{cases} x_0 \in [a,b] \ \text{ arbitrario} \\ x_n = x_{n-1} - \dfrac{F(x_{n-1})}{\lambda},\ n \in \mathbb{N}. \end{cases}$$

Además, por la proposición 8.1, la convergencia de la sucesión es, al menos, lineal. □

Observación 8.13.

1. En términos geométricos, el teorema 8.3 asegura que el método de Whittaker es convergente para todos los valores de la pendiente λ que sean mayores que la mayor pendiente de las tangentes a la curva $y = F(x)$ en el intervalo $[a, b]$.

2. Puede demostrarse (véase el problema 8.9) que el método de Whittaker converge más rápidamente cuanto más cerca de $F'(c)$ esté λ. Por tanto, el método de Whittaker *óptimo*, aquel en el que los términos de la sucesión $\{|x_n - \xi|\}_{n=0}^{\infty}$ son lo menor posible, se obtiene eligiendo

$$\lambda = F'(c),$$

y a veces se denomina *método de Newton modificado.*

3. En el caso general en que F verifique $(\mathcal{H})$, si $c \in [a, b]$ es el extremo de $[a, b]$ tal que

$$\operatorname{sign} F(c) = \operatorname{sign} F''(c)$$

y d es el otro extremo del intervalo, entonces la función

$$f(x) = x - \frac{F(x)}{\lambda},\ x \in [a, b]$$

es contractiva en $[a, b]$ de constante

$$k = 1 - \frac{F'(d)}{\lambda}.$$

Como λ y $F'(d)$ tienen igual signo, el teorema del Punto Fijo permite obtener la siguiente estimación del error: para cada $n \in \mathbb{N}$ se verifica que

$$|x_n - \xi| \leq \frac{\left(1 - \dfrac{F'(d)}{\lambda}\right)^n}{\dfrac{F'(d)}{\lambda}} |f(x_0) - x_0| = \left(\frac{\lambda - F'(d)}{\lambda}\right)^n \frac{|F(x_0)|}{|F'(d)|}. \quad \square$$

Ejemplo 8.5. Sea $F : [0, 1] \to \mathbb{R}$ dada por

$$F(x) = e^x - 2\cos x,\ x \in [0, 1].$$

Como

$$F(0) = -1 < 0 < e - 2\cos 1 = F(1)$$

y para todo $x \in [0, 1]$ se verifica que

$$F'(x) = e^x + 2\operatorname{sen} x > 0 \ \text{ y } \ F''(x) = e^x + 2\cos x > 0$$

entonces $c = 1$, $d = 0$ y $F'(1) = e + 2\,\mathrm{sen}\,1$. Tomando $\lambda = 5$ y $x_0 = 1$, los primeros términos de la sucesión del método de Whittaker son los dados en la siguiente tabla:

n	x_n
0	1
1	0.67246455665545
2	0.59356817154295
3	0.56306088171033
4	0.55010311807329
5	0.54440501432519
6	0.54186273656146

siendo $\xi = 0.53978516080928\ldots$ □

8.5.2. Método de las cuerdas

En este método se toma, en lugar de $F'(x)$, la función

$$\frac{F(c) - F(x)}{c - x},\ x \in [a, b],$$

donde c va a ser uno de los extremos del intervalo $[a, b]$. A partir de (8.22), esta elección equivale a tomar

$$\phi(x) = \frac{c - x}{F(c) - F(x)},\ x \in [a, b]$$

en (8.17). Así, en este caso

$$\boxed{f(x) = x - \frac{F(x)}{F(c) - F(x)}(c - x),\ x \in [a, b]}$$

a partir de la cual se obtiene el *método de las cuerdas*

$$\begin{cases} x_0 \in [a, b] \text{ dado} \\ x_n = x_{n-1} - \dfrac{F(x_{n-1})}{F(c) - F(x_{n-1})}(c - x_{n-1}),\ n \in \mathbb{N}. \end{cases} \tag{8.32}$$

Observación 8.14 (Interpretación geométrica). Tomando $c = b$, la ecuación de la recta que pasa por los puntos $(x_n, F(x_n))$ y $(b, F(b))$ es

$$y = F(x_n) + \frac{F(b) - F(x_n)}{b - x_n}(x - x_n)$$

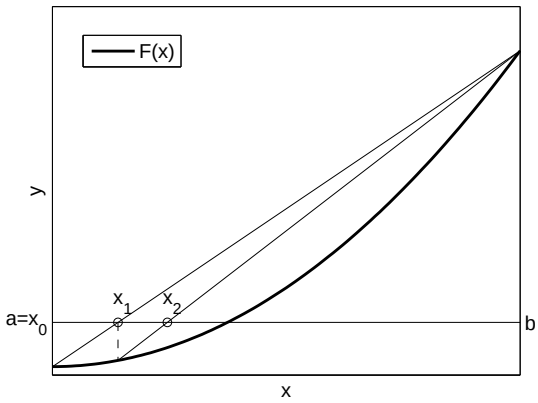


Figura 8.12: Método de las cuerdas.

y la intersección de esta recta con el eje de abscisas es

$$x = x_n - \frac{F(x_n)}{F(b) - F(x_n)}(b - x_n),$$

es decir, la iteración siguiente x_{n+1} en el método de las cuerdas. □

Veamos un resultado de convergencia para este método bajo las mismas hipótesis que en el método de Newton.

Teorema 8.4. *Sea F verificando $(\mathcal{H})$, $c \in [a, b]$ el extremo de $[a, b]$ tal que*

$$\operatorname{sign} F(c) = \operatorname{sign} F''(c)$$

y d el otro extremo del intervalo. Entonces el método de las cuerdas para F con $x_0 = d$ converge, al menos linealmente, a la única raíz ξ de F en $[a, b]$.

Demostración. Por la observación 8.10 sabemos que F tiene una única raíz ξ en $[a, b]$ y que, sin pérdida de generalidad, podemos suponer que F verifica la hipótesis $(\mathcal{H})_1$ lo que hace que $c = b$ y $d = a$. Veamos, por inducción, que

$$x_n \in [a, \xi),\ n \in \mathbb{N} \cup \{0\}.$$

i) Para $n = 0$ el resultado es obvio, pues $x_0 = d = a$.

ii) Suponiendo cierto el resultado para n, es decir,

$$a \leq x_n < \xi \tag{8.33}$$

lo mostramos para $n+1$. Al ser $F' > 0$, la función F es estrictamente creciente en $[a,b]$ y, por tanto,

$$F(a) \leq F(x_n) < F(\xi) = 0.$$

De esta forma, como

$$F(b) - F(x_n) > -F(x_n) > 0 \ \text{ y } \ b - x_n > b - \xi > 0, \tag{8.34}$$

entonces

$$x_{n+1} = x_n - \frac{F(x_n)}{F(b) - F(x_n)}(b - x_n) > x_n \geq a. \tag{8.35}$$

La hipótesis

$$F''(x) > 0,\ x \in [a,b]$$

determina que la función F es convexa en el intervalo $[a,b]$; así, en particular, la gráfica de F en $[x_n,b]$ queda por debajo de la cuerda de ecuación

$$y(x) = F(x_n) + \frac{F(b) - F(x_n)}{b - x_n}(x - x_n)$$

que une los puntos $(x_n, F(x_n))$ y $(b, F(b))$. De esta forma, evaluando en el punto $\xi \in (x_n, b)$ se tiene que

$$0 = F(\xi) < y(\xi) = F(x_n) + \frac{F(b) - F(x_n)}{b - x_n}(\xi - x_n),$$

de donde se obtiene que

$$\xi > x_n - \frac{F(x_n)}{F(b) - F(x_n)}(b - x_n) = x_{n+1}.$$

Por tanto, por (8.35), la sucesión $\{x_n\}_{n=0}^{\infty}$ es estrictamente creciente y, además,

$$a = x_0 < x_n < x_{n+1} < \xi < b,\ n \in \mathbb{N}, \tag{8.36}$$

luego existe $\eta \in [a,\xi]$ tal que

$$\eta = \lim_{n \to +\infty} x_n.$$

Como $F \in \mathcal{C}([a,b])$ y

$$x_n = x_{n-1} - \frac{F(x_{n-1})}{F(b) - F(x_{n-1})}(b - x_{n-1}),\ n \in \mathbb{N}$$

haciendo tender $n \to +\infty$, se obtiene que

$$\eta = \eta - \frac{F(\eta)}{F(b) - F(\eta)}(b - \eta),$$

es decir,

$$\frac{F(\eta)}{F(b) - F(\eta)}(b - \eta) = 0$$

de donde se concluye que $F(\eta) = 0$ ya que, por (8.36), $\eta \neq b$. De esta forma, por la unicidad de raíces de F en $[a, b]$, se llega a que $\eta = \xi$.

Finalmente, utilizando (8.33) y (8.35), para todo $n \in \mathbb{N}$, se tiene que

$$0 > x_{n+1} - \xi > x_n - \xi.$$

Por tanto,

$$e_{n+1} = |x_{n+1} - \xi| < |x_n - \xi| = e_n,$$

de donde se deduce la convergencia, al menos lineal, del método. □

Además se tiene el siguiente resultado relativo a la estimación del error:

Proposición 8.3. *Sea F verificando $(\mathcal{H})$ y denotemos por*

$$m_1 = \min_{a \leq x \leq b} |F'(x)| \;\; y \;\; M_1 = \max_{a \leq x \leq b} |F'(x)|. \tag{8.37}$$

La sucesión del método de las cuerdas comenzando en $x_0 = d$ verifica que

$$|x_n - \xi| \leq \frac{M_1 - m_1}{m_1} |x_n - x_{n-1}|, \; n \in \mathbb{N}.$$

Demostración. Nuevamente, por la observación 8.10, F tiene una única raíz ξ en $[a, b]$ y, sin pérdida de generalidad, podemos suponer que F verifica la hipótesis $(\mathcal{H})_1$ lo que hace que $c = b$ y $d = a$. Para cada $n \in \mathbb{N}$, por el teorema del Valor Medio, se verifica que

$$F(\xi) - F(x_{n-1}) = F'(\eta_{n-1})(\xi - x_{n-1}) \;\; \text{con} \;\; \eta_{n-1} \in (x_{n-1}, \xi)$$

y

$$F(b) - F(x_{n-1}) = F'(\nu_{n-1})(b - x_{n-1}) \;\; \text{con} \;\; \nu_{n-1} \in (x_{n-1}, b).$$

De esta forma, como $F(\xi) = 0$, se tiene que

$$\begin{aligned} F'(\eta_{n-1})(\xi - x_{n-1}) = -F(x_{n-1}) &= \frac{F(b) - F(x_{n-1})}{b - x_{n-1}}(x_n - x_{n-1}) \\ &= F'(\nu_{n-1})(x_n - x_{n-1}). \end{aligned}$$

Como la función F' es positiva en $[a,b]$ entonces

$$\xi - x_{n-1} = \frac{F'(\nu_{n-1})}{F'(\eta_{n-1})}(x_n - x_{n-1})$$

y, por tanto,

$$\begin{aligned}\xi - x_n &= (\xi - x_{n-1}) + (x_{n-1} - x_n) = \left(\frac{F'(\nu_{n-1})}{F'(\eta_{n-1})} - 1\right)(x_n - x_{n-1}) \\ &= \frac{F'(\nu_{n-1}) - F'(\eta_{n-1})}{F'(\eta_{n-1})}(x_n - x_{n-1}).\end{aligned}$$

Tomando valores absolutos se concluye, a partir de (8.37), el resultado. □

8.5.3. Método de la secante

Este método es una versión del método de Newton en el que se discretiza la derivada mediante la fórmula de derivación numérica más simple, la de dos puntos. Es decir, en lugar de $F'(x_{n-1})$ tomamos el cociente incremental

$$\frac{F(x_{n-1}) - F(x_{n-2})}{x_{n-1} - x_{n-2}}$$

para $n \geq 2$. Se obtiene así el *método de la secante*

$$\begin{cases} x_0,\ x_1 \in [a,b] \ \text{ dados} \\ x_n = x_{n-1} - \dfrac{F(x_{n-1})}{F(x_{n-1}) - F(x_{n-2})}(x_{n-1} - x_{n-2}),\ n \geq 2. \end{cases} \tag{8.38}$$

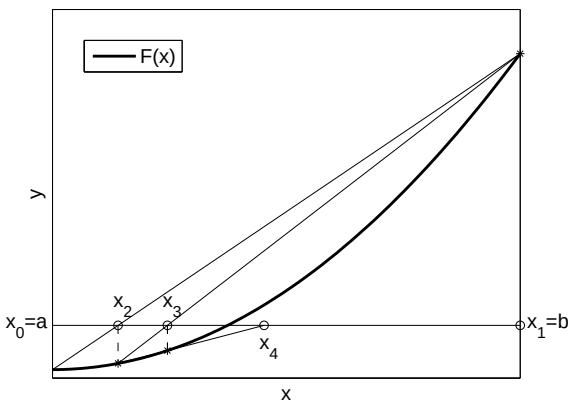


Figura 8.13: Método de la secante.

Observación 8.15 (Interpretación geométrica). La recta que une los puntos $(x_{n-1}, F(x_{n-1}))$ y $(x_n, F(x_n))$ viene dada por

$$y = F(x_n) + \frac{F(x_n) - F(x_{n-1})}{x_n - x_{n-1}}(x - x_n)$$

y la intersección de esta recta con el eje de abscisas es

$$x = x_n - \frac{F(x_n)}{F(x_n) - F(x_{n-1})}(x_n - x_{n-1}),$$

es decir, la iteración siguiente x_{n+1} en el método de la secante. □

A continuación se presenta un resultado que asegura que el método de la secante es, de todas las variantes del método de Newton consideradas, el que converge más rápido.

Proposición 8.4. *Si $F \in \mathcal{C}^2([a,b])$ y denotamos por*

$$e_n = |x_n - \xi|,\ n \in \mathbb{N} \cup \{0\}$$

siendo ξ una raíz de F en $[a,b]$, entonces

$$e_n \leq \frac{M_2}{2m_1} e_{n-1} e_{n-2},\ n \geq 2 \tag{8.39}$$

donde

$$m_1 = \min_{a \leq x \leq b} |F'(x)| \ \ y \ \ M_2 = \max_{a \leq x \leq b} |F''(x)|.$$

Por tanto, cuando el método de la secante converge, se verifica que

$$\lim_{n \to +\infty} \frac{e_n}{e_{n-1}} = 0$$

por lo que la convergencia es superlineal.

Demostración. Por definición se tiene que

$$\begin{aligned}
\xi - x_n &= \xi - x_{n-1} + \frac{F(x_{n-1})}{F(x_{n-1}) - F(x_{n-2})}(x_{n-1} - x_{n-2}) \\
&= (\xi - x_{n-1}) \left(1 - \frac{F(\xi) - F(x_{n-1})}{\xi - x_{n-1}} \frac{x_{n-1} - x_{n-2}}{F(x_{n-1}) - F(x_{n-2})}\right) \\
&= (\xi - x_{n-1}) \left(1 - F[x_{n-1}, \xi] \frac{1}{F[x_{n-1}, x_{n-2}]}\right) \\
&= (\xi - x_{n-1}) \Big(F[x_{n-2}, x_{n-1}] - F[x_{n-1}, \xi]\Big) \frac{1}{F[x_{n-1}, x_{n-2}]}
\end{aligned}$$

(véase la definición 6.2, donde se introducen las diferencias divididas). Por tanto,

$$\begin{aligned}\xi - x_n &= -(\xi - x_{n-1})(\xi - x_{n-2})\frac{F[x_{n-2}, x_{n-1}] - F[x_{n-1}, \xi]}{x_{n-2} - \xi}\frac{1}{F[x_{n-1}, x_{n-2}]} \\ &= -(\xi - x_{n-1})(\xi - x_{n-2})\frac{F[x_{n-2}, x_{n-1}, \xi]}{F[x_{n-1}, x_{n-2}]}.\end{aligned}$$

Ahora bien, la propiedad

$$f[x_{m-k}, x_{m-(k-1)}, \ldots, x_m, x] = \frac{f^{k+1)}(\eta_x)}{(k+1)!}$$

(véase (6.12)) determina la existencia de η_ξ y η tales que

$$\xi - x_n = -(\xi - x_{n-1})(\xi - x_{n-2})\frac{F''(\eta_\xi)}{2F'(\eta)}$$

de donde, tomando valores absolutos, se obtiene (8.39). □

Observación 8.16. El resultado de la proposición 8.4 se puede mejorar, puesto que puede demostrarse que existe $M > 0$ tal que

$$\lim_{n \to +\infty} \frac{e_{n+1}}{(e_n)^p} = M$$

siendo p la *razón áurea*

$$p = \frac{1 + \sqrt{5}}{2} \simeq 1.61803$$

(véase la práctica 8.7). □

8.5.4. Método de la Falsa Posición (o Regula Falsi)

Este método, sin dejar de ser una variante del método de Newton, puede interpretarse también como una generalización del método de la bisección. Para aplicarlo, basta con que la función $F \in \mathcal{C}([a, b])$ y no es necesario que F sea cóncava o convexa en $[a, b]$ (en cuyo caso este método coincide con el método de las cuerdas).

Para describirlo, consideremos $F \in \mathcal{C}([a, b])$ con $F(a)F(b) < 0$ de forma que existe un único $\xi \in (a, b)$ tal que $F(\xi) = 0$. La ecuación de la recta que une los puntos $(a, F(a))$ y $(b, F(b))$ viene dada por

$$y - F(a) = \frac{F(b) - F(a)}{b - a}(x - a)$$

y la intersección de dicha recta con el eje de abscisas es

$$x = a - \frac{F(a)}{F(b) - F(a)}(b - a) = c.$$

Como hacíamos en el método de la bisección, tomamos

$$\begin{cases} a_1 = a,\ b_1 = c & \text{si} \quad F(a)F(c) < 0 \\ a_1 = c,\ b_1 = b & \text{si} \quad F(c)F(b) < 0. \end{cases}$$

La siguiente iteración se obtendría aplicando la misma estrategia al intervalo $[a_1, b_1]$ y, así, sucesivamente (véase la figura 8.14).

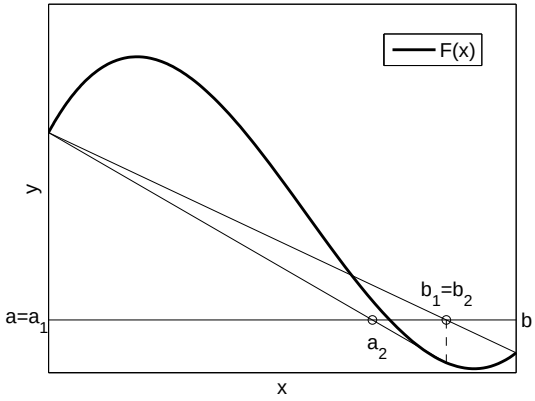


Figura 8.14: Método de la *Regula Falsi*.

8.6. Consideraciones finales

8.6.1. Test de parada de las iteraciones

Supongamos que $\xi \in [a, b]$ es la única raíz de la ecuación

$$F(x) = 0 \ \text{ en } \ [a, b].$$

En la práctica, suelen considerarse como buenas aproximaciones de ξ valores η tales que $|F(\eta)|$ sea pequeño, mientras que valores η para los cuales $|F(\eta)|$ sea grande suelen ser considerados malas aproximaciones de la raíz. No obstante, esta forma de proceder no es correcta: nótese que las ecuaciones $F(x) = 0$ y $\sigma F(x) = 0$ (con $\sigma \neq 0$) son equivalentes, por lo que el valor $|\sigma F(\eta)|$ puede hacerse tan grande o tan pequeño como queramos en función de la elección de σ. Veamos algunos ejemplos patológicos.

Ejemplo 8.6.

1. La única raíz real de la función

$$F(x) = (x-1)\left((x-4)^4 + 10^{-6}\right)^2$$

es $\xi = 1$. No obstante, para valores $x \simeq 4$ el valor de $F(x)$ es pequeño. Así, por ejemplo, $F(4) = 3 \times 10^{-12}$, lo que podría inducir a pensar, erróneamente, que una raíz de la función F está cercana al punto $x = 4$ (véase la figura 8.15).

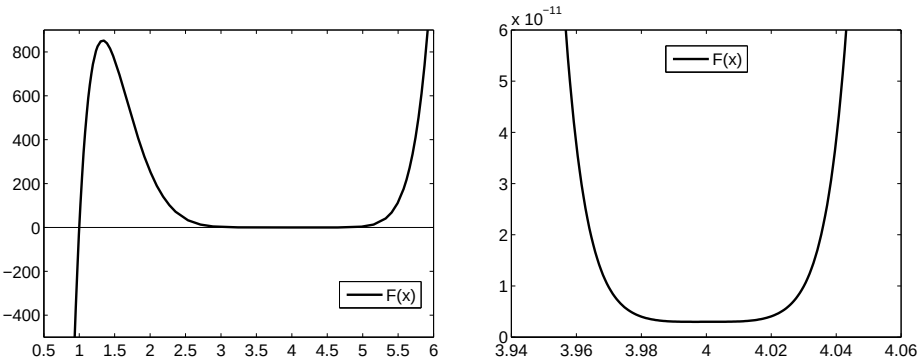


Figura 8.15: Función $F(x) = (x-1)\left((x-4)^4 + 10^{-6}\right)^2$.

2. La función

$$F(x) = \frac{(1-x)^2}{(x-0.95)^6}$$

(véase la figura 8.16) tiene una única raíz real doble que es $\xi = 1$.

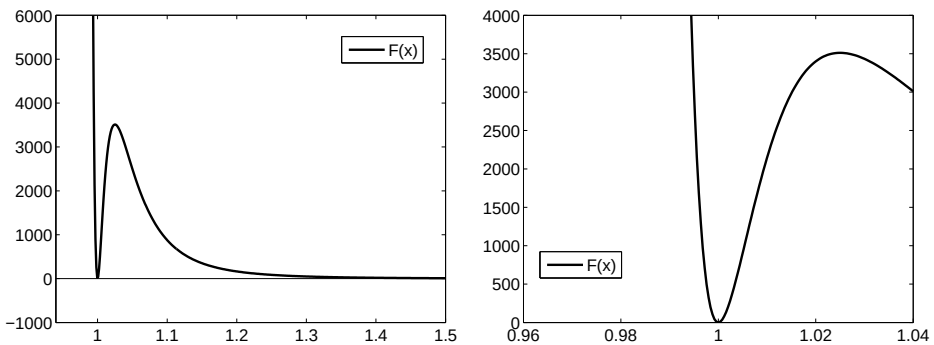


Figura 8.16: Función $F(x) = \dfrac{(1-x)^2}{(x-0.95)^6}$.

En cambio, para valores $x \simeq 1$ el valor de $F(x)$ es grande. En particular,

$$F(0.99) = 2.441406249999991 \times 10^4$$

lo que podría llevar a pensar, erróneamente, que el punto $x = 0.99$ está alejado de la raíz de la función F. □

El test más usual es parar las iteraciones cuando

$$|x_n - x_{n-1}| < \varepsilon |x_{n-1}|$$

siendo $\varepsilon > 0$ la *tolerancia* admisible en el error relativo.

a) Este procedimiento funciona especialmente bien cuando la derivada de la función F es pequeña en un entorno de la raíz.

b) Cuando la derivada de F es grande en entornos de la raíz, puede ocurrir que a valores próximos entre sí la función F les asigne valores cuya diferencia sea grande. Teniendo en cuenta que

$$F(x_n) = F(x_n) - F(\xi) \simeq F'(\xi)(x_n - \xi)$$

entonces

$$|F(x_n)| \simeq |F'(\xi)|\, |x_n - \xi|.$$

Por tanto, en el caso de que $|F'(\xi)|$ no sea pequeño, una forma adecuada de ver si x_n está próximo a ξ es comprobar si $|F(x_n)|$ es pequeño.

c) En el método de Newton y en el de las cuerdas sabemos, por las proposiciones 8.2 y 8.3 respectivamente, que si dos iteraciones seguidas están cerca, entonces estamos cerca, módulo las cotas de las derivadas primera y segunda, de la raíz.

En resumen, para actuar con cautela será bueno utilizar, para detener las iteraciones, los dos tests siguientes

$$|x_n - x_{n-1}| < \varepsilon |x_{n-1}| \;\text{ y }\; |F(x_n)| < \varepsilon$$

y no detener el proceso si alguna de ellas falla.

8.6.2. Raíces múltiples

Hasta ahora hemos venido suponiendo que las raíces que queríamos aproximar eran simples. En el caso de que ξ sea raíz de multiplicidad m de la ecuación $F(x) = 0$ entonces podemos escribir

$$F(x) = (x - \xi)^m G(x) \;\text{ con }\; G(\xi) \neq 0. \tag{8.40}$$

De esta forma:

a) Si ξ es una raíz de multiplicidad m de la ecuación $F(x) = 0$ se puede considerar la ecuación

$$(F(x))^{\frac{1}{m}} = 0$$

de la cual ξ es raíz simple.

b) Raíces múltiples en el método de Newton. Si $F \in \mathcal{C}^m([a,b])$, $m \geq 2$ y $\xi \in [a,b]$ es un cero de multiplicidad m de F, podemos expresar F en la forma (8.40). Por tanto, la función

$$f(x) = x - \frac{F(x)}{F'(x)}$$

del método de Newton es continua y diferenciable en $[a,b]$, $f(\xi) = \xi$ (basta aplicar reiteradas veces la regla de L'Hôpital) y

$$f'(\xi) = 1 - \frac{1}{m}. \tag{8.41}$$

En efecto, de la relación (8.40) es inmediato calcular (se deja como ejercicio al lector) las dos primeras derivadas de la función F, que son

$$F'(x) = (mG(x) + (x-\xi)G'(x))\,(x-\xi)^{m-1}$$

y

$$F''(x) = \big(m(m-1)G(x) + 2m(x-\xi)G'(x) + (x-\xi)^2 G''(x)\big)\,(x-\xi)^{m-2}.$$

De esta forma, de la expresión

$$\begin{aligned} f'(x) &= 1 - \frac{(F'(x))^2 - F(x)F''(x)}{(F'(x))^2} = \frac{F(x)F''(x)}{(F'(x))^2} \\ &= \frac{\big(m(m-1)G(x) + 2m(x-\xi)G'(x) + (x-\xi)^2 G''(x)\big)\,G(x)}{m^2(G(x))^2 + 2m(x-\xi)G(x)G'(x) + (x-\xi)^2(G'(x))^2} \end{aligned}$$

se deduce que

$$f'(\xi) = \frac{m(m-1)(G(\xi))^2}{m^2(G(\xi))^2} = \frac{m-1}{m} = 1 - \frac{1}{m}$$

como queríamos ver. Por tanto, al ser $f'(\xi) \neq 0$, la convergencia del método ya no será cuadrática (véase el problema 8.3). No obstante, es posible recuperar la convergencia cuadrática del método considerando la función

$$\widetilde{f}(x) = x - m\frac{F(x)}{F'(x)}. \tag{8.42}$$

En efecto, expresando $\widetilde{f}$ en términos de f se tiene que

$$\widetilde{f}(x) = x - m(x - f(x)) = mf(x) - (m-1)x.$$

De esta forma

$$\widetilde{f}(\xi) = mf(\xi) - (m-1)\xi = \xi,$$

pues $f(\xi) = \xi$. De la relación

$$\widetilde{f}'(x) = mf'(x) - (m-1)$$

y de (8.41) se obtiene que

$$\widetilde{f}'(\xi) = mf'(\xi) - (m-1) = m\left(1 - \frac{1}{m}\right) - (m-1) = 0.$$

Por tanto, la sucesión

$$\begin{cases} x_0 \in [a,b] \text{ arbitrario} \\ x_n = \widetilde{f}(x_{n-1}) = x_{n-1} - m\dfrac{F(x_{n-1})}{F'(x_{n-1})},\ n \in \mathbb{N} \end{cases}$$

converge, al menos cuadráticamente, a la raíz ξ de F en $[a,b]$ (véase, nuevamente, el problema 8.3).

c) En general, si ξ es raíz múltiple de $F(x) = 0$ con multiplicidad desconocida entonces ξ es raíz simple de la ecuación

$$\frac{F(x)}{F'(x)} = 0.$$

En efecto, basta observar que si m es la multiplicidad de ξ entonces, a partir de (8.40), se tiene que

$$\frac{F(x)}{F'(x)} = \frac{(x-\xi)^m G(x)}{m(x-\xi)^{m-1}G(x) + (x-\xi)^m G'(x)} = \frac{(x-\xi)G(x)}{mG(x) + (x-\xi)G'(x)}. \quad \square$$

8.7. Problemas

8.7.1. Problemas resueltos

8.1. Utilizar el método de la bisección para aproximar una raíz de la ecuación

$$\sqrt{x}\,\mathrm{sen}\,x - x^3 + 2 = 0$$

en el intervalo $[1,2]$ con un error menor que $\dfrac{1}{30}$.

SOLUCIÓN. Consideramos la función

$$F(x) = \sqrt{x}\,\operatorname{sen} x - x^3 + 2,\ x > 0.$$

Como $F \in \mathcal{C}([1,2])$ y $F(1) > 0 > F(2)$, por el teorema de Bolzano se sabe que existe $\xi \in (1,2)$ tal que $F(\xi) = 0$. Según la fórmula (8.7) se verifica que

$$\left|\xi - \frac{a_n + b_n}{2}\right| \leq \frac{b_n - a_n}{2} \leq \frac{b-a}{2^{n+1}} = \frac{1}{2^{n+1}}$$

por lo que basta tomar $n = 4$ para que se cumpla que

$$\left|\xi - \frac{a_4 + b_4}{2}\right| \leq \frac{1}{2^5} < \frac{1}{30}.$$

De esta forma podemos formar la tabla:

n	a_n	$\frac{a_n+b_n}{2}$	b_n	$F(a_n)$	$F\left(\frac{a_n+b_n}{2}\right)$	$F(b_n)$
0	1	1.5	2	1.841471	−0.153323	−4.714059
1	1	1.25	1.5	1.841471	1.107872	−0.153323
2	1.25	1.375	1.5	1.107872	0.550590	−0.153323
3	1.375	1.4375	1.5	0.550590	0.217863	−0.153323
4	1.4375	1.46875	1.5	0.217863	0.037189	−0.153323

Por tanto, para $n = 4$ podemos tomar como valor aproximado $\widetilde{\xi} = 1.46875$. Compárese este resultado con el valor de $\xi = 1.47497971328\ldots$ □

8.2. Determinar un intervalo y una función para poder aplicar el método del Punto Fijo a las siguientes ecuaciones:

a) $x^3 - x - 1 = 0$.

b) $4 - x - \tan x = 0$.

Determinar, en cada caso, el número de iteraciones necesario para que el error cometido sea inferior a 10^{-5}.

SOLUCIÓN.

a) Consideramos la función

$$F(x) = x^3 - x - 1.$$

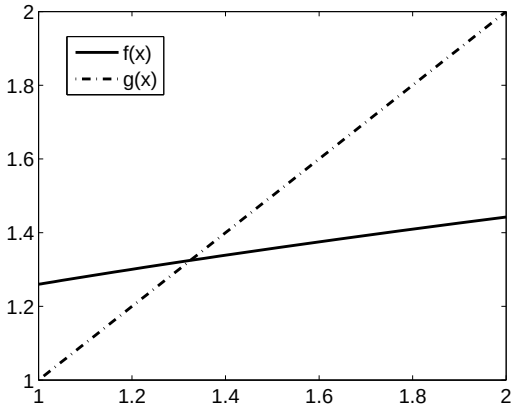


Figura 8.17: Funciones $f(x) = \sqrt[3]{1+x}$ y $g(x) = x$.

Como $F \in \mathcal{C}([1,2])$ y $F(1) = -1 < 0 < 5 = F(2)$, por el teorema de Bolzano existe $\xi \in (1,2)$ tal que $F(\xi) = 0$. Por otra parte,

$$F(x) = 0 \Leftrightarrow x^3 - x - 1 = 0 \Leftrightarrow x^3 = 1 + x \Leftrightarrow x = \sqrt[3]{1+x} \Leftrightarrow f(x) = x$$

siendo

$$f(x) = \sqrt[3]{1+x}.$$

Como $f \in \mathcal{C}^1([1,2])$ y

$$f'(x) = \frac{1}{3\sqrt[3]{(1+x)^2}} > 0,\ x \in [1,2]$$

se verifica que la función f es estrictamente creciente en $[1,2]$; por otra parte,

$$f(1) = \sqrt[3]{2} \simeq 1.2599 > 1 \ \text{ y } \ f(2) = \sqrt[3]{3} \simeq 1.4422 < 2$$

lo que hace que se tenga que

$$f([1,2]) = [\sqrt[3]{2}, \sqrt[3]{3}] \subset [1,2].$$

Finalmente,

$$|f'(x)| = \frac{1}{3\sqrt[3]{(1+x)^2}} \leq \frac{1}{3\sqrt[3]{4}} < 1,\ x \in [1,2]$$

por lo que f es contractiva en $[1,2]$ de constante $k = \dfrac{1}{3\sqrt[3]{4}} \simeq 0.2100$. Por tanto, por el teorema del Punto Fijo, la función f tiene un único punto

fijo en el intervalo $[1,2]$ que, por construcción, es el punto ξ. Además, si consideramos la sucesión

$$\begin{cases} x_0 \in [1,2] \\ x_n = \sqrt[3]{1+x_{n-1}},\ n \in \mathbb{N} \end{cases}$$

se verifica que

$$\lim_{n\to+\infty} x_n = \xi.$$

Para dar un valor aproximado de ξ de forma que el error cometido sea inferior a 10^{-5}, a la vista de (8.13), basta tomar $n \in \mathbb{N}$ de forma que

$$\frac{k^n}{1-k}|x_1 - x_0| < 10^{-5}.$$

Por tanto, si se comienza en $x_0 = 1$, tomando $n \in \mathbb{N}$ de forma que

$$\frac{\left(\dfrac{1}{3\sqrt[3]{4}}\right)^n}{1-\dfrac{1}{3\sqrt[3]{4}}}\left(\sqrt[3]{2}-1\right) < 10^{-5} \Leftrightarrow n > 1 + \frac{\ln\left(\dfrac{10^5(\sqrt[3]{2}-1)}{3\sqrt[3]{4}-1}\right)}{\ln\left(3\sqrt[3]{4}\right)} \simeq 6.6644,$$

el término $x_7 = 1.324715$ verifica que $|x_7 - \xi| < 10^{-5}$.

b) Como se observa en la figura 8.18, la función

$$G(x) = 4 - x - \tan x$$

tiene infinitas raíces.

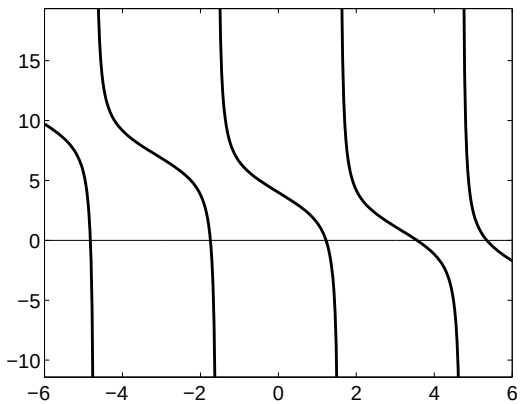


Figura 8.18: Función $G(x) = 4 - x - \tan x$.

Lo que haremos será dar un valor aproximado de la primera raíz positiva de G que, como veremos, se encuentra en $\left(1, \frac{\pi}{2}\right)$. Dado que, si $x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$,

$$G(x) = 0 \Leftrightarrow \tan x = 4 - x \Leftrightarrow x = \arctan(4 - x) \Leftrightarrow x = f(x)$$

siendo

$$f(x) = \arctan(4 - x),$$

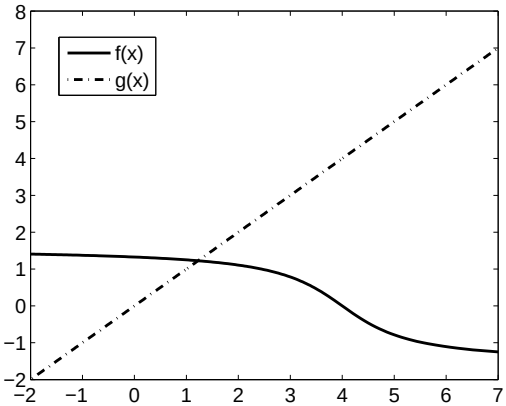


Figura 8.19: Funciones $f(x) = \arctan(4 - x)$ y $g(x) = x$.

vamos a considerar la función

$$F(x) = x - f(x) = x - \arctan(4 - x).$$

Como $F \in \mathcal{C}\left(\left[1, \frac{\pi}{2}\right]\right)$ y $F(1) < 0 < F\left(\frac{\pi}{2}\right)$, por el teorema de Bolzano existe $\xi \in \left(1, \frac{\pi}{2}\right)$ tal que $F(\xi) = 0$. Por otra parte, $f \in \mathcal{C}^1(\mathbb{R})$ y

$$f'(x) = -\frac{1}{1 + (4 - x)^2} < 0,\ x \in \mathbb{R}$$

por lo que la función f es estrictamente decreciente en $\mathbb{R}$; además,

$$f(1) = \arctan 3 \simeq 1.2490 < \frac{\pi}{2}$$

y

$$f\left(\frac{\pi}{2}\right) = \arctan\left(4 - \frac{\pi}{2}\right) \simeq 1.1803 > 1,$$

lo que hace que se tenga que

$$f\left(\left[1,\frac{\pi}{2}\right]\right)=\left[\arctan\left(4-\frac{\pi}{2}\right),\arctan 3\right]\subset\left[1,\frac{\pi}{2}\right].$$

Finalmente, para todo $x\in\left[1,\dfrac{\pi}{2}\right]$,

$$|f'(x)|=\frac{1}{1+(4-x)^2}\leq\frac{1}{1+\left(4-\dfrac{\pi}{2}\right)^2}\simeq 0.1449<1$$

y, en consecuencia, f es contractiva en $\left[1,\dfrac{\pi}{2}\right]$ de constante $k=\dfrac{1}{1+\left(4-\dfrac{\pi}{2}\right)^2}$.

Por tanto, la sucesión

$$\begin{cases} x_0\in\left[1,\dfrac{\pi}{2}\right] \\ x_n=\arctan(4-x_{n-1}),\ n\in\mathbb{N} \end{cases}$$

converge a la única raíz ξ de la ecuación $F(x)=0$ en el intervalo $\left[1,\dfrac{\pi}{2}\right]$. Para dar un valor aproximado de ξ de forma que el error cometido sea inferior a 10^{-5}, si comenzamos en $x_0=1$, basta tomar $n\in\mathbb{N}$ de forma que

$$\frac{k^n}{1-k}(\arctan 3-1)<10^{-5}\ \Leftrightarrow\ n>\frac{\ln\left(\dfrac{10^{-5}(1-k)}{\arctan 3-1}\right)}{\ln k}\simeq 5.3215.$$

Luego el término $x_6\simeq 1.224929$ verifica que $|x_6-\xi|<10^{-5}$. □

8.3. Sea $f\in\mathcal{C}([a,b])$, f derivable en (a,b) tal que $f([a,b])\subset[a,b]$ y existe $\lim\limits_{n\to+\infty}x_n=\xi$ siendo

$$\begin{cases} x_0\in[a,b]\ \text{ arbitrario} \\ x_n=f(x_{n-1}),\ n\in\mathbb{N} \end{cases}$$

(por tanto, $f(\xi)=\xi$). Denotando por

$$e_n=|x_n-\xi|,\ n\in\mathbb{N}\cup\{0\}$$

demostrar los siguientes resultados:

a) Si $x_0\neq\xi$ y

$$f'(x)\neq 0,\ x\in(a,b) \tag{8.43}$$

entonces

$$e_n \neq 0,\ n \in \mathbb{N} \cup \{0\}$$

(es decir, o la sucesión empieza en el punto fijo ξ de f –en cuyo caso se mantiene constante– o nunca converge en un número finito de pasos).

b) Si existe $M \geq 0$ tal que

$$|f'(x)| \leq M,\ x \in (a, b)$$

entonces

$$\frac{e_n}{e_{n-1}} \leq M,\ n \in \mathbb{N}$$

por lo que la sucesión $\{x_n\}_{n=0}^{\infty}$ converge, al menos linealmente, a ξ.

c) Si $f \in \mathcal{C}^2([a, b])$ entonces la convergencia de la sucesión $\{x_n\}_{n=0}^{\infty}$ es, al menos, cuadrática si y solo si $f'(\xi) = 0$.

Solución.

a) Si para algún $n \in \mathbb{N}$ se verificara

$$0 = e_n = |x_n - \xi|,$$

entonces x_n sería punto fijo de f en $[a, b]$ y, por tanto,

$$f(x_n) = x_n = f(x_{n-1}).$$

De esta forma, por el teorema del Valor Medio, se tendría que

$$0 = |x_n - x_n| = |f(x_n) - f(x_{n-1})| = |f'(\eta_n)|\,|x_n - x_{n-1}|$$

de donde, por (8.43), se concluye que

$$x_{n-1} = x_n = \xi.$$

Reiterando este argumento se llegaría a

$$x_n = x_{n-1} = x_{n-2} = \cdots = x_1 = x_0 = \xi,$$

obteniendo la contradicción $x_0 = \xi$.

b) Por ser ξ punto fijo de f, aplicando nuevamente el teorema del Valor Medio, se obtiene, para cada $n \in \mathbb{N}$, que

$$\begin{aligned} e_n = |x_n - \xi| &= |f(x_{n-1}) - f(\xi)| = |f'(\eta_{n-1})|\,|x_{n-1} - \xi| \\ &= |f'(\eta_{n-1})|\,e_{n-1} \leq M e_{n-1}, \end{aligned}$$

de donde se sigue el resultado.

c) A partir de un desarrollo de Taylor de segundo orden, teniendo nuevamente en cuenta que ξ es punto fijo de f, para todo $n \in \mathbb{N}$ se verifica que

$$\begin{aligned} e_n &= |x_n - \xi| = |f(x_{n-1}) - f(\xi)| \\ &= \left| f'(\xi)\,(x_{n-1} - \xi) + \frac{f''(\nu_{n-1})}{2}\,(x_{n-1} - \xi)^2 \right| \\ &= \left| f'(\xi)\,\frac{x_{n-1} - \xi}{(e_{n-1})^2} + \frac{f''(\nu_{n-1})}{2} \right| (e_{n-1})^2. \end{aligned}$$

De esta forma,

$$\frac{e_n}{(e_{n-1})^2} = \left| f'(\xi)\frac{x_{n-1} - \xi}{(e_{n-1})^2} + \frac{f''(\nu_{n-1})}{2} \right|$$

de donde se sigue el resultado. □

8.4. Se considera la ecuación $x^2 - 1 - \operatorname{sen} x = 0$.

a) Probar que dicha ecuación tiene, al menos, una raíz positiva.

b) Encontrar un intervalo en el cual la iteración

$$x_n = \sqrt{1 + \operatorname{sen} x_{n-1}},\ n \in \mathbb{N}$$

converja, para cualquier valor inicial x_0 de dicho intervalo, a una raíz positiva de la ecuación anterior. ¿Cuántos pasos deben darse, a partir de $x_0 = \dfrac{\pi}{2}$, para obtener una aproximación de la raíz con un error inferior a la milésima?

SOLUCIÓN.

a) Como la función

$$F(x) = x^2 - 1 - \operatorname{sen} x$$

verifica que $F \in \mathcal{C}([1,2])$ y

$$F(1) = -\operatorname{sen} 1 < 0 < 3 - \operatorname{sen} 2 = F(2),$$

aplicando el teorema de Bolzano, existe $\xi \in (1,2)$ tal que $F(\xi) = 0$.

b) Claramente, para $x \in [1,2]$, se cumple que

$$F(x) = 0 \Leftrightarrow x^2 = 1 + \operatorname{sen} x \Leftrightarrow f(x) = x$$

siendo

$$f(x) = \sqrt{1 + \operatorname{sen} x}.$$

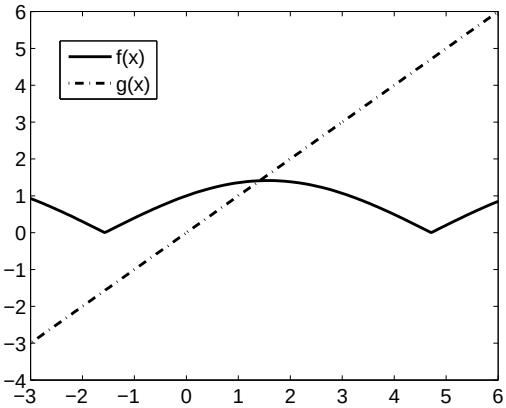


Figura 8.20: Funciones $f(x) = \sqrt{1 + \operatorname{sen} x}$ y $g(x) = x$.

Como para todo $x \in [1, 2]$ se verifica que

$$f'(x) = \frac{\cos x}{2\sqrt{1 + \operatorname{sen} x}}$$

y

$$f''(x) = \frac{-\operatorname{sen} x\sqrt{1 + \operatorname{sen} x} - \dfrac{\cos^2 x}{2\sqrt{1 + \operatorname{sen} x}}}{2(1 + \operatorname{sen} x)} = -\frac{2\operatorname{sen} x(1 + \operatorname{sen} x) + \cos^2 x}{4\sqrt{(1 + \operatorname{sen} x)^3}}$$
$$= -\frac{\operatorname{sen}^2 x + 2\operatorname{sen} x + 1}{4\sqrt{(1 + \operatorname{sen} x)^3}} = -\frac{(1 + \operatorname{sen} x)^2}{4(1 + \operatorname{sen} x)^{\frac{3}{2}}} = -\frac{\sqrt{1 + \operatorname{sen} x}}{4} < 0,$$

la función f' es estrictamente decreciente en el intervalo $[1, 2]$ y, por tanto,

$$-0.1506 \simeq f'(2) \leq f'(x) \leq f'(1) \simeq 0.1991,\ x \in [1, 2].$$

De esta forma, la función f es contractiva en el intervalo $[1, 2]$ de constante

$$k = \max_{1 \leq x \leq 2} |f'(x)| = f'(1) = \frac{\cos 1}{2\sqrt{1 + \operatorname{sen} 1}} \simeq 0.1991.$$

Por otra parte,

$$f'(x) \begin{cases} > 0, & x \in \left[1, \dfrac{\pi}{2}\right) \\ = 0, & x = \dfrac{\pi}{2} \\ < 0, & x \in \left(\dfrac{\pi}{2}, 2\right] \end{cases}$$

por lo que la función f es creciente en el intervalo $\left[1, \dfrac{\pi}{2}\right)$ y decrececiente en $\left(\dfrac{\pi}{2}, 2\right]$, siendo $\widetilde{x} = \dfrac{\pi}{2}$ un máximo relativo de la función f. Como

$$\begin{cases} f(1) = \sqrt{1 + \operatorname{sen} 1} \simeq 1.3570 \\ f\left(\dfrac{\pi}{2}\right) = \sqrt{2} \simeq 1.4142 \\ f(2) = \sqrt{1 + \operatorname{sen} 2} \simeq 1.3818 \end{cases}$$

entonces

$$\max_{1 \le x \le 2} f(x) = f\left(\frac{\pi}{2}\right) \quad \text{y} \quad \min_{1 \le x \le 2} f(x) = f(1).$$

De esta forma,

$$1 < f(1) \le f(x) \le f\left(\frac{\pi}{2}\right) < 2, \; x \in [1, 2],$$

es decir,

$$f([1, 2]) \subset [1, 2].$$

Por tanto, la sucesión

$$\begin{cases} x_0 \in [1, 2] \\ x_n = \sqrt{1 + \operatorname{sen} x_{n-1}}, \; n \in \mathbb{N} \end{cases}$$

verifica que

$$\lim_{n \to +\infty} x_n = \xi$$

siendo $\xi \in [1, 2]$ la única raíz de la función F en el intervalo $[1, 2]$. Tomando como valor inicial $x_0 = \dfrac{\pi}{2}$ se tiene que

$$|x_n - \xi| \le \frac{k^n}{1 - k} |x_1 - x_0| = \frac{\left(\dfrac{\cos 1}{2\sqrt{1 + \operatorname{sen} 1}}\right)^n}{1 - \dfrac{\cos 1}{2\sqrt{1 + \operatorname{sen} 1}}} \left|\sqrt{2} - \frac{\pi}{2}\right| < 10^{-3} \;\text{ si }\; n \ge 4.$$

Los primeros términos de la sucesión anterior vienen dados en la siguiente tabla:

n	x_n
0	1.570796
1	1.414214
2	1.409881
3	1.409639
4	1.409625

8.5. Se considera la función

$$g(x) = \operatorname{sen}^2 x,\ x \in [0, \pi].$$

Aplicar el teorema del Punto Fijo a la función

$$f(x) = 2 + \frac{\operatorname{sen} 2x}{2}$$

en el intervalo $[1.6, 2]$ para determinar el valor del punto ζ para el que se verifica que

$$\int_0^{\zeta} g(t)\, dt = 1 \tag{8.44}$$

de forma que el error cometido sea inferior a 10^{-4}.

Solución. Claramente

$$\begin{aligned}\int_0^x g(t)\, dt &= \int_0^x \operatorname{sen}^2 t\, dt = \frac{1}{2}\int_0^x (1 - \cos 2t)\, dt \\ &= \frac{1}{2}\left(t - \frac{\operatorname{sen} 2t}{2}\right)\Big|_0^x = \frac{1}{2}\left(x - \frac{\operatorname{sen} 2x}{2}\right).\end{aligned}$$

Puesto que

$$\frac{1}{2}\left(1.6 - \frac{\operatorname{sen} 3.2}{2}\right) \simeq 0.814594 < 1 < 1.189201 \simeq \frac{1}{2}\left(2 - \frac{\operatorname{sen} 4}{2}\right)$$

el valor de ζ buscado estará en el intervalo $[1.6, 2]$ (véase la figura 8.21).

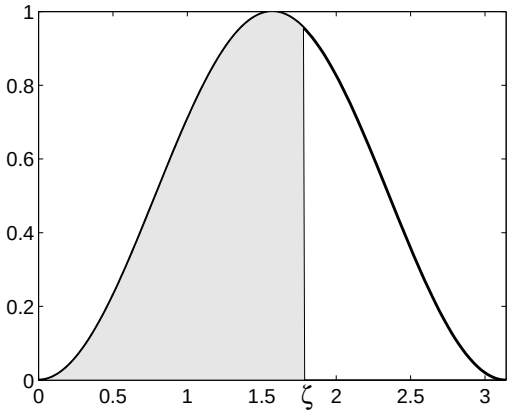


Figura 8.21: Función $g(x) = \operatorname{sen}^2 x$.

De esta forma, para que se satisfaga la propiedad (8.44), el punto ζ debe cumplir

$$\frac{1}{2}\left(\zeta - \frac{\operatorname{sen} 2\zeta}{2}\right) = 1 \Leftrightarrow \zeta - \frac{\operatorname{sen} 2\zeta}{2} = 2 \Leftrightarrow \zeta = f(\zeta)$$

siendo

$$f(x) = 2 + \frac{\operatorname{sen} 2x}{2},\ x \in [1.6, 2].$$

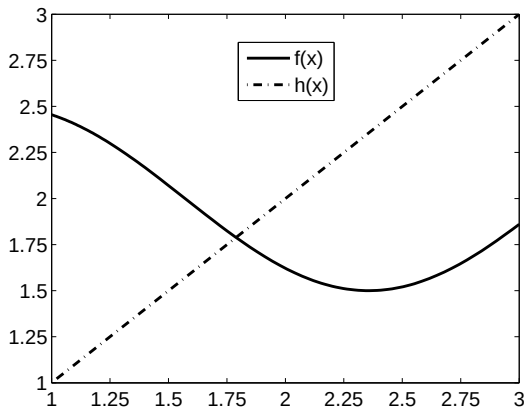


Figura 8.22: Funciones $f(x) = 2 + \dfrac{\operatorname{sen} 2x}{2}$ y $h(x) = x$.

Veamos que la función f verifica las hipótesis del teorema del Punto Fijo en dicho intervalo:

i) Como

$$f'(x) = \cos 2x < 0,\ x \in [1.6, 2]$$

se verifica que la función f es estrictamente decreciente en ese intervalo, por lo que

$$f([1.6, 2]) = [f(2), f(1.6)] = \left[2 + \frac{\operatorname{sen} 4}{2}, 2 + \frac{\operatorname{sen} 3.2}{2}\right]$$
$$\simeq [1.621599, 1.970813] \subset [1.6, 2].$$

ii) La propiedad

$$|f'(x)| \leq -\cos 3.2,\ x \in [1.6, 2]$$

hace que la función f sea contractiva en el intervalo $[1.6, 2]$ de constante $k = -\cos 3.2 \simeq 0.998295$.

Por tanto, por el teorema del Punto Fijo se tiene que existe un único punto $\zeta \in [1.6, 2]$ tal que $\zeta = f(\zeta)$. Además, la sucesión

$$\begin{cases} x_0 \in [1.6, 2] \\ x_n = f(x_{n-1}) = 2 + \dfrac{\operatorname{sen} 2x_{n-1}}{2},\ n \in \mathbb{N} \end{cases}$$

verifica que

$$\lim_{n \to +\infty} x_n = \zeta.$$

Finalmente, para obtener la aproximación con la precisión deseada, si se comienza en $x_0 = 2$, debemos tomar $n \in \mathbb{N}$ de forma que

$$\frac{(-\cos 3.2)^n}{1 + \cos 3.2} \left(-\frac{\operatorname{sen} 4}{2} \right) < 10^{-4},$$

es decir,

$$n > \frac{\ln\left(-\dfrac{2(1 + \cos 3.2)}{\operatorname{sen} 4 \times 10^4} \right)}{\ln(-\cos 3.2)} \simeq 8562.0040.$$

Un valor aproximado es $\zeta \simeq 1.78882000599$. Como se observa, la convergencia es muy lenta (hacen falta 8563 iteraciones) y esto es debido a que la constante de contractividad de f está muy próxima a 1. Nótese también que la convergencia no puede mejorarse sustancialmente cambiando el intervalo de trabajo, puesto que

$$|f'(\zeta)| = -\cos 2\zeta \simeq 0.906428$$

que es un valor muy próximo a 1. □

8.6. El objetivo de este problema es determinar los valores del parámetro $\mu > 0$ para los cuales las gráficas de las funciones

$$f(x) = e^x \ \text{ y } \ g(x) = \mu - x^2$$

son tangentes. Para ello:

a) Determinar el número de raíces reales de la ecuación

$$e^x = \lambda x$$

en función del parámetro $\lambda \in \mathbb{R}$.

b) Utilizar el teorema del Punto Fijo para aproximar la única solución de la ecuación

$$e^x = -2x$$

de forma que el error cometido sea inferior a 10^{-6}.

c) Concluir el resultado.

SOLUCIÓN.

a) Consideremos la función $F(x) = e^x - \lambda x$. Obviamente,

$$F'(x) = e^x - \lambda \ \text{ y } \ F''(x) = e^x > 0.$$

Por tanto, F' es estrictamente creciente en $\mathbb{R}$. Analicemos el número de raíces de F en función de λ:

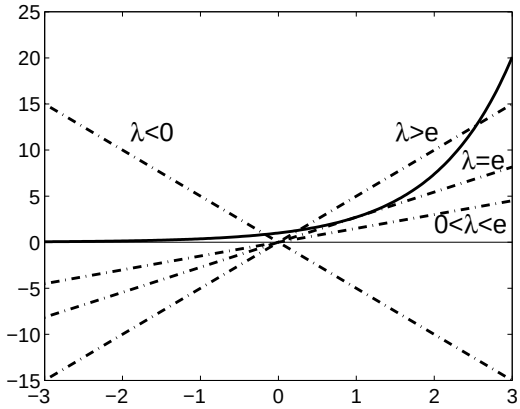


Figura 8.23: Gráficas de $f(x) = e^x$ y $\varphi(x) = \lambda x$ para $\lambda \in \mathbb{R}$.

i) Si $\lambda < 0$ se verifica que $F' > 0$ en $\mathbb{R}$, por lo que F es estrictamente creciente. Como

$$\lim_{x \to \pm\infty} F(x) = \pm\infty,$$

la función F tendrá una única raíz.

ii) Si $\lambda = 0$ es evidente que F no se anula nunca.

iii) Si $\lambda > 0$, puesto que

$$F'(x) \begin{cases} > 0 & \text{si} \quad x > \ln \lambda \\ < 0 & \text{si} \quad x < \ln \lambda, \end{cases}$$

F decrece en $(-\infty, \ln \lambda)$ y crece en $(\ln \lambda, +\infty)$; es decir, la función F tiene un mínimo en $\ln \lambda$. Por otra parte, como

$$F(\ln \lambda) = \lambda(1 - \ln \lambda)$$

podemos distinguir los siguientes casos para el signo de $F(\ln \lambda)$:

- $\boxed{0 < \lambda < e}$ F no tiene raíces pues su mínimo es positivo.
- $\boxed{\lambda = e}$ F se anula únicamente en $x = 1$, que será una raíz doble, al ser $F(1) = F'(1) = 0$.
- $\boxed{\lambda > e}$ Como $F(\ln \lambda) < 0$ y

$$\lim_{x \to -\infty} F(x) = +\infty$$

y F es monótona en $(-\infty, \ln \lambda)$, tiene una única raíz en dicho intervalo. Análogamente, por ser

$$\lim_{x \to +\infty} F(x) = +\infty$$

y F monótona, tiene una única raíz en el intervalo $(\ln \lambda, +\infty)$. Así pues, F tendrá dos raíces reales.

En la figura 8.23 se representan gráficas de las funciones e^x y λx para diversos valores de λ.

b) Escribimos la ecuación como $F(x) = 0$ siendo

$$F(x) = e^x + 2x.$$

Por el apartado *a*), esta ecuación tiene una única raíz. Como

$$F(-1) = \frac{1}{e} - 2 < 0 \ \text{ y } \ F(0) = 1,$$

la única raíz se encuentra en el intervalo $(-1, 0)$. Consideremos la función

$$h(x) = -\frac{e^x}{2}.$$

Al ser

$$h'(x) = -\frac{e^x}{2} < 0,$$

la función h es estrictamente decreciente en dicho intervalo. Como

$$h(-1) = -\frac{1}{2e} < 0 \ \text{ y } \ h(0) = -\frac{1}{2} > -1,$$

se verifica que

$$h([-1, 0]) = \left[-\frac{1}{2}, -\frac{1}{2e}\right] \subset [-1, 0].$$

Por otra parte, cuando $x \in [-1, 0]$,

$$|h'(x)| = \frac{e^x}{2} < \frac{1}{2},$$

por lo que h es contractiva en $[-1, 0]$ de constante $k = \dfrac{1}{2}$. En consecuencia, la sucesión

$$\begin{cases} x_0 \in [-1, 0] \\ x_n = -\dfrac{e^{x_{n-1}}}{2},\ n \in \mathbb{N} \end{cases}$$

converge a la única raíz de la ecuación. Comenzando en $x_0 = 0$, para obtener la precisión deseada, debemos tomar $n \in \mathbb{N}$ tal que

$$\frac{(0.5)^n}{1 - 0.5} \left| -\frac{e^0}{2} - 0 \right| < 10^{-6} \Leftrightarrow n > \frac{\ln 10^6}{\ln 2} \simeq 19.93,$$

es decir, basta tomar $n = 20$ iteraciones. Si así se hace, se obtiene

$$x_{20} \simeq -0.351734.$$

c) Si denominamos ζ el punto de tangencia entre las gráficas de las funciones f y g, se tendrá que:

$$\begin{cases} f(\zeta) = g(\zeta) & \Rightarrow \quad \mu = e^{\zeta} + \zeta^2 \\ f'(\zeta) = g'(\zeta) & \Rightarrow \quad e^{\zeta} = -2\zeta. \end{cases}$$

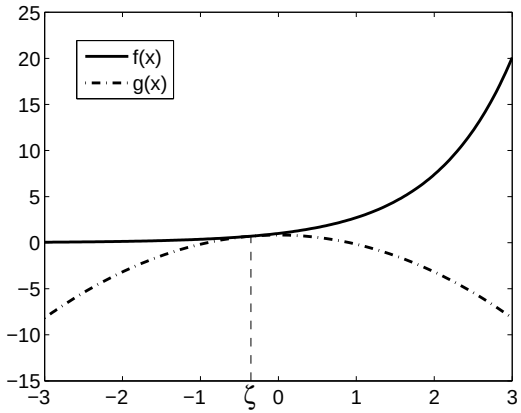


Figura 8.24: Gráficas de $f(x) = e^x$ y $g(x) = 0.827184 - x^2$.

Aplicando el apartado b) se obtiene el valor aproximado

$$\zeta \simeq -0.351734$$

a partir del cual se determina

$$\mu \simeq 0.827184.$$

La figura 8.24 muestra las gráficas de estas funciones para el valor anterior de μ. □

8.7. Aproximar, mediante el método de Punto Fijo, la raíz real de la ecuación

$$x^3 - x^2 - x - 1 = 0$$

con una precisión del orden de 10^{-6}.

SOLUCIÓN. Claramente, para $x \neq 0$,

$$x^3 - x^2 - x - 1 = 0 \Leftrightarrow x = 1 + \frac{1}{x} + \frac{1}{x^2},$$

por lo que vamos a considerar las funciones

$$F(x) = x^3 - x^2 - x - 1 \text{ y } f(x) = 1 + \frac{1}{x} + \frac{1}{x^2}.$$

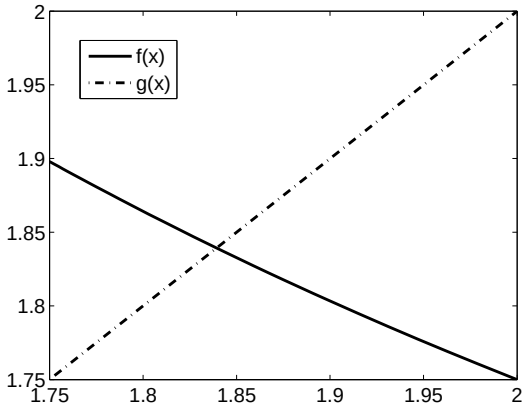


Figura 8.25: Funciones $f(x) = 1 + \frac{1}{x} + \frac{1}{x^2}$ y $g(x) = x$.

Como $F \in \mathcal{C}\left(\left[\frac{7}{4}, 2\right]\right)$ y

$$F\left(\frac{7}{4}\right) = -\frac{29}{64} < 0 < 1 = F(2)$$

por el teorema de Bolzano existe $\xi \in \left[\frac{7}{4}, 2\right]$ tal que

$$F(\xi) = 0 \Leftrightarrow f(\xi) = \xi.$$

Debido a que

$$f'(x) = -\frac{1}{x^2} - \frac{2}{x^3} < 0,\ x \in \left[\frac{7}{4}, 2\right]$$

la función f es estrictamente decreciente, por lo que

$$f\left(\left[\frac{7}{4}, 2\right]\right) = \left[f(2), f\left(\frac{7}{4}\right)\right] = \left[\frac{7}{4}, \frac{93}{49}\right] \subset \left[\frac{7}{4}, 2\right]$$

y, además,

$$|f'(x)| = \frac{1}{x^2} + \frac{2}{x^3} \leq \left(\frac{4}{7}\right)^2 + 2\left(\frac{4}{7}\right)^3 = \frac{240}{343},\ x \in \left[\frac{7}{4}, 2\right].$$

Por lo tanto, la función f es contractiva en $\left[\frac{7}{4}, 2\right]$ de constante

$$k = \frac{240}{343} < 1.$$

Así, la sucesión

$$\begin{cases} x_0 = \dfrac{7}{4} \\ x_n = 1 + \dfrac{1}{x_{n-1}} + \dfrac{1}{x_{n-1}^2},\ n \in \mathbb{N} \end{cases}$$

verifica que

$$\lim_{n \to +\infty} x_n = \xi.$$

A continuación consideramos $n \in \mathbb{N}$ verificando

$$\frac{\left(\frac{240}{343}\right)^n}{1 - \frac{240}{343}} |1.897959 - 1.75| < 10^{-6} \Leftrightarrow n > 36.706816.$$

De esta forma, basta tomar $n = 37$ para que el término

$$x_{37} = 1.83928675683076$$

aproxime a ξ con un error inferior a 10^{-6}. □

8.8. Interpolación inversa. Sea $f : [a, b] \to \mathbb{R}$ inyectiva y $\xi \in [a, b]$ tal que $f(\xi) = 0$. Si $\{x_0, x_1, \ldots, x_n\}$ son $n + 1$ puntos distintos del intervalo $[a, b]$, denotando por

$$y_i = f(x_i)$$

para $i = 0, 1, \ldots, n$, se puede aproximar la raíz ξ de f mediante el polinomio de interpolación de la función f^{-1} en los puntos $\{y_0, y_1, \ldots, y_n\}$: teniendo en cuenta que

$$f^{-1}(y_i) = x_i$$

para $i = 0, 1, \ldots, n$, basta evaluar el polinomio anterior en $y = 0$ para obtener un valor aproximado de

$$f^{-1}(0) = \xi.$$

Como aplicación, utilizar la interpolación inversa para aproximar la raíz de la ecuación

$$2x - \cos x = 0$$

a partir de los datos de la tabla

x_i	$\cos x_i$
0.2	0.98006657784124
0.3	0.95533648912561
0.4	0.92106099400289
0.5	0.87758256189037
0.6	0.82533561490968

SOLUCIÓN. Como la función $f(x) = 2x - \cos x$ verifica que

$$f'(x) = 2 + \operatorname{sen} x > 0,\ x \in \mathbb{R}$$

se tiene que f es inyectiva en todo $\mathbb{R}$, por lo que existe f^{-1}. Además, como

$$f(0) < 0 < f\left(\frac{\pi}{2}\right),$$

la función f tiene una única raíz real ξ que se encuentra en el intervalo $\left[0, \frac{\pi}{2}\right]$ (aquí se han aplicado los teoremas de Bolzano y de Rolle). A partir de la tabla anterior podemos construir esta otra

$y_i = f(x_i)$	x_i
−0.58006657784124	0.2
−0.35533648912561	0.3
−0.12106099400289	0.4
0.12241743810963	0.5
0.37466438509032	0.6

El polinomio de interpolación de la función f^{-1} en los puntos

$$x_i = f^{-1}(y_i)$$

para $i = 0, 1, 2, 3, 4$, es (véase la figura 8.26):

$$P_4(y) = -0.00141103096947y^4 + 0.00683725410099y^3 - 0.03115164458429y^2 + 0.41065496852299y + 0.45018328385781$$

y el valor que toma en $y = 0$ (valor buscado)

$$P_4(0) = 0.45018328385781 \simeq \xi.$$

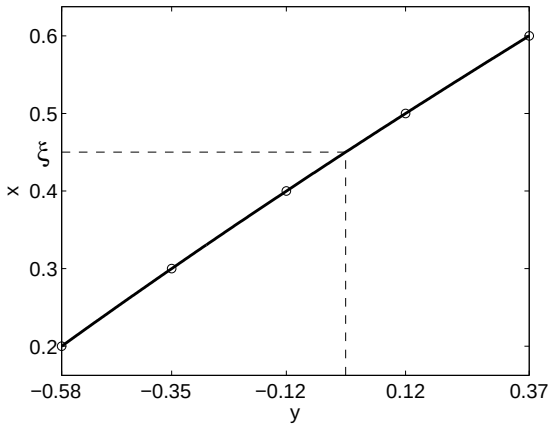


Figura 8.26: Interpolación inversa.

Si se compara con el valor exacto de

$$\xi = 0.45018361129487\ldots$$

vemos que el error que se ha cometido es inferior a la millonésima. □

8.9. Método de Whittaker con parámetro óptimo. Sea F una función que verifica $(\mathcal{H})_1$ y $\xi \in [a, b]$ la única raíz de F en $[a, b]$. Dados μ, $\nu \in \mathbb{R}$ tales que

$$\mu > \nu \geq F'(b) > 0, \tag{8.45}$$

se consideran, a partir de $x_0 \in [a,b]\backslash\{\xi\}$, las sucesiones

$$\begin{cases} y_0 = x_0 \\ y_n = y_{n-1} - \dfrac{F(y_{n-1})}{\mu},\ n \in \mathbb{N} \end{cases} \quad \text{y} \quad \begin{cases} z_0 = x_0 \\ z_n = z_{n-1} - \dfrac{F(z_{n-1})}{\nu},\ n \in \mathbb{N}. \end{cases}$$

En las condiciones anteriores demostrar que se verifica que

$$|z_n - \xi| < |y_n - \xi|,\ n \in \mathbb{N}. \tag{8.46}$$

Concluir que la mejor elección de λ en el método de Whittaker es $\lambda = F'(b)$ (*método de Newton modificado*).

Solución. Supongamos que $\xi < x_0 \leq b$ y probemos que, entonces,

$$\xi < z_n < y_n \leq b,\ n \in \mathbb{N}$$

(de forma análoga se puede probar que si $a \leq x_0 < \xi$ entonces se verifica que $a \leq y_n < z_n < \xi$, $n \in \mathbb{N}$, por lo que uniendo ambos resultados se obtiene la relación (8.46) cualquiera que sea el dato inicial $x_0 \in [a,b]\backslash\{\xi\}$). Para ello, razonamos por inducción:

i) $n = 1$. Puesto que $F > 0$ en $(\xi, b]$

$$z_1 = z_0 - \frac{F(z_0)}{\nu} = y_0 - \frac{F(y_0)}{\nu} < y_0 - \frac{F(y_0)}{\mu} = y_1 < y_0 \leq b.$$

Por otra parte, argumentando como en la demostración del teorema 8.2,

$$0 = F(\xi) = F(z_0) + F'(z_0)(\xi - z_0) + \frac{F''(\eta_0)}{2}(\xi - z_0)^2$$

y, así, como por la hipótesis de inducción $\xi < z_0$,

$$F(z_0) + F'(z_0)(\xi - z_0) < 0.$$

Despejando,

$$\xi < z_0 - \frac{F(z_0)}{F'(z_0)} \leq z_0 - \frac{F(z_0)}{\nu} = z_1.$$

ii) Supongamos cierto el resultado para n. Así, usando que F es estrictamente creciente, tenemos

$$z_{n+1} = z_n - \frac{F(z_n)}{\nu} < z_n - \frac{F(z_n)}{\mu} < y_n - \frac{F(y_n)}{\mu} = y_{n+1} < y_n \leq b,$$

donde se ha utilizado que la función

$$f(x) = x - \frac{F(x)}{\mu}$$

verifica que

$$f'(x) = 1 - \frac{F'(x)}{\mu} > 0$$

y, por tanto, es estrictamente creciente. La demostración de que $\xi < z_{n+1}$ se hace igual que en el caso i).

De esta forma, si μ, $\nu \in \mathbb{R}$ verifican (8.45) entonces el método de Whittaker para ν es más rápido que el método de Whittaker para μ para resolver la ecuación

$$F(x) = 0 \ \text{ en } \ [a, b]$$

cuando comenzamos en un mismo punto $x_0 \neq \xi$. Por tanto, el método de Whittaker óptimo es cuando $\lambda = F'(b)$. □

8.10. Se considera la función $F : \mathbb{R}\backslash\{0\} \to \mathbb{R}$ definida como

$$F(x) = \frac{x-1}{x} - e^{-x}.$$

a) Dibujar la gráfica de F y determinar el número de raíces reales de la ecuación $F(x) = 0$, localizando cada una de sus raíces.

b) Para cada una de las funciones siguientes

$$f_1(x) = 1 + xe^{-x},\ f_2(x) = \ln\left(\frac{x}{x-1}\right),\ f_3(x) = (x-1)e^x$$

estudiar cuáles de las iteraciones sucesivas de f_i, $i = 1, 2, 3$, convergen hacia alguna de las raíces de la ecuación $F(x) = 0$.

c) Elegir intervalos y puntos iniciales adecuados para que el método de Newton converja a cada una de las raíces.

Solución.

a) En primer lugar, tengamos en cuenta que

$$\lim_{x\to-\infty} F(x) = -\infty,\ \lim_{x\to+\infty} F(x) = 1,\ \lim_{x\to0^-} F(x) = +\infty \text{ y } \lim_{x\to0^+} F(x) = -\infty.$$

Por otra parte, la función $F \in \mathcal{C}^2(\mathbb{R}\backslash\{0\})$ y

$$\begin{cases} F'(x) = \dfrac{1}{x^2} + e^{-x}, & x \neq 0 \\ F''(x) = -\dfrac{2}{x^3} - e^{-x}, & x \neq 0. \end{cases}$$

Dado que

$$F'(x) > 0,\ x \neq 0$$

se verifica que la función F es estrictamente creciente en $(-\infty, 0) \cup (0, +\infty)$. Por otra parte, como

$$\begin{cases} F''(x) < 0, & x \in (-\infty, \widetilde{x}) \cup (0, +\infty) \\ F''(x) > 0, & x \in (\widetilde{x}, 0), \end{cases}$$

la función F es cóncava en $(-\infty, \widetilde{x}) \cup (0, +\infty)$ y convexa en $(\widetilde{x}, 0)$, por lo que $\widetilde{x} \in (-\infty, 0)$ es un punto de inflexión de F, es decir, verifica $F''(\widetilde{x}) = 0$. Como

$$\begin{cases} F(-1) = 2 - e \simeq -0.7183 < 0 \\ F(-0.5) = 3 - \sqrt{e} \simeq 1.3513 > 0 \end{cases}$$

el teorema de Bolzano implica que existe $\xi_1 \in (-1, -0.5)$ tal que $F(\xi_1) = 0$. Por otra parte, al ser

$$\begin{cases} F(1) = -\dfrac{1}{e} \simeq -0.3679 < 0 \\ F(2) = \dfrac{1}{2} - \dfrac{1}{e^2} \simeq 0.3647 > 0, \end{cases}$$

nuevamente el teorema de Bolzano implica que existe $\xi_2 \in (1, 2)$ tal que $F(\xi_2) = 0$. Además, la monotonía estricta de la función F en los intervalos $(-\infty, 0)$ y $(0, +\infty)$ hace que ξ_1 y ξ_2 sean las dos únicas raíces reales de la ecuación $F(x) = 0$. Veamos, por otra parte, la localización del punto de inflexión $\widetilde{x}$; como

$$\begin{cases} F''(-1) = 2 - e \simeq -0.7183 < 0 \\ F''(-0.5) = 16 - \sqrt{e} \simeq 14.3513 > 0 \end{cases}$$

el teorema de Bolzano implica que $\widetilde{x} \in (-1, -0.5)$. Como en este mismo intervalo se encuentra también la raíz ξ_1 de $F(x) = 0$ es muy importante saber la posición relativa de ξ_1 y $\widetilde{x}$ con vistas a aplicar el método de Newton. De esta forma,

$$\begin{cases} F''(-1) = 2 - e \simeq -0.7183 < 0 \\ F''(-0.9) = \dfrac{2}{0.9^3} - e^{0.9} \simeq 0.2839 > 0 \end{cases}$$

por lo que, por el teorema de Bolzano, $\widetilde{x} \in (-1, -0.9)$. Estamos ya en condiciones de esbozar la gráfica de la función F como se muestra en la figura 8.27.

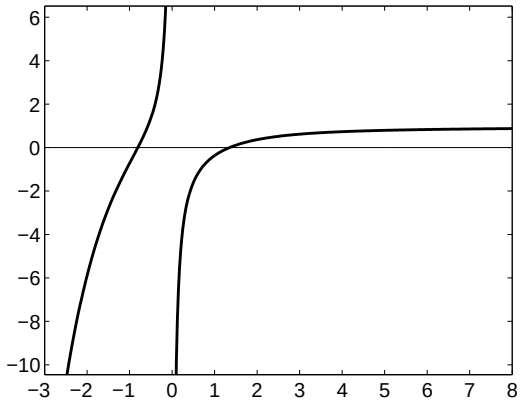


Figura 8.27: Función $F(x) = \dfrac{x-1}{x} - e^{-x}$.

b) En primer lugar observemos que si $x \in (-1, -0.5) \cup (1, 2)$ entonces

$$F(x) = 0 \;\Leftrightarrow\; f_i(x) = x$$

para $i = 1, 2, 3$. En efecto,

$$F(x) = 0 \;\Leftrightarrow\; \begin{cases} x - 1 - xe^{-x} = 0 & \Leftrightarrow \quad x = 1 + xe^{-x} = f_1(x) \\ e^x = \dfrac{x}{x-1} & \Leftrightarrow \quad x = \ln\left(\dfrac{x}{x-1}\right) = f_2(x) \\ xe^{-x} = x - 1 & \Leftrightarrow \quad x = (x-1)e^x = f_3(x). \end{cases}$$

Analicemos cada uno de los casos:

i) $f_1(x) = 1 + xe^{-x} \;\Rightarrow\; \begin{cases} f_1'(x) = (1-x)e^{-x}, & x \in \mathbb{R} \\ f_1''(x) = (x-2)e^{-x}, & x \in \mathbb{R}. \end{cases}$

α) Intervalo $[1, 2]$. Dado que

$$f_1'(x) \leq 0 \;\text{ y }\; f_1''(x) \leq 0 \;\text{ para }\; x \in [1, 2],$$

las funciones f_1 y f_1' son no crecientes en el intervalo $[1, 2]$ y, por tanto,

$$1.2707 \simeq 1 + \frac{2}{e^2} = f_1(2) \leq f_1(x) \leq f_1(1) = 1 + \frac{1}{e} \simeq 1.3679, x \in [1, 2]$$

y

$$-0.1353 \simeq -\frac{1}{e^2} = f_1'(2) \leq f_1'(x) \leq f_1'(1) = 0,\ x \in [1,2].$$

Consecuentemente, la función $f_1 \in \mathcal{C}^1([1,2])$ verifica

$$f_1([1,2]) = \left[1 + \frac{2}{e^2}, 1 + \frac{1}{e}\right] \subset [1,2]$$

y

$$|f_1'(x)| \leq \frac{1}{e^2},\ x \in [1,2]$$

por lo que f_1 es contractiva en $[1,2]$ de constante $k = \dfrac{1}{e^2} \simeq 0.1353$. Por tanto, por el teorema del Punto Fijo, la sucesión

$$\begin{cases} x_0 \in [1,2] \\ x_n = 1 + x_{n-1}e^{-x_{n-1}},\ n \in \mathbb{N} \end{cases}$$

converge a la raíz de F en $[1,2]$. De hecho, podemos aproximar con esta sucesión la raíz, obteniendo

$$\xi_2 \simeq 1.349976.$$

β) Intervalo $[-1,-0.5]$. Como

$$f_1''(x) < 0,\ x \in [-1,-0.5]$$

se verifica que la función f_1' es estrictamente decreciente en el intervalo $[-1,-0.5]$ por lo que

$$|f_1'(x)| = f_1'(x) \geq f_1'(-0.5) = \frac{3}{2}\sqrt{e} > \frac{3}{2} > 1,\ x \in [-1,-0.5].$$

Aplicando el teorema del Valor Medio se tiene que

$$|f_1(x) - f_1(y)| = |f_1'(\eta)||x-y| > \frac{3}{2}\sqrt{e}|x-y|$$

por lo que la función f_1 no es contractiva en $[-1,-0.5]$.

Puede comprobarse (véase la figura 8.28) que comenzando a iterar a la derecha de la raíz ξ_1, los valores de la iteración "saltan" hacia la otra raíz y la sucesión converge a ξ_2; si se comienza a la izquierda de ξ_1, la sucesión tiende a $-\infty$.

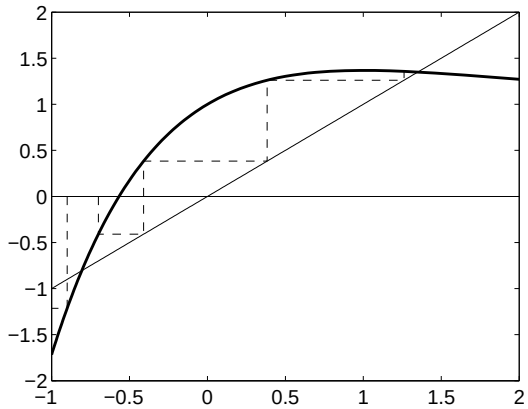


Figura 8.28: Función $f_1(x) = 1 + xe^{-x}$.

ii) $f_2(x) = \ln\left(\dfrac{x}{x-1}\right)$. Como

$$x(x-1) > 0 \;\Leftrightarrow\; x \in (-\infty, 0) \cup (1, +\infty),$$

el dominio de definición de la función f_2 es $(-\infty, 0) \cup (1, +\infty)$. Además,

$$\begin{cases} f_2'(x) = -\dfrac{1}{x(x-1)}, & x \in (-\infty, 0) \cup (1, +\infty) \\ f_2''(x) = \dfrac{2x-1}{x^2(x-1)^2}, & x \in (-\infty, 0) \cup (1, +\infty). \end{cases}$$

α) Intervalo $[-1, -0.5]$. Como ahora

$$f_2'(x) < 0 \;\text{ y }\; f_2''(x) < 0 \;\text{ para }\; x \in [-1, -0.5],$$

las funciones f_2 y f_2' son estrictamente decrecientes en el intervalo $[-1, -0.5]$ y, por tanto, para puntos $x \in [-1, -0.5]$ se verifican las desigualdades

$$\begin{aligned} -1.0986 \simeq \ln\left(\frac{1}{3}\right) = f_2(-0.5) &\le f_2(x) \\ &\le f_2(-1) = \ln\left(\frac{1}{2}\right) \simeq -0.6931 \end{aligned}$$

y

$$-\frac{4}{3} = f_2'(-0.5) \le f_2'(x) \le f_2'(-1) = -0.5.$$

Como no se cumplen ninguna de las dos hipótesis requeridas, debemos considerar un nuevo intervalo con vistas a aplicar el teorema del Punto Fijo. En ambos casos, los problemas se plantean cerca del punto -0.5, por lo que vamos a acortar el intervalo por la derecha. Tomando el intervalo $[-1, -\ln 2]$ se tiene que

$$f_2'(x) < 0 \ \text{ y } \ f_2''(x) < 0 \ \text{ para } \ x \in [-1, -\ln 2]$$

por lo que las funciones f_2 y f_2' son decrecientes en el intervalo $[-1, -\ln 2]$ y, por tanto, para valores $x \in [-1, -\ln 2]$ se verifica que

$$\begin{aligned} -0.8931 \simeq \ln\left(\frac{\ln 2}{\ln 2 + 1}\right) = f_2(-\ln 2) \leq f_2(x) \\ \leq f_2(-1) = \ln\left(\frac{1}{2}\right) \simeq -0.6931 \end{aligned}$$

y

$$-0.8521 \simeq -\frac{1}{\ln 2(\ln 2 + 1)} = f_2'(-\ln 2) \leq f_2'(x) \leq f_2'(-1) = -0.5.$$

Consecuentemente la función $f_2 \in \mathcal{C}^1([-1, -\ln 2])$ verifica que

$$f_2([-1, -\ln 2]) = \left[\ln\left(\frac{\ln 2}{\ln 2 + 1}\right), -\ln 2\right] \subset [-1, -\ln 2]$$

y

$$|f_2'(x)| \leq \frac{1}{\ln 2(\ln 2 + 1)}, \ x \in [-1, -\ln 2]$$

por lo que f_2 es contractiva en el intervalo $[-1, -\ln 2]$ de constante $k = \dfrac{1}{\ln 2(\ln 2 + 1)} \simeq 0.8521$. Por tanto, por el teorema del Punto Fijo se tiene que existe un único $\xi_1 \in [-1, -\ln 2]$ tal que

$$\xi_1 = f_2(\xi_1) = \ln\left(\frac{\xi_1}{\xi_1 - 1}\right).$$

Además, la sucesión

$$\begin{cases} x_0 \in [-1, -\ln 2] \\ x_n = \ln\left(\dfrac{x_{n-1}}{x_{n-1} - 1}\right), \ n \in \mathbb{N} \end{cases}$$

converge a la raíz negativa de F. Aplicando estas iteraciones sucesivas podemos aproximar dicha raíz como

$$\xi_1 \simeq -0.806466.$$

β) Si consideramos el intervalo $[1.2, 1.5]$ se tiene que

$$\begin{cases} F(1.2) = \dfrac{0.2}{1.2} - e^{-1.2} \simeq -0.1345 < 0 \\ F(1.5) = \dfrac{0.5}{1.5} - e^{-1.5} \simeq 0.1102 > 0 \end{cases}$$

por lo que la raíz ξ_2 de F se encuentra en él. Por otra parte, como

$$f_2''(x) > 0,\ x \in [1.2, 1.5],$$

la función f_2' es estrictamente creciente en dicho intervalo. De esta forma, al ser

$$f_2'(x) < 0,\ x \in [1.2, 1.5],$$

se verifica que

$$|f_2'(x)| \geq |f_2'(1.5)| = \frac{1}{1.5(1.5-1)} = \frac{4}{3} > 1,\ x \in [1.2, 1.5].$$

Aplicando el teorema del Valor Medio como en el apartado β) de i) se llega a que la función f_2 no es contractiva en el intervalo $[1.2, 1.5]$.

Un examen gráfico de las iteraciones (véase la figura 8.29) nos permite concluir que si comienzan en un entorno de ξ_2 la sucesión cae, en algún momento, en el intervalo $(0, 1)$, zona donde la función f_2 no está definida.

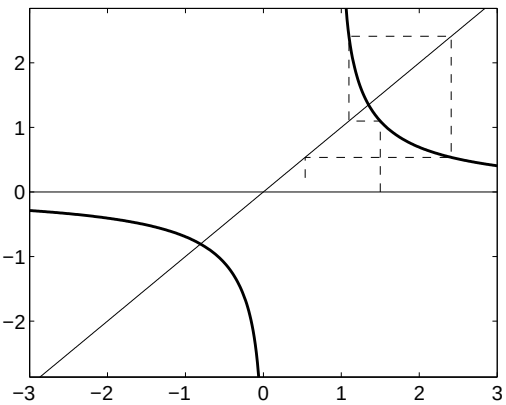


Figura 8.29: Función $f_2(x) = \ln\left(\dfrac{x}{x-1}\right)$.

iii) $f_3(x) = (x-1)e^x \Rightarrow \begin{cases} f_3'(x) = xe^x, & x \in \mathbb{R} \\ f_3''(x) = (1+x)e^x, & x \in \mathbb{R}. \end{cases}$

α) Intervalo $[-1, -0.5]$. Como

$$f_3'(x) < 0 \text{ y } f_3''(x) \geq 0 \text{ para } x \in [-1, -0.5]$$

se verifica que las funciones f_3 y f_3' son, respectivamente, estrictamente decreciente y no decreciente en el intervalo $[-1, -0.5]$. Por tanto, para todo $x \in [-1, -0.5]$ se verifica que

$$-0.9098 \simeq -\frac{3}{2\sqrt{e}} = f_3(-0.5) \leq f_3(x) \leq f_3(-1) = -\frac{2}{e} \simeq -0.7358$$

y

$$-0.3679 \simeq -\frac{1}{e} = f_3'(-1) \leq f_3'(x) \leq f_3'(-0.5) = -\frac{1}{2\sqrt{e}} \simeq -0.3033.$$

Consecuentemente la función $f_3 \in \mathcal{C}^1([-1, -0.5])$ verifica que

$$f_3([-1, -0.5]) = \left[-\frac{3}{2\sqrt{e}}, -\frac{2}{e}\right] \subset [-1, -0.5]$$

y

$$|f_3'(x)| \leq \frac{1}{e},\ x \in [-1, -0.5]$$

por lo que f_3 es contractiva en el intervalo $[-1, -0.5]$ de constante $k = \dfrac{1}{e} \simeq 0.3679$. Por tanto, por el teorema del Punto Fijo, la sucesión

$$\begin{cases} x_0 \in [-1, -0.5] \\ x_n = (x_{n-1} - 1)e^{x_{n-1}},\ n \in \mathbb{N} \end{cases}$$

converge a la raíz negativa de F.

β) Para la raíz positiva, como

$$f_3''(x) > 0,\ x \in [1, 2]$$

se verifica que la función f_3' es creciente en $[1, 2]$ y, por tanto,

$$|f_3'(x)| = f_3'(x) \geq f_3'(1) = e > 1,\ x \in [1, 2].$$

Nuevamente el teorema del Valor Medio nos asegura que la función f_3 no es contractiva en el intervalo $[1, 2]$.

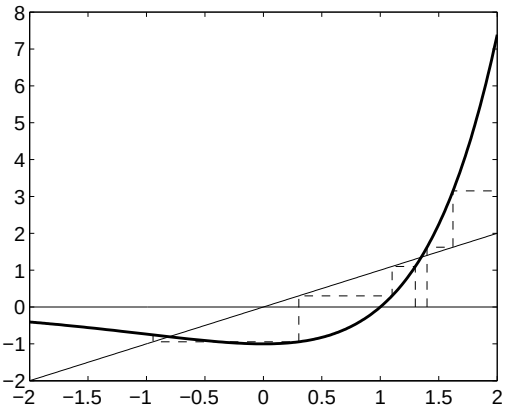


Figura 8.30: Función $f_3(x) = (x-1)e^x$.

Aquí también podemos detallar (véase la figura 8.30) el comportamiento de las iteraciones sucesivas de la función f_3 cuando se comienza en un entorno de ξ_2. Si se toma como dato inicial $x_0 > \xi_2$ la sucesión diverge a $+\infty$; si se elige $x_0 < \xi_2$ la sucesión "salta", de nuevo, a la otra raíz.

c) Vamos a elegir intervalos y datos iniciales adecuados en los que se den las hipótesis de convergencia del método de Newton. En primer lugar, según se ha probado en el apartado *a*),

$$F'(x) > 0 > F''(x),\ x \in [1,2].$$

Por tanto, como $F(1) < 0$, la sucesión

$$\begin{cases} x_0 = 1 \\ x_n = x_{n-1} - \dfrac{F(x_{n-1})}{F'(x_{n-1})},\ n \in \mathbb{N} \end{cases}$$

converge a la raíz positiva ξ_2 de F.

Por otra parte, y usando de nuevo el apartado *a*), se tiene que

$$F'(x) > 0 \ \text{ y } \ F''(x) > 0 \ \text{ para } \ x \in [-0.9, -0.5].$$

Así, al ser $F(-0.5) > 0$, la sucesión

$$\begin{cases} x_0 = -0.5 \\ x_n = x_{n-1} - \dfrac{F(x_{n-1})}{F'(x_{n-1})},\ n \in \mathbb{N} \end{cases}$$

tiene como límite la raíz negativa ξ_1 de F. □

8.11. Dados un número natural n y un número positivo α, una forma de calcular las raíces reales n–ésimas de α sin usar radicales es aplicar el método de Newton a la ecuación

$$x^n - \alpha = 0.$$

Encontrar un intervalo y un valor inicial para los que el método sea convergente. Aplicar la técnica anterior para dar valores aproximados de $\sqrt{2}$, $\sqrt[3]{2}$ y $\sqrt[5]{0.7}$.

SOLUCIÓN. La función $F(x) = x^n - \alpha$ con $n \geq 2$ y $\alpha > 0$ verifica $F \in \mathcal{C}^\infty(\mathbb{R})$ y

$$\begin{cases} F'(x) = nx^{n-1} > 0, & x > 0 \\ F''(x) = n(n-1)x^{n-2} > 0, & x > 0. \end{cases}$$

Elijamos adecuadamente el intervalo de aplicación del método y el dato inicial. Para ello distinguimos dos casos (excluyendo el caso trivial $\alpha = 1$) basándonos en que las potencias de un número menor que uno son menores que el número y las de un número mayor que uno son mayores que él:

a) $\boxed{0 < \alpha < 1}$ Tomamos el intervalo $[\alpha, 1]$. Como

$$F(\alpha) = \alpha^n - \alpha < 0,\ F(1) = 1 - \alpha > 0 \ \text{ y } \ F''(1) > 0$$

la sucesión del método de Newton converge a partir de $x_0 = 1$.

b) $\boxed{\alpha > 1}$ Análogamente, si consideramos el intervalo $[1, \alpha]$ tenemos que

$$F(1) = 1 - \alpha < 0,\ F(\alpha) = \alpha^n - \alpha > 0 \ \text{ y } \ F''(\alpha) > 0$$

y, por tanto, la sucesión del método de Newton comenzando en $x_0 = \alpha$ es convergente.

La sucesión del método de Newton para F viene dada por

$$x_k = x_{k-1} - \frac{x_{k-1}^n - \alpha}{nx_{k-1}^{n-1}} = \frac{(n-1)x_{k-1}^n + \alpha}{nx_{k-1}^{n-1}},\ k \in \mathbb{N}.$$

Para los casos concretos del cálculo de $\sqrt{2}$, $\sqrt[3]{2}$ y $\sqrt[5]{0.7}$ consideramos, respectivamente, las sucesiones $\{x_k\}_{k=0}^\infty$, $\{y_k\}_{k=0}^\infty$ y $\{z_k\}_{k=0}^\infty$ dadas por

$$x_k = \frac{x_{k-1}^2 + 2}{2x_{k-1}},\ y_k = \frac{2y_{k-1}^3 + 2}{3y_{k-1}^2} \ \text{ y } \ z_k = \frac{4z_{k-1}^5 + 0.7}{5z_{k-1}^4} \ \text{ para } \ k \in \mathbb{N}$$

con $x_0 = 2$, $y_0 = 2$ y $z_0 = 1$. Los primeros términos de las sucesiones anteriores vienen dados en la siguiente tabla:

k	x_k	y_k	z_k
0	2	2	1
1	1.50000000000000	1.50000000000000	0.94000000000000
2	1.41666666666667	1.29629629629630	0.93131500030432
3	1.41421568627451	1.26093222474175	0.93114997361058
4	1.41421356237469	1.25992186056593	0.93114991509484
5	1.41421356237310	1.25992104989539	0.93114991509484

8.12. Demostrar que la ecuación

$$e^x \ln x + x^3 - 2 = 0$$

tiene una única raíz positiva. Determinar un intervalo y un valor inicial para los que el método de Newton converja a dicha raíz.

SOLUCIÓN. Consideremos la función

$$F(x) = e^x \ln x + x^3 - 2,\ x > 0.$$

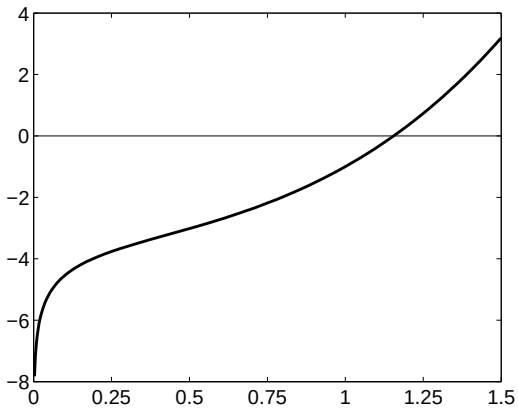


Figura 8.31: Función $F(x) = e^x \ln x + x^3 - 2$.

Claramente la función $F \in \mathcal{C}^2(0, +\infty)$ y

$$\begin{cases} F'(x) = e^x \left(\ln x + \dfrac{1}{x} \right) + 3x^2, & x > 0 \\ F''(x) = e^x \left(\ln x + \dfrac{2}{x} - \dfrac{1}{x^2} \right) + 6x, & x > 0. \end{cases}$$

Como $F(1) < 0 < F(2)$ el teorema de Bolzano asegura la existencia de $\xi \in (1, 2)$ tal que $F(\xi) = 0$. Veamos que ésta es la única raíz positiva demostrando que

$$F'(x) > 0 \ \text{ si } \ x > 0. \tag{8.47}$$

Como $\ln x \geq 0$ si $x \geq 1$ es obvio que

$$F'(x) > 0 \text{ si } x \geq 1.$$

Para $0 < x \leq 1$ consideramos la función auxiliar

$$h(x) = \ln x + \frac{1}{x},\ x \in (0,1].$$

Como la función h es derivable en $(0,1]$ y

$$h'(x) = \frac{1}{x} - \frac{1}{x^2} = \frac{x-1}{x^2} \leq 0,\ x \in (0,1]$$

se tiene que h es una función no creciente en el intervalo $(0,1]$ y, como $h(1) = 1$, la función h es positiva. Por tanto,

$$F'(x) = e^x h(x) + 3x^2 \geq h(x) > 0,\ x \in (0,1].$$

De esta forma se obtiene la propiedad (8.47) y, en particular,

$$F'(x) > 0,\ x \in [1,2].$$

Por otra parte, para todo $x \in [1,2]$, se verifica que

$$F''(x) = e^x \left(\ln x + \frac{2}{x} - \frac{1}{x^2} \right) + 6x \geq \frac{2}{x} - \frac{1}{x^2} = \frac{2x-1}{x^2} > \frac{2x-1}{4} \geq \frac{1}{4} > 0.$$

Así pues, el método de Newton

$$\begin{cases} x_0 = 2 \\ x_n = x_{n-1} - \dfrac{e^{x_{n-1}} \ln x_{n-1} + x_{n-1}^3 - 2}{e^{x_{n-1}} \left(\ln x_{n-1} + \dfrac{1}{x_{n-1}} \right) + 3x_{n-1}^2},\ n \in \mathbb{N} \end{cases}$$

es convergente a la única raíz ξ de F. Los primeros términos de la sucesión anterior vienen dados en la siguiente tabla:

n	x_n
0	2
1	1.46571965113289
2	1.21202860187979
3	1.15750854860195
4	1.15525662059539
5	1.15525291938671
6	1.15525291937673

8.7.2. Problemas propuestos

8.13. Comprobar que se puede aplicar el teorema del Punto Fijo a las siguientes funciones en los intervalos dados:

a) $f(x) = \dfrac{\cos x}{8} + \dfrac{x^2}{4}$ en $\left[0, \dfrac{\pi}{2}\right]$.

b) $g(x) = \dfrac{x - x^2 + 1}{5}$ en $[0, 1]$.

8.14. Determinar un intervalo y una función para poder aplicar el método del Punto Fijo a las siguientes ecuaciones:

a) $x - \ln(1 + x) - 0.2 = 0$.

b) $x = -\ln x$.

Determinar, en cada caso, el número de iteraciones necesario para que el error cometido sea inferior a 10^{-5}.

8.15. El resultado del problema 8.3 puede generalizarse a órdenes de convergencia superiores. Concretamente, si $f : [a, b] \to \mathbb{R}$ verifica:

i) $f \in \mathcal{C}^{m-1}([a, b])$ y existe $f^{m)}$ en (a, b) para algún $m \in \mathbb{N}$.

ii) $f([a, b]) \subset [a, b]$.

iii) $f'(\xi) = f''(\xi) = \cdots = f^{m-1)}(\xi) = 0$.

iv) Existe $M \geq 0$ tal que $|f^{m)}(x)| \leq M$, $x \in (a, b)$,

demostrar que

$$\frac{e_n}{(e_{n-1})^m} \leq M,\ n \in \mathbb{N}$$

por lo que la convergencia es, al menos, de orden m.

8.16. Dado $\alpha > 0$ se considera la sucesión

$$x_n = \sqrt{\alpha + x_{n-1}},\ n \in \mathbb{N}.$$

a) Encontrar un intervalo I en el que la sucesión $\{x_n\}_{n=0}^{\infty}$ sea convergente para cualquier dato inicial $x_0 \in I$ y calcular $\lim\limits_{n \to +\infty} x_n$.

b) Hallar el valor de

$$\sqrt{6 + \sqrt{6 + \sqrt{6 + \cdots}}}$$

8.17. Demostrar que la ecuación

$$\cos x - 3x = \frac{\pi}{2}.$$

tiene una única raíz real. Encontrar un intervalo, un valor inicial y un valor de λ para los que el método de Whittaker converja a dicha raíz. Determinar el número de iteraciones necesario para que el error cometido sea inferior a 10^{-3}.

8.18. Aplicar el teorema del Punto Fijo para aproximar una raíz ξ de la ecuación

$$x^2 + \frac{\operatorname{sen} x}{2} - 1 = 0$$

en el intervalo $\left[0, \frac{\pi}{2}\right]$, demostrando las hipótesis de convergencia del método, y determinar una sucesión $\{x_n\}_{n=0}^{\infty}$ que converja a ξ. ¿Qué términos de la sucesión anterior distan de ξ una cantidad inferior a 10^{-3}?

8.19. Aplicar el método del Punto Fijo para aproximar la menor raíz positiva de la ecuación

$$\cos x + 3x^2 - 6x = 0$$

determinando una sucesión que converja a dicha raíz y justificando las hipótesis de convergencia.

8.20. Se considera la ecuación

$$(x+1)\tan x - 1 = 0.$$

a) Probar que la ecuación anterior tiene infinitas raíces positivas y determinar intervalos que contengan a cada una de ellas.

b) Si ξ es la menor raíz positiva, demostrar que se puede aproximar mediante el método del Punto Fijo.

c) Comenzando en $x_0 = 0$ construir una sucesión $\{x_n\}_{n=0}^{\infty}$ que converja a ξ. Determinar un valor de $n \in \mathbb{N}$ a partir del cual todos los términos de la sucesión distan de ξ una cantidad inferior a 10^{-4}.

8.21. Convergencia local del método de Newton. Sea $F \in \mathcal{C}^2([a,b])$ y $\xi \in (a,b)$ tal que

$$F(\xi) = 0 \ \text{ y } \ F'(\xi) \neq 0.$$

Demostrar que existe $\delta > 0$ de forma que la aplicación

$$f(x) = x - \frac{F(x)}{F'(x)},\ x \in [a,b]$$

verifica las hipótesis del teorema del Punto Fijo en el intervalo $[\xi - \delta, \xi + \delta]$.

8.22. Se considera la función

$$F(x) = e^{-x} - \operatorname{sen} x - 2.$$

a) Demostrar que

$$F(x) < 0,\ x \geq 0.$$

b) Probar que la ecuación $F(x) = 0$ tiene una única raíz negativa.

c) Determinar un intervalo donde se pueda aplicar el método de Newton para aproximar dicha raíz, así como los primeros términos de la sucesión $\{x_n\}_{n=0}^{\infty}$ definida mediante dicho método.

8.23. Consideremos la ecuación

$$x^3 + 2x^2 + 10x - 20 = 0.$$

a) Demostrar que la ecuación anterior tiene una única raíz positiva ξ.

b) Determinar un intervalo donde se pueda aplicar el método de Newton; aproximar ξ de forma que el error cometido sea inferior a 10^{-6}.

c) Comparar el resultado anterior con la siguiente aproximación que *Leonardo de Pisa*, más conocido como *Fibonacci*, dio en el año 1224:

$$\xi \simeq 1 + \frac{22}{60} + \frac{7}{60^2} + \frac{42}{60^3} + \frac{33}{60^4} + \frac{4}{60^5} + \frac{40}{60^6}.$$

8.8. Prácticas

8.1. Escribir un programa que implemente el método de la bisección. Aplicarlo a la resolución del problema 8.1 con una precisión de 10^{-4}.

8.2. Programar el método del Punto Fijo. Resolver las ecuaciones del problema 8.14 con un error inferior a 10^{-5}.

8.3. Programar el método de Newton como una función cuyas variables sean la función F que define la ecuación y el valor inicial. Utilizar el comando `diff` de `MATLAB` para calcular F'.

8.4. Aproximar, mediante el método de Newton, $\sqrt{2}$ y $\sqrt[3]{3}$ con una tolerancia en el test de parada de 10^{-6}, resolviendo la ecuación correspondiente del problema 8.11.

8.5. Hallar la solución del problema 8.17 con un error inferior a la milésima, mediante un programa que implemente el método de Whittaker.

8.6. Programar el método de la secante como una función cuyas variables sean la función F que define la ecuación y los dos valores iniciales.

8.7. Orden de convergencia del método de la secante. Se considera la ecuación

$$x - \cos x = 0.$$

a) Demostrar que la ecuación anterior tiene una única raíz real ξ.

b) Determinar las siete primeras iteraciones que se obtienen al aplicar el método de la secante comenzando en $x_0 = 0.8$ y $x_1 = 0.7$.

c) Sabiendo que

$$\xi = 0.7390851332151606\ldots$$

construir una tabla, para $n = 1, 2, \ldots, 7$, con los cocientes

$$\frac{|x_n - \xi|}{|x_{n-1} - \xi|^p}$$

siendo $p = \dfrac{1+\sqrt{5}}{2}$ la razón áurea. Comprobar que dichos valores permanecen, prácticamente, constantes a partir de uno dado.

8.8. Programar el método de las cuerdas. Aplicarlo a la resolución de la ecuación de la práctica 8.7.

8.9. Ídem para el método de la Falsa Posición (*Regula Falsi*).

9 Resolución de sistemas no lineales

9.1. Introducción

En este breve capítulo vamos a estudiar métodos que sirven para aproximar las raíces de un sistema de ecuaciones no lineales. Este tipo de problemas presenta una gran complejidad, tanto desde el punto de vista teórico, como para el cálculo efectivo de las raíces. Por esta razón, vamos a realizar una presentación meramente descriptiva, sin analizar resultados de convergencia (el lector interesado puede acudir al completo libro [**Or–Rh**]).

De entre las distintas familias de métodos que sirven para abordar estos problemas, hemos elegido dos que resultan conceptualmente sencillos, una vez familiarizados con los métodos de resolución de sistemas lineales y los del cálculo de raíces de ecuaciones no lineales. En la primera familia, agrupada en el epígrafe *método de Newton*, la idea básica es la misma que en el caso de una sola ecuación: linealizar el problema, sustituyendo la función por su tangente (en este caso, hiperplanos tangentes) y resolver una sucesión de sistemas lineales. Lo que agrupamos bajo el nombre de *generalización de métodos lineales* es una familia de métodos que se basan en la idea de calcular, como en los métodos iterativos de sistemas lineales, cada coordenada de la incógnita usando, en las restantes coordenadas, valores obtenidos previamente. Esto lleva, para cada coordenada, a una ecuación no lineal con una sola incógnita que podrá resolverse mediante alguno de los métodos considerados en el capítulo 8.

9.2. Método de Newton

Recordemos que en el método de Newton estudiado en el capítulo anterior, si ξ es raíz de la ecuación $F(x) = 0$, entonces

$$0 = F(\xi) = F(x + h) \simeq F(x) + hF'(x),$$

es decir, aproximamos la función en un entorno de x por la recta tangente a $y = F(x)$ en dicho punto; la intersección de esta recta con el eje de abscisas se obtiene cuando

$$h = -\frac{F(x)}{F'(x)}.$$

A partir de este valor se define el método de Newton como

$$\begin{cases} x^0 \in \mathbb{R} \text{ dado} \\ x^k = x^{k-1} + h^{k-1} = x^{k-1} - \dfrac{F(x^{k-1})}{F'(x^{k-1})},\ k \in \mathbb{N}. \end{cases}$$

En el caso de un sistema no lineal de dos ecuaciones

$$\begin{cases} F_1(x_1, x_2) = 0 \\ F_2(x_1, x_2) = 0 \end{cases} \tag{9.1}$$

al emplear el método de Newton para la resolución de (9.1) argumentaremos como en el caso de una sola ecuación, linealizando el sistema (9.1). En concreto, si

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} x_1 + h_1 \\ x_2 + h_2 \end{pmatrix} \in \mathbb{R}^2$$

es solución de (9.1), un desarrollo de Taylor de primer orden determina

$$\begin{cases} 0 = F_1(x_1 + h_1, x_2 + h_2) \simeq F_1(x_1, x_2) + h_1 \dfrac{\partial F_1}{\partial x_1}(x_1, x_2) + h_2 \dfrac{\partial F_1}{\partial x_2}(x_1, x_2) \\ 0 = F_2(x_1 + h_1, x_2 + h_2) \simeq F_2(x_1, x_2) + h_1 \dfrac{\partial F_2}{\partial x_1}(x_1, x_2) + h_2 \dfrac{\partial F_2}{\partial x_2}(x_1, x_2). \end{cases}$$

Esto lleva a considerar el sistema lineal

$$J_{(F_1,F_2)}(x_1, x_2) \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = - \begin{pmatrix} F_1(x_1, x_2) \\ F_2(x_1, x_2) \end{pmatrix} \tag{9.2}$$

donde

$$J_{(F_1,F_2)}(x_1, x_2) = \begin{pmatrix} \dfrac{\partial F_1}{\partial x_1}(x_1, x_2) & \dfrac{\partial F_1}{\partial x_2}(x_1, x_2) \\ \dfrac{\partial F_2}{\partial x_1}(x_1, x_2) & \dfrac{\partial F_2}{\partial x_2}(x_1, x_2) \end{pmatrix}$$

es la *matriz jacobiana* de las funciones F_1 y F_2. Para poder resolver el sistema (9.2) es necesario que la matriz $J_{(F_1,F_2)}$ sea no singular, en cuyo caso la solución viene dada por

$$\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = - \left(J_{(F_1,F_2)}(x_1, x_2) \right)^{-1} \begin{pmatrix} F_1(x_1, x_2) \\ F_2(x_1, x_2) \end{pmatrix}.$$

De esta forma, se puede definir el *método de Newton* para el sistema (9.1) como

$$\begin{cases} \begin{pmatrix} x_1^0 \\ x_2^0 \end{pmatrix} \in \mathbb{R}^2 \text{ dado} \\ \begin{pmatrix} x_1^k \\ x_2^k \end{pmatrix} = \begin{pmatrix} x_1^{k-1} \\ x_2^{k-1} \end{pmatrix} - \left(J_{(F_1,F_2)}(x_1^{k-1}, x_2^{k-1})\right)^{-1} \begin{pmatrix} F_1(x_1^{k-1}, x_2^{k-1}) \\ F_2(x_1^{k-1}, x_2^{k-1}) \end{pmatrix}, \ k \in \mathbb{N} \end{cases}$$

o, equivalentemente,

$$\begin{cases} \begin{pmatrix} x_1^0 \\ x_2^0 \end{pmatrix} \in \mathbb{R}^2 \text{ dado} \\ J_{(F_1,F_2)}(x_1^{k-1}, x_2^{k-1}) \begin{pmatrix} x_1^k \\ x_2^k \end{pmatrix} = J_{(F_1,F_2)}(x_1^{k-1}, x_2^{k-1}) \begin{pmatrix} x_1^{k-1} \\ x_2^{k-1} \end{pmatrix} - \begin{pmatrix} F_1(x_1^{k-1}, x_2^{k-1}) \\ F_2(x_1^{k-1}, x_2^{k-1}) \end{pmatrix}. \end{cases}$$

Ejemplo 9.1. Vamos a aplicar el método de Newton para resolver el sistema no lineal

$$(\mathcal{S}) \begin{cases} x_1^2 - 10x_1 + x_2^2 + 8 &= 0 \\ x_1 x_2^2 + x_1 - 10x_2 + 8 &= 0. \end{cases}$$

En este caso

$$\begin{cases} F_1(x_1, x_2) = x_1^2 - 10x_1 + x_2^2 + 8 \\ F_2(x_1, x_2) = x_1 x_2^2 + x_1 - 10x_2 + 8 \end{cases}$$

por lo que la matriz jacobiana de F_1 y F_2 es

$$J_{(F_1,F_2)}(x_1, x_2) = \begin{pmatrix} \dfrac{\partial F_1}{\partial x_1}(x_1, x_2) & \dfrac{\partial F_1}{\partial x_2}(x_1, x_2) \\ \dfrac{\partial F_2}{\partial x_1}(x_1, x_2) & \dfrac{\partial F_2}{\partial x_2}(x_1, x_2) \end{pmatrix} = \begin{pmatrix} 2x_1 - 10 & 2x_2 \\ x_2^2 + 1 & 2x_1 x_2 - 10 \end{pmatrix}.$$

Siempre que el determinante

$$\det J_{(F_1,F_2)}(x_1, x_2) = 2\big(2x_1^2 x_2 - 10x_1(1 + x_2) - x_2(1 + x_2^2) + 50\big)$$

sea no nulo, el sistema

$$J_{(F_1,F_2)}(x_1^{k-1}, x_2^{k-1}) \begin{pmatrix} x_1^k \\ x_2^k \end{pmatrix} = J_{(F_1,F_2)}(x_1^{k-1}, x_2^{k-1}) \begin{pmatrix} x_1^{k-1} \\ x_2^{k-1} \end{pmatrix} - \begin{pmatrix} F_1(x_1^{k-1}, x_2^{k-1}) \\ F_2(x_1^{k-1}, x_2^{k-1}) \end{pmatrix}$$

tendrá solución para cada $k \in \mathbb{N}$. De esta forma, partiendo del vector

$$\begin{pmatrix} x_1^0 \\ x_2^0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \mathbb{R}^2,$$

el primer sistema que debemos resolver es

$$\begin{pmatrix} -10 & 0 \\ 1 & -10 \end{pmatrix} \begin{pmatrix} x_1^1 \\ x_2^1 \end{pmatrix} = - \begin{pmatrix} 8 \\ 8 \end{pmatrix}$$

cuya solución es

$$\begin{pmatrix} x_1^1 \\ x_2^1 \end{pmatrix} = \begin{pmatrix} 0.8 \\ 0.88 \end{pmatrix};$$

el sistema correspondiente a la segunda iteración se escribe como

$$\begin{pmatrix} -8.4000 & 1.7600 \\ 1.7744 & -8.5920 \end{pmatrix} \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix} = \begin{pmatrix} -8.4000 & 1.7600 \\ 1.7744 & -8.5920 \end{pmatrix} \begin{pmatrix} 0.8 \\ 0.88 \end{pmatrix} - \begin{pmatrix} 1.4144 \\ 0.6195 \end{pmatrix}$$
$$= \begin{pmatrix} 9.4144 \\ -6.7610 \end{pmatrix}$$

y tiene por solución

$$\begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix} = \begin{pmatrix} 0.99178824065904 \\ 0.99171660314541 \end{pmatrix}.$$

De forma análoga, se obtienen los vectores

$$\begin{pmatrix} x_1^3 \\ x_2^3 \end{pmatrix} = \begin{pmatrix} 0.99997548056617 \\ 0.99996889851052 \end{pmatrix} \quad \text{y} \quad \begin{pmatrix} x_1^k \\ x_2^k \end{pmatrix} \simeq \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \ k \geq 4.$$

Nótese que $\xi = (1,1)^{\mathrm{T}}$ es una solución exacta de $(\mathcal{S})$. □

Siguiendo la idea anterior podemos considerar sistemas generales de ecuaciones de la forma

$$\begin{cases} F_1(x_1, x_2, \ldots, x_n) = 0 \\ F_2(x_1, x_2, \ldots, x_n) = 0 \\ \cdots \\ F_n(x_1, x_2, \ldots, x_n) = 0. \end{cases}$$

Usando la notación vectorial

$$F = (F_1, F_2, \ldots, F_n)^{\mathrm{T}} \ \text{ y } \ x = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$$

el sistema anterior puede escribirse en la forma

$$F(x) = 0. \tag{9.3}$$

El método de Newton aplicado a (9.3) se corresponde entonces con la expresión

$$\boxed{\begin{cases} x^0 \in \mathbb{R}^n \text{ dado} \\ J_F(x^{k-1})x^k = J_F(x^{k-1})x^{k-1} - F(x^{k-1}),\ k \in \mathbb{N} \end{cases}}$$

donde

$$J_F(x) = \begin{pmatrix} \dfrac{\partial F_1}{\partial x_1}(x_1, x_2, \ldots, x_n) & \cdots & \dfrac{\partial F_1}{\partial x_n}(x_1, x_2, \ldots, x_n) \\ \dfrac{\partial F_2}{\partial x_1}(x_1, x_2, \ldots, x_n) & \cdots & \dfrac{\partial F_2}{\partial x_n}(x_1, x_2, \ldots, x_n) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \dfrac{\partial F_n}{\partial x_1}(x_1, x_2, \ldots, x_n) & \cdots & \dfrac{\partial F_n}{\partial x_n}(x_1, x_2, \ldots, x_n) \end{pmatrix}$$

es la matriz jacobiana de la función F. En cada iteración del método de Newton debe resolverse el sistema lineal

$$A_{k-1}u = b_{k-1}$$

donde

$$A_{k-1} = J_F(x^{k-1}) \text{ y } b_{k-1} = J_F(x^{k-1})x^{k-1} - F(x^{k-1}).$$

Nótese que, en cada iteración, la matriz del sistema que hay que resolver es distinta, por lo que los métodos directos estudiados en el capítulo 4 sólo son de aplicación si el número de ecuaciones no es demasiado elevado. Para sistemas grandes son preferibles los métodos iterativos del capítulo 5; si se usa uno de estos métodos se introduce una nueva iteración (que denominaremos *secundaria*) la cual, tomando como valor inicial de la iteración el de x^{k-1}, aproximará a x^k. En general, basta considerar un número reducido de pasos en la iteración secundaria para conseguir la convergencia del método. Dependiendo del método iterativo elegido y del número de pasos que se den en la iteración secundaria se obtienen los diversos métodos de resolución. Empleando la notación

$$A_{k-1} = D_{k-1} - E_{k-1} - F_{k-1}$$

para la descomposición $D - E - F$ por puntos de A_{k-1} se tiene:

9.2.1. Método de Newton–Jacobi de m pasos

Una vez elegido un vector inicial $x^0 \in \mathbb{R}^n$, el vector x^k se obtiene de x^{k-1} de la siguiente manera: se toma como valor x^k el vector u^m obtenido mediante la aplicación de m pasos del método de Jacobi a partir de x^{k-1}.

De esta forma, una iteración del método se describe como

$$\boxed{\begin{cases} u^0 = x^{k-1} \\ D_{k-1}u^p = (E_{k-1} + F_{k-1})u^{p-1} + b_{k-1},\ 1 \leq p \leq m \\ x^k = u^m \end{cases}}$$

En particular, cuando $m = 1$, se tiene que

$$\begin{aligned} x^k = u^1 &= (D_{k-1})^{-1}(E_{k-1} + F_{k-1})u^0 + (D_{k-1})^{-1}b_{k-1} \\ &= (D_{k-1})^{-1}(D_{k-1} - A_{k-1})x^{k-1} + (D_{k-1})^{-1}(J_F(x^{k-1})x^{k-1} - F(x^{k-1})) \\ &= \left(I - (D_{k-1})^{-1}J_F(x^{k-1})\right)x^{k-1} + (D_{k-1})^{-1}(J_F(x^{k-1})x^{k-1} - F(x^{k-1})), \end{aligned}$$

es decir,

$$x^k = x^{k-1} - (D_{k-1})^{-1}F(x^{k-1}),\ k \in \mathbb{N}.$$

Expresando la relación anterior coordenada a coordenada obtenemos el método de Newton–Jacobi de un paso:

$$x_i^k = x_i^{k-1} - \frac{F_i(x^{k-1})}{\dfrac{\partial F_i}{\partial x_i}(x^{k-1})} \tag{9.4}$$

para $i = 1, 2, \ldots, n$.

9.2.2. Método de Newton–relajación de m pasos

Comenzando nuevamente con un vector inicial $x^0 \in \mathbb{R}^n$, se toma ahora como valor x^k el vector u^m obtenido tras m iteraciones del método de relajación aplicadas a partir de x^{k-1}. Es decir,

$$\boxed{\begin{cases} u^0 = x^{k-1} \\ \left(\dfrac{D_{k-1}}{w} - E_{k-1}\right)u^p = \left(\dfrac{1-w}{w}D_{k-1} + F_{k-1}\right)u^{p-1} + b_{k-1},\ 1 \leq p \leq m \\ x^k = u^m \end{cases}}$$

Para el caso particular $m = 1$ se obtiene

$$\begin{aligned}\left(\frac{D_{k-1}}{w} - E_{k-1}\right) x^k &= \left(\frac{D_{k-1}}{w} - E_{k-1}\right) u^1 = \left(\frac{1-w}{w} D_{k-1} + F_{k-1}\right) u^0 + b_{k-1} \\ &= \left(\frac{D_{k-1}}{w} - D_{k-1} + D_{k-1} - E_{k-1} - A_{k-1}\right) x^{k-1} + b_{k-1} \\ &= \left(\frac{D_{k-1}}{w} - E_{k-1} - J_F(x^{k-1})\right) x^{k-1} \\ &\quad + J_F(x^{k-1}) x^{k-1} - F(x^{k-1}) \\ &= \left(\frac{D_{k-1}}{w} - E_{k-1}\right) x^{k-1} - F(x^{k-1})\end{aligned}$$

para cada $k \in \mathbb{N}$. Considerando coordenada a coordenada en la expresión anterior se tiene que

$$\begin{aligned}\sum_{j=1}^{i-1} \frac{\partial F_i}{\partial x_j}(x^{k-1}) x_j^k + \frac{1}{w} \frac{\partial F_i}{\partial x_i}(x^{k-1}) x_i^k &= \sum_{j=1}^{i-1} \frac{\partial F_i}{\partial x_j}(x^{k-1}) x_j^{k-1} \\ &\quad + \frac{1}{w} \frac{\partial F_i}{\partial x_i}(x^{k-1}) x_i^{k-1} - F_i(x^{k-1})\end{aligned}$$

para $i = 1, 2, \ldots, n$, y con ello, el método de Newton–relajación de un paso:

$$x_i^k = x_i^{k-1} - \frac{w}{\dfrac{\partial F_i}{\partial x_i}(x^{k-1})} \left(F_i(x^{k-1}) - \sum_{j=1}^{i-1} \frac{\partial F_i}{\partial x_j}(x^{k-1}) \left(x_j^{k-1} - x_j^k\right) \right)$$

para $i = 1, 2, \ldots, n$.

Ejemplo 9.2. Vamos a aplicar los métodos de Newton–Jacobi y Newton–Gauss–Seidel de un paso al sistema no lineal

$$(\mathcal{S}) \begin{cases} 3x_1 + x_1 x_2 x_3 &= 4 \\ x_1 + 4x_2 + x_3 &= -4 \\ (x_1 + x_2)x_2 + (1 + 2x_3)x_3 &= 1. \end{cases}$$

En este caso

$$\begin{cases} F_1(x_1, x_2, x_3) = 3x_1 + x_1 x_2 x_3 - 4 \\ F_2(x_1, x_2, x_3) = x_1 + 4x_2 + x_3 + 4 \\ F_3(x_1, x_2, x_3) = x_1 x_2 + x_2^2 + x_3 + 2x_3^2 - 1, \end{cases}$$

por lo que

$$J_F(x_1,x_2,x_3)=\begin{pmatrix}3+x_2x_3 & x_1x_3 & x_1x_2\\ 1 & 4 & 1\\ x_2 & x_1+2x_2 & 1+4x_3\end{pmatrix}.$$

a) Método de Newton–Jacobi de un paso. La expresión de cada iteración del método viene dada por

$$\begin{cases} x_1^k=\dfrac{4}{3+x_2^{k-1}x_3^{k-1}}\\[2ex] x_2^k=-\dfrac{x_1^{k-1}+x_3^{k-1}+4}{4}\\[2ex] x_3^k=\dfrac{2(x_3^{k-1})^2-x_1^{k-1}x_2^{k-1}-(x_2^{k-1})^2+1}{1+4x_3^{k-1}}, \end{cases}$$

obteniéndose, a partir de $x^0=(0,0,0)^{\mathrm{T}}$, los resultados de la siguiente tabla

k	x_1^k	x_2^k	x_3^k
1	1.3333	−1.0000	1.0000
10	2.6772	−1.8211	0.8198
20	1.9875	−1.7406	0.8002
30	−0.0739	−3.1089	2.8178
40	1.0015	−1.0004	−1.0009
41	0.9997	−1.0002	−1.0004
42	0.9999	−0.9998	−0.9998
43	1.0001	−1.0000	−1.0000
44	1.0000	−1.0000	−1.0000

b) Método de Newton–Gauss–Seidel de un paso. En este caso, cada iteración del método toma la forma

$$\begin{cases} x_1^k=\dfrac{4}{3+x_2^{k-1}x_3^{k-1}}\\[2ex] x_2^k=-\dfrac{x_3^{k-1}+4+x_1^k}{4}\\[2ex] x_3^k=\dfrac{x_1^{k-1}x_2^{k-1}+(x_2^{k-1})^2+2(x_3^{k-1})^2+1-(x_1^{k-1}+2x_2^{k-1})x_2^k-x_2^{k-1}x_1^k}{1+4x_3^{k-1}}. \end{cases}$$

Comenzando nuevamente en $x^0=(0,0,0)^{\mathrm{T}}$ las iteraciones sucesivas son:

k	x_1^k	x_2^k	x_3^k
1	1.3333	−1.3333	1.0000
20	2.4767	−1.8135	0.8113
40	2.7718	−1.9036	0.8971
60	2.1764	−1.7145	0.7250
80	1.0033	−1.0044	−0.9986
81	0.9992	−1.0001	−1.0011
82	0.9997	−0.9996	−1.0002
83	1.0000	−1.0000	−0.9999
84	1.0000	−1.0000	−1.0000

Las gráficas de la figura 9.1 representan el error cometido con cada uno de estos dos métodos en las distintas iteraciones, sabiendo que una solución de $(\mathcal{S})$ es $(1, -1, -1)^{\mathrm{T}}$. □

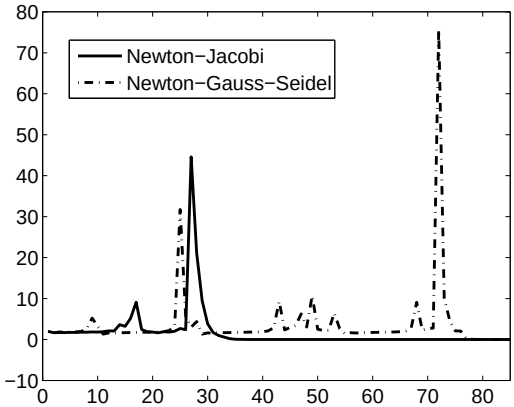


Figura 9.1: Comparación entre los dos métodos.

9.3. Generalización de métodos lineales

A continuación vamos a estudiar una serie de métodos que, basados en las ideas desarrolladas para los métodos iterativos de resolución de sistemas lineales, nos permitirán resolver sistemas de ecuaciones no lineales de la forma

$$\begin{cases} F_1(x_1, x_2, \ldots, x_n) = 0 \\ F_2(x_1, x_2, \ldots, x_n) = 0 \\ \qquad \cdots \\ F_n(x_1, x_2, \ldots, x_n) = 0 \end{cases} \tag{9.5}$$

empleando como herramienta auxiliar los métodos estudiados en el capítulo 8 de resolución de ecuaciones no lineales (especialmente, el método de Newton).

Para resolver el sistema (9.5) lo que haremos será "congelar" $n-1$ variables en cada ecuación y resolver la ecuación resultante (que se trata de una ecuación con una única incógnita) mediante alguno de los métodos iterativos de resolución de ecuaciones no lineales. Al igual que ocurría en la sección anterior, también aquí habrá una iteración secundaria (la de la resolución de la ecuación no lineal) y bastará considerar un número reducido de pasos en esta iteración para obtener convergencia; dependiendo de la estrategia seguida en la elección de las $n-1$ variables fijadas y del número de pasos de la iteración se obtendrán los diversos métodos.

9.3.1. Método de Jacobi no lineal

Análogamente a como se hizo en la resolución de sistemas lineales (véase (5.8)), la idea del *método de Jacobi no lineal* consiste en resolver cada ecuación del sistema (9.5) respecto a la variable que se encuentra en la diagonal, dando a las $n-1$ variables restantes el valor que tenían en la iteración anterior. De esta forma, a partir del vector x^{k-1}, la coordenada i–ésima de x^k se obtiene como solución de la ecuación (con una única incógnita) en u

$$F_i(x_1^{k-1}, x_2^{k-1}, \ldots, x_{i-1}^{k-1}, u, x_{i+1}^{k-1}, \ldots, x_n^{k-1}) = 0. \tag{9.6}$$

Obviamente, al ser ésta una ecuación no lineal no podremos, en general, obtener el valor exacto de u y tendremos que aplicar para su resolución alguno de los métodos del capítulo 8. Lo más habitual es elegir para esta iteración secundaria el método de Newton con un número pequeño m de iteraciones; esto da lugar al *método de Jacobi–Newton* de m pasos: partiendo de un vector inicial $x^0 \in \mathbb{R}^n$ el vector x^k se obtiene de x^{k-1} de la siguiente forma: se toma como valor de la componente i–ésima de x^k el valor u^m correspondiente a la m–ésima iteración del método de Newton aplicado a la ecuación (9.6) partiendo de x_i^{k-1}. Es decir, para cada $i \in \{1, 2, \ldots, n\}$

$$\left\{\begin{array}{l} u^0 = x_i^{k-1} \\ u^p = u^{p-1} - \dfrac{F_i(x_1^{k-1}, x_2^{k-1}, \ldots, x_{i-1}^{k-1}, u^{p-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}{\dfrac{\partial F_i}{\partial x_i}(x_1^{k-1}, x_2^{k-1}, \ldots, x_{i-1}^{k-1}, u^{p-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}, \ 1 \le p \le m \\ x_i^k = u^m \end{array}\right.$$

En el caso particular de $m = 1$ se tiene el método de Jacobi–Newton de un paso:

$$x_i^k = u^1 = x_i^{k-1} - \frac{F_i(x^{k-1})}{\dfrac{\partial F_i}{\partial x_i}(x^{k-1})}$$

para $i = 1, 2, \ldots, n$, expresión que coincide con la del *método de Newton–Jacobi* de un paso (véase (9.4)).

9.3.2. Método de Gauss–Seidel no lineal

Recordando cómo se dedujo el método de Gauss–Seidel para sistemas lineales a partir del método de Jacobi, podemos usar la misma idea para deducir el *método de Gauss–Seidel no lineal* a partir del método de Jacobi no lineal. La estrategia seguida era la de utilizar las coordenadas del nuevo vector desde el momento en que se van obteniendo; esto lleva, en el paso k–ésimo, a definir la coordenada i–ésima de x^k como la solución de la ecuación

$$F_i(x_1^k, x_2^k, \ldots, x_{i-1}^k, u, x_{i+1}^{k-1}, \ldots, x_n^{k-1}) = 0$$

(nótese que las coordenadas de x^k van obteniéndose sucesivamente y en el orden "natural"). Nuevamente, cada una de estas ecuaciones debe resolverse mediante uno de los métodos iterativos del capítulo 8. Si el método elegido para esta iteración secundaria es el de Newton con un número m de iteraciones se obtiene el *método de Gauss–Seidel–Newton* de m pasos: a partir de $x_0 \in \mathbb{R}^n$, para cada $i \in \{1, 2, \ldots, n\}$ se toma como valor de la componente i–ésima de x^k el valor obtenido mediante

$$\begin{cases} u^0 = x_i^{k-1} \\ u^p = u^{p-1} - \dfrac{F_i(x_1^k, x_2^k, \ldots, x_{i-1}^k, u^{p-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}{\dfrac{\partial F_i}{\partial x_i}(x_1^k, x_2^k, \ldots, x_{i-1}^k, u^{p-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}, \ 1 \le p \le m \\ x_i^k = u^m \end{cases}$$

Eligiendo $m = 1$ se obtiene el método de Gauss–Seidel–Newton de un paso:

$$x_i^k = u^1 = x_i^{k-1} - \frac{F_i(x_1^k, x_2^k, \ldots, x_{i-1}^k, x_i^{k-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}{\dfrac{\partial F_i}{\partial x_i}(x_1^k, x_2^k, \ldots, x_{i-1}^k, x_i^{k-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}$$

para $i = 1, 2, \ldots, n$.

9.3.3. Método de relajación no lineal

También la idea que dio lugar al método de relajación para sistemas lineales se puede generalizar al caso no lineal. Tomaremos como valor de la coordenada i–ésima de x^k el resultado de una media ponderada entre x_i^{k-1} y la coordenada i–ésima que se obtendría aplicando el método de Gauss–Seidel no lineal. Más explícitamente, se define

$$x_i^k = (1-w)x_i^{k-1} + wu \tag{9.7}$$

donde u es la solución de la ecuación

$$F_i(x_1^k, x_2^k, \ldots, x_{i-1}^k, u, x_{i+1}^{k-1}, \ldots, x_n^{k-1}) = 0. \tag{9.8}$$

Obsérvese que lo que se propone es resolver las ecuaciones (9.8) sucesivamente y, antes de resolver la ecuación (9.8) para $i+1$, calcular (9.7) para i. Una vez más, el método iterativo más habitual para resolver (9.8) es el de Newton, dando lugar al *método de relajación–Newton* de m pasos: se toma como valor de la componente i–ésima de x^k la cantidad

$$x_i^k = (1-w)x_i^{k-1} + wu^m$$

donde u^m se obtiene a partir de

$$\left\{\begin{array}{l} u^0 = x_i^{k-1} \\ \\ u^p = u^{p-1} - \dfrac{F_i(x_1^k, x_2^k, \ldots, x_{i-1}^k, u^{p-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}{\dfrac{\partial F_i}{\partial x_i}(x_1^k, x_2^k, \ldots, x_{i-1}^k, u^{p-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}, \ 1 \le p \le m \end{array}\right.$$

En particular, para $m = 1$

$$\begin{aligned} x_i^k &= (1-w)x_i^{k-1} + wu^1 \\ &= (1-w)x_i^{k-1} + wx_i^{k-1} - w\frac{F_i(x_1^k, x_2^k, \ldots, x_{i-1}^k, x_i^{k-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}{\dfrac{\partial F_i}{\partial x_i}(x_1^k, x_2^k, \ldots, x_{i-1}^k, x_i^{k-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}, \end{aligned}$$

es decir,

$$x_i^k = x_i^{k-1} - w\frac{F_i(x_1^k, x_2^k, \ldots, x_{i-1}^k, x_i^{k-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}{\dfrac{\partial F_i}{\partial x_i}(x_1^k, x_2^k, \ldots, x_{i-1}^k, x_i^{k-1}, x_{i+1}^{k-1}, \ldots, x_n^{k-1})}$$

para $i = 1, 2, \ldots, n$, fórmula que proporciona las iteraciones del método de relajación–Newton de un paso.

Observación 9.1. Adelantábamos ya en la introducción que tan sólo íbamos a realizar una descripción de los distintos métodos. En general, tanto el estudio teórico de las ecuaciones no lineales (existencia, unicidad y localización de solución) como el de los métodos de resolución (esencialmente, su convergencia y la acotación del error) necesitan técnicas que exceden los objetivos de este libro.

Por otra parte, puede imaginarse la gran cantidad de métodos que se pueden diseñar para resolver este tipo de problemas. Sólo considerando la estrategia aquí seguida de combinar métodos de resolución de sistemas lineales con métodos para ecuaciones no lineales, se obtendría un amplísimo catálogo. La adecuación de un método dependerá fuertemente del problema que se necesite resolver, aunque los aquí presentados son los de uso más frecuente. □

9.4. Prácticas

9.1. Implementar con `MATLAB` el método de Newton–Jacobi de un paso. Aproximar con este método la solución del sistema

$$\begin{cases} x(x^2+\cos\pi x+6)-y+z(\operatorname{sen}\pi y-2) &= 1.1250 \\ -x+y(7e^y+y^4)+z(\cos\pi x-4) &= -4.5 \\ x(\operatorname{sen}\pi x-3)-y+z(z^2+\operatorname{sen}\pi z+8) &= 8 \end{cases}$$

comenzando en $x_0=(0,0,0)^{\mathrm{T}}$.

9.2. Programar en `MATLAB` el método de Newton–Gauss–Seidel de un paso. Utilizar este programa para aproximar la solución del sistema del ejemplo 9.2 partiendo de $x_0=(0,0,0)^{\mathrm{T}}$.

9.3. Escribir un programa en `MATLAB` que implemente el método de Gauss–Seidel–Newton de un paso. Aproximar con él la solución del ejemplo 9.2 cuando se toma como valor inicial $x_0=(0,0,0)^{\mathrm{T}}$. Comparar los resultados obtenidos con los de la práctica 9.2.

9.4. Utilizar los programas de las prácticas 9.1, 9.2 y 9.3 para aproximar la solución del sistema no lineal

$$\begin{cases} x^3+6x-y-2z &= 40 \\ -x+7y+y^5-4z &= -3 \\ -3x-y+z^3+8z &= 14. \end{cases}$$

9.5. Utilizar el método de relajación–Newton con los valores de $w=0:0.1:2$ para aproximar la solución del sistema del ejemplo 9.2. ¿Cuál es el valor óptimo del parámetro w?

10 Cálculo de raíces de polinomios

10.1. Introducción

En este capítulo nos disponemos a estudiar métodos diseñados para aproximar las raíces de una ecuación

$$F(x) = 0$$

en el caso particular de que la función F sea un polinomio. Si bien todos los métodos presentados en el capítulo 8 pueden ser usados para abordar este problema, existen otros más indicados para este tipo de ecuaciones, que utilizan propiedades específicas de los polinomios.

Se recoge aquí una selección de métodos más bien clásicos; además de su probada eficacia para aproximar raíces de polinomios, alguno de ellos tiene relevancia desde el punto de vista teórico, por su aplicabilidad en contextos más generales (en especial, el *método de Sturm*).

Sin embargo, debemos hacer notar que, si se necesita aproximar todas las raíces de un polinomio

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \ \text{ con } \ a_n \neq 0,$$

la estrategia más adecuada suele ser construir la *matriz asociada al polinomio* (*companion matrix*, en inglés)

$$\begin{pmatrix} -\frac{a_{n-1}}{a_n} & -\frac{a_{n-2}}{a_n} & \cdots & -\frac{a_2}{a_n} & -\frac{a_1}{a_n} & -\frac{a_0}{a_n} \\ 1 & 0 & \cdots & 0 & 0 & 0 \\ & 1 & \cdots & 0 & 0 & 0 \\ & & \ddots & \cdots & \cdots & \cdots \\ & & & 1 & 0 & 0 \\ & & & & 1 & 0 \end{pmatrix},$$

matriz que tiene a $P(x)$ por polinomio característico, y calcular sus autovalores. Por supuesto, para ello habrá que contar con un buen método de aproximación de autovalores de una matriz; el comando `roots` de `MATLAB` aplica, precisamente, esta estrategia.

Comenzamos recordando algunos resultados y dando las definiciones básicas.

10.2. Algunas propiedades de los polinomios

Definición 10.1. Si $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{R}[x]$ entonces

$$P(x) = 0$$

es una *ecuación algebraica*. □

Observación 10.1. Este tipo de ecuaciones pueden resolverse por métodos algebraicos cuando $\partial P \leq 4$, obteniéndose las raíces de forma directa (aunque los cálculos pueden llegar a ser altamente complicados). *Galois* demostró que la ecuación anterior no puede resolverse, en general, por métodos algebraicos para polinomios de grado mayor que 4. □

Recordemos, sin demostración, el resultado que define la división de polinomios.

Teorema 10.1. *Sean $P(x)$, $Q(x) \in \mathbb{R}[x]$ con $\partial P \geq \partial Q$. Entonces existen unos únicos polinomios $C(x)$, $R(x) \in \mathbb{R}[x]$ con $\partial C = \partial P - \partial Q$ y $\partial R < \partial Q$ tales que*

$$P(x) = C(x)Q(x) + R(x).$$

$C(x)$ es el cociente de la división de $P(x)$ entre $Q(x)$ y $R(x)$ es el resto. □

Cuando el resto de dividir $P(x)$ entre $Q(x)$ es cero se tiene la siguiente noción:

Definición 10.2. Sean $P(x)$, $Q(x) \in \mathbb{R}[x]$. Se dice que $Q(x)$ es un *divisor* de (o *divide* a) $P(x)$, y se denota

$$Q(x)|P(x),$$

si existe $C(x) \in \mathbb{R}[x]$ tal que

$$P(x) = C(x)Q(x). \quad \square$$

A partir del teorema 10.1 se deduce el siguiente resultado:

Corolario 10.1. *Si $P(x) \in \mathbb{R}[x]$ y $\xi \in \mathbb{R}$ entonces existe un único polinomio $C(x) \in \mathbb{R}[x]$ con $\partial C = \partial P - 1$ tal que*

$$P(x) = (x - \xi)C(x) + P(\xi).$$

Por tanto,

$$\xi \text{ es raíz de } P \Leftrightarrow (x - \xi) | P(x). \quad \square$$

Para las raíces múltiples se tiene la siguiente caracterización:

Proposición 10.1. *Sea $P(x) \in \mathbb{R}[x]$. $\xi \in \mathbb{R}$ es raíz de multiplicidad m de P si y sólo si*

$$P(\xi) = P'(\xi) = \cdots = P^{m-1)}(\xi) = 0 \ \text{ y } \ P^{m)}(\xi) \neq 0.$$

Demostración.

$\boxed{\Rightarrow}$ Por definición,

$$P(x) = (x - \xi)^m Q(x)$$

con $\partial Q = n - m$ y $Q(\xi) \neq 0$. Derivando sucesivamente se obtiene

$$P^{k)}(x) = m(m-1)\cdots(m-k+1)(x-\xi)^{m-k}Q(x) + (x-\xi)^{m-k+1}\psi_k(x)$$

para $k = 1, 2, \ldots, m-1$, donde las funciones $\{\psi_1(x), \psi_2(x), \ldots, \psi_{m-1}(x)\}$ son polinomios que incluyen las sucesivas derivadas de $Q(x)$. Claramente,

$$P^{k)}(\xi) = 0, \ k = 0, 1, \ldots, m-1 \ \text{ y } \ P^{m)}(\xi) = m!Q(\xi) \neq 0.$$

$\boxed{\Leftarrow}$ Como P es una función analítica y $\partial P = n$ entonces $P^{k)} \equiv 0$, $k \geq n+1$. Así, desarrollando en serie de Taylor se obtiene

$$\begin{aligned} P(x) &= \sum_{k=0}^{m-1} \frac{P^{k)}(\xi)}{k!}(x-\xi)^k + \sum_{k=m}^{n} \frac{P^{k)}(\xi)}{k!}(x-\xi)^k + \sum_{k=n+1}^{\infty} \frac{P^{k)}(\xi)}{k!}(x-\xi)^k \\ &= \sum_{k=m}^{n} \frac{P^{k)}(\xi)}{k!}(x-\xi)^k = (x-\xi)^m Q(x) \end{aligned}$$

siendo

$$Q(x) = \sum_{k=m}^{n} \frac{P^{k)}(\xi)}{k!}(x-\xi)^{k-m} = \sum_{j=0}^{n-m} \frac{P^{m+j)}(\xi)}{(m+j)!}(x-\xi)^j$$

con $\partial Q = n - m$ y $Q(\xi) = \dfrac{P^{m)}(\xi)}{m!} \neq 0$. $\quad \square$

Definición 10.3. Sean $P(x), Q(x) \in \mathbb{R}[x]$. Un polinomio $S(x) \in \mathbb{R}[x]$ es el *máximo común divisor* de $P(x)$ y $Q(x)$ (y lo notaremos $S(x) = \text{MCD}\{P(x), Q(x)\}$) si $S(x)|P(x)$, $S(x)|Q(x)$ y para cualquier polinomio $\widetilde{S}(x) \in \mathbb{R}[x]$ tal que $\widetilde{S}(x)|P(x)$, $\widetilde{S}(x)|Q(x)$ se cumple que $\widetilde{S}(x)|S(x)$. □

Observación 10.2. Dada una ecuación algebraica $P(x) = 0$ siempre es posible encontrar un polinomio $\widetilde{P}(x)$ con las mismas raíces que $P(x)$ pero todas ellas simples. En efecto, supongamos que $\{\xi_1, \xi_2, \ldots, \xi_r\}$ son las raíces de la ecuación

$$P(x) = 0$$

con multiplicidad respectiva $\{m_1, m_2, \ldots, m_r\}$ y consideremos la descomposición factorial del polinomio

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{R}[x] \text{ con } a_n \neq 0,$$

es decir,

$$P(x) = a_n(x - \xi_1)^{m_1}(x - \xi_2)^{m_2} \cdots (x - \xi_r)^{m_r}$$

con

$$m_1 + m_2 + \cdots + m_r = n.$$

De esta forma su derivada toma la expresión (compruébese)

$$P'(x) = a_n(x - \xi_1)^{m_1 - 1}(x - \xi_2)^{m_2 - 1} \cdots (x - \xi_r)^{m_r - 1} Q(x)$$

donde

$$Q(\xi_k) \neq 0$$

para $k = 1, 2, \ldots, r$. Así pues, el máximo común divisor de $P(x)$ y $P'(x)$ es

$$\text{MCD}\{P(x), P'(x)\} = a_n(x - \xi_1)^{m_1 - 1}(x - \xi_2)^{m_2 - 1} \cdots (x - \xi_r)^{m_r - 1}$$

y, por tanto, basta tomar

$$\widetilde{P}(x) = \frac{P(x)}{\text{MCD}\{P(x), P'(x)\}} = (x - \xi_1)(x - \xi_2) \cdots (x - \xi_r)$$

que tiene por raíces simples los valores $\{\xi_1, \xi_2, \ldots, \xi_r\}$. Por lo tanto, para calcular las raíces de P bastará aproximar las de $\widetilde{P}$, con la ventaja de que ahora son todas simples. Así, por ejemplo, la ecuación algebraica

$$x^5 - 3x^3 - 2x^2 = 0$$

tiene por raíces $\xi_1 = 0$ (doble), $\xi_2 = -1$ (doble) y $\xi_3 = 2$ (simple). El máximo común divisor de $P(x)$ y $P'(x)$ es $x(x+1)$ y entonces

$$\widetilde{P}(x) = \frac{P(x)}{\text{MCD}\{P(x), P'(x)\}} = x^3 - x^2 - 2x.$$

Como se observa, las raíces de la ecuación

$$x^3 - x^2 - 2x = 0$$

son $\widetilde{\xi}_1 = 0$, $\widetilde{\xi}_2 = -1$ y $\widetilde{\xi}_3 = 2$ todas ellas simples. □

10.3. Algoritmo de Horner

En los métodos que presentamos para el cálculo de las raíces de una ecuación algebraica $P(x) = 0$ será necesario evaluar el polinomio P en una gran cantidad de puntos (las aproximaciones sucesivas de las raíces). Por ello será necesario utilizar algoritmos que permitan realizar estas evaluaciones de la forma más rápida y sencilla posible. El algoritmo más apropiado es el de Horner: para evaluar el polinomio

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \ \text{ con } \ a_n \neq 0$$

en un punto ξ dividimos $P(x)$ entre $x - \xi$ mediante la regla de Ruffini, es decir,

	a_n	a_{n-1}	a_{n-2}	$\cdots$	a_2	a_1	a_0
ξ		ξb_{n-1}	ξb_{n-2}	$\cdots$	ξb_2	ξb_1	ξb_0
	b_{n-1}	b_{n-2}	b_{n-3}	$\cdots$	b_1	b_0	b_{-1}

donde

$$\begin{cases} b_{n-1} = a_n \\ b_k = \xi b_{k+1} + a_{k+1}, \ k = n-2, n-3, \ldots, 0, -1. \end{cases} \tag{10.1}$$

Como

$$P(x) = (x - \xi)Q(x) + b_{-1}$$

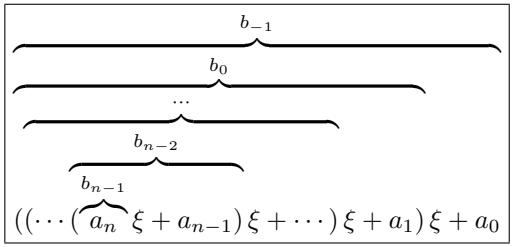
siendo

$$Q(x) = b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + \cdots + b_1 x + b_0,$$

entonces

$$\boxed{P(\xi) = b_{-1}}$$

Esquemáticamente, las iteraciones en el método de Horner (también conocido como *multiplicación anidada* o *división sintética*) vienen dadas por



donde los coeficientes $\{b_{n-1}, b_{n-2}, \ldots, b_0, b_{-1}\}$ son los dados en (10.1).

Observación 10.3. Si $\xi \in \mathbb{R}$ es raíz de $P(x) = 0$ entonces

$$P(x) = (x - \xi)Q(x)$$

por lo que

$$P'(x) = Q(x) + (x - \xi)Q'(x).$$

De esta forma, como

$$P'(\xi) = Q(\xi),$$

podemos aplicar el algoritmo de Horner a $Q(x)$ para hallar el valor de $P'(\xi)$. Este argumento se puede aplicar sucesivamente para saber si una raíz es múltiple y para determinar su multiplicidad. □

10.4. Métodos de acotación de raíces

La acotación de raíces consiste en la determinación de conjuntos fuera de los cuales podemos asegurar que no existen raíces del polinomio.

Existe una gran variedad de métodos de acotación de raíces, pero nos restringiremos al *método de McLaurin* por la sencillez de su implementación. Otros métodos (como el de *Laguerre* o *Newton*) acotan las raíces reales con mayor efectividad pero presentan el inconveniente de que las cotas que proponen hay que hallarlas mediante el procedimiento de prueba y error (véanse los problemas 10.1 y 10.2).

El *método de McLaurin* nos va a permitir acotar en módulo todas las raíces (ya sean reales o complejas) de la ecuación algebraica

$$P(x) = 0$$

donde

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{R}[x] \;\text{ con }\; a_n \neq 0$$

como se explica a continuación:

Teorema 10.2. *Si* $\{\xi_1, \xi_2, \ldots, \xi_n\} \subset \mathbb{C}$ *son las* n *raíces de* P *entonces*

$$|\xi_k| < 1 + \lambda$$

para $k = 1, 2, \ldots, n$, *donde*

$$\lambda = \max_{0 \le k \le n-1} \left| \frac{a_k}{a_n} \right|. \tag{10.2}$$

Es decir, las raíces de P *se encuentran en el interior de la bola abierta*

$$\mathbf{B}_{1+\lambda}(0) = \{z \in \mathbb{C} : \ |z| < 1 + \lambda\}.$$

Demostración. Como $\lambda \ge 0$, pueden presentarse dos casos:

a) $\lambda = 0$. En esta situación

$$a_0 = a_1 = \cdots = a_{n-1} = 0$$

por lo que $P(x) = a_n x^n$ sólo tiene la raíz $x = 0$ con multiplicidad n y el resultado se tiene trivialmente.

b) $\lambda > 0$. Supongamos que existiera una raíz $\xi_k \in \mathbb{C}$ de P tal que $|\xi_k| \ge 1 + \lambda$ y derivemos una contradicción. Como

$$|\xi_k| \ge 1 + \lambda > 1 \ \text{ y } \ |\xi_k| - 1 \ge \lambda > 0$$

entonces

$$|\xi_k|^n - 1 > 0 \ \text{ y } \ \frac{\lambda}{|\xi_k| - 1} \le 1. \tag{10.3}$$

Por ser ξ_k raíz de P entonces

$$0 = P(\xi_k) = a_n \xi_k^n + a_{n-1} \xi_k^{n-1} + \cdots + a_1 \xi_k + a_0$$

y, como $a_n \neq 0$, despejando obtenemos

$$-\xi_k^n = \frac{a_0}{a_n} + \frac{a_1}{a_n} \xi_k + \cdots + \frac{a_{n-1}}{a_n} \xi_k^{n-1}.$$

De esta forma, tomando módulos en la expresión anterior, se tiene que

$$\begin{aligned} |\xi_k|^n &= \left| \frac{a_0}{a_n} + \frac{a_1}{a_n} \xi_k + \cdots + \frac{a_{n-1}}{a_n} \xi_k^{n-1} \right| \\ &\le \left| \frac{a_0}{a_n} \right| + \left| \frac{a_1}{a_n} \right| |\xi_k| + \cdots + \left| \frac{a_{n-1}}{a_n} \right| |\xi_k|^{n-1} \\ &\le \lambda \left(1 + |\xi_k| + \cdots + |\xi_k|^{n-1}\right) = \lambda \sum_{j=0}^{n-1} |\xi_k|^j = \lambda \frac{|\xi_k|^n - 1}{|\xi_k| - 1} \end{aligned}$$

donde hemos utilizado (10.2). Así, a partir de (10.3), se llega a la contradicción

$$|\xi_k|^n \leq \frac{\lambda}{|\xi_k|-1}\left(|\xi_k|^n-1\right) \leq |\xi_k|^n - 1. \quad \square$$

Corolario 10.2. *Si $a_0 \neq 0$ y $\{\xi_1, \xi_2, \ldots, \xi_n\} \subset \mathbb{C}$ son las n raíces de P entonces*

$$|\xi_k| > \frac{1}{1+\mu}$$

para $k = 1, 2, \ldots, n$, donde

$$\mu = \max_{1\leq k\leq n} \left|\frac{a_k}{a_0}\right|.$$

Por tanto, las raíces de P se encuentran en el exterior de la bola cerrada

$$\overline{\mathbf{B}}_{\frac{1}{1+\mu}}(0) = \left\{ z \in \mathbb{C} : \ |z| \leq \frac{1}{1+\mu} \right\}.$$

Demostración. Haciendo el cambio de variable

$$y = \frac{1}{x} \tag{10.4}$$

podemos escribir

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = \frac{a_n}{y^n} + \frac{a_{n-1}}{y^{n-1}} \cdots + \frac{a_1}{y} + a_0 = \frac{Q(y)}{y^n}$$

siendo

$$Q(y) = a_0 y^n + a_1 y^{n-1} + \cdots + a_{n-1} y + a_n.$$

De esta forma,

$$P(x) = 0, \ x \neq 0 \ \Leftrightarrow \ Q(y) = 0, \ y \neq 0. \tag{10.5}$$

Nótese que al ser $a_0 \neq 0$ entonces $x = 0$ no es raíz de P y además, por el teorema 10.2, sabemos que las raíces $\{\eta_1, \eta_2, \ldots, \eta_n\} \subset \mathbb{C}$ de $Q(y) = 0$ verifican

$$|\eta_k| < 1 + \mu \tag{10.6}$$

para $k = 1, 2, \ldots, n$. Por tanto, a partir de (10.4), (10.5) y (10.6), para todo índice $k \in \{1, 2, \ldots, n\}$ se tiene que

$$|\xi_k| = \frac{1}{|\eta_k|} > \frac{1}{1+\mu}. \quad \square$$

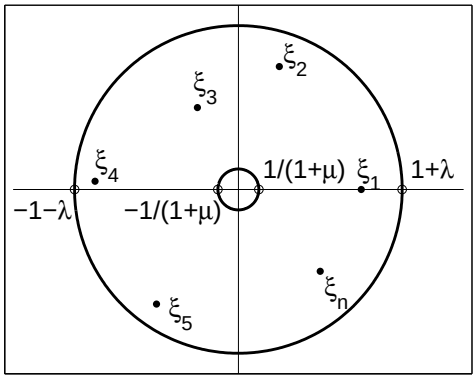


Figura 10.1: Acotación de McLaurin.

Observación 10.4. Con las notaciones anteriores, cuando $a_0 \neq 0$, las n raíces de P se encuentran en el interior de la corona circular

$$C = \left\{ z \in \mathbb{C} : \ \frac{1}{1+\mu} < |z| < 1 + \lambda \right\}$$

(véase la figura 10.1). □

Ejemplo 10.1. Acotando las raíces de la ecuación

$$2x^4 + 4x^3 - 59x^2 - 61x + 30 = 0$$

mediante el método de McLaurin se obtiene que

$$\lambda = \max\left\{\frac{4}{2}, \frac{59}{2}, \frac{61}{2}, \frac{30}{2}\right\} = \frac{61}{2} \ \text{ y } \ \mu = \max\left\{\frac{2}{30}, \frac{4}{30}, \frac{59}{30}, \frac{61}{30}\right\} = \frac{61}{30}.$$

Por tanto, si $\{\xi_1, \xi_2, \xi_3, \xi_4\} \subset \mathbb{C}$ son las raíces de dicha ecuación, se verifica que

$$0.329670 \simeq \frac{30}{91} = \frac{1}{1+\mu} < |\xi_k| < 1 + \lambda = \frac{63}{2} = 31.5$$

para $k = 1, 2, 3, 4$. Las raíces de P son

$$\xi_1 = 5, \ \xi_2 = -6, \ \xi_3 = \frac{-1+\sqrt{3}}{2} \simeq 0.3660 \ \text{ y } \ \xi_4 = \frac{-1-\sqrt{3}}{2} \simeq -1.3660. \quad \square$$

10.5. Separación de raíces reales

El objetivo de los métodos de separación de raíces reales es la determinación del número de raíces reales y distintas que tiene un polinomio y de intervalos en los cuales podamos asegurar que existe una y sólo una raíz. Recordemos algunas peculiaridades que presentan los polinomios:

a) Si $P(a)P(b) < 0$, por el teorema de Bolzano se tiene que P tiene un número impar de raíces en (a, b) (contadas con su multiplicidad).

b) Si $P(a)P(b) > 0$ o bien P no tiene raíces reales en (a, b) o bien tiene un número par de raíces en (a, b) (contadas con su multiplicidad).

10.5.1. Regla de los signos de Descartes

Se trata de un resultado clásico que aporta información previa acerca del número de raíces positivas y negativas de una ecuación algebraica y que deberá ser complementado con otro tipo de técnicas.

Definición 10.4. Se denomina *número de cambios de signo* en la secuencia $\{c_1, c_2, \ldots, c_k\} \subset \mathbb{R}$ al número total de cambios de signo en cada par de elementos consecutivos de la secuencia que resulta al eliminar los eventuales elementos nulos en $\{c_1, c_2, \ldots, c_k\}$. □

Ejemplo 10.2. En la secuencia $\{1, 0, -2, 4, 7, 0, -21\}$ hay tres cambios de signo mientras que en $\{-1, -2, 4, 5, -7\}$ sólo hay dos. □

Teorema 10.3 (Regla de los signos de Descartes). *Sea un polinomio*

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{R}[x] \ \text{ con } \ a_n \neq 0.$$

El número de raíces positivas de la ecuación $P(x) = 0$ (*contando cada raíz con su multiplicidad*) *es igual al número de cambios de signo en la secuencia de los coeficientes* $\{a_n, a_{n-1}, \ldots, a_0\}$ *o menor que dicho número en un entero par.*

Demostración. Véase, por ejemplo, [**Ko**]. □

Observación 10.5.

1. Este resultado también es válido para las raíces negativas sin más que aplicarlo a la ecuación $P(-x) = 0$, dado que las raíces negativas de $P(x) = 0$ se corresponden con las raíces positivas de la ecuación $P(-x) = 0$.

2. La regla de los signos es un caso particular del teorema de Budan–Fourier (véase [**De–Ma**]). □

Ejemplo 10.3. Vamos a determinar el número de raíces reales positivas y negativas del polinomio

$$P(x) = x^5 + x^3 - x^2 - 10x + 1.$$

En este caso la secuencia de coeficientes del polinomio P

$$\{1, 1, -1, -10, 1\}$$

presenta dos cambios de signo, por lo que el polinomio P tiene dos raíces positivas o ninguna. Para discernir lo anterior, como

$$P(0) = 1 \ \text{ y } \ P(1) = -8$$

entonces, por el teorema de Bolzano, existe $\xi \in (0, 1)$ tal que $P(\xi) = 0$; consecuentemente, P tiene dos raíces positivas. Por otra parte,

$$P(-x) = -x^5 - x^3 - x^2 + 10x + 1$$

cuya secuencia de coeficientes

$$\{-1, -1, -1, 10, 1\}$$

presenta un cambio de signo; esto hace que P tenga una raíz negativa. Por tanto, las otras dos raíces de P son complejas. □

10.5.2. Método de Sturm

Ahora nos proponemos determinar el número de raíces reales distintas que tiene la ecuación algebraica $P(x) = 0$, donde

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{R}[x] \ \text{ con } \ a_n \neq 0,$$

así como separar dichas raíces. El *método de Sturm*, de hecho, nos va a dar el número de raíces reales distintas del polinomio P en cualquier intervalo. Este método consiste en construir una secuencia de polinomios, que surge al calcular el máximo común divisor de $P(x)$ y $P'(x)$ mediante el algoritmo de Euclides, y utilizar, posteriormente, las propiedades de dicha secuencia. Denotando por

$$P_1(x) = P'(x),$$

dividimos el polinomio $P(x)$ entre $P_1(x)$, luego dividimos $P_1(x)$ entre $P_2(x)$ definido como el resto cambiado de signo de la división anterior; a continuación,

$P_2(x)$ entre $P_3(x)$ que es el resto, cambiado de signo, de la división previa y así sucesivamente:

$$\begin{cases} P(x) = P_1(x)Q_1(x) - P_2(x), & \partial P_2 < \partial P_1 \\ P_1(x) = P_2(x)Q_2(x) - P_3(x), & \partial P_3 < \partial P_2 \\ \dots & \\ P_{m-2}(x) = P_{m-1}(x)Q_{m-1}(x) - P_m(x), & \partial P_m < \partial P_{m-1} \\ P_{m-1}(x) = P_m(x)Q_m(x). & \end{cases} \tag{10.7}$$

Nótese que el proceso se detiene al llegar a una división exacta.

Definición 10.5. La secuencia $\{P(x), P_1(x), \dots, P_m(x)\}$ dada en (10.7) se denomina *secuencia de Sturm* para el polinomio P. □

Observación 10.6. En el proceso anterior obtenemos el máximo común divisor de $P(x)$ y $P'(x)$ que es, precisamente, $P_m(x)$ (véase el problema 10.5). Por tanto:

a) Si $P_m(x)$ es una constante entonces todas las raíces de P son simples.

b) En el caso de que

$$P_m(x) = c(x-\xi_1)^{m_1}(x-\xi_2)^{m_2}\cdots(x-\xi_r)^{m_r}$$

entonces cada ξ_k es raíz de P de multiplicidad $m_k + 1$ y

$$P(x) = (x-\xi_1)^{m_1+1}(x-\xi_2)^{m_2+1}\cdots(x-\xi_r)^{m_r+1}Q(x)$$

donde

$$Q(\xi_k) \neq 0$$

para $k = 1, 2, \dots, r$, y las raíces de $Q(x) = 0$ son las raíces simples de $P(x) = 0$ (véase la observación 10.2). □

Notación 10.1. Vamos a denotar por:

- $N(\alpha)$ al número de cambios de signo en la secuencia de Sturm para el polinomio $P(x)$ particularizada en $x = \alpha$, es decir, en la secuencia

$$\{P(\alpha), P_1(\alpha), \dots, P_m(\alpha)\}.$$

- $N(a, b)$ al número de raíces reales distintas de la ecuación $P(x) = 0$ en el intervalo (a, b) (sin contar la multiplicidad). □

Teorema 10.4 (Sturm). *Consideremos un polinomio*

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{R}[x] \ \ con \ \ a_n \neq 0.$$

Si $P(a)P(b) \neq 0$ *entonces*

$$N(a,b) = N(a) - N(b),$$

es decir, el número de raíces reales distintas de P *en el intervalo* (a,b) *es igual a la diferencia entre el número de cambios de signo en la secuencia de Sturm del polinomio* P *en* $x = a$ *y en* $x = b$.

DEMOSTRACIÓN. Trataremos únicamente el caso en que todas las raíces reales del polinomio P son simples (el caso en que P tiene raíces múltiples es considerado en el problema 10.6). Denotando por $\{P_0, P_1, \ldots, P_m\}$ la secuencia de Sturm del polinomio $P \equiv P_0$ se tiene que $P_m(x)$ es una constante no nula y, por tanto,

$$P_m(x) \neq 0, \ x \in \mathbb{R}. \tag{10.8}$$

Consideramos la función $N : \mathbb{R} \to \{0, 1, \ldots, n\}$ que a cada número real x le asocia el número de cambios de signo de la secuencia anterior. Claramente, la función $N(x)$ sólo podrá cambiar de valor en un punto $x = \alpha$ si α anula alguno de los polinomios $\{P_0, P_1, \ldots, P_m\}$.

TABLA 10.1:
Signo de los polinomios P_0 *y* P_1 *en* α^- *y* α^+ *siendo* $P_0(\alpha) = 0$

	Signo de P_0	Signo de P_1
α^-	$-$	$+$
α^+	$-$	$-$

	Signo de P_0	Signo de P_1
α^-	$-$	$+$
α^+	$+$	$+$

	Signo de P_0	Signo de P_1
α^-	$+$	$-$
α^+	$-$	$-$

	Signo de P_0	Signo de P_1
α^-	$+$	$-$
α^+	$+$	$+$

a) Estudiemos, en primer lugar, el comportamiento de $N(x)$ al pasar por una raíz α de P_0. Las situaciones que pueden presentarse vienen recogidas en la tabla 10.1 y en la figura 10.2 y se corresponden con el hecho de que para pasar P_0 de un valor positivo (respectivamente, negativo) a cero, su derivada P_1 tendrá que ser negativa (respectivamente, positiva) y, análogamente, para pasar de cero a un valor no nulo. Por tanto, hemos demostrado que cuando $N(x)$ pasa por una raíz $x = \alpha$ de P_0 su valor disminuye, al menos, en una unidad (a la izquierda de α la secuencia $\{P_0, P_1\}$ tiene un cambio de signo mientras que a la derecha de α no lo tiene).

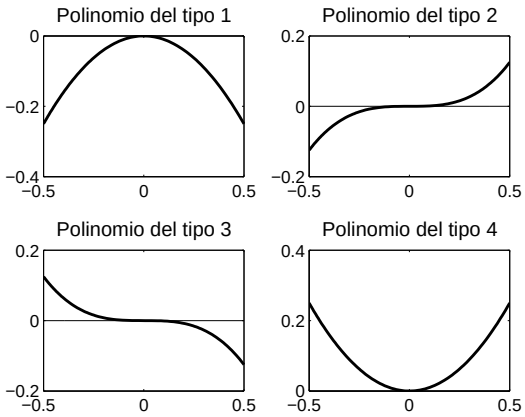


Figura 10.2: Diversos polinomios P_0 en torno a la raíz $\alpha = 0$.

b) Analizamos a continuación el comportamiento de $N(x)$ cuando pasa por una raíz α de algún polinomio $P_i(x)$ para $i \in \{1, 2, \ldots, m\}$. Como $P_i(\alpha) = 0$ entonces, por la construcción de la secuencia de Sturm, se tiene que

$$P_{i-1}(\alpha) = P_i(\alpha)Q_i(\alpha) - P_{i+1}(\alpha) = -P_{i+1}(\alpha).$$

Veamos que

$$P_{i-1}(\alpha) = -P_{i+1}(\alpha) \neq 0.$$

En efecto, si $P_{i-1}(\alpha) = -P_{i+1}(\alpha) = 0$ entonces, por la construcción de la secuencia de Sturm, se llega a que

$$P_1(\alpha) = P_2(\alpha) = \cdots = P_m(\alpha) = 0$$

lo que contradice (10.8). De esta forma, puesto que los polinomios $P_{i-1}(x)$ y $P_{i+1}(x)$ no se anulan en α, tendrán signo constante (y opuesto) en un entorno de α. Por tanto, independientemente del signo que tome $P_i(x)$ en α^- y α^+, la secuencia

$$\{P_{i-1}(\alpha^-), P_i(\alpha), P_{i+1}(\alpha^-)\} = \{P_{i-1}(\alpha^-), 0, -P_{i-1}(\alpha^-)\}$$

tendrá exactamente un cambio de signo, al igual que la secuencia

$$\{P_{i-1}(\alpha^+), P_i(\alpha), P_{i+1}(\alpha^+)\} = \{P_{i-1}(\alpha^+), 0, -P_{i-1}(\alpha^+)\}.$$

De esta forma, $N(x)$ no cambia su valor al pasar por la raíz α de $P_i(x)$ salvo que α sea también raíz de $P_0(x)$, en cuyo caso disminuye exactamente en una unidad (como se ha visto en el apartado *a*)).

Así pues, hemos demostrado que la función $N(x)$ disminuye exactamente en una unidad al pasar por una raíz de $P_0(x)$ y no varía al pasar por cualquier otro valor. Se puede concluir, por tanto, que al pasar de a a b la función $N(x)$ habrá disminuido en tantas unidades como raíces reales (simples) tenga $P_0(x)$ en el intervalo (a, b); esto es, $N(a) - N(b) = N(a, b)$. □

Observación 10.7. Como lo único que nos interesa es el número de cambios de signo en la secuencia de Sturm $\{P(x), P_1(x), \ldots, P_m(x)\}$, podemos multiplicar los polinomios $P_k(x)$ por constantes positivas con vistas a simplificar los cálculos. De hecho, con un abuso de notación, cuando nos refiramos a "la" secuencia de Sturm del polinomio $P(x)$, estaremos considerando una cualquiera de las así construidas. □

Como consecuencia del teorema 10.4 se tiene:

Corolario 10.3. *Sea* $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{R}[x]$ *con* $a_n \neq 0$ *y denotemos*

$$N(\pm\infty) = \lim_{x \to \pm\infty} N(x).$$

a) Si $P(0) \neq 0$ *entonces*

$$\begin{cases} N_+ = N(0) - N(+\infty) & \text{es el número de raíces positivas de } P \\ N_- = N(-\infty) - N(0) & \text{es el número de raíces negativas de } P. \end{cases}$$

b) Las n raíces de P son reales y simples si y sólo si $N(-\infty) - N(+\infty) = n$. □

Ejemplo 10.4. Vamos a determinar el número de raíces reales positivas y negativas del polinomio

$$P(x) = x^4 - 4x + 1.$$

Como

$$P'(x) = 4x^3 - 4$$

podemos tomar

$$P_1(x) = \frac{P'(x)}{4} = x^3 - 1.$$

La división de $P(x)$ entre $P_1(x)$ determina que

$$x^4 - 4x + 1 = (x^3 - 1)x + (-3x + 1)$$

por lo que tomamos $P_2(x) = 3x - 1$. Finalmente, como al dividir $P_1(x)$ entre $P_2(x)$ se obtiene que

$$x^3 - 1 = (3x - 1)\left(\frac{1}{3}x^2 + \frac{1}{9}x + \frac{1}{27}\right) - \frac{26}{27},$$

tomamos $P_3(x) = 1$ y, de esta forma, la secuencia de Sturm para el polinomio P viene dada por

$$\{P(x), P_1(x), P_2(x), P_3(x)\} = \left\{x^4 - 4x + 1, x^3 - 1, 3x - 1, 1\right\}.$$

Nótese que las raíces del polinomio $P(x)$ serán todas simples al ser $\partial P_3 = 0$. Como

x	$P(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$N(x)$
$-\infty$	$+$	$-$	$-$	$+$	2
0	$+$	$-$	$-$	$+$	2
$+\infty$	$+$	$+$	$+$	$+$	0

entonces

$$\begin{cases} N_- = N(-\infty) - N(0) = 0 \\ N_+ = N(0) - N(+\infty) = 2 \end{cases}$$

por lo que P tiene dos raíces positivas, ninguna negativa y dos complejas conjugadas (dado que $\partial P = 4$ y las raíces de P son simples). □

El teorema de Sturm, combinado con un proceso de bisección, proporciona un método (el *método de Sturm*) para separar las raíces de un polinomio:

Supongamos que todas las raíces están en un intervalo (a, b) (por ejemplo, si no hay otra información, puede tomarse $(a, b) = (-1 - \lambda, 1 + \lambda)$, siendo $1 + \lambda$ la cota dada por el método de McLaurin).

a) Si $N(a) - N(b) = 1$, el proceso está acabado.

b) Si $N(a) - N(b) > 1$ se toma $c = \dfrac{a+b}{2}$ y se consideran los intervalos (a, c) y (c, b), calculando $N(a) - N(c)$ y $N(c) - N(b)$. Si alguno es mayor que 1, se repite el proceso en el intervalo correspondiente.

Se llega así a tantos intervalos como raíces reales y distintas tenga el polinomio, intervalos en los que tan sólo hay una raíz.

Ejemplo 10.5. Para el polinomio del ejemplo 10.4 las raíces positivas estarán en el intervalo $(0, 1 + \lambda) = (0, 5)$. Sabemos que $N(0) = 2$ y $N(5) = N(+\infty) = 0$. Puede comprobarse que $N(2.5) = 0$, por lo que ambas raíces estarán en el intervalo $(0, 2.5)$. Calculando $N(1.25) = 1$, deducimos que una raíz está en $(0, 1.25)$ y la otra en $(1.25, 2.5)$. □

10.6. Ecuaciones con coeficientes racionales

En esta sección presentamos algunos resultados específicos para la obtención de raíces enteras y racionales de polinomios con coeficientes racionales de la forma

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{Q}[x] \text{ con } a_n \neq 0.$$

En el caso en que el polinomio tenga algún coeficiente irracional, no hay resultados que permitan estudiar separadamente la aproximación de raíces enteras y racionales.

Sin pérdida de generalidad podemos suponer que los coeficientes del polinomio anterior son enteros. En efecto, si los coeficientes de $P(x)$ se escriben como

$$a_i = \frac{p_i}{q_i}.$$

con $p_i \in \mathbb{Z}$ y $q_i \in \mathbb{N}$, considerando el mínimo común múltiplo de los denominadores

$$m = \text{mcm}\,\{q_n, q_{n-1}, \ldots, q_0\},$$

se tiene que el polinomio

$$Q(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$$

donde

$$b_i = m\frac{p_i}{q_i} \in \mathbb{Z}$$

para $i = 0, 1, \ldots, n$, verifica

$$Q(x) = mP(x),$$

por lo que las ecuaciones $P(x) = 0$ y $Q(x) = 0$ son equivalentes.

Así pues, en lo que resta de sección, supondremos que

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{Z}[x] \text{ con } a_n \neq 0.$$

Comenzamos recordando el resultado que muestra que las raíces enteras de P han de ser divisores del término independiente.

Proposición 10.2. *Si $\xi \in \mathbb{Z}$ es raíz del polinomio P entonces $\xi | a_0$.*

Demostración. Como $\xi \in \mathbb{Z}$ es raíz de P entonces

$$0 = P(\xi) = a_n\xi^n + a_{n-1}\xi^{n-1} + \cdots + a_0,$$

de donde

$$a_0 = -(a_n\xi^{n-1} + a_{n-1}\xi^{n-2} + \cdots + a_1)\xi. \quad \square$$

Las raíces racionales del polinomio P pueden ser determinadas a partir del siguiente resultado:

Proposición 10.3. *Si* $\frac{p}{q} \in \mathbb{Q}$ (*fracción irreducible*) *es raíz del polinomio* P *entonces* $p|a_0$ *y* $q|a_n$.

DEMOSTRACIÓN. Como $\frac{p}{q} \in \mathbb{Q}$ es raíz de P entonces

$$0 = P\left(\frac{p}{q}\right) = a_n\left(\frac{p}{q}\right)^n + a_{n-1}\left(\frac{p}{q}\right)^{n-1} + \cdots + a_1\left(\frac{p}{q}\right) + a_0$$

y, por tanto,

$$a_0 q^n + a_1 p q^{n-1} + \cdots + a_{n-1} p^{n-1} q + a_n p^n = 0.$$

De esta forma

$$\begin{cases} a_0 q^n = -(a_1 q^{n-1} + \cdots + a_{n-1} p^{n-2} q + a_n p^{n-1}) p & \Rightarrow \quad p|a_0 q^n \\ a_n p^n = -(a_0 q^{n-1} + a_1 p q^{n-2} + \cdots + a_{n-1} p^{n-1}) q & \Rightarrow \quad q|a_n p^n \end{cases}$$

de donde $p|a_0$ y $q|a_n$ ya que p y q son primos entre sí, al ser la fracción irreducible. □

Observación 10.8.

1. De los posibles candidatos a raíces enteras o racionales de la ecuación excluiremos aquellos valores que caigan fuera de los intervalos de acotación de las raíces (obtenidos previamente) y a los restantes valores se les aplicará el algoritmo de Horner para comprobar si son raíces.
2. Cuando $a_n = \pm 1$ las posibles raíces racionales se reducen a las enteras. □

Ejemplo 10.6. Vamos a resolver la ecuación

$$10x^4 - 11x^3 - 41x^2 + x + 6 = 0$$

calculando sus cuatro raíces $\{\xi_1, \xi_2, \xi_3, \xi_4\}$.

1. Número de raíces positivas y negativas. A partir de

$$\begin{cases} P(x) = 10x^4 - 11x^3 - 41x^2 + x + 6 \\ P(-x) = 10x^4 + 11x^3 - 41x^2 - x + 6 \end{cases}$$

se obtienen las secuencias de coeficientes

$$\begin{cases} \{10, -11, -41, 1, 6\} & \Rightarrow \quad 2 \text{ cambios de signo} \\ \{10, 11, -41, -1, 6\} & \Rightarrow \quad 2 \text{ cambios de signo.} \end{cases}$$

Por tanto, por la regla de los signos de Descartes, el polinomio P puede tener

$$\begin{cases} 2 \text{ raíces positivas o ninguna} \\ 2 \text{ raíces negativas o ninguna.} \end{cases}$$

2. Acotación de las raíces. Aplicando el método de McLaurin se tiene que

$$\lambda = \max\left\{\frac{11}{10}, \frac{41}{10}, \frac{1}{10}, \frac{3}{5}\right\} = \frac{41}{10} \;\Rightarrow\; 1+\lambda = \frac{51}{10}$$

y

$$\mu = \max\left\{\frac{5}{3}, \frac{11}{6}, \frac{41}{6}, \frac{1}{6}\right\} = \frac{41}{6} \;\Rightarrow\; 1+\mu = \frac{47}{6} \;\Rightarrow\; \frac{1}{1+\mu} = \frac{6}{47}$$

por lo que para cada $k \in \{1,2,3,4\}$ se verifica que

$$0.127 \simeq \frac{6}{47} < |\xi_k| < \frac{51}{10} = 5.1.$$

3. Raíces enteras. Los divisores de $a_0 = 6$ son $\{\pm 1, \pm 2, \pm 3, \pm 6\}$. Claramente 6 y -6 no pertenecen al intervalo de acotación y, por tanto, las excluimos. Mediante el algoritmo de Horner se comprueba que el polinomio P no tiene raíces enteras.

4. Raíces racionales. Como los divisores de $a_4 = 10$ son $\{\pm 1, \pm 2, \pm 5, \pm 10\}$, las posibles raíces racionales de P (excluidas $\pm\dfrac{1}{10}$, por no verificar la acotación, y las enteras) expresadas como fracciones irreducibles son

$$\left\{\pm\frac{1}{2}, \pm\frac{1}{5}, \pm\frac{2}{5}, \pm\frac{3}{2}, \pm\frac{3}{5}, \pm\frac{3}{10}, \pm\frac{6}{5}\right\}.$$

Por el algoritmo de Horner se comprueba que $-\dfrac{3}{2}$ es una raíz de P, por lo que factorizamos P en la forma

$$P(x) = 2\left(x+\frac{3}{2}\right)(5x^3 - 13x^2 - x + 2)$$

y continuamos trabajando con las raíces del polinomio

$$Q(x) = 5x^3 - 13x^2 - x + 2.$$

Los candidatos a raíces racionales de Q son

$$\left\{\pm\frac{1}{5}, \pm\frac{2}{5}\right\}.$$

Aplicando nuevamente algoritmo de Horner a Q se comprueba que $-\dfrac{2}{5}$ es raíz de Q, por lo que

$$\begin{aligned} Q(x) &= 5\left(x+\frac{2}{5}\right)(x^2 - 3x + 1) \\ &= 5\left(x+\frac{2}{5}\right)\left(x - \frac{3+\sqrt{5}}{2}\right)\left(x - \frac{3-\sqrt{5}}{2}\right) \end{aligned}$$

y, de esta forma,

$$\begin{aligned} P(x) &= 2\left(x+\frac{3}{2}\right)Q(x) \\ &= 10\left(x+\frac{3}{2}\right)\left(x+\frac{2}{5}\right)\left(x-\frac{3+\sqrt{5}}{2}\right)\left(x-\frac{3-\sqrt{5}}{2}\right). \end{aligned}$$

Así, las raíces del polinomio P son:

$$\begin{cases} \xi_1 = -\dfrac{3}{2} = -1.5, & \xi_2 = -\dfrac{2}{5} = -0.4, \\ \xi_3 = \dfrac{3+\sqrt{5}}{2} \simeq 2.6180, & \xi_4 = \dfrac{3-\sqrt{5}}{2} \simeq 0.3819. \end{cases} \quad \square$$

10.7. Proceso de separación y cálculo de raíces reales de un polinomio

Para calcular, de forma aproximada, las raíces reales de un polinomio $P(x)$ utilizando las técnicas expuestas hasta ahora, deben llevarse a cabo las siguientes tareas:

1. Aplicación de la regla de los signos de Descartes.
2. Acotación de raíces mediante el método de McLaurin.
3. Si los coeficientes del polinomio son racionales, buscar las raíces enteras y racionales aplicando el método de Horner a los candidatos, utilizando la información obtenida en los dos apartados anteriores para evaluar en el menor número posible de puntos.
4. En el caso general, cálculo de la secuencia de Sturm del polinomio $P(x)$. Si $\partial P_m \geq 1$ (caso de raíces múltiples) entonces se comienza todo el proceso aplicado a P_m. Una vez se tengan halladas o aproximadas sus raíces, se divide $P(x)$ entre
$$(x-\eta_1)^{r_1+1}(x-\eta_2)^{r_2+1}\cdots(x-\eta_s)^{r_s+1}$$
donde r_i es la multiplicidad de η_i como raíz de $P_m(x)$ (véase la observación 10.6). El cociente será un polinomio con raíces simples.
5. Separación de las raíces mediante el método de Sturm hasta obtener un intervalo en el que se pueda aplicar alguno de los métodos iterativos de aproximación estudiados en el capítulo 8 (especialmente, el método de Newton).
6. Aplicación, en el intervalo obtenido, del método iterativo en cuestión.

Observación 10.9 (Deflación de polinomios). En el proceso de hallar las raíces de la ecuación algebraica $P(x) = 0$, desde el primer momento en que se conozca una raíz ξ de P debemos, antes de continuar, dividir $P(x)$ entre $x - \xi$ obteniendo

$$P(x) = (x - \xi)Q(x) \text{ con } \partial Q = \partial P - 1 \tag{10.9}$$

y seguir trabajando con la ecuación $Q(x) = 0$. Esto no sólo es importante por la reducción del grado del polinomio, sino también porque, cuando (una vez separadas las raíces) se aplique un método iterativo al polinomio Q, tendremos asegurado que las iteraciones no se desviarán hacia una de las raíces ya encontradas.

Si tan sólo se conoce una aproximación $\widetilde{\xi}$ de la raíz ξ de P entonces, al efectuar la división de $P(x)$ entre $x - \widetilde{\xi}$, se obtiene

$$P(x) \simeq (x - \widetilde{\xi})\widetilde{Q}(x),$$

donde los coeficientes del polinomio $\widetilde{Q}(x)$ no son exactamente los del polinomio $Q(x)$ definido en (10.9), sino una aproximación. En este caso, al trabajar con la ecuación $\widetilde{Q}(x) = 0$ se obtendrá una aproximación $\widehat{\eta}$ de una raíz $\widetilde{\eta}$ de la ecuación $\widetilde{Q}(x) = 0$ que, a su vez, será una aproximación de una raíz η de $Q(x) = 0$ (y, por tanto, también de $P(x) = 0$). Llegados a este punto, en lugar de dividir $\widetilde{Q}(x)$ entre $x - \widehat{\eta}$ y continuar el proceso (con la consiguiente propagación de los errores), lo que se debe hacer es tomar $\widehat{\eta}$ como valor inicial de un método iterativo que resuelva $P(x) = 0$ y *depurarla* así, hasta conseguir un valor más próximo $\breve{\eta}$ a la raíz de $P(x)$; será éste el valor que se tome para llevar a cabo la deflación, dividiendo $\widetilde{Q}(x)$ entre $x - \breve{\eta}$. Este proceso se continúa, depurando sucesivamente las raíces. □

Ejemplo 10.7. Vamos a resolver la ecuación $P(x) = 0$ siendo

$$P(x) = x^4 - 4x^3 - x^2 + 12x - 6.$$

Denotemos por $\{\xi_1, \xi_2, \xi_3, \xi_4\} \subset \mathbb{C}$ las raíces de P.

1. Número de raíces positivas y negativas. A partir de

$$\begin{cases} P(x) = x^4 - 4x^3 - x^2 + 12x - 6 \\ P(-x) = x^4 + 4x^3 - x^2 - 12x - 6 \end{cases}$$

se obtienen las secuencias de coeficientes

$$\begin{cases} \{1, -4, -1, 12, -6\} & \Rightarrow \quad 3 \text{ cambios de signo} \\ \{1, 4, -1, -12, -6\} & \Rightarrow \quad 1 \text{ cambio de signo.} \end{cases}$$

Por tanto, por la regla de los signos de Descartes, el polinomio P tiene

$$\begin{cases} 3 \text{ o } 1 \text{ raíces reales positivas} \\ 1 \text{ raíz real negativa.} \end{cases}$$

2. Acotación de las raíces. Aplicando el método de McLaurin se obtiene:

$$\lambda = \max\{4, 1, 12, 6\} = 12 \;\Rightarrow\; 1 + \lambda = 13$$

y

$$\mu = \max\left\{\frac{1}{6}, \frac{2}{3}, \frac{1}{6}, 2\right\} = 2 \;\Rightarrow\; 1 + \mu = 3 \;\Rightarrow\; \frac{1}{1+\mu} = \frac{1}{3}$$

por lo que para cada $k \in \{1, 2, 3, 4\}$

$$\frac{1}{3} < |\xi_k| < 13.$$

3. Raíces enteras. Los divisores de $a_0 = -6$ son $\{\pm 1, \pm 2, \pm 3, \pm 6\}$. Mediante el método de Horner se comprueba que ninguno de ellos es raíz de P. Por tanto, P no tiene raíces enteras.

4. Raíces racionales. Como $a_4 = 1$, las posibles raíces racionales de P se reducen a las enteras; por tanto, P tampoco tiene raíces racionales.

5. Separación de las raíces reales. Puesto que

$$P'(x) = 4x^3 - 12x^2 - 2x + 12 = 2(2x^3 - 6x^2 - x + 6),$$

podemos tomar

$$P_1(x) = 2x^3 - 6x^2 - x + 6.$$

Como

$$P(x) = P_1(x)\left(\frac{x}{2} - \frac{1}{2}\right) + \left(-\frac{7}{2}x^2 + \frac{17}{2}x - 3\right)$$

tomamos

$$P_2(x) = 7x^2 - 17x + 6$$

y efectuamos la división de $P_1(x)$ entre $P_2(x)$, obteniendo

$$P_1(x) = P_2(x)\left(\frac{2}{7}x - \frac{8}{49}\right) + \left(-\frac{269}{49}x + \frac{342}{49}\right);$$

de esta forma, elegimos

$$P_3(x) = 269x - 342.$$

Dividiendo $P_2(x)$ entre $P_3(x)$ se obtiene

$$P_2(x) = P_3(x)\left(\frac{7}{269}x - \frac{2179}{72361}\right) - \frac{311052}{72361}$$

y podemos considerar

$$P_4(x) = 1.$$

De esta forma, la secuencia de Sturm de P es

$$\{x^4 - 4x^3 - x^2 + 12x - 6, 2x^3 - 6x^2 - x + 6, 7x^2 - 17x + 6, 269x - 342, 1\}.$$

Por tanto, considerando la secuencia de signos obtenemos

x	$P(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$N(x)$
$-\infty$	$+$	$-$	$+$	$-$	$+$	4
0	$-$	$+$	$+$	$-$	$+$	3
$+\infty$	$+$	$+$	$+$	$+$	$+$	0

Así pues, por el teorema de Sturm, se tiene que

$$\begin{cases} N_+ = N(0) - N(+\infty) = 3 \\ N_- = N(-\infty) - N(0) = 1 \end{cases}$$

lo que indica que P tiene 3 raíces positivas y 1 negativa (esto último, como ya se sabía). Aplicando el método de Sturm (y trabajando con intervalos de extremos enteros, para una mayor claridad en la exposición) se obtiene la siguiente tabla:

x	$P(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$N(x)$
-6	$+$	$-$	$+$	$-$	$+$	4
-3	$+$	$-$	$+$	$-$	$+$	4
-2	$+$	$-$	$+$	$-$	$+$	4
-1	$-$	$-$	$+$	$-$	$+$	3
0	$-$	$+$	$+$	$-$	$+$	3
1	$+$	$+$	$-$	$-$	$+$	2
2	$-$	$-$	0	$+$	$+$	1
3	$-$	$+$	$+$	$+$	$+$	1
6	$+$	$+$	$+$	$+$	$+$	0

Por tanto, el polinomio P tiene, exactamente, una raíz en cada intervalo $(-2, -1)$, $(0, 1)$, $(1, 2)$ y $(3, 6)$.

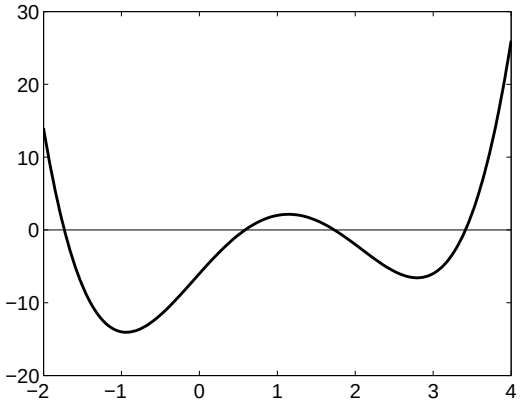


Figura 10.3: Polinomio $x^4 - 4x^3 - x^2 + 12x - 6$.

6. Aproximación de las raíces reales. Claramente,

$$P(x) = x^4 - 4x^3 - x^2 + 12x - 6,$$

$$P'(x) = 4x^3 - 12x^2 - 2x + 12 = 2(2x^3 - 6x^2 - x + 6)$$

y

$$P''(x) = 12x^2 - 24x - 2 = 2(6x^2 - 12x - 1).$$

Como las raíces de la ecuación

$$6x^2 - 12x - 1 = 0$$

son

$$\alpha_1 = 1 - \frac{\sqrt{42}}{6} \simeq -0.0801232 \text{ y } \alpha_2 = 1 + \frac{\sqrt{42}}{6} \simeq 2.08012$$

entonces se verifica que

$$\begin{cases} P''(x) > 0, & x \in (-\infty, \alpha_1) \cup (\alpha_2, +\infty) \\ P''(x) < 0, & x \in (\alpha_1, \alpha_2). \end{cases}$$

Por tanto, la función P' es estrictamente creciente en $(-\infty, \alpha_1) \cup (\alpha_2, +\infty)$ y estrictamente decreciente en (α_1, α_2). Aplicamos el método de Newton para aproximar cada una de las raíces; como

$$x - \frac{P(x)}{P'(x)} = \frac{(x-1)(3x^3 - 5x^2 - 6x - 6)}{2(2x^3 - 6x^2 - x + 6)}$$

entonces consideramos la sucesión

$$\begin{cases} x_0 \in \mathbb{R} \\ x_n = \dfrac{(x_{n-1}-1)(3x_{n-1}^3 - 5x_{n-1}^2 - 6x_{n-1} - 6)}{2(2x_{n-1}^3 - 6x_{n-1}^2 - x_{n-1} + 6)},\ n \in \mathbb{N}. \end{cases} \tag{10.10}$$

a) Intervalo $[-2,-1]$: $\begin{cases} P(-2) > 0 > P(-1) \\ P'(x) \leq P'(-1) < 0, & x \in [-2,-1] \\ P''(x) > 0, & x \in [-2,-1]. \end{cases}$

Tomando $x_0 = -2$ en (10.10) se tiene que

$$\lim_{n\to+\infty} x_n = \xi_1 \simeq -1.732051.$$

b) Intervalo $[0,1]$: $\begin{cases} P(0) < 0 < P(1) \\ P'(x) \geq P'(1) > 0, & x \in [0,1] \\ P''(x) < 0, & x \in [0,1]. \end{cases}$

Tomando $x_0 = 0$ en (10.10) se tiene que

$$\lim_{n\to+\infty} x_n = \xi_2 \simeq 0.585786.$$

c) Intervalo $[1,2]$: aquí la función $P'(x)$ cambia de signo (compruébese) por lo que debemos considerar un intervalo más pequeño como, por ejemplo, $[1.5,2]$. Como en este intervalo

$$\begin{cases} P(1.5) > 0 > P(2) \\ P'(x) \leq P'(1.5) < 0, & x \in [1.5,2] \\ P''(x) < 0, & x \in [1.5,2], \end{cases}$$

tomando $x_0 = 2$ en (10.10) se verifica que

$$\lim_{n\to+\infty} x_n = \xi_3 \simeq 1.732051.$$

d) Intervalo $[3,6]$: $\begin{cases} P(3) < 0 < P(6) \\ P'(x) \geq P'(3) > 0, & x \in [3,6] \\ P''(x) > 0, & x \in [3,6]. \end{cases}$

Tomando $x_0 = 6$ en (10.10) se tiene que

$$\lim_{n\to+\infty} x_n = \xi_4 \simeq 3.414214.$$

7. Observación. Las raíces exactas del polinomio P son:

$$\begin{cases} \xi_1 = -\sqrt{3} = -1.732050807\ldots, & \xi_2 = 2 - \sqrt{2} = 0.585786437\ldots, \\ \xi_3 = \sqrt{3} = 1.732050807\ldots, & \xi_4 = 2 + \sqrt{2} = 3.414213562\ldots \end{cases} \quad \square$$

10.8. Raíces complejas: método de Bairstow

El método de Sturm sirve para separar y, con la ayuda de los métodos del capítulo 8, calcular las raíces reales de un polinomio; nos proponemos aproximar ahora las eventuales raíces complejas de una ecuación algebraica. Para ello, emplearemos el *método de Bairstow*, que sirve para aproximar los coeficientes u y v de un trinomio de la forma

$$x^2 - ux - v.$$

La idea de partida es que si $\alpha \pm i\beta \in \mathbb{C}$ son raíces de un polinomio P entonces éste es divisible por

$$(x - (\alpha + i\beta))(x - (\alpha - i\beta)) = (x - \alpha)^2 + \beta^2 = x^2 - 2\alpha x + (\alpha^2 + \beta^2).$$

Por tanto, una vez obtenidos, de forma aproximada, los valores u y v, a partir del sistema

$$\begin{cases} 2\alpha = u \\ \alpha^2 + \beta^2 = -v \end{cases}$$

se obtienen las raíces complejas

$$\boxed{\alpha \pm i\beta = \frac{u}{2} \pm i\frac{\sqrt{-u^2 - 4v}}{2}} \tag{10.11}$$

Con vistas a obtener los valores de u y v efectuamos la división de $P(x)$ entre $x^2 - ux - v$. Esto dará lugar a un cociente $Q(x)$ y a un resto $R(x)$ que dependerán de u y v. Claramente

$$P(x) = (x^2 - ux - v)Q(x) + R(x) \ \text{ con } \ \partial Q = n - 2 \ \text{ y } \ \partial R \leq 1 \tag{10.12}$$

donde

$$Q(x) = b_n x^{n-2} + b_{n-1}x^{n-3} + \cdots + b_3 x + b_2$$

y

$$R(x) = b_1(x - u) + b_0.$$

Insistimos en que los coeficientes $\{b_n, b_{n-1}, \ldots, b_0\}$ dependen de u y v.

Si u y v fueran los valores correspondientes a las raíces exactas $\alpha \pm i\beta$ se tendría que $b_0 = b_1 = 0$. La idea es, por tanto, encontrar valores de u y v solución de

$$(\mathcal{S}) \begin{cases} b_0(u,v) = 0 \\ b_1(u,v) = 0. \end{cases}$$

El sistema $(\mathcal{S})$, que es no lineal y consta de dos ecuaciones con dos incógnitas, lo resolveremos mediante el método de Newton a partir de dos valores iniciales u^0 y v^0. Como veremos, en este caso particular, la aplicación del método de Newton puede hacerse de una forma especialmente sencilla. Comenzamos probando el siguiente resultado:

Proposición 10.4.

a) Los coeficientes de los polinomios $Q(x)$ y $R(x)$ vienen dados por la siguiente ley de recurrencia:

$$\begin{cases} b_n = a_n \\ b_{n-1} = a_{n-1} + ub_n \\ b_k = a_k + ub_{k+1} + vb_{k+2},\ k = n-2, n-3, \ldots, 0. \end{cases} \tag{10.13}$$

b) Denotando por

$$\begin{cases} c_k = \dfrac{\partial b_k}{\partial u}, & k = 0, 1, \ldots, n \\ d_k = \dfrac{\partial b_{k-1}}{\partial v}, & k = 1, 2, \ldots, n \end{cases}$$

se verifica que

$$\begin{cases} c_n = 0 \\ c_{n-1} = b_n \\ c_k = b_{k+1} + uc_{k+1} + vc_{k+2},\ k = n-2, n-3, \ldots, 0 \end{cases}$$

y

$$\begin{cases} d_n = 0 \\ d_{n-1} = b_n \\ d_k = b_{k+1} + ud_{k+1} + vd_{k+2},\ k = n-2, n-3, \ldots, 1. \end{cases}$$

En consecuencia, para todo $k \in \{1, 2, \ldots, n\}$, se tiene que

$$c_k = d_k.$$

DEMOSTRACIÓN.

a) La relación (10.12) determina que

$$\begin{aligned}\sum_{k=0}^{n} a_k x^k &= (x^2 - ux - v)\left(\sum_{k=2}^{n} b_k x^{k-2}\right) + b_1(x-u) + b_0 \\ &= \sum_{k=2}^{n} b_k x^k - u\sum_{k=2}^{n} b_k x^{k-1} - v\sum_{k=2}^{n} b_k x^{k-2} + b_1 x + b_0 - ub_1 \\ &= \sum_{k=2}^{n} b_k x^k - u\sum_{k=1}^{n-1} b_{k+1} x^k - v\sum_{k=0}^{n-2} b_{k+2} x^k + b_1 x + b_0 - ub_1 \\ &= (b_0 - ub_1 - vb_2) + (b_1 - ub_2 - vb_3)x \\ &\quad + \sum_{k=2}^{n-2}(b_k - ub_{k+1} - vb_{k+2})x^k + (b_{n-1} - ub_n)x^{n-1} + b_n x^n \\ &= \sum_{k=0}^{n-2}(b_k - ub_{k+1} - vb_{k+2})x^k + (b_{n-1} - ub_n)x^{n-1} + b_n x^n\end{aligned}$$

de donde, identificando coeficientes, se obtiene (10.13).

b) A partir de (10.13) se verifica:

$$c_n = \frac{\partial b_n}{\partial u} = \frac{\partial a_n}{\partial u} = 0,$$

$$c_{n-1} = \frac{\partial b_{n-1}}{\partial u} = \frac{\partial a_{n-1}}{\partial u} + \frac{\partial (ub_n)}{\partial u} = 0 + b_n + u\frac{\partial b_n}{\partial u} = b_n$$

y

$$\begin{aligned}c_k = \frac{\partial b_k}{\partial u} &= \frac{\partial a_k}{\partial u} + \frac{\partial (ub_{k+1})}{\partial u} + \frac{\partial (vb_{k+2})}{\partial u} \\ &= 0 + b_{k+1} + u\frac{\partial b_{k+1}}{\partial u} + v\frac{\partial b_{k+2}}{\partial u} = b_{k+1} + uc_{k+1} + vc_{k+2}\end{aligned}$$

para $k = n-2, n-3, \ldots, 0$. Por otra parte,

$$d_n = \frac{\partial b_{n-1}}{\partial v} = \frac{\partial a_{n-1}}{\partial v} + \frac{\partial (ub_n)}{\partial v} = 0 + u\frac{\partial b_n}{\partial v} + u\frac{\partial b_n}{\partial u} = 0,$$

$$\begin{aligned}d_{n-1} = \frac{\partial b_{n-2}}{\partial v} &= \frac{\partial a_{n-2}}{\partial v} + \frac{\partial (ub_{n-1})}{\partial v} + \frac{\partial (vb_n)}{\partial v} \\ &= 0 + u\frac{\partial b_{n-1}}{\partial v} + b_n + v\frac{\partial b_n}{\partial v} = b_n\end{aligned}$$

y

$$\begin{aligned} d_k &= \frac{\partial b_{k-1}}{\partial v} = \frac{\partial a_{k-1}}{\partial v} + \frac{\partial (ub_k)}{\partial v} + \frac{\partial (vb_{k+1})}{\partial v} \\ &= 0 + u\frac{\partial b_k}{\partial v} + b_{k+1} + v\frac{\partial b_{k+1}}{\partial v} = b_{k+1} + ud_{k+1} + vd_{k+2} \end{aligned}$$

para $k = n-2, n-3, \ldots, 1$. □

Con la notación y los resultados obtenidos en la proposición 10.4 podemos escribir la matriz jacobiana

$$J_{(b_0,b_1)}(u,v) = \begin{pmatrix} \dfrac{\partial b_0}{\partial u}(u,v) & \dfrac{\partial b_0}{\partial v}(u,v) \\ \dfrac{\partial b_1}{\partial u}(u,v) & \dfrac{\partial b_1}{\partial v}(u,v) \end{pmatrix}$$

en la forma

$$J_{(b_0,b_1)}(u,v) = \begin{pmatrix} c_0(u,v) & d_1(u,v) \\ c_1(u,v) & d_2(u,v) \end{pmatrix} = \begin{pmatrix} c_0(u,v) & c_1(u,v) \\ c_1(u,v) & c_2(u,v) \end{pmatrix}.$$

Cuando esta matriz es no singular, podemos aplicar el método de Newton

$$\begin{cases} \begin{pmatrix} u^0 \\ v^0 \end{pmatrix} \in \mathbb{R}^2 \text{ dado} \\ \begin{pmatrix} u^{p+1} \\ v^{p+1} \end{pmatrix} = \begin{pmatrix} u^p \\ v^p \end{pmatrix} - \left(J_{(b_0,b_1)}(u^p, v^p)\right)^{-1} \begin{pmatrix} b_0(u^p, v^p) \\ b_1(u^p, v^p) \end{pmatrix} \end{cases}$$

es decir,

$$\begin{cases} \begin{pmatrix} u^0 \\ v^0 \end{pmatrix} \in \mathbb{R}^2 \text{ dado} \\ \begin{pmatrix} u^{p+1} \\ v^{p+1} \end{pmatrix} = \begin{pmatrix} u^p \\ v^p \end{pmatrix} - \dfrac{1}{c_0c_2 - c_1^2} \begin{pmatrix} c_2 & -c_1 \\ -c_1 & c_0 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} u^p + \dfrac{c_1b_1 - c_2b_0}{c_0c_2 - c_1^2} \\ v^p + \dfrac{c_1b_0 - c_0b_1}{c_0c_2 - c_1^2} \end{pmatrix}. \end{cases}$$

Así pues, para calcular una iteración a partir de la iteración anterior basta conocer los valores de b_0, b_1, c_0, c_1 y c_2. Éstos se obtienen, de forma recursiva, mediante las iteraciones definidas en la proposición 10.4. En concreto, a partir del valor $(u^p, v^p)^{\mathrm{T}}$ se calculan los valores

$$\boxed{b_k = a_k + u^p b_{k+1} + v^p b_{k+2}}$$

para $k = n, n-1, \ldots, 0$ ($b_{n+2} = b_{n+1} = 0$) y, una vez calculados éstos, los valores

$$\boxed{c_k = b_{k+1} + u^p c_{k+1} + v^p c_{k+2}}$$

para $k = n-1, n-2, \ldots, 0$ ($c_{n+1} = c_n = 0$). De esta forma se obtienen, en particular, los números b_0, b_1, c_0, c_1 y c_2 y, con ellos, el vector $(u^{p+1}, v^{p+1})^{\mathrm{T}}$. Cuando se verifique el test de parada con la precisión requerida se toman como valores de u y v los de u^p y v^p y, a partir de ellos, se calculan las raíces complejas mediante la fórmula (10.11).

Ejemplo 10.8. Vamos a utilizar el método de Bairstow para aproximar las raíces de la ecuación

$$x^3 + x + 1 = 0.$$

Comenzando en los valores iniciales $u = v = 0$ y trabajando con un test de parada de tolerancia 10^{-5}, se obtienen las siguientes iteraciones del método

n	u	v
1	-1	-1
2	0.5	1
3	0.36364	-1.11364
4	0.79635	-1.44693
5	0.67714	-1.44430
6	0.68241	-1.46566
7	0.68233	-1.46557

siendo, por tanto, las raíces complejas aproximadas

$$\xi_1 = 0.34116 + 1.16154i \;\text{ y }\; \xi_2 = 0.34116 - 1.16154i.$$

Si ahora se divide el polinomio de partida por $x^2 - ux - v$ con los últimos valores de u y v obtenidos, se obtiene una aproximación de la tercera raíz

$$\xi_3 = -0.68233. \quad \square$$

Observación 10.10. El método de Bairstow sirve, en realidad, para calcular las raíces de un polinomio por pares. Si se aplica sin haber eliminado previamente todas las raíces reales, puede ocurrir que el trinomio $x^2 - ux - v$ tenga dos raíces reales (no puede tener una real y una compleja puesto que u y v son siempre reales). Esta situación aparece en el ejemplo 10.9. $\square$

Ejemplo 10.9. Consideremos el polinomio

$$P(x) = x^5 - 9x^4 + 37x^3 - 103x^2 + 144x - 70$$

cuyas raíces son $1 \pm 3i$, 1 y $3 \pm \sqrt{2}$.

Si utilizamos el método de Bairstow, partiendo de los valores $u = v = 2$ y con una tolerancia de 10^{-5}, se obtienen los valores

$$u = 2.58579 \ \text{ y } \ v = -1.58579,$$

que dan lugar a las raíces

$$\xi_1 = 1.58579 \ \text{ y } \ \xi_2 = 1,$$

aproximaciones de $3 - \sqrt{2}$ y 1, respectivamente. Nótese que ambas son reales. Dividiendo $P(x)$ entre $x^2 - ux - v$, con los valores de u y v anteriores, obtenemos el polinomio

$$Q(x) = x^3 - 6.41421x^2 + 18.82841x - 44.14210.$$

Aplicando otra vez el método de Bairstow, a partir de los mismos valores iniciales, se obtienen las aproximaciones de las raíces

$$\xi_3 = 0.99999 + 3i \ \text{ y } \ \xi_4 = 0.99999 - 3i.$$

Finalmente, dividiendo por el trinomio correspondiente, se obtiene la aproximación de la última raíz

$$\xi_5 = 4.41421. \quad \square$$

10.9. Problemas

10.9.1. Problemas resueltos

10.1. Método de Laguerre para la acotación de raíces. Sea $L > 0$ de forma que tanto los coeficientes del cociente como el resto de dividir el polinomio

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{R}[x] \ \text{ con } \ a_n \neq 0$$

entre $x - L$ son no negativos (incluso nulos). Demostrar que entonces L es una cota superior de las raíces de la ecuación $P(x) = 0$. Como aplicación, acotar superiormente las raíces reales de la ecuación

$$2x^4 + 4x^3 - 59x^2 - 61x + 30 = 0.$$

Solución. Por hipótesis,

$$P(x) = (x - L)Q(x) + P(L)$$

donde $P(L) \geq 0$ y

$$Q(x) = c_{n-1}x^{n-1} + c_{n-2}x^{n-2} + \cdots + c_1x + c_0 \ \text{ con } \ c_i \geq 0$$

para $i = 0, 1, \ldots, n-1$. Como $\partial P = n$, entonces $\partial Q = n-1$ y, por tanto,

$$c_{n-1} > 0.$$

Consecuentemente,

$$P(x) > 0 \ \text{ si } \ x > L,$$

lo que hace que ningún valor $x > L$ sea raíz de P.

Para acotar superiormente las raíces del polinomio

$$P(x) = 2x^4 + 4x^3 - 59x^2 - 61x + 30$$

aplicamos la regla de Ruffini con valores crecientes de L:

	2	4	−59	−61	30
3		6	30	−87	−444
	2	10	−29	−148	−414

	2	4	−59	−61	30
4		8	48	−44	−420
	2	12	−11	−105	−390

	2	4	−59	−61	30
5		10	70	55	−30
	2	14	11	−6	0

	2	4	−59	−61	30
6		12	96	222	966
	2	16	37	161	996

Por tanto, $L = 6$ es una cota superior de las raíces de P. □

10.2. Método de Newton para la acotación de raíces. Consideremos el polinomio

$$P(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in \mathbb{R}[x].$$

Si $a_n > 0$ y existe $L > 0$ verificando

$$P^{k)}(L) \geq 0 \tag{10.14}$$

para $k = 0, 1, \ldots, n-1$, demostrar que L es una cota superior de las raíces de la ecuación $P(x) = 0$. Como aplicación, hallar una cota superior de las raíces de

$$3x^4 - 18x^3 + 24x^2 - 18x + 73 = 0.$$

SOLUCIÓN. Veamos que si $\xi > L$ entonces $P(\xi) > 0$ y, por tanto, ξ no es raíz de P. En efecto, desarrollando por Taylor el polinomio P, teniendo en cuenta la relación (10.14) y que $\partial P = n$, para $x \geq L$ se tiene que

$$\begin{aligned} P(x) &= \sum_{k=0}^{n-1} \frac{P^{k)}(L)}{k!}(x-L)^k + \frac{P^{n)}(L)}{n!}(x-L)^n + \sum_{k=n+1}^{\infty} \frac{P^{k)}(L)}{k!}(x-L)^k \\ &\geq \frac{P^{n)}(L)}{n!}(x-L)^n = \frac{n!a_n}{n!}(x-L)^n = a_n(x-L)^n. \end{aligned}$$

Al particularizar en $x = \xi$ se concluye que

$$P(\xi) \geq a_n(\xi - L)^n > 0.$$

Como aplicación, el polinomio P y sus sucesivas derivadas son:

$$\begin{cases} P(x) = 3x^4 - 18x^3 + 24x^2 - 18x + 73 \\ P'(x) = 12x^3 - 54x^2 + 48x - 18 = 6(2x^3 - 9x^2 + 8x - 3) \\ P''(x) = 6(6x^2 - 18x + 8) = 12(3x^2 - 9x + 4) \\ P'''(x) = 12(6x - 9) = 36(2x - 3). \end{cases}$$

Vamos tomando valores crecientes de L:

L	$P(L)$	$P'(L)$	$P''(L)$	$P'''(L)$
2	73	-18		
3	-737			
4	1	78	192	180

Consecuentemente, $L = 4$ es una cota superior de las raíces de P. □

10.3. Calcular las raíces de la ecuación $x^3 - x^2 + 3x = 3$.

SOLUCIÓN. Denotemos por

$$P(x) = x^3 - x^2 + 3x - 3.$$

Por la regla de los signos de Descartes, las secuencias de signos son

$$\begin{cases} P(x) \longrightarrow \{1, -1, 3, -3\} \Rightarrow 3 \text{ cambios de signo} \\ P(-x) \longrightarrow \{-1, -1, -1, -3\} \Rightarrow 0 \text{ cambios de signo} \end{cases}$$

por lo que el polinomio P no tiene raíces negativas y puede tener

$$\begin{cases} 3 \text{ raíces positivas} \\ 1 \text{ raíz positiva.} \end{cases}$$

Los candidatos a raíces enteras del polinomio P se encuentran entre los divisores del término independiente -3. Es decir, de existir raíces enteras, éstas deben encontrarse entre los números $\{1, 3\}$. Se comprueba que $\xi_1 = 1$ es la única raíz entera de P, por lo que podemos factorizar el polinomio en la forma

$$P(x) = x^3 - x^2 + 3x - 3 = (x-1)(x^2+3) = (x-1)\left(x - \sqrt{3}i\right)\left(x + \sqrt{3}i\right).$$

Así pues, las tres raíces de P son $\xi_1 = 1$, $\xi_2 = \sqrt{3}i$ y $\xi_3 = \overline{\xi}_2 = -\sqrt{3}i$. □

10.4. Hallar las raíces reales, determinando su multiplicidad, de la ecuación

$$x^{13} - x^{11} + x^2 - 1 = 0.$$

SOLUCIÓN. En primer lugar recordemos que todo número complejo $z \in \mathbb{C}$ de *módulo* $r > 0$ y *argumento* $\vartheta \in [0, 2\pi)$ escrito en la forma

$$z = re^{i\vartheta} = r(\cos\vartheta + i\,\mathrm{sen}\,\vartheta)$$

tiene n raíces n–ésimas de la forma

$$z_k = \sqrt[n]{r}e^{i\frac{\vartheta+2k\pi}{n}}$$

para $k = 0, 1, \ldots, n-1$. Como se observa, las n raíces están situadas en la circunferencia de centro 0 y radio $\sqrt[n]{r}$ en los vértices de un polígono regular (de una raíz z_k se obtiene la siguiente z_{k+1} incrementando el argumento de z_k en $\frac{2\pi}{n}$).

Apliquemos lo anterior para calcular todas las raíces del polinomio

$$P(x) = x^{13} - x^{11} + x^2 - 1.$$

Los candidatos a raíces enteras de P son los divisores del término independiente -1, es decir, $\{-1, 1\}$. Se comprueba que $\xi_5 = -1$ y $\xi_{11} = 1$ son raíces de P, por lo que podemos factorizar éste en la forma

$$P(x) = (x-1)(x+1)(x^{11}+1).$$

Por otra parte, como

$$x^{11} + 1 = 0 \;\Rightarrow\; x^{11} = -1 = e^{i\pi} \;\Rightarrow\; x = \sqrt[11]{e^{i\pi}},$$

se verifica que

$$\xi_k = e^{i\frac{(2k+1)\pi}{11}} = \cos\frac{(2k+1)\pi}{11} + i\,\mathrm{sen}\,\frac{(2k+1)\pi}{11}$$

para $k = 0, 1, \ldots, 10$, son las raíces undécimas del número -1 (véase la figura 10.4). Nótese que $\xi_5 = -1$ es una raíz de P con multiplicidad dos.

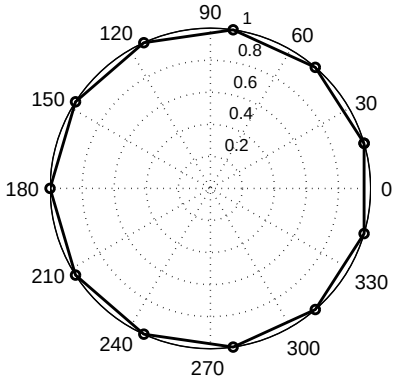


Figura 10.4: Distribución de las raíces undécimas de -1.

De esta forma, las 13 raíces del polinomio P son $\xi_5 = -1$ (doble), $\xi_{11} = 1$ y

$$\xi_k = \cos\frac{(2k+1)\pi}{11} + i\,\mathrm{sen}\,\frac{(2k+1)\pi}{11}$$

para $k = 0, 1, 2, 3, 4, 6, 7, 8, 9, 10$. □

10.5. Si $\{P, P_1, \ldots, P_m\}$ es la secuencia de Sturm del polinomio $P(x)$ probar que

$$P_m(x) = \mathrm{MCD}\{P(x), P'(x)\}.$$

Solución. Por construcción, como $P_m(x)|P_{m-1}(x)$ y

$$P_{m-2}(x) = P_{m-1}(x)Q_{m-1}(x) - P_m(x)$$

entonces $P_m(x)|P_{m-2}(x)$. Reiterando este argumento, se llega a que $P_m(x)|P_1(x)$ y $P_m(x)|P(x)$.

Por otra parte, si un polinomio $Q(x) \in \mathbb{R}[x]$ es tal que $Q(x)|P(x)$ y $Q(x)|P'(x)$, por construcción, $Q(x)|P_2(x)$. Nuevamente, reiterando este argumento, se concluye que $Q(x)|P_m(x)$, obteniéndose así el resultado. □

10.6. Demostrar el teorema de Sturm en el caso de que el polinomio P tenga raíces reales múltiples.

Solución. Con la notación empleada en la demostración del teorema 10.4, puesto que la construcción de la secuencia de Sturm no es más que el algoritmo de Euclides

para hallar el máximo común divisor de P y P', se tiene que P_m divide a todos los polinomios P_i, $i = 0, 1, \ldots, m$ y, por tanto, podemos considerar los polinomios

$$q_i(x) = \frac{P_i(x)}{P_m(x)}$$

para $i = 0, 1, \ldots, m$. Nótese que si $x = \alpha$ no es raíz de P_0 entonces $P_m(\alpha) \neq 0$ y, así, el número de cambios de signo $N(\alpha)$ de la secuencia $\{P_0(\alpha), P_1(\alpha), \ldots, P_m(\alpha)\}$ coincide con el número de cambios de signo $M(\alpha)$ de $\{q_0(\alpha), q_1(\alpha), \ldots, q_m(\alpha)\}$. Como los polinomios P_0 y q_0 tienen las mismas raíces reales (sin tener en cuenta la multiplicidad) bastará probar que si $q_0(a)q_0(b) \neq 0$ el número de raíces reales del polinomio q_0 en el intervalo (a, b) es $M(a) - M(b)$. Queremos, pues, probar el resultado del teorema de Sturm para la secuencia $\{q_0, q_1, \ldots, q_m\}$ que, aunque no es una secuencia de Sturm para q_0 (pues $q_1 \neq q_0'$), tiene propiedades parecidas. De hecho, intentaremos adaptar en lo posible la demostración del teorema de Sturm que se hizo para el caso de raíces reales simples.

a) Comenzamos estudiando el comportamiento de $M(x)$ al pasar por una raíz $x = \alpha$ de q_0. Si el polinomio q_1 fuera la derivada de q_0 podríamos argumentar como en el caso de raíces simples; en realidad nos basta con que q_1 tenga, en un entorno de α, el mismo signo que q_0'. Esta propiedad se cumple puesto que, como

$$q_0'(x) = \frac{P_0'(x)P_m(x) - P_0(x)P_m'(x)}{(P_m(x))^2} \quad \text{y} \quad P_0(\alpha) = 0$$

entonces

$$q_0'(\alpha) = \frac{P_0'(\alpha)}{P_m(\alpha)}$$

y, por tanto,

$$q_1(\alpha)q_0'(\alpha) = \left(\frac{P_0'(\alpha)}{P_m(\alpha)}\right)^2 > 0$$

ya que si α fuera también raíz de P_0' lo sería de P_m, con la misma multiplicidad, y el cociente $\dfrac{P_0'(\alpha)}{P_m(\alpha)}$ sería no nulo. Así pues, la función $M(x)$ disminuye en una unidad al pasar por una raíz $x = \alpha$ de q_0.

b) Como la secuencia $\{q_1, q_2, \ldots, q_m\}$ verifica la ley de recurrencia

$$\begin{cases} q_{i-2}(x) = q_{i-1}(x)Q_{i-1}(x) - q_i(x), & i = 2, 3, \ldots, m \\ q_m(x) = 1 & (q_{m+1}(x) \equiv 0) \end{cases}$$

(véase (10.7)) argumentando como se hizo en el caso de raíces simples, se obtiene que la función $M(x)$ no cambia su valor al pasar por una raíz $x = \alpha$

de q_i para $i \in \{1, 2, \ldots, m\}$ salvo que α sea también raíz de q_0 en cuyo caso disminuirá en una unidad.

Por tanto, al pasar de a a b la función $M(x)$ disminuye en tantas unidades como raíces reales simples tiene $q_0(x)$ en el intervalo (a, b). De esta forma, $M(a) - M(b)$ determina el número de raíces reales de q_0 en el intervalo (a, b). □

10.7. Dados dos números, $a > 0$ y $b \in \mathbb{R}$, se considera el polinomio

$$P(x) = x^3 - bx^2 + ax - ab.$$

a) Encontrar una relación entre a y b que garantice que la secuencia de Sturm de P tenga sólo tres términos $\{P_0(x), P_1(x), P_2(x)\}$.

b) Decidir, en el caso en que a y b verifiquen la relación anterior, el número de raíces reales y distintas de P. ¿Son simples?

SOLUCIÓN. Dividiendo $P(x)$ entre $P_1(x) = P'(x)$ obtenemos

$$P(x) = \left(\frac{x}{3} - \frac{b}{9}\right) P_1(x) + \frac{2}{9}((3a - b^2)x - 4ab).$$

a) Basta tomar $b = \pm\sqrt{3a}$ para que $b^2 = 3a > 0$ y, por tanto, la secuencia de Sturm se reduzca a los términos

$$\left\{x^3 - bx^2 + ax - ab, 3x^2 - 2bx + a, \frac{8ab}{9}\right\}.$$

b) Distinguimos los dos posibles casos que pueden darse:

i) $\boxed{b = \sqrt{3a}}$ Como

x	$P(x)$	$P_1(x)$	$P_2(x)$	$N(x)$
$-\infty$	$-$	$+$	$+$	1
0	$-$	$+$	$+$	1
$+\infty$	$+$	$+$	$+$	0

entonces

$$\begin{cases} N_- = N(-\infty) - N(0) = 0 \\ N_+ = N(0) - N(+\infty) = 1 \end{cases}$$

por lo que P tiene una raíz positiva y dos complejas conjugadas (dado que $\partial P = 3$ y las raíces de P son simples).

ii) $\boxed{b=-\sqrt{3a}}$ Como ahora

x	$P(x)$	$P_1(x)$	$P_2(x)$	$N(x)$
$-\infty$	$-$	$+$	$-$	2
0	$+$	$+$	$-$	1
$+\infty$	$+$	$+$	$-$	1

entonces

$$\begin{cases} N_- = N(-\infty) - N(0) = 1 \\ N_+ = N(0) - N(+\infty) = 0 \end{cases}$$

por lo que P tiene una raíz negativa y dos complejas conjugadas (dado que $\partial P = 3$ y las raíces de P son simples).

Observación:

1. En principio podría ocurrir que la secuencia de Sturm tuviera sólo tres términos en el caso de que P tuviera raíces múltiples. Veamos si esto es posible. Suponiendo que
$$b^2 \neq 3a \tag{10.15}$$
consideraríamos
$$P_2(x) = \left(b^2 - 3a\right) x + 4ab$$
y efectuaríamos la división entre $P_1(x)$ y $P_2(x)$ para luego escoger una relación entre a y b (si es que existe) que haga cero el resto de esa división. Como puede comprobarse (se deja como ejercicio al lector)
$$P_1(x) = \left(\frac{3}{b^2 - 3a} x - \frac{2\left(3a + b^2\right)}{\left(b^2 - 3a\right)^2} b \right) P_2(x) + R(a,b)$$
siendo el resto de la división anterior
$$R(a,b) = -\frac{9a\left(a + b^2\right)^2}{\left(3a - b^2\right)^2}$$
(recuérdese que estamos bajo la hipótesis (10.15)). Por tanto,
$$R(a,b) = 0 \;\Leftrightarrow\; 9a\left(a + b^2\right)^2 \;\Leftrightarrow\; a = -b^2 < 0.$$
Luego $R(a,b)$ no se anula nunca y, por tanto, no puede darse esta segunda opción.

2. Como fácilmente se comprueba, se puede factorizar P en la forma

$$P(x) = (x - b)(x^2 + a)$$

por lo que las tres raíces de P son $\xi_1 = b$, $\xi_2 = \sqrt{a}i$ y $\xi_3 = \overline{\xi}_2 = -\sqrt{a}i$. □

10.8. Consideremos el polinomio

$$P(x) = 9x^3 + 9x^2 + 9\lambda x + \lambda$$

donde $\lambda \in \mathbb{R}$.

a) Estudiar, en función del parámetro λ, el número de raíces (reales y complejas) de la ecuación $P(x) = 0$. ¿Para qué valores de λ las raíces de P son múltiples? Hallar todas las raíces de P para esos valores de λ.

b) Fijado $\lambda = \sqrt{3}$ encontrar un intervalo donde pueda aplicarse el método de Newton para calcular una raíz negativa de P. Determinar los primeros términos de la sucesión definida por dicho método.

Solución.

a) Hallemos la secuencia de Sturm del polinomio P. Como

$$P'(x) = 27x^2 + 18x + 9\lambda = 9(3x^2 + 2x + \lambda),$$

tomamos

$$P_1(x) = 3x^2 + 2x + \lambda.$$

Al dividir $P(x)$ entre $P_1(x)$ se obtiene

$$P(x) = (3x + 1)P_1(x) + 2(3\lambda - 1)x,$$

por lo que elegimos

$$P_2(x) = (1 - 3\lambda)x \text{ si } \lambda \neq \frac{1}{3}.$$

En esta situación, como al efectuar la división de $P_1(x)$ entre $P_2(x)$ resulta

$$P_1(x) = \frac{3x + 2}{1 - 3\lambda}P_2(x) + \lambda,$$

distinguimos los diversos casos que pueden presentarse en función de λ:

i) Si $\lambda = \dfrac{1}{3}$ la secuencia de Sturm de P se reduce a $\{P(x), P_1(x)\}$. Ahora bien, como

$$P_1(x) = 3x^2 + 2x + \frac{1}{3} = 3\left(x + \frac{1}{3}\right)^2,$$

se tiene que $\xi = -\dfrac{1}{3}$ es raíz triple del polinomio

$$P(x) = 9x^3 + 9x^2 + 3x + \frac{1}{3} = 9\left(x + \frac{1}{3}\right)^3.$$

ii) Si $\lambda = 0$ la secuencia de Sturm del polinomio P es $\{P(x), P_1(x), P_2(x)\}$. En este caso, como $P_2(x) = x$ entonces $\xi_1 = 0$ es raíz doble de P. Por tanto,

$$P(x) = 9x^3 + 9x^2 = 9x^2(x+1)$$

de donde se deduce que $\xi_2 = -1$ es raíz simple de P.

iii) Para valores $\lambda \notin \left\{0, \dfrac{1}{3}\right\}$ la secuencia de Sturm del polinomio P es $\{P(x), P_1(x), P_2(x), P_3(x)\}$ siendo

$$P_3(x) = -\lambda,$$

por lo que todas las raíces del polinomio $P(x)$ son simples. Distinguimos, a su vez, las tres posibilidades que pueden presentarse:

α) Si $\lambda < 0$ obtenemos los siguientes valores

x	$P(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$N(x)$
$-\infty$	$-$	$+$	$-$	$+$	3
0	$-$	$-$	0	$+$	1
$+\infty$	$+$	$+$	$+$	$+$	0

de donde

$$\begin{cases} N_- = N(-\infty) - N(0) = 2 \\ N_+ = N(0) - N(+\infty) = 1 \end{cases}$$

por lo que P tiene dos raíces negativas y una positiva.

β) Si $0 < \lambda < \dfrac{1}{3}$ se obtiene

x	$P(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$N(x)$
$-\infty$	$-$	$+$	$-$	$-$	2
0	$+$	$+$	0	$-$	1
$+\infty$	$+$	$+$	$+$	$-$	1

de donde

$$\begin{cases} N_- = N(-\infty) - N(0) = 1 \\ N_+ = N(0) - N(+\infty) = 0 \end{cases}$$

por lo que P tiene una raíz negativa y las otras dos son complejas conjugadas (por ser $\partial P = 3$ y tener P las raíces simples).

γ) Finalmente, si $\lambda > \dfrac{1}{3}$ entonces

x	$P(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$N(x)$
$-\infty$	$-$	$+$	$+$	$-$	2
0	$+$	$+$	0	$-$	1
$+\infty$	$+$	$+$	$-$	$-$	1

por lo que, nuevamente

$$\begin{cases} N_- = N(-\infty) - N(0) = 1 \\ N_+ = N(0) - N(+\infty) = 0 \end{cases}$$

y, por tanto, P tiene una raíz negativa y dos complejas conjugadas.

b) Cuando $\lambda = \sqrt{3}$ se sabe, por el apartado anterior, que el polinomio

$$P(x) = 9x^3 + 9x^2 + 9\sqrt{3}x + \sqrt{3}$$

tiene una raíz negativa y dos complejas conjugadas. Busquemos un intervalo donde podamos aplicar el método de Newton para aproximar la raíz negativa ξ de $P(x) = 0$. Claramente,

$$\begin{cases} P'(x) = 9(3x^2 + 2x + \sqrt{3}) > 0, \ x \in \mathbb{R} \\ \\ P''(x) = 18(3x+1) \begin{cases} < 0, \ x < -\dfrac{1}{3} \\ = 0, \ x = -\dfrac{1}{3} \\ > 0, \ x > -\dfrac{1}{3} \end{cases} \end{cases}$$

De esta forma, como

$$P\left(-\frac{1}{6}\right) < 0 < P(0)$$

y

$$P'(x) > 0 \ \text{ y } \ P''(x) > 0, \ x \in \left[-\frac{1}{6}, 0\right]$$

la sucesión del método de Newton

$$x_n = x_{n-1} - \frac{9x_{n-1}^3 + 9x_{n-1}^2 + 9\sqrt{3}x_{n-1} + \sqrt{3}}{9(3x_{n-1}^2 + 2x_{n-1} + \sqrt{3})},\ n \in \mathbb{N}$$

comenzando en $x_0 = 0$ converge a la raíz negativa de P. Los primeros términos de la sucesión anterior vienen dados en la siguiente tabla:

n	x_n
0	0
1	-0.11111111111111
2	-0.11820541635719
3	-0.11822700626632
4	-0.11822700646197

10.9. Calcular las raíces del polinomio

$$P(x) = 2x^5 - \pi x^4 - 8x^3 + 4\pi x^2 + 8x - 4\pi.$$

SOLUCIÓN. Hallemos la secuencia de Sturm del polinomio P:

i) $P(x) = 2x^5 - \pi x^4 - 8x^3 + 4\pi x^2 + 8x - 4\pi$.

ii) $P_1(x) = \dfrac{P'(x)}{2} = 5x^4 - 2\pi x^3 - 12x^2 + 4\pi x + 4$.

iii) Al efectuar la división de $P(x)$ entre $P_1(x)$ se obtiene (compruébese como ejercicio) que

$$P(x) = \left(\frac{2x}{5} - \frac{\pi}{25}\right) P_1(x) + \frac{2}{25}\left(-(\pi^2+40)x^3 + 24\pi x^2 + 2(\pi^2+40)x - 48\pi\right)$$

por lo que podemos tomar

$$P_2(x) = (\pi^2+40)x^3 - 24\pi x^2 - 2(\pi^2+40)x + 48\pi.$$

iv) Al dividir $P_1(x)$ entre $P_2(x)$ se llega a (compruébese también como ejercicio)

$$P_1(x) = \left(\frac{5}{\pi^2+40}x + \frac{2\pi(20-\pi^2)}{(\pi^2+40)^2}\right) P_2(x) + 50\frac{\pi^4 - 16\pi^2 + 64}{(\pi^2+40)^2}(2 - x^2).$$

Como $\pi^4 - 16\pi^2 + 64 > 0$ podemos tomar

$$P_3(x) = x^2 - 2.$$

v) Al efectuar la división entre $P_2(x)$ y $P_3(x)$ se obtiene (compruébese nuevamente)

$$P_2(x) = \left((\pi^2 + 40)x - 24\pi\right) P_3(x)$$

por lo que $P_4(x) = 0$.

Así pues, la secuencia de Sturm del polinomio P es

$$\{P(x), P_1(x), P_2(x), P_3(x)\}.$$

Por otro lado, como

$$P_3(x) = x^2 - 2 = (x + \sqrt{2})(x - \sqrt{2}) = \text{MCD}\,\{P(x), P'(x)\}$$

se obtiene que $\xi_1 = \sqrt{2}$ y $\xi_2 = -\sqrt{2}$ son raíces dobles de P. Por tanto, el polinomio P puede factorizarse en la forma

$$P(x) = (x + \sqrt{2})^2(x - \sqrt{2})^2(2x - \pi),$$

de donde se obtiene que $\xi_3 = \dfrac{\pi}{2}$ es la tercera raíz de P. □

10.10. Encontrar una aproximación de las raíces de la ecuación algebraica

$$2x^4 - x^3 + 2x^2 - 7x + 3 = 0.$$

SOLUCIÓN. Denotemos por $\{\xi_1, \xi_2, \xi_3, \xi_4\} \subset \mathbb{C}$ las raíces de la ecuación.

1. Número de raíces positivas y negativas. A partir de

$$\begin{cases} P(x) = 2x^4 - x^3 + 2x^2 - 7x + 3 \\ P(-x) = 2x^4 + x^3 + 2x^2 + 7x + 3 \end{cases}$$

se obtienen las secuencias de coeficientes

$$\begin{cases} P(x) & \longrightarrow & \{2, -1, 2, -7, 3\} & \Rightarrow & 4 \text{ cambios de signo} \\ P(-x) & \longrightarrow & \{2, 1, 2, 7, 3\} & \Rightarrow & 0 \text{ cambios de signo.} \end{cases}$$

Por tanto, por la regla de los signos de Descartes, el polinomio P no tiene raíces negativas y puede tener

$$\begin{cases} 4 \text{ raíces positivas} \\ 2 \text{ raíces positivas} \\ \text{Ninguna raíz positiva.} \end{cases}$$

2. Acotación de las raíces. Aplicando el método de McLaurin se tiene que

$$\lambda = \max\left\{\frac{3}{2}, \frac{7}{2}, 1, \frac{1}{2}\right\} = \frac{7}{2} \Rightarrow 1 + \lambda = \frac{9}{2}$$

y

$$\mu = \max\left\{\frac{2}{3}, \frac{1}{3}, \frac{2}{3}, \frac{7}{3}\right\} = \frac{7}{3} \Rightarrow 1 + \mu = \frac{10}{3} \Rightarrow \frac{1}{1+\mu} = \frac{3}{10}$$

por lo que para cada $k \in \{1, 2, 3, 4\}$ se tiene que

$$0.3 = \frac{3}{10} < |\xi_k| < \frac{9}{2} = 4.5.$$

3. Raíces enteras. Los divisores positivos de $a_0 = 3$ son $\{1, 3\}$. Mediante el algoritmo de Horner se comprueba que el polinomio P no tiene raíces enteras.

4. Raíces racionales. Como los divisores positivos de $a_4 = 2$ son $\{1, 2\}$, las posibles raíces racionales positivas de P son

$$\left\{\frac{1}{2}, \frac{3}{2}\right\}.$$

Mediante el algoritmo de Horner se comprueba que $\xi_1 = \dfrac{1}{2}$ es la única raíz racional del polinomio P.

5. Deflación. A la vista del apartado anterior podemos factorizar el polinomio P en la forma

$$P(x) = \left(x - \frac{1}{2}\right)(2x^3 + 2x - 6) = 2\left(x - \frac{1}{2}\right)(x^3 + x - 3).$$

De esta forma, en lo sucesivo trabajaremos con la ecuación $Q(x) = 0$ siendo

$$Q(x) = x^3 + x - 3.$$

La secuencia de signos en el polinomio Q es $\{+, +, -\}$ por lo que, por la regla de los signos de Descartes, Q tiene una raíz positiva (ya se sabe que Q no puede tener raíces negativas pues, en ese caso, las tendría también P). Por tanto el polinomio P tiene dos raíces positivas (una de ellas es $\xi_1 = \frac{1}{2}$), ninguna negativa y, por tanto, las otras dos raíces son complejas conjugadas.

6. Separación de raíces. Como $Q(1) = -1 < 0 < 7 = Q(2)$ por el teorema de Bolzano existe $\xi_2 \in (1, 2)$ tal que $Q(\xi_2) = 0$. Por tanto, la única raíz positiva de Q se encuentra en el intervalo $(1, 2)$.

7. Método de Newton. Como

$$\begin{cases} Q'(x) = 3x^2 + 1 > 0, & x \in [1,2] \\ Q''(x) = 6x > 0, & x \in [1,2] \end{cases}$$

si consideramos la sucesión

$$\begin{cases} x_0 = 2 \\ x_n = x_{n-1} - \dfrac{Q(x_{n-1})}{Q'(x_{n-1})} = \dfrac{2x_{n-1}^3 + 3}{3x_{n-1}^2 + 1}, \ n \in \mathbb{N} \end{cases}$$

se verifica que

$$\lim_{n \to +\infty} x_n = \xi_2.$$

Los primeros términos de la sucesión anterior vienen dados en la siguiente tabla:

n	x_n
0	2
1	1.46153846153846
2	1.24778815433768
3	1.21418456499057
4	1.21341206394728
5	1.21341166276234
6	1.21341166276223

8. Deflación. Si tomamos $\xi_2 \simeq 1.21341166276223$ y escribimos el polinomio Q como

$$Q(x) \simeq (x - 1.21341166276223)\widetilde{Q}(x) \ \text{ con } \ \partial\widetilde{Q} = 2$$

se obtiene que

$$\widetilde{Q}(x) = x^2 + 1.21341166278x + 2.47236765561,$$

polinomio que tiene por raíces

$$\alpha \pm \beta i \simeq -0.606705831781 \pm 1.45061217736i.$$

Se obtiene así una aproximación de las raíces complejas de P. □

10.11. Aproximar las raíces reales de la ecuación algebraica

$$2x^5 - 100x^2 + 2x - 1 = 0.$$

SOLUCIÓN. Consideremos el polinomio

$$P(x) = 2x^5 - 100x^2 + 2x - 1$$

y denotemos por $\{\xi_1, \xi_2, \xi_3, \xi_4, \xi_5\} \subset \mathbb{C}$ sus raíces.

1. Número de raíces positivas y negativas. Puesto que

$$\begin{cases} P(x) = 2x^5 - 100x^2 + 2x - 1 \\ P(-x) = -2x^5 - 100x^2 - 2x - 1, \end{cases}$$

el método de Descartes asegura que P no tiene raíces negativas y puede tener 1 o 3 positivas.

2. Acotación de las raíces. Como

$$\lambda = \max\left\{\frac{1}{2}, 1, 50\right\} = 50 \text{ y } \mu = \max\{2, 100, 2\} = 100$$

entonces

$$0.00\overbrace{9900} = \frac{1}{101} = \frac{1}{1+\mu} < |\xi_k| < 1 + \lambda = 51$$

para $k = 1, 2, 3, 4, 5$. De esta forma, puesto que P no tiene raíces negativas, podemos asegurar que todas sus raíces reales se encuentran en el intervalo $(0.00\overbrace{9900}, 51)$.

3. Raíces enteras. El único candidato a raíz entera, $x = 1$, claramente no lo es.

4. Raíces racionales. Como $a_5 = 2$, el único candidato a raíz racional es $x = \frac{1}{2}$; mediante el algoritmo de Horner se comprueba que $P\left(\frac{1}{2}\right) \neq 0$.

5. Separación de raíces. La secuencia de Sturm del polinomio $P(x)$, calculada mediante un programa como el que se propone en la práctica 10.3, viene dada por

$$\begin{cases} P(x) = 2x^5 - 100x^2 + 2x - 1 \\ P_1(x) = 10x^4 - 200x^3 + 2 \\ P_2(x) = 60x^2 - 1.6x + 1 \\ P_3(x) = 200.0087x - 2.0027 \\ P_4(x) = -0.9900. \end{cases}$$

A partir de ella, deducimos la tabla:

x	$P(x)$	$P_1(x)$	$P_2(x)$	$P_3(x)$	$P_4(x)$	$N(x)$
0	$-$	$+$	$+$	$-$	$-$	2
$+\infty$	$+$	$+$	$+$	$+$	$-$	1

por lo que el polinomio P tiene una única raíz real ξ que será simple (por ser $\partial P_4 = 0$). Aplicando reiteradamente el proceso de bipartición, se llega a que la raíz ξ se encuentra en el intervalo $(3, 4)$.

6. Aproximación de la raíz. Veamos que en el intervalo $[3, 4]$ se satisfacen las hipótesis de convergencia del método de Newton. Claramente,

$$P(3) = -409 < 0 < 455 = P(4),$$

$$P'(x) = 10x^4 - 200x + 2 \text{ y } P''(x) = 40x^3 - 200.$$

Como

$$P''(x) > 0,\ x \in [3, 4]$$

se verifica que la función P' es estrictamente creciente en el intervalo $[3, 4]$, por lo que

$$P'(x) \geq P'(3) = 212 > 0,\ x \in [3, 4].$$

De esta forma, aplicando el método de Newton, la sucesión dada por

$$\begin{cases} x_0 = 4 \\ x_n = x_{n-1} - \dfrac{P(x_{n-1})}{P'(x_{n-1})} = \dfrac{8x_{n-1}^5 - 100x_{n-1}^2 + 1}{2(5x_{n-1}^4 - 100x_{n-1} + 1)},\ n \in \mathbb{N} \end{cases}$$

converge a la raíz ξ de P. Los primeros términos de la sucesión anterior vienen dados en la siguiente tabla:

n	x_n
0	4
1	3.74177071509648
2	3.68134434305652
3	3.67826072652546
4	3.67825295053744
5	3.67825295048808

10.9.2. Problemas propuestos

10.12. Hallar las raíces reales y complejas, determinando su multiplicidad, de la ecuación

$$x^{13} - x^{12} - x^{11} + x^{10} - x^3 + x^2 + x - 1 = 0.$$

10.13. Acotar las raíces de las siguientes ecuaciones:

a) $x^5 - x^4 + x^3 - x^2 + 1 = 0$.

b) $x^7 - 5x^6 - 27x^5 + 3x^3 + 4x^2 + 7x - 2 = 0$.

10.14. Separar en intervalos de longitud uno, mediante el método de Sturm, las raíces reales de las ecuaciones:

a) $x^5 - 3x^4 + 2x^3 - 3x^2 + 4x + 1 = 0$.

b) $x^5 - x^4 + x^3 - x^2 + 1 = 0$.

10.15. Aproximar las raíces del polinomio

$$P(x) = 5x^5 - 17x^4 - 79x^3 + 269x^2 - 34x - 24.$$

10.16. Calcular las raíces de la ecuación

$$x^5 - x^4 + x^3 - x^2 + 1 = 0.$$

10.17. Consideremos el polinomio

$$P(x) = x^3 + \sqrt{6}x^2 + 2x + \lambda$$

donde $\lambda > 0$.

a) Estudiar, en función del parámetro λ, el número de raíces (reales y complejas) de la ecuación $P(x) = 0$.

b) ¿Para qué valores de λ las raíces de P son múltiples? Hallar todas las raíces de P para esos valores de λ.

c) Fijado $\lambda = 1$, encontrar un intervalo donde pueda aplicarse el método de Newton para calcular una raíz negativa de P. Determinar los primeros términos de la sucesión definida por dicho método.

10.18. Dados λ, $\mu > 0$ se considera el polinomio

$$P(x) = x^3 + \lambda x^2 + \frac{\lambda^2}{3}x + \mu.$$

a) Hallar la secuencia de Sturm para el polinomio $P(x)$, distinguiendo los diversos casos que pueden presentarse en función de los valores que tomen los parámetros λ y μ.

b) Determinar, en función de λ y μ, el tipo de raíces (reales y complejas) que tiene $P(x)$.

c) Encontrar intervalos y valores iniciales en que se pueda aplicar el método de Newton para aproximar las raíces reales de la ecuación

$$x^3 + 3x^2 + 3x + 3 = 0$$

determinando los primeros términos de la sucesión que define dicho método.

10.19. Se considera la ecuación algebraica

$$x^5 + x^4 + 5x^3 + 2x^2 - 13x - 10 = 0.$$

a) Determinar el número de raíces positivas.

b) Encontrar una raíz racional negativa.

c) Hallar el número de raíces reales y complejas de la ecuación anterior.

d) Determinar un intervalo donde se pueda aplicar el método de Newton para aproximar la raíz positiva más pequeña, así como los primeros términos de la sucesión $\{x_n\}_{n=0}^{\infty}$ que determina dicho método.

10.10. Prácticas

10.1. Escribir un programa que, a partir de un polinomio dado, calcule las cotas de sus raíces mediante el método de McLaurin.

10.2. Programar la evaluación de un polinomio en un punto mediante el algoritmo de Horner. Comparar con el comando `polyval` de `MATLAB`.

10.3. Escribir una función en `MATLAB` que calcule la secuencia de Sturm de un polinomio dado.

10.4. Programar el método de Sturm para el cálculo de las raíces reales y distintas que tiene un polinomio en un intervalo dado.

10.5. Adaptar el programa de la práctica 8.3 para el caso particular de que la función F sea un polinomio.

10.6. Programar el método de Bairstow. Aplicarlo a los polinomios de los ejemplos 10.8 y 10.9.

11 Apéndice: Introducción al programa MATLAB

En este apéndice se pretende dar un primer paso en el aprendizaje del uso del programa `MATLAB`. Todas las prácticas propuestas en este libro pueden ser desarrolladas a partir de la edición del estudiante (véase [**Ha**–**Li**]). Estas notas no pretenden, ni mucho menos, equipararse a un manual del programa; simplemente pretenden introducir al lector en el manejo del mismo. El lector puede encontrar eventuales actualizaciones de estas notas en las siguientes páginas WEB:

`http://www.mat.ucm.es/~infante` o `http://www.mat.ucm.es/~jrey`

Antes de comenzar, hagamos algunas consideraciones generales:

a) `MATLAB` distingue entre mayúsculas y minúsculas.

b) La comilla `'` es la que, en un teclado estándar, se encuentra en la tecla de la interrogación.

c) Los comentarios deben ir precedidos por `%` o, lo que es lo mismo, `MATLAB` ignora todo lo que vaya precedido por el símbolo `%`.

d) La ayuda de `MATLAB` es bastante útil; para acceder a la misma basta teclear `help`. Es recomendable usarla para obtener una información más precisa sobre la sintaxis y diversas posibilidades de uso de los comandos.

A continuación se detallan los comandos básicos con los que conviene familiarizarse con vistas a realizar las prácticas de este libro. Se entiende que el usuario tecletaría lo que aparece después del símbolo `>>`, que se supone que es el *prompt* de la máquina, (de hecho, lo óptimo sería que el aprendiz de `MATLAB` reprodujera estos y parecidos ejemplos por sí mismo) y, a continuación, aparece la respuesta que `MATLAB` daría a la instrucción tecleada. Los comentarios aparecen después del símbolo `%`.

11.1. Generalidades

Los cálculos que no se asignan a una variable en concreto se asignan a la variable de respuesta por defecto que es `ans` (del inglés, *answer*):

```
>> 2+3
ans =
5
```

Sin embargo, si el cálculo se asigna a una variable, el resultado queda guardado en ella:

```
>> x=2+3
x =
5
```

Para conocer el valor de una variable, basta teclear su nombre:

```
>> x
x =
5
```

Si se añade un punto y coma (`;`) al final de la instrucción, la máquina no muestra la respuesta ...

```
>> y=5*4;
```

... pero no por ello deja de realizarse el cálculo.

```
>> y
y =
20
```

Las operaciones se evalúan por orden de prioridad: primero las potencias, después la multiplicaciones y divisiones y, finalmente, las sumas y restas. Las operaciones de igual prioridad se evalúan de izquierda a derecha:

```
>> 2/4*3
ans =
1.5000

>> 2/(4*3)
ans =
0.1667
```

Se pueden utilizar las funciones matemáticas habituales. Así, por ejemplo, la función coseno,

```
>> cos(pi)       % pi es una variable con valor predeterminado
ans =            % 3.14159...
-1
```

o la función exponencial.

```
>> exp(1)        % Función exponencial evaluada en 1, es decir,
ans =            % el número e
2.7183
```

Además de la variable `pi`, `MATLAB` tiene otras variables con valor predeterminado; éste se pierde si se les asigna otro valor distinto. Por ejemplo:

```
>> eps           % épsilon de la máquina. Obsérvese que MATLAB
ans =            % trabaja en doble precisión.
2.2204e-016
```

pero ...

```
>> eps=7
eps =
7
```

Otro ejemplo de función matemática: la raíz cuadrada; como puede verse, trabajar con números complejos no da ningún tipo de problema. La unidad imaginaria se representa en `MATLAB` como `i` o `j`, variables con dicho valor como predeterminado.

```
>> sqrt(-4)
ans =
0+2.0000i
```

El usuario puede controlar el número de decimales con que aparece en pantalla el valor de las variables, sin olvidar que ello no está relacionado con la precisión con la que se hacen los cálculos, sino con el aspecto con que éstos se muestran:

```
>> 1/3
ans =
0.3333

>> format long
>> 1/3
ans =
0.33333333333333

>> format        % Vuelve al formato estándar que es el de
                 % 4 cifras decimales.
```

Para conocer las variables que se han usado hasta el momento:

```
>> who
Your variables are:
ans eps x y
```

o, si se quiere más información (obsérvese que todas las variables son *arrays*):

```
>> whos
Name  Size  Bytes       Class
ans   1x1     8     double array
eps   1x1     8     double array
 x    1x1     8     double array
 y    1x1     8     double array
Grand total is 4 elements using 32 bytes
```

Para deshacerse de una variable:

```
>> clear y
>> who
Your variables are:
ans eps x
```

11.2. Vectores y matrices

Para definir un vector fila, basta introducir sus coordenadas entre corchetes:

```
>> v=[1 2 3]     % Vector de 3 coordenadas.
v=
1 2 3

>> w=[4 5 6];
```

El operador ' es el de trasposición (en realidad trasposición y conjugación):

```
>> w'
ans =
4
5
6
```

Si queremos declarar un vector de coordenadas equiespaciadas entre dos dadas, por ejemplo, que la primera valga 0, la última 20 y la distancia entre coordenadas sea 2, basta poner:

```
>> vect1=0:2:20
vect1 =
0 2 4 6 8 10 12 14 16 18 20
```

Equivalentemente, si lo que conocemos del vector es que la primera coordenada vale 0, la última 20 y que tiene 11 en total, escribiremos:

```
>> vect2=linspace(0,20,11)
vect2 =
0 2 4 6 8 10 12 14 16 18 20
```

A las coordenadas de un vector se accede sin más que escribir el nombre del vector y, entre paréntesis, su índice:

```
>> vect2(3)
ans =
4
```

y se pueden extraer subvectores, por ejemplo:

```
>> vect2(2:5)
ans=
2 4 6 8
```

o

```
>> vect1(:)
ans=
0
2
4
6
8
10
12
14
16
18
20
```

Las matrices se escriben como los vectores, pero separando las filas mediante un punto y coma; así una matriz 3×3:

```
>> M=[1 2 3;4 5 6;7 8 9]
M =
```

```
1 2 3
4 5 6
7 8 9

>> M'              % Su traspuesta (su adjunta).
ans =
1 4 7
2 5 8
3 6 9

>> mat=[v;w;0 0 1]
mat =              % También es una matriz 3x3.
1 2 3
4 5 6
0 0 1
```

A los elementos de una matriz se accede sin más que escribir el nombre de la matriz y, entre paréntesis, los respectivos índices:

```
>> mat(1,3)        % Elemento de la primera fila y tercera
ans =              % columna de la matriz mat.
3
```

También se puede acceder a un fila o columna completas,

```
>> mat(:,2)        % Segunda columna de mat.
ans =
2
5
0

>> mat(2,:)        % Su segunda fila.
ans =
4 5 6
```

acceder a la matriz como si fuera una columna,

```
>> M(2:7)          % Los elementos segundo a séptimo de la
ans =              % matriz como columna.
4
7
2
5
8
3
```

o acceder a cualquiera de sus submatrices:

```
>> mat(2:3,[1 3])
ans =              % Submatriz formada por los elementos que
4 6                % están en "todas" las filas que hay entre
0 1                % la segunda y la tercera y en las columnas
                   % primera y tercera.
```

Existen algunas matrices definidas previamente; por ejemplo, la matriz identidad,

```
>> eye(5)          % eye se pronuncia en inglés como I.
ans =
1 0 0 0 0
0 1 0 0 0
0 0 1 0 0
0 0 0 1 0
0 0 0 0 1
```

la matriz nula,

```
>> zeros(3)
ans =
0 0 0
0 0 0
0 0 0
```

o la matriz cuyos elementos valen todos 1:

```
>> ones(4)
ans =
1 1 1 1
1 1 1 1
1 1 1 1
1 1 1 1
```

Se puede conocer el tamaño de una matriz y la longitud de un vector:

```
>> size(mat)       % Dimensiones de la matriz mat (número de
ans =              % filas y de columnas).
3 3

>> size(v)
ans =
1 3
```

```
>> length(v)      % Longitud del vector (número de coordenadas)
ans =
3
```

Existen comandos que permiten crear de forma sencilla matrices. Por ejemplo:

```
>> diag(v)        % Matriz diagonal cuya diagonal
ans =             % es el vector v.
1 0 0
0 2 0
0 0 3

>> diag(diag(M))  % Matriz diagonal con la diagonal de M.
ans =             % La sentencia diag(M) da el vector formado
1 0 0             % por la diagonal de la matriz M.
0 5 0
0 0 9

>> diag(ones(1,4),1)+diag(ones(1,4),-1)
ans =             % Matriz tridiagonal 5x5 con 0 en la diagonal
0 1 0 0 0         % principal y 1 en la sub y superdiagonal.
1 0 1 0 0
0 1 0 1 0
0 0 1 0 1
0 0 0 1 0

>> tril(M)        % Matriz formada por la parte triangular
ans =             % inferior de M.
1 0 0
4 5 0
7 8 9

>> triu(M)        % Matriz formada por la parte triangular
ans =             % superior de M.
1 2 3
0 5 6
0 0 9
```

11.3. Operaciones con vectores y matrices

Las funciones matemáticas elementales están definidas de forma que se pueden aplicar sobre *arrays*. El resultado es el *array* formado por la aplicación de la función

a cada elemento del *array*. Así:

```
>> log(v)
ans =
0 0.6931 1.0986

>> p=(0:0.1:1)*pi     % Vector definido como el producto de un
p =                   % vector por un escalar.
Columns 1 through 7
0 0.3142 0.6283 0.9425 1.2566 1.5708 1.8850
Columns 8 through 11
2.1991 2.5133 2.8274 3.1416

>> x=sin(p)
x =
Columns 1 through 7
0 0.3090 0.5878 0.8090 0.9511 1.0000 0.9511
Columns 8 through 11
0.8090 0.5878 0.3090 0.0000
```

Las operaciones habituales entre *arrays* (suma, resta y producto escalar de vectores; suma, resta, producto y potencia de matrices) se representan con los operadores habituales:

```
>> v,w            % Recordamos los valores de v y w.
v =
1 2 3
w =
4 5 6

>> z=v*w'         % Producto escalar.
z =               % (Producto de matrices 1x3 por 3x1).
32

>> Z=w'*v         % Producto de matrices 3x1 por 1x3 =
Z =               % matriz 3x3.
4 8 12
5 10 15
6 12 18

>> v*w            % Los vectores v y w no se pueden multiplicar.
??? Error using ==> *
Inner matrix dimensions must agree.
```

```
>> mat            % Recordamos el valor de la matriz mat.
mat =
1 2 3
4 5 6
0 0 1

>> mat^2          % Matriz mat elevada al cuadrado.
ans =
9 12 18
24 33 48
0  0  1
```

También pueden efectuarse multiplicaciones, divisiones y potencias de *arrays*, entendiéndolas como elemento a elemento (como, de hecho, se realizan la suma y la resta). El operador utilizado para ellas es el habitual precedido por un punto; es decir:

```
>> v.*w           % Vector formado por los productos de
ans =             % las respectivas coordenadas:
4 10 18           % ans(i)=v(i)*w(i).

>> w./v           % Vector formado por el cociente de cada
ans =             % coordenada de w entre la coordenada corres-
4.0000 2.5000 2.0000       % pondiente de v: ans(i)=w(i)/v(i).

>> mat.^2         % Matriz cuyos elementos son los de mat
ans =             % elevados al cuadrado: ans(i,j)=mat(i,j)^2.
1  4  9
16 25 36
0  0  1
```

Finalmente, pueden calcularse determinantes:

```
>> det(mat)
ans=
-3
```

y resolverse sistemas de ecuaciones lineales con el versátil comando \:

```
>> mat\v'
ans=
2.6667
-5.3333
3.0000
```

11.4. Variables lógicas

También existen variables lógicas que toman los valores 0 (falso) o 1 (verdadero). Por ejemplo:

```
>> abs(v)>=2      % Vector lógico cuyas coordenadas valen 1 si
ans =             % la coordenada correspondiente de v es >= 2
0 1 1             % y 0 si no lo es.

>> vector=v(abs(v)>=2)
vector =          % Vector formado por la coordenadas de v que
2 3               % verifican la desigualdad.

>> v2=[3 2 1]
v2 =
3 2 1

>> logica=v==v2   % Asignación de un valor lógico (el signo
logica =          % igual doble es el igual lógico).
0 1 0

>> logic2=v~=v2   % Distinto (~ es el operador de negación)
logic2 =
1 0 1
```

11.5. Polinomios

Se puede trabajar con polinomios: basta tener en cuenta que un polinomio no es más que un vector. El orden de los coeficientes es de mayor a menor grado, por ejemplo:

```
>> p=[1 0 2 0 3]  % Polinomio x^4+2*x^2+3
p =
1 0 2 0 3

>> q=[2 1 0]      % Polinomio 2*x^2+x
q =
2 1 0
```

MATLAB tiene funciones específicas para polinomios como:

```
>> polyval(p,-1)  % Evaluación del polinomio x^4+2x^2+3 en x=-1.
ans =
```

```
6

>> pro=conv(p,q)  % Producto de los polinomios p y q.
pro =
2 1 4 2 6 3 0

>> deconv(pro,p)  % Cociente entre pro y p; obviamente,
ans =             % el resultado es q.
2 1 0

>> roots(pro)     % Raíces del polinomio pro.
ans =
0
0.6050+1.1688i
0.6050-1.1688i
-0.6050+1.1688i
-0.6050-1.1688i
-0.5000

>> poly([i -i 1/2 pi])   % Polinomio mónico que tiene por raíces
ans =                    % a los números i,-i, 0.5 y pi.
1.0000 -3.6416 2.5708 -3.6416 1.5708
```

11.6. Derivadas y primitivas

Dentro del módulo (*toolbox*) de matemática simbólica, se utiliza el programa de cálculo simbólico `MAPLE`. Con estas herramientas, se puede trabajar con funciones,

```
>> f='sin(x)'            % Función sin(x) definida mediante una
f =                      % cadena de caracteres.
sin(x)
```

calcular derivadas,

```
>> diff(sym(f))
ans =
cos(x)

>> diff(sym(f),2)        % Derivada segunda de f.
ans =
-sin(x)
```

o encontrar primitivas:

```
>> int(sym('log(x)'))       % Primitiva de la función logaritmo.
ans =
x*log(x)-x

>> diff(sym('x*log(x)-x'))
ans =                       % Comprobación.
log(x)
```

11.7. Gráficas de funciones

MATLAB tiene un gran potencial de herramientas gráficas. Se pueden dibujar los valores de un vector frente a otro (de la misma longitud).

```
>> x=pi*(-1:0.1:1);
>> y=x.*sin(x);
>> plot(x,y)      % Por defecto une los puntos (x(i),y(i))
                  % mediante una poligonal.
                  % (Ver la figura 11.1)
```

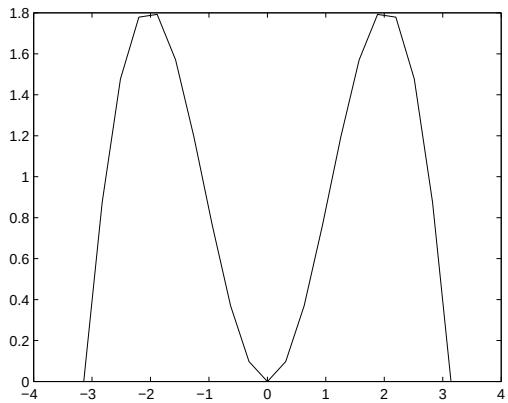


Figura 11.1: Función $f(x) = x \operatorname{sen} x$ con paso $h = 0.1$.

Como se ve, con pocos puntos la gráfica tiene un aspecto demasiado lineal a trozos. Para "engañar" al ojo, basta tomar más puntos.

```
>> x=pi*(-1:0.01:1);
>> y=x.*sin(x);
>> plot(x,y)      % Ver la figura 11.2
```

También pueden dibujarse funciones. Así:

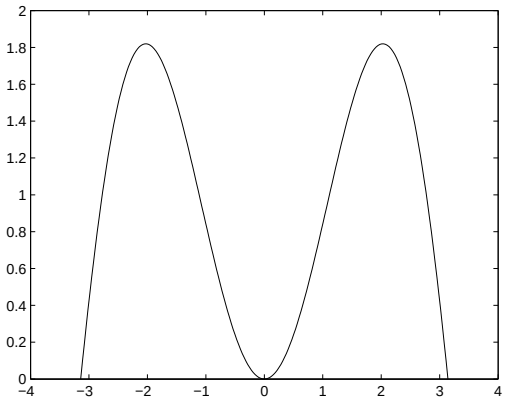


Figura 11.2: Función $f(x) = x \operatorname{sen} x$ con paso $h = 0.01$.

```
>> fplot('sin(x)',[0 2*pi])   % Dibuja la función seno en
                              % el intervalo [0,2*pi].
                              % (Ver la figura 11.3)
```

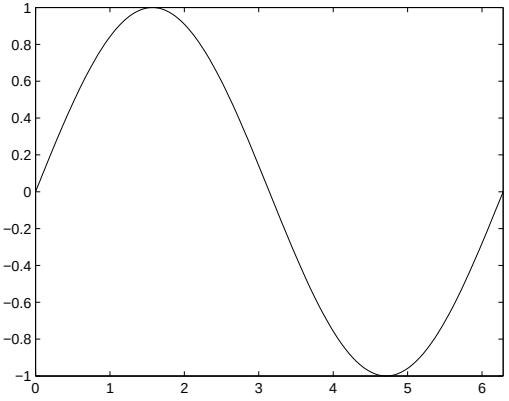


Figura 11.3: Función $f(x) = \operatorname{sen} x$ en $[0, 2\pi]$.

```
>> hold on         % Mantiene en la ventana gráfica los
                   % dibujos anteriores.
>> fplot('cos(x)',[0 2*pi])  % Dibuja sobre la gráfica anterior
                             % la función cos(x).
                             % (Ver la figura 11.4)
```

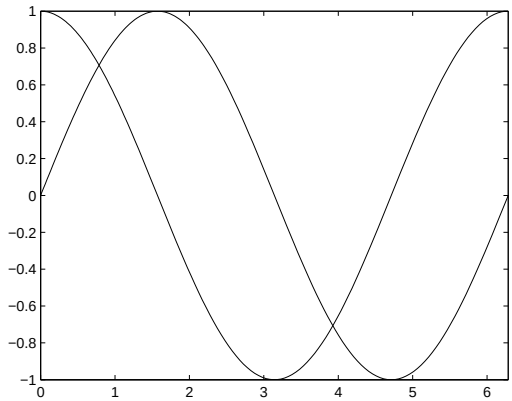


Figura 11.4: Funciones $f(x) = \operatorname{sen} x$ y $g(x) = \cos x$ en $[0, 2\pi]$.

```
>> hold off       % Con esto olvida los dibujos anteriores
                  % y dibuja en una ventana nueva.
>> fplot('x^2*sin(1/x)',[-0.05 0.05]) % f(x)=x^2*sin(1/x).
                                      % (Ver la figura 11.5)
```

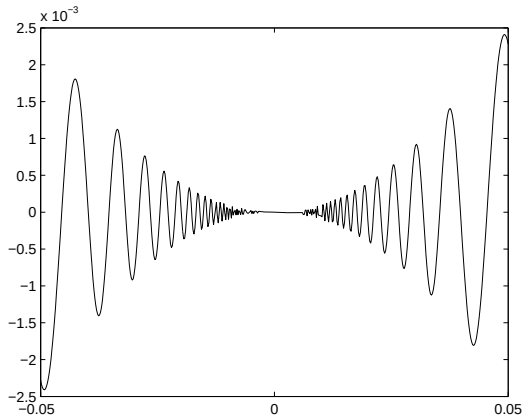


Figura 11.5: Función $f(x) = x^2 \operatorname{sen} \dfrac{1}{x}$ en $[-0.05, 0.05]$.

También puede usarse el versátil comando `ezplot` (se lee como *easy plot*) que permite dibujar funciones,

```
>> ezplot('exp(x)')  % Dibuja la función exponencial en un
                     % intervalo adecuado a la función
                     % (Ver la figura 11.6)
```

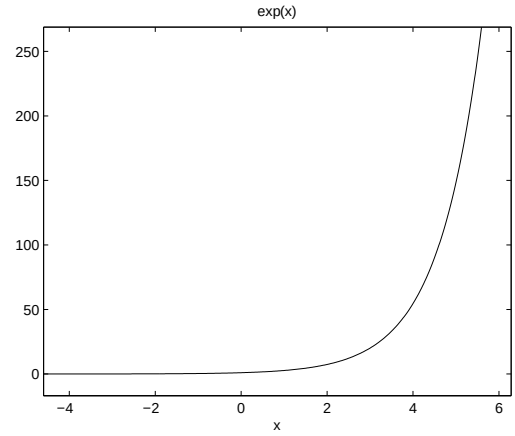


Figura 11.6: Función $f(x) = e^x$.

curvas en paramétricas,

```
>> ezplot('sin(t)','cos(t)',[0 pi])   % (Ver la figura 11.7)
```

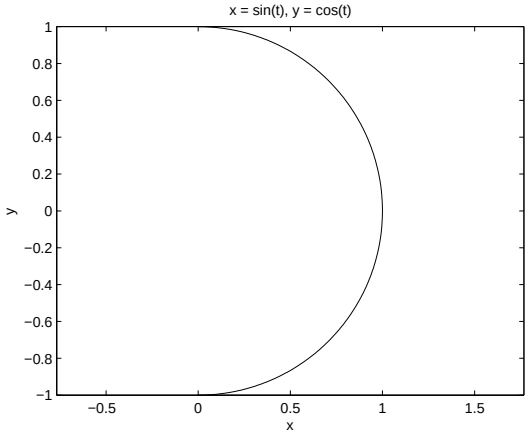


Figura 11.7: Curva $(x, y) = (\text{sen}(t), \cos(t))$ para $t \in [0, \pi]$.

e implícitas

```
>> ezplot('x^2 - y^2 - 1')            % (Ver la figura 11.8)
```

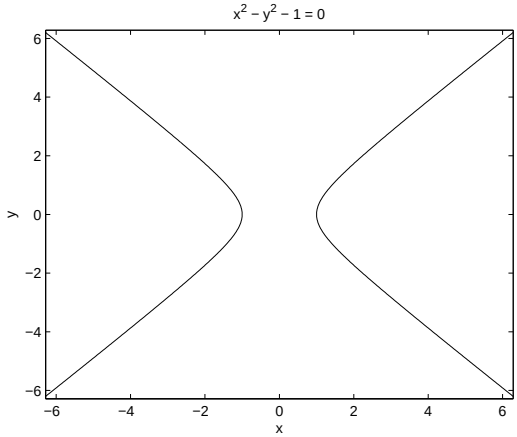


Figura 11.8: Curva implícita $x^2 - y^2 - 1 = 0$.

También permite dibujar superficies. La forma más sencilla es mediante el comando `ezsurf`,

```
>> ezsurf('sin(x*y)',[-2 2 -2 2])     % (Ver la figura 11.9)
```

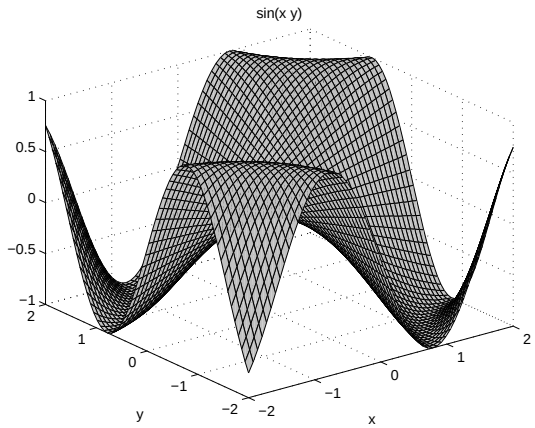


Figura 11.9: Superficie $z = \text{sen}(xy)$ con $x, y \in [-2, 2]$.

aunque se pueden realizar gráficas más sofisticadas:

```
>> t=0:0.001:0.009;
>> v=900:1025;
```

```
>> [T V]=meshgrid(t,v);
>> aux1=16*pi^2*(T.^2).*((V-918).^2).*((V-1011).^2);
>> aux2=aux1+(2*V-1929).^2;
>> w=T./aux2;
>> z=35000000*w;
>> surfl(t,v,z);      % Este comando dibuja la superficie creada
>> shading interp;    % mediante las órdenes anteriores. Los si-
>> colormap(pink);    % guientes sirven para modificar el dibujo.
>> rotate3d;          % Sirve para girar la figura con el ratón.
                      % Ver la figura 11.10.
```

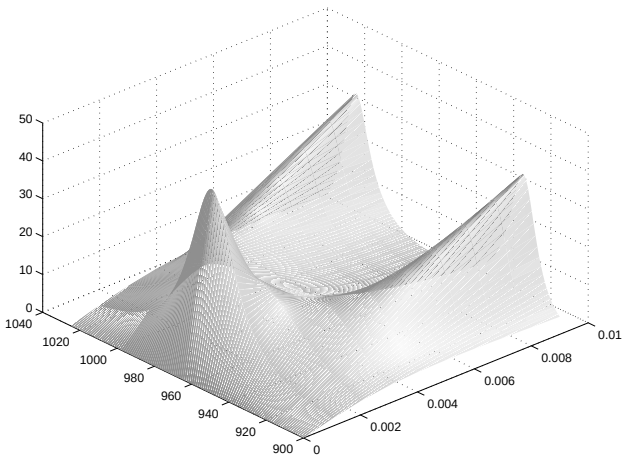


Figura 11.10: Gráfica de una superficie.

11.8. Programación con MATLAB

Para escribir un programa con `MATLAB` habrá que crear un fichero que tenga extensión `.m` y contenga las instrucciones. Esto se puede hacer con cualquier editor de textos, pero tiene algunas ventajas usar el editor propio de `MATLAB` llamado `medit`.

`MATLAB` trabaja con memoria dinámica, por lo que no es necesario declarar las variables que se van a usar. Por esta misma razón, habrá que tener especial cuidado y cerciorarse de que entre las variables del espacio de trabajo no hay ninguna que se llame igual que las de nuestro programa (proveniente, por ejemplo, de un programa previamente ejecutado en la misma sesión), porque esto podría provocar conflictos. A menudo, es conveniente reservar memoria para las variables (por ejemplo, si se

van a utilizar matrices muy grandes); para ello, basta con asignarles cualquier valor. Del mismo modo, si se está usando mucha memoria, puede ser conveniente liberar parte de ella borrando (`clear`) variables que no se vayan a usar más.

Un programa escrito en `MATLAB` admite la mayoría de las estructuras de programación al uso y su sintaxis es bastante estándar. En los siguientes ejemplos se muestra la sintaxis de algunas de estas estructuras (`if`, `for`, `while`,...).

a) Calcular la suma de los n primeros términos de la sucesión $1, 2x, 3x^2, 4x^3, \ldots$, para un valor de x dado:

```
n=input('¿Cuántos términos quieres sumar?  ');
x=input('Dame el valor del número x  ');
suma=1;
for i=2:n
   suma=suma+i*x^(i-1);
end
disp('El valor pedido es')
disp(suma)
```

b) Decidir si un número natural es primo:

```
n=input('Número natural que deseas saber si es primo ');
i=2;
primo=1;
while i<=sqrt(n)
   if rem(n,i)==0  % Resto de dividir n entre i.
      primo=0;
      break
   end
   i=i+1;
end
if primo
   disp('El número dado es primo.')
else
   disp('El número dado no es primo.')
   disp('De hecho, es divisible por:')
   disp(i)
end
```

c) Escribir un número natural en una base dada (menor que diez):

```
n=input('Dame el número que quieres cambiar de base ');
base=input('¿En qué base quieres expresarlo? ');
i=1;
while n>0
   c(i)=rem(n,base);
   n=fix(n/base);   % Parte entera de n/base.
   i=i+1;
end
disp('La expresión en la base dada es:')
i=i-1;
disp(c(i:-1:1))
```

Por último, también pueden programarse funciones. La primera instrucción de un fichero que contenga una función de nombre `fun` debe ser:

```
function [argumentos de salida]=fun(argumentos de entrada)
```

Es conveniente que el fichero que contenga la función se llame como ella; así, la función anterior debería guardarse en el fichero `fun.m`; por ejemplo, si se desea programar una función que calcule, mediante el algoritmo de Euclides, el máximo común divisor de dos números naturales, basta escribir un fichero `euclides.m` cuyo contenido sea:

```
function m=euclides(a,b)
% Cálculo del máximo común divisor de dos números naturales
% mediante el algoritmo de Euclides.
if a<b
   c=b;
   b=a;
   a=c;
end
while b>0
   c=rem(a,b);
   a=b;
   b=c;
end
m=a;
```

Si, una vez escrito el fichero anterior, en el espacio de trabajo o en un programa se escribe la instrucción

```
mcd=euclides(33,121)
```

en la variable `mcd` se almacenará el valor 11.

Las variables de una función son siempre locales. Por tanto, aunque en el seno de la función se modifiquen los argumentos de entrada, el valor de las variables correspondientes queda inalterado. Por ejemplo, en la función `euclides.m` se modifica el valor de los argumentos de entrada, pero, sin embargo:

```
>> x=15;
>> mcd=euclides(x,3);
>> x
x =
 15
```

Si se pretende que las modificaciones de un argumento de entrada afecten a la variable correspondiente, deberá situarse dicho argumento, además, en la lista de argumentos de salida.

Bibliografía básica

[**Ap**] T. M. Apostol: *Calculus*. Reverté. 1985.

[**Au–Be–De**] A. Aubanel, A. Bensey y A. Delshams: *Utiles básicos de Cálculo Numérico*. Labor. 1993.

[**Bu–Fa**] R. Burden y J. D. Faires: *Análisis Numérico*. Grupo Editorial Iberoamérica. 1996.

[**Ci**] P. G. Ciarlet: *Introduction à l'Analyse Numérique Matricielle et à l'Optimization*. Masson. 1982.

[**Co–Bo**] S. D. Conte y C. de Boor: *Elementary Numerical Analysis. An Algorithmic Approach*. McGraw–Hill. 1972 (segunda edición).

[**De–Ma**] B. P. Demidovich e I. A. Maron: *Cálculo Numérico Fundamental*. Paraninfo. 1988.

[**G**] H. H. Goldstine: *A History of Numerical Analysis from the* 16*th through* 19*th Century*. Springer–Verlag. 1977.

[**Go–Lo**] G. H. Golub y C. F. Van Loan: *Matrix Computations*. North Oxford Academic. 1989 (segunda edición).

[**Hä–Ho**] G. Hämmerlin y K.–H. Hoffmann: *Numerical Mathematics*. Springer–Verlag. 1991.

[**Ha–Li**] D. Hanselman y B. Littlefield: *The Student Edition of* `MATLAB`. *The Language of Technical Computing*. Version 5. User's Guide. Prentice–Hall. 1997.

[**Ki–Ch**] D. Kincaid y W. Cheney: *Análisis Numérico: las matemáticas del cálculo científico*. Addison–Wesley Iberoamericana. 1994.

[**Is–Ke**] E. Isaacson y H. B. Keller: *Analysis of Numerical Methods*. John Wiley. 1966.

[**La–Th**] P. Lascaux y R. Théodor: *Analyse Numérique Matricielle Appliquée a l'Art de l'Ingénieur*. Masson. 1987.

[**Sa**] J. M. Sanz Serna: *Diez Lecciones de Cálculo Numérico*. Universidad de Valladolid. 1998.

[**Si–Ma**] M. Sibony y J. C. Mardon: *Analyse Numérique* (dos tomos). Hermann. 1984.

[**St–Bu**] J. Stoer y R. Bulirsch: *Introduction to Numerical Analysis*. Springer–Verlag. 1980.

[**Th**] R. Théodor: *Initiation à l'Analyse Numérique*. Masson. 1986 (segunda edición).

Bibliografía de consulta

[**Ca–Lu–Wi**] B. Carnahan, H. A. Luther y J. O. Wilkes: *Métodos Numéricos Aplicados*. Castillo. 1978.

[**Ci–Mi**] P. G. Ciarlet, B. Miara y J. M. Thomas: *Exercices d'Analyse Numérique Matricielle et d'Optimization avec Solutions*. Masson. 1986 (segunda edición).

[**Ci–Li**] P. G. Ciarlet y J. L. Lions: *Handbook of Numerical Analysis. Vol. I: Difference methods. Solutions of equations in* $\mathbb{R}^N$. North Holland. 1990.

[**Coh**] A. M. Cohen: *Análisis Numérico*. Reverté. 1977.

[**Fu**] J. L. de la Fuente: *Tecnologías Computacionales para Sistemas de Ecuaciones, Optimización Lineal y Entera*. Reverté. 1993.

[**Ga**] M. Gasca: *Cálculo Numérico*. UNED. 1986.

[**He**] P. Henrici: *Elementos de Análisis Numérico*. Trillas. 1972.

[**Hi**] F. B. Hildebrand: *Introduction to Numerical Analysis*. McGraw–Hill. 1974 (segunda edición).

[**Ho**] A. S. Householder: *The Theory of Matrices in Numerical Analysis*. Dover. 1964.

[**Ko**] A. I. Kostrikin: *Introduction to Algebra*. Springer–Verlag. 1982.

[**No**] J. P. Nougier: *Méthodes de Calcul Numérique*. Masson. 1985.

[**Or–Rh**] J. M. Ortega y W. C. Rheinboldt: *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press. 1970.

[**Ra**] A. Ralston: *Introducción al Análisis Numérico*. Limusa. 1970.

[**Sa**] A. A. Samarski: *Introducción a los Métodos Numéricos*. MIR. 1986.

[**Tr–Ba**] L. N. Trefethen y D. Bau III: *Numerical Linear Algebra*. SIAM. 1997.

[**Va**] R. S. Varga: *Matrix Iterative Analysis*. Prentice–Hall. 1962.

[**Wi**] J. H. Wilkinson: *The Algebraic Eigenvalue Problem*. Oxford University Press. 1965.

[**Yo**] D. Young: *Iterative Solutions of Large Linear Systems*. Academic Press. 1971.

Índice alfabético

TÍTULOS PUBLICADOS

ÁLGEBRA LINEAL (vol. 2), *A. Gutiérrez Gómez y F. García Castro.*
ANÁLISIS DE DATOS EN LAS CIENCIAS DE LA ACTIVIDAD FÍSICA Y DEL DEPORTE, *M.ª I. Barriopedro y C. Muniesa.*
CIENCIA DE MATERIALES, *P. Coca Rebollero y J. Rosique Jiménez.*
CURSO DE GENÉTICA MOLECULAR E INGENIERÍA GENÉTICA, *M. Izquierdo Rojo*
ECOLOGÍA, *J. Rodríguez.*
ECUACIONES DIFERENCIALES II, *C. Fernández Pérez y J. M. Vegas Montaner.*
ENZIMOLOGÍA, *I. Núñez de Castro.*
FÍSICA CUÁNTICA, *C. Sánchez del Río (coord.).*
FISIOLOGÍA VEGETAL, J. *Barceló Coll, G. Nicolás Rodrigo, B. Sabater García y R. Sánchez Tamés.*
INTEGRACIÓN DE FUNCIONES DE VARIAS VARIABLES, *J. A. Facenda Aguirre, F. J. Freniche Ibáñez.*
INTRODUCCIÓN A LA ESTADÍSTICA Y SUS APLICACIONES, *R. Cao Abad, M. Francisco Fernández, S. Naya Fernández, M. A. Presedo Quindimil, M. Vázquez Brage, J. A. Vilar Fernández, J. M. Vilar Fernández.*
MÉTODOS NUMÉRICOS. Teoría, problemas y prácticas con MATLAB, *J. A. Infante del Río y J. M.ª Cabezas.*
PROBLEMAS, CONCEPTOS Y MÉTODOS DEL ANÁLISIS MATEMÁTICO. 1. Números reales, sucesiones y series, *M. de Guzmán y B. Rubio.*
PROBLEMAS, CONCEPTOS Y MÉTODOS DEL ANÁLISIS MATEMÁTICO. 2. Funciones, integrales, derivadas, *M. de Guzmán y B. Rubio.*
SERIES DE FOURIER Y APLICACIONES. Un tratado elemental, con notas históricas y ejercicios resueltos, *A. Cañada Villar.*
TABLAS DE COMPOSICIÓN DE ALIMENTOS, *O. Moreiras, A. Carbajal, L. Cabrera y C. Cuadrado.*
TECNOLOGÍA MECÁNICA Y METROTECNIA, *P. Coca Rebollero y J. Rosique Jiménez.*