UNPA
Universidad Nacional de la Patagonia Austral

PROF. SEBASTIÁN A. HERNÁNDEZ(*)

APUNTES DE CÁLCULO NUMÉRICO

CON APLICACIONES SOBRE *EULER MATH TOOLBOX*

```
function map Trunc(x:real, n:natural)
## Truncamiento de numeros reales
## * Parametros de entrada:
##     x: elemento real o vector de reales
##     n: cantidad de decimales de mantisa
## * Parametro de salida:
##     valor o valores de x truncados a n decimales
s=sign(x);
x=s*x;
j=0;
if x<10^n && x<>0 then
   repeat while x<10^n
      x=x*10;
      j=j+1;
   end
   x=10^(-j+1)*floor(x/10);
else
   repeat while x>10^n
      x=x/10;
      j=j+1;
   end
x=10^(j)*floor(x);
endif
return {s*x}
endfunction
```

UNIVERSIDAD NACIONAL DE LA PATAGONIA AUSTRAL
UNIDAD ACADÉMICA RÍO GALLEGOS

Ed. UNPA Ediciones UNPAedita
Universidad Nacional de la Patagonia Austral

Hernández, Sebastián L.
Apuntes de cálculo numérico : con aplicaciones sobre Euler Math Toolbox / Sebastián L. Hernández. - 1a ed . - Río Gallegos : Universidad Nacional de la Patagonia Austral, 2018.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-987-3714-49-8

1. Matemática. 2. Cálculos. I. Título.
CDD 515

Primera edición: Marzo 2018

Diagramación | Sebastián Hernández / Rogelio Corvalán

| Puesta en Página | Rogelio Corvalán

Hecho el depósito que establece la ley 11.723

Ediciones UNPAedita
Universidad Nacional de la Patagonia Austral

UNPA
Universidad Nacional de la Patagonia Austral

RECTORADO |

Av. Lisandro de la Torre Nº 860 CP. Z9400JZR
Río Gallegos . Santa Cruz . Patagonia Austral Argentina
TE +54 02966 442686 _ FAX +54 02966 442377 / 76
rectorad@unpa.edu.ar . www.unpa.edu.ar

ISBN 978-987-3714-49-8
9 789873 714498

Índice General

1 Conceptos Básicos del Cálculo Numérico

1.1. Estimación de errores

Uno de los objetivos principales de la computación científica es desarrollar métodos precisos y eficientes para calcular aproximaciones de modelos que son imposibles ó muy costosos de resolver por métodos analíticos. Sin embargo, es necesario poder controlar las diferentes fuentes de error a fin de no modificar los resultados calculados.

Los resultados numéricos se ven afectados por muchos tipos de error. Algunas fuentes de error son muy difíciles de eliminar, otras pueden ser reducidas ó eliminadas por la reescritura de fórmulas ó bien al realizar cambios en la secuencia computacional utilizada, por citar un par de ejemplos clásicos.

Los errores se propagan desde el inicio del cómputo hasta el resultado final, a veces con una considerable amplificación, otras veces observándose oscilaciones. Es importante poder distinguir entre el nuevo error producido durante cada cálculo en particular (error de procesamiento) y el error heredado (propagado) desde los datos durante todos los cálculos.

Las siguientes son los errores más comunes:

A **Error en los datos iniciales**. Los datos iniciales ó de entrada pueden ser el resultado de mediciones influenciadas por errores sistemáticos ó por perturbaciones temporales. Un **error de redondeo** ocurre, por ejemplo, cada vez que un número irracional es recortado para utilizar una cantidad fija de decimales. También puede ocurrir cuando una fracción es convertida al tipo de números usados por la PC.

B **Error de redondeo**. La limitación de los números de punto flotante en una PC lleva asociada una pérdida frecuente de información que, de acuerdo al contexto, puede ó no ser importante. Dos casos típicos son:

Si el procesador utilizado no tiene la capacidad de operar con números de más de s dígitos, entonces el producto exacto de dos números de s dígitos de longitud (que tiene $2s$ ó $2s-1$ dígitos) no puede ser utilizado en cálculos siguientes porque el resultado debe ser recortado.

Si, en un procesador que opera con punto flotante, un número relativamente pequeño b es sumado a otro número a, entonces algunos dígitos de b se *pierden* y no tendrán efecto en futuros cálculos que dependan del valor de $a+b$. El efecto de redondeos por la capacidad del procesador puede ser notorio durante un cálculo de muchos pasos, ó en un algoritmo numéricamente inestable.

C **Error de truncamiento**. Este tipo de error ocurre cuando un proceso límite es truncado antes de llegar al valor límite. Por ejemplo, cuando una serie infinita es truncada después de un número finito de términos, ó cuando una derivada es aproximada a través de un cociente en diferencias[1]. Otro ejemplo es cuando una función no lineal es aproximada con una función lineal en un intervalo definido.

D **Simplificaciones en el modelo matemático**. En la mayoría de las aplicaciones de la matemática se hacen idealizaciones. En un problema de mecánica, por ejemplo, es habitual asumir que la cuerda que sostiene un péndulo carece de masa. En otros tipos de problemas, puede ser una complicación considerar un cuerpo como un objeto homogéneo y lleno completamente de materia, en vez de estar compuesto por átomos. Este tipo de error es más complicado de estimar que los descritos anteriormente.

E **Error humano**. En cualquier trabajo que se necesite operar con números, son comunes los errores de administración de información, errores en cálculos manuales y discrepancias en los datos procesados. A veces las rutinas computacionales también contienen errores, asociados al programador ó a la falta de consideración del procesador a utilizar.

Los resultados intermedios durante un proceso largo de cómputo pueden mostrar errores que no serán visibles al llegar al resultado final. Debe, para esto, considerarse qué tipo de comprobaciones pueden ser realizadas. Dos de las comprobaciones más simples son revisar los órdenes de magnitud obtenidos en cálculos intermedios y que los resultados intermedios sean regulares en procesos iterativos.

Desde otro punto de vista, es posible distinguir entre los errores *controlables* y *no controlables*. Errores del tipo A y D son considerados como *no controlables* dentro del proceso numérico, aunque reconstruir los modelos y revisar sistemáticamente los datos de entrada a veces puede resultar en grandes beneficios. Errores del tipo C son generalmente *controlables*, por ejemplo, a través de la cantidad de iteraciones realizadas para obtener un resultado ó eligiendo la longitud del paso en una simulación. Los errores del tipo B son *controlables* débilmente. Para ello se introdujo la *doble precisión* durante los cálculos, aunque es mejor reescribir las fórmulas ó realizar un proceso inteligente, que considere la limitación del procesador involucrado.

1.1.1. Error relativo y absoluto

El concepto de aproximación es central en la mayoría de las aplicaciones matemáticas. Lo ideal es trabajar con valores aproximados que satisfagan los requerimientos establecidos. Con el fin de poder analizar las aproximaciones, se introduce la siguiente definición:

Definición 1. *Sea $\tilde{x}$ una aproximación del valor correcto x. Entonces se define como error absoluto a:*

$$\epsilon_A = |\tilde{x} - x|,$$

y, si $x \neq 0$ el error relativo se define como:

$$\epsilon_R = \left| \frac{\tilde{x} - x}{x} \right|$$

En algunos libros clásicos de análisis numérico, se definen los errores con un respectivo signo. La aplicación del *valor absoluto* ofrece una estimación más grosera pero

[1] en este caso una expresión más correcta es **error de discretización**

también más realista. Es muy difícil poder estimar si el error cometido es un *error por exceso* ó un *error por defecto*. Por la definición anterior, la notación $x = \tilde{x} \pm \epsilon$ significa $|\tilde{x} - x| \leq \epsilon$. Por ejemplo, si $x = 0{,}5678 \pm 0{,}0014$ entonces $0{,}5862 \leq x \leq 0{,}5890$, y $|\tilde{x} - x| \leq 0{,}0014$. En el caso de que $\mathbf{x}$ sea un vector, se utiliza la misma definición pero en vez de aplicar valor absoluto debe calcularse la diferencia y luego aplicar alguna norma vectorial.

1.1.2. Redondeo y truncamiento

Cuando se cuenta el *número de dígitos* en un valor numérico, no deben incluirse los ceros al comienzo del número, ya que estos ceros sólo marcan la posición del punto decimal. Sin embargo, al contar el *número de decimales*, sí deben incluirse los ceros ubicados a la derecha del punto decimal. Por ejemplo, el número 0,00147 está expresado con tres dígitos pero tiene cinco decimales. El número 12,34 está expresado con cuatro dígitos y con sólo dos decimales.

Si la magnitud del error en $\tilde{a}$ no excede $\frac{1}{2}10^{-t}$, entonces se dice que $\tilde{a}$ tiene t **decimales correctos**. Los dígitos de $\tilde{a}$ que ocupan posiciones donde la unidad es mayor ó igual a 10^{-t} son llamados **dígitos significativos** (los ceros iniciales no se cuentan). Siguiendo esto, el número $0{,}001234 \pm 0{,}000004$ tiene cinco decimales correctos y tres dígitos significativos, mientras que $0{,}001234 \pm 0{,}000006$ tiene cuatro decimales correctos y sólo dos dígitos significativos. El número de decimales correctos da una idea de la magnitud del *error absoluto*, mientras que el número de cifras significativas da una idea aproximada de la magnitud del *error relativo*.

Existen dos formas de redondear un número x a t decimales. En el proceso denominado **truncamiento** (o redondeo hacia el cero) simplemente se eliminan todos los números ubicados a la derecha del t-ésimo decimal. La magnitud del error de este proceso puede ser tan grande como 10^{-t}. El **redondeo al más cercano** (también llamado *redondeo óptimo*), debe elegirse el número con s decimales que sea más cercano a x. Aquí, si p es la parte del número ubicada a la derecha del s-ésimo decimal el decimal t-ésimo permanece sin cambios si y sólo si $|p| < \frac{1}{2}10^{-s}$. En el otro caso, debe aumentarse el s-ésimo decimal en 1.

1.2. Sistemas Numéricos en PC

1.2.1. El sistema posicional

Para representar números, se utiliza a diario el sistema posicional con base 10, es decir el **sistema decimal**. Así, para representar los números se usan diez diferentes caracteres, y la magnitud con que el dígito a contribuye al valor del número depende de la posición de dicho dígito dentro del número. Si el dígito está ubicado a n pasos hacia la derecha del punto decimal, entonces su valor de contribución es de $a \cdot 10^{-n}$. Siguiendo esto, la secuencia de dígitos 4711,303 significa:

$$4 \cdot 10^3 + 7 \cdot 10^2 + 1 \cdot 10^1 + 1 \cdot 10^0 + 3 \cdot 10^{-1} + 0 \cdot 10^{-2} + 3 \cdot 10^{-3}.$$

Cada número real tiene una única representación de acuerdo al esquema presentado, excepto por la posibilidad de una secuencia infinita de nueves. Por ejemplo, el número con infinitos decimales $0{,}319999\ldots$ representa el mismo número que 0,32.

Es posible considerar otros sistemas posicionales con base diferente de 10. Cualquier número natural $\beta \geq 2$ puede ser usado como base. Es posible mostrar que todo número

real positivo a tiene una única representación de la forma[2]:

$$a = d_n\beta^n + d_{n-1}\beta^{n-1} + \ldots + d_1\beta^1 + d_0\beta^0 + d_{-1}\beta^{-1} + d_{-2}\beta^{-2} + \ldots,$$

ó en forma más compacta:

$$a = (d_n d_{n-1} + \ldots + d_1 d_0 d_{-1} d_{-2} \ldots)_\beta\,,$$

donde los coeficientes d_i, los *dígitos* en el sistema de base β, son enteros positivos tales que $0 \le d_i \le \beta - 1$.

Una de las grandes ventajas del sistema posicional es que se pueden dar reglas simples y generales para las operaciones aritméticas. Cuanto más pequeña es la base, más simples se vuelven las reglas. Ésta es una de las razones por las que la mayoría de las computadoras opera en base 2, el **sistema binario**. La adición y multiplicación tienen la forma:

$$0+0=0; \qquad 0+1=1+0=1; \qquad 1+1=10;$$

$$0\cdot 0=0; \qquad 0\cdot 1=1\cdot 0=0; \qquad 1\cdot 1=1;$$

En el sistema binario, el número *diecisiete* se escribe como 10001, ó en notación más compacta $(10001)_2 = (17)_{10}$ ya que:

$$1\cdot 2^4 + 0\cdot 2^3 + 0\cdot 2^2 + 0\cdot 2^1 + 1\cdot 2^0 = dieciseis + uno = diecisiete.$$

Los números se vuelven más largos de escribir en el sistema binario; los enteros *grandes* ocupan 3,3 más caracteres que su versión decimal. Por esto, N dígitos binarios son suficientes para representar enteros menores que $2^N = 10^{N\log_{10} 2} \approx 10^{N/3{,}3}$.

Dos sistemas de numeración muy utilizados en computación son las denominadas **base octal** y **base hexadecimal**. El sistema octal opera con los dígitos del 0 al 7; en el hexadecimal se utilizan los dígitos del 0 al 9 y las letras A, B, C, D, E y F que representan desde el *diez* hasta el *quince*.

Ejemplo 1.

$$(17)_{10} = (10001)_2 = (21)_8 = (11)_{16}$$
$$(13{,}25)_{10} = (1101{,}01)_2 = (15{,}2)_8 = (D{,}4)_{16}$$
$$(0{,}1)_{10} = (0{,}000110011001\ldots)_2 = (0{,}199999\ldots)_{16}$$

1.2.2. Conversión entre sistemas de numeración

Sea a un entero dado en un sistema de numeración con base α. Es posible determinar su representación en un sistema numérico con base β:

$$a = b_n\beta^n + b_{n-1}\beta^{n-1} + \ldots + b_0, \qquad 0 \le b_i < \beta \tag{1.1}$$

Los cálculos de conversión deben ser realizados en el sistema con base α para que también β esté expresado en esta representación. La conversión se realiza por divisiones sucesivas de a con β: Sea $q_0 = a$, y

$$q_k = q_{k+1}\beta + b_k, \qquad k = 0, 1, 2, \ldots, \tag{1.2}$$

donde q_{k+1} es el cociente y r_k el resto en la división entera. Si a no es un entero, debe escribirse $a = b + c$, donde b es la parte entera y

$$c = b_{-1}\beta^{-1} + b_{-2}\beta^{-2} + b_{-3}\beta^{-3} + \ldots \tag{1.3}$$

[2] con excepción de las secuencias de nueves antes mencionadas

es la parte fraccional, donde $b_1, b_2, \ldots$ deben ser determinados. Estos dígitos se obtienen de la parte entera en las multiplicaciones sucesivas de c con β: Sea $p_{-1} = c$, y

$$p_k\beta = b_k\beta + p_{k-1}, \qquad k = -1, -2, -3, \ldots. \tag{1.4}$$

Dado que una parte fraccionara finita en una base α generalmente no corresponde con una parte fraccionaria finita en otra base β, el redondeo en la conversión es necesariamente utilizado.

Ejemplo 2. *Convertir el número decimal* 176,524 *a la base ternaria* ($\beta = 3$). *Para la parte entera, se realizan las siguientes operaciones:* $176/3 = 58$ *con resto* 2; $58/3 = 19$ *con resto* 1; $19/3 = 6$ *con resto* 1; $6/3 = 2$ *con resto* 0 *y* $2/3 = 0$ *con resto* 2. *Por lo tanto,* $(176)_{10} = (20112)_3$. *Para la parte fraccionaria, se realizan las siguientes operaciones:* $0{,}524 \cdot 3 = 1{,}572$; $0{,}572 \cdot 3 = 1{,}716$; $0{,}716 \cdot 3 = 2{,}148$ *y así en forma consecutiva. Entonces una buena aproximación de* $(0{,}524)_{10}$ *es* $(0{,}112010222\ldots)_3$. *En este caso, un número cuya parte fraccionara es finita en base 10 se torna infinita en base 3.*

Comandos de EMT. *Los comandos para realizar cambios de base son:*

- `printbase(x:número, base=#, digits=#, integer=#)`, *donde* ***x*** *es un número en base 10;* ***base*** *es un entero que representa la base a la que se convertirá (valor por defecto: 16);* ***digits*** *es la cantidad de decimales del número cuasi-normalizado que se mostrarán luego de la conversión (valor por defecto: 13) y* ***integer*** *acepta los valores booleanos 0 y 1 para indicar si la salida mostrará la parte fraccionara ó no (valor por defecto: 0).*

- `baseinput(s:string, b=#)`, *donde* ***s*** *es un string de caracteres y* ***b*** *es la base a la que se convertirá el string (valor por defecto: 16).*

Ejemplo en EMT 1. *Convertir los números a las bases indicadas de acuerdo a lo solicitado:*

- $(187{,}2)_{10}$ *a las bases 2, 5 y 18. Usar 20 dígitos como máximo.*

```
>printbase(187.2,base=2,digits=20)
 1.01110110011001100110*2^7
>printbase(187.2,base=5,digits=20)
 1.22204444444444444444*5^3
>printbase(187.2,base=18,digits=20)
 A.73AE73AE73ACDAH480D1*18^1
```

- $(9782{,}2A)_{11}$ *a las bases 2, 4 y 16. Usar 25 dígitos como máximo.*

```
>num=baseinput("9782.2A",b=11)
 12916.2644628
>printbase(num,base=2,digits=25)
 1.1001001110100010000111011*2^13
>printbase(num,base=4,digits=25)
 3.0213101003230331112233212*4^6
>printbase(num,digits=25)
 3.27443B3D5AF9A000000000000*16^3
```

1.2.3. Punto fijo y punto flotante

Una computadora es construida para manejar piezas de información de tamaño fijo llamadas **palabras**. El número de dígitos en una palabra (usualmente cadenas binarias) define la **longitud de palabra** de la computadora. Entre las longitudes más comunes están 32, 48 ó 64 bits[3]. Los enteros pueden ser representados en forma exacta siempre que la longitud de la palabra sea suficiente para almacenar todos los dígitos necesarios para su representación.

En la primera generación de computadoras, los cálculos se realizaban a través del sistema de numeración de **punto fijo**. Esto es, números reales que se representan con una cantidad fija t de dígitos binarios. Si el largo de la palabra de la computadora es $s+1$ bits (incluyendo el bit del signo), entonces sólo es posible escribir números en el intervalo $I = [-2^{s-t}; 2^{s-t}]$. Algunas convenciones comunes de punto fijo son $t = s$ (convención de fracciones) ó $t = 0$ (convención de enteros). Esta limitación causó dificultades, porque cuando $x \in I$, $y \in I$, es posible que $x - y \notin I$ ó $x/y \notin I$. Para que un sistema de punto fijo tenga una implementación exitosa es necesario que todos los números, inclusive los resultados intermedios, permanezcan dentro de I. Esto puede ser logrado multiplicando las variables por **factores de escala** apropiados y luego transformando las ecuaciones de acuerdo a ellos, aunque es un proceso tedioso. Más aún, es complicado por el riesgo de que si los factores de escala no son escogidos cuidadosamente, ciertos resultados intermedios pueden tener demasiados ceros iniciales y afectar la precisión del resultado final. Como consecuencia de esto, la notación de punto fijo rara vez es utilizada para cálculos con números reales.

Por **representación normalizada de punto flotante** de un número real a, se entiende una expresión de la forma:

$$a = \pm m \cdot \beta^q, \qquad \beta^{-1} \leq m < 1, \qquad q \in \mathbb{Z}. \tag{1.5}$$

Alternativamente, la representación puede ser normalizada utilizando la condición $1 \leq m < \beta$. Es posible representar, de forma única, todos los números reales siempre que $a \neq 0$. La parte fraccional m es denominada **mantisa**, q es el **exponente** y β la **base**.

Dentro de la computadora, el número de dígitos para q y para m está limitado por la longitud de palabra del procesador. Suponiendo que p dígitos son usados para representar a m, entonces sólo es posible representar números de punto flotante de la forma:

$$\overline{a} = \pm \overline{m} \cdot \beta^e, \qquad \overline{m} = (.d_1 d_2 \ldots d_p)_\beta, \qquad 0 \leq d_i < \beta, \tag{1.6}$$

donde $\overline{m}$ es la mantisa m redondeada a p dígitos, y el exponente está limitado a un rango finito:

$$e_{min} \leq e \leq e_{max}. \tag{1.7}$$

Un sistema de punto flotante F se caracteriza a través de la base β, la precisión p (también llamada mantisa), y los números e_{min} y e_{max}. Sólo un conjunto finito F de números racionales puede ser representado en la forma (1.7). Los números de este conjunto son denominados **números de punto flotante**. Como $d_1 \neq 0$ este conjunto contiene exactamente $2(\beta - 1)\beta^{p-1}(e_{max} - e_{min} + 1) + 1$ elementos. La cantidad finita de dígitos en el exponente implica que a está limitada en magnitud a un intervalo llamado **rango** del sistema de punto flotante. Si a es mayor en magnitud que el número más grande del conjunto F, entonces a no puede ser representado y ocurre ***overflow***. Algo similar ocurre cuando, el número a representar es más pequeño que el menor número distinto de cero en F. En este caso el error se denomina ***underflow***.

[3]bits: *binary digits*

Ejemplo 3. *Sea el sistema de punto flotante donde* $\beta = 2$, $p = 3$, $e_{min} = -1$ *y* $e_{max} = 2$. *El conjunto* F *contiene exactamente* $2 \cdot 16 + 1 = 33$ *números. Para este ejemplo el número de magnitud más pequeña, distinto de cero, es* $(0{,}100)_2 \cdot 2^{-1} = \frac{1}{4}$ *y el de mayor magnitud es* $(0{,}111)_2 \cdot 2^2 = \frac{7}{2}$.

Ejercicio 1. *Desarrollar por extensión el conjunto* F *del ejemplo anterior.*

Es importante notar que los números de punto flotante no están equiespaciados en la recta numérica. La separación entre cada par de números es un factor de β para cada potencia de β. El espaciamiento entre los números de punto flotante se caracteriza por el **epsilon de máquina**, que es la distancia ϵ_M desde 1,0 al siguiente número mayor que él.

Incluso si los operandos en una operación aritmética son números de punto flotante de F, el resultado *exacto* de la operación puede no pertenecer a F. Por ejemplo, el producto exacto de dos números de punto flotante con p-dígitos tiene $2p$ ó $2p - 1$ dígitos.

Si un número real a está en el rango de un sistema de punto flotante, la forma obvia de representar a es $\overline{a} = fl(a)$, donde $fl(a)$ denota el número en F más cercano a a. Esto corresponde al redondeo de la mantisa, y de acuerdo a (1.6) se tiene:

$$|\overline{m} - m| \leq \frac{1}{2}\beta^{-p}. \tag{1.8}$$

Como $m \geq 0{,}1$ esto significa que la magnitud del error relativo en $\overline{a}$ es como máximo igual a:

$$\frac{|\overline{a} - a|}{|a|} = \frac{|(\overline{m} - m) \cdot \beta^e|}{|m \cdot \beta^e|} \leqslant \frac{\frac{1}{2}\beta^{-p} \cdot \beta^e}{m \cdot \beta^e} \leqslant \frac{1}{2}\beta^{-p+1}. \tag{1.9}$$

Teorema 1. *En un sistema de punto flotante* $F = F(\beta, p, e_{min}, e_{max})$ *cada número real dentro del rango de* F *puede ser representado con un error relativo, que no excede la* ***unidad de redondeo*** u, *que se define como:*

$$u = \begin{cases} \frac{1}{2}\beta^{-p+1}, \text{(redondeo)} \\ \beta^{-p+1}, \text{(truncamiento)} \end{cases} \tag{1.10}$$

1.3. Precisión y redondeo de errores

1.3.1. Aritmética de punto flotante

Es útil conocer la forma en que se *transmiten* los errores entre operaciones de punto flotante. Si x e y son dos números de punto flotante, entonces:

$$fl(x + y), \quad fl(x - y), \quad fl(x \cdot y), \quad fl(x/y)$$

son las operaciones de suma, resta, multiplicación y división en punto flotante, que la máquina almacena en memoria luego de redondear ó truncar. Se asume que se cumple el siguiente **modelo estándar** de aritmética:

Definición 2. *Asumiendo que tanto* x *como* y *pertenecen a* F, *entonces se cumple para cualquier operación:*

$$fl(x \diamond y) = (x \diamond y)(1 + \delta), \quad |\delta| \leq u,$$

donde u *es la unidad de redondeo y* $\diamond$ *es cualquiera de las operaciones elementales.*

A veces la computación en punto flotante es más precisa que lo que el modelo estándar asume. Un ejemplo obvio es cuando el valor exacto de $x \diamond y$ puede ser representado en un número de punto flotante que no contenga errores. Algunas computadoras pueden realizar una operación *fusionada* entre la suma y la multiplicación[4], donde una expresión del tipo $a \cdot x + y$ puede ser evaluada con una única instrucción y, en consecuencia, ocurre sólo un error de redondeo:

$$fl(a \cdot x + y) = (a \cdot x + y)(1 + \delta), \qquad |\delta| \leq u.$$

Esta operación *fusionada* representa una ventaja en muchos algoritmos. Por ejemplo, la forma de Horner de evaluar un polinomio de grado n, necesita sólo n operaciones fusionadas.

Es importante darse cuenta de que estas operaciones de punto flotante poseen, hasta cierto punto, propiedades distintas que las operaciones aritméticas exactas. Por ejemplo, la suma y multiplicación en punto flotante son conmutativas, pero no asociativas y la ley distributiva también falla para ambas. Esto torna al análisis de cálculos en punto flotante más difícil de realizar.

Ejemplo 4. *Sean los números a, b y c escritos en base 10 con una mantisa de 7 dígitos tales que:*

$$a = 0{,}1234567 \cdot 10^0; \qquad b = 0{,}4711325 \cdot 10^4; \qquad c = -b.$$

El siguiente esquema muestra cómo se resuelve la suma, asociando en forma distinta:

$$fl(a + fl(b + c)) = fl(0{,}1234567 \cdot 10^0 + 0) = 0{,}1234567 \cdot 10^0$$
$$fl(fl(a + b) + c) = fl(0{,}4711448 \cdot 10^4 + (-0{,}4711325 \cdot 10^4)) = 0{,}1230000 \cdot 10^0.$$

El caso mostrado en el ejemplo anterior tiene como punto central una resta que, debido a las diferencias de orden de magnitud de los números involucrados, devuelve un resultado poco preciso. En general, este tipo de operaciones se conocen como **cancelación catastrófica**.

1.3.2. Suma compensada

Para reducir los efectos de los errores de redondeo durante una suma $\sum_{i=0}^{n}$, es posible usar una **suma compensada**. En este algoritmo, el error de redondeo en cada adición es estimado y compensado a través de un término de corrección. La suma compensada puede ser muy útil cuando una gran cantidad de pequeños términos debe ser sumada, como por ejemplo en las cuadraturas numéricas ó en la solución numérica de ecuaciones diferenciales ordinarias. Es importante notar que los términos deben ser sumados en el mismo orden en el cual son generados. La compensación está basada en la posibilidad de simular la suma en doble precisión de punto flotante bajo una aritmética de simple precisión. Retomando el ejemplo 4, puede definirse:

$$s = fl(a + b) = 0{,}4711448 \cdot 10^4,$$

entonces el término de corrección es:

$$c = fl(fl(b - s) + a) = -0{,}1230000 \cdot 10^0 + 0{,}1234567 \cdot 10^0 = 0{,}4567000 \cdot 10^{-3}.$$

Ejercicio 2. *Desarrollar la estructura general del algoritmo de suma compensada y analizar su desempeño con:*

$$0{,}1234567 \cdot 10^0 + 0{,}6485130 \cdot 10^3 + 0{,}1010202 \cdot 10^2 - 0{,}1234567 \cdot 10^0 - 0{,}6485130 \cdot 10^3$$

[4] *fused multiply-add*

1.3.3. Evitando el *overflow*

En aquellos casos raros donde la entrada y la salida de datos son demasiado grandes ó demasiado pequeños en magnitud de forma tal que el rango de operación de la máquina no es suficiente, es posible utilizar mayor precisión ó también operar con logaritmos ó alguna transformación de los datos. Sin embargo, es necesario tener en mente el riesgo de que los resultados intermedios en un cálculo pueden producir un exponente extremo (*overflow* ó *underflow*) para el sistema de punto flotante de la máquina. Ocasionalmente, *errores inexplicables* en los datos de salida son la conclusión de algún desbordamiento ó supresión en los cálculos intermedios, casi siempre del tipo *underflow*.

El teorema de Pitágoras es de simple aplicación:

$$c = \sqrt{a^2 + b^2},$$

pero el posible *overflow* ó *underflow* puede ocurrir al elevar los valores de a y b al cuadrado, incluso si a, b y el resultado c están bien definidos dentro del rango del sistema de punto flotante utilizado. Este inconveniente puede ser evitado definiendo, para a y b distintos de cero, los valores:

$$p = \text{máx}\left\{|a|, |b|\right\}, \qquad q = \text{mín}\left\{|a|, |b|\right\}, \qquad \rho = \frac{q}{p}$$

y luego:

$$c = p\sqrt{1 + \rho^2}.$$

Precauciones similares a las consideradas para la suma pitagórica son necesarias para calcular la norma euclidiana de un vector:

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

En este caso, debe identificarse $x_{\max} = \text{máx}_{1 \le i \le n} |x_i|$ y luego:

$$s = \sum_{i=1}^{n} \left(\frac{x_i}{x_{\max}}\right)^2,$$

entonces:

$$\|\mathbf{x}\|_2 = x_{\max}\sqrt{s}.$$

El esquema presentado tiene un inconveniente: es necesario recorrer dos veces los datos. El algoritmo de Hammarling requiere un único paso por los datos:

- Hacer $t = 0$ y $s = 1$.
- Iterar n veces: si $|x_i| > t$ entonces $s = 1 + s\left(\frac{t}{x_i}\right)^2$ y $t = |x_i|$; sino $s = 1 + s\left(\frac{x_i}{t}\right)^2$.
- La norma euclidiana del vector $\mathbf{x}$ es $t\sqrt{s}$.

Ejercicio 3. *Implementar el algoritmo de Hammarling en EMT. Probar su funcionamiento con:*

$$\mathbf{x} = \left[3{,}14 \times 10^{180}; 6{,}55 \times 10^{181}; 4{,}16 \times 10^{182}\right].$$

1.3.4. Precisión de máquina

Se puede conocer en forma práctica la precisión de máquina que se está utilizando, equivalente a dar el valor del **epsilon de máquina**. Éste se define como el valor más pequeño que se le puede agregar a la unidad de forma tal que la máquina no devuelva un resultado distinto de la unidad. En los sistemas de punto flotante se opera habitualmente en base 10, aunque en forma interna se opera en base 2. Como la expresión de 0,1 en base 2 es infinita, entonces se utiliza esta propiedad para calcular el valor numérico del epsilon de máquina.

Existen dos algorimos principales para identificar el epsilon de máquina, el primero es:

- Hacer $a = 1$.
- Repetir (mientras $1 + a > 1$): $a = \frac{a}{2}$.
- El valor $2a$ coincide con el epsilon de máquina.

Mientras que el segundo es:

- Hacer $a = 0$.
- Iterar 10 veces: $a = a + 0{,}1$.
- El valor $2(1 - a)$ coincide con el epsilon de máquina.

Ejercicio 4. *Bajo EMT:*

1. *Implementar ambos códigos.*
2. *Calcular el epsilon de máquina.*
3. *¿Cuál de los dos es más eficiente desde el punto de vista operacional? ¿Por qué?*

Nota. *Dentro de* EMT *está definida la constante interna* `epsilon`*, que no representa el epsilon de máquina, sino que es un valor constante utilizado como tolerancia de terminación en los algoritmos iterativos.*

1.4. Propagación de errores y número de condición

En computación científica, es común que los datos de entrada de un problema sean imprecisos. Estos errores se propagan a través del proceso de cálculo y surgen errores en la salida. Para varios de los algoritmos más comunes de cálculo numérico el análisis del error se desarrolla de manera eficaz, permitiendo ajustar parámetros propios de cada algoritmo con el fin de obtener un resultado tan preciso como se desee[5]. El efecto de los errores se resume en las dos siguientes definiciones.

Definición 3. *En suma y resta, una cota para los errores absolutos en el resultado está dada por la suma de las cotas de los errores absolutos de los operandos:*

$$y = \sum_{i=1}^{n} \pm x_i, \qquad |\Delta y| \leq \sum_{i=1}^{n} |\Delta x_i|.$$

[5] ó tan preciso como la aritmética utilizada lo permita

Pero para obtener el resultado correspondiente para el error de propagación en multiplicación y división, es importante notar que para $y = \ln(x)$ se tiene que:

$$\Delta(\ln(x)) \approx \frac{\Delta(x)}{x},$$

es decir que **el error relativo de una cantidad es aproximadamente igual al error absoluto de su logaritmo natural**. Esto está relacionado al hecho de que desplazamientos de la misma longitud en diferentes lugares en una escala logarítmica, significan el mismo cambio relativo en el valor obtenido. A partir de esto se sigue la definición que falta.

Definición 4. *En multiplicación y división, una cota aproximada para el error relativo se obtiene a través de la suma de los errores relativos de los operandos. En forma más general, para* $y = x_1^{m_1} x_2^{m_2} \cdots x_n^{m_n}$*:*

$$\left|\frac{\Delta y}{y}\right| \leqq \sum_{i=1}^{n} |m_i| \left|\frac{\Delta x_i}{x_i}\right|.$$

Para estudiar la propagación de errores en una forma más general, se centrará el estudio en expresiones no lineales. Si se tiene una función de variable simple y real $y = f(x)$, ¿cuál es el error de x propagado a y? Sea $\tilde{x} - x = \Delta x$, entonces una forma *natural* de aproximar $\Delta y = \tilde{y} - y$ es con el diferencial de y. Por el teorema del valor medio:

$$\Delta y = f(x + \Delta x) - f(x) = f'(\xi)\Delta x,$$

donde ξ es un número entre x y $x + \Delta x$. Suponiendo que $|\Delta x| \leq \varepsilon$, se sigue que:

$$|\Delta y| \leq \max_{\xi} |f'(\xi)|\varepsilon, \qquad \xi \in [x - \varepsilon, x + \varepsilon]. \tag{1.11}$$

En la práctica, usualmente es suficiente reemplazar ξ con el valor estimado de x. Por el teorema de la función implícita se obtiene un resultado similar si y es una función implícita de x denotada por $g(x, y) = 0$. Si $\frac{\partial g}{\partial y} \neq 0$, entonces:

$$|\Delta y| \leq \max_{\xi} \left|\frac{\frac{\partial g}{\partial x}(\xi)}{\frac{\partial g}{\partial y}(\xi)}\right| \varepsilon, \qquad \xi \in [x - \varepsilon, x + \varepsilon]. \tag{1.12}$$

Ejemplo 5. *Calcular las cotas de error para* $f = x_1^2 - x_2$*, donde* $x_1 = 1{,}03 \pm 0{,}01$ *y* $x_2 = 0{,}45 \pm 0{,}01$*. Entonces:*

$$\left|\frac{\partial f}{\partial x_1}\right| = |2x_1| \leq 2{,}1; \qquad \left|\frac{\partial f}{\partial x_2}\right| = |-1| = 1,$$

y por lo tanto $|\Delta f| \leq 2{,}1 \cdot 0{,}01 + 1 \cdot 0{,}01 = 0{,}031$*. Esto nos lleva a:*

$$f = 1{,}06 - 0{,}450 \pm 0{,}031 = 0{,}610 \pm 0{,}031.$$

Es normal que en un algoritmo los errores iniciales se propaguen con las operaciones aritméticas realizadas. Si el algoritmo que opera aritméticamente produce pequeños errores en la salida a partir de pequeños errores de entrada, el algoritmo se denomina **estable**. En cambio, si en la salida se obtienen grandes errores el algoritmo se denomina **inestable**.

1.4.1. Número de condición

Es útil tener, para un problema dado, una idea de cuán sensible son los datos de salida con respecto a variaciones en los datos de entrada. En general, si pequeños cambios en los datos de entrada resultan en *grandes* cambios en los datos de salida, se dice que el problema está **mal condicionado**, si esto no ocurre, el problema está **bien condicionado**.

Nota. *La definición de* grande *puede diferir de un problema a otro, dependiendo de la precisión de los datos y la precisión necesaria en la solución.*

Ya se mostró que $|f'(x)|$ puede ser interpretado como una **medida de la sensibilidad de $f(x)$ con respecto a una perturbación Δx de x**. En la mayoría de los contextos, la proporción de la perturbación relativa en $f(x)$ y x proporciona más información al respecto.

Definición 5. *Asumiendo que $x \neq 0$ y $f(x) \neq 0$, entonces el **número de condición** $\mathcal{K}$ para el problema numérico de calcular $y = f(x)$ es:*

$$\mathcal{K} = \lim_{|\Delta x| \to 0} \frac{\dfrac{|f(x+\Delta x) - f(x)|}{|f(x)|}}{\dfrac{|\Delta x|}{|x|}} = \left| \frac{x f'(x)}{f(x)} \right|, \tag{1.13}$$

que indica, para un valor de $\mathcal{K}$ grande*, que la función f está mal condicionada en los alrededores de x.*

Es importante recalcar que el número de condición es una propiedad de un problema numérico y no depende del algoritmo utilizado en la resolución. Un problema mal condicionado es intrínsecamente difícil de resolver en forma precisa utilizando cualquier algoritmo numérico. Incluso si los datos de entrada son exactos, los errores de redondeo generados durante los cálculos en aritmética de punto flotante pueden causar grandes perturbaciones en el resultado final. No debe confundirse un algoritmo inestable con un problema mal condicionado, aunque sea difícil operar con ambos.

Ejemplo 6. *Considerar el sistema lineal:*

$$\begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

donde $\alpha \neq 1$ es el dato de entrada. La solución exacta es:

$$x = \frac{1}{1-\alpha^2}; \qquad y = \frac{-\alpha}{1-\alpha^2}.$$

La matriz del sistema es singular para $\alpha = 1$, y el problema de calcular x e y es mal condicionado cuando $\alpha \approx 1$. Utilizando la ecuación (1.13), el número de condición para calcular x es:

$$\mathcal{K} = \frac{\alpha x'(\alpha)}{x(\alpha)} = \frac{2\alpha^2}{|1-\alpha^2|}.$$

Para $\alpha = 0{,}9950$, y por medio del algoritmo de eliminación de Gauss con una aritmética de 4 dígitos, se obtiene:

$$\tilde{y} = \frac{-0{,}9950}{1 - 0{,}9900} = -99{,}50$$
$$\tilde{x} = 1 + 0{,}9950 \cdot 99{,}50 = 100{,}0,$$

en lugar de los valores correctos[6] *$y = -99{,}7494$, $x = 100{,}2506$. El número de condición $\mathcal{K} = 198$ indica que es esperable perder algunos dígitos significativos en la salida.*

[6] excediendo la precisión de la aritmética utilizada

El tener un problema bien condicionado no implica que la aritmética utilizada no genere perturbaciones en los datos de salida. Para problemas bien condicionados, incluso con doble precisión, la interacción de las fuentes de error es un ítem importante a considerar.

Ejemplo 7. *El polinomio* $f_1(x) = (x-1)^5$ *posee un número de condición alto en cercanías de* $x = 1$*:*

$$\mathcal{K}(f_1) = \left|\frac{x5(x-1)^4}{(x-1)^5}\right| = \left|\frac{5x}{x-1}\right|,$$

y el gráfico de la función del número de condición se muestra como figura 1.1. En cambio, su expresión expandida es:

$$f_2(x) = x^5 - 5x^4 + 10x^3 - 10x^2 + 5x - 1,$$

y su número de condición se calcula como:

$$\mathcal{K}(f_2) = \left|\frac{x\left(5x^4 - 20x^3 + 30x^2 - 20x + 5\right)}{x^5 - 5x^4 + 10x^3 - 10x^2 + 5x - 1}\right|$$

de forma que su gráfico es la figura 1.2. De los gráficos, parece mejor condicionada la expresión de $f_2(x)$ *que* $f_1(x)$ *en* $x \approx 1$*. Pero al graficar los polinomios* $f_1(x)$*, en color azul, y* $f_2(x)$*, en color rojo de la figura 1.3, se observa un comportamiento diferente.*

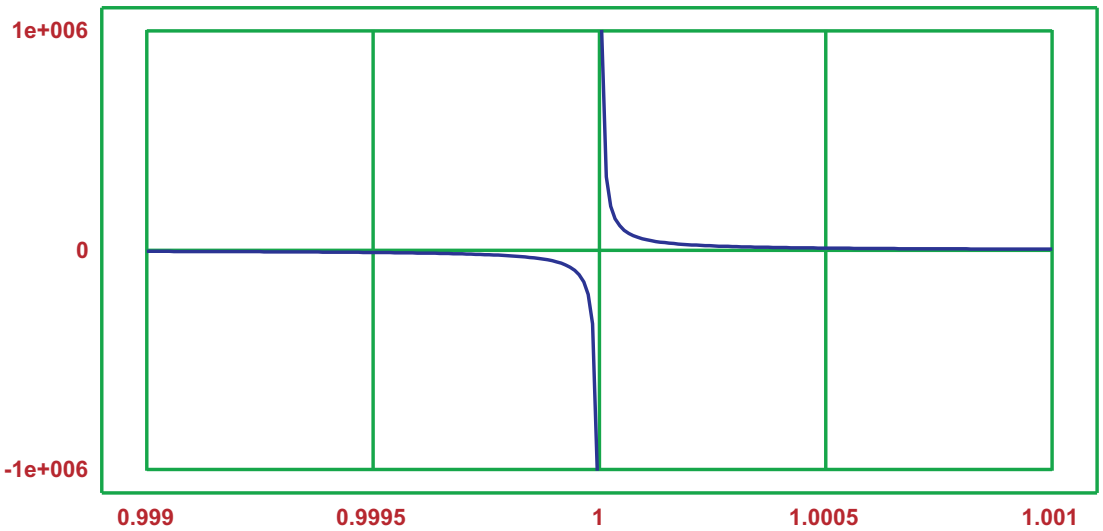

Figura 1.1: Gráfico de $\mathcal{K}(f_1)$ en $[0{,}999; 1{,}001]$.

Ejercicio 5. *Dado los polinomios equivalentes:*

$$f_1(x) = (2-x)^3; \qquad f_2(x) = -x^3+6x^2-12x+8; \qquad f_3(x) = x(x(-x+6)-12)+8,$$

calcular con una aritmética de 4 dígitos y truncamiento $f_i(-3{,}12)$ *y* $f_i(1{,}97)$*. ¿Qué se observa?*

1.5. Ejercicios

1. Construir los siguientes algoritmos en PC:

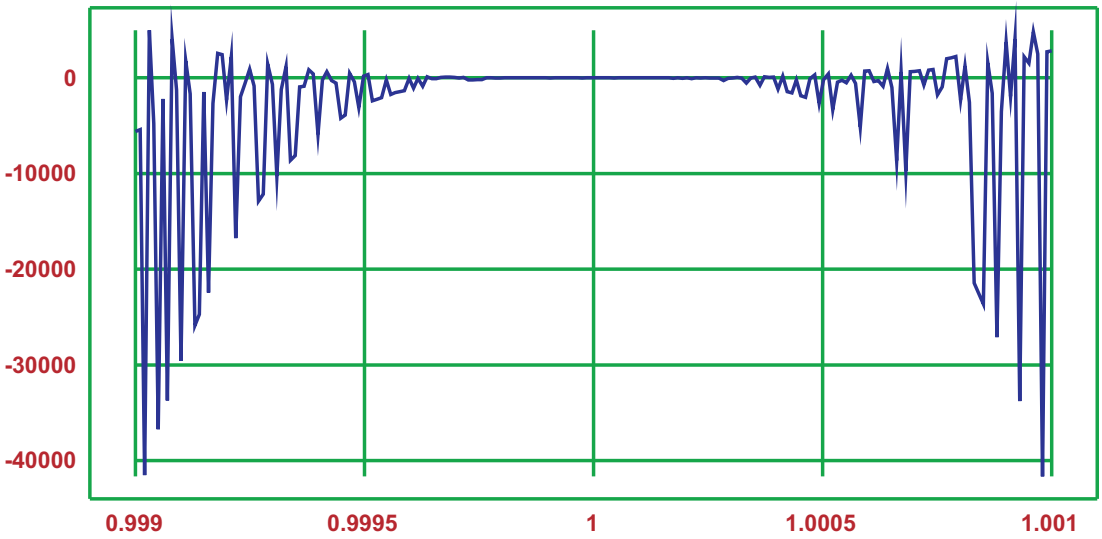

Figura 1.2: Gráfico de $\mathcal{K}(f_2)$ en $[0{,}999; 1{,}001]$.

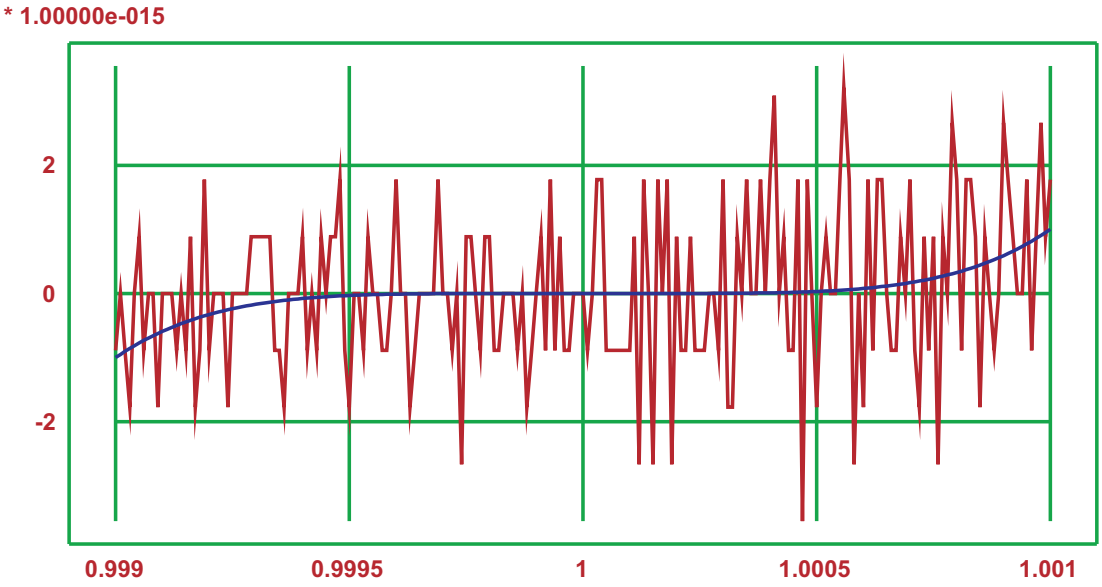

Figura 1.3: Gráficos de $f_1(x) = (x-1)^5$, en color azul, y $f_2(x) = x^5 - 5x^4 + 10x^3 - 10x^2 + 5x - 1$, en color rojo.

a) **Aritmética de punto flotante con truncamiento**. Entrada: un número decimal; la cantidad de dígitos a mantener, k. Salida: un número con k dígitos.

b) **Conversión decimal a binario**. Entrada: un número decimal entero. Salida: una cadena binaria (*string*) que represente al número antes ingresado. Opcional: permitir la conversión de números no enteros.

c) **Conversión base 5 a base 7**. Entrada: un número en base 5. Salida: un número en base 7.

2. Dados $a = 1, b = 10^{-16}, c = 10^{-16}$ comparar los resultados de las siguientes operaciones en *EMT*, utilizando la salida máxima de decimales por pantalla:

$$(c + b + a) - (a + b + c)$$

y

$$c + b + a - a - b - c.$$

3. El área de un triángulo T con lados a, b y c se calcula con la *fórmula de Herón*:

$$A(T) = \sqrt{p(p-a)(p-b)(p-c)},$$

donde p es el semiperímetro del triángulo. Mostrar que en el caso de triángulos deformados ($a \approx b + c$) esta fórmula pierde precisión.

4. ¿Qué constante de máquina se obtiene al realizar la siguiente operación sobre PC[7]?

$$|3 \cdot (4/3 - 1) - 1|.$$

5. [*EMT*] Es posible ver que el error asociado a sumas y restas dentro de una PC puede computarse experimentalmente como

$$((a + b) - b) - a,$$

verificar qué ocurre al efectuar esas operaciones con:

- $a = 3$; $b = 0{,}1$
- $a = \sqrt{2}$; $b = \pi$
- $a = 1/3$; $b = 0{,}1$
- $a = 0{,}1$; $b = 1/3$

y la máxima precisión de pantalla en *EMT*.

6. ¿A qué conclusión, ya explicada anteriormente, puede llegarse al analizar los incisos c) y d) del ejercicio anterior?

7. Aplicar el **algoritmo de Horner** a los polinomios dados a continuación. Calcular la cantidad de operaciones en punto flotante antes y después de anidarlos, en caso de evaluarlos en $x = x_0$.

a) $P_1(x) = 3x^2 - 5x + 2 - 4x^3$

b) $P_2(x) = \left(\frac{7}{2}x + 1\right)^2 - 2x^3$

c) $P_3(x) = x^4 - 5 - 3x^2 - 9x$

[7] Probar en *Euler Math* con diferentes configuraciones de salida y *Maxima*

8. Dados los siguientes números, convertirlos a base 10: 1323_4; 203_7; 463_9 y $1B59_{12}$.

9. Al número decimal 37191, convertirlo a base 6, 8, 11 y 15.

10. Convertir al número decimal e, en base 2, teniendo en cuenta la aritmética deseada:

    a) PF(10,4,2,R) en PF(2,4,2,R).
    b) PF(10,6,2,T) en PF(2,8,2,T).
    c) PF(10,5,2,T) en PF(2,10,2,T).
    d) PF(10,5,2,R) en PF(2,10,2,R).

11. Calcular el error absoluto cometido en la conversión de base 10 a base 2 del ejercicio anterior.

12. Estimar en qué valores de abscisa las funciones:

    a) $f_1(x) = \dfrac{\sin(x)}{\cos(x) + 1{,}01}$, $f_1 : [0, 20] \to \mathbb{R}$.
    b) $f_2(x) = \dfrac{3x^2 - 2x + 1}{5x^2 - 2}$, $f_2 : [-5, 5] \to \mathbb{R}$.
    c) $f_3(x) = \dfrac{1}{2x}$, $f_3 : (0, 5) \to \mathbb{R}$.
    d) $f_4(x) = \sqrt{x-1} - \sqrt{x+1}$, $f_4 : [1, 10] \to \mathbb{R}$.

    pueden considerarse mal condicionadas.

13. Generar un algoritmo en *EMT* que grafique el número de condición de funciones dentro de un intervalo preestablecido.

14. El número de condición de la función $f(x) = x^\alpha$ es constante e independiente del valor de x. ¿Por qué ocurre esto?

15. Analizar la estabilidad, con respecto a la propagación de errores, de las siguientes dos expresiones para calcular $f(x) = (e^x - 1)/x$ con $|x| < 1 \times 10^{-7}$:

    - `if x==0 then f=1, else f=(exp(x)-1)/x, endif`
    - `y=exp(x); if y==1 then f=1, else f=(y-1)/log(y), endif`

16. Graficar el resultado que se obtiene al restar las dos expresiones anteriores.

17. Sea $y = \sqrt{2} - 1$. En forma equivalente, se puede escribir como $y = \left(\sqrt{2} + 1\right)^{-1}$. Analizando los números de condición, establecer cuál de las dos formas es más sensible al error cuando $f(2)$ se aproxima a través de un número en punto flotante.

18. La expresión $x^2 - y^2$ exhibe **cancelación catastrófica** si $|x| \approx |y|$. Mostrar que es más apropiado evaluar la expresión original como $(x + y)(x - y)$.

19. Considerar la identidad trigonométrica $\sin^2(x) + \cos^2(x) = 1$ para calcular:

$$\cos(x) = \sqrt{1 - \sin^2(x)}.$$

    ¿Para qué argumentos del intervalo $0 \leq x \leq \pi/4$ esta fórmula presenta la peor precisión? Analizar con el número de condición ó bien tomando la diferencia entre las dos expresiones equivalentes, con una partición mínima de 3000 elementos.

20. Si se aplica la conocida fórmula de Baskhara para resolver la ecuación cuadrática $ax^2 + bx + c = 0$ es posible obtener problemas de precisión numérica, de acuerdo a la aritmética utilizada, si los órdenes de magnitud de b^2 y $4ac$ difieren en forma significativa.

    a) Mostrar que esto ocurre cuando $a = 1{,}00$, $b = 50{,}1$ y $c = 0{,}100$.

    b) Proponer otra forma de resolver este problema, sabiendo que la ecuación cuadrática puede escribirse como $a(x - r_1)(x - r_2) = 0$, donde r_1 y r_2 son las raíces buscadas.

21. Sugerir alguna forma de calcular:

$$f(x) = \frac{e^x - 1}{x}$$

    en cercanías de $x = 0$ cuando se opera con alguna aritmética reducida[8]. Calcular $f(0{,}0001)$, con una mantisa de tres dígitos, con la fórmula original y luego con la sugerencia. Comparar resultados.

22. Calcular xy/z es un proceso poco preciso. De acuerdo a la literatura, se sugiere:

    - $(xy)/z$, cuando x e y son muy diferentes en magnitud;
    - $x(y/z)$, cuando y y z son cercanos en magnitud;
    - $(x/z)y$, cuando x y z son cercanos en magnitud.

    Probar esto con una aritmética reducida y valores convenientes de x, y y z.

23. [*EMT*] Se quiere calcular el valor de:

$$\frac{300^{125}}{125!}e^{-300},$$

    con lo que se ingresa `(300^125/125!)*exp(-300)` en *EMT*, pero devuelve un error.

    a) ¿Qué tipo de error de desbordamiento es la salida del comando anterior?

    b) ¿Será de ayuda cambiar el orden de las operaciones?

    c) Como alternativa, crear una rutina que evalúe la expresión:

$$p(k) = \frac{\lambda^k}{k!}e^{-1},$$

    para $\lambda = 300$ y $k \in \mathbb{Z}$ en forma recursiva, es decir que:

$$p(k+1) = p(k)\frac{\lambda}{k}\frac{1}{e}.$$

    d) Utilizar la rutina creada en el inciso anterior y calcular el valor pedido.

24. [*EMT*] Es sabido que la unidad de redondeo está relacionada de manera especial con el epsilon de máquina. Si, para los procesadores numéricos se cumple que:

$$u = \frac{\epsilon_M}{2},$$

    entonces es posible deducir cuántos dígitos binarios se utilizan en la mantisa para la representación numérica en *EMT*. Si el primer *bit* se reserva para el signo, los intermedios para la mantisa y los restantes para la potencia; además, en forma interna, *EMT* utiliza redondeo en vez de truncamiento:

[8] Sugerencia: serie de Taylor

a) ¿Cuántos dígitos binarios se utilizan para la mantisa?

b) ¿Cuántos para el exponente?

25. Sean los polinomios[9]

$$f(x) = \prod_{k=1}^{20}(x-k) = (x-1)(x-2)\cdots(x-20)$$
$$g(x) = x^{20},$$

Las raíces de $f(x)$ son los enteros $1, 2, \ldots, 20$. ¿Cuán sensible es la raíz $x_0 = 20$ cuando se perturba $f(x)$ de la forma $f(x) + \varepsilon g(x)$ Sugerencia: generar una serie de Taylor truncada y analizar para h pequeño.

Bibliografía

- *A theoretical introduction to numerical analysis**, V. RYABEN'KII y S. TSYNKOV, Cap.1
- *An introduction to numerical analysis*, Kendall ATKINSON, Cap.1
- *Análisis numérico*, R. BURDEN y J. FAIRES, Cap.1
- *Análisis numérico - Primer curso*, Hernán GONZÁLEZ, Cap.1
- *Numerical mathematics**, A. QUARTERONI, R. SACCO y F. SALERI, Cap.2
- *Numerical methods*, G. DAHLQUIST y A. BJÖRK, Cap.2

[9] ejemplo clásico desarrollado originalmente por Wilkinson

2 Resolución de Ecuaciones No Lineales

En este capítulo se considera el problema de encontrar x que permita resolver la ecuación:

$$f(x) = 0, \tag{2.1}$$

para una f arbitraria, pero de forma tal que $f(x) \in \mathbb{R}$. De esta forma, cualquier valor x que verifique la ecuación (2.1) se denomina **raíz** de la función. Se desarrollarán varios métodos de solución, tales como *bisección*, *punto fijo* y *Newton-Raphson* (junto con sus modificaciones). Todos ellos son iterativos y construyen una sucesión de puntos que, se espera, converjan a la raíz. Sin embargo estos procesos iterativos pueden fallar, en un proceso llamado *breakdown*. Es decir que la sucesión numérica obtenida puede oscilar, diverger ó mostrar un comportamiento caótico.

Todos los métodos iterativos requieren un valor inicial para comenzar la sucesión, generalmente una estimación de la raíz. Esta estimación inicial es crucial, ya que una mala elección puede generar una sucesión que no converge, ó puede ser convergente a una raíz *incorrecta*. Una de las formas más comunes de elegir esta estimación inicial es por medio de un gráfico.

2.1. Método de Bisección

Es el método más simple de aplicar, con convergencia segura y se basa en el siguiente teorema:

Teorema 2 (del Valor Intermedio)**.** *Si $f : [a, b] \to \mathbb{R}$ es continua en el intervalo cerrado y existe $y_0 \in \mathbb{R}$ tal que $f(a) \leq y_0 \leq f(b)$, entonces existe $x_0 \in [a, b]$ tal que $f(x_0) = y_0$. En decir que una función continua en un intervalo cerrado $[a, b]$ toma todos los valores entre $f(a)$ y $f(b)$ como mínimo una vez.*

Ahora, suponiendo que la función f, continua, está definida en $[a, b]$, y se verifica que $f(a) \cdot f(b) < 0$, entonces por el teorema 2 se asegura la existencia de una raíz p en el intervalo antes mencionado. Por lo tanto se dice que
$[a, b] = [a_0, b_0]$ contiene a la raíz de la función f. Ahora:

$$p_0 = \frac{a_0 + b_0}{2}, \tag{2.2}$$

es el punto medio del intervalo $[a, b]$. Existen tres posibilidades:

1. Si $f(p_0) = 0$, entonces $p = p_0$ y la raíz ya fue hallada. Termina el algoritmo.
2. Si $f(a) \cdot f(p_0) < 0$, entonces $p \in [a, p_0]$ y se define $[a_1, b_1] = [a, p_0]$. De esta forma se asegura un intervalo más pequeño donde está la raíz y se sigue iterando.

3. Si $f(p_0) \cdot f(b) < 0$, entonces $p \in [p_0, b]$ y se define $[a_1, b_1] = [p_0, b]$. De esta forma se asegura un intervalo más pequeño donde está la raíz y se sigue iterando.

Este proceso se repite considerando el punto medio del nuevo intervalo:

$$p_1 = \frac{a_1 + b_1}{2},$$

obteniendo una de las tres posibilidades antes mencionadas y así en adelante. Lo que en realidad se hace es construir una secuencia de números $\{p_n\} = p_0, p_1, p_2, \ldots$ de forma tal que:

$$\lim_{n\to\infty} p_n = p,$$

donde p_n es el punto medio del intervalo $[a_n, b_n]$ y $f(p) = 0$. La justificación formal de este proceso, es decir la convergencia hacia un único p, está dada por el siguiente teorema.

Teorema 3 (de Intersección de Cantor)**.** *Sea* $\{[a_k, b_k]\}$ *una secuencia de intervalos cerrados y encajados de forma tal que:*

$$[a_0, b_0] \supset [a_1, b_1] \supset \cdots \supset [a_n, b_n] \supset \cdots,$$

además:

$$\lim_{n\to\infty} (b_n - a_n) = 0.$$

Entonces existe un único punto $p \in [a_n, b_n]$ *para todo* $n \in \mathbb{N}_0$ *de forma tal que:*

$$\bigcap_{n=0}^{\infty} [a_n, b_n] = \{p\}$$

El método de bisección produce la sucesión $\{p_n\}$ de forma tal que $a_n < p_n < b_n$ y $p \in [a_n, b_n]$ para todo $n \in \mathbb{N}_0$. Consecuentemente, como $p_n = \frac{a_n + b_n}{2}$ entonces:

$$|p_n - p| \leq |b_n - a_n| \leq \frac{b - a}{2^n},$$

para todo $n \in \mathbb{N}_0$, asi $\lim_{n\to+\infty} p_n = p$. Si, además se considera que f es continua en $[a, b]$, entonces $\lim_{n\to\infty} f(p_n) = f(p)$. Por lo tanto:

$$|p_n - a_n| \leq \frac{1}{2^n} |b - a|$$

y

$$|b_n - p_n| \leq \frac{1}{2^n} |b - a|.$$

De esta manera y aplicando la desigualdad triangular:

$$|x - y| = |(x - z) - (y - z)| \leq |x - z| + |y - z|,$$

se obtiene:

$$|p - a_n| \leq |p - p_n| + |p_n - a_n| \leq \frac{1}{2^n} (b - a) + \frac{1}{2^n} (b - a) = \frac{1}{2^{n-1}} (b - a).$$

De la misma forma:

$$|p - b_n| \leq |p - p_n| + |p_n - b_n| \leq \frac{1}{2^{n-1}} (b - a).$$

Así:

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = p.$$

En cada paso, el nuevo intervalo sigue conteniendo a la raíz p. Esto implica que existe una subsecuencia de $\{p_n\}$ denotada $\{x_n\}$ que es convergente a p de forma tal que $f(x_n) > 0$ para todo $n \in \mathbb{N}_0$. De la misma manera, existe una subsecuencia $\{y_n\}$ convergente a p de forma tal que $f(y_n) < 0$. Así:

$$f(p) = \lim_{n\to\infty} f(x_n) \geq 0$$

y

$$f(p) = \lim_{n\to\infty} f(y_n) \leq 0,$$

lo que implica que $f(p) = 0$. Se puede concluir que el método de bisección produce una secuencia $\{p_n\}$ que converge a p de forma tal que $f(p) = 0$. Por lo tanto, el método de bisección tiene convergencia segura.

Ahora que se demostró que el método converge, se debe analizar cuándo terminar este proceso. En teoría, termina cuando se obtiene el caso 1 de los antes mencionados en la obtención de p_n. En la práctica, esto es casi imposible debido a los efectos de los errores de redondeo y la aritmética finita. Es necesario entonces un criterio práctico para detener el proceso. Se procederá a analizar algunas posibilidades:

$$\left|\frac{1}{2}(b_n - a_n)\right| < \varepsilon, \tag{2.3}$$

$$|p_n - p| < \varepsilon, \tag{2.4}$$

$$f(p_n) < \varepsilon, \tag{2.5}$$

$$\left|\frac{p_n - p_{n-1}}{p_n}\right| < \varepsilon \qquad (p_n \neq 0). \tag{2.6}$$

Se debería iterar hasta que las inecuaciones se satisfagan. Si bien las cuatro opciones mencionadas son válidas bajo ciertas condiciones, es posible analizarlas más a fondo para elegir la óptima. La condición (2.3) no es muy buena como criterio de parada puesto que depende del tamaño del *n-ésimo* intervalo, mientras que la precisión en la estimación de p es lo más importante. La condición (2.4) requiere conocer p, lo que no es razonable puesto que el valor de p es lo que se está tratando de determinar. La condición (2.5) está basada en $f(p_n)$, y nuevamente el interés se centra en cómo p_n aproxima p. Por lo tanto, la mejor opción es la (2.6). Esta condición indica que el algoritmo termina cuando el valor de p_n es lo suficientemente cercano al valor de p_{n-1}.

En la figura 2.1 se muestra el intervalo inicial $[a_0, b_0]$ y el punto medio de dicho intervalo: p_0. En la figura 2.2 continúa el proceso iterativo, donde $[a_0, p_0] = [a_1, b_1]$ y p_1 es el punto medio del $[a_1, b_1]$. Nuevamente, la raíz queda encerrada dentro del nuevo intervalo. Este proceso continúa hasta que se verifica la condición (2.6) y se obtiene una aproximación aceptable a la raíz.

Denotando $\epsilon_k = |p_k - p|$ el error absoluto cometido en el paso k, se sigue que $|\epsilon_k| \leq (b-a)/2^k, k \geq 0$, lo que implica $\lim_{k\to\infty} |\epsilon_k| = 0$. El método de bisección tiene por tanto *convergencia global*. Más aún, para lograr $|p_m - p| \leq \varepsilon$, se debe tener en cuenta que:

$$m \geq \log_2(b-a) - \log_2(\varepsilon) = \frac{\log((b-a)/\varepsilon)}{\log(2)} \simeq \frac{\log((b-a)/\varepsilon)}{0{,}6931}$$

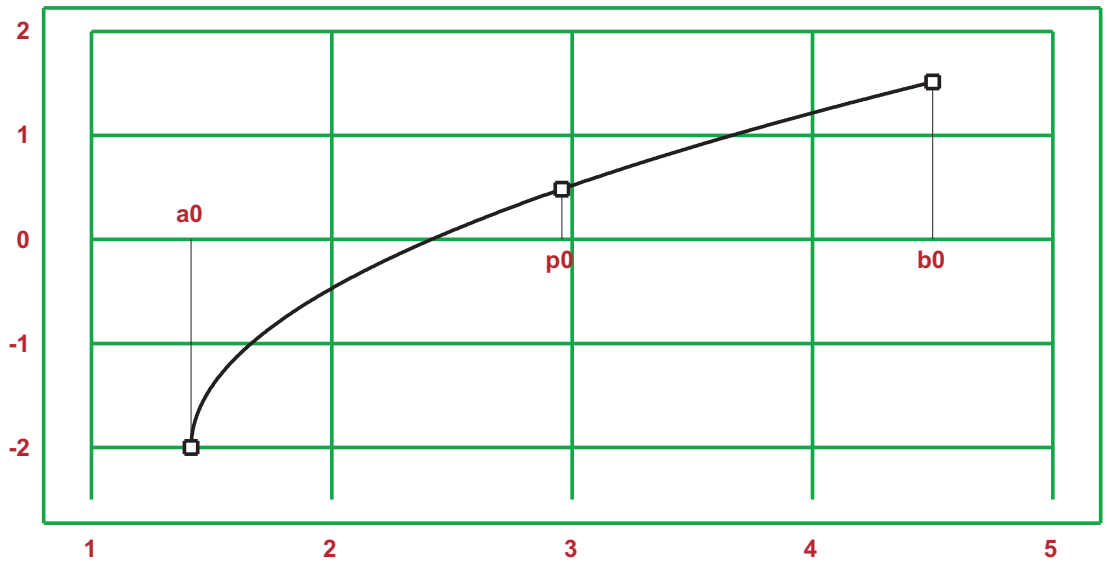

Figura 2.1: Primera iteración genérica del método de bisección.

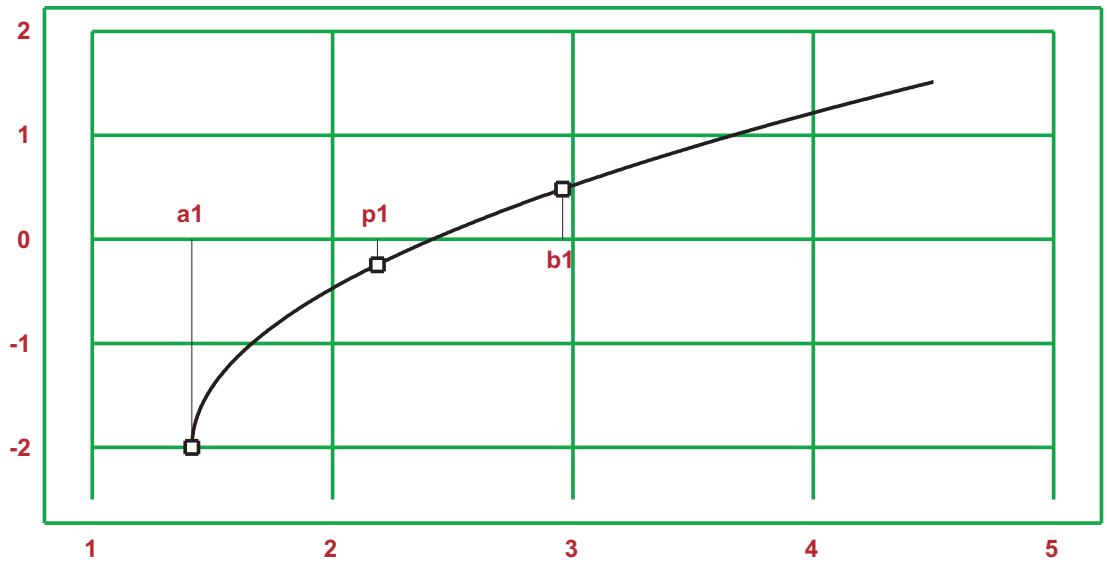

Figura 2.2: Segunda iteración genérica del método de bisección.

i	a_i	b_i	p_i	ϵ_a	ϵ_r
0	2,0000000	4,0000000	3,0000000	-	-
1	3,0000000	4,0000000	3,5000000	5,00E-01	1,43E-01
2	3,0000000	3,5000000	3,2500000	2,50E-01	7,69E-02
3	3,2500000	3,5000000	3,3750000	1,25E-01	3,70E-02
4	3,2500000	3,3750000	3,3125000	6,25E-02	1,89E-02
5	3,3125000	3,3750000	3,3437500	3,13E-02	9,35E-03
6	3,3125000	3,3437500	3,3281250	1,56E-02	4,69E-03
7	3,3125000	3,3281250	3,3203125	7,81E-03	2,35E-03
8	3,3125000	3,3203125	3,3164062	3,91E-03	1,18E-03
9	3,3164062	3,3203125	3,3183593	1,95E-03	5,89E-04
10	3,3183593	3,3203125	3,3193358	9,77E-04	2,94E-04
11	3,3193358	3,3203125	3,3198241	4,88E-04	1,47E-04
12	3,3198241	3,3203125	3,3200683	2,44E-04	7,36E-05
13	3,3198241	3,3200683	3,3199461	1,22E-04	3,68E-05
14	3,3199461	3,3200683	3,3200071	6,10E-05	1,84E-05
15	3,3199461	3,3200071	3,3199766	3,05E-05	9,19E-06
16	3,3199766	3,3200071	3,3199918	1,52E-05	4,58E-06
17	3,3199918	3,3200071	3,3199994	7,60E-06	2,29E-06

Tabla 2.1: Salida del algoritmo de bisección del ejemplo 8.

Ejemplo 8. *La función* $f(x) = x^2 - 2x + \cos(x+1) - 4$ *posee dos raíces en el intervalo* [-2; 4] *, denominadas* r_1 *y* r_2*, donde* $r_1 < r_2$*. Utilizando el método de bisección es posible obtener ambas, sin embargo, se calculará sólo el valor de* r_1*. Ahora* $f(2) < 0$*,* $f(4) > 0$ *y como* f *es continua en* $[2, 4]$ *se asegura la existencia de una raíz en ese intervalo. La tabla 2.1 muestra la aplicación del método con un valor de tolerancia de* $\varepsilon = 1 \times 10^{-5}$ *en error absoluto para la sucesión convergente y las figuras (2.3) y (2.4) muestran la convergencia hacia la raíz y los errores. Es de notar que en escala logarítmica, se visualiza como constante el descenso de los errores, tanto absoluto como relativo.*

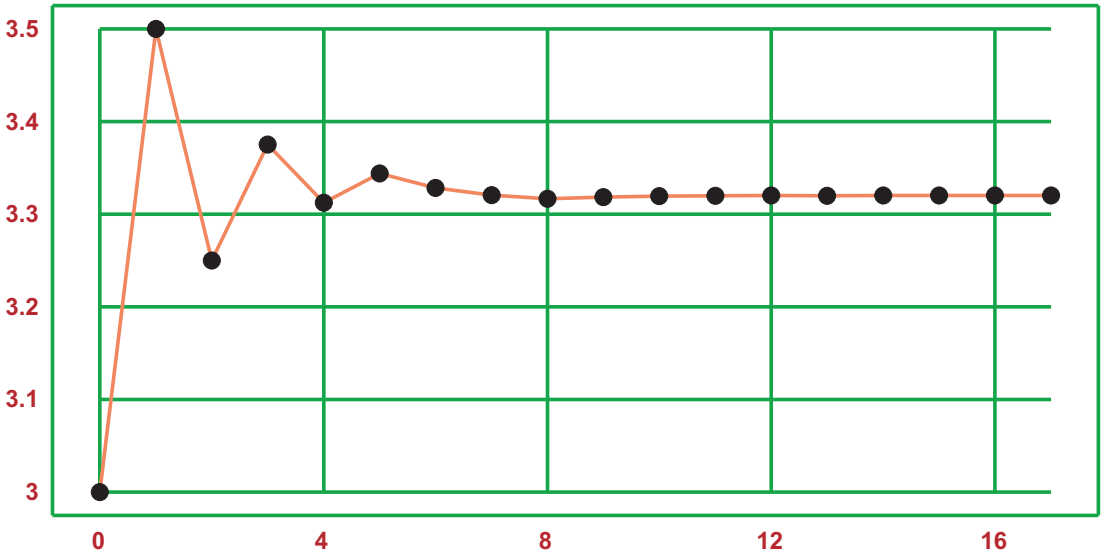

Figura 2.3: Convergencia del método de bisección del ejemplo 8.

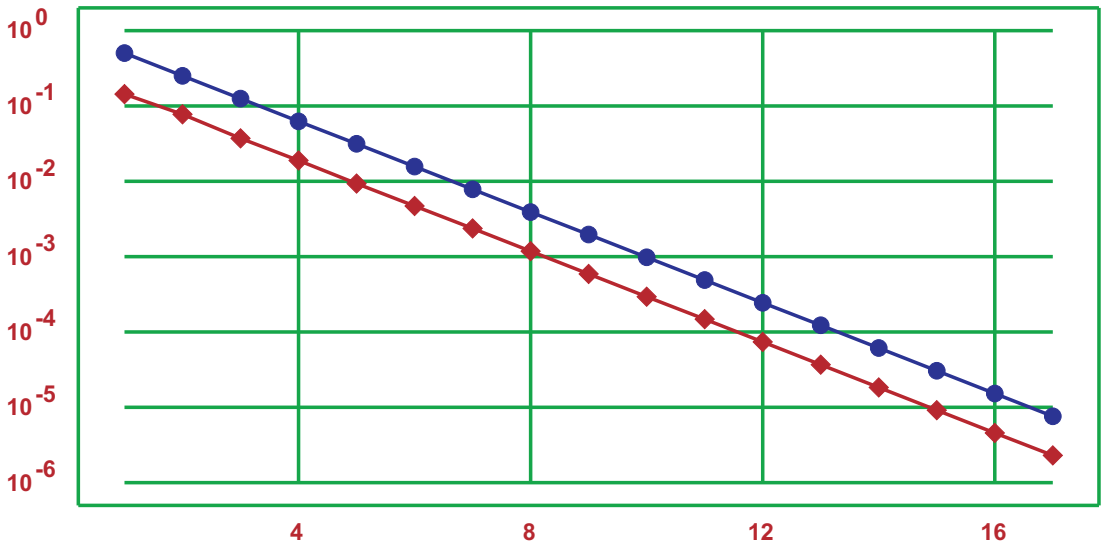

Figura 2.4: Error absoluto, marcado con círculos azules, y error relativo, identificado con rombos rojos, de los iterados del algoritmo de bisección del ejemplo 8. Escala logarítmica en el eje de las ordenadas.

Comandos de EMT. *El comando para aplicar bisección es:*

- `bisect(f$:string, a:número, b:número, y:número)`, *donde* ***f$*** *es una función de* x *expresada como* string*;* ***a*** *y* ***b*** *forman el intervalo de búsqueda;* ***y*** *es un parámetro opcional para resolver* $f(x) = y$ *en vez de* $f(x) = 0$.

Ejemplo en EMT 2. *Calcular la raíz de la función* $f(x) = \cos(x) + x$ *que está en el intervalo* $[-2; 2]$*, utilizando el algoritmo de bisección.*

```
>bisect("cos(x)+x",-2,2)
 -0.739085133215
```

2.2. Método de Punto Fijo

Este método se basa en el hecho de que, para una función dada $f : [a, b] \to \mathbb{R}$, siempre es posible transformar el problema $f(x) = 0$ en el problema equivalente $x - \phi(x) = 0$, donde la función auxiliar $\phi : [a, b] \to \mathbb{R}$ debe ser elegida de cierta forma para que $\phi(\alpha) = \alpha$ siempre que $f(\alpha) = 0$. Aproximar los ceros de una función se transforma en el problema de encontrar los puntos fijos del mapeo ϕ, lo que se realiza siguiendo el siguiente algoritmo iterativo: dado x_0, sea:

$$x_{n+1} = \phi(x_n), \tag{2.7}$$

para $n \in \mathbb{N}_0$.

Se dice que (2.7) es una *iteración de punto fijo* y ϕ es su *función de iteración* asociada. La elección de ϕ no es única. Por ejemplo, cualquier función de la forma $\phi(x) = x + F(f(x))$, donde F es una función continua tal que $F(0) = 0$ es admisible como función de iteración. Los dos resultados siguientes proveen condiciones suficientes para que el método de punto fijo converja a la raíz α del problema.

Teorema 4 (Existencia y unicidad del punto fijo)**.** *Sea el mapeo* $\phi \in C^1[a, b]$*, entonces:*

1. *Si* $\phi(x) \in [a, b]$, $\forall x \in [a, b]$*, entonces* ϕ *tiene al menos un punto fijo* $\alpha \in [a, b]$.

2. *Si $\phi'(x)$ está definida en (a,b) y se verifica que $|\phi'(x)| < 1$, $\forall x \in (a,b)$, entonces el punto fijo α de ϕ en $[a,b]$ es único.*

Demostración. Si $\phi(a) = a$ ó $\phi(b) = b$, el primer consecuente del teorema está demostrado. Si no se da esa condición, entonces se cumple que $\phi(a) \in (a,b]$ y $\phi(b) \in [a,b)$. Por lo tanto la función $f(x) = x - \phi(x)$ tiene la propiedad:

$$f(a) = a - \phi(a) < 0, \qquad f(b) = b - \phi(b) > 0.$$

Al aplicar el teorema del valor intermedio, se asegura la existencia de un $\alpha \in (a,b)$ tal que $f(\alpha) = 0$ entonces $\alpha = \phi(\alpha)$ con lo que se asegura la existencia del punto fijo. Para demostrar la unicidad se supondrán dos puntos fijos $\alpha_1, \alpha_2 \in [a,b]$. Por el teorema del valor medio, es posible identificar un $c \in (\alpha_1, \alpha_2)$ tal que:

$$\begin{aligned} \phi'(c) &= \frac{\phi(\alpha_1) - \phi(\alpha_2)}{\alpha_1 - \alpha_2} \\ &= \frac{\alpha_1 - \alpha_2}{\alpha_1 - \alpha_2} \\ &= 1, \end{aligned}$$

pero esto contradice la hipótesis planteada. Por lo tanto, queda demostrado el segundo consecuente. □

Teorema 5 (Convergencia de las iteraciones de punto fijo)**.** *Dada la secuencia $x_{n+1} = \phi(x_n)$, para $n \in \mathbb{N}_0$, dado x_0 y asumiendo que:*

- $\phi : [a,b] \to [a,b]$;
- $\phi \in C^1([a,b])$;
- $\exists K < 1 : |\phi'(x)| \leq K, \forall x \in [a;b]$.

Entonces ϕ tiene un único punto fijo α en $[a;b]$ y la secuencia $\{x_n\}$ converge a α para todo $x_0 \in [a;b]$. Además, se verifica que:

$$\lim_{n \to +\infty} \frac{x_{n+1} - \alpha}{x_n - \alpha} = \phi'(\alpha) \tag{2.8}$$

El teorema 5 asegura la convergencia de la secuencia x_n a la raíz α para cualquier punto que se elija como semilla dentro del intervalo $[a;b]$. Es decir que se asegura la convergencia global del método dentro de las condiciones anteriores. Sin embargo, en la práctica, es a menudo muy difícil determinar a priori los extremos del intervalo de convergencia $[a;b]$. En dichos casos, es muy útil el siguiente resultado:

Teorema 6 (de Ostrowski)**.** *Sea α un punto fijo de la función ϕ, la que es continua y diferenciable en una vecindad de α. Si $|\phi'(\alpha)| < 1$ entonces existe $\delta > 0$ tal que la secuencia x_n converge a α para cualquier valor x_0 que cumpla que $|x_0 - \alpha| < \delta$.*

Nota. *Si $|\phi'(\alpha)| > 1$ y x_n es lo suficientemente cercano a α, entonces, de acuerdo a (2.8) $|\phi'(x_n)| > 1$, entonces $|x_{n+1} - \alpha| > |x_n - \alpha|$ y la convergencia es imposible. En el caso de que $|\phi'(\alpha)| = 1$, no hay conclusión con respecto a la convergencia.*

Ejemplo 9. *Sea $\phi(x) = x - x^3$, que admite $\alpha = 0$ como punto fijo. A pesar de que $\phi'(\alpha) = 1$, si $x_0 \in [-1,1]$ entonces $x_k \in (-1,1)$ para $k \geq 1$ y converge (muy lentamente) a α. Si $x_0 = \pm 1$, también se cumple que $x_k = \alpha$ para cualquier $k \geq 1$. Eligiendo $x_0 = \frac{1}{2}$ el error absoluto después de 2000 iteraciones, con una aritmética de 8 dígitos, es $3{,}93 \times 10^{-6}$. Sea ahora $\phi(x) = x + x^3$, donde también $\alpha = 0$ es punto fijo. Nuevamente, $\phi'(\alpha) = 1$ pero en este caso la secuencia x_k diverge para cualquier $x_0 \neq 0$.*

i	x_i	$\phi(x_i)$	ϵ_a	ϵ_r
0	3,0000000	4,6536436	-	-
1	4,6536436	1,2027623	1,65E+00	3,55E-01
2	1,2027623	-5,7582979	3,45E+00	2,87E+00
3	-5,7582979	-0,5096617	6,96E+00	1,21E+00
4	-0,5096617	-1,2423293	5,25E+00	1,03E+01
5	-1,2423293	-0,9342723	7,33E-01	5,90E-01
6	-0,9342723	-1,0231359	3,08E-01	3,30E-01
7	-1,0231359	-0,9924356	8,89E-02	8,69E-02
8	-0,9924356	-1,0025374	3,07E-02	3,09E-02
9	-1,0025374	-0,9991560	1,01E-02	1,01E-02
10	-0,9991560	-1,0002815	3,38E-03	3,38E-03
11	-1,0002815	-0,9999062	1,13E-03	1,13E-03
12	-0,9999062	-1,0000312	3,75E-04	3,75E-04
13	-1,0000312	-0,9999896	1,25E-04	1,25E-04
14	-0,9999896	-1,0000034	4,16E-05	4,16E-05
15	-1,0000034	-0,9999989	1,38E-05	1,38E-05
16	-0,9999989	-1,0000003	4,54E-06	4,54E-06

Tabla 2.2: Salida del algoritmo de punto fijo para $\phi(x)$ del ejemplo 10.

Ejemplo 10. *La función* $f(x) = x^2 - 2x + \cos(x+1) - 4$ *posee dos raíces en el intervalo* $[-2; -4]$, *denominadas* r_1 *y* r_2, *donde* $r_1 < r_2$. *Utilizando el método de punto fijo es posible obtener* r_1, *sin embargo, se intentará calcular* r_2. *Utilizando:*

$$\phi(x) = \frac{4 - \cos(x+1)}{x - 2},$$

y $x_0 = 3$ *(valor cercano a* r_2*) se obtienen los datos volcados en la tabla 2.2. Se utilizó* $\varepsilon = 1 \times 10^{-5}$, *una aritmética de 8 dígitos y truncamiento. En las figuras 2.5 y 2.6 se muestran la convergencia del método y los errores, respectivamente.*

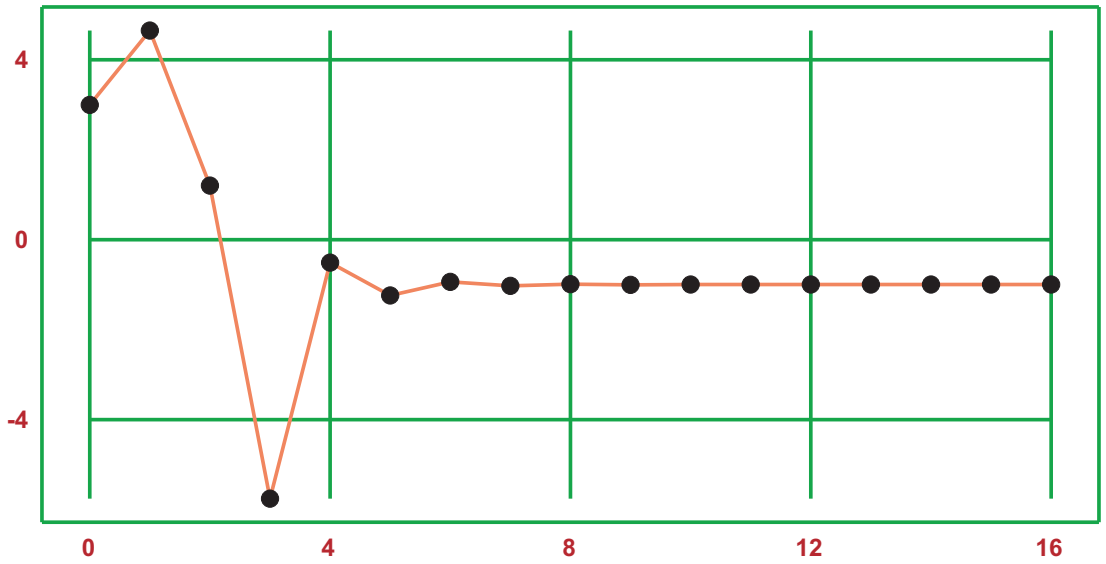

Figura 2.5: Convergencia a la raíz de la función del ejemplo 10.

Comandos de EMT. *No está implementado el algoritmo de punto fijo en* EMT, *sin embargo es posible realizar diagramas de Verhlust.*

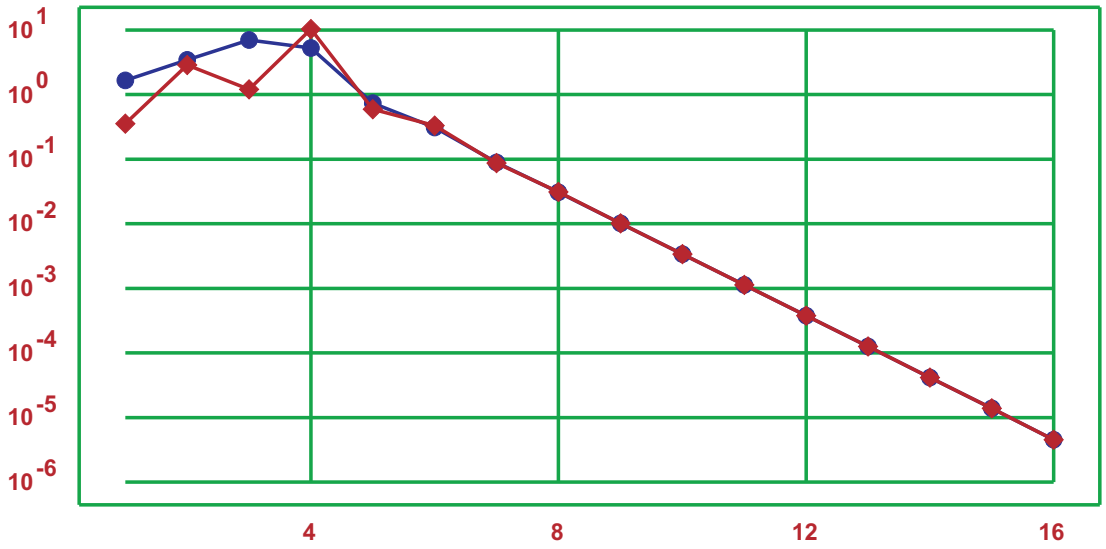

Figura 2.6: Error absoluto, marcado con círculos azules, y error relativo, identificado con rombos rojos, de los iterados del algoritmo de punto fijo del ejemplo 10. Escala logarítmica en el eje de las ordenadas.

- `fwebplot(f$:string, a:número, b:número, xstart:número, n:natural, color)`, *donde* ***f$*** *es el mapeo de* x *expresado como* string*;* ***a*** *y* ***b*** *forman el intervalo de visualización gráfica;* ***xstart*** *es la semilla;* ***n*** *es la cantidad de iteraciones a realizar;* ***color*** *es un parámetro optativo y permite modificar el color de visualización del mapeo.*

Ejemplo en EMT 3. *Obtener la raíz de* $f(x) = x - 2\cos(x)$*, a través de dos mapeos diferentes:* $\phi_1(x) = 2\cos(x)$ *y* $\phi_2(x) = \arccos(x/2)$*. Para ambos esquemas, se utilizará la semilla* $x_0 = 1$*. Mostrar, con los diagramas de Verhlust ó* cobweb *y 8 iteraciones, que el primero de ellos es divergente, mientras que el segundo es convergente.*

```
>fwebplot("2*cos(x)",0.25,1.5,1.04,8);
>fwebplot("arccos(x/2)",0.5,1.5,0.55,8);
```

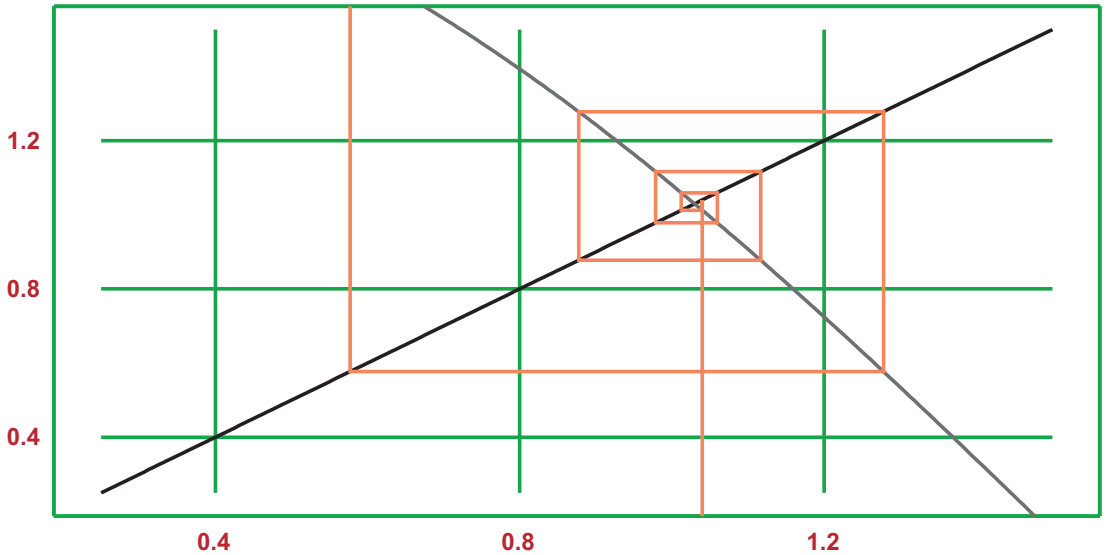

Figura 2.7: *Cobweb* del mapeo $\phi_1(x)$.

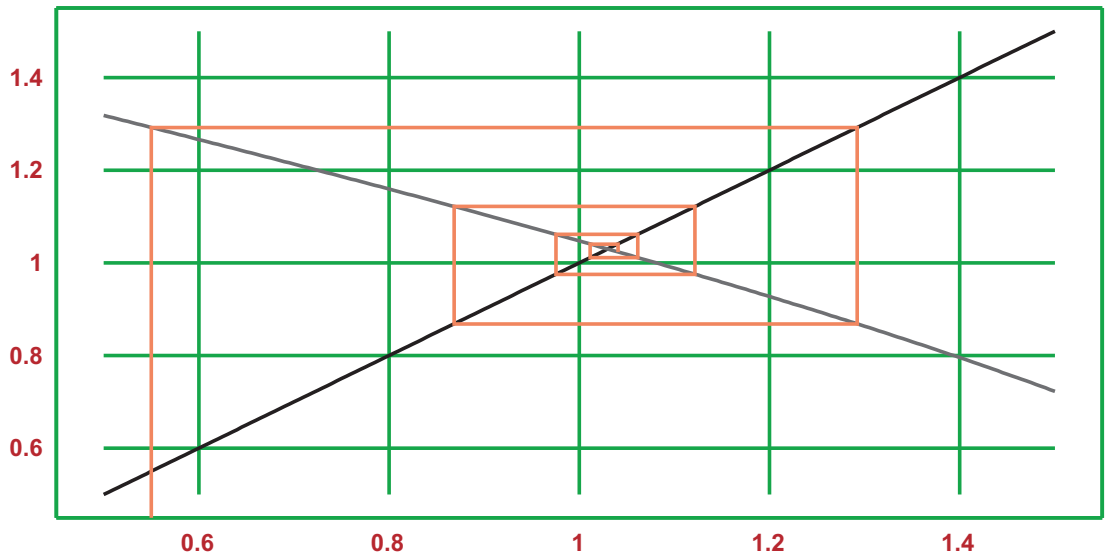

Figura 2.8: *Cobweb* del mapeo $\phi_2(x)$.

2.3. Método de Newton-Raphson

El método de Newton-Raphson es uno de los algoritmos más utilizados para encontrar raíces por una buena razón: es rápido y simple. El único inconveniente de este método es la utilización de la función $f(x)$ y su derivada $f'(x)$. Más aún, el método de Newton-Raphson se usa sólo en aquellos problemas en los que es fácilmente calculable la derivada $f'(x)$.

La fórmula iterativa de Newton-Raphson puede ser derivada de la expansión por series de Taylor de $f(x)$ en x_{i+1}:

$$f(x_{i+1}) = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \mathcal{O}(x_{i+1} - x_i)^2.$$

Si x_{i+1} es una raíz de $f(x)$ entonces:

$$0 = f(x_i) + f'(x_i)(x_{i+1} - x_i) + \mathcal{O}(x_{i+1} - x_i)^2, \tag{2.9}$$

donde $\mathcal{O}(x_{i+1}-x_i)^2$ representa un término despreciable de la forma $K(x_{i+1}-x_i)^2$, para K constante. Ahora, si x_{i+1} es lo suficientemente cercano a x_i, entonces los términos de orden superior se vuelven despreciables y se puede resolver la ecuación (2.9) para x_{i+1}. El resultado es la fórmula iterativa de Newton-Raphson:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}. \tag{2.10}$$

Escrito como una relación general $x_{n+1} = g(x_n)$, no es difícil ver que la ecuación (2.10) usa una función g definida como:

$$g(x) = x - \frac{f(x)}{f'(x)}.$$

Recordando que la derivada de g proporciona la descripción de la convergencia a través del método de punto fijo, queda:

$$\begin{aligned} g'(x) &= 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2} \\ &= 1 - \frac{[f'(x)]^2}{[f'(x)]^2} + \frac{f(x)f''(x)}{[f'(x)]^2} \\ &= \frac{f(x)f''(x)}{[f'(x)]^2} \end{aligned}$$

En la raíz $x = \alpha$ el valor de $f(\alpha)$ es cero por definición. Por lo tanto, para el proceso iterativo de Newton-Raphson, $g'(\alpha) = 0$ lo que implica convergencia óptima. Sin embargo sería útil analizar el error de convergencia con más detenimiento.

Cuando se analiza la convergencia a través del método de punto fijo, es necesario calcular cuán alejado del valor α se encuentra x_n, por lo tanto el análisis recaerá sobre:

$$x_{n+1} - \alpha = g(x_n) - g(\alpha),$$

donde $\alpha = g(\alpha)$ (α es punto fijo de g) y el error cometido es $\epsilon_n = x_n - \alpha$. Entonces $x_n = \alpha + \epsilon_n$ y la ecuación anterior se convierte en:

$$\epsilon_{n+1} = g(\alpha + \epsilon_n) - g(\alpha).$$

Aplicando el Teorema de Taylor:

$$\begin{aligned}\epsilon_{n+1} &= g(\alpha) + \epsilon_n g'(\alpha) + \frac{1}{2}{\epsilon_n}^2 g''(\alpha) + \ldots - g(\alpha) \\ &= \epsilon_n g'(\alpha) + \frac{1}{2}{\epsilon_n}^2 g''(\alpha) + \ldots\end{aligned}$$

Como $g'(\alpha) = 0$ se elimina el primer término del lado derecho de la igualdad anterior, y en las cercanías de α se pueden despreciar los términos de orden superior, con lo que:

$$\epsilon_{n+1} \propto {\epsilon_n}^2 \tag{2.11}$$

y se muestra que el error en cada paso es proporcional al cuadrado del error cometido en el paso anterior. Como consecuencia de esto, el número de cifras significativas se dobla en cada iteración, siempre que x_i esté cercano a la raíz.

Ejemplo 11. *Se quiere calcular la raíz del polinomio $p(x) = 2x^2 - 1$ aplicando el método de Newton-Raphson. Se elige $x_0 = 0{,}5$ y la función de iteración es:*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Se obtiene la siguiente secuencia:

$$\begin{aligned}x_1 &= 0{,}750000000000000 \\ x_2 &= \underline{0{,}708333333333333} \\ x_3 &= \underline{0{,}707107843137255} \\ x_4 &= \underline{0{,}707106781187345} \\ x_5 &= \underline{0{,}707106781186548},\end{aligned}$$

donde se subrayan los dígitos correctos y se nota la rápida convergencia del método.

Al igual que el método de bisección, éste método termina su proceso iterativo cuando se cumple la condición (2.6). En la figura 2.9 se muestra cómo el método converge a una raíz luego de cuatro iteraciones para una función genérica.

Ejemplo 12. *La función $f(x) = x^2 - 2x + \cos(x+1) - 4$ posee dos raíces en el intervalo $[-2; -4]$, denominadas r_1 y r_2, donde $r_1 < r_2$. Utilizando el método de Newton Raphson es posible obtener ambas, sin embargo, se calculará sólo el valor de r_2. En la tabla 2.3 se muestra la secuencia iterativa que converge a r_2. Se iteró hasta que el error absoluto de tolerancia para la secuencia de valores fue menor a 1×10^{-5}. En las figuras 2.10 y 2.11 se muestran la convergencia del método y los errores, respectivamente, para los datos de la tabla 2.3.*

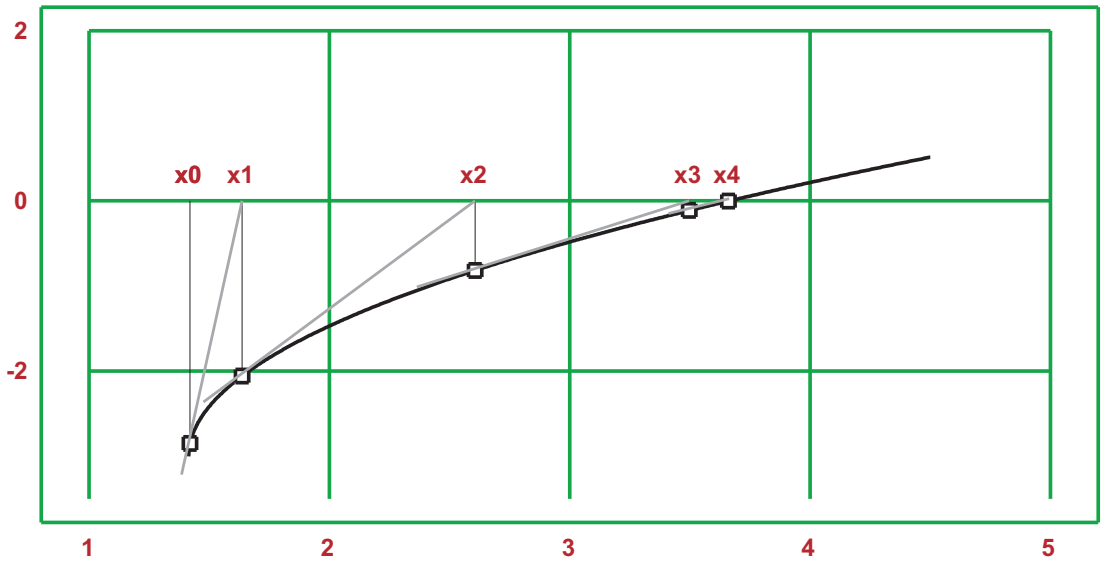

Figura 2.9: Obtención de la raíz de una función genérica a través del método de Newton-Raphson.

i	x_i	x_{i+1}	ϵ_a	ϵ_r
0	3,0000000	3,347638	-	-
1	3,3476380	3,3201599	3,48E-01	1,04E-01
2	3,3201599	3,3199994	2,75E-02	8,28E-03
3	3,3199994	3,3199994	1,61E-04	4,83E-05
4	3,3199994	3,3199994	0,00E+00	0,00E+00

Tabla 2.3: Salida del algoritmo de Newton-Raphson para $x_0 = 3$ del ejemplo 12.

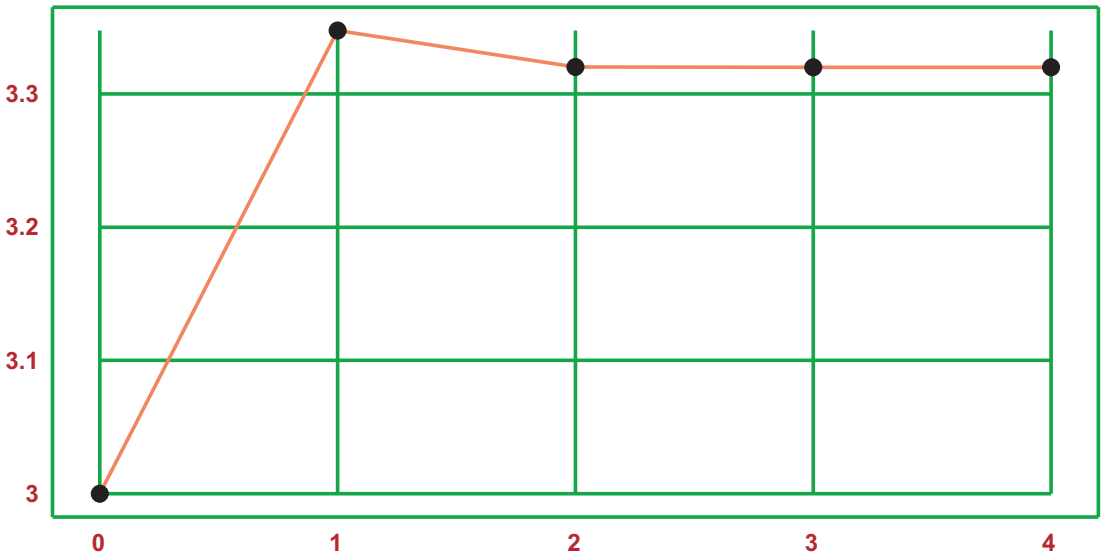

Figura 2.10: Convergencia del método de Newton-Raphson para la raíz r_2 del ejemplo 12.

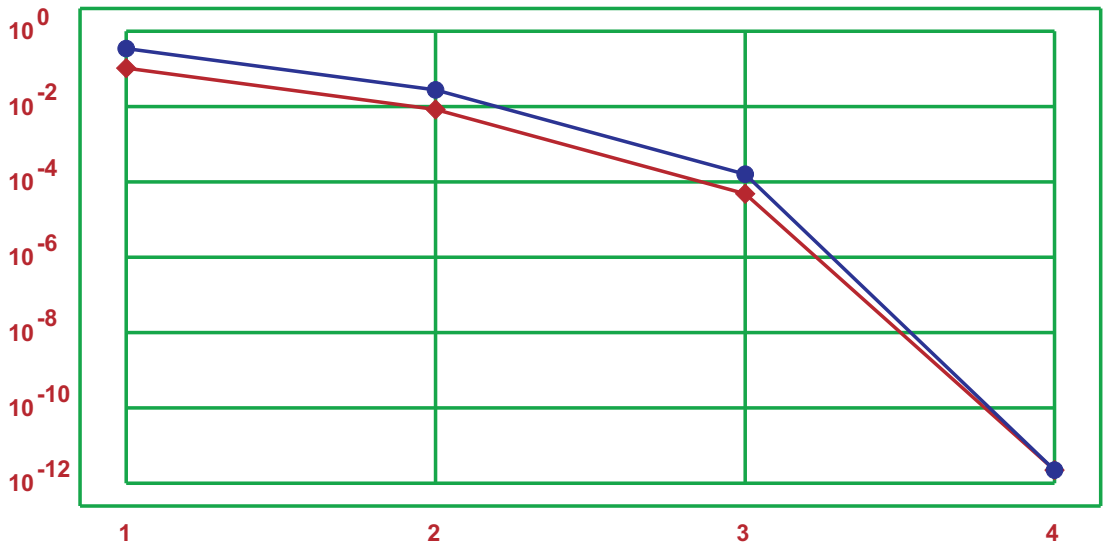

Figura 2.11: Error absoluto, marcado con círculos azules, y error relativo, identificado con rombos rojos, de los iterados del algoritmo de Newton-Raphson del ejemplo 12. Escala logarítmica en el eje de las ordenadas.

Comandos de EMT. *El comando para aplicar el método de Newton-Raphson es:*

- `newton(f$:string, df$:string, x:número, y:número)`, *donde* ***f$*** *es la función de* x *a resolver, expresada como* string*;* ***df$*** *es la derivada de la función* f*, expresada como* string*;* $\boldsymbol{x}$ *es la semilla;* $\boldsymbol{y}$ *es un parámetro optativo y permite resolver* $f(x) = y$ *en vez de* $f(x) = 0$*.*

Ejemplo en EMT 4. *Obtener la primera raíz positiva de* $f(x) = x^2\cos(x) + x$*. A través de un gráfico se observa que es cercana a* $x = 2$*, por lo que se utilizará ese valor como semilla:*

```
>newton("x^2*cos(x)+x","2*x*cos(x)-x^2*sin(x)+1",2)
 2.07393280909
```

2.4. Método de la Secante

Una de las principales complicaciones del método de Newton-Raphson es la obtención analítica de $f'(x)$ y a esto se suma que se debe calcular en cada paso el valor de $f'(x_n)$. Sin embargo este proceso se simplifica si en vez de tomar la derivada en cada punto se utiliza una aproximación numérica (recta secante en vez de recta tangente). Puesto que:

$$f'(x_n) = \lim_{x_{n-1}\to x_n} \frac{f(x_{n-1}) - f(x_n)}{x_{n-1} - x_n},$$

entonces:

$$f'(x_n) \approx \frac{f(x_{n-1}) - f(x_n)}{x_{n-1} - x_n}$$

y esta aproximación es la que se utilizará en (2.10):

$$\begin{aligned} x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \\ &= x_n - \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)} f(x_n) \\ &= \frac{x_n[f(x_{n-1}) - f(x_n)] - [x_{n-1} - x_n]f(x_n)}{f(x_{n-1}) - f(x_n)} \\ &= \frac{x_n f(x_{n-1}) - x_{n-1} f(x_n)}{f(x_{n-1}) - f(x_n)}, \end{aligned}$$

sumando 1 a cada subíndice para definir x_{n+2} en forma iterativa se obtiene lo que se denomina método de la secante:

$$x_{n+2} = \frac{x_{n+1} f(x_n) - x_n f(x_{n+1})}{f(x_n) - f(x_{n+1})} \tag{2.12}$$

Ahora, es necesario contar con x_n y x_{n+1} para obtener x_{n+2}. Luego, con x_{n+1} y x_{n+2} se repite el proceso (2.12) y se obtiene x_{n+3}. El método termina cuando se cumple la condición (2.6). En la figura 2.12 se muestran los valores x_0, x_1, x_2, x_3, x_4 de la secuencia iterativa. Es de notar que el proceso de obtención de raíz aún no terminó, pero se suspendió para no sobrecargar el gráfico.

Es de notar que la ecuación (2.12) lleva involucrada una carga algebraica imporante. La versión original, donde la aproximación a $f'(x)$ se calcula por separado, presenta una pequeña diferencia con esta versión *algebraica*. Sin embargo, ambas convergen con la misma rapidez hacia la raíz, bajo las condiciones iniciales adecuadas.

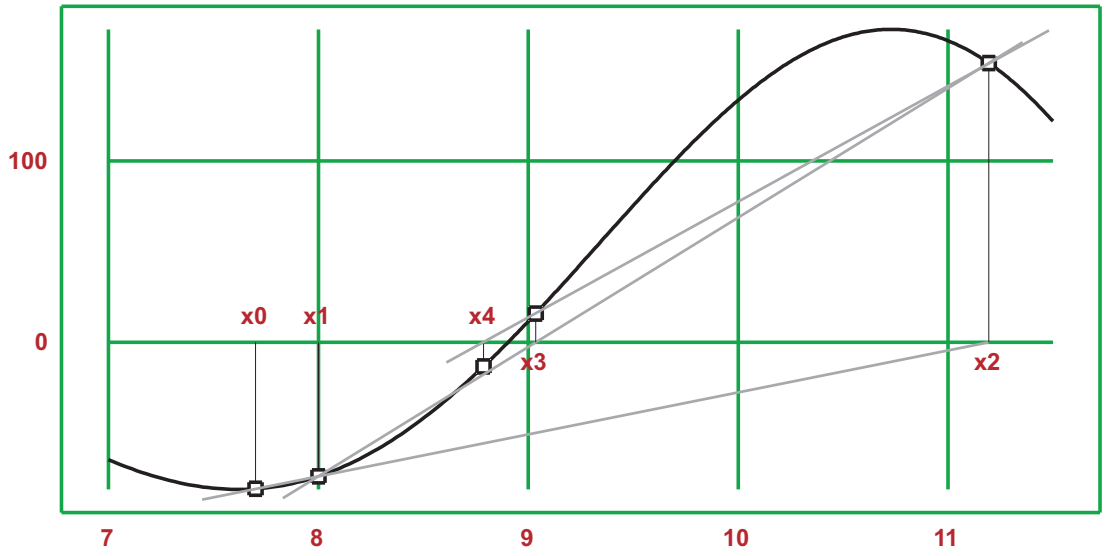

Figura 2.12: Obtención de la raíz de una función genérica, a través del método de la secante.

Ejemplo 13. *La función* $f(x) = x^2 - 2x + \cos(x+1) - 4$ *posee dos raíces en el intervalo* $[-2; -4]$, *denominadas* r_1 *y* r_2, *donde* $r_1 < r_2$. *Utilizando el método de la secante es posible obtener ambas, sin embargo, se calculará sólo el valor de* r_2. *La tabla de iteración para* r_2 *es 2.4. En la figura 2.13 se aprecia la convergencia hacia la raíz y en la figura 2.14 se muestran los dos errores: absoluto (círculos) y relativo (rombos).*

Comandos de EMT. *El comando para aplicar el método de la secante es:*

i	x_i	x_{i+1}	x_{i+2}	ϵ_a	ϵ_r
0	2,8000000	3,1000000	3,3521002	-	-
1	3,1000000	3,3521002	3,3183849	3,37E-02	1,02E-02
2	3,3521002	3,3183849	3,3199884	1,60E-03	4,83E-04
3	3,3183849	3,3199884	3,3199994	1,10E-05	3,31E-06
4	3,3199884	3,3199994	3,3199994	0,00E+00	0,00E+00

Tabla 2.4: Salida del algoritmo de la secante para $x_0 = 2{,}8$ y $x_1 = 3{,}1$ del ejemplo 13.

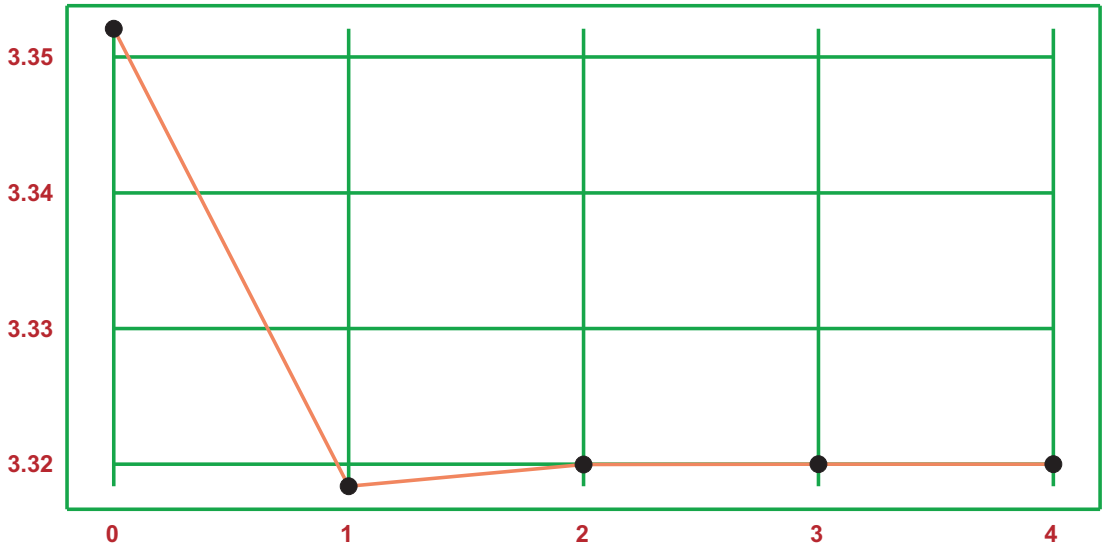

Figura 2.13: Convergencia hacia la raíz de la función del ejemplo 13, a través del método de la secante.

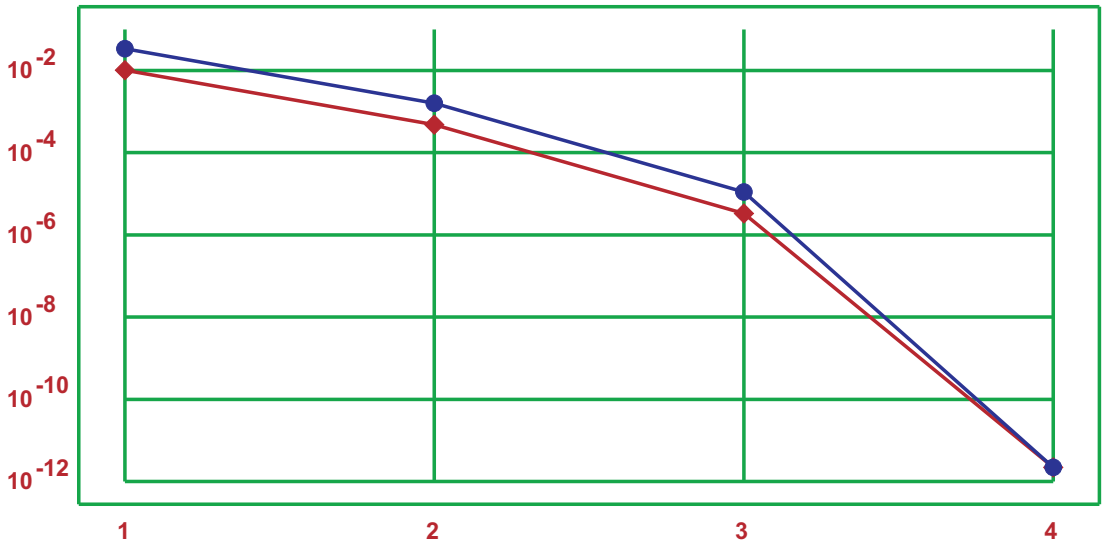

Figura 2.14: Error absoluto, marcado con círculos azules, y error relativo, identificado con rombos rojos, de los iterados del algoritmo de la secante del ejemplo 13. Escala logarítmica en el eje de las ordenadas.

- `secant(f$:string, a:número, b:número, y:número)`, *donde* ***f$*** *es la función de* x *a resolver, expresada como* string*;* ***a*** *y* ***b*** *forman el intervalo de exploración, aunque* ***b*** *es un parámetro optativo (si no se declara se utiliza un entorno centrado en* ***a*** *como intervalo de exploración);* ***y*** *es un parámetro optativo y permite resolver* $f(x) = y$ *en vez de* $f(x) = 0$.

Ejemplo en EMT 5. *Obtener la primera raíz positiva de* $f(x) = x^2\cos(x) + x$*, utilizando el método de la secante. A través de un gráfico se observa que es cercana a* $x = 2$*, por lo que se utilizará como semilla:*

```
>secant("x^2*cos(x)+x",2)
 2.07393280909
```

2.5. Método de Regula-Falsi

Este método, también llamado *de la Falsa Posición*, aprovecha la secuencia de rectas generadas por el método de la secante pero para la selección de los puntos se basa en el método de bisección.

Dada $f(x)$ continua y dados x_0 y x_1 tales que $f(x_0)f(x_1) < 0$ se construye la recta secante que pasa por ellos. Si x_2 es la raíz de la secante, entonces la raíz de f se encuentra en el intervalo (x_0, x_2) siempre y cuando $f(x_0)f(x_2) < 0$. Si $f(x_0)f(x_2) > 0$ entonces se selecciona el intervalo (x_2, x_1) para seguir aplicando el proceso iterativo.

Es decir que:

$$x_{n+1} = \frac{x_{n'}f(x_n) - x_n f(x_{n'})}{f(x_n) - f(x_{n'})}, \tag{2.13}$$

donde $x_{n'}$ depende de $f(x)$ puesto que $f(x_{n'})f(x_n) < 0$.

En la figura 2.15 se aprecia cómo el método opera: se forma la secante que pasa por x_0 y x_1, su raíz x_2 se utiliza como abscisa. En este caso, como $f(x_1)f(x_2) > 0$ entonces x_2 reemplaza a x_1. Luego se forma la secante que pasa por x_0 y x_2, su raíz es x_3. Como $f(x_0)f(x_3) > 0$ entonces x_3 reemplaza a x_0. De la secante entre x_2 y x_3 se obtiene x_4. Este proceso se repite hasta lograr la condición de convergencia. A veces puede ocurrir que uno de los extremos del intervalo de exploración permanezca constante a lo largo de todo el proceso iterativo. Esto no tiene relación con la distancia de los extremos del intervalo a la raíz, es una cuestión netamente geométrica.

Ejemplo 14. *La función* $f(x) = x^2 - 2x + \cos(x+1) - 4$ *posee dos raíces en el intervalo* $[-2; -4]$*, denominadas* r_1 *y* r_2*, donde* $r_1 < r_2$*. Utilizando el método de regula falsi es posible obtener ambas, sin embargo, se calculará sólo el valor de* r_2*. La tabla de iteración que converge en* r_2 *es 2.5. En la figura 2.16 se muestra la convergencia del método y en la figura 2.17 se visualizan los errores.*

2.6. Análisis de convergencia

¿Cómo caracterizar la eficiencia computacional de un algoritmo iterativo? Los algoritmos secuenciales se caracterizan generalmente por la cantidad de *flops*[1] que se deben realizar para una cantidad ya definida de pasos. Para comparar la eficiencia de los distintos métodos iterativos, se debe medir la rapidez con la que los iterados $x_1, x_2, \ldots, x_n, \ldots$, convergen ó no a la raíz. Como la solución exacta se desconoce, se comparan los diferentes cocientes que surgen de tomar valores sucesivos $\Delta x_{k+1} = x_{k+1} - x_k$ y $\Delta x_{k+2} = x_{k+2} - x_{k+1}$:

[1] *floating point operations*

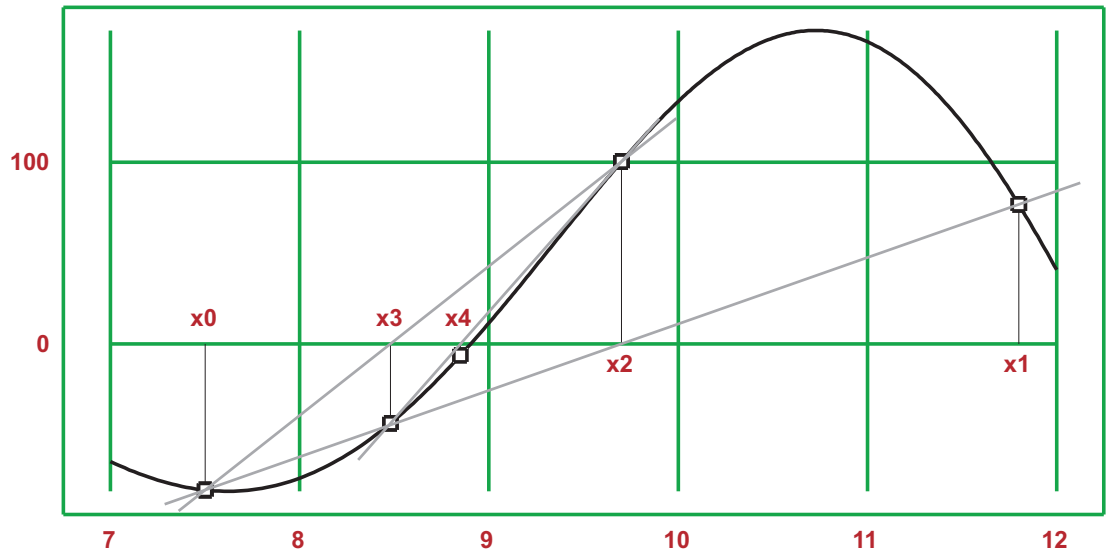

Figura 2.15: Cuatro pasos en la obtención de la raíz de una función genérica, a través del método de regula-falsi.

i	a_i	b_i	p_i	ϵ_a	ϵ_r
0	2,0000000	4,0000000	3,0761652	-	-
1	3,0761652	4,0000000	3,2891793	2,13E-01	6,48E-02
2	3,2891793	4,0000000	3,3163652	2,72E-02	8,20E-03
3	3,3163652	4,0000000	3,3195746	3,21E-03	9,67E-04
4	3,3195746	4,0000000	3,3199498	3,75E-04	1,13E-04
5	3,3199498	4,0000000	3,3199936	4,38E-05	1,32E-05
6	3,3199936	4,0000000	3,3199987	5,10E-06	1,54E-06

Tabla 2.5: Salida del algoritmo de la regula-falsi para $a_0 = 2$ y $b_0 = 4$ del ejemplo 14.

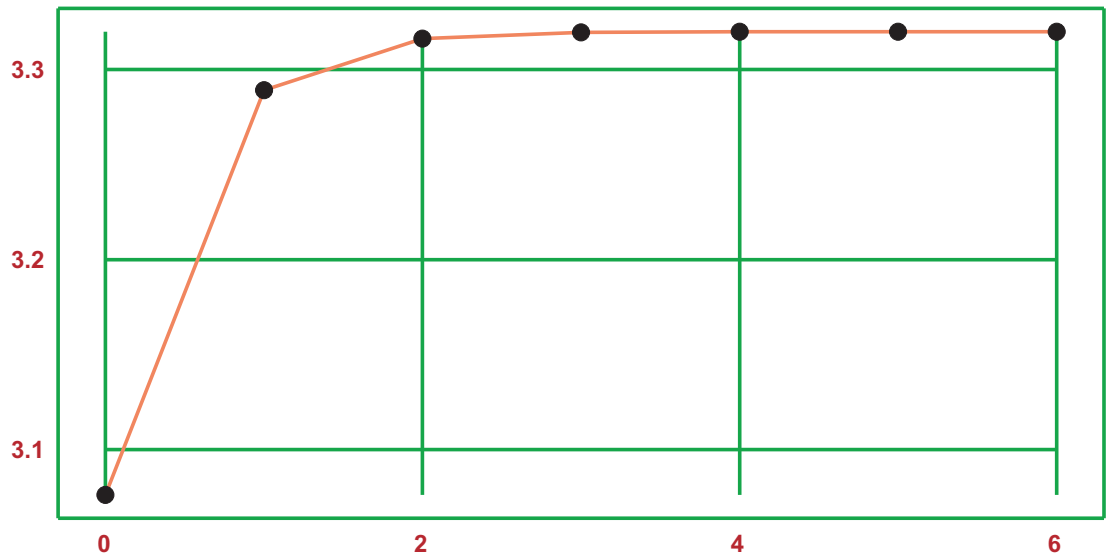

Figura 2.16: Convergencia hacia la raíz de la función del ejemplo 14, a través del método de regula falsi.

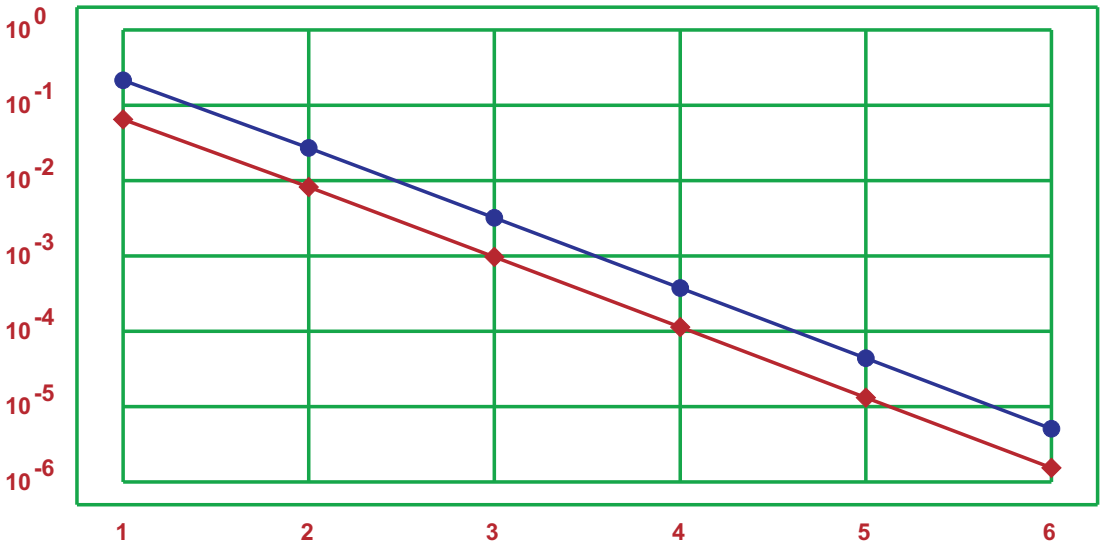

Figura 2.17: Error absoluto, marcado con círculos azules, y error relativo, identificado con rombos rojos, de los iterados del algoritmo de regula-falsi del ejemplo 14. Escala logarítmica en el eje de las ordenadas.

- Si los cocientes $\Delta x_{k+2}/\Delta x_{k+1}$ tienden a estabilizarse en torno a un valor constante C_L y $0 < C_L < 1$, entonces el método tiene **convergencia lineal** y la tasa de convergencia lineal es C_L. Cuanto más cercano a 0 sea el valor de C_L, más rápida será la convergencia y cuando más próximo a 1, más lenta. Si $C_L > 1$, entonces los iterados divergen.

- Si los cocientes $\Delta x_{k+2}/\Delta x_{k+1}$ tienden a cero y los cocientes $\Delta x_{k+2}/\Delta {x_{k+1}}^{\alpha}$, para cierto $1 < \alpha < 2$ tienden a estabilizarse, entonces el método tiene **convergencia superlineal**.

- Si los cocientes $\Delta x_{k+2}/\Delta x_{k+1}$ tienden a cero y los cocientes $\Delta x_{k+2}/\Delta {x_{k+1}}^{2}$ tienden a estabilizarse, entonces el método tiene **convergencia cuadrática**.

En la práctica, el tipo de convergencia indica el trabajo necesario para alcanzar determinada precisión. Con un método que converge linealmente, cada cierta cantidad de iteraciones (por ejemplo cada 3), se obtiene un dígito adicional exacto en la solución. En cambio si hay convergencia cuadrática, con la misma cantidad de iteraciones que antes se duplica el número de decimales exactos.

Los métodos de *bisección* y *regula-falsi* convergen linealmente. El método de *punto fijo*, si converge, lo suele hacer linealmente, aunque puede tener convergencia cuadrática ó superior en casos especiales. El método de *Newton-Raphson* converge cuadráticamente, si se dan las condiciones adecuadas (cálculo de una raíz simple). El método de la *secante* tiene, generalmente, convergencia superlineal.

Ejemplo 15. *Los iterados de los ejemplos donde se calculó alguna raíz de* $f(x) = x^2 - 2x + \cos(x+1) - 4$ *tienen tasas de convergencias acordes a lo planteado anteriormente: bisección, regula-falsi y punto fijo convergen linealmente, siendo el más rápido regula-falsi, con una tasa de convergencia de* 0,116 *aproximadamente; Newton-Raphson y secante fallan en el testeo de convergencia lineal, teniendo Newton-Raphson convergencia cuadrática y secante convergencia superlineal de tasa* 1,63575 *aproximadamente. En la tabla 2.6 se muestran las sucesivas divisiones de iterados para el testeo de convergencia lineal; en la tabla 2.7 el testo de convergencia cuadrática y la confirmación de convergencia superlineal para el método de la secante.*

k	Bisección	Punto Fijo	Newton-Raphson	Secante	Regula-Falsi
0	0,50000000	2,08683490	0,07904122	0,04756001	0,12762488
1	0,50000000	2,01718330	0,00584110	0,00685999	0,11805384
2	0,50000000	0,75399954	0,00000000	0,00000000	0,11690658
3	0,50000000	0,13959199	-	-	0,11673773
4	0,50000000	0,42045947	-	-	0,11643835
5	0,50000000	0,28846488	-	-	-
6	0,50000640	0,34547684	-	-	-
7	0,49998720	0,32904630	-	-	-
8	0,49997439	0,33473340	-	-	-
9	0,50005120	0,33285424	-	-	-
10	0,50010239	0,33346364	-	-	-
11	0,50040950	0,33310241	-	-	-
12	0,49918166	0,33274676	-	-	-
13	0,49999999	0,33173076	-	-	-
14	0,49836065	0,32898550	-	-	-
15	0,50000000	-	-	-	-

Tabla 2.6: Cálculo de $\Delta x_{k+2}/\Delta x_{k+1}$ para los iterados de la solución de $f(x)$.

k	**Newton-Raphson** ($\alpha = 2$)	**Secante** ($\alpha = 2$)	**Secante** ($\alpha = 1{,}63575$)
0	0,22736671	1,41063580	0,41037028
1	0,21257602	4,27813760	0,41039411
2	0,00000000	0,00000000	0,00000000

Tabla 2.7: Cálculo de $\Delta x_{k+2}/\Delta x_{k+1}^{\alpha}$ para los iterados de la solución de $f(x)$.

k	Bisección	Newton-Raphson	Secante	Regula-Falsi
0	0,50000000	0,66666664	1,06872130	0,97301773
1	0,50000000	0,66666668	0,66666031	0,97351291
2	0,50000000	0,66666664	0,79278687	0,97398877
3	0,49999998	0,66666666	0,74091541	0,97444960
4	0,50000000	0,66666667	0,76034065	0,97489180
5	-	0,66666668	0,75278883	0,97531951
6	-	0,66666665	0,75568333	0,97573376

Tabla 2.8: Cálculo de $\Delta x_{k+2}/\Delta x_{k+1}$ para los iterados de la solución del ejemplo 16.

Es sabido que el método más rápido presentado hasta ahora, Newton-Raphson, pierde su convergencia cuadrática cuando busca una raíz real que no sea simple. sin embargo, esto no afecta el comportamiento de los otros métodos. Debe tenerse en cuenta al aplicar el método de punto fijo que nunca se sabe de antemano la velocidad de convergencia para el mapeo $\phi(x)$ elegido.

Ejemplo 16. *El polinomio $f(x) = x^3$ posee una raíz múltiple en $x = 0$. Al ejecutar los distintos algoritmos, todos logran convergencia lineal con algunas diferencias. El algoritmo de bisección se aplicó en el intervalo $[-0{,}11; 0{,}1]$, logrando una tasa de $0{,}5$ en 4 iteraciones; con Newton-Raphson y $x_0 = 0{,}1$ se obtiene una tasa $0{,}666666$ en 23 iteraciones; con secante y el intervalo $[-0{,}11; 0{,}1]$ la convergencia se logra en 29 iteraciones con una tasa de $0{,}754877$; con Regula-Falsi se obtiene la raíz en 184 iteraciones y una tasa de $0{,}993773$ al utilizar el intervalo $[-0{,}11; 0{,}1]$ como intervalo inicial. En todos los casos, se utilizó una aritmética de 8 dígitos de mantisa y truncamiento. En la tabla 2.8 se muestran algunas de las divisiones realizas para calcular la tasa de convergencia. Para este ejemplo y con las condiciones dadas, el método más rápido es bisección.*

2.7. Fallos en la aplicación de los métodos iterativos

Los métodos iterativos ofrecen una manera simple de encontrar raíces de funciones continuas generando sucesiones de puntos que convergen a la raíz deseada. Sin embargo, esto no siempre se cumple. De los métodos vistos, el único que asegura la convergencia es el método de bisección, pues su desarrollo se basa en el concepto de completitud de $\mathbb{R}$. Por eso, se recomienda siempre acotar el intervalo donde se encuentra la raíz por medio de algunas iteraciones por bisección, para luego continuar con alguno de los otros métodos.

2.7.1. Divergencia en Punto Fijo

Cuando se presentó este método, se dieron ciertas condiciones para asegurar la convergencia de la secuencia iterativa. Uno de los problemas que se plantean es verificar el Teorema de Ostrowski en cada paso del proceso. Sin embargo, la principal dificultad radica en cómo se crea la función de iteración ó mapeo $\phi(x)$. Se puede hacer por medio de despejes, sumando el mismo término a ambos miembros, multiplicando ambos miembros por una constante y/o variable, entre otros procesos algebraicos:

Ejemplo 17. *Sea $f(x) = -2x^2 + 3x + 1$, $x_0 = 1{,}2$ y $\varepsilon = 1 \times 10^{-5}$. Existen varias opciones para crear $\phi(x)$, todas con diferentes resultados:*

- $x = \frac{-1+2x^2}{3}$, *converge a* $-0{,}280776$ *luego de* 14 *iteraciones.*
- $x = \frac{-1}{-2x+3}$, *converge a* $-0{,}280776$ *tras* 9 *iteraciones.*
- $x = \sqrt{\frac{3x+1}{2}}$, *converge a* $1{,}78078$ *después de* 14 *iteraciones.*
- $x = \frac{3x+1}{2x}$, *converge a* $1{,}78078$ *luego de* 8 *iteraciones.*
- $x = 2x^2 - 1 - 2x$, *no converge, no cumple con el Teorema de Ostrowski.*
- $x = \frac{2x^3}{3x+1}$, *converge a* 0 *tras* 6 *iteraciones. No es una raíz del problema original, la creación de* $\phi(x)$ *afectó los puntos de convergencia del método.*

2.7.2. Fallas en Newton-Raphson

El método de Newton-Raphson puede fallar de varias formas. Una de las formas más básicas en las que falla es la *finalización prematura* ó *breakdown*, que ocurre cuando para algún x_n se cumple que $f'(x_n) = 0$. Algebraicamente es imposible dividir por cero, pero geométricamente se trata de una recta tangente a la curva que es paralela al eje de las abscisas. Esto se observa en la figura 2.18, donde $f(x) = x^3 - 4x - 1$ y $x_0 = 0{,}838917$. Por lo tanto $x_1 = -1{,}15470$ y no se puede determinar x_2.

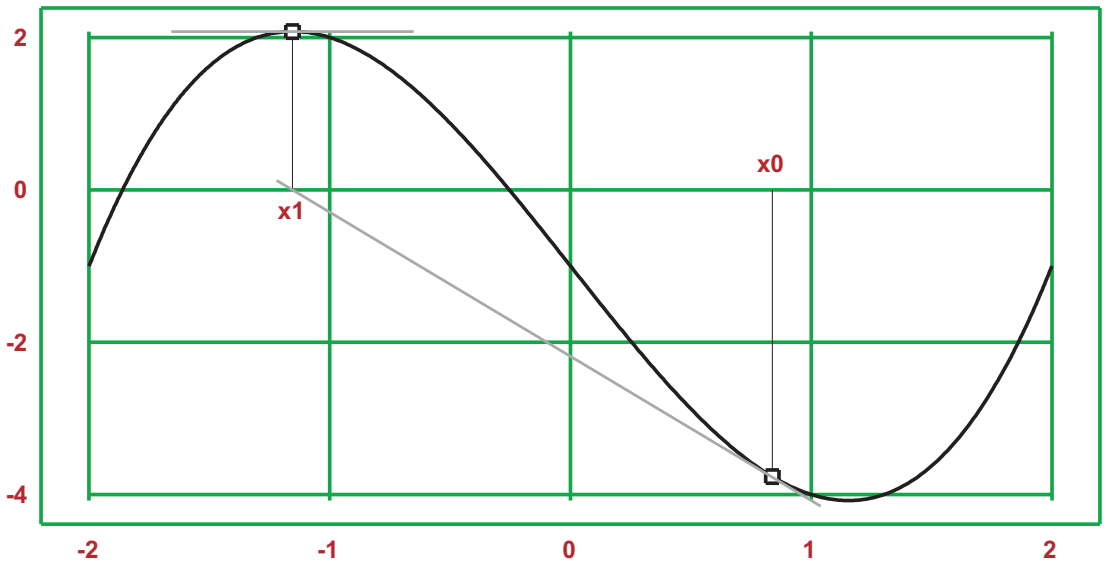

Figura 2.18: Finalización prematura del algoritmo Newton-Raphson, no se puede determinar x_2.

Otra las fallas que puede presentar este método es que la secuencia de iterados $\{x_n\}$ se aleje cada vez más de la raíz α, como se aprecia en la figura 2.19. En este caso, $f(x) = xe^{-x}$; $x_0 = 1{,}3$; $x_1 = 5{,}6333333$; $x_2 = 6{,}8491606$; $x_3 = 8{,}0201253$ y $x_4 = 9{,}1625728$ y así continúa la sucesión, siempre en forma creciente.

La última de las formas en la que el algoritmo de Newton-Raphson puede fallar al buscar una raíz es cuando el proceso iterativo genera una sucesión oscilante. En este caso, $f(x) = x^3 - 2x + 2$; $x_0 = 0$; $x_1 = 1$ y $x_2 = 0 = x_0$, con lo que el proceso iterativo no termina nunca. El gráfico de esta función está representado en la figura 2.20.

2.8. Ejercicios

1. Construir los siguientes algoritmos en PC:

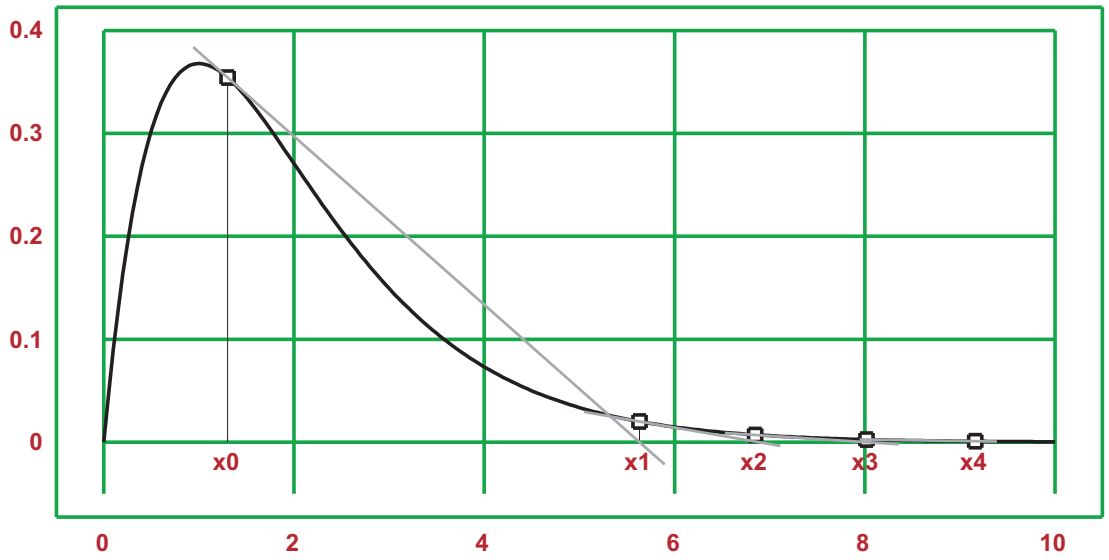

Figura 2.19: Divergencia del algoritmo Newton-Raphson, la sucesión crece indefinidamente.

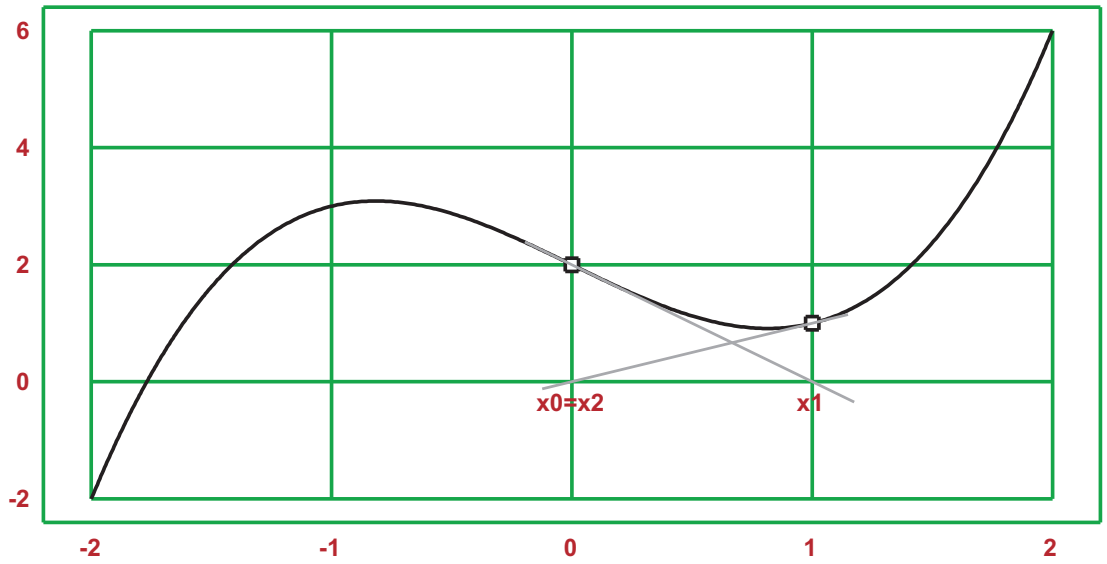

Figura 2.20: Oscilación de los sucesión generada por Newton-Raphson.

a) **Método de bisección**. Entrada: una función; dos valores de abscisa entre los cuales se encuentra la raíz; cantidad máxima de iteraciones; tolerancia para el stop del algoritmo. Salida: la raíz buscada ó un cartel que indique la no convergencia del método. Opcional: mostrar la tabla de convergencia y el gráfico de las iteraciones.

b) **Método de punto fijo**. Entrada: una función de iteración; un valor de abscisa inicial; cantidad máxima de iteraciones; tolerancia para detener el algoritmo. Salida: la raíz buscada ó un cartel que indique la no convergencia del método. Opcional: graficar las iteraciones.

c) **Método de Newton-Raphson**. Entrada: una función; un valor de abscisa para iniciar las iteraciones; cantidad máxima de iteraciones; error para detener el algoritmo. Salida: la raíz encontrada ó un cartel que anuncie la falla en la convergencia del método. Opcional: método de la secante.

2. Utilizando el método de Newton-Raphson y una calculadora básica (sólo operaciones de $+, -, \times$ y $\div$), calcular el valor de $\sqrt{75}$ con cuatro decimales correctos.

3. Encontrar la raíz real más pequeña de $x^3 - 3{,}23x^2 - 5{,}54x + 9{,}84 = 0$ a través del método de bisección en $[-3; 0]$. Iterar hasta que el error sea menor a 10^{-3}.

4. Probar el código del método de bisección creado en el ejercicio 1 con las siguientes funciones:

   a) $x^{-1} - \tan(x)$ en $[0; \pi/2]$
   b) $x^{-1} - 2^x$ en $[0; 1]$
   c) $2^{-x} + e^x + 2\cos(x) - 6$ en $[1; 3]$
   d) $(x^3 + 4x^2 + 3x + 5)/(2x^3 - 9x^2 + 18x - 2)$ en $[0; 4]$

5. Una de las raíces de $\cosh(x)\cos(x) - 1 = 0$ pertenece al intervalo $[4; 5]$:

   a) Calcularla utilizando los métodos de la secante y Newton-Raphson. En ambos casos usar como criterio de *stop* $\varepsilon = 10^{-2}$.
   b) Comparar los resultados.
   c) Mostrar gráficamente que el método de Newton-Raphson no converge a esta raíz si se utiliza $x_0 = 4$.

6. Una raíz de $\tan(x) - \tanh(x) = 0$ pertenece al intervalo $(7{,}0; 7{,}4)$. Encontrar dicha raíz con tres decimales correctos iterando a través del método de bisección y regula falsi.

7. Verificar los tipos de convergencia cuando se calcula la raíz de $f(x) = \cos(x)$, en el intervalo $[0; 4]$, utilizando $\varepsilon = 10^{-3}$, con los métodos de Newton-Raphson con $x_0 = 3$ y bisección con el intervalo inicial: $[1; 2]$.

8. Aproximar, con una tolerancia de $\varepsilon = 10^{-2}$, la raíz que $f(x) = \sqrt{2x+1} - \sqrt{3x}$ posee en el intervalo $[0, 2]$. Utilizar el método de la secante.

9. [**EMT**] Dada la ecuación no lineal $8x - \cos(x) - 2x^2 = 0$:

   a) Representarla gráficamente para averiguar el número de soluciones.
   b) Resolverla aplicando bisección y regula-falsi, $\varepsilon = 10^{-5}$.
   c) Transformarla en una ecuación de punto fijo de dos formas diferentes a fin de compararla con las resoluciones anteriores.

d) Calcular la tasa de convergencia de los métodos utilizados.

10. Dado el sistema no lineal:

$$x^2 + x - y^2 = 1$$
$$y - \sin(x^2) = 0$$

a) Determinar, gráficamente, las soluciones del sistema.

b) Transformarlo en una ecuación no lineal e iterar mediante el método de punto fijo hasta que $\varepsilon = 10^{-3}$. Contabilizar los puntos iniciales y la cantidad de iteraciones hasta llegar a la solución.

c) Repetir el inciso anterior utilizando el método de Newton-Raphson. Comprobar la convergencia cuadrática del método.

11. [**EMT**] Sea $f(x) = \cos^2(2x) - x^2$ una función definida en el intervalo $[0; 1{,}5]$. Tomando como tolerancia $\varepsilon = 1 \times 10^{-10}$ para el error absoluto, determinar en forma exploratoria para qué valores converge a la solución interna al intervalo al utilizar el método de Newton Raphson. **Sugerencia**: probar cerca de los extremos de definición de la función.

12. Investigar los puntos fijos de la función:

$$g(x) = \frac{1}{e^{-x-1}} + \frac{1}{e^{x-1}} - 10.$$

¿Cuál es el valor de $g'(x)$ en esos valores?

13. Aplicar el método de Newton-Raphson para determinar una de las raíces complejas de $z^2 + 1 = 0$. Utilizar $z_0 = 1 + i$.

14. Mostrar que el algoritmo de bisección obtiene una mejor precisión cuando calcula la raíz de $f(x) = (x - 1)^5$ utilizando la expresión original de f en vez de la expresión de Horner.

15. Encontrar la raíz de:

$$x^8 - 36x^7 + 546x^6 - 4536x^5 + 22449x^4 - 67284x^3 + 118124x^2 - 109584x + 40320,$$

utilizando el método de bisección en el intervalo $[5{,}5; 6{,}4]$. Luego cambiar el coeficiente -36 a $-36{,}001$ y repetir el ejercicio.

16. Encontrar el valor positivo más pequeño que puede utilizarse como semilla en el método de Newton-Raphson de forma tal que $f(x) = \tan^{-1}(x)$ diverge.

17. Crear una función iterativa para el método de Newton-Raphson con el fin de calcular $\sqrt[3]{R}$, donde $R > 0$. Por medio de un rápido análisis gráfico de la función creada, determinar algunos valores iniciales para los cuales el algoritmo es convergente.

18. El polinomio $p(x) = x^3 + 96x^2 - 199x - 294$ tiene raíces -1, 3 y -98. El punto $x_0 = 1$ debería ser un buen inicial para las raíces cercanas. Explicar qué pasa al ejecutar el algoritmo de Newton-Raphson.

19. [**EMT**] Para las siguientes funciones e intervalos, utilizar algoritmo que no dependan de la derivada para obtener una raíz:

a) $x^3 - 1$, en $[0; 10]$

b) $\tan(x) - 30x$, en $[1; 1{,}57]$
c) $x^2 - (1-x)^{10}$, en $[0; 1]$
d) $x^3 + 10^{-4}$, en $[-0{,}75; 0{,}5]$
e) xe^{-x^2}, en $[-1; 4]$

20. Mostrar que las siguientes funciones son contractivas en los intervalos indicados:

    a) $(1+x^2)^{-1}$ en un intervalo cerrado arbitrario
    b) $x/2$ en $[1; 5]$
    c) $\tan^{-1}(x)$ en un intervalo cerrado arbitrario que no contenga al 0

21. Con el método de punto fijo y la semilla $x_0 = 2{,}999$ iterar con una aritmética de 5 dígitos y hasta obtener uno de los puntos fijo del mapeo $F(x) = 2 + (x-2)^4$, en este caso $\alpha = 2$. ¿Qué orden de convergencia se obtiene?

22. Mostrar que la función $F(x) = 4x(1-x)$ mapea el intervalo $[0; 1]$ en sí misma y no es una contracción. Encontrar su punto fijo. ¿Por qué esto no contradice al Teorema de Banach?

23. Mostrar que la función $f(x) = 2 + x - \tan^{-1}(x)$ tiene la propiedad $|f'(x)| < 1$. Probar que f no tiene un punto fijo. Explicar por qué esto no contradice al Teorema del Mapeo Contractivo.

24. Comparar los algoritmos desarrollados en *EMT* para resolver la ecuación $p(x) = x^{10} - 1 = 0$, con una tolerancia de 10^{-6}. Utilizar, cuando la semilla sea un intervalo, al $[0; 1{,}5]$; cuando la semilla sea un valor inicial, a $x_0 = 0{,}5$.

25. La ecuación logística $x = ax\,(1-x)$, $a > 0$ es una ecuación de punto fijo que modela una población cuyo crecimiento está limitado. Es muy conocida porque los iterados pueden presentar un comportamiento caótico.

    a) Determinar analíticamente los puntos fijos de la ecuación logística.
    b) ¿Para qué valores de a es atrayente cada uno de los puntos fijos?
    c) ¿Para qué valores de a las iteraciones convergen cuadráticamente?
    d) Mostrar que, para valores del parámetro a entre 0 y 4, la función de punto fijo $g(x) = ax(1-x)$ aplica el intervalo $[0; 1]$ en sí mismo.
    e) Comprobar que, para $a = 0{,}5$, con cualquier estimación inicial en $[0; 1]$, los iterados convergen a 0. Verificar que la convergencia es lineal.
    f) Comprobar que, para $a = 2$, la convergencia es cuadrática.

Bibliografía

- *A theoretical introduction to numerical analysis**, V. RYABEN'KII y S. TSYNKOV, Cap.8
- *An introduction to numerical analysis*, Kendall ATKINSON, Cap.2
- *Análisis numérico - Un enfoque práctico*, M. MARON y R. LÓPEZ, Cap.2
- *Análisis numérico con aplicaciones*, C. GERALD y P. WHEATLEY, Cap.1
- *Numerical mathematics**, A. QUARTERONI, R. SACCO y F. SALERI, Cap.6
- *Numerical methods*, G. DAHLQUIST y A. BJÖRK, Cap.6

3 S.E.L. Métodos Directos

En este capítulo se trata de dar solución a un sistema de n ecuaciones lineales con n incógnitas. Es una de las ramas más importantes del cálculo numérico, puesto que es casi imposible desarrollar métodos discretos sin plantear sistemas de ecuaciones lineales. Más aún, los sistemas que se forman a partir de problemas físicos y/o químicos son generalmente grandes en dimensión, necesitando una gran cantidad de recursos como para ser resueltos en forma exacta. Para este tipo de problemas, generalmente involucrando matrices *sparse*[1], existen métodos especiales que se agrupan en una de las rutinas computacionales más conocidas: *LAPACK* (*Linear Algebra PACKage*), que fue programada en *Fortran77*.

Antes de comenzar a describir los métodos directos de resolución, que involucran una cantidad finita de pasos a desarrollar, se darán ciertas definiciones necesarias para comprender los conceptos en los que se basan.

3.1. Conceptos básicos

3.1.1. Unicidad de las soluciones

Un sistema de n ecuaciones lineales con n incógnitas:

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + \ldots + A_{1n}x_n &= b_1 \\ A_{21}x_1 + A_{22}x_2 + \ldots + A_{2n}x_n &= b_2 \\ A_{31}x_1 + A_{32}x_2 + \ldots + A_{3n}x_n &= b_3 \\ &\vdots \\ A_{n1}x_1 + A_{n2}x_2 + \ldots + A_{nn}x_n &= b_n, \end{aligned}$$

se puede escribir en forma matricial de la siguiente manera:

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

ó $\mathbf{Ax} = \mathbf{b}$, donde $\mathbf{A}$ es la matriz de coeficientes, $\mathbf{x}$ es el vector de incógnitas y $\mathbf{b}$ es el vector de constantes. Dicho sistema tiene solución única, si y sólo si el determinante de la matriz de coeficientes es *no singular*, es decir que $|\mathbf{A}| \neq 0$. Entonces las filas y

[1] una matriz se denomina *sparse* o *rala* si más de la mitad de sus elementos son ceros

columnas de dicha matriz son *linealmente independientes*, es decir que ninguna fila (o columna) es *combinación lineal* de otras filas (o columnas).

Si la matriz de coeficientes es *singular*, entonces el sistema tiene infinitas soluciones, ó ninguna solución, dependiendo del vector de constantes.

Ejemplo 18. *El sistema de ecuaciones:*

$$\begin{aligned} 2x + y &= 3 \\ 4x + 2y &= 6 \end{aligned}$$

tiene infinitas soluciones. La matriz de coeficientes es singular y una de las ecuaciones es combinación lineal de la otra. Es decir que cualquier solución de la primera de ellas es también solución de la otra y viceversa.

Ejemplo 19. *El sistema de ecuaciones:*

$$\begin{aligned} 2x + y &= 3 \\ 4x + 2y &= 0 \end{aligned}$$

no posee soluciones. La matriz de coeficientes es singular pero ninguna ecuación es combinación lineal de la otra. Es decir que ninguna de las soluciones que posee la primera ecuación satisface la segunda.

Ejercicio 6. *Graficar los sistemas de ecuaciones de los ejemplos anteriores para confirmar la existencia ó no de las soluciones.*

3.1.2. Normas vectoriales y matriciales

Una cuestión obvia que surge de analizar la unicidad de las soluciones en un sistema es: ¿qué pasa si la matriz de coeficientes es *casi* singular, es decir que $|\mathbf{A}|$ es muy pequeño? Un planteo lógico es ¿cuándo se dice que un determinante es pequeño? Es necesario tener un patrón de medida, un valor para poder comparar y decidir si el determinante es pequeño o no. Para ello, se lo puede comparar con alguna de las normas matriciales existentes y decir que el determinante es pequeño si:

$$|\mathbf{A}| \ll \|\mathbf{A}\|.$$

Las siguientes son las normas matriciales más utilizadas.

Definición 6. *Sea $\mathcal{V}$ un espacio vectorial sobre el campo $\mathbb{R}$ de los números reales. La función de valores no negativos $\|\cdot\|$ se denomina* ***norma del espacio*** *$\mathcal{V}$ si se cumplen los siguientes axiomas:*

- $\|\mathbf{v}\| = 0$, *si y sólo si* $\mathbf{v} = 0$ *y* $\mathbf{v} \in \mathcal{V}$.
- $\|\lambda\mathbf{v}\| = |\lambda|\|\mathbf{v}\|$, *para todo* $\lambda \in \mathbb{R}$ *y todo* $\mathbf{v} \in \mathcal{V}$.
- $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$, *para todos* $\mathbf{u}, \mathbf{v} \in \mathcal{V}$.

Un espacio vectorial $\mathcal{V}$, con una norma asociada, se denomina ***espacio vectorial normado****.*

Cualquier norma de un espacio vectorial $\mathcal{V} = \mathbb{R}^n$ se llama norma vectorial. Tres normas vectoriales son comunes en el álgebra numérica: la norma 1, la norma 2 (o norma euclideana) y la norma infinito.

Definición 7. *La **norma 1** del vector* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^T \in \mathbb{R}^n$ *se define como:*

$$\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|.$$

Definición 8. *La **norma 2** del vector* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^T \in \mathbb{R}^n$ *se define como:*

$$\|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^{n} |v_i|^2}.$$

En general, es posible definir la **norma p**, utilizando la misma noción aplicada en las dos normas anteriores:

Definición 9. *La **norma** $\mathbf{p}$ del vector* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^T \in \mathbb{R}^n$ *se define como:*

$$\|\mathbf{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p}.$$

En particular, si se toma el límite sobre p, cuando $p \to \infty$, de define la **norma infinito**, que es una de las más utilizadas debido a lo simple que es calcularla:

Definición 10. *La **norma infinito** del vector* $\mathbf{v} = (v_1, v_2, \ldots, v_n)^T \in \mathbb{R}^n$ *se define como:*

$$\|\mathbf{v}\|_\infty = \max_{i=1}^{n} |v_i|.$$

Cuando $n = 1$, cada una de estas normas se transforman en el valor absoluto, $|\cdot|$, el ejemplo más simple de una norma en $\mathbb{R}$.

Cualquier norma del espacio vectorial $\mathbb{R}^{n\times n}$ de matrices cuadradas de orden n será llamada **norma matricial**. En particular, se considerarán normas matriciales inducidas por normas vectoriales.

Definición 11. *Dada cualquier norma* $\|\cdot\|$ *en el espacio* $\mathbb{R}^n$ *de vectores n-dimensionales, la **norma matricial subordinada** ó **inducida** del espacio* $\mathbb{R}^{n\times n}$ *se define como:*

$$\|\mathbf{A}\| = \max_{\mathbf{v}\in\mathbb{R}^n_*} \frac{\|\mathbf{A}\mathbf{v}\|}{\|\mathbf{v}\|}.$$

Es fácil verificar que las normas matriciales verifican los axiomas planteados en la definición 6.

Dado cualquier vector $\mathbf{v}$ en $\mathbb{R}^n$, es un ejercicio trivial evaluar tres de las normas definidas, $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$ o $\|\mathbf{v}\|_\infty$. Sin embargo, no es tan obvio el cómo calcular las normas matriciales subordinadas correspondientes para una matriz $\mathbf{A} \in \mathbb{R}^{n\times n}$. La definición 11 no es útil para esta tarea, pues si se quiere calcular $\|\mathbf{A}\|$ es necesario maximizar la función $\mathbf{v} \to \frac{\|\mathbf{A}\mathbf{v}\|}{\|\mathbf{v}\|}$ sobre $\mathbb{R}^n_*$ (o lo que es equivalente, maximizar $\mathbf{w} \to \|\mathbf{A}\mathbf{w}\|$ sobre la esfera unitaria $\{\mathbf{w} \in \mathbb{R}^n : \|\mathbf{w}\| = 1\}$). Esta dificultad se elimina con los siguientes teoremas.

Teorema 7. *La norma matricial subordinada a la norma vectorial* $\|\cdot\|_1$ *puede ser expresada, para cualquier matriz* $n \times n$, $\mathbf{A} = (A_{ij})_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$, *como:*

$$\|\mathbf{A}\|_1 = \max_{j=1}^{n} \sum_{i=1}^{n} |A_{ij}|. \tag{3.1}$$

Este resultado es expresado en forma ligera diciendo que la norma 1 de una matriz es igual al mayor valor de las sumas de los valores absolutos de los elementos de las columnas.

Teorema 8. *La norma matricial subordinada a la norma vectorial* $\|\cdot\|_\infty$ *puede ser expresada, para cualquier matriz* $n \times n$, $\mathbf{A} = (A_{ij})_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$, *como:*

$$\|\mathbf{A}\|_\infty = \max_{i=1}^{n} \sum_{j=1}^{n} |A_{ij}|. \tag{3.2}$$

Este resultado es expresado en forma ligera diciendo que la norma infinito de una matriz es igual al mayor valor de las sumas de los valores absolutos de los elementos de las filas.

Así como se definieron axiomas para las normas vectoriales, existe uno para las normas matriciales.

Teorema 9. *Dada* $\|\cdot\|$, *una norma matricial subordinada en* $\mathbb{R}^{n\times n}$, *se cumple que:*

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|,$$

para cualquier par de matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$.

Comandos de EMT. *Los comandos para calcular normas son:*

- `norm(A:vector numérico, p:real no negativo)`, *donde* ***A*** *es vector;* ***p*** *es la norma que se desea calcular. Tener en cuenta que* ***p=0*** *no calcula la norma infinito, sino que identifica el mayor elemento del vector.*

- `Norm(A:matriz numérica, p:entero)`, *donde* ***A*** *es una matriz cuadrada;* ***p*** *es la norma que se desea calcular, los valores permitidos son 0 (norma infinito), 1 ó 2. El paquete* Calculo *debe cargarse previamente en memoria.*

Ejemplo en EMT 6. *Calcular, para el vector* $[1; -2; 3; -4]$, *las normas 1, 2 e infinito.*

```
>norm([1;-2;3;-4],1)
 10
>norm([1;-2;3;-4])
 5.47722557505
>norm(abs([1;-2;3;-4]),0)
 4
```

Ejemplo en EMT 7. *Calcular, para la matriz:*

$$\begin{bmatrix} 1 & 2 & 3 \\ -4 & 5 & -6 \\ 7 & 8 & 9 \end{bmatrix},$$

las normas 1, 2 e infinito.

```
>Norm([1,2,3;-4,5,-6;7,8,9],1)
 18
>Norm([1,2,3;-4,5,-6;7,8,9],2)
 14.841377283
>Norm([1,2,3;-4,5,-6;7,8,9],0)
 24
```

3.1.3. Condicionamiento de matrices

En la subsección anterior se definieron las normas matriciales. Su principal utilidad es poder calcular el número de condición de una matriz, es decir, qué tan sensible es a las perturbaciones cuando se trabaja con ella.

Sea $\mathbf{A} \in \mathbb{R}^{n\times n}$ tal que $\mathbf{A}^{-1}$ exista. El error entre la solución calculada $\tilde{\mathbf{x}}$ de $\mathbf{Ax} = \mathbf{b}$ y $\mathbf{x}$ es:

$$\mathbf{e} = \mathbf{x} - \tilde{\mathbf{x}}. \tag{3.3}$$

Es decir que $\mathbf{Ax} = \mathbf{b}$, pero generalmente $\mathbf{A}\tilde{\mathbf{x}} \neq \mathbf{b}$. Entonces se define el residuo:

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}. \tag{3.4}$$

Se ve que:

$$\mathbf{Ae} = \mathbf{Ax} - \mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}} = \mathbf{r}. \tag{3.5}$$

Así, $\mathbf{e} = \mathbf{A}^{-1}\mathbf{r}$. Es de notar que si $\mathbf{e} = \mathbf{0}$, entonces $\mathbf{r} = \mathbf{0}$, pero si los valores de $\mathbf{r}$ son pequeños, entonces $\mathbf{e}$ no necesariamente tiene valores pequeños puesto que $\mathbf{A}^{-1}$ debería tener valores grandes, haciendo que $\mathbf{A}^{-1}\mathbf{r}$ tenga valores grandes. En otras palabras, un residuo $\mathbf{r}$ pequeño no garantiza que $\tilde{\mathbf{x}}$ sea cercano a $\mathbf{x}$. A veces, $\mathbf{r}$ es calculado como un testeo superficial para comprobar si $\tilde{\mathbf{x}}$ es *razonable*. La principal ventaja de $\mathbf{r}$ es que siempre puede ser calculado mientras que $\mathbf{x}$ puede ser desconocido y por lo tanto $\mathbf{e}$ no puede ser calculado en forma exacta. Más allá de esto, puede mostrarse que considerar $\mathbf{r}$ en combinación con un número de condición es el método más efectivo de verificar cuán cercano es $\tilde{\mathbf{x}}$ a $\mathbf{x}$.

Como $\mathbf{e} = \mathbf{A}^{-1}\mathbf{r}$, se puede asegurar que $\|\mathbf{e}\|_p \leq \left\|\mathbf{A}^{-1}\right\|_p \|\mathbf{r}\|_p$. De la misma forma, $\mathbf{r} = \mathbf{Ae}$, por lo que $\|\mathbf{r}\|_p = \|\mathbf{Ae}\|_p \leq \|\mathbf{A}\|_p \|\mathbf{e}\|_p$. Así:

$$\frac{\|\mathbf{r}\|_p}{\|\mathbf{A}\|_p} \leq \|\mathbf{e}\|_p \leq \left\|\mathbf{A}^{-1}\right\|_p \|\mathbf{r}\|_p \tag{3.6}$$

De la misma forma, $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, entonces:

$$\frac{\|\mathbf{b}\|_p}{\|\mathbf{A}\|_p} \leq \|\mathbf{x}\|_p \leq \left\|\mathbf{A}^{-1}\right\|_p \|\mathbf{b}\|_p. \tag{3.7}$$

Si $\|\mathbf{x}\|_p \neq 0$, y $\|\mathbf{b}\|_p \neq 0$, entonces tomando recíprocos en (3.7) se obtiene:

$$\frac{1}{\|\mathbf{A}^{-1}\|_p \|\mathbf{b}\|_p} \leq \frac{1}{\|\mathbf{x}\|_p} \leq \frac{\|\mathbf{A}\|_p}{\|\mathbf{b}\|_p} \tag{3.8}$$

Ahora, multiplicando términos correspondientes en (3.8) y (3.6):

$$\frac{1}{\|\mathbf{A}^{-1}\|_p \|\mathbf{A}\|_p} \frac{\|\mathbf{r}\|_p}{\|\mathbf{b}\|_p} \leq \frac{\|\mathbf{e}\|_p}{\|\mathbf{x}\|_p} \leq \left\|\mathbf{A}^{-1}\right\|_p \|\mathbf{A}\|_p \frac{\|\mathbf{r}\|_p}{\|\mathbf{b}\|_p}. \tag{3.9}$$

Recordando que:

$$\epsilon_r = \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_p}{\|\mathbf{x}\|} = \frac{\|\mathbf{e}\|_p}{\|\mathbf{x}\|},$$

entonces:

$$\frac{1}{\|\mathbf{A}^{-1}\|_p \|\mathbf{A}\|_p} \frac{\|\mathbf{r}\|_p}{\|\mathbf{b}\|_p} \leq \epsilon_r \leq \left\|\mathbf{A}^{-1}\right\|_p \|\mathbf{A}\|_p \frac{\|\mathbf{r}\|_p}{\|\mathbf{b}\|_p}.$$

Se denomina **residuo relativo** a:

$$\frac{\|\mathbf{r}\|_p}{\|\mathbf{b}\|_p} = \frac{\|\mathbf{b} - \mathbf{A}\overline{\mathbf{x}}\|_p}{\|\mathbf{b}\|_p}, \tag{3.10}$$

y a partir de la expresión recién vista se define el número de condición, asociado a la norma p, de una matriz cuadrada como:

$$\mathcal{K}_p(\mathbf{A}) = \|\mathbf{A}\|_p \left\|\mathbf{A}^{-1}\right\|_p \tag{3.11}$$

Es cierto que $\mathcal{K}_p(\mathbf{A}) \geq 1$ para cualquier matriz cuadrada $\mathbf{A}$ y cualquier p válido. También puede verse que ϵ_r está entre $1/\mathcal{K}_p(\mathbf{A})$ y $\mathcal{K}_p(\mathbf{A})$ veces el residuo relativo. En particular, si $\mathcal{K}_p(\mathbf{A}) \gg 1$, incluso si el residuo relativo es pequeño, entonces ϵ_r debería ser grande. Por otro lado, si $\mathcal{K}_p(\mathbf{A})$ es cercano a la unidad, entonces ϵ_r deberá ser pequeño si el residuo relativo es pequeño. En conclusión, si $\mathcal{K}_p(\mathbf{A})$ es grande, debe prestarse atención a pequeñas perturbaciones en $\mathbf{A}$ y $\mathbf{b}$ que pueden generar un $\tilde{\mathbf{x}}$ muy distinto de $\mathbf{x}$. En forma equivalente, si $\mathcal{K}_p(\mathbf{A})$ es grande, entonces un valor pequeño de $\mathbf{r}$ no implica que $\tilde{\mathbf{x}}$ sea cercano a $\mathbf{x}$.

Ejemplo 20. *Sea* $\mathbf{A} \in \mathbb{R}^{2\times 2}$*, tal que:*

$$\mathbf{A} = \begin{bmatrix} 1 & 1-\varepsilon \\ 1 & 1 \end{bmatrix},$$

donde $|\varepsilon| \ll 1$*. Para calcular el número de condición es necesario conocer* $\mathbf{A}^{-1}$*. Es claro que:*

$$\mathbf{A}^{-1} = \frac{1}{\varepsilon}\begin{bmatrix} 1 & -1+\varepsilon \\ -1 & 1 \end{bmatrix}.$$

Se calculará la norma 1 de ambas matrices, resultado que puede conseguirse tomando la mayor suma, en valor absoluto, de las columnas:

$$\|\mathbf{A}\|_1 = \max\{2, |1-\varepsilon|+1\} \approx 2,$$

$$\left\|\mathbf{A}^{-1}\right\|_1 = \max\left\{\frac{2}{|\varepsilon|}, \left|1-\frac{1}{\varepsilon}\right| + \frac{1}{|\varepsilon|}\right\} \approx \frac{2}{|\varepsilon|},$$

entonces $\mathcal{K}_1(\mathbf{A}) \approx \dfrac{4}{|\varepsilon|}$.

Es de notar que, si $\varepsilon = 0$*, entonces* $\mathbf{A}^{-1}$ *no existe, por lo que el resultado anterior es razonable:*

$$\lim_{|\varepsilon|\to 0} \mathcal{K}_1(\mathbf{A}) = \infty$$

Si $\mathcal{K}(\mathbf{A}) \gg 1$, la matriz se denomina **mal condicionada**. Evidentemente, el número de condición de una matriz no se ve afectado por un rescalamiento al multiplicar sus elementos por una constante no nula. Tal cual se intuye, el número de condición de una matriz $\mathbf{A} \in \mathbb{R}^{n\times n}$ es fuertemente dependiente de la elección de norma a utilizar. Esto se muestra en el siguiente ejemplo.

Ejemplo 21. *Sea* $\mathbf{A}$ *una matriz triangular inferior de la siguiente forma:*

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

mientras que su inversa, $\mathbf{A}^{-1}$ *es:*

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Ahora, como $\|\mathbf{A}\|_1 = n$ *y* $\|\mathbf{A}^{-1}\|_1 = n$, *entonces* $\mathcal{K}_1(\mathbf{A}) = n^2$. *Por otro lado,* $\|\mathbf{A}\|_\infty = 2$ *y* $\|\mathbf{A}^{-1}\|_\infty = 2$, *entonces* $\mathcal{K}_\infty(\mathbf{A}) = 4 \ll n^2 = \mathcal{K}_1(\mathbf{A})$ *cuando* $n \gg 1$.

Ahora que se estableció cómo calcular el número de condición de una matriz y cuál es su significado, es tiempo de analizar cuánto afecta una perturbación en el vector de constantes a la solución de un sistema lineal.

Teorema 10. *Sea* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *una matriz no singular,* $\mathbf{b} \in \mathbb{R}^n_*$, $\mathbf{x} \in \mathbb{R}^n_*$, $\mathbf{Ax} = \mathbf{b}$ *y* $\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$, *con* $\delta\mathbf{x}$, $\delta\mathbf{b} \in \mathbb{R}^n$. *Entonces:*

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \mathcal{K}(\mathbf{A})\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

Demostración. Evidentemente, $\mathbf{b} = \mathbf{Ax}$ y:

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

Como $\mathbf{b} \neq 0$ y $\mathbf{A}$ es no singular, esto implica que $\mathbf{x} \neq 0$. Además:

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

y

$$\|\delta\mathbf{x}\| \leq \|\mathbf{A}^{-1}\|\|\delta\mathbf{b}\|.$$

El resultado buscado se consigue inmediatamente al multiplicar las desigualdades anteriores. □

Debido al efecto de los errores de redondeo durante los cálculos, la solución numérica de $\mathbf{Ax} = \mathbf{b}$ no será exacta. La solución numérica puede ser escrita como $\mathbf{x} + \delta\mathbf{x}$, y se desea que ese vector satisfaga la ecuación $\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$, donde los elementos de $\delta\mathbf{b}$ son muy pequeños. Si la matriz $\mathbf{A}$ tiene un número de condición grande, los elementos de $\delta\mathbf{x}$ pueden no ser pequeños.

Ejercicio 7. *Construir un script en Euler Math Toolbox que genere matrices de Hilbert y calcule su número de condición.*

Ejemplo 22. *La siguiente es la matriz aumentada de un sistema de ecuaciones lineales:*

$$\left[\begin{array}{cc|c} 0{,}69641 & 7{,}5585 & 41{,}275 \\ 5{,}2230 & 56{,}689 & 309{,}57 \end{array}\right]$$

Si se resuelve con la PC o trabajando en forma manual con una aritmética de 5 dígitos, con redondeo, se obtiene el siguiente resultado:

$$\begin{aligned} x_1 &= -66{,}198 \\ x_2 &= 11{,}560 \end{aligned}.$$

Sin embargo, la solución exacta es:

$$x_1 = 5$$

$$x_2 = 5.$$

Es imposible que un sistema de ecuaciones compatible determinado posea dos soluciones diferentes. Esto se debe al mal condicionamiento de la matriz del sistema, en este caso $\mathcal{K}(\mathbf{A}) = 4{,}45 \times 10^6$ *y* $|\mathbf{A}| = 7{,}41 \times 10^{-4}$ *lo que justifica el resultado incorrecto anterior.*

Comandos de EMT. *El comando para calcular el número de condición es:*

- `Cond(A:matriz numérica, p:entero)`, *donde* $\boldsymbol{A}$ *es una matriz cuadrada;* $\boldsymbol{p}$ *es la norma que se desea utilizar en el cálculo de condición, los valores válidos son 0 (norma infinito), 1 ó 2. El paquete* Calculo *debe cargarse previamente en memoria.*

Ejemplo en EMT 8. *Calcular, para la matriz:*

$$\begin{bmatrix} 1 & 2 & 3 \\ -4 & 5 & -6 \\ 7 & 8 & 9 \end{bmatrix},$$

el número de condición asociado a las normas 2 e infinito.

```
>Cond([1,2,3;-4,5,-6;7,8,9],2)
 14.6508186761
>Cond([1,2,3;-4,5,-6;7,8,9],0)
 25.2
```

3.2. Eliminación gaussiana

Una de las técnicas más conocidas para la resolución de sistemas de ecuaciones lineales, fue publicado por Carl Gauss en su *Theoria motus corporum coelestium in section-ibus conicis solem ambientium* (1809), sobre el movimiento de los cuerpos celestes. Sin embargo este método se encontró en escritos chinos de más de dos mil años de antigüedad. Este método ejecuta eliminaciones sucesivas sobre los elementos que están por debajo de la diagonal principal de la matriz de coeficientes. Una vez que se ejecutaron n eliminaciones, la matriz se vuelve triangular superior, con lo que el sistema se resuelve por sustitución hacia atrás. La técnica antes mencionada se denomina **eliminación gaussiana**.

Dado el sistema de ecuaciones lineales:

$$\begin{aligned} E_1 &= A_{11}x_1 + A_{12}x_2 + \ldots + A_{1n}x_n = b_1 \\ E_2 &= A_{21}x_1 + A_{22}x_2 + \ldots + A_{2n}x_n = b_2 \\ E_3 &= A_{31}x_1 + A_{32}x_2 + \ldots + A_{3n}x_n = b_3 \\ &\vdots \\ E_n &= A_{n1}x_1 + A_{n2}x_2 + \ldots + A_{nn}x_n = b_n \end{aligned}$$

se construye la **matriz aumentada del sistema**, que consta de la matriz de coeficientes a la que se anexa una última columna, correspondiente al vector de constantes:

$$[\mathbf{A}|\mathbf{b}] = \left[\begin{array}{cccc|c} A_{11} & A_{12} & \cdots & A_{1n} & b_1 \\ A_{21} & A_{22} & \cdots & A_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} & b_n \end{array}\right]$$

Para resolver el sistema se utilizan tres operaciones sobre filas (ecuaciones):

- La ecuación E_i puede multiplicarse por una constante $\lambda \neq 0$ y la ecuación resultante se emplea en vez de E_i. Esta operación se denota por:

$$E_i^* = \lambda E_i.$$

- La ecuación E_i puede sumarse a cualquier ecuación E_j y la ecuación resultante se emplea en vez de E_i. Esta operación se denota por:

$$E_i^* = E_i + E_j.$$

- El orden de las ecuaciones E_i y E_j puede intercambiarse. Esta operación se denota por:

$$E_i^* = E_j$$
$$E_j^* = E_i.$$

Ejemplo 23. *Sea el sistema de ecuaciones lineales:*

$$\begin{aligned} x_1 + x_2 + x_3 &= 6 \\ 2x_1 + 4x_2 + 3x_3 &= 16 \\ -x_1 + 5x_2 - 4x_3 &= -3 \end{aligned},$$

su matriz aumentada es:

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 2 & 4 & 3 & 16 \\ -1 & 5 & -4 & -3 \end{array}\right].$$

Para eliminar el primer coeficiente de la segunda fila, se hará $F_2^* = F_2 - 2F_1$. Es necesario hacer $F_3^* = F_3 + F_1$ para que el primer coeficiente de la tercera fila se convierta en cero. En la eliminación de los dos coeficientes antes nombrados al elemento A_{11} se lo denomina pivote. La matriz aumentada del sistema queda:

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ \boldsymbol{0} & 2 & 1 & 4 \\ \boldsymbol{0} & 6 & -3 & 3 \end{array}\right].$$

Para eliminar el segundo coeficiente de la tercera fila, se hará $F_3^* = F_3 - 3F_2$. La matriz aumentada del sistema queda:

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & 2 & 1 & 4 \\ 0 & \boldsymbol{0} & -6 & -9 \end{array}\right].$$

Si se reescribe el sistema de ecuaciones a partir de la matriz aumentada del sistema (ahora triangular superior), se obtiene:

$$\begin{aligned} x_1 + x_2 + x_3 &= 6 \\ 2x_2 + x_3 &= 4 \\ -6x_3 &= -9 \end{aligned},$$

este nuevo sistema, equivalente al original, puede ser resuelto por una simple sustitución hacia atrás. De dicho proceso se concluye que:

$$x_3 = \frac{3}{2}, \quad x_2 = \frac{5}{4}, \quad x_1 = \frac{13}{4}.$$

El método de eliminación gaussiana, si es desarrollado en forma algebraica, siempre obtiene resultados exactos. Sin embargo, al desarrollarlo en forma numérica, la precisión de la solución depende de la aritmética y principalmente del condicionamiento de la matriz de coeficientes. Existe un factor más que es fundamental al momento de resolver un sistema por medio del método de Gauss: la elección de los pivotes.

Ejemplo 24. *La siguiente es la matriz aumentada de un sistema de ecuaciones lineales:*

$$\left[\begin{array}{cc|c} 0{,}00031 & 5{,}171 & 3{,}002 \\ -7{,}123 & 9{,}874 & 4{,}896 \end{array}\right]$$

El sistema se resolverá utilizando una aritmética de 4 dígitos y truncamiento. Aplicando la operación de Gauss $F_2^* = F_2 - \frac{-7{,}123}{0{,}00031}F_1 \approx F_2 + 22970F_1$ *se elimina el primer elemento de la segunda fila:*

$$\left[\begin{array}{cc|c} 0{,}00031 & 5{,}171 & 3{,}002 \\ 0^* & 118700 & 68960 \end{array}\right]$$

Resolviendo ahora por sustitución hacia atrás, se obtiene:

$$\begin{aligned} x_2 &= 0{,}5809 \\ x_1 &= -5{,}915. \end{aligned}$$

Pero:

$$\mathbf{Ax} = \left[\begin{array}{c} 3{,}002 \\ 47{,}86 \end{array}\right] \neq \left[\begin{array}{c} 3{,}002 \\ 4{,}896 \end{array}\right] = \mathbf{b}$$

La *solución* obtenida no es tal. La causa no es el mal condicionamiento de la matriz de coeficientes, ya que $\mathcal{K}(\mathbf{A}) = 4{,}52$, sino que el factor usado en la operación de Gauss es demasiado grande para el sistema ($\frac{-7{,}123}{0{,}00031} \approx 22970$). Esto se corrige eligiendo como pivote al mayor elemento en valor absoluto de cada columna, eventualmente intercambiando filas para este cometido. Esta técnica de trabajo se denomina **pivoteo parcial**.

Nota. *El número* 0^* *representa un valor, muy pequeño en valor absoluto, cercano al epsilon de máquina. Es el redondeo forzado a cero para que el sistema se vuelva escalonado. De otra manera, con la aritmética empleada, nunca se podrá eliminar el primer elemento de la segunda fila.*

Ejemplo 25. *Para resolver el mismo sistema que en el ejemplo 24, se aplicará pivoteo parcial. Se intercambiarán* F_1 *y* F_2*. La matriz aumentada del sistema queda:*

$$\left[\begin{array}{cc|c} -7{,}123 & 9{,}874 & 4{,}896 \\ 0{,}00031 & 5{,}171 & 3{,}002 \end{array}\right]$$

Aplicando la operación de Gauss $F_2^* = F_2 - \frac{0{,}00031}{-7{,}123}F_1 \approx F_2 + 4{,}352 \times 10^{-5}F_1$ *se elimina el primer elemento de la segunda fila:*

$$\left[\begin{array}{cc|c} -7{,}123 & 9{,}874 & 4{,}896 \\ 0^* & 5{,}171 & 3{,}002 \end{array}\right]$$

Resolviendo ahora por sustitución hacia atrás, se obtiene la solución correcta (aceptable, en términos más estrictos de la aritmética) del sistema:

$$\begin{aligned} x_2 &= 0{,}5805 \\ x_1 &= 0{,}1173. \end{aligned}$$

El aplicar pivoteo es necesario cuando el sistema depende de un parámetro pequeño, puesto que las soluciones pueden variar dentro de un amplio rango, como lo muestra el siguiente ejemplo:

Ejemplo 26. *Sea el sistema de ecuaciones:*

$$\begin{aligned} -x_2 + x_3 &= 0 \\ -x_1 + 2x_2 - x_3 &= 0, \\ 2x_1 - x_2 &= 1 \end{aligned}$$

cuya solución es $x_1 = x_2 = x_3 = 1$. *Si se agrega el parámetro positivo* $\varepsilon \ll 1$ *en lugar de la primera variable de la primera ecuación, la matriz aumentada es:*

$$\left[\begin{array}{ccc|c} \varepsilon & -1 & 1 & 0 \\ -1 & 2 & -1 & 0 \\ 2 & -1 & 0 & 1 \end{array}\right],$$

donde aplicando operaciones de Gauss sin pivotear se obtiene:

$$\left[\begin{array}{ccc|c} \varepsilon & -1 & 1 & 0 \\ 0 & 2-\frac{1}{\varepsilon} & -1+\frac{1}{\varepsilon} & 0 \\ 0 & -1+\frac{2}{\varepsilon} & -\frac{2}{\varepsilon} & 1 \end{array}\right],$$

pero como un procesador opera con un largo de palabra de longitud finita, todos los números son redondeados. Como ε *es muy pequeño, entonces* $\frac{1}{\varepsilon}$ *es muy grande y, por ejemplo,* $2 - \frac{1}{\varepsilon} \approx -\frac{1}{\varepsilon}$ *obteniendo:*

$$\left[\begin{array}{ccc|c} \varepsilon & -1 & 1 & 0 \\ 0 & -\frac{1}{\varepsilon} & \frac{1}{\varepsilon} & 0 \\ 0 & \frac{2}{\varepsilon} & -\frac{2}{\varepsilon} & 1 \end{array}\right].$$

El sistema anterior no tiene solución, puesto que las ecuaciones representadas en las filas 2 y 3 de la matriz aumentada son combinación lineal.

Ejercicio 8. *Pivotear sobre el sistema anterior y resolverlo correctamente. Aplicar los redondeos que sean necesarios en las operaciones.*

Comandos de EMT. *Los comandos para efectuar operaciones de Gauss sobre una matriz son:*

- `swapRows(A:matriz numérica, i1:natural, i2:natural)`, *donde* ***A*** *es una matriz;* ***i1*** *e* ***i2*** *son las filas de la matriz que serán intercambiadas.*
- `pivotize(A:matriz numérica, i:entero, j:entero)`, *donde* ***A*** *es una matriz;* ***i*** *y* ***j*** *representan la posición del elemento que se transformará en el valor 1, el resto de la columna se transformará en 0.*
- `echelon(A:matriz numérica)`, *donde* ***A*** *es una matriz cuadrada que será transformada en una matriz escalonada reducida por filas.*

Ejemplo en EMT 9. *Dado el sistema de ecuaciones:*

$$\begin{aligned} 3x + 2y &= 1 \\ -5x - y &= 2, \end{aligned}$$

intercambiar las filas 1 y 2 de la matriz aumentada del sistema, pivotear sobre el elemento A_{11} *y luego sobre* A_{22}.

```
>A=swapRows([3,2,1;-5,-1,2],1,2)
            -5            -1             2
             3             2             1
>A=pivotize(A,1,1)
             1           0.2          -0.4
             0           1.4           2.2
>A=pivotize(A,2,2)
             1             0     -0.714286
             0             1       1.57143
```

Ejemplo en EMT 10. *Resolver el sistema de ecuaciones anterior a través de la reducción escalonada por filas, sin aplicar pivoteo.*

```
>echelon([3,2,1;-5,-1,2])
             1             0     -0.714286
             0             1       1.57143
```

3.3. Descomposición LU

Es posible mostrar que cualquier matriz cuadrada $\mathbf{A}$, bajo ciertas condiciones, puede ser expresada como el producto de una matriz triangular inferior $\mathbf{L}$ y una matriz triangular superior $\mathbf{U}$:

$$\mathbf{A} = \mathbf{LU}. \tag{3.12}$$

El proceso de computar $\mathbf{L}$ y $\mathbf{U}$ para una matriz dada $\mathbf{A}$ es conocido como **descomposición LU** o **factorización LU**. Esta descomposición no es única, a menos que se requieran ciertas condiciones extras para $\mathbf{L}$ o $\mathbf{U}$. Dichas restricciones permiten identificar una descomposición de otra. Las tres más comunes son:

- Si $L_{ii} = 1$, para $i = 1, 2, \ldots, n$, se denomina **descomposición de Doolittle**.
- Si $U_{ii} = 1$, para $i = 1, 2, \ldots, n$, se denomina **descomposición de Crout**.
- Si $\mathbf{L} = \mathbf{U}^T$, se denomina **descomposición de Choleski**.

Luego de descomponer la matriz $\mathbf{A}$, es muy fácil resolver el sistema $\mathbf{Ax} = \mathbf{b}$. Lo primero es rescribir el sistema de ecuaciones:

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{LUx} = \mathbf{b}.$$

Ahora, si se define $\mathbf{Ux} = \mathbf{y}$, el sistema a resolver es:

$$\mathbf{Ly} = \mathbf{b},$$

que es de resolución simple (por sustitución hacia adelante) ya que la matriz $\mathbf{L}$ es triangular inferior. Una vez conocido el vector $\mathbf{y}$, sólo resta resolver:

$$\mathbf{Ux} = \mathbf{y},$$

lo que se hace por sustitución hacia atrás, similar al desarrollo del método de eliminación gaussiana.

La principal ventaja de la descomposición LU con respecto al método de eliminación gaussiana es que, una vez descompuesta $\mathbf{A}$, se puede utilizar para resolver el sistema $\mathbf{Ax} = \mathbf{b}$ con tantos vectores de constantes $\mathbf{b}$ como se requiera. El costo de modificar el vector $\mathbf{b}$ es mínimo, ya que sólo queda hacer una sustitución hacia adelante y luego una hacia atrás.

3.3.1. Descomposición de Doolittle

Esta factorización opera con una matriz **U** similar a la que se obtiene por Gauss. Para mostrar esta relación, se ejemplificará con matrices de orden 3, sin perder generalidad.

Ejemplo 27. *Dada la matriz* **A** *de orden 3, que se puede descomponer por Doolittle, y dadas las matrices:*

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{bmatrix}$$

y

$$\mathbf{U} = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{bmatrix}$$

tal que $\mathbf{A} = \mathbf{LU}$*, con lo que:*

$$\mathbf{A} = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ U_{11}L_{21} & U_{12}L_{21} + U_{22} & U_{13}L_{21} + U_{23} \\ U_{11}L_{31} & U_{12}L_{31} + U_{22}L_{32} & U_{13}L_{31} + U_{23}L_{32} + U_{33} \end{bmatrix}. \tag{3.13}$$

Si se quiere reducir **A** *a través del método de eliminación gaussiana, es necesario aplicar operaciones sobre las filas. En este caso, para eliminar el primer coeficiente de* F_2 *y* F_3*, se hacen las siguientes operaciones:*

$$F_2^* = F_2 - L_{21}F_1$$

y

$$F_3^* = F_3 - L_{31}F_1.$$

El resultado es:

$$\mathbf{A}_1 = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & U_{22}L_{32} & U_{23}L_{32} + U_{33} \end{bmatrix}.$$

En el siguiente paso, se elije al elemento $A_{22} = U_{22}$ *como pivote y se efectúa la siguiente operación para eliminar el segundo coeficiente de la tercera fila:*

$$F_3^* = F_3 - L_{32}F_2.$$

Se obtiene $\mathbf{A}_2$*:*

$$\mathbf{A}_2 = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{bmatrix} = \mathbf{U}.$$

Del ejemplo anterior se observan dos hechos de la descomposición Doolittle:

- La matriz **U** es idéntica a la triangular superior que resulta de la eliminación gaussiana.
- Los elementos que se ubican por debajo de la diagonal principal de **L** son los factores por los que se multiplican los pivotes en la eliminación gaussiana de **A**. Es decir que $-L_{ij}$ es el factor utilizado para eliminar a A_{ij}.

Es común que se almacenen los factores utilizados por el método de Gauss en la parte triangular inferior de **U**. Los elementos de la diagonal principal de **L** no se almacenan, puesto que son todos 1. De acuerdo al ejemplo anterior y utilizando la notación habitual de muchos procesadores matemáticos, la descomposición queda:

$$[\mathbf{L}\backslash\mathbf{U}] = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ L_{21} & U_{22} & U_{23} \\ L_{31} & L_{32} & U_{33} \end{bmatrix}.$$

Ejercicio 9. *Utilizando descomposición LU resolver los sistemas de ecuaciones:*

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 4 \\ x_1 + 4x_2 + 9x_3 &= 5 \\ x_1 + 8x_2 + 27x_3 &= 6 \end{aligned}$$

y

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 1 \\ x_1 + 4x_2 + 9x_3 &= 2. \\ x_1 + 8x_2 + 27x_3 &= 3 \end{aligned}$$

Comandos de EMT. *El comando para realizar una descomposición LU es:*

- `{L,U,P}=LU(A:matriz numérica)`, *donde* ***A*** *es una matriz cuadrada;* ***L*** *es una matriz triangular inferior, su diagonal principal está compuesta por unos;* ***U*** *es una matriz triangular superior;* ***P*** *es una matriz de permutación tal que* $\mathbf{LU} = \mathbf{PA}$

Ejemplo en EMT 11. *Descomponer la matriz:*

$$\begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix}$$

de acuerdo al algoritmo de Doolitte. Pivotear si es necesario.

```
>{L,U,P}=LU([1,2;3,-4]);
>L
          1          0
  0.333333          1
>U
          3         -4
          0    3.33333
>P
          0          1
          1          0
```

3.3.2. Descomposición de Crout

El método de descomposición LU de Crout es similar al de Doolittle, salvo que en este caso la matriz que tiene su diagonal principal compuesta de números 1 es la **U**. Se mostrará en un ejemplo genérico de una matriz 3×3 cómo aplicar la descomposición.

Ejemplo 28. *Dada la matriz* **A** *de orden 3, que se puede descomponer según Crout, y dadas las matrices:*

$$\mathbf{L} = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix}$$

y

$$\mathbf{U} = \begin{bmatrix} 1 & U_{12} & U_{13} \\ 0 & 1 & U_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

tal que: $\mathbf{A} = \mathbf{LU}$*, con lo que:*

$$\mathbf{A} = \begin{bmatrix} L_{11} & L_{11}U_{12} & L_{11}U_{13} \\ L_{21} & L_{21}U_{12} + L_{22} & L_{21}U_{13} + L_{22}U_{23} \\ L_{31} & L_{31}U_{12} + L_{32} & L_{31}U_{13} + L_{32}U_{23} + L_{33} \end{bmatrix}.$$

Observando la estructura de $\mathbf{A} = \mathbf{LU}$*, se puede plantear una forma simple en la construcción de* $\mathbf{L}$ *y* $\mathbf{U}$*:*

- *Primera columna de* $\mathbf{L}$*:* $L_{i1} = A_{i1}$*,* $i = 1, 2, 3$*.*
- *Primera fila de* $\mathbf{U}$*:* $U_{1j} = \dfrac{A_{1j}}{L_{11}}$*,* $j = 2, 3$*.*
- *Segunda columna de* $\mathbf{L}$*:* $L_{i2} = A_{i2} - L_{i1}U_{12}$*,* $i = 2, 3$*.*
- *Segunda fila de* $\mathbf{U}$*:* $U_{2j} = \dfrac{A_{2j} - L_{21}U_{1j}}{L_{22}}$*,* $j = 3$*.*
- *Tercera fila de* $\mathbf{L}$*:* $L_{i3} = A_{i3} - \sum_{j=1}^{i-1} L_{ij}U_{ji}$*,* $i = 3$*.*

Ejercicio 10. *Desarrollar las fórmulas para obtener los coeficientes de las matrices* $\mathbf{L}$ *y* $\mathbf{U}$ *para una matriz* $\mathbf{A}$ *de orden* n*.*

3.4. Ejercicios

1. Construir los siguientes algoritmos en PC:

   a) **Normas vectoriales**. Entrada: un vector de longitud n; el tipo de norma elegida. Salida: la norma del vector.

   b) **Eliminación gaussiana**. Entrada: una matriz de orden n; un vector columna de longitud n. Salida: un vector columna, solución del sistema. Opcional: implementar pivoteo.

   c) **Descomposición LU - Doolittle**. Entrada: una matriz cuadrada. Salida: dos matrices triangulares, inferior y superior.

   d) **Número de condición**. Entrada: una matriz cuadrada. Salida: el número de condición de la matriz ingresada. Opcional: mostrar el cálculo de la inversa por eliminación gaussiana.

2. Analizar el condicionamiento de los sistemas de ecuaciones planteados y resolverlos utilizando descomposición LU (Crout y Doolittle):

   a)
   $$\begin{aligned} 3x_1 + x_2 &= 7 \\ 3x_1 + 1,0001x_2 &= 7,0001 \end{aligned}$$

   b)
   $$\begin{aligned} 0,003x_1 + x_2 &= 1,006 \\ 3x_1 + x_2 &= 7 \end{aligned}$$

3. Calcular la inversa de la matriz de Pascal de orden 3, utilizando eliminación Gaussiana.

4. Sea:
$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix},$$
graficar los siguientes conjuntos:

   a) $\{\mathbf{Ax}/\mathbf{x} \in \mathbb{R}^2 \wedge \|\mathbf{x}\|_1 = 1\}$.
   b) $\{\mathbf{Ax}/\mathbf{x} \in \mathbb{R}^2 \wedge \|\mathbf{x}\|_\infty = 1\}$.
   c) $\{\mathbf{Ax}/\mathbf{x} \in \mathbb{R}^2 \wedge \|\mathbf{x}\|_2 = 1\}$.

5. Dada la matriz:
$$\mathbf{A} = \begin{bmatrix} 1 & \delta \\ 0 & 1 \end{bmatrix},$$
donde $\delta > 0$. Verificar que $\mathcal{K}_\infty(\mathbf{A}) = \mathcal{K}_1(\mathbf{A}) = (1+\delta)^2$. ¿Es una matriz bien o mal condicionada?

6. Determinar cuáles de las siguientes expresiones definen normas en $\mathbb{R}^n$:

   a) $\text{máx}\{|x_2|, |x_3|, \ldots, |x_n|\}$
   b) $\sum_{i=1}^n |x_i|^3$
   c) $\{\sum_{i=1}^n |x_i|^{1/2}\}^2$
   d) $\sum_{i=1}^n 2^{-i}|x_i|$

7. Calcular el número de condición de las siguientes matrices. Para invertirlas utilizar el algoritmo de Gauss:

   a)
$$\mathbf{A} = \begin{bmatrix} 0{,}2436 & 0{,}4830 \\ 0{,}5361 & 0{,}2108 \end{bmatrix}$$
   b)
$$\mathbf{B} = \begin{bmatrix} 0{,}9423 & 0{,}1756 \\ 0{,}8626 & 0{,}6604 \end{bmatrix}$$
   c)
$$\mathbf{C} = \begin{bmatrix} 0{,}2339 & 0{,}0070 \\ 0{,}0035 & 0{,}2990 \end{bmatrix}$$
   d)
$$\mathbf{D} = \begin{bmatrix} 0{,}8045 & 0{,}3754 \\ 0{,}0189 & 0{,}0089 \end{bmatrix}$$

8. Sea $\mathbf{Ax} = \mathbf{b}$, donde:
$$0{,}485x_1 + 0{,}068x_2 = 0{,}621$$
$$0{,}729x_1 + 0{,}102x_2 = 0{,}933$$
cuya solución exacta es $\mathbf{x} = [1; 2]^T$. Resolver el sistema $(\mathbf{A} + \delta\mathbf{A})x = b$, tal que:
$$\delta\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0{,}001 & 0{,}002 \end{bmatrix}$$

9. Sea $\mathbf{Ax} = \mathbf{b}$, donde:

$$0{,}780x_1 + 0{,}563x_2 = 0{,}217$$
$$0{,}913x_1 + 0{,}659x_2 = 0{,}254$$

cuya solución exacta es $\mathbf{x} = [1; -1]^T$. Sean dos soluciones aproximadas $\mathbf{x_1} = [0{,}999; -1{,}001]$ y $\mathbf{x_2} = [1{,}01; -1{,}001]$. ¿Cuál tiene menor residuo?

10. Demostrar que una matriz estrictamente diagonal dominante por columnas tiene descomposición LU sin necesidad de pivoteo.

11. Calcular la inversa de las siguientes matrices utilizando descomposición LU:

a)

$$\mathbf{A} = \begin{bmatrix} 0{,}3959 & 0{,}6252 & 0{,}2368 \\ 0{,}5778 & 0{,}1467 & 0{,}3385 \\ 0{,}7697 & 0{,}7992 & 0{,}9020 \end{bmatrix}$$

b)

$$\mathbf{B} = \begin{bmatrix} 0{,}9061 & 0{,}7827 & 0{,}8737 \\ 0{,}8263 & 0{,}0453 & 0{,}2004 \\ 0{,}7405 & 0{,}7947 & 0{,}9935 \end{bmatrix}$$

c)

$$\mathbf{C} = \begin{bmatrix} 0{,}3517 & 0{,}3487 & 0{,}0223 \\ 0{,}4306 & 0{,}9512 & 0{,}8261 \\ 0{,}9450 & 0{,}6205 & 0{,}8545 \end{bmatrix}$$

12. Calcular las normas $\|\cdot\|_1$, $\|\cdot\|_2$ y $\|\cdot\|_\infty$ de la matriz:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ -1 & -5 \end{bmatrix}$$

13. ¿Son las matrices

a)

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

b)

$$\mathbf{B} = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 5 & 2 \\ 1 & 2 & 4 \end{bmatrix}$$

definidas positivas? Recordar que es necesario que $\mathbf{x}^T\mathbf{Ax} > 0$, para cualquier vector $\mathbf{x}$ distinto del vector nulo.

14. Dar un ejemplo de una matriz simétrica $\mathbf{A}$ que tenga todos sus elementos positivos pero que $\mathbf{x}^T\mathbf{Ax}$ sea, en ciertos valores, negativo.

15. Mostrar que la matriz:

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

no tiene factorización LU.

16. Mostrar que las matrices de la forma:

$$\mathrm{A} = \begin{bmatrix} 0 & 0 \\ a & b \end{bmatrix}$$

tienen factorización LU.

17. Si la matriz $\mathbf{A}$ es definida positiva, ¿puede asegurarse que $\mathbf{A}^{-1}$ es también definida positiva?

18. Calcular los números de condición de las siguientes matrices, utilizando norma 1, 2 e infinito:

a)

$$\mathbf{A} = \begin{bmatrix} a+1 & a \\ a & a-1 \end{bmatrix}$$

b)

$$\mathbf{B} = \begin{bmatrix} 0 & 1 \\ -2 & 0 \end{bmatrix}$$

c)

$$\mathbf{C} = \begin{bmatrix} c & 1 \\ 1 & 1 \end{bmatrix}$$

19. Demostrar que el número de condición de una matriz tiene la propiedad:

$$\mathcal{K}(\alpha\mathbf{A}) = \mathcal{K}(\mathbf{A}),$$

siempre y cuando $\alpha \neq 0$, para toda norma matricial.

20. Demostrar que, para cualquier vector $\mathbf{v} \in \mathbb{R}^n$, se verifica que $\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2$ y también $\|\mathbf{v}\|_2^2 \leq \|\mathbf{v}\|_1 \|\mathbf{v}\|_\infty$

21. Resolver el sistema de ecuaciones lineales $\mathbf{Ax} = \mathbf{b}$, donde:

$$\mathbf{A} = \begin{bmatrix} 1 & -0{,}01 \\ 2 & 0{,}01 \end{bmatrix}$$

y

$$\mathbf{b} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Luego, resolver $(\mathbf{A} + \delta\mathbf{A})\,\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$, donde:

a)

$$\delta\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0{,}005 \end{bmatrix}, \quad \delta\mathbf{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

b)

$$\delta\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & -0{,}03 \end{bmatrix}, \quad \delta\mathbf{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

c)

$$\delta\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & -0{,}02 \end{bmatrix}, \quad \delta\mathbf{b} = \begin{bmatrix} 0{,}10 \\ -0{,}05 \end{bmatrix}$$

22. Comprobar que la matriz $\mathbf{A} \in \mathbb{R}^{n\times n}$, tal que $A_{ii} = 1$, $A_{ij} = -1$ si $i < j$ y $A_{ij} = 0$ si $i > j$, tiene determinante igual a 1, pero $\mathcal{K}_\infty(\mathbf{A}) = n2^{n-1}$.

23. Determinar los coeficientes del polinomio $P(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ que pasa a través de los puntos $(0; 10)$, $(1; 35)$, $(3; 31)$ y $(4; 2)$.

24. [**EMT**] Sea la matriz:
$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 2 & \varepsilon \end{bmatrix},$$
donde $0 < \varepsilon \ll 1$. Calcular $\mathcal{K}_\infty(\mathbf{A})$.

25. ¿Qué condiciones pueden establecerse sobre los coeficientes a, b y c en:
$$\mathbf{A} = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix},$$
para que $\mathbf{A}$ sea definida positiva?

Bibliografía

- *A theoretical introduction to numerical analysis**, V. RYABEN'KII y S. TSYNKOV, Cap.5
- *An introduction to numerical analysis*, Kendall ATKINSON, Cap.8
- *Análisis numérico*, R. BURDEN y J. FAIRES, Cap.6
- *Análisis numérico - Un enfoque práctico*, M. MARON y R. LÓPEZ, Cap.3
- *Numerical methods*, G. DAHLQUIST y A. BJÖRK, Cap.5

4 S.E.L. Métodos Iterativos

4.1. Consideraciones generales

Una matriz $\mathbf{A} \in \mathbb{R}^{n\times n}$ se define como *sparse* si más de la mitad de sus n^2 elementos tienen el valor cero. Esa clase de matrices son utilizadas al momento de resolver problemas complejos, tales como la solución numérica de ecuaciones diferenciales en derivadas parciales. En el capítulo anterior se presentó la descomposición LU (en sus diferentes versiones) como una herramienta para resolver $\mathbf{Ax} = \mathbf{b}$ (considerando que la matriz $\mathbf{A}$ es no singular). Sin embargo, este procedimiento no supone mejora alguna en el caso de que $\mathbf{A}$ sea una matriz *sparse*. En este capítulo se consideran los métodos iterativos para determinar $\mathbf{x}$ en $\mathbf{Ax} = \mathbf{b}$. Siempre asumiendo que $\mathbf{A}^{-1}$ existe, con los métodos iterativos se creará una sucesión de Cauchy de vectores $\mathbf{x}^{(k)}$ que convergen a $\mathbf{x}$. Los métodos iterativos son particularmente ventajosos cuando $\mathbf{A}$ es *sparse* y/o de gran tamaño. Esto es porque los métodos directos a menudo necesitan un movimiento considerable de información a través del sistema de memoria de las computadoras y esto puede tornar muy lentos los procesos de cómputo. Una buena implementación de los métodos iterativos puede evitar este inconveniente.

El objetivo de los métodos iterativos es crear una metodología que permita generar una sucesión de vectores $\mathbf{x}^{(k)}$ de forma tal que:

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x}, \tag{4.1}$$

donde $\mathbf{x} = [x_0, x_1, \ldots, x_{n-1}]^T \in \mathbb{R}$. La idea básica es encontrar un operador T de forma tal que $\mathbf{x}^{(k+1)} = T(\mathbf{x}^{(k)})$, para $k = 0, 1, 2, \ldots$. Como la sucesión formada es de Cauchy, entonces para todo $\varepsilon > 0$ existe un $m \in \mathbb{N}$ tal que $\|\mathbf{x}^{(m)} - \mathbf{x}\| < \varepsilon$. El operador T se define de acuerdo a:

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{f}, \tag{4.2}$$

donde $\mathbf{x}^0 \in \mathbb{R}^n$ es el valor inicial o *semilla*, $\mathbf{B} \in \mathbb{R}^{n\times n}$ se denomina *matriz de iteración* y $\mathbf{f} \in \mathbb{R}^n$ se deriva de $\mathbf{A}$ y $\mathbf{b}$. Como se debe mantener (4.1), a partir de (4.2) se buscarán $\mathbf{B}$ y f de forma tal que $\mathbf{x} = \mathbf{Bx} + \mathbf{f}$ o $\mathbf{A}^{-1}\mathbf{b} = \mathbf{BA}^{-1}\mathbf{b} + \mathbf{f}$. Entonces:

$$\mathbf{f} = (\mathbf{I} - \mathbf{B})\mathbf{A}^{-1}\mathbf{b}. \tag{4.3}$$

El vector de error en el paso k se define como:

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}, \tag{4.4}$$

y naturalmente se desea que $\lim_{k\to\infty} \mathbf{e}^{(k)} = 0$. La convergencia será dentro de alguna norma seleccionada.

El problema es que no hay garantía de que (4.1) se mantenga con el paso de las iteraciones. La convergencia sólo será posible a través de una selección apropiada de la matriz $\mathbf{B}$, y para las matrices $\mathbf{A}$ que cumplan ciertas condiciones.

Definición 12 (Radio Espectral)**.** *Sea $s(\mathbf{A})$ el conjunto de autovalores de la matriz $\mathbf{A} \in \mathbb{R}^{n\times n}$. El radio espectral de $\mathbf{A}$ es:*

$$\rho(\mathbf{A}) = \max_{\lambda \in s(\mathbf{A})} |\lambda|.$$

Definición 13 (Convergencia de Matrices)**.** *La secuencia de matrices $(\mathbf{A}^{(k)})$, con $\mathbf{A}^{(k)} \in \mathbb{R}^{n\times n}$, converge a $\mathbf{A} \in \mathbb{R}^{n\times n}$ si y sólo si:*

$$\lim_{k\to\infty} \|\mathbf{A} - \mathbf{A}^{(k)}\| = 0. \tag{4.5}$$

Definición 14 (Convergencia de Matrices II)**.** *La matriz $\mathbf{A}$ es convergente si, para alguna norma matricial, se cumple que $\|\mathbf{A}\| < 1$.*

La norma utilizada en (4.5) es arbitraria debido a la equivalencia de normas.

Teorema 11. *Sea $\mathbf{A} \in \mathbb{R}^{n\times n}$, entonces:*

$$\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0} \Leftrightarrow \rho(\mathbf{A}) < 1. \tag{4.6}$$

También, la serie geométrica de matrices $\sum_{k=0}^{\infty} \mathbf{A}^k$ converge si y sólo si $\rho(\mathbf{A}) < 1$. En este caso:

$$\sum_{k=0}^{\infty} \mathbf{A}^k = (\mathbf{I} - \mathbf{A})^{-1}. \tag{4.7}$$

Así, si $\rho(\mathbf{A}) < 1$, entonces la matriz $\mathbf{I} - \mathbf{A}$ es invertible, y también:

$$\frac{1}{1 + \|\mathbf{A}\|} \leq \|(\mathbf{I} - \mathbf{A})^{-1}\| \leq \frac{1}{1 - \|\mathbf{A}\|}, \tag{4.8}$$

donde $\|\cdot\|$ es una norma inducida de matriz de forma tal que $\|\mathbf{A}\| < 1$.

Demostración. Primero se demostrará que (4.6) se cumple. Sea $\rho(\mathbf{A}) < 1$ por lo que existe $\varepsilon > 0$ de forma tal que $\rho(\mathbf{A}) < 1 - \varepsilon$ y por lo tanto habrá una norma matricial tal que:

$$\|\mathbf{A}\| \leq \rho(\mathbf{A}) + \varepsilon < 1.$$

Como $\|\mathbf{A}^k\| \leq \|\mathbf{A}\|^k < 1$, y aplicando la definición de convergencia, como $k \to \infty$, se tiene que $\mathbf{A}^k \to 0 \in \mathbb{R}^{n\times n}$. De forma inversa, si se asume que $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$, y λ es un autovalor de $\mathbf{A}$, para cualquier autovector $\mathbf{x} \neq 0$ de $\mathbf{A}$ asociado al autovalor λ, se tiene que $\mathbf{A}^k\mathbf{x} = \lambda^k\mathbf{x}$ y entonces $\lim_{k\to\infty} \lambda^k = 0$. Así, $|\lambda| < 1$, y por lo tanto $\rho(\mathbf{A}) < 1$. Esto prueba (4.7).

Si λ es un autovalor de $\mathbf{A}$, entonces $1 - \lambda$ es un autovalor de $\mathbf{I} - \mathbf{A}$. Se observa entonces que:

$$(\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \ldots + \mathbf{A}^{n-1} + \mathbf{A}^n) = \mathbf{I} - \mathbf{A}^{n+1}. \tag{4.9}$$

Como $\rho(\mathbf{A}) < 1$, entonces $\mathbf{I} - \mathbf{A}$ tiene inversa, y tomando $n \to \infty$ en (4.9) se sigue que:

$$(\mathbf{I} - \mathbf{A}) \sum_{k=0}^{\infty} \mathbf{A}^k = \mathbf{I}.$$

Esto prueba (4.7).

Ahora, sabiendo que cualquier norma matricial cumple que $\|\mathbf{I}\| = 1$, se tiene que:

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

lo que lleva directamente a la primera inecuación de (4.8). Ahora como $\mathbf{I} = (\mathbf{I}-\mathbf{A})+\mathbf{A}$, entonces:

$$(\mathbf{I} - \mathbf{A})^{-1} = \mathbf{I} + \mathbf{A}(\mathbf{I} - \mathbf{A})^{-1},$$

lo que lleva a:

$$(\mathbf{I} - \mathbf{A})^{-1} \leq 1 + \|\mathbf{A}\|\|(\mathbf{I} - \mathbf{A})^{-1}\|.$$

La condición $\|\mathbf{A}\| < 1$ implica que se cumple la segunda condición de (4.8). Esto demuestra el teorema. □

Ahora sí están dadas las condiciones para establecer bajo qué condiciones de $\mathbf{B}$ el sistema iterativo (4.2) es convergente.

Teorema 12. *Sea $f \in \mathbb{R}^n$ tal que cumple (4.3). Entonces $(\mathbf{x}^{(k)})$ converge a $\mathbf{x}$ satisfaciendo $\mathbf{Ax} = \mathbf{b}$ para cualquier $\mathbf{x}^{(0)}$ si y sólo si $\rho(\mathbf{B}) < 1$.*

Demostración. De (4.2), (4.3) y (4.4) se tiene que:

$$\begin{aligned}
\mathbf{e}^{(k+1)} &= \mathbf{x}^{(k+1)} - \mathbf{x} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{f} - \mathbf{x} = \mathbf{B}\mathbf{x}^{(k)} + (\mathbf{I} - \mathbf{B})\mathbf{A}^{-1}\mathbf{b} - \mathbf{x} \\
&= \mathbf{B}\mathbf{e}^{(k)} + \mathbf{B}\mathbf{x} + (\mathbf{I} - \mathbf{B})\mathbf{A}^{-1}\mathbf{b} - \mathbf{x} \\
&= \mathbf{B}\mathbf{e}^{(k)} + \mathbf{B}\mathbf{x} + \mathbf{x} - \mathbf{B}\mathbf{x} - \mathbf{x} \\
&= \mathbf{B}\mathbf{e}^{(k)}.
\end{aligned}$$

A partir de la última igualdad se sigue que:

$$\mathbf{e}^{(k)} = \mathbf{B}^k\mathbf{e}^{(0)} \tag{4.10}$$

para todo $k \in \mathbb{N}$. Del teorema 11 se sabe que:

$$\lim_{k\to\infty} \mathbf{B}^k\mathbf{e}^{(0)} = 0$$

para todo $\mathbf{e}^{(0)} \in \mathbb{R}^n$ si y sólo si $\rho(\mathbf{B}) < 1$. Por otra parte, si se supone que $\rho(\mathbf{B}) \geq 1$, entonces existe al menos un autovalor λ de $\mathbf{B}$ tal que $|\lambda| \geq 1$. Sea $\mathbf{e}^{(0)}$ el autovector asociado a λ, entonces $\mathbf{B}\mathbf{e}^{(0)} = \lambda\mathbf{e}^{(0)}$, lo que implica que $\mathbf{e}^{(k)} = \lambda^k\mathbf{e}^{(0)}$. Pero esto implica que $\mathbf{e}^{(k)} \not\to 0$ cuando $k \to \infty$ puesto que $|\lambda| \geq 1$. □

El teorema 12 brinda una condición general sobre $\mathbf{B}$ de forma tal que el proceso iterativo (4.2) converja. Sin embargo, el problema que continúa sin solución es cómo encontrar la matriz $\mathbf{B}$. Lo único que se asegura según el teorema 11 es que una condición suficiente para la convergencia es que $\|\mathbf{B}\| < 1$, para cualquier norma matricial. Una forma general de construcción de matrices para usar en métodos iterativos es la *división aditiva* de la matriz $\mathbf{A}$ de acuerdo a:

$$\mathbf{A} = \mathbf{P} - \mathbf{N}, \tag{4.11}$$

donde $\mathbf{P}, \mathbf{N} \in \mathbb{R}^{n\times n}$ son matrices especiales y particularmente $\mathbf{P}^{-1}$ existe. La matriz $\mathbf{P}$ a veces es llamada *matriz precondicionante*[1]. Para ser específicos, se reescribe (4.2) como:

$$\mathbf{x}^{(k+1)} = \mathbf{P}^{-1}\mathbf{N}\mathbf{x}^{(k)} + \mathbf{P}^{-1}\mathbf{b},$$

[1] la explicación de este nombre se desarrolla en el libro de Quarteroni

esto es, para $k \in \mathbb{N}_0$:

$$\mathbf{P}\mathbf{x}^{(k+1)} = \mathbf{N}\mathbf{x}^{(k)} + \mathbf{b}, \tag{4.12}$$

de forma tal que $f = \mathbf{P}^{-1}\mathbf{b}$, y $\mathbf{B} = \mathbf{P}^{-1}\mathbf{N}$. De forma alternativa:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{P}^{-1}[\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}], \tag{4.13}$$

donde $\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)} = \mathbf{r}^{(k)}$ es el vector residual en el paso k. A partir de (4.13) se ve que para obtener $\mathbf{x}^{(k+1)}$ es necesario resolver un sistema de ecuaciones lineales donde se involucra a $\mathbf{P}$. Claramente, $\mathbf{P}$ debe ser no singular, y fácil de invertir para no volver los cálculos más complejos aún.

4.1.1. Método de Jacobi o de Iteraciones simultáneas

Sumado a lo planteado anteriormente para las matrices iterativas, se agregará una suposición: todos los elementos en la diagonal principal de $\mathbf{A}$ son distintos de cero. En este caso, es posible además expresar $\mathbf{A}\mathbf{x} = \mathbf{b}$ como:

$$x_i = \frac{1}{A_{ii}} \left[b_i - \sum_{\substack{j=0 \\ j \neq i}}^{n-1} A_{ij} x_j \right] \tag{4.14}$$

para $i = 0, 1, \ldots, n-1$. La expresión anterior se puede suponer como una secuencia iterativa, de forma que:

$$x_i^{(k+1)} = \frac{1}{A_{ii}} \left[b_i - \sum_{\substack{j=0 \\ j \neq i}}^{n-1} A_{ij} x_j^{(k)} \right], \tag{4.15}$$

para $i = 0, 1, \ldots, n-1$. Dicha expresión se denomina **Método de Jacobi**. Es fácil mostrar que este algoritmo implementa la siguiente división:

$$\mathbf{P} = \mathbf{D}, \quad \mathbf{N} = \mathbf{D} - \mathbf{A} = \mathbf{L} + \mathbf{U}, \tag{4.16}$$

donde $\mathbf{D} = diag(A_{11}, A_{22}, \ldots, A_{nn})$, $\mathbf{L}$ es la matriz triangular inferior tal que $L_{ij} = -A_{ij}$ si $i > j$ y $L_{ij} = 0$ si $i \leq j$, y $\mathbf{U}$ es la matriz triangular superior de forma tal que $U_{ij} = -A_{ij}$ si $j > i$, y $U_{ij} = 0$ si $j \leq i$. De aquí que la matriz de iteración está dada por:

$$\mathbf{B} = \mathbf{B}_J = \mathbf{P}^{-1}\mathbf{N} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}) = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}. \tag{4.17}$$

Ejemplo 29. *Sea el sistema de ecuaciones* $\mathbf{A}x = \mathbf{b}$*, donde:*

$$\mathbf{A} = \begin{bmatrix} 12 & 3 & 4 \\ 5 & 17 & 4 \\ 3 & 1 & 9 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}.$$

Si se resuelve el sistema en forma matricial, entonces:

$$\mathbf{D} = \begin{bmatrix} 12 & 0 & 0 \\ 0 & 17 & 0 \\ 0 & 0 & 9 \end{bmatrix},$$

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 \\ -5 & 0 & 0 \\ -3 & -1 & 0 \end{bmatrix}$$

y

$$\mathbf{U} = \begin{bmatrix} 0 & -3 & -4 \\ 0 & 0 & -4 \\ 0 & 0 & 0 \end{bmatrix}.$$

Entonces, $\mathbf{B} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$*, con lo que:*

$$\mathbf{B} = \begin{bmatrix} 0 & -0{,}2500 & -0{,}3333 \\ -0{,}2941 & 0 & -0{,}2353 \\ -0{,}3333 & -0{,}1111 & 0 \end{bmatrix}$$

y

$$\mathbf{f} = \begin{bmatrix} 0{,}08333 \\ -0{,}1177 \\ 0{,}3333 \end{bmatrix}.$$

Ejercicio 11. *Implementar en PC las iteraciones para obtener la solución del sistema utilizado en el ejemplo 29, utilizar* $\epsilon = 0{,}0001$ *como condición de finalización.*

Ejemplo 30. *Sea el sistema de ecuaciones lineales utilizado en el ejemplo 29. Ahora se resolverá utilizando la versión lineal del método de Jacobi. De acuerdo a (4.15), se tiene:*

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{12}\left[1 - \left(3x_2^{(k)} + 4x_3^{(k)}\right)\right] \\ x_2^{(k+1)} &= \frac{1}{17}\left[-2 - \left(5x_1^{(k)} + 4x_3^{(k)}\right)\right] \\ x_3^{(k+1)} &= \frac{1}{9}\left[3 - \left(3x_1^{(k)} + x_2^{(k)}\right)\right] \end{aligned}.$$

Ejercicio 12. *Implementar en PC las iteraciones para obtener la solución del sistema utilizado en el ejemplo 30, utilizar* $\epsilon = 0{,}0001$ *como condición de finalización.*

4.1.2. Método de Gauss-Seidel o de Iteraciones sucesivas

Una alternativa al método de Jacobi es el método de Gauss-Seidel. La idea central es la misma que se utilizó en (4.15), pero esta vez se utilizan las versiones *actualizadas* de las variables. Es decir que para calcular la variable j en el paso $k+1$-ésimo se utilizarán las versiones $k+1$-ésimas de las variables cuyo subíndice sea menor que j. Si en subíndice es mayor que j, se debe usar la versión k-ésima. Entonces, las iteraciones sobre x_i que definen al método de Gauss-Seidel son:

$$x_i^{(k+1)} = \frac{1}{A_{ii}}\left[b_i - \sum_{j=0}^{i-1} A_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n-1} A_{ij}x_j^{(k)}\right], \tag{4.18}$$

donde $i = 0, 1, \ldots, n-1$. Aprovechando la notación matricial que se definió en la sección anterior, (4.18) puede ser expresado como:

$$\mathbf{D}x^{(k+1)} = \mathbf{b} + \mathbf{L}x^{(k+1)} + \mathbf{U}x^{(k)}, \tag{4.19}$$

donde $\mathbf{D}$, $\mathbf{L}$ y $\mathbf{U}$ son las mismas matrices que las asociadas al método de Jacobi. En el método de Gauss-Seidel se implementa la siguiente división:

$$\mathbf{P} = \mathbf{D} - \mathbf{L}, \quad \mathbf{N} = \mathbf{U}, \tag{4.20}$$

con la matriz de iteración:

$$\mathbf{B} = \mathbf{B}_{GS} = (\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}. \tag{4.21}$$

Ejemplo 31. *Sea el sistema de ecuaciones* $\mathbf{Ax} = \mathbf{b}$ *y las matrices* $\mathbf{D}$, $\mathbf{L}$ *y* $\mathbf{U}$ *definidas en el ejemplo 29. Entonces, por el método de Gauss-Seidel,* $\mathbf{B} = (\mathbf{D} - \mathbf{L})^{-1}\,\mathbf{U}$, *por lo que:*

$$\mathbf{B} = \begin{bmatrix} 0 & -0{,}2500 & -0{,}3333 \\ 0 & 0{,}07353 & -0{,}1373 \\ 0 & 0{,}07516 & 0{,}1264 \end{bmatrix}$$

y

$$\mathbf{f} = \begin{bmatrix} 0{,}08333 \\ -0{,}1422 \\ 0{,}3214 \end{bmatrix}$$

Ejercicio 13. *Modificar la ecuaciones planteadas en el ejercicio 30 para adaptarlas al método de Gauss-Seidel e iterar hasta que el error relativo sea menor a* 0,0001.

Comandos de EMT. *El comando para resolver un sistema de ecuaciones a través del método de Gauss-Seidel es:*

- `seidel(A:real, b:vector columna, x:vector columna, om:número)`, *donde* ***A*** *es vector;* ***p*** *es la norma que se desea calcular. Tener en cuenta que* ***p=0*** *no calcula la norma infinito, sino que identifica el mayor elemento del vector.*

- `Norm(A,p)`, *donde* ***A*** *es una matriz cuadrada;* ***p*** *es la norma que se desea calcular, los valores permitidos son 0 (norma infinito), 1 ó 2. El paquete* Calculo *debe cargarse previamente en memoria.*

Ejemplo en EMT 12. *Calcular, para el vector* $[1; -2; 3; -4]$, *las normas 1, 2 e infinito.*

```
>norm([1;-2;3;-4],1)
 10
>norm([1;-2;3;-4])
 5.47722557505
>norm(abs([1;-2;3;-4]),0)
 4
```

4.1.3. Sobre la matriz de iteración

Para los dos métodos planteados anteriormente se desarrollaron ejemplos, primero utilizando la matriz de iteración $\mathbf{B}$ y luego con la versión lineal de cada método. Es de notar que ambos esquemas repiten los iterados en cada paso y, como es lógico, convergen al mismo vector solución con el paso de las iteraciones. Sin embargo, el esquema que depende de la matriz de iteración $\mathbf{B}$, es poco realista desde el punto de vista de la implementación. En la construcción del vector estacionario $\mathbf{f}$ interviene $\mathbf{A}^{-1}$, sin embargo, si se dispone de esa matriz inversa, la resolución del sistema es directamente $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. Por lo tanto, la creación de la matriz $\mathbf{B}$ se restringe sólo a su utilización en el análisis de convergencia. Al momento de la implementación, deben seguirse el esquema general 4.13, donde la matriz precondicionante $\mathbf{P}$ es simple de obtener teniendo en cuenta la división aditiva.

Ejercicio 14. *Verificar que los iterados en la resolución de un sistema son los mismos, sin importar el uso de la matriz precondicionante; la matriz de iteración* $\mathbf{B}$ *ó la versión lineal de los métodos vistos.*

4.2. Control de los métodos iterativos

De acuerdo a lo visto en las secciones anteriores, los métodos de Jacobi y Gauss-Seidel se utilizan para resolver sistemas de ecuaciones lineales efectuando iteraciones sobre un vector inicial. La convergencia del sistema ocurre de acuerdo a (4.15) y (4.18). Ahora, se puede controlar la convergencia de los iterados por medio de la utilización de constantes especiales llamadas *parámetros de relajación.*

El método de Jacobi puede controlarse por medio de:

$$x_i^{(k+1)} = \frac{\omega}{A_{ii}} \left[b_i - \sum_{\substack{j=0 \\ j \neq i}}^{n-1} A_{ij} x_j^{(k)} \right] + (1-\omega)\, x_i^{(k)}, \tag{4.22}$$

donde $i = 0, 1, \ldots, n-1$, y ω es el parámetro de relajación introducido dentro del proceso iterativo para controlar el orden de convergencia. El algoritmo (4.22) se denomina *Método de Sobrerrelajación de Jacobi* o JOR. En este algoritmo, la matriz de iteración $\mathbf{B}$ es:

$$\mathbf{B} = \mathbf{B}_J(\omega) = \omega \mathbf{B}_J + (1-\omega)\mathbf{I}, \tag{4.23}$$

y (4.22) puede ser expresado en la forma de (4.13) como:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + (\omega \mathbf{D})^{-1}\, \mathbf{r}^{(k)}. \tag{4.24}$$

El método JOR satisface (4.3) puesto que $\omega \neq 0$. También puede reducirse fácilmente al método de Jacobi cuando $\omega = 1$.

Así como existe un método de sobrerrelajación para la iteración de Jacobi, la misma idea se aplica al caso del método de Gauss-Seidel. El *método de sobrerrelajaciones sucesivas de Gauss-Seidel* se define como:

$$x_i^{(k+1)} = \frac{\omega}{A_{ii}} \left[b_i - \sum_{j=0}^{i-1} A_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n-1} A_{ij} x_j^{(k)} \right] + (1-\omega)\, x_i^{(k)}, \tag{4.25}$$

donde $i = 0, 1, \ldots, n-1$. En forma matricial se expresa como:

$$\mathbf{D}\mathbf{x}^{(k+1)} = \omega \left[\mathbf{b} + \mathbf{L}\mathbf{x}^{(k+1)} + \mathbf{U}\mathbf{x}^{(k)} \right] + (1-\omega)\, \mathbf{D}\mathbf{x}^{(k)},$$

o

$$\left[\mathbf{I} - \omega \mathbf{D}^{-1}\mathbf{L} \right] \mathbf{x}^{(k+1)} = \omega \mathbf{D}^{-1}\mathbf{b} + \left[(1-\omega)\, \mathbf{I} + \omega \mathbf{D}^{-1}\mathbf{U} \right] \mathbf{x}^{(k)}, \tag{4.26}$$

donde la matriz de iteración es ahora:

$$\mathbf{B} = \mathbf{B}_{GS}\,(\omega) = \left[\mathbf{I} - \omega \mathbf{D}^{-1}\mathbf{L} \right]^{-1} \left[(1-\omega)\, \mathbf{I} + \omega \mathbf{D}^{-1}\mathbf{U} \right]. \tag{4.27}$$

Se observa que a partir de (4.26), multiplicando ambos lados de la igualdad por $\mathbf{D}$ queda:

$$\left[\mathbf{D} - \omega \mathbf{L} \right] \mathbf{x}^{(k+1)} = \omega \mathbf{b} + \left[(1-\omega)\, \mathbf{D} + \omega \mathbf{U} \right] \mathbf{x}^{(k)},$$

pero sabiendo que $\mathbf{A} = \mathbf{D} - (\mathbf{L} + \mathbf{U})$ entonces:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \left[\frac{1}{\omega}\mathbf{D} - \mathbf{L} \right]^{-1} \mathbf{r}^{(k)}. \tag{4.28}$$

La condición (4.3) se mantiene si $\omega \neq 0$. En caso de $\omega = 1$ se tiene el método de Gauss-Seidel. Si $\omega \in (0, 1)$, esta técnica es conocida como método de *subrrelajación* y sirve para obtener la convergencia de algunos sistemas que no son convergentes con el

método de Gauss-Seidel. Mientras que si $\omega \in (1, \infty)$ a la ecuación (4.28) se la conoce como método de *sobrerrelajación* o SOR y se utiliza para acelerar la convergencia de sistemas que son convergentes con el método de Gauss-Seidel.

Como no hay condiciones concretas definidas sobre cómo identificar el valor de ω ideal, se enuncian dos teoremas que serán útiles en algunos casos.

Teorema 13 (Kahan)**.** *Si $A_{ii} \neq 0$ para $i = 1, 2, \ldots, n$, entonces el método de relajación puede converger sólo si $0 < \omega < 2$.*

Teorema 14 (Ostrowski-Reich)**.** *Si $\mathbf{A}$ es una matriz definida positiva y si $0 < \omega < 2$, entonces el método de relajación converge para cualquier elección del vector inicial $\mathbf{x}^{(0)}$.*

4.3. Condiciones de convergencia y terminación

Las condiciones establecidas para la convergencia de (4.2) fueron dadas en función de la matriz $\mathbf{B}$. Si se requiere conocer qué sistema será convergente con alguno de los métodos iterativos, sólo es necesario analizar la forma de la matriz de coeficientes, independientemente de los valores de $\mathbf{b}$.

Definición 15 (Matriz Diagonal Dominante)**.** *La matriz $\mathbf{A} \in \mathbb{R}^{n \times n}$ es diagonal dominante si:*

$$|A_{ii}| > \sum_{\substack{j=0 \\ j \neq i}}^{n-1} |A_{ij}|,$$

para $i = 0, 1, \ldots, n-1$.

Se puede demostrar, por medio del Teorema de Punto Fijo, que si la matriz de coeficientes $\mathbf{A}$ es diagonal dominante, entonces los métodos de Jacobi y Gauss-Seidel convergen.

Para terminar las secuencias iterativas desarrolladas anteriormente, se utiliza el error relativo. En este caso, no se comparan las normas de $\mathbf{x}^{(k-1)}$ y $\mathbf{x}^{(k)}$, sino que se intenta estabilizar la convergencia de $\mathbf{A}\mathbf{x}^{(k)} = \mathbf{b}^{(k)}$ con respecto a $\mathbf{b}$. Es decir que la condición de terminación[2] es:

$$\frac{\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\|}{\|\mathbf{b}\|} < \varepsilon, \tag{4.29}$$

para un cierto ε pequeño definido de acuerdo al problema.

4.4. Refinamiento iterativo

También conocido como el *método de iteración directa*, es uno de los más básicos dentro de los métodos iterativos aunque es el de mayor potencia de cálculo.

Asumiendo que $\mathbf{x}^{(0)}$ es el vector solución aproximado del sistema $\mathbf{A}\mathbf{x} = \mathbf{b}$, obtenido por algún método directo o por un método iterativo con convergencia lenta y sin llegar a la condición de convergencia, se puede ver que:

$$\mathbf{A}\mathbf{x}^{(0)} = \mathbf{b}^{(0)}.$$

Ahora restando a la segunda expresión, la primera (sistema inicial):

$$\mathbf{A}\mathbf{x}^{(0)} - \mathbf{A}\mathbf{x} = \mathbf{b}^{(0)} - \mathbf{b}$$
$$\mathbf{A}\left(\mathbf{x}^{(0)} - \mathbf{x}\right) = \mathbf{r}^{(0)},$$

[2] no es la única, pero sí una de las más utilizadas

donde se utilizan el *vector de error*, ya definido en (4.4), y el *vector residual*: $\mathbf{r}^{(0)} = \mathbf{b}^{(0)} - \mathbf{b}$. Ahora se puede resolver el sistema:

$$\mathbf{A}\mathbf{e}^{(0)} = \mathbf{r}^{(0)} \tag{4.30}$$

para calcular el error del vector solución inicial. Se obtiene una nueva solución del sistema o una solución refinada con respecto al vector solución inicial. La expresión (4.30) se puede suponer como una secuencia iterativa, entonces:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{e}^{(k)}, \tag{4.31}$$

donde $\mathbf{e}^{(k)}$ es la solución del sistema $\mathbf{A}\mathbf{e}^{(k)} = \mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$.

La finalización de este método ocurre cuando se cumple alguna de las siguientes condiciones:

- La norma del vector residual en la iteración k es mayor que la norma del vector residual en la iteración $k-1$. Es decir que no se puede refinar más, ya sea por la aritmética utilizada o por el sistema en sí.
- Se verifica que el cociente entre la norma del vector residual en la iteración k y la norma del vector de constantes es menor a un cierto ε pequeño definido de antemano. Esta es la misma condición que (4.29).
- Se alcanzó la cantidad máxima de iteraciones, esta condición se aplica siempre sobre el desarrollo en computadora.

Ejemplo 32. *Se desea mejorar la solución aproximada* $\mathbf{x}^{(0)}$*, del sistema de ecuaciones lineales* $\mathbf{A}\mathbf{x} = \mathbf{b}$*, donde:*

$$\mathbf{x}^{(0)} = \begin{bmatrix} 0{,}6000 \\ -0{,}3000 \\ -0{,}2000 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 1{,}921 & 1{,}605 & 1{,}347 \\ 2{,}977 & 2{,}095 & 2{,}404 \\ 2{,}029 & 1{,}608 & 2{,}662 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0{,}3922 \\ 0{,}6554 \\ 0{,}1711 \end{bmatrix}.$$

Para eso se aplicará el algoritmo de refinamiento iterativo con la condición de stop de $\varepsilon = 1 \times 10^{-3}$*. Con el fin de resolver los sucesivos sistemas de ecuaciones que surgen a través de la aplicación del algoritmo, se descompondrá* $\mathbf{A}$ *según Doolittle:*

$$\begin{bmatrix} 1{,}921 & 1{,}605 & 1{,}347 \\ 2{,}977 & 2{,}095 & 2{,}404 \\ 2{,}029 & 1{,}608 & 2{,}662 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1{,}550 & 1 & 0 \\ 1{,}056 & 0{,}2224 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1{,}921 & 1{,}605 & 1{,}347 \\ 0 & -0{,}3923 & 0{,}3165 \\ 0 & 0 & 1{,}169 \end{bmatrix}.$$

Entonces:

$$\mathbf{A}\mathbf{x}_0 = \mathbf{b}_0 = \begin{bmatrix} 0{,}4017 \\ 0{,}6769 \\ 0{,}2026 \end{bmatrix}, \quad \mathbf{b} - \mathbf{b}_0 = \mathbf{r}_0 = \begin{bmatrix} -0{,}0095 \\ -0{,}02149 \\ -0{,}03149 \end{bmatrix},$$

$$\|\mathbf{r}_0\|_\infty = 0{,}03149, \quad \frac{\|\mathbf{r}_0\|_\infty}{\|\mathbf{b}\|} = \frac{0{,}03149}{0{,}6554} = 0{,}0480,$$

$$\mathbf{L}\mathbf{y}_0 = \mathbf{r}_0 \Rightarrow \mathbf{y}_0 = \begin{bmatrix} -0{,}0095 \\ -0{,}006765 \\ -0{,}01995 \end{bmatrix}, \quad \mathbf{U}\mathbf{e}_0 = \mathbf{y}_0 \Rightarrow \mathbf{e}_0 = \begin{bmatrix} 0{,}004116 \\ 0{,}003476 \\ -0{,}01706 \end{bmatrix},$$

$$\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{e}_0 \Rightarrow \mathbf{x}_1 = \begin{bmatrix} 0{,}6041 \\ -0{,}2965 \\ -0{,}2170 \end{bmatrix}.$$

Realizando la segunda iteración:

$$\mathbf{A}\mathbf{x}_1 = \mathbf{b}_1 \Rightarrow \mathbf{b}_1 = \begin{bmatrix} 0{,}3922 \\ 0{,}6555 \\ 0{,}1712 \end{bmatrix}, \quad \mathbf{b} - \mathbf{b}_1 = \mathbf{r}_1 = \begin{bmatrix} 0 \\ 0{,}0001 \\ 0{,}0001 \end{bmatrix},$$

$$\|\mathbf{r}_1\|_\infty = 0{,}0001, \quad \frac{\|\mathbf{r}_1\|_\infty}{\|\mathbf{b}\|} = \frac{0{,}0001}{0{,}6554} = 0{,}0001525.$$

Debido a que se cumple la condición de stop, la solución refinada del sistema de ecuaciones es $\mathbf{x}_1$.

4.5. Ejercicios

1. Construir los siguientes algoritmos en PC:

   a) **Método de Jacobi**. Entrada: una matriz cuadrada de orden n; un vector columna de longitud n; un vector inicial x_0, la cantidad máxima de iteraciones; la tolerancia para terminar el algoritmo. Salida: el vector solución del sistema o un cartel que indique la no convergencia del algoritmo.

   b) **Método de Gauss-Seidel**. Utilizar las mismas entradas y salidas dadas en el algoritmo de Jacobi.

   c) **Refinamiento iterativo**. Utilizar las mismas entradas y salidas dadas en el algoritmo de Jacobi.

2. Resolver los siguientes sistemas de ecuaciones utilizando el método de Jacobi y Gauss-Seidel. Iterar hasta obtener un error relativo menor a 0,001:

   a)
   $$\begin{aligned} 3x_1 + x_2 &= 7 \\ 3x_1 + 1,0001x_2 &= 7,0001 \end{aligned}$$

   b)
   $$\begin{aligned} 0,003x_1 + x_2 &= 1,006 \\ 3x_1 + x_2 &= 7 \end{aligned}$$

   c)
   $$\begin{aligned} 5x_1 - 4x_2 &= 1 \\ -9x_1 + 10x_2 &= 1 \end{aligned}$$

   d)
   $$\begin{aligned} 2x_1 - x_2 &= 1 \\ -x_1 + 4x_2 &= 3 \end{aligned}$$

3. [*EMT*] Generar una matriz $\mathbf{A}$ de dimensión 2 que cumpla con $\mathcal{K}(\mathbf{A}) > 10^4$ en alguna norma matricial. Resolver $\mathbf{A}\mathbf{x} = \mathbf{b}$, donde $\mathbf{b} = [-2; 1]^T$ utilizando Gauss con aritmética reducida y mejorar la solución con refinamiento iterativo.

4. Mejorar, si es posible, las soluciones obtenidas en el ejercicio 2 utilizando Refinamiento Iterativo.

5. Resolver el sistema $\mathbf{H}\mathbf{x} = \mathbf{b}$, donde $\mathbf{H}$ es la matriz de Hilbert de orden 3 y $\mathbf{b} = [1; 2; 3]^T$. Utilizar $\mathbf{x_0} = [30; -190; 200]$.

6. [*EMT*] Generar una matriz $\mathbf{A}$ tal que $5 < \mathcal{K}(\mathbf{A}) < 10$. Resolver el sistema $\mathbf{Ax} = \mathbf{b}$, donde $\mathbf{b} = [1; 1; 1]^T$, utilizando relajación. Probar con $\omega = 0{,}4;\ 0{,}8;\ 1{,}2$ y $1{,}6$.

7. Generar dos planillas de cálculo que permita resolver sistemas de ecuaciones lineales de dimensión 3. Una utilizando el método de Jacobi y la otra Gauss-Seidel.

8. Demostrar que $\rho(\mathbf{A}) < 1$ si y sólo si $\lim_{k\to\infty} \mathbf{A}^k\mathbf{x} = \mathbf{0}$ para todo $\mathbf{x}$.

9. ¿Cuáles de los axiomas necesarios para definir normas se cumplen por la función de radio espectral ρ y cuáles no?

10. Utilizando una aritmética de punto flotante de 4 dígitos y truncamiento, obtener una aproximación a la solución con tres dígitos correctos del sistema:

$$\begin{aligned} 0{,}8647x_1 + 0{,}5766x_2 &= 0{,}2885 \\ 0{,}4322x_1 + 0{,}2882x_2 &= 0{,}1442 \end{aligned}$$

    Calcular el residuo una vez finalizado el algoritmo iterativo.

11. Resolver el sistema:

$$\begin{aligned} 8x + 3y + 2z &= 20{,}00 \\ 16x + 6y + 4{,}001z &= 40{,}02 \\ 4x + 1{,}501y + z &= 10{,}01 \end{aligned}$$

    utilizando los método de Gauss y Gauss-Seidel (semilla: $x_0 = y_0 = z_0 = 1$). ¿Cuál es más eficiente? ¿Por qué?

12. Sea $\mathbf{B}$ una matriz diagonal de orden n. Demostrar que para cualquier matriz $\mathbf{A}$ de orden n se cumple que:

$$\text{diag}(\mathbf{BA}) = \mathbf{B}\text{diag}(\mathbf{A})$$

13. Sea $\mathbf{B}$ una matriz diagonal de orden n. Demostrar que las iteraciones de Jacobi para resolver $\mathbf{Ax} = \mathbf{b}$ y $\mathbf{BAx} = \mathbf{Bb}$ generan la misma secuencia de iterados $\mathbf{x}^n$.

14. [*EMT*] Establecer una condición suficiente sobre el parámetro β de forma tal que los métodos de Jacobi y Gauss-Seidel converjan cuando se aplican para obtener una solución del sistema cuya matriz de coeficientes es:

$$\mathbf{A} = \begin{bmatrix} -10 & 2 \\ \beta & 5 \end{bmatrix}.$$

15. [*EMT*] Para el sistema:

$$\begin{aligned} 0{,}96326x_1 + 0{,}81321x_2 &= 0{,}88824 \\ 0{,}81321x_1 + 0{,}68654x_2 &= 0{,}74988 \end{aligned}$$

    a) Determinar el radio espectral de la matriz de iteración $\mathbf{P}^{-1}$ para el método de Gauss-Seidel.

    b) Utilizando el método de Gauss-Seidel y la semilla $\mathbf{x} = [0{,}33116; 0{,}70000]$ resolver el sistema. ¿Cómo se explica el resultado?

16. Considerar el sistema:

$$\begin{aligned} x_1 - \tfrac{1}{4}x_3 - \tfrac{1}{4}x_4 &= \tfrac{1}{2} \\ x_2 - \tfrac{1}{4}x_3 - \tfrac{1}{4}x_4 &= \tfrac{1}{2} \\ -\tfrac{1}{4}x_1 - \tfrac{1}{4}x_2 + x_3 &= \tfrac{1}{2} \\ -\tfrac{1}{4}x_1 - \tfrac{1}{4}x_2 + x_4 &= \tfrac{1}{2} \end{aligned}$$

a) Utilizando como semilla a $\mathbf{x} = \mathbf{0}$ realizar cuatro iteraciones del método de Jacobi.

b) Con la misma semilla que el ítem anterior, hacer cuatro iteraciones del método de Gauss-Seidel.

c) ¿Cuál es la solución verdadera del sistema?

d) Calcular la matriz $\mathbf{B}$ para los métodos utilizados y encontrar su radio espectral.

17. Mostrar que la iteración matricial $\mathbf{B}^{(i+1)} = \mathbf{B}^{(i)}\left(2\mathbf{I} - \mathbf{A}\mathbf{B}^{(i)}\right)$, utilizada para obtener $\mathbf{A}^{-1}$, donde $\mathbf{B}^{(0)}$ es una aproximación a $\mathbf{A}^{-1}$, es análoga al método de Newton Raphson para encontrar el inverso de un número. ¿Qué condición debe cumplir $\mathbf{I} - \mathbf{A}\mathbf{B}^{(0)}$ para que el esquema planteado sea convergente?

18. Analizar la convergencia de los métodos de Jacobi y Gauss-Seidel para la matriz de segundo orden:

$$\mathbf{A} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

donde $|\rho| < 1$ y el vector semilla es distinto del vector nulo.

19. [*EMT*] Aplicar la técnica de relajación sobre el método de Gauss-Seidel con diferentes valores para el parámetro $\omega = 0{,}2; 0{,}4; \ldots; 1{,}8$ con los dos sistemas dados a continuación. Identificar cuáles de los valores de ω son mejores para cada problema, iterando hasta que $\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| / \left\|\mathbf{x}^{(k)}\right\| < 10^{-6}$:

a)

$$\begin{aligned} 5x_1 - 4x_2 &= 1 \\ -9x_1 + 10x_2 &= 1 \end{aligned}$$

b)

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

20. [*EMT*] Una forma de refinar una aproximación a la inversa de una matriz $\mathbf{A}$ se presenta en el siguiente esquema iterativo:

$$\mathbf{C}^{(m+1)} = \mathbf{C}^{(m)}\left(\mathbf{I} + \mathbf{R}^{(m)}\right), \qquad \mathbf{R}^{(m+1)} = \mathbf{I} - \mathbf{A}\mathbf{C}^{(m+1)},$$

donde $\mathbf{R}^{(0)} = \mathbf{I} - \mathbf{A}\mathbf{C}^{(0)}$ y $\mathbf{C}^{(0)}$ es la aproximación a $\mathbf{A}^{-1}$. Implementar este esquema en PC y utilizarlo para refinar las inversas de la matriz de Hilbert que calcula *EMT*.

21. [*EMT*] Para resolver el sistema de ecuaciones lineales en bloque:

$$\begin{bmatrix} \mathbf{A}_1 & \mathbf{B} \\ \mathbf{B} & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix},$$

Quarteroni propone el siguiente esquema iterativo:

$$\mathbf{A}_1\mathbf{x}^{(k+1)} + \mathbf{B}\mathbf{y}^{(k)} = \mathbf{b}_1, \qquad \mathbf{B}\mathbf{x}^{(k)} + \mathbf{A}_2\mathbf{y}^{(k+1)} = \mathbf{b}_2,$$

que es convergente para cualquier $\mathbf{x}^{(0)}$, $\mathbf{y}^{(0)}$ si $\rho(\mathbf{A}_1^{-1}\mathbf{B}) < 1$ y $\rho(\mathbf{A}_2^{-1}\mathbf{B}) < 1$. Implementar este código en PC, generar un sistema y resolverlo. Comparar lo obtenido con los resultados al aplicar los métodos de Jacobi y Gauss-Seidel. Utilizar $\mathbf{x}^{(0)} = \mathbf{y}^{(0)} = [1; 1]^T$.

22. [*EMT*] Sea $\mathbf{A}$ una matriz simétrica de orden n cuyos autovalores son reales y pertenecen al intervalo $[\alpha, \beta]$, $0 < \alpha \leq \beta$. Entonces es posible resolver el sistema lineal $\mathbf{Ax} = \mathbf{b}$ a través de la iteración de Richardson:

$$\mathbf{x}^{(k+1)} = (\mathbf{I} - \omega\mathbf{A})\,\mathbf{x}^{(k)} + \omega\mathbf{b},$$

donde $\mathbf{e}^{(k+1)} = (\mathbf{I} - \omega\mathbf{A})\,\mathbf{e}^{(k)}$.

a) Implementar este código en PC y verificar que la elección óptima del parámetro es $\omega = \dfrac{2}{\alpha + \beta}$.

b) Probar el código anterior con el sistema $\mathbf{Ax} = \mathbf{b}$ donde:

$$\mathbf{A} = \begin{bmatrix} 0{,}864 & 0{,}369 & 0{,}601 & 0{,}618 \\ 0{,}369 & 0{,}831 & 0{,}367 & 0{,}102 \\ 0{,}601 & 0{,}367 & 0{,}641 & 0{,}130 \\ 0{,}618 & 0{,}102 & 0{,}130 & 0{,}850 \end{bmatrix}$$

y $\mathbf{b}$ es un vector definido libremente.

23. Demostrar que, para cualquier matriz $\mathbf{A} \in \mathbb{R}^{n\times n}$ y para cualquier norma consistente, se verifica que:

$$\rho(\mathbf{A}) \leq \|\mathbf{A}\|\,.$$

24. Dado el sistema de ecuaciones:

$$\begin{aligned} x_1 + 2x_2 - 2x_3 &= 7 \\ x_1 + x_2 + x_3 &= 2 \\ 2x_1 + 2x_2 + x_3 &= 5 \end{aligned}$$

a) Calcular el radio espectral de las matrices $\mathbf{B}$ de iteración para los métodos de Jacobi y Gauss-Seidel.

b) Resolver el sistema por ambos métodos.

25. Dar un ejemplo de una matriz $\mathbf{A}$ que no sea diagonal dominante, pero que el sistema $\mathbf{Ax} = \mathbf{b}$ igual sea convergente al utilizar algún método iterativo (no refinamiento).

Bibliografía

- *A theoretical introduction to numerical analysis**, V. RYABEN'KII y S. TSYNKOV, Cap.6
- *An introduction to numerical analysis*, Kendall ATKINSON, Cap.8
- *Análisis numérico*, R. BURDEN y J. FAIRES, Cap.7
- *Análisis numérico con aplicaciones*, C. GERALD y P. WHEATLEY, Cap.2
- *Numerical mathematics**, A. QUARTERONI, R. SACCO y F. SALERI, Cap.4
- *Numerical methods*, G. DAHLQUIST y A. BJÖRK, Cap.5

5 Autovalores

Autovalores y autovectores son herramientas estándar dentro de la computación científica. Los autovalores proporcionan información acerca del comportamiento de sistemas modelados por medio de una matriz u operador. El problema de calcular autovalores y autovectores de una matriz aparece en la mayoría de desarrollos en física e ingeniería. Los autovalores son de mucha utilidad para cálculos de análisis de resonancia, inestabilidad y tasas de crecimiento ó decrecimiento de sistemas, por ejemplo, sistemas con vibraciones, alas de avión, edificios, puentes y hasta moléculas. Además, la descomposición por autovalores juega un importante rol dentro de los métodos numéricos, tales como resolución de SEL por iteraciones ó ecuaciones diferenciales.

5.1. Introducción

Un punto central dentro del estudio de matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ son los vectores especiales cuyas direcciones no cambian cuando son multiplicados por una matriz $\mathbf{A}$, sino que sólo sufren estiramientos ó contracciones.

Definición 16. *Si* $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$, $\lambda \in \mathbb{R}$ *y además:*

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \tag{5.1}$$

entonces el escalar real λ *se denomina* ***autovalor*** *(*eigenvalue *ó valor propio) de* $\mathbf{A}$ *y* $\mathbf{x}$ *es el* ***autovector*** *(*eigenvector *ó vector propio) de* $\mathbf{A}$ *asociado a* λ.

Cuando un autovalor λ es conocido, el autovector asociado se obtiene resolviendo el sistema lineal homogéneo:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0. \tag{5.2}$$

Por lo tanto, λ es un autovalor de $\mathbf{A}$ sólo si $\mathbf{A} - \lambda\mathbf{I}$ es una matriz singular. Claramente, un autovector $\mathbf{x}$ se puede determinar por resolución del sistema asociado, salvo una constante multiplicativa $\alpha \neq 0$. Una solución no trivial de (5.2) existe sólo si el determinante de la matriz de coeficientes es cero, por lo tanto:

$$|\mathbf{A} - \lambda\mathbf{I}| = 0. \tag{5.3}$$

La expansión de (5.3) lleva a una ecuación polinomial denominada **ecuación característica** ó **polinomio característico**:

$$p(\lambda) = a_1\lambda^n + a_2\lambda^{n-1} + \ldots + a_n\lambda + a_{n+1}$$

donde las raíces $\lambda_i, i = 1, 2, \ldots, n$, son los autovalores de la matriz $\mathbf{A}$.

La ecuación característica puede ser reescrita como:

$$p(\lambda) = a_1(\lambda_1 - \lambda)(\lambda_2 - \lambda)\cdots(\lambda_n - \lambda).$$

Utilizando la relación entre raíces y coeficientes de una ecuación algebraica se obtiene:

$$p(0) = \lambda_1\lambda_2\cdots\lambda_n = |\mathbf{A}|, \tag{5.4}$$

además, también se verifica que:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = \sum_{i=1}^{n} A_{ii}. \tag{5.5}$$

Ambas relaciones son útiles para comprobar la precisión del espectro de autovalores.

Ejemplo 33. *La matriz:*

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix}$$

determina una ecuación característica:

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} 1-\lambda & 1 & 0 \\ 2 & 1-\lambda & 0 \\ -3 & 0 & 1-\lambda \end{vmatrix} = -\lambda^3 + 3\lambda^2 - \lambda - 1$$

cuyas soluciones son $\lambda_1 = 1$, $\lambda_2 = -0{,}4142$ *y* $\lambda_3 = 2{,}414$. *Reemplazando* λ_3 *en (5.2) y resolviendo el sistema:*

$$\begin{bmatrix} -1{,}414 & 1{,}000 & 0{,}000 \\ 2{,}000 & -1{,}414 & 0{,}000 \\ -3{,}000 & 0{,}000 & -1{,}414 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0{,}000 \\ 0{,}000 \\ 0{,}000 \end{bmatrix}$$

se obtiene uno de los autovectores asociados a λ_3.

Algunas propiedades importantes de autovalores y autovectores, dadas sin demostración:

- Todos los autovalores de una matriz simétrica son reales.
- Todos los autovalores de una matriz simétrica y definida positiva son reales y positivos.
- Los autovectores de una matriz simétrica son ortonomales.
- Si los autovalores de $\mathbf{A}$ son λ_i, entonces los autovalores de $\mathbf{A}^{-1}$ son λ_i^{-1}.

Generalmente, las matrices que se originan a partir de problemas físicos son simétricas. Esto es un beneficio, pues es mucho más simple calcular los autovalores de matrices simétricas que de aquellas que no lo son. Los problemas de análisis de vibraciones y estabilidad se resuelven a través del cálculo de autovalores.

Comandos de EMT. *Los comandos útiles para calcular autovalores y autovectores son:*

- `charpoly(A:matriz numérica)`. *Calcula el polinomio característico, donde* ***A*** *es una matriz cuadrada. La salida es un vector con los coeficientes del polinomio, ordenados de la menor potencia a la mayor.*

- `{l,x}=eigen(A:matriz numérica, usekernel:booleano, usecharpoly:boolean, check:booleano)`. *Calcula autovalores y autovectores, donde* $\boldsymbol{A}$ *es una matriz cuadrada;* ***usekernel*** *desactiva el uso de AlgLib y LAPACK, el valor por defecto es 0;* ***usecharpoly*** *permite calcular los autovalores como raíces del polinomio característico, el valor por defecto es 0;* ***check****, habilita el testeo luego de calcular por AlgLib, el valor por defecto es 1. Las salidas son* $\boldsymbol{l}$*, un vector con los autovalores y* $\boldsymbol{x}$ *una base de autovectores, organizados en columnas.*

Ejemplo en EMT 13. *Dada la matriz:*

$$\begin{bmatrix} 1 & 2 & 3 \\ -4 & 5 & -6 \\ 7 & 8 & 9 \end{bmatrix},$$

calcular el polinomio característico, los autovalores y una base de los autovectores.

```
>charpoly([1,2,3;-4,5,-6;7,8,9])
 [120,  94,  -15,  1]
>{l,x}=eigen([1,2,3;-4,5,-6;7,8,9]); l, x
 [ -1.07788+0i   8.03894+6.83414i  8.03894-6.83414i ]
               -1+0i 0.190618+0.0920379i 0.190618-0.0920379i
         0.015455+0i -0.568435+0.400091i -0.568435-0.400091i
         0.682322+0i    0.61654+0.38346i    0.61654-0.38346i
```

5.2. Teoremas de Gerschgorin

Los teoremas de Gerschgorin proporcionan una manera simple de determinar regiones en las cuales se ubican los autovalores de una matriz. Es general y no asume condiciones especiales para las matrices tales como la simetría ó la tridiagonalidad, además opera desde y hacia valores complejos.

Definición 17. *Sea* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *y* $n \geq 2$*. Los* ***discos de Gerschgorin*** $D_i, i = 1, 2, \ldots, n$*, de la matriz* $\mathbf{A}$ *se definen como las regiones circulares cerradas:*

$$D_i = \{z \in \mathbb{C} : |z - A_{ii}| \leq R_i\} \tag{5.6}$$

en el plano complejo, donde:

$$R_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |A_{ij}| \tag{5.7}$$

es el radio de D_i*.*

Teorema 15 (Primer Teorema de Gerschgorin)**.** *Sea* $n \geq 2$ *y* $\mathbf{A} \in \mathbb{C}^{n\times n}$*. Todos los autovalores de la matriz* $\mathbf{A}$ *se ubican en la región* $D = \bigcup_{i=1}^{n} D_i$*, donde* $D_i, i = 1, 2, \ldots, n$*, son los discos de Gerschgorin.*

Demostración. Suponiendo que $\lambda \in \mathbb{C}$ y $\mathbf{x} \in \mathbb{C}^n_*$ son un autovalor de $\mathbf{A}$ y su autovector asociado, entonces:

$$\sum_{j=1}^{n} A_{ij}x_j = \lambda x_i, \quad i = 1, 2, \ldots, n.$$

Sea x_k, con $k \in \{1, 2, \ldots, n\}$, la componente de $\mathbf{x}$ que tiene mayor módulo ó una de sus componentes si más de una tienen el mismo módulo. De las hipótesis se sigue que $x_k \neq 0$, puesto que $\mathbf{x} \neq 0$. Por lo tanto:

$$|x_j| \leq |x_k|, \quad j = 1, 2, \ldots, n.$$

Esto significa que:

$$\begin{aligned}|\lambda - A_{kk}||x_k| &= |\lambda x_k - A_{kk}x_k| \\ &= \left|\sum_{j=1}^{n} A_{kj}x_j - A_{kk}x_k\right| \\ &= \left|\sum_{\substack{j=1 \\ j\neq k}}^{n} A_{kj}x_j\right| \\ &\leq |x_k|R_k,\end{aligned}$$

lo cual, dividiendo ambas expresiones por $|x_k|$, muestra que λ se encuentra en el disco de Gerschgorin D_k de radio R_k centrado en A_{kk}. De aquí, $\lambda \in D = \bigcup_{i=1}^{n} D_i$. □

Teorema 16 (Segundo Teorema de Gerschgorin)**.** *Si la unión $\mathcal{M}$ de k discos de Gerschgorin D_i es disjunta de los discos restantes, entonces $\mathcal{M}$ contiene exactamente k autovalores de $\mathbf{A}$.*

Demostración. Considerar para $t \in [0,1]$ la familia de matrices:

$$\mathbf{A}(t) = t\mathbf{A} + (1-t)\text{diag}(A_{ii}).$$

Los coeficientes en el polinomio característico son funciones continuas de t, y por lo tanto los autovalores $\lambda(t)$ de $\mathbf{A}(t)$ son también funciones continuas de t. Como es cierto que $\mathbf{A}(0) = \text{diag}(A_{ii})$ y $\mathbf{A}(1) = \mathbf{A}$ entonces $\lambda_i(0) = A_{ii}$ y $\lambda_i(1) = \lambda_i$. Para $t = 0$, hay exactamente k autovalores en $\mathcal{M}$. Por razones de continuidad un autovalor $\lambda_i(t)$ no puede saltar a un subconjunto que no tiene una conexión continua con A_{ii} para $t = 1$. De ahí que k autovalores de $\mathbf{A} = \mathbf{A}(1)$ se ubican en $\mathcal{M}$. □

Ejemplo 34. *Sea:*

$$\mathbf{A} = \begin{bmatrix} 3{,}432 & 0{,}1672 & 0{,}1981 \\ 0{,}01552 & 1{,}106 & 0{,}4897 \\ 0{,}9841 & 0{,}3724 & 0{,}9395 \end{bmatrix}.$$

Se generan tres discos de Gerschgorin:

$$\begin{aligned}D_1 &= \{|z - 3{,}432| \leq 0{,}3653\}, \\ D_2 &= \{|z - 1{,}106| \leq 0{,}5052\}, \\ D_3 &= \{|z - 0{,}9395| \leq 1{,}356\}.\end{aligned}$$

Esos discos se dividen en dos áreas:

$$\begin{aligned}\mathcal{M}_1 &= D_1, \\ \mathcal{M}_2 &= D_2 \cup D_3 = D_3.\end{aligned}$$

Por lo tanto, en $\mathcal{M}_2$ deben existir 2 autovalores y el restante esta en $\mathcal{M}_1$. Los autovalores de $\mathbf{A}$ son $\lambda_1 = 3{,}524$, $\lambda_2 = 0{,}5789$ y $\lambda_3 = 1{,}374$. Entonces $\lambda_1 \in \mathcal{M}_1$ y λ_2, $\lambda_3 \in \mathcal{M}_2$.

5.3. Métodos de la potencia

Los métodos de la potencia son útiles para aproximar los autovalores extremos de una matriz, es decir, los autovalores que tienen mayor y menor módulo, denotados

como λ_1 y λ_n respectivamente. También permiten obtener los autovectores asociados a dichos autovalores. La resolución de estos problemas son de interés para aplicaciones de la vida real (geosísmica, vibraciones de máquinas y estructuras, análisis de redes eléctricas, entre otras) donde el cómputo de λ_n determina la frecuencia adecuada de un sistema físico.

5.3.1. Aproximación del autovalor dominante

Sea $\mathbf{A} \in \mathbb{R}^{n\times n}$ una matriz diagonalizable y sea $\mathbf{X} \in \mathbb{R}^{n\times n}$ la matriz de sus autovectores $\mathbf{x}_i, i = 1, 2, \ldots, n$, suponiendo que los autovalores de $\mathbf{A}$ pueden ser ordenados en forma descendente de acuerdo a su módulo:

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|\,,$$

donde λ_1 tiene multiplicidad algebraica igual a 1. Bajo estas condiciones, λ_1 es denominado el **autovalor dominante** de la matriz $\mathbf{A}$. Dado un vector inicial arbitrario $\mathbf{q}^{(0)} \in \mathbb{R}^n$, se realizan las iteraciones $k = 1, 2, \ldots$ basándose en potencia de matrices. Dicho método se conoce como el **método de la potencia**:

$$\begin{aligned} \mathbf{z}^{(k)} &= \mathbf{A}\mathbf{q}^{(k-1)} \\ \mathbf{q}^{(k)} &= \frac{\mathbf{z}^{(k)}}{\left\|\mathbf{z}^{(k)}\right\|_2} \\ \mathcal{L}_1^{(k)} &= (\mathbf{q}^{(k)})^T \mathbf{A}\mathbf{q}^{(k)}. \end{aligned} \tag{5.8}$$

La convergencia del método se demuestra por inducción sobre k. Sea $\mathbf{q}^{(0)}$ arbitrario:

$$\begin{aligned} \mathbf{z}^{(k)} &= \mathbf{A}\mathbf{q}^{(k-1)}, \\ \mathbf{q}^{(k)} &= \frac{\mathbf{z}^{(k)}}{\left\|\mathbf{z}^{(k)}\right\|_2}, \end{aligned}$$

entonces:

$$\begin{aligned} \mathbf{z}^{(k)} = \mathbf{A}\mathbf{q}^{(k-1)} &= \mathbf{A}\frac{\mathbf{z}^{(k-1)}}{\left\|\mathbf{z}^{(k-1)}\right\|_2} \\ &= \mathbf{A}^2\mathbf{q}^{(k-2)} = \mathbf{A}^2\frac{\mathbf{z}^{(k-2)}}{\left\|\mathbf{z}^{(k-2)}\right\|_2} \\ &= \mathbf{A}^3\mathbf{q}^{(k-3)} = \mathbf{A}^3\frac{\mathbf{z}^{(k-3)}}{\left\|\mathbf{z}^{(k-3)}\right\|_2} \\ &= \cdots \\ &= \mathbf{A}^k\mathbf{q}^{(0)}, \\ \mathbf{q}^{(k)} &= \frac{\mathbf{A}^k\mathbf{q}^{(0)}}{\left\|\mathbf{A}^k\mathbf{q}^{(0)}\right\|_2}. \end{aligned} \tag{5.9}$$

Esta relación explica el papel jugado por las potencias de $\mathbf{A}$ en el método. Como $\mathbf{A}$ es diagonalizable, sus autovectores $\mathbf{x}_i$ forman una base de $\mathbb{R}^{n\times n}$; entonces es posible representar $\mathbf{q}^{(0)}$ como:

$$\mathbf{q}^{(0)} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i, \qquad \alpha_i \in \mathbb{R}, \qquad i = 1, 2, \ldots, n. \tag{5.10}$$

Más aún, como $\mathbf{A}^n\mathbf{x}_i = \lambda_i^n\mathbf{x}_i$, entonces:

$$\begin{aligned}\mathbf{A}^k\mathbf{q}^{(0)} &= \mathbf{A}^k\sum_{i=1}^{n}\alpha_i\mathbf{x}_i \\ &= \mathbf{A}^k\left(\alpha_1\mathbf{x}_1 + \sum_{i=2}^{n}\alpha_i\mathbf{x}_i\right) \\ &= \alpha_1\lambda_1^k\mathbf{x}_1 + \sum_{i=2}^{n}\alpha_i\lambda_i^k\mathbf{x}_i \\ &= \alpha_1\lambda_1^k\left(\mathbf{x}_1 + \sum_{i=2}^{n}\frac{\alpha_i}{\alpha_1}\left(\frac{\lambda_i}{\lambda_1}\right)^k\mathbf{x}_i\right),\end{aligned} \tag{5.11}$$

Como $|\lambda_i/\lambda_1| < 1$ para $i = 2, 3, \ldots, n$, a medida que k crece el vector $\mathbf{A}^k\mathbf{q}^{(0)}$ tiende a converger su componente principal hacia la dirección del autovector $\mathbf{x}_1$, mientras que sus componentes en otras direcciones $\mathbf{x}_j$ decrecen. Utilizando (5.9) y (5.11):

$$\mathbf{q}^{(k)} = \frac{\alpha_1\lambda_1^k\left(\mathbf{x}_1 + \mathbf{y}^{(k)}\right)}{\left\|\alpha_1\lambda_1^k\left(\mathbf{x}_1 + \mathbf{y}^{(k)}\right)\right\|_2} = \mu_k\frac{\mathbf{x}_1 + \mathbf{y}^{(k)}}{\left\|\mathbf{x}_1 + \mathbf{y}^{(k)}\right\|_2},$$

donde μ_k es el signo de $\alpha_1\lambda_1^k$ y $\mathbf{y}^{(k)}$ denota un vector que tiende a cero cuando $k \to \infty$. Es decir que el vector $\mathbf{q}^{(k)}$ se alinea a través de la dirección del autovector $\mathbf{x}_1$.

Como la sucesión $\{\mathcal{L}_1^{(k)}\}$ converge en λ_1 se puede usar esta condición para terminar el algoritmo iterativo (5.9):

$$\left\|\mathbf{A}\mathbf{q}^{(k)} - \mathcal{L}_1^{(k)}\mathbf{q}^{(k)}\right\|_2 < \epsilon,$$

cumplida la condición, el autovalor dominante es $\mathcal{L}_1^{(k)}$ y uno de sus autovectores asociados es $\mathbf{q}^{(k)}$.

Ejemplo 35. *Sea la matriz:*

$$\mathbf{A} = \begin{bmatrix} 0.81472368 & 0.91337585 & 0.27849821 \\ 0.90579193 & 0.63235924 & 0.54688151 \\ 0.12698681 & 0.09754040 & 0.95750683 \end{bmatrix},$$

utilizando el método de la potencia se aproximará su autovalor dominante. Eligiendo $\mathbf{q}^{(0)} = [1, 2, 3]^T$*, la tabla 5.1 muestra las iteraciones realizadas hasta lograr la condición de stop (cantidad máxima de pasos: 50,* $\epsilon = 0{,}001$*).*

5.3.2. Aproximación del autovalor mínimo

La aplicación del método de la potencia, (5.8), a la matriz $(\mathbf{M}_\mu)^{-1} = (\mathbf{A} - \mu\mathbf{I})^{-1}$, se denomina **iteración inversa** ó **método de la potencia inversa**. Al número μ se lo llama *desplazamiento*.

Los autovalores de $\mathbf{M}_\mu^{-1}$ son $\xi_i = (\lambda_i - \mu)^{-1}$, asumiendo que existe un entero m tal que:

$$|\lambda_m - \mu| < |\lambda_i - \mu|, \quad \forall i = 1, 2, \ldots, n \quad i \neq m. \tag{5.12}$$

Esto equivale a exigir que el autovalor λ_m, que es cercano a μ, tenga multiplicidad igual a 1. Más aún, (5.12) muestra que ξ_m es el autovalor de $\mathbf{M}_\mu^{-1}$ con el mayor módulo. En particular, si $\mu = 0$, λ_m resulta ser el autovalor de $\mathbf{A}$ con el menor módulo.

k	$\mathbf{z}^{(k)}$			$\mathcal{L}_1^{(k)}$	ϵ
1	0,57300884	0,62808292	0,52647196	1,77248545	3,53E-01
2	0,65685084	0,66624455	0,35307955	1,79331812	1,61E-01
3	0,68981371	0,67168168	0,27018654	1,77827953	7,28E-02
4	0,70276070	0,67274556	0,23138888	1,76614809	3,39E-02
5	0,70845526	0,67281780	0,21309001	1,75938575	1,60E-02
6	0,71105753	0,67276994	0,20439618	1,75594489	7,59E-03
7	0,71227719	0,67272721	0,20024810	1,75425221	3,62E-03
8	0,71285528	0,67270239	0,19826457	1,75343125	1,73E-03
9	0,71313085	0,67268950	0,19731506	**1,75303561**	8,30E-04

Tabla 5.1: Salida del algoritmo del método de la potencia del ejemplo 35.

Dado un vector arbitrario $\mathbf{q}^{(0)} \in \mathbb{R}^n$, para $k = 1, 2, \ldots$ se construye la siguiente secuencia:

$$\begin{aligned} (\mathbf{A} - \mu\mathbf{I})\,\mathbf{z}^{(k)} &= \mathbf{q}^{(k-1)} \\ \mathbf{q}^{(k)} &= \frac{\mathbf{z}^{(k)}}{\|\mathbf{z}^{(k)}\|_2} \\ \sigma^{(k)} &= \left(\mathbf{q}^{(k)}\right)^T \mathbf{A}\mathbf{q}^{(k)}. \end{aligned} \tag{5.13}$$

Debe notarse que los autovectores de $\mathbf{M}_\mu$ son los mismos que los de $\mathbf{A}$. La principal diferencia con el algoritmo (5.8) es que en cada paso k debe ser resuelto un sistema lineal cuya matriz de coeficientes es $\mathbf{M}_\mu = \mathbf{A} - \mu\mathbf{I}$. Por conveniencia numérica, la factorización LU de $\mathbf{M}_\mu$ es recomendada. De esta forma, sólo se deben resolver dos sistemas triangulares en cada iteración.

A pesar de que es mucho más pesado computacionalmente que el método de la potencia, el método de las iteraciones inversas tiene una gran ventaja: converge a cualquier autovalor de $\mathbf{A}$ (el más cercano al desplazamiento μ). Este método es ideal para refinar cualquier estimación inicial de un autovalor, por ejemplo, obtenida con el teorema 15.

Para asegurar la convergencia de la iteración (5.13) se asume que la matriz $\mathbf{A}$ es diagonalizable, así que $\mathbf{q}^{(0)}$ puede ser representada en la forma (5.10). Procediendo de la misma forma que en el método de las potencias, se tiene:

$$\tilde{\mathbf{q}}^{(k)} = \mathbf{x}_m + \sum_{\substack{i=1 \\ i\neq m}}^{n} \frac{\alpha_i}{\alpha_m} \left(\frac{\xi_i}{\xi_m}\right)^k \mathbf{x}_i,$$

donde $\mathbf{x}_i$ son los autovectores de $\mathbf{M}_\mu^{-1}$ (y también de $\mathbf{A}$), mientras α_i tienen la forma planteada en (5.10). Como consecuencia, volviendo a la definición de ξ_i y usando (5.12) se tiene:

$$\lim_{k\to\infty} \tilde{\mathbf{q}}^{(k)} = \mathbf{x}_m, \qquad \lim_{k\to\infty} \sigma^{(k)} = \lambda_m.$$

La convergencia se vuelve más rápida cuando μ es cercano a λ_m.

Ejemplo 36. *Sea:*

$$\mathbf{A} = \begin{bmatrix} 0{,}6554 & 0{,}2009 & 0{,}8936 \\ 0{,}2818 & 0{,}5250 & 0{,}3141 \\ 0{,}4446 & 0{,}2994 & 0{,}2826 \end{bmatrix}$$

una matriz diagonalizable. Aplicando el teorema 15 se definirán intervalos donde se ubican los autovalores de $\mathbf{A}$. *Entonces* $\lambda_a \in [-0{,}4391; 1{,}749]$, $\lambda_b \in [-0{,}0709; 1{,}120]$ *y* $\lambda_c \in [-0{,}4614; 1{,}026]$. *Como no son tres regiones disjuntas, se utilizará el intervalo*

$[-0{,}4614; 1{,}749]$ *para localizar los tres autovalores. Dividiendo la región antes mencionada en tres subregiones y tomando el punto medio de cada región como desplazamiento queda* $\mu_a = -0{,}093$, $\mu_b = 0{,}6437$ *y* $\mu_c = 1{,}380$. *En la tablas 5.2, 5.3 y 5.4 se muestran las iteraciones que se realizaron para obtener las aproximaciones a los autovalores de* $\mathbf{A}$.

k	$\mathbf{z}^{(k)}$			$\sigma_a^{(k)}$	ϵ
1	2,64942827	1,09403353	-1,34581854	0,14610423	1,32E+00
2	-6,49578040	0,31625849	6,30547717	-0,19056200	3,76E-01
3	7,21280324	0,27736074	-6,90564637	-0,19290645	6,26E-02
4	-7,38568389	-0,16010847	7,02960229	-0,19105556	1,21E-02
5	7,33629572	0,18526657	-6,99638208	-0,19162600	2,57E-03
6	-7,34845875	-0,17969594	7,00452066	**-0,19148729**	5,74E-04

Tabla 5.2: Salida del algoritmo del método de la potencia inversa para μ_a del ejemplo 36.

5.3.3. Problemas de implementación

El análisis de convergencia de las secciones anteriores muestra que la efectividad de los métodos dependen fuertemente de la separación entre el autovalor dominante del resto ($|\lambda_2|/|\lambda_1| \ll 1$). Se analizará el comportamiento de la iteración (5.8) cuando existen dos autovalores de igual módulo. Tres casos pueden ser identificados

- $\lambda_2 = \lambda_1$: los dos autovalores dominantes son coincidentes. El método aún converge, para un k suficientemente grande, la relación (5.11) se transforma en:

$$\mathbf{A}^k \mathbf{q}^{(0)} \simeq \lambda_1^k \left(\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2\right)$$

  lo que es un autovector de $\mathbf{A}$. Para $k \to \infty$, la secuencia $\tilde{\mathbf{q}}^{(k)}$ (luego de una adecuada redefinición) converge a un vector que se encuentra en el subespacio generado por los autovectores $\mathbf{x}_1$ y $\mathbf{x}_2$, mientras que la secuencia $\mathcal{L}^{(k)}$ converge a λ_1.

- $\lambda_2 = -\lambda_1$: los dos autovalores dominantes son opuestos. En este caso el autovalor dominante puede ser aproximado aplicando el método de la potencia a la matriz $\mathbf{A}^2$. En efecto, para $i = 1, 2, \ldots, n$, $\lambda_i\left(\mathbf{A}^2\right) = \left[\lambda_i\left(\mathbf{A}\right)\right]^2$, de esa forma $\lambda_1^2 = \lambda_2^2$ y el análisis continúa en el caso anterior, donde la matriz es ahora $\mathbf{A}^2$.

- $\lambda_2 = \overline{\lambda}_1$: los dos autovalores dominantes son complejos conjugados. Aquí surgen oscilaciones amortiguadas en la secuencia de vectores $\mathbf{q}^{(k)}$ y el método de la potencia no es convergente.

Ejercicio 15. *Calcular los autovalores, en forma exacta, de las siguientes matrices:*

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & -1 \\ 0 & -2 & 0 \\ 1 & 2 & 3 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 2 & 2 & -1 \\ 0 & -2 & 0 \\ 1 & 2 & 4 \end{bmatrix}$$

k	$\mathbf{z}^{(k)}$			$\sigma_b^{(k)}$	ϵ
1	2,68745083	0,51644235	0,96777480	1,13881276	7,44E-01
2	0,40789721	1,37750431	0,72105056	0,97775407	7,90E-01
3	2,07883976	-1,00625369	0,48298148	0,62247881	8,36E-01
4	-0,91226195	2,55035834	0,42451366	0,40363956	6,62E-01
5	2,10956724	-2,43707382	0,14792618	0,36073872	2,64E-01
6	-1,65775979	2,77184183	0,13017111	0,34016562	1,51E-01
7	1,90490276	-2,68772033	0,00538688	0,34132562	6,31E-02
8	-1,79981970	2,74318308	0,05392959	0,33919441	3,02E-02
9	1,84897697	-2,72216084	-0,02601755	0,33987508	1,33E-02
10	-1,82697741	2,73278258	0,03829141	0,33952442	6,14E-03
11	1,83699046	-2,72827832	-0,03259008	0,33968049	2,74E-03
12	-1,83245964	2,73041354	0,03512142	0,33961138	1,25E-03
13	1,83451194	-2,72947687	-0,03395570	**0,33964380**	5,64E-04

Tabla 5.3: Salida del algoritmo del método de la potencia inversa para μ_b del ejemplo 36.

k	$\mathbf{z}^{(k)}$			$\sigma_c^{(k)}$	ϵ
1	-20,28173697	-12,45611532	-12,52653653	1,31742981	2,69E-01
2	11,72018858	7,00711608	7,08443481	1,31499998	7,74E-03
3	-11,70664100	-6,98807132	-7,06900232	**1,31488478**	4,66E-04

Tabla 5.4: Salida del algoritmo del método de la potencia inversa para μ_c del ejemplo 36.

$$\mathbf{C} = \begin{bmatrix} 1 & -3 & -1 \\ 0 & -2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Ejercicio 16. *Aplicar el método de la potencia a las matrices del ejercicio anterior. Elegir los parámetros de stop en forma conveniente para verificar lo desarrollado sobre problemas de convergencia.*

5.4. Ejercicios

1. Construir los siguientes algoritmos en PC:

   a) **Teorema de Gerschgorin**. Entrada: una matriz cuadrada de orden n. Salida: un gráfico con n círculos. Opcional: graficar los autovalores dentro de los círculos.

   b) **Método de la potencia para autovalor dominante**. Entrada: una matriz cuadrada de orden n; un vector de n elementos; la cantidad máxima de iteraciones; el error de terminación. Salida: el autovalor dominante. Opcional: la lista de convergencia del método; un autovector asociado al autovalor dominante.

   c) **Método de la potencia inversa**. Entrada: una matriz cuadrada de orden n; un vector de n elementos; el desplazamiento; la cantidad máxima de iteraciones; el error de terminación. Salida: el autovalor más cercano al desplazamiento. Opcional: la lista de convergencia del método; un autovector asociado al autovalor obtenido.

2. Graficar los círculos de Gerschgorin para las matrices:

   a)
   $$\mathbf{A} = \begin{bmatrix} 8 & 3 & -3 \\ 1 & 2 & 7 \\ 9 & -1 & 4 \end{bmatrix}$$

   b)
   $$\mathbf{B} = \begin{bmatrix} 6 & 6 & 5 \\ 0 & 5 & 9 \\ -6 & 0 & 7 \end{bmatrix}$$

   c)
   $$\mathbf{C} = \begin{bmatrix} 7 & 2 & -1 \\ 1 & -6 & 1 \\ 2 & 1 & 0 \end{bmatrix}$$

3. Obtener los autovalores de las matrices del ejercicio 2 utilizando la definición.

4. Estimar el autovalor dominante, utilizando el método de la potencia, de las matrices del ejercicio 2. Detener la iteración cuando tres lugares decimales queden estables.

5. Estimar el autovalor mínimo, de las siguientes matrices:

   a)
   $$\mathbf{A} = \begin{bmatrix} 0{,}0603 & 0{,}3875 & 0{,}4970 \\ 0{,}2542 & 0{,}0861 & 0{,}8659 \\ 0{,}1197 & 0{,}9146 & 0{,}4674 \end{bmatrix}$$

b)

$$\mathbf{B} = \begin{bmatrix} 0{,}2972 & 0{,}4980 & 0{,}6083 \\ 0{,}7698 & 0{,}2778 & 0{,}9477 \\ 0{,}8995 & 0{,}5259 & 0{,}6375 \end{bmatrix}$$

c)

$$\mathbf{C} = \begin{bmatrix} 0{,}3437 & 0{,}6982 & 0{,}5686 \\ 0{,}5991 & 0{,}2987 & 0{,}0485 \\ 0{,}3273 & 0{,}3960 & 0{,}4786 \end{bmatrix}$$

6. Construir matrices que muestren los problemas de implementación planteados en el apunte y resolverlos utilizando *Euler Math Toolbox* para mostrar cuál es el comportamiento del algoritmo.

7. Demostrar que $\mathbf{I} - \mathbf{AB}$ tiene los mismos autovalores que $\mathbf{I} - \mathbf{BA}$, siempre que $\mathbf{A}$ ó $\mathbf{B}$ sea nosingular.

8. Sean $\lambda_1, \lambda_2, \ldots, \lambda_n$ los autovalores de la matriz $\mathbf{A} \in \mathbb{R}^{n \times x}$. Calcular los autovalores de $\mathbf{A} + \alpha\mathbf{I}$, donde $\mathbf{I}$ es la matriz identidad de orden n y $\alpha \in \mathbb{R}$.

9. Sea $\mathbf{A} = \mathbf{LU}$, donde $\mathbf{L}$ es triangular inferior con los elementos de la diagonal principal iguales a 1 y $\mathbf{U}$ es triangular superior. Sea $\mathbf{B} = \mathbf{UL}$. Mostrar que $\mathbf{A}$ y $\mathbf{B}$ tienen los mismos autovalores.

10. Una matriz es denominada *nilpotente* si $\mathbf{A}^k = \mathbf{0}$ para algún $k > 0$. Mostrar que una matriz nilpotente sólo puede tener autovalores que verifican $|\lambda_i| < 1$.

11. Sea $\mathbf{A}$ una matriz real, donde todos los discos de Gerschgorin son disjuntos. Demostrar que todos los autovalores de la matriz $\mathbf{A}$ son reales.

12. Verificar que el método de la potencia no puede obtener el autovalor dominante de la matriz:

$$\mathbf{A} = \begin{bmatrix} 1/3 & 2/3 & 2 & 3 \\ 1 & 0 & -1 & 2 \\ 0 & 0 & -5/3 & -2/3 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

y explicar por qué.

13. Por medio de los círculos de Gerschgorin, estimar la cantidad máxima de autovalores complejos que pueden tener las siguientes matrices:

a)

$$\mathbf{A} = \begin{bmatrix} 2 & -1/2 & 0 & -1/2 \\ 0 & 4 & 0 & 2 \\ -1/2 & 0 & 6 & 1/2 \\ 0 & 0 & 1 & 9 \end{bmatrix}$$

b)

$$\mathbf{B} = \begin{bmatrix} -5 & 0 & 1/2 & 1/2 \\ 1/2 & 2 & 1/2 & 0 \\ 0 & 1 & 0 & 1/2 \\ 0 & 1/4 & 1/2 & 3 \end{bmatrix}$$

14. Dadas las matrices $\mathbf{A}$ y $\mathbf{E}$, calcular los autovalores de $\mathbf{A}$, luego de $\mathbf{A} + \mathbf{E}$ y determinar por qué existe tanta diferencia entre ellos cuando $\epsilon \to 0$:

$$\mathbf{A} = \begin{bmatrix} 101 & -90 \\ 110 & -98 \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} -\varepsilon & -\varepsilon \\ 0 & 0 \end{bmatrix}$$

15. Dada la matriz:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & -12 \\ 1 & 0 & 11 \\ 0 & 1 & 2 \end{bmatrix},$$

¿por qué si se utiliza el método de la potencia y $\mathbf{x}^{(0)} = [4; -5; 1]^T$, la convergencia es a $\lambda = -3$, pero si se cambia la semilla a $\mathbf{x}^{(0)} = [4; -5; 1+1\times10^{-10}]^T$ se converge a $\lambda = 4$?

16. Considerar la matriz:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Mostrar que esta matriz tiene cuatro autovalores con módulo igual a 1. Justificar por qué el método de la potencia para esta matriz presenta un comportamiento que depende del vector semilla. Probar con varios casos: (1) $x_1 = x_2 = 0$, (2) $x_3 = x_4 = 0$ y (3) $x_1 = x_2 = x_3 = 0$.

17. [*EMT*] Modificar el código del método de la potencia para que opere bajo norma infinito, en vez de la norma 2 utilizada por definición.

18. Es posible demostrar que:

$$\rho(\mathbf{A}) = \lim_{k\to\infty} \left(\left\| \mathbf{A}^k \right\| \right)^{1/k}.$$

¿Qué relación existe con el método de la potencia? ¿Por qué es mejor aplicar el método de la potencia en vez de la secuencia iterativa presentada?

19. Dada una matriz $\mathbf{A}$ se verifica que:

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1},$$

donde $\mathbf{P}$ es la matriz formada con los autovectores de $\mathbf{A}$ ubicados en columnas y $\mathbf{D}$ es la matriz diagonal tal que sus elementos son los autovalores de $\mathbf{A}$. Demostrar por inducción que $\mathbf{A}^n = \mathbf{P}\mathbf{D}^n\mathbf{P}^{-1}$.

20. Uno de los métodos de deflación, conocido como **Método de Hotelling**, genera una matriz $\mathbf{B}$ a partir de la definición:

$$\mathbf{B} = \mathbf{A} - \lambda_i \mathbf{x}_i \mathbf{y},$$

donde λ_i es el i-ésimo autovalor de $\mathbf{A}$, $\mathbf{x}_i$ es un autovector asociado a λ_i y el vector $\mathbf{y}$ cumple que $\mathbf{x}^T\mathbf{y} = 1$. Los autovalores de $\mathbf{B}$ son los mismos que los de la matriz $\mathbf{A}$ a excepción de λ_i. Para las matrices del ejercicio 2, calcular el segundo autovalor en magnitud.

21. Dada una matriz $\mathbf{A}$, tridiagonal y simétrica, es posible calcular su polinomio característico sin aplicar la definición. Para ello se utiliza la secuencia de polinomios de Sturn:

$$\begin{aligned} P_0(\lambda) &= 1 \\ P_1(\lambda) &= d_1 - \lambda \\ P_i(\lambda) &= (d_i - \lambda)P_{i-1}(\lambda) - c_{i-1}^2 P_{i-2}(\lambda), \end{aligned}$$

para $i = 2, 3, \ldots, n$, donde los elementos d_i se ubican en la diagonal principal y los c_i se ubican fuera de ella.

a) Implementar este código en PC.

b) Comprobar su eficiencia con la matriz:

$$\mathbf{A} = \begin{bmatrix} 4 & 9 & 0 \\ 9 & 3 & -2 \\ 0 & -2 & -1 \end{bmatrix}$$

22. Una aplicación interesante de la secuencia de Sturn es que permite calcular la cantidad de autovalores menores que un cierto valor real α, de acuerdo a la siguiente regla: *La cantidad de cambios de signo en la secuencia* $P_0(\alpha), P_1(\alpha), \ldots, P_n(\alpha)$ *es igual al número de autovalores menores que* α. *Si* $P_i(\alpha) = 0$, *entonces debe tomarse el signo opuesto a* $P_{i-1}(\alpha)$.

a) Para la matriz del ejercicio anterior, calcular la cantidad de autovalores menores que $\alpha = 2$.

b) ¿Existirán problemas de implementación para autovalores complejos? Justificar.

23. Por medio de la secuencia de Sturn, y dada la matriz:

$$\mathbf{A} = \begin{bmatrix} 5 & -2 & 0 & 0 \\ -2 & 4 & -1 & 0 \\ 0 & -1 & 4 & -2 \\ 0 & 0 & -2 & 5 \end{bmatrix}$$

a) Mostrar que $\mathbf{A}$ posee un autovalor en el intervalo $[2; 4]$.

b) Calcular el polinomio característico.

c) Calcular el autovalor que está en el intervalo $[2; 4]$ utilizando el método de bisección sobre el polinomio del inciso anterior.

24. Se denomina *matriz compañera* a la matriz:

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -c_n \\ 1 & 0 & 0 & \cdots & 0 & -c_{n-1} \\ 0 & 1 & 0 & \cdots & 0 & -c_{n-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -c_2 \\ 0 & 0 & 0 & \cdots & 1 & -c_1 \end{bmatrix}$$

y su polinomio característico es:

$$P(\lambda) = \lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \ldots + c_{n-1}\lambda + c_n$$

a) Generar una matriz tal que sus autovalores sean $\lambda_1 = -5$; $\lambda_2 = 2$ y $\lambda_3 = 1$.

b) ¿Puede generarse una matriz que tenga como autovalores a $\lambda_1 = i$, $\lambda_2 = -i$?

c) ¿Puede generarse una matriz que tenga autovalores complejos no conjugados?

25. [*EMT*] Dada la matriz:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ -2 & 4 & 5 \\ -3 & -5 & 4 \end{bmatrix},$$

aplicar el método de la potencia, con $\mathbf{x}^{(0)} = [1; 1; 1]^T$ y:

a) 300 iteraciones, graficar la evolución de $\mathcal{L}_1$ a través de las iteraciones.

b) $1{,}5 \times 10^5$ iteraciones, calcular el promedio de las salidas $\mathcal{L}_1$.

c) Mostrar que, al utilizar el método de la potencia inversa con 5×10^6 iteraciones y $\mu = 13{,}817$ el resultado oscila en forma amortiguada alrededor del autovalor real de $\mathbf{A}$. Sin embargo, si se elige $\mu = 13{,}818$, los iterados de $\mathcal{L}$ oscilan alrededor de la parte real del par de autovalores complejos conjugados. ¿A qué se debe esto?

Bibliografía

- *An introduction to numerical analysis*, Kendall ATKINSON, Cap.9
- *Análisis numérico*, R. BURDEN y J. FAIRES, Cap.9
- *Métodos numéricos con MATLAB*, J. MATHEWS y K. FINK, Cap.11
- *Numerical analysis**, Larkin SCOTT, Cap.15
- *Numerical mathematics**, A. QUARTERONI, R. SACCO y F. SALERI, Cap.5

6 Sistemas de Ecuaciones No Lineales

La resolución de sistemas de ecuaciones no lineales es un problema que surge frecuentemente. En general se trata de evitar desarrollos que contengan este tipo de esquemas, tomando aproximaciones lineales en un proceso llamado *linealización*. Cuando, por la exactitud requerida o la complejidad del modelo, no se puede evitar el tener que resolver un sistema de ecuaciones no lineales se recurre a dos métodos numéricos clásicos: punto fijo y Newton-Raphson. Cuando el sistema a resolver posee una gran cantidad de variables, es conveniente recurrir a otras técnicas entre las que se destacan metaheurísticas como *evolución diferencial* o *particle swarm optimization*.

La expresión general de un sistema de ecuaciones no lineales es:

$$\begin{array}{c} f_1(x_1, x_2, \ldots, x_n) = 0 \\ f_2(x_1, x_2, \ldots, x_n) = 0 \\ \vdots \\ f_n(x_1, x_2, \ldots, x_n) = 0 \end{array},$$

donde cada $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, 2, \ldots, n$ puede ser una una función no lineal con respecto a cualquiera de sus variables. El sistema también puede expresarse de la forma $F(\mathbf{x}) = 0$, donde F es la función vectorial $F : \mathbb{R}^n \to \mathbb{R}^n$ tal que:

$$F(x_1, x_2, \ldots, x_n) = [f_1(x_1, x_2, \ldots, x_n), f_2(x_1, x_2, \ldots, x_n), \ldots, f_n(x_1, x_2, \ldots, x_n)].$$

Las funciones $f_1, f_2, \ldots, f_n$ reciben el nombre de **funciones coordenadas** de F. Por otra parte, si $F \in C^1(\mathbb{R}^n)$, se denomina **matriz jacobiana** de F en $\mathbf{x} \in \mathbb{R}^n$ a la matriz real de tamaño $n \times n$:

$$\mathbf{J_F}(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{bmatrix}$$

Se procederá a extender los métodos de punto fijo y Newton-Raphson a fin de poderse aplicar a sistemas con varias variables. Sin pérdida de generalidad, se trabajará con sistemas bidimensionales puesto que es simple extender estos métodos a sistemas con más variables.

6.1. Método multimensional de Punto fijo

Si se desea resolver el sistema:

$$\begin{aligned} \hat{f}_1(x,y) &= 0 \\ \hat{f}_2(x,y) &= 0 \end{aligned}$$

por medio del método de punto fijo es necesario poder escribirlo como:

$$\begin{aligned} x - f_1(x,y) &= 0 \\ y - f_2(x,y) &= 0 \end{aligned}, \tag{6.1}$$

para poder generar las iteraciones que llevarán a la solución. Es muy frecuente el caso en que los sistemas no lineales presentan más de una solución. La notación vectorial servirá para una descripción más compacta de este método.

Se define $\mathbf{x} = [x,y]^T \in \mathbb{R}^2$, $f_1(\mathbf{x}) = f_1(x,y)$, $f_2(\mathbf{x}) = f_2(x,y)$ y:

$$F(\mathbf{x}) = \begin{bmatrix} f_1(x,y) \\ f_2(x,y) \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \end{bmatrix}.$$

Entonces el sistema no lineal (6.1) se transforma en:

$$\mathbf{x} = F(\mathbf{x}), \tag{6.2}$$

y el punto fijo $\mathbf{p} \in \mathbb{R}^2$ de F satisface:

$$\mathbf{p} = F(\mathbf{p}). \tag{6.3}$$

Sabiendo que $\mathbb{R}^2$ es un espacio normado, entonces:

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^{2} x_i^2 = x_1^2 + x_2^2 = x^2 + y^2. \tag{6.4}$$

Ahora, considerando una sucesión de vectores $\{\mathbf{x}^{(n)}\}$, se dice que ésta converge a $\mathbf{x}$ si y sólo si:

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

Además, como $\mathbb{R}^2$ con la norma 2 es un espacio completo, toda sucesión de Cauchy será convergente.

Del mismo modo que para las funciones escalares, se considerará la secuencia de vectores $\mathbf{p}^{(n)}$ tales que:

$$\mathbf{p}^{(n+1)} = F(\mathbf{p}^{(n)}), \qquad n \in \mathbb{N}.$$

Teorema 17. *Sea $\mathcal{R}$ es un subconjunto cerrado de $\mathbb{R}^2$, $F : \mathcal{R} \to \mathcal{R}$, y F es contractiva en $\mathcal{R}$. Entonces:*

$$\mathbf{x} = F(\mathbf{x})$$

tiene una única solución $\mathbf{p} \in \mathbb{R}^2$. La sucesión $\{\mathbf{p}^{(n)}\}$, donde:

$$\mathbf{p}^{(n+1)} = F(\mathbf{p}^{(n)}), \qquad \mathbf{p}^{(0)} \in \mathcal{R}, \qquad n \in \mathbb{N} \tag{6.5}$$

es tal que:

$$\lim_{n\to\infty} \|\mathbf{p}^{(n)} - \mathbf{p}\| = 0$$

y

$$\|\mathbf{p}^{(n)} - \mathbf{p}\| \leq \frac{\alpha^n}{1-\alpha} \|\mathbf{p}^{(b)} - \mathbf{p}^{(a)}\|, \tag{6.6}$$

donde $\|F(\mathbf{p}^{(b)}) - F(\mathbf{p}^{(a)})\| \leq \alpha \|\mathbf{p}^{(b)} - \mathbf{p}^{(a)}\|$ para cualquier $\mathbf{p}^{(a)}, \mathbf{p}^{(b)} \in \mathcal{R}$, y $0 < \alpha < 1$.

Una elección típica para $\mathcal{R}$ es la región rectangular limitada y cerrada:

$$\mathcal{R} = \{[x,y]^T / a_1 \leq x \leq b_1, \quad a_2 \leq y \leq b_2\}. \tag{6.7}$$

El siguiente teorema aplica la desigualdad de Schwarz a la estimación de α. Es cierto que aplicar este teorema es a menudo complicado en la práctica, y es más simple utilizar el método y fallar experimentalmente que comprobar las condiciones necesarias. Sin embargo, hay muchas excepciones por lo que conocer el teorema es útil.

Teorema 18. *Sea $\mathcal{R} \subset \mathbb{R}^2$ como se definió en (6.7), entonces si:*

$$\alpha = \max_{\mathbf{x}\in R} \left[\left(\frac{\partial f_1}{\partial x_1}\right)^2 + \left(\frac{\partial f_1}{\partial x_2}\right)^2 + \left(\frac{\partial f_2}{\partial x_1}\right)^2 + \left(\frac{\partial f_2}{\partial x_2}\right)^2 \right]^{1/2} \tag{6.8}$$

se tiene que:

$$\left\| F(\mathbf{p}^{(b)}) - F(\mathbf{p}^{(a)}) \right\| \leq \alpha \|\mathbf{p}^{(b)} - \mathbf{p}^{(a)}\|$$

para cualquier $\mathbf{p}^{(a)}, \mathbf{p}^{(b)} \in \mathcal{R}$.

Demostración. Dados $\mathbf{x}_1 = [x_1, y_1]^T$, $\mathbf{x}_2 = [x_2, y_2]^T$, existe un punto $\xi \in \mathcal{R}$ sobre el segmento de línea que une $\mathbf{x}_1$ con $\mathbf{x}_2$ tal que:

$$\begin{aligned} F(\mathbf{x}_1) &= F(\mathbf{x}_2) + F^{(1)}(\xi)(\mathbf{x}_1 - \mathbf{x}_2) \\ &= F(\mathbf{x}_2) + \begin{bmatrix} \dfrac{\partial f_1(\xi)}{\partial x} & \dfrac{\partial f_1(\xi)}{\partial y} \\ \dfrac{\partial f_2(\xi)}{\partial x} & \dfrac{\partial f_2(\xi)}{\partial y} \end{bmatrix} \begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \end{bmatrix} \end{aligned}$$

por lo que

$$\begin{aligned} \|F(\mathbf{x}_1) - F(\mathbf{x}_2)\|^2 &= \left[\frac{\partial f_1(\xi)}{\partial x}(x_1 - x_2) + \frac{\partial f_1(\xi)}{\partial y}(y_1 - y_2) \right]^2 \\ &\quad + \left[\frac{\partial f_2(\xi)}{\partial x}(x_1 - x_2) + \frac{\partial f_2(\xi)}{\partial y}(y_1 - y_2) \right]^2 \\ &\leq \left[\left(\frac{\partial f_1(\xi)}{\partial x}\right)^2 + \left(\frac{\partial f_1(\xi)}{\partial y}\right)^2 \right] \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \\ &\quad + \left[\left(\frac{\partial f_2(\xi)}{\partial x}\right)^2 + \left(\frac{\partial f_2(\xi)}{\partial y}\right)^2 \right] \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \end{aligned}$$

a través de la desigualdad de Schwarz. En consecuencia:

$$\begin{aligned} \|F(\mathbf{x}_1) - F(\mathbf{x}_2)\|^2 &\leq \left[\left(\frac{\partial f_1(\xi)}{\partial x}\right)^2 + \left(\frac{\partial f_1(\xi)}{\partial y}\right)^2 \right] \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \\ &\quad + \left[\left(\frac{\partial f_2(\xi)}{\partial x}\right)^2 + \left(\frac{\partial f_2(\xi)}{\partial y}\right)^2 \right] \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \\ &\leq \left\{ \max_{\mathbf{x}\in\mathcal{R}} \left[\left(\frac{\partial f_1(\mathbf{x})}{\partial x}\right)^2 + \left(\frac{\partial f_1(\mathbf{x})}{\partial y}\right)^2 \right. \right. \\ &\quad \left. \left. + \left(\frac{\partial f_2(\mathbf{x})}{\partial x}\right)^2 + \left(\frac{\partial f_2(\mathbf{x})}{\partial y}\right)^2 \right] \right\} \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \\ &= \alpha^2 \|\mathbf{x}_1 - \mathbf{x}_2\|^2 . \end{aligned}$$

□

Para aplicar este teorema, lo ideal es seleccionar F y $\mathcal{R}$ de forma tal que $0 < \alpha < 1$. Como ya se remarcó, esto es a menudo totalmente experimental. El proceso iterativo (6.5) puede ser utilizado como algoritmo siempre que se establezca de antemano un criterio de *stop*. Una buena elección, para $\epsilon > 0$, es terminar el algoritmo cuando:

$$\frac{\left\|\mathbf{p}^{(n)} - \mathbf{p}^{(n-1)}\right\|}{\left\|\mathbf{p}^{(n)}\right\|} < \epsilon, \qquad \mathbf{p}^{(n)} \neq 0. \tag{6.9}$$

Es permitido, en la condición (6.9), usar cualquier norma. Como se parte de una elección arbitraria de F y $\mathcal{R}$, es posible que dicha elección sea incorrecta. Es decir, la convergencia puede no ocurrir. Por esto, es necesario agregarle al algoritmo implementado en computadora, una condición extra para *stop*: que finalice luego de m iteraciones.

El siguiente ejemplo muestra la utilización correcta del teorema de punto fijo. Sin embargo, en la práctica no ocurre así, es más simple iniciar las iteraciones y fallar que comprobar las condiciones solicitadas.

Ejemplo 37. *Sea el sistema de ecuaciones:*

$$x^2 + y^2 - 1 = 0$$
$$5x^2 + 21y^2 - 9 = 0,$$

que debe resolverse para valores reales dentro del primer cuadrante. El mapeo multidimensional escogido es:

$$x_{n+1} = \sqrt{1 - y_n^2}$$
$$y_{n+1} = \sqrt{\frac{9 - 5x_n^2}{21}}.$$

De acuerdo a las condiciones de dominio real, es necesario que:

$$|y| < 1$$
$$|x| < \frac{3}{\sqrt{5}} \approx 1{,}34164.$$

La matriz jacobiana del sistema es:

$$\begin{bmatrix} \dfrac{\partial g_1}{\partial x} & \dfrac{\partial g_1}{\partial y} \\ \dfrac{\partial g_2}{\partial x} & \dfrac{\partial g_2}{\partial y} \end{bmatrix} = \begin{bmatrix} 0 & -y\left(1 - y^2\right)^{-1/2} \\ -\dfrac{5x}{21}\left(\dfrac{9 - 5x^2}{21}\right)^{-1/2} & 0 \end{bmatrix}.$$

Eligiendo por ejemplo el punto inicial $\mathbf{x}^{(0)} = [0{,}5; 0{,}9]$ *el sistema converge. Es importante notar que este punto no cumple con las condiciones del teorema 17, por lo que no se pudo asegurar su convergencia antes de realizar las cuentas.*

La generación del mapeo multidimensional es algo artesanal. A continuación su muestran algunos ejemplos de mapeos convergentes y divergentes.

Ejemplo 38. *Sea el sistema de ecuaciones no lineales:*

$$\hat{f}_1(x, y) = x - x^2 - \frac{y^2}{4} = 0$$
$$\hat{f}_2(x, y) = y - x^2 + y^2 = 0,$$

donde se utilizará el método de punto fijo multidimensional con una tolerancia de $\epsilon = 0{,}0001$ *y un máximo de 50 iteraciones. Se elige la siguiente función de iteración vectorial:*

$$f\left(\mathbf{x}^{(k+1)}\right) = \begin{bmatrix} f_1\left(\mathbf{x}^{(k)}\right) \\ f_2\left(\mathbf{x}^{(k)}\right) \end{bmatrix},$$

de forma tal que:

$$f_1\left(\mathbf{x}^{(k)}\right) = (x_k)^2 + \tfrac{1}{4}(y_k)^2$$

y

$$f_2\left(\mathbf{x}^{(k)}\right) = (x_k)^2 - (y_k)^2 .$$

Si el vector semilla es $\mathbf{x}^{(0)} = [1;1]^T$*, entonces el método no converge. Se muestra en la tabla 6.1 las primeras 6 iteraciones.*

k	$\mathbf{x}^{(k)}$		$f(\mathbf{x}^{(k)})$		ϵ_a	ϵ_r
0	1,00000000	1,00000000	1,25000000	0,000000000	-	-
1	1,25000000	0,00000000	1,56250000	1,562500000	1,56E+00	1,00E+00
2	1,56250000	1,56250000	3,05175781	0,000000000	1,56E+00	1,00E+00
3	3,05175781	0,00000000	9,31322575	9,313225746	9,31E+00	1,00E+00
4	9,31322575	9,31322575	108,420217	0,000000000	9,91E+01	1,06E+01
5	108,420217	0,00000000	11754,9435	11754,94351	1,18E+04	1,00E+00
6	11754,9435	11754,9435	172723371	0,000000000	1,73E+08	1,47E+04

Tabla 6.1: Salida del algoritmo de punto fijo multidimensional, sin convergencia, para el ejemplo 38.

Sin embargo, modificando la función de iteración vectorial:

$$f_1\left(\mathbf{x}^{(k)}\right) = \frac{x_k^2}{x_k^2 + \frac{1}{4}y_k^2}$$

y

$$f_2\left(\mathbf{x}^{(k)}\right) = \frac{x_k^2}{1 + y_k^2},$$

y utilizando los mismos parámetros que en las iteraciones anteriores, el algoritmo converge luego de 14 iteraciones. La salida se muestra en la tabla 6.2.

k	$\mathbf{x}^{(k)}$		$f(\mathbf{x}^{(k)})$		ϵ_a	ϵ_r
0	1,00000000	1,00000000	0,80000000	0,50000000	-	-
1	0,80000000	0,50000000	0,91103203	0,42666667	1,11E-01	2,22E-01
2	0,91103203	0,42666667	0,94801644	0,58176123	1,55E-01	2,67E-01
3	0,94801644	0,58176123	0,91395554	0,56818637	3,41E-02	5,85E-02
4	0,91395554	0,56818637	0,91189208	0,53266292	3,55E-02	6,25E-02
5	0,91189208	0,53266292	0,92140272	0,54255059	9,89E-03	1,82E-02
6	0,92140272	0,54255059	0,92023367	0,55037609	7,83E-03	1,42E-02
7	0,92023367	0,55037609	0,91791473	0,54620940	4,17E-03	7,57E-03
8	0,91791473	0,54620940	0,91867646	0,54492454	1,28E-03	2,35E-03
9	0,91867646	0,54492454	0,91915104	0,54628328	1,36E-03	2,49E-03
10	0,91915104	0,54628328	0,91885718	0,54636731	2,94E-04	5,38E-04
11	0,91885718	0,54636731	0,91878654	0,54598834	3,79E-04	6,94E-04
12	0,91878654	0,54598834	0,91887857	0,54603821	9,20E-05	1,69E-04
13	0,91887857	0,54603821	0,91887988	0,54612998	9,18E-05	1,68E-04
14	0,91887988	0,54612998	**0,91885504**	**0,54609913**	3,09E-05	5,65E-05

Tabla 6.2: Salida del algoritmo de punto fijo multidimensional, solución obtenida del ejemplo 38.

Ejercicio 17. *Modificar los parámetros iniciales para que el algoritmo de punto fijo multidimensional converja a la otra raíz real del sistema de ecuaciones del ejemplo anterior. ¿Qué conclusión se puede obtener después de probar con valores iniciales cercanos a la otra raíz real?*

6.2. Método multidimensional de Newton-Raphson

Dado el sistema de ecuaciones:

$$\begin{aligned} F_1(x,y) &= F_1(\mathbf{x}) = 0 \\ F_2(x,y) &= F_2(\mathbf{x}) = 0 \end{aligned}, \tag{6.10}$$

se resolverá por medio del método de Newton-Raphson multivariable. Si se asume que una aproximación de la solución del sistema anterior es igual a:

$$\mathbf{x}^{(0)} = [x_0, y_0]^T$$

entonces resolver utilizando el método de Newton consiste en encontrar correcciones $\Delta\mathbf{x} = [\Delta x, \Delta y]$, definidas para las incógnitas de modo que el vector:

$$\begin{aligned} \mathbf{x}^{(1)} &= [x_1, y_1] \\ &= [x_0 + \Delta x_0, y_0 + \Delta y_0] \\ &= \mathbf{x}^{(0)} + \mathbf{\Delta x}^{(0)} \end{aligned} \tag{6.11}$$

sea la solución que se busca. Desarrollando las funciones en el lado izquierdo del sistema (6.10) por medio de una serie de Taylor alrededor del vector conocido $\mathbf{x}^{(0)}$:

$$\begin{aligned} F_1(\mathbf{x}) &\approx F_1\left(\mathbf{x}^{(0)}\right) + \frac{\partial F_1\left(\mathbf{x}^{(0)}\right)}{\partial x}\Delta x_0 + \frac{\partial F_1\left(\mathbf{x}^{(0)}\right)}{\partial y}\Delta y_0 + \ldots \\ F_2(\mathbf{x}) &\approx F_2\left(\mathbf{x}^{(0)}\right) + \frac{\partial F_2\left(\mathbf{x}^{(0)}\right)}{\partial x}\Delta x_0 + \frac{\partial F_2\left(\mathbf{x}^{(0)}\right)}{\partial y}\Delta y_0 + \ldots \end{aligned} \tag{6.12}$$

y sabiendo que $\mathbf{x}$ es la solución del sistema (6.10), entonces $F_i(\mathbf{x}) = 0$. Luego de aplicar esto y truncar la serie de Taylor dejando sólo los términos lineales, se obtiene:

$$\begin{aligned} \frac{\partial F_1\left(\mathbf{x}^{(0)}\right)}{\partial x}\Delta x_0 + \frac{\partial F_1\left(\mathbf{x}^{(0)}\right)}{\partial y}\Delta y_0 &= -F_1\left(\mathbf{x}^{(0)}\right) \\ \frac{\partial F_2(\mathbf{x}^{(0)})}{\partial x}\Delta x_0 + \frac{\partial F_2\left(\mathbf{x}^{(0)}\right)}{\partial y}\Delta y_0 &= -F_2\left(\mathbf{x}^{(0)}\right) \end{aligned} \tag{6.13}$$

lo que puede ser expresado en forma matricial como:

$$\begin{bmatrix} \dfrac{\partial F_1\left(\mathbf{x}^{(0)}\right)}{\partial x} & \dfrac{\partial F_1\left(\mathbf{x}^{(0)}\right)}{\partial y} \\ \dfrac{\partial F_2\left(\mathbf{x}^{(0)}\right)}{\partial x} & \dfrac{\partial F_2\left(\mathbf{x}^{(0)}\right)}{\partial y} \end{bmatrix} \cdot \begin{bmatrix} \Delta x_0 \\ \Delta y_0 \end{bmatrix} = \begin{bmatrix} -F_1\left(\mathbf{x}^{(0)}\right) \\ -F_2\left(\mathbf{x}^{(0)}\right) \end{bmatrix}, \tag{6.14}$$

o en forma vectorial como:

$$\mathbf{J_F}\left(\mathbf{x}^{(0)}\right) \cdot \mathbf{\Delta x}^{(0)} = -F\left(\mathbf{x}^{(0)}\right) \tag{6.15}$$

Ahora, $\mathbf{\Delta x}^{(0)}$ representa la mejora de la aproximación $\mathbf{x}^{(0)}$ con respecto a la solución $\mathbf{x}$. Pero $\mathbf{\Delta x}^{(0)} = \mathbf{x}^{(1)} - \mathbf{x}^{(0)}$ sin que $\mathbf{x}^{(1)}$ sea la solución de (6.10) ya que la serie de Taylor

fue truncada. Como $\mathbf{x}^{(1)}$ está más cerca de la solución que $\mathbf{x}^{(0)}$ se puede repetir este proceso para obtener una mejor aproximación, llamada $\mathbf{x}^{(2)}$. Generando un algoritmo iterativo, se obtiene lo que se denomina método de Newton multidimensional:

$$\mathbf{J_F}\left(\mathbf{x}^{(k)}\right) \cdot \mathbf{\Delta x}^{(k)} = -F\left(\mathbf{x}^{(k)}\right) \tag{6.16}$$

donde:

$$\mathbf{\Delta x}^{(k)} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}.$$

Al igual que el método de punto fijo multidimensional, se recomienda utilizar como condición de *stop*:

$$\frac{\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|}{\left\|\mathbf{x}^{(k+1)}\right\|} < \epsilon,$$

con alguna norma conocida y $\mathbf{x}^{(k+1)} \neq 0$.

Algunas consideraciones que deben ser tenidas en cuenta se refieren a la estabilidad del sistema de ecuaciones lineales (6.16):

- Es necesario utilizar un vector inicial que sea cercano a la solución. De otra forma, la convergencia no se asegura.
- Debe verificarse el condicionamiento de la matriz jacobiana en cada iteración, puesto que cambia en cada paso y al estar cerca de la solución su determinante puede tender a cero.
- Se recomienda resolver el sistema (6.16) utilizando el método de Gauss con pivoteo parcial con el fin de no agregar mucho error de propagación.

Ejemplo 39. *Sea el sistema de ecuaciones no lineales:*

$$F\left(\mathbf{x}\right) = \begin{bmatrix} x^2 + y^2 - 5 \\ x^2 - y^2 + 3 \end{bmatrix} = 0,$$

entonces:

$$\mathbf{J_F}\left(\mathbf{x}\right) = \begin{bmatrix} 2x & 2y \\ 2x & -2y \end{bmatrix}.$$

Aplicando la fórmula iterativa (6.16) con $\epsilon = 0{,}001$ *el sistema de ecuaciones converge luego de 4 iteraciones. La tabla 6.3 muestra la sucesión generada.*

k	$\mathbf{x}^{(k)}$		$F(\mathbf{x}^{(k)})$		$\mathbf{J_F}\left(\mathbf{x}^{(k)}\right)$	ϵ_a	ϵ_r
0	0,500000	1,000000	-3,750000	2,250000	-4,00	-	-
1	1,250000	2,500000	2,812500	-1,687500	-25,00	1,50E+00	6,00E-01
2	1,025000	2,050000	0,253125	-0,151875	-16,81	4,50E-01	2,20E-01
3	1,000304	2,000609	0,003049	-0,001829	-16,01	4,94E-02	2,47E-02
4	**1,00000**	**2,00000**	0,000000	-0,000000	-16,00	6,10E-04	3,05E-04

Tabla 6.3: Iteraciones del algoritmo de Newton-Raphson multidimensional del ejemplo 39.

Ejercicio 18. *El ejemplo anterior muestra un sistema de ecuaciones no lineales y una solución. Encontrar las tres soluciones restantes, utilizando diferentes vectores iniciales y el método de Newton-Raphson multidimensional.*

La selección de la semilla al momento de aplicar el método de Newton-Raphson afecta la estabilidad de la matriz jacobiana a lo largo del proceso iterativo. El proceso de inversión matricial (o la falta de pivoteo en el proceso de Gauss) genera matrices mal condicionadas o inestables debido a la aritmética. Esto se aprecia en el siguiente ejemplo.

Ejemplo 40. *El sistema de ecuaciones no lineales:*

$$e^{x^2+y^2} - 1 = 0$$
$$e^{x^2-y^2} - 1 = 0$$

tiene por matriz jacobiana a:

$$\mathbf{J_F} = \begin{bmatrix} 2xe^{x^2+y^2} & 2ye^{x^2+y^2} \\ 2xe^{x^2-y^2} & -2ye^{x^2-y^2} \end{bmatrix}$$

Generando las iteraciones de acuerdo a:

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} - (\mathbf{J_F})^{-1}(\mathbf{x}^{(n)})\mathbf{F}\left(\mathbf{x}^{(n)}\right),$$

se verifica que el sistema converge rápidamente (17 iteraciones) si $\mathbf{x}^{(0)} = [0{,}1; 0{,}1]^T$ *pero converge en forma lenta (221 iteraciones) si* $\mathbf{x}^{(0)} = [10; 10]^T$*. Sin embargo, no converge si* $\mathbf{x}^{(0)} = [20; 20]^T$*, debido a* overflow*.*

6.2.1. Condiciones de convergencia

La convergencia del método multidimensional de Newton-Raphson depende fuertemente de la semilla utilizada. Es por ello que se considera un método local y no global. A través del siguiente teorema se establece formalmente cuál es la condición de convergencia para una semilla determinada:

Teorema 19. *Sea* $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ *una función de clase* C^1 *en un conjunto convexo abierto* $D \subset \mathbb{R}^n$ *que contiene a* $\mathbf{x}^*$*. Suponiendo que* $J_{\mathbf{F}}^{-1}(\mathbf{x}^*)$ *existe y también existen tres constantes positivas* R*,* C *y* L*, de forma tal que* $\|J_{\mathbf{F}}^{-1}(\mathbf{x}^*)\| \leq C$ *y:*

$$\|\mathbf{J_F}(\mathbf{x}) - \mathbf{J_F}(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in B(\mathbf{x}^*; R),$$

donde $\|\cdot\|$ *denota una única norma vectorial y matricial. Entonces, existe* $r > 0$ *de forma tal que, para cualquier* $\mathbf{x}^{(0)} \in B(\mathbf{x}^*; r)$*, la secuencia de iterados de Newton-Raphson es única y converge a* $\mathbf{x}^*$ *con:*

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| \leq CL\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2.$$

Comandos de EMT. *El comando para aplicar el método de Newton-Raphson multidimensional es:*

- `newton2(f$:string, df$:string, x:numerical)`*, donde* ***f$*** *es una función vectorial de* x *expresada como* string*;* ***df$*** *es el jacobiano de la función vectorial;* ***x*** *es un punto cercano a la solución, utilizado como semilla. La salida es un vector con la solución del sistema.*

Ejemplo en EMT 14. *Calcular la solución del sistema:*

$$f_1(x, y) = \cos(x) - y^2 = 0$$
$$f_2(x, y) = \sin(y) - 2x - 2 = 0$$

utilizando el algoritmo de Newton-Raphson multidimensional y la semilla $[-1; 1]^T$*.*

```
>function f([x,y])&=[cos(x)-y^2,sin(y)-2*x-2];
>function Df([x,y])&=jacobian(f(x,y),[x,y]);
>newton2("f","Df",[-1,1])
 [-0.606324,  0.906504]
```

6.3. Análisis de convergencia

La velocidad de convergencia es el factor que ayudará en la decisión por uno u otro método a la hora de resolver un problema concreto.

Sea $\{\mathbf{x}^{(k)}\}$ una sucesión en $\mathbb{R}^n$ que converge a $\mathbf{x}$. Se dice entonces que la convergencia es:

- **lineal**, si:

$$\exists M, 0 < M < 1, \text{ y } \exists k_0 \text{ tal que } \|\mathbf{x}^{(k+1)} - \mathbf{x}\| \leq M\|\mathbf{x}^{(k)} - \mathbf{x}\|, \forall k \geq k_0;$$

- **cuadrática**, si:

$$\exists M, M > 0, \text{ y } \exists k_0 \text{ tal que } \|\mathbf{x}^{(k+1)} - \mathbf{x}\| \leq M\|\mathbf{x}^{(k)} - \mathbf{x}\|^2, \forall k \geq k_0;$$

- **superlineal**, si:

$$\exists M, M > 0, \text{ y } \exists k_0 \text{ tal que } \|\mathbf{x}^{(k+1)} - \mathbf{x}\| \leq M\|\mathbf{x}^{(k)} - \mathbf{x}\|^P, \forall k \geq k_0, 1 < P < 2.$$

Como no es posible conocer el valor exacto de $\mathbf{x}$, se analiza entonces la *tasa de convergencia*. Al analizar el cociente:

$$\frac{\left\|\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}\right\|}{\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^p},$$

a partir de cierto k grande se puede determinar con qué velocidad p converge la sucesión, siempre y cuando el cociente anterior se estabilice. Esta es la versión extendida de la tasa de convergencia planteada para las sucesiones generadas por los métodos para resolver ecuaciones no lineales.

Ejercicio 19. *Analizar la tasa de convergencia de los dos ejemplos planteados anteriormente.*

6.4. Ejercicios

1. Implementar en *EMT* el algoritmo del **Método multidimensional de Newton**. Entrada: una función de dos dimensiones; un vector semilla; cantidad máxima de iteraciones; tolerancia para stop. Salida: la solución del sistema de ecuaciones.

2. Dado el sistema no lineal:

$$\begin{aligned} x^2 + y^2 &= 4 \\ xy &= 1 \end{aligned}$$

   a) Representar ambas curvas para identificar las soluciones del sistema en forma gráfica.

   b) Resolverlo numéricamente en *EMT* utilizando el método de punto fijo. Utilizar como error de tolerancia 10^{-4}.

   c) Mostrar la convergencia lineal del método.

   d) Transformar el sistema en una ecuación no lineal y resolverla utilizando el método de punto fijo. Comparar el resultado con el obtenido en los incisos anteriores.

3. [EMT] Considerar el sistema no lineal:

$$x^2 - 2x - y + 0{,}5 = 0$$
$$x^2 + 4y^2 - 4 = 0$$

   a) Graficar el sistema.

   b) Resolverlo utilizando el método de Newton, implementado en EMT. Utilizar como error de tolerancia 10^{-15}.

   c) Mostrar la convergencia, superior a cuadrática, del método.

4. Resolver el sistema de ecuaciones:

$$2x = \sin\left(\frac{x+y}{2}\right)$$
$$2y = \cos\left(\frac{x-y}{2}\right)$$

   utilizando el método de punto fijo y las siguientes semillas:

   a) $\mathbf{x}^{(0)} = (0{,}1; 0{,}1)$

   b) $\mathbf{x}^{(0)} = (10; 10)$

   c) $\mathbf{x}^{(0)} = (20; 20)$

5. Dado el sistema de ecuaciones no lineales:

$$x^2 - x - y = -1$$
$$x^2 + y^2 = 1$$

   resolverlo utilizando el método de Newton y la semilla $\mathbf{x}^{(0)} = [0{,}63; 0{,}77]^T$.

6. Dado el sistema no lineal:

$$e^x e^y + x\cos(y) = 0$$
$$x + y - 1 = 0$$

   a) Estimar, gráficamente, las raíces del sistema para $x \in [-6; 6]$.

   b) Transformarlo en una única ecuación no lineal y utilizar las estimaciones del punto anterior como semillas para el método de Newton con $\epsilon = 10^{-5}$.

   c) Comparar la salida del inciso anterior con el método multidimensional de Newton.

7. Para cada una de las raíces del sistema de ecuaciones:

$$x^2 - 2x - y + 1 = 0$$
$$x^2 + y^2 - 1 = 0,$$

   determinar si los siguientes esquemas iterativos son o no localmente convergentes:

   a) $x_{k+1} = \sqrt{1 - y_k^2}$; $y_{k+1} = (x_k - 1)^2$

   b) $x_{k+1} = \sqrt{y_k} + 1$; $y_{k+1} = \sqrt{1 - x_k^2}$

8. Considerar el sistema de ecuaciones no lineales:

$$x^2 - y + \alpha = 0$$
$$-x + y^2 + \alpha = 0.$$

Mostrar gráficamente que, para $\alpha > 0{,}25$ no existe solución, para $\alpha = 0{,}25$ la solución es única (calcularla si es posible a través de punto fijo) y para $\alpha < 0{,}25$ existen dos soluciones.

9. En el siglo XII Omar Khyyam resolvió, a través de métodos geométricos, una ecuación cúbica de la forma:

$$x^3 - cx^2 + b^2x + a^3 = 0.$$

Las raíces de esta ecuación deben estar entre los valores x de las intersecciones de la circunferencia:

$$x^2 + y^2 - \left(c - \frac{a^3}{b^2}\right)x + 2by - b^2 - c\frac{a^3}{b^2} = 0$$

y la hipérbola:

$$xy = \frac{a^3}{b}.$$

a) Resolver, en forma gráfica, el sistema no lineal asociado para los parámetros $a = -1$; $b = 2$ y $c = 4$.

b) ¿Qué ocurre cuando los parámetros son $a = -1$; $b = 1$ y $c = 1$?

10. [*EMT*] Dado el sistema no lineal:

$$f_1(x, y) = 0; \qquad f_2(x, y) = 0,$$

es posible resolverlo a través de la iteración de Newton Jacobi:

$$x_{i+1} = x_i - \frac{\partial f_1}{\partial x}(x_i, y_i)^{-1} f_1(x_i, y_i)$$
$$y_{i+1} = y_i - \frac{\partial f_2}{\partial y}(x_i, y_i)^{-1} f_2(x_i, y_i)$$

a) Graficar el sistema:

$$f_1(x, y) = x^2 - 3y + \cos(xy) = 0$$
$$f_2(x, y) = x + y - \sin(x)\sin(y) = 0,$$

y estimar las soluciones utilizando el gráfico.

b) Aplicar la iteración de Newton Jacobi con $x_0 = -2$; $y_0 = 1$ que es un punto cercano a una de las soluciones. ¿A qué solución converge?

c) Graficar las iteraciones de x_{i+1} e y_{i+1}. ¿Cuál parece lograr convergencia en forma más rápida?

11. La iteración de Newton Gauss Seidel:

$$x_{i+1} = x_i - \frac{\partial f_1}{\partial x}(x_i, y_i)^{-1} f_1(x_i, y_i)$$
$$y_{i+1} = y_i - \frac{\partial f_2}{\partial y}(x_{i+1}, y_i)^{-1} f_2(x_{i+1}, y_i)$$

logra, en la mayoría de los casos, una convergencia más rápida que la iteración de Newton Jacobi. Repetir el ejercicio anterior con este nuevo esquema iterativo.

12. [EMT] Dado el sistema de ecuaciones no lineales:

$$f_1(x,y) = x^2 + y^2 - 2; \qquad f_2(x,y) = x + y - 2,$$

a) Mostrar gráficamente que una de las soluciones del sistema es $[1;1]^T$.

b) Mostrar que, para cualquier semilla $\mathbf{x}^{(0)} = [x_0; y_0]^T$, tal que $x_0 \neq y_0$, se puede asegurar que $\mathbf{x}^{(1)} = [x_1; y_1]^T$ donde $x_1 + y_1 = 2$.

c) Determinar $\mathbf{x}^{(1)}$ cuando $x_1 = 1 + \alpha$, $y_1 = 1 - \alpha$, donde $\alpha \neq 0$.

13. Dado el mapeo multidimensional:

$$x_{k+1} = \frac{0{,}5}{1 + (x_k + y_k)^2}$$
$$y_{k+1} = \frac{0{,}5}{1 + (x_k - y_k)^2}$$

a) Elegir una región que cumpla con las condiciones del teorema 17.

b) Representar gráficamente el sistema de ecuaciones original (no el mapeo).

c) Resolverlo con alguna semilla que esté dentro de la región identificada en el primer inciso.

14. [EMT] Sea el sistema de ecuaciones no lineales:

$$f_1(x,y) = x^2 + y^2 - 1 = 0$$
$$f_2(x,y) = 2x + y - 1 = 0$$

cuyas soluciones son $\mathbf{x}_1^* = [0;1]^T$ y $\mathbf{x}_2^* = [4/5; -3/5]^T$. Es posible definir las dos funciones de iteración:

$$\mathbf{G}_1(\mathbf{x}) = \begin{bmatrix} (1-y)/2 \\ \sqrt{1-x^2} \end{bmatrix}, \qquad \mathbf{G}_2(\mathbf{x}) = \begin{bmatrix} (1-y)/2 \\ -\sqrt{1-x^2} \end{bmatrix}$$

a) Comprobar que $\mathbf{G}_i(\mathbf{x}_i^*) = \mathbf{x}_i^*$ para $i = 1, 2$.

b) Calcular $\rho\left(J_{\mathbf{G}_1(\mathbf{x}_1^*)}\right)$ y $\rho\left(J_{\mathbf{G}_2(\mathbf{x}_2^*)}\right)$. ¿Qué significa la diferencia entre los radios espectrales?

c) Utilizando una tolerancia de 1×10^{-10}, resolver el primer esquema con la semilla $\mathbf{x}^{(0)} = [-0{,}9; 0{,}9]^T$ y el segundo esquema con $\mathbf{x}^{(0)} = [0{,}9; 0{,}9]^T$. ¿A qué se debe la diferencia entre la cantidad de iteraciones?

15. Dadas las curvas $C_1 : y = x^2$ y $C_2 : v = \cos(u)$:

a) Plantear el sistema no lineal que permita determinar la distancia mínima entre C_1 y C_2.

b) Resolver el sistema planteado a través de Newton-Raphson, utilizando la semilla $[0{,}25; 0{,}5]^T$.

16. Considerar la función de dos variables:

$$g(x,y) = \cos(x+y) + \sin(x) + \cos(y).$$

a) Representar gráficamente la función $g(x,y)$ y de ahí obtener semillas para calcular los puntos críticos.

b) Plantear el sistema de ecuaciones no lineales que se debe resolver para determinar los puntos críticos de $g(x,y)$.

c) Resolver el sistema planteado en el inciso anterior a través del método de Newton-Raphson.

d) Determinar la convergencia del método con todas las raíces obtenidas.

17. Utilizar cualquier método para encontrar todas las raíces reales en $0 < x < 1{,}5$ del sistema de ecuaciones:

$$\begin{aligned} \tan(x) - y &= 1 \\ \cos(x) - 3\sin(y) &= 0 \end{aligned}$$

18. La ecuación de una circunferencia es:

$$(x-a)^2 + (y-b)^2 = r^2,$$

donde r es el radio y (a, b) son las coordenadas del centro. Si las coordenadas de tres puntos de la circunferencia se muestran en la tabla 6.4, determinar r, a y b.

x	$8,21$	$0,34$	$5,96$
y	$0,00$	$6,62$	$-1,12$

Tabla 6.4

19. [*EMT*] Si algunas de las ecuaciones en el sistema $F(\mathbf{x}) = \mathbf{0}$ son lineales, entonces el método de Newton-Raphson detecta esta característica. Mostrar con los sistemas dados a continuación que, si $f_i(\mathbf{x})$ es lineal, entonces para $k \geq 1$ se cumple que $f_i(\mathbf{x}^{(k)}) = 0$:

a)

$$\begin{aligned} x + y - 1 &= 0 \\ \sqrt{x}\cos(y) + y - x &= 0 \end{aligned}$$

b)

$$\begin{aligned} 3x + 4y + z + 2 &= 0 \\ e^x + \cos(yz) - 3z &= 0 \\ -x + y - z + 1 &= 0 \end{aligned}$$

20. Siguiendo el razonamiento del ejercicio anterior, ¿qué ocurre cuando se intenta resolver un sistema lineal con el esquema iterativo para sistemas no lineales? ¿Se llega a la solución en un solo paso? ¿Por qué?

21. Dado el sistema:

$$\begin{aligned} f_1(x, y) &= x^3 - xy + x - 1 \\ f_2(x, y) &= y - x^2 \end{aligned}$$

a) Mostrar que $\mathbf{J_F}(\mathbf{x})$ es no singular en la región $D = \left\{\mathbf{x} \in \mathbb{R}^2 / y \neq x^2 + 1\right\}$.

b) Mostrar que, para cualquier vector semilla $\mathbf{x}^{(0)} \in D$, se cumple que $\mathbf{x}^{(k)} \in D$ y finalmente converge a la raíz.

22. [*EMT*] El sistema de ecuaciones lineales:

$$\begin{aligned} f_1(x, y) &= \sin(x) + 3x - y \\ f_2(x, y) &= \sin(y) + 3y - x \end{aligned}$$

tiene como solución única al vector nulo. Sin embargo, al aplicar el método de Newton-Raphson, converge a ciertos valores que no son la solución verdadera.

a) Verificar gráficamente que el vector nulo es solución del sistema planteado.

b) Utilizar como semilla a $\mathbf{x} = [\pi; \pi]^T$ y verificar a qué valores converge.

c) Identificar qué región del plano no converge a la raíz verdadera del sistema.

23. Demostrar que el sistema no lineal:

$$f(x, y) = 0$$
$$g(x, y) = 0$$

puede resolverse a través del método de Newton-Raphson, pero en vez de utilizar la iteración matricial clásica, usando el esquema iterativo:

$$x_{n+1} = x_n - \frac{fg_y - gf_y}{f_x g_y - g_x f_y}; \qquad y_{n+1} = y_n - \frac{f_x g - g_x f}{f_x g_y - g_x f_y},$$

donde todas las funciones involucradas son evaluadas en $[x_n; y_n]$.

24. [*EMT*] Dado el sistema no lineal:

$$f_1(x, y) = x^2 + y^2 - 25 = 0$$
$$f_2(x, y) = x^2 - y - 2 = 0$$

a) Graficar ambas expresiones e identificar las soluciones en forma gráfica.

b) Dado que el sistema planteado es simétrico, analizar qué ocurre cuando se utiliza como semilla algún vector ubicado a la misma distancia de ambas soluciones reales.

c) ¿En qué región del plano es posible escoger semillas que tengan convergencia segura?

25. El sistema no lineal definido por:

$$f_1(x, y) = x^2 + y^2 - 25 = 0$$
$$f_2(x, y) = x^2 - y^2 - 2 = 0$$

posee cuatro soluciones reales, una en cada cuadrante. Repetir los incisos del ejercicio anterior.

Bibliografía

- *A theoretical introduction to numerical analysis**, V. RYABEN'KII y S. TSYNKOV, Cap.8
- *Análisis numérico*, R. BURDEN y J. FAIRES, Cap.10
- *Numerical analysis**, Larkin SCOTT, Cap.7
- *Numerical mathematics**, A. QUARTERONI, R. SACCO y F. SALERI, Cap.7
- *Numerical methods*, G. DAHLQUIST y A. BJÖRK, Cap.6

7 Interpolación

Los conjuntos discretos de datos son muy usuales en computación matemática. El origen de estos datos puede ser observaciones, experimentos o cómputos numéricos. Existe una gran diferencia entre interpolación y ajuste de datos. En interpolación se construye una curva que, necesariamente, pasa a través de todos los datos discretos (puntos). Para hacer esto, debe suponerse que todos los puntos son exactos y distintos. El ajuste de datos se aplica a una gran cantidad de puntos que pueden contener errores, casi siempre generados durante la adquisición de datos. La idea entonces es encontrar una curva que pase lo más cerca de todos los puntos, por lo que no se exige que necesariamente pase por los puntos. Esta es la gran diferencia entre interpolación y ajuste de datos. Un ejemplo gráfico se ve en la figura 7.1.

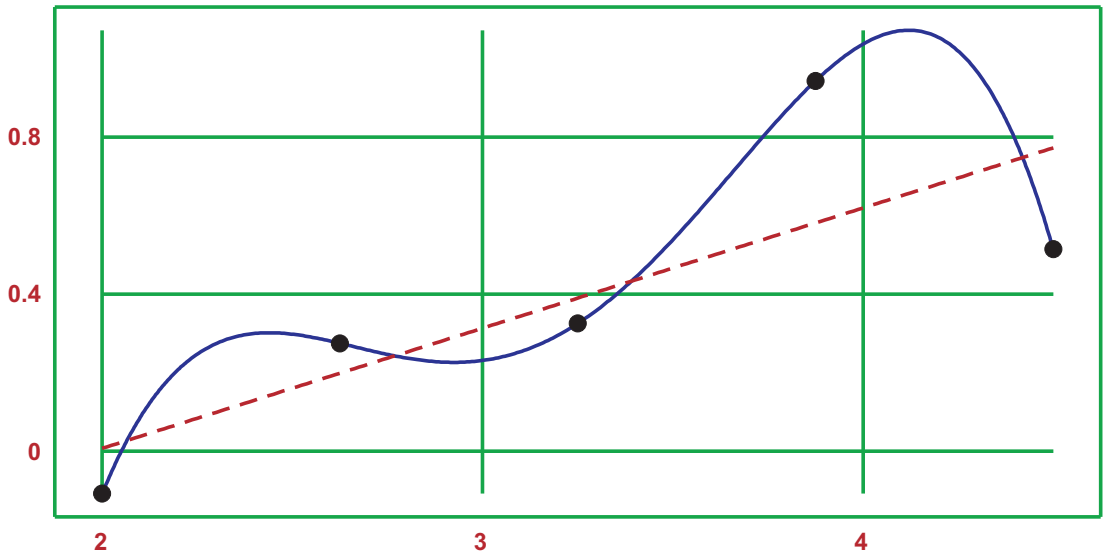

Figura 7.1: Interpolación cúbica (línea continua de color azul) y ajuste de datos lineal (línea discontinua de color rojo) para un conjunto genérico de 5 datos (círculos de color negro).

7.1. Interpolación polinómica

Ante la diversidad de funciones interpoladoras que se pueden construir, es lógico elegir las de desarrollo más simple, en este caso polinomios. Dados $n+1$ puntos de abscisas distintas (x_1, y_1), (x_2, y_2), ..., (x_{n+1}, y_{n+1}), existe un único polinomio $p(x)$ de grado menor o igual que n, que pasa por todos ellos.

7.1.1. Forma normal

El planteamiento más directo del problema de interpolación se obtiene expresando el polinomio con respecto a la base canónica:

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$$

e imponiendo las condiciones de interpolación, $p(x_i) = y_i$, $i = 1, 2, \ldots, n, n+1$. Se obtiene así el siguiente sistema de ecuaciones lineales:

$$\begin{aligned}
a_n x_1^n + a_{n-1}x_1^{n-1} + \ldots + a_1 x_1 + a_0 &= y_1 \\
a_n x_2^n + a_{n-1}x_2^{n-1} + \ldots + a_1 x_2 + a_0 &= y_2 \\
a_n x_3^n + a_{n-1}x_3^{n-1} + \ldots + a_1 x_3 + a_0 &= y_3 \\
&\vdots \\
a_n x_{n+1}^n + a_{n-1}x_{n+1}^{n-1} + \ldots + a_1 x_{n+1} + a_0 &= y_{n+1}
\end{aligned}$$

Este sistema es compatible determinado, puesto que su determinante es no nulo. Pero la resolución del mismo está afectada por el mal condicionamiento de la matriz de coeficientes, llamada **matriz de Vandermonde**, por lo que el resultado puede ser poco fiable. Sin embargo, ese condicionamiento puede mejorarse mediante un desplazamiento del origen, es decir, eligiendo otra base para expresar el polinomio:

$$p(x) = b_n(x-c)^n + b_{n-1}(x-c)^{n-1} + \ldots + b_1(x-c) + b_0.$$

El sistema resultante al imponer las condiciones de interpolación resulta ligeramente mejor condicionado que el original, pero la matriz sigue teniendo la misma estructura y, al aumentar el grado del polinomio, volverá el mal condicionamiento.

7.1.2. Interpolación de Lagrange

La fórmula de Lagrange permite obtener en forma rápida una expresión polinómica de grado $n-1$ para interpolar exitosamente n puntos distintos:

$$P_{n-1}(x) = \sum_{i=1}^{n} y_i \ell_i(x) \tag{7.1}$$

donde:

$$\begin{aligned}
\ell_i(x) &= \frac{x-x_1}{x_i-x_1} \cdot \frac{x-x_2}{x_i-x_2} \cdots \frac{x-x_{i-1}}{x_i-x_{i-1}} \cdot \frac{x-x_{i+1}}{x_i-x_{i+1}} \cdots \frac{x-x_n}{x_i-x_n} \\
&= \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}, \qquad i = i, 2, \ldots, n
\end{aligned} \tag{7.2}$$

son llamadas **funciones cardinales**. Las funciones cardinales son polinomios de grado $n-1$ y tienen la propiedad:

$$\ell_i(x_j) = \left\{ \begin{array}{ll} 0, & \text{si } i \neq j \\ 1, & \text{si } i = j \end{array} \right\} = \delta_{ij} \tag{7.3}$$

donde δ_{ij} es el **Delta de Kronecker**.

Ejemplo 41. *Las funciones cardinales para tres puntos cuyas abscisas son* $x_1 = -2$, $x_2 = 3$ *y* $x_3 = 5$ *son:*

$$\begin{aligned}
\ell_1(x) &= \frac{x - x_2}{x_1 - x_2} \cdot \frac{x - x_3}{x_1 - x_3} = \frac{x^2}{35} - \frac{8x}{35} + \frac{3}{7}, \\
\ell_2(x) &= \frac{x - x_1}{x_2 - x_1} \cdot \frac{x - x_3}{x_2 - x_3} = -\frac{x^2}{10} + \frac{3x}{10} + 1, \\
\ell_3(x) &= \frac{x - x_1}{x_3 - x_1} \cdot \frac{x - x_2}{x_3 - x_2} = \frac{x^2}{14} - \frac{x}{14} - \frac{3}{7}.
\end{aligned} \tag{7.4}$$

El gráfico de las funciones cardinales se muestra en la figura 7.2. Con línea punteada de color rojo se graficó $\ell_1(x)$*; con línea alternada de color azul* $\ell_2(x)$ *y con línea continua de color naranja* $\ell_3(x)$*.*

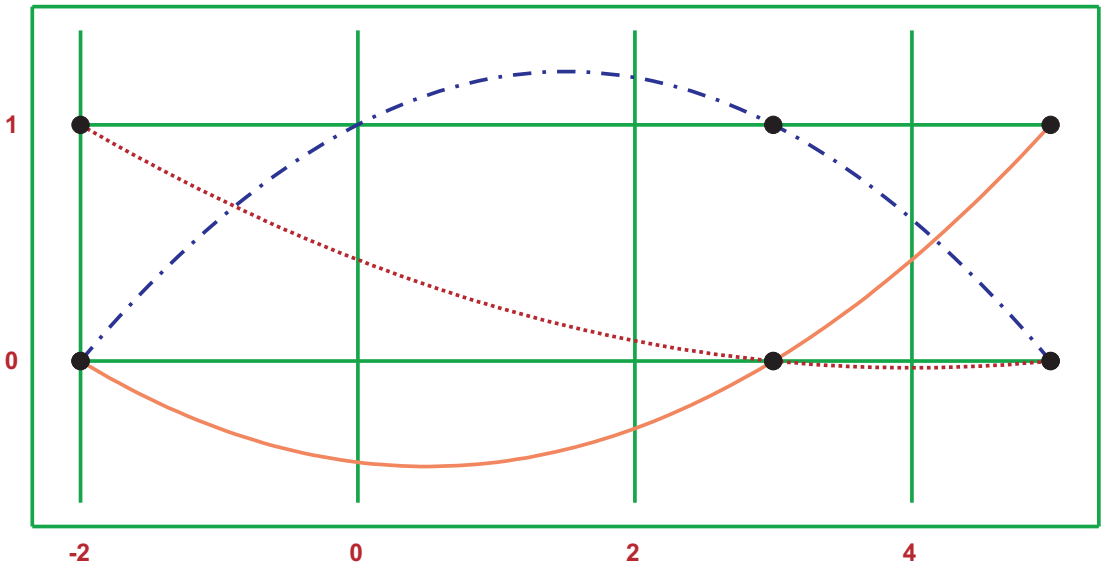

Figura 7.2: Funciones cardinales para las abscisas $x_1 = -2$, $x_2 = 3$ y $x_3 = 5$ del ejemplo 41.

Para probar que el polinomio interpolante pasa a través de los puntos dados, se debe sustituir $x = x_j$ en la ecuación (7.1) y utilizar la ecuación (7.2). El resultado es:

$$P_{n-1}(x_j) = \sum_{i=1}^{n} y_i \ell_i(x_j) = \sum_{i=1}^{n} y_i \delta_{ij} = y_j.$$

Es posible demostrar que el error en el polinomio de interpolación es:

$$f(x) - P_{n-1}(x) = \frac{(x - x_1)(x - x_2)\ldots(x - x_n)}{n!} f^{(n)}(\xi), \tag{7.5}$$

donde $\xi \in (x_1; x_n)$, su valor en cualquier otro punto es desconocido. Es importante notar que cuanto más lejos de x se toma un valor, mayor será su error.

Ejemplo 42. *Se desea interpolar la función* $f(x) = \cos(x) + 2x$ *que está definida en el intervalo* $[-1, 11]$*, pero sólo en el conjunto de datos* $x_1 = 0{,}1$*;* $x_2 = 2{,}3$*;* $x_3 = 4{,}1$*;* $x_4 = 5{,}9$ *y* $x_5 = 9{,}2$*. Para ello se construyen las funciones cardinales, resultando de la*

siguiente manera:

$$\begin{aligned}
\ell_1(x) &= 0{,}0021530x^4 - 0{,}046289x^3 + 0{,}34523x^2 - 1{,}0545x + 1{,}1020\\
\ell_2(x) &= -0{,}010166x^4 + 0{,}19620x^3 - 1{,}2007x^2 + 2{,}3805x - 0{,}22624\\
\ell_3(x) &= 0{,}015129x^4 - 0{,}26476x^3 + 1{,}3730x^2 - 2{,}0234x + 0{,}18888\\
\ell_4(x) &= -0{,}0080627x^4 + 0{,}12658x^3 - 0{,}56334x^2 + 0{,}75456x - 0{,}069949\\
\ell_5(x) &= 0{,}00094629x^4 - 0{,}011734x^3 + 0{,}045819x^2 - 0{,}057114x + 0{,}0052648
\end{aligned}$$

Ahora, se construirá el polinomio de Lagrange, de acuerdo a (7.1):

$$\begin{aligned}
P_4(x) &= 1{,}1950\ell_1(x) + 3{,}9337\ell_2(x) + 7{,}6252\ell_3(x) + 12{,}727\ell_4(x) + 17{,}425\ell_5(x)\\
&= -0{,}0081816x^4 + 0{,}10430x^3 - 0{,}21287x^2 + 1{,}2830x + 1{,}0690.
\end{aligned}$$

En la figura 7.3 se muestra: la función a interpolar, en línea punteada de color azul; el polinomio interpolante, en línea continua de color rojo y los datos para la interpolación, círculos negros. Debe notarse que el polinomio interpolante no siempre posee un error pequeño.

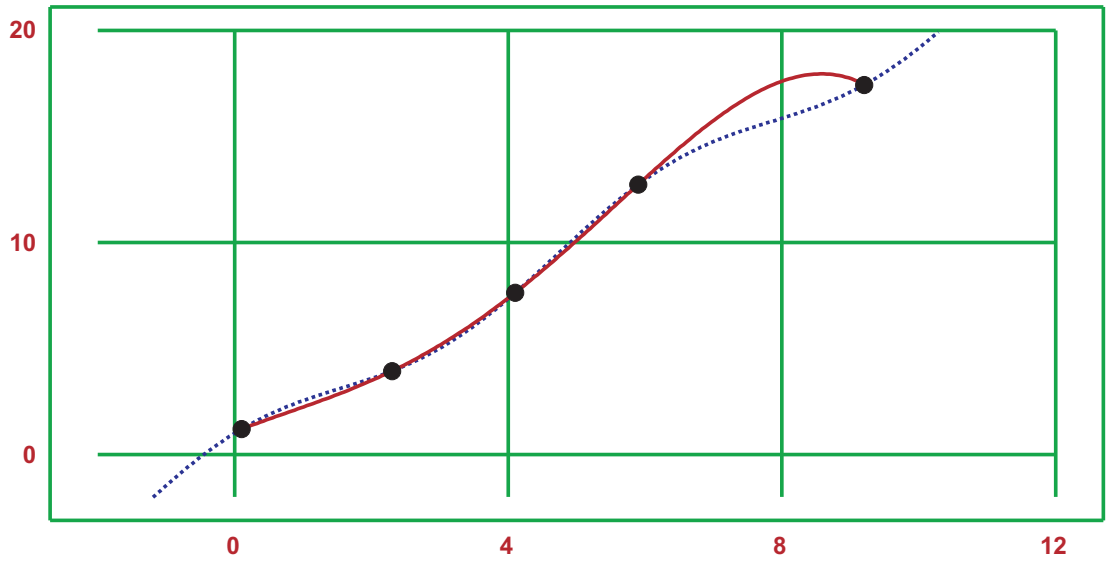

Figura 7.3: Polinomio interpolante de $f(x) = \cos(x) + 2x$ del ejemplo 42.

Ejercicio 20. *Calcular el error máximo cometido en la interpolación del ejemplo anterior, primero de acuerdo a la fórmula de error dada y luego con los datos reales del enunciado.*

7.1.3. Interpolación de Newton

A pesar de que el método de Lagrange es conceptualmente simple, no ofrece un algoritmo eficiente, puesto que si se desean agregar más puntos (o quitar algunos), debe hacerse todo el procedimiento de nuevo. Una mejora computacional se plantea en el método de Newton, donde el polinomio interpolante es escrito de la siguiente forma:

$$P_{n-1}(x) = a_1 + (x - x_1)a_2 + (x - x_1)(x - x_2)a_3 + \ldots + (x - x_1)(x - x_2)\cdots(x - x_n)\,a_n + 1. \tag{7.6}$$

El polinomio escrito en la forma (7.6) brinda directamente un procedimiento eficiente desde el punto de vista computacional. Por ejemplo, si se utilizan cuatro puntos, el

polinomio interpolante es:

$$\begin{aligned} P_3(x) &= a_1 + (x - x_1)a_2 + (x - x_1)(x - x_2)a_3 + (x - x_1)(x - x_2)(x - x_3)a_4 \\ &= a_1 + (x - x_1)\left(a_2 + (x - x_2)\left(a_3 + (x - x_3)a_4\right)\right), \end{aligned}$$

que puede ser evaluado iterando hacia atrás con la siguiente relación de recurrencia:

$$\begin{aligned} P_0(x) &= a_4 \\ P_1(x) &= a_3 + (x - x_3)P_0(x) \\ P_2(x) &= a_2 + (x - x_2)P_1(x) \\ P_3(x) &= a_1 + (x - x_1)P_2(x). \end{aligned}$$

Para un n arbitrario se tiene:

$$P_0 = a_n, \qquad P_k(x) = a_{n-k} + (x - x_{n-k})P_{k-1}(x), \qquad k = 1, 2, \ldots, n-1. \tag{7.7}$$

Cómputo manual de coeficientes

Los coeficientes de $P_{n-1}(x)$ se determinan forzando al polinomio a pasar a través de cada punto de los datos: $y_i = P_{n-1}(x_i)$, $i = 1, 2, \ldots, n$. Esto lleva a las ecuaciones simultáneas:

$$\begin{aligned} y_1 &= a_1 \\ y_2 &= a_1 + (x_2 - x_1)a_2 \\ y_3 &= a_1 + (x_3 - x_1)a_2 + (x_3 - x_1)(x_3 - x_2)a_3 \\ &\vdots \\ y_n &= a_1 + (x_n - x_1)a_2 + \ldots + (x_n - x_1)(x_n - x_2)\cdots(x_n - x_{n-1})a_n. \end{aligned} \tag{7.8}$$

Introduciendo las **diferencias divididas**:

$$\begin{aligned} \nabla y_i &= \frac{y_i - y_1}{x_i - x_1}, \qquad i = 2, 3, \ldots, n \\ \nabla^2 y_i &= \frac{\nabla y_i - \nabla y_2}{x_i - x_2}, \qquad i = 3, 4, \ldots, n \\ \nabla^3 y_i &= \frac{\nabla^2 y_i - \nabla^2 y_3}{x_i - x_3}, \qquad i = 4, 5, \ldots, n \\ &\vdots \\ \nabla^n y_n &= \frac{\nabla^{n-1} y_n - \nabla^{n-1} y_{n-1}}{x_n - x_{n-1}}, \end{aligned} \tag{7.9}$$

entonces la solución del conjunto de ecuaciones (7.8) es:

$$a_1 = y_1, \qquad a_2 = \nabla y_2, \qquad a_3 = \nabla^2 y_3 \qquad \cdots \qquad a_n = \nabla^{n-1} y_n. \tag{7.10}$$

Si los coeficientes son calculados en forma manual, es conveniente trabajar con el formato de la tabla 7.1, donde se muestra un ejemplo esquemático para $n = 5$.

Los valores ubicados sobre la diagonal de la tabla son los coeficientes del polinomio de Newton. Si los datos son copiados en diferente orden, los valores de la tabla cambiarán, pero el polinomio resultante será el mismo. Recordar que el polinomio de grado $n - 1$ que interpola n puntos distintos es único.

Ejercicio 21. *Realizar la tabla de diferencias divididas para los datos del ejemplo 42.*

x_1	y_1				
x_2	y_2	∇y_2			
x_3	y_3	∇y_3	$\nabla^2 y_3$		
x_4	y_4	∇y_4	$\nabla^2 y_4$	$\nabla^3 y_4$	
x_5	y_5	∇y_5	$\nabla^2 y_5$	$\nabla^3 y_5$	$\nabla^4 y_5$

Tabla 7.1: Esquema de coeficientes del polinomio de Newton, $n = 5$.

Comandos de EMT. *Los comandos para interpolar datos son:*

- `interp(x:vector numérico, y:vector numérico)`, *donde* $\boldsymbol{x}$ *e* $\boldsymbol{y}$ *son los datos para interpolar. La salida es un vector con los coeficientes finales de la tabla de diferencias divididas.*
- `polytrans(x:vector numérico, d:vector numérico)`, *donde* $\boldsymbol{x}$ *son las abscisas para interpolar;* $\boldsymbol{d}$ *es el vector con los coeficientes finales de la tabla de diferencias divididas. La salida es un vector con los coeficientes del polinomio interpolador, ordenados de la menor a la mayor potencia.*

Ejemplo en EMT 15. *Calcular los coeficientes del polinomio que interpola los datos:*

$$x = [1; 1{,}35; 1{,}7; 2{,}05; 2{,}4; 2{,}75]^T$$
$$y = [1{,}40887; 1{,}70009; 2{,}04741; 2{,}21126; 2{,}98832; 3{,}58884]^T$$

```
>x=[1,1.35,1.7,2.05,2.4,2.75];
>y=[1.40887,1.70009,2.04741,2.21126,2.98832,3.58884];
>DD=interp(x,y)
 [1.40887,  0.832057,  0.22898,  -0.931273,  2.87727,  -4.16125]
>P=polytrans(x,DD)
 [63.5464,  -194.742,  234.712,  -136.194,  38.2479,  -4.16125]
```


7.1.4. Interpolación de Neville

El método de interpolación polinómica involucra dos pasos: cálculo de los coeficientes y luego evaluación del polinomio en el punto a interpolar. Este procedimiento es útil cuando se necesita interpolar varios valores y se utilizará siempre el mismo polinomio. Si sólo un punto debe ser interpolado, es mejor utilizar un método que calcule directamente esto, en vez de generar un polinomio. Este es el atractivo del método de Neville.

Sea $P_k[x_i, x_{i+1}, \ldots, x_{i+k}]$ el polinomio de grado k que pasa a través de $k+1$ puntos $(x_i, y_i), (x_{i+1}, y_{i+1}), \ldots, (x_{i+k}, y_{i+k})$. Para un solo punto, se tiene que:

$$P_0[x_i] = y_i.$$

El interpolante basado en dos puntos es:

$$P_1[x_i, x_{i+1}] = \frac{(x - x_{i+1})P_0[x_i] + (x_i - x)P_0[x_{i+1}]}{x_i - x_{i+1}}.$$

Es fácil verificar que $P_1[x_i, x_{i+1}]$ pasa a través de los puntos dados. Es decir $P_1[x_i, x_{i+1}] = y_i$ cuando $x = x_i$, y $P_1[x_i, x_{i+1}] = y_{i+1}$ cuando $x = x_{i+1}$. El interpolante que opera a través de tres puntos es:

$$P_2 = [x_i, x_{i+1}, x_{i+2}] = \frac{(x - x_{i+2})P_1[x_i, x_{i+1}] + (x_i - x)P_1[x_{i+1}, x_{i+2}]}{x_i - x_{i+2}}.$$

Para mostrar que este interpolante pasa por los puntos solicitados, primero se debe sustituir $x = x_i$, con lo que:

$$P_2[x_i, x_{i+1}, x_{i+2}] = P_1[x_i, x_{i+1}] = y_i,$$

luego, $x = x_{i+2}$:

$$P_2[x_i, x_{i+1}, x_{i+2}] = P_1[x_{i+1}, x_{i+2}] = y_{i+2},$$

finalmente, $x = x_{i+1}$:

$$P_1[x_i, x_{i+1}] = P_1[x_{i+1}, x_{i+2}] = y_{i+1}.$$

Entonces el polinomio interpola correctamente los tres puntos dados.

Considerando el desarrollo anterior, es posible construir una fórmula general recursiva:

$$P_k[x_i,x_{i+1}, \ldots, x_{i+k}] = \frac{(x - x_{i+k})P_{k-1}[x_i, x_{i+1}, \ldots, x_{i+k-1}] + (x_i - x)P_{k-1}[x_{i+1}, x_{i+1}, \ldots, x_{i+k}]}{x_i - x_{i+k}}. \tag{7.11}$$

Dados los datos $x_i, i = 1, 2, 3, 4$, los cálculos pueden ser tabulados de acuerdo al esquema presentado en la tabla 7.2.

	$k=0$	$k=1$	$k=2$	$k=3$
x_1	$P_0[x_1] = y_1$	$P_1[x_1, x_2]$	$P_2[x_1, x_2, x_3]$	$P_3[x_1, x_2, x_3, x_4]$
x_2	$P_0[x_2] = y_2$	$P_1[x_2, x_3]$	$P_2[x_2, x_3, x_4]$	
x_3	$P_0[x_3] = y_3$	$P_1[x_3, x_4]$		
x_4	$P_0[x_4] = y_4$			

Tabla 7.2: Esquema de coeficientes para la interpolación de Neville, $n = 4$.

Ejemplo 43. *Se desea interpolar el valor de $f(x)$ cuando $x = 5$, utilizando el conjunto de datos $x_1 = 0{,}1$; $x_2 = 2{,}3$; $x_3 = 4{,}1$; $x_4 = 5{,}9$ y $x_5 = 9{,}2$. Para ello se construye la tabla 7.3, que muestra los coeficientes de Neville.*

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
0,1	1,1950	7,2948	9,9605	9,9936	10,080
2,3	3,9337	9,4709	9,9997	10,155	
4,1	7,6252	10,176	10,399		
5,9	12,727	11,445			
9,2	17,425				

Tabla 7.3: Esquema de coeficientes para la interpolación de Neville, $n = 4$, con los datos del ejemplo 43.

Ejercicio 22. *Comparar el resultado de la interpolación puntual del ejercicio anterior con la que se obtiene al evaluar alguno de los polinomios interpoladores en el punto $x = 5$.*

7.1.5. Limitaciones de la interpolación polinómica

La interpolación polinomial debe realizarse con la menor cantidad de puntos posibles. El caso lineal, utilizando los dos puntos más cercanos, es en la mayoría de las veces suficientemente buena si los datos son cercanos. De tres a seis puntos cercanos producen resultados en la mayoría de los casos. Si un polinomio interpolante tiene bajo error y se utilizaron más de seis puntos, debe revisarse con cuidado. Esto es debido a que los puntos más alejados de aquel que se debe interpolar, no aportan demasiada información.

El problema de utilizar demasiados puntos se ilustra en la figura 7.4. Se visualizan 13 puntos y el polinomio interpolador. En ambos extremos muestra oscilaciones de gran magnitud, que no se corresponden con los datos iniciales, además de pequeñas oscilaciones en la parte central del gráfico.

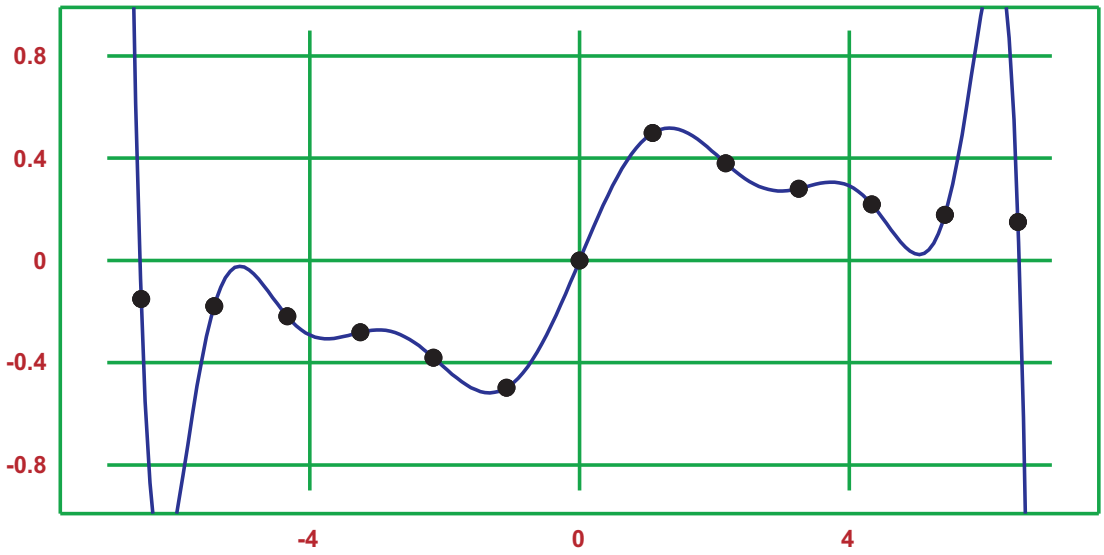

Figura 7.4: Datos y polinomio interpolador de grado 12. Grandes oscilaciones en los extremos.

La extrapolación polinómica es peligrosa. Por ejemplo, la figura 7.5 muestra 6 datos y el polinomio que los interpola. Los datos fueron extraídos de una tabla de *tiempo vs porcentaje de pacientes infectados* dentro de las pruebas de un tratamiento médico. Si bien la interpolación relaciona correctamente los datos tabulados, es imposible tener una cantidad negativa de pacientes infectados. Si no puede evitarse la extrapolación, deben tenerse en cuenta los siguientes ítems:

- Graficar los datos y verificar visualmente que los valores extrapolados tienen sentido.
- Utilizar un polinomio interpolador de bajo orden cerca de los datos a extrapolar.
- Realizar los gráficos con escalas $\log(x)$ vs $\log(y)$, lo que usualmente brinda una curva mucho más suave que las curvas x vs y y puede servir para extrapolar (tomando las precauciones antes mencionadas).

7.2. Interpolación segmentaria

Hasta ahora, el foco de la discusión ha sido la cuestión de aproximar una función f (o datos extraídos de de ella), definida en el intervalo $[a, b]$ a través de un polinomio utilizando la interpolación de Lagrange o Newton. Dicha interpolación era de

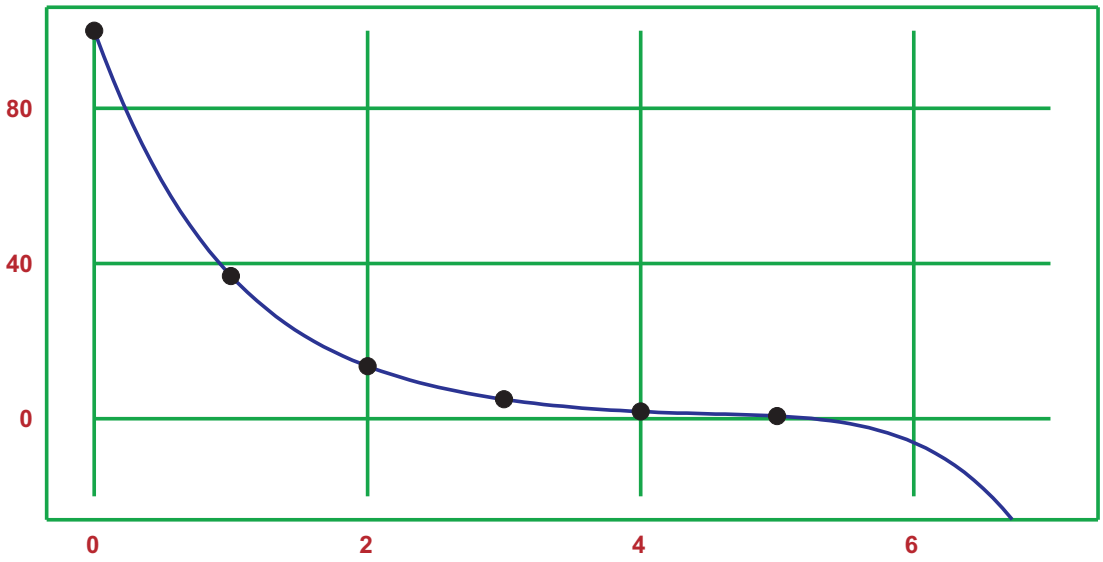

Figura 7.5: Datos y polinomio interpolador de grado 5. Problemas de coherencia para extrapolar.

característica global, es decir que se generó utilizando la misma definición analítica para todo el intervalo. Una alternativa más flexible, que evita oscilaciones y un grado polinómico alto, es la aproximación de la función f (o datos extraídos de ella), es la división del intervalo $[a, b]$ en un número de subintervalos y utilizar en cada uno de ellos una interpolación polinómica de grado bajo. Este tipo de aproximaciones polinomiales a tramos se denominan ***splines*** y los puntos que se utilizan se denominan **nodos**.

Más específicamente, un *spline* de grado n, $n \geq 1$, es una función polinómica de grado n o menor en cada subintervalo, que tiene un grado prestablecido de suavidad. Se espera además de los *splines* que sean como mínimo continuos, aunque se les pide derivadas continuas hasta el orden k, para cierto $0 \leq k < n$. Claramente, si se requiere que la derivada de orden n sea continua en todo punto, el *spline* será un único polinomio. Es imposible que dos polinomios diferentes tengan el mismo valor en cada nodo y además sus derivadas coincidan en los nodos. Se desarrollarán dos casos de *splines*: lineales y cúbicos.

7.2.1. *Splines* lineales

Definición 18. *Sea f una función real, definida y continua en el intervalo $[a, b]$. Además, sea $K = \{x_0, x_1, \ldots, x_m\}$ un subconjunto de $[a, b]$, eligiendo $a = x_0 < x_1 < \cdots < x_m = b$, $m \geq 2$. El* ***spline lineal*** *S_L, que interpola f en los puntos x_i se define como:*

$$S_L(x) = \frac{x_i - x}{x_i - x_{i-1}} f(x_{i-1}) + \frac{x - x_{i-1}}{x_i - x_{i-1}} f(x_i), \quad x \in [x_{i-1}, x_i], \quad i = 1, 2, \ldots, m. \tag{7.12}$$

Los puntos $x_i, i = 0, 1, \ldots, m$, son los ***nodos*** *del spline y a K se lo conoce como el* ***conjunto de nodos****.*

Como la función S_L interpola la función f en los nodos, entonces $S_L(x_i) = f(x_i)$, para $i = 0, 1, \ldots, m$, y sobre cada intervalo $[x_{i-1}, x_i]$, para $i = 0, 1, \ldots, m$, la función S_L es un polinomio lineal (y por lo tanto, continuo). Entonces S_L es una función lineal a tramos sobre el intervalo $[a, b]$.

Dada una serie de nodos $K = \{x_0, \ldots, x_m\}$, se utilizará la notación $h_i = x_i - x_{i-1}$, y $h = \text{máx}_i\{h_i\}$. Para destacar la precisión de la interpolación por *splines* lineales, se analizará el error con la norma infinito sobre el intervalo de definición $[a, b]$.

Teorema 20. *Sea $f \in C^2[a,b]$ y sea S_L el spline lineal que interpola f en los nodos $a = x_0 < x_1 < \ldots < x_m = b$, entonces la interpolación fue realizada con la siguiente cota de error:*

$$\|f - S_L\|_\infty \leq \frac{1}{8} h^2 \|f''\|_\infty,$$

donde $h = \max_i\{h_i\} = \max_i\{x_i - x_{i-1}\}$.

Ejemplo 44. *Sea $f(x) = e^{-3x}$, definida en $[0,1]$. Se quiere interpolar la función utilizando cuatro puntos equiespaciados, entonces:*

$$S_L(x) = \begin{cases} 1 - 1{,}896x & si \quad 0 \leq x \leq 0{,}3333 \\ 0{,}6004 - 0{,}6977x & si \quad 0{,}3333 \leq x \leq 0{,}6667 \\ 0{,}3064 - 0{,}2566x & si \quad 0{,}6667 \leq x \leq 1 \end{cases}$$

En la figura 7.6 se muestra el gráfico de los nodos, en línea continua de color rojo la función S_L y en línea punteada de color azul la función f.

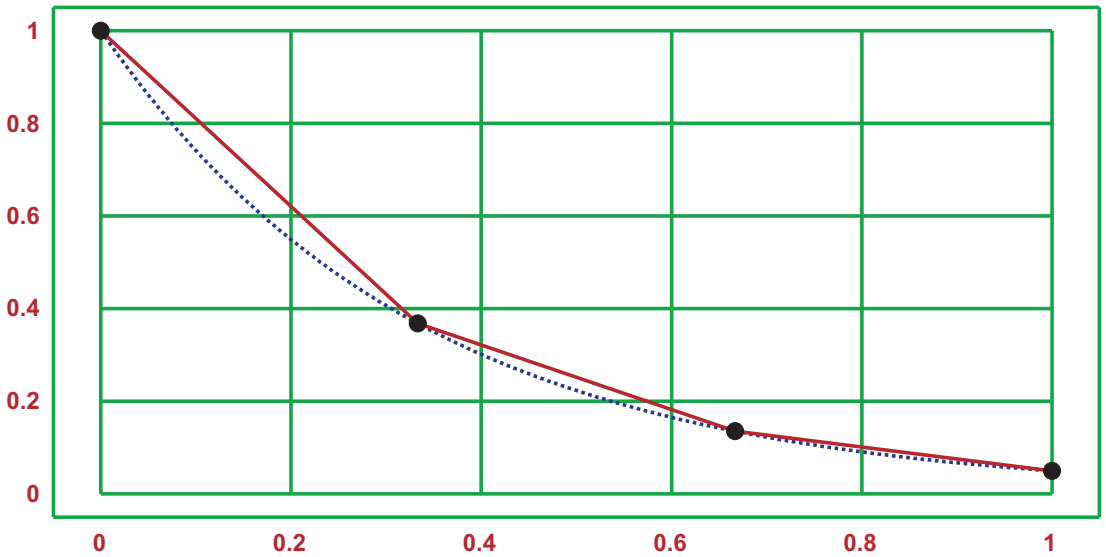

Figura 7.6: *Spline* lineal que interpola a $f(x)$ del ejemplo 44.

Ejercicio 23. *Calcular el error asociado a la interpolación de spline lineal con respecto a la función del ejemplo anterior.*

7.2.2. *Splines* cúbicos

Sea $f \in C[a,b]$ y sea $K = \{x_0, x_1, \ldots, x_m\}$ un conjunto de $m+1$ nodos en el intervalo $[a,b]$, $a = x_0 < x_1 < \cdots < x_m = b$. Se considera entonces el conjunto $\mathcal{S}$ de todas las funciones $s \in C^2[a,b]$ tal que

- $s(x_i) = f(x_i), \quad i = 0, 1, \ldots, m.$
- s es un polinomio cúbico sobre $[x_{i-1}, x_i], \quad i = 1, 2, \ldots, m.$

Cualquier elemento de $\mathcal{S}$ será referido como un ***spline* interpolador cúbico**. La principal diferencia con los *splines* interpoladores lineales es que existe más de un *spline* $s \in C^2[a,b]$ que satisface las dos condiciones mencionadas anteriormente. Es más, existen $4m$ coeficientes de polinomios cúbicos (cuatro en cada subintervalo $[x_{i-1}, x_i]$,

$i = 1, 2, \ldots, m$), y sólo $m + 1$ condiciones de interpolación y $3(m - 1)$ condiciones de continuidad. Como s pertenece a $C^2[a, b]$ esto significa que s, s' y s'' son continuos en los nodos internos $x_1, x_2, \ldots, x_{m-1}$. De aquí que hay $4m - 2$ condiciones para los $4m$ coeficientes desconocidos. Dependiendo de la elección de las dos condiciones remanentes, se pueden construir varios *splines* interpoladores cúbicos.

Definición 19. *El **spline natural cúbico**, denotado por s_2, es el elemento del conjunto $\mathcal{S}$ que satisface las dos condiciones terminales:*

$$s_2''(x_0) = 0 = s_2''(x_m).$$

Construcción del *spline* natural cúbico

Aplicando la definición $\sigma_i = s_2''(x_i)$, $i = 0, 1, \ldots, m$, y teniendo en cuenta que s_2'' es una función lineal en cada subintervalo $[x_{i-1}, x_i]$, se puede expresar s_2'' como:

$$s_2''(x) = \frac{x_i - x}{h_i}\sigma_{i-1} + \frac{x - x_{i-1}}{h_i}\sigma_i, \qquad x \in [x_{i-1}, x_i].$$

Integrando dos veces con respecto a x, se obtiene:

$$s_2(x) = \frac{(x_i - x)^3}{6h_i}\sigma_{i-1} + \frac{(x - x_{i-1})^3}{6h_i}\sigma_i + \alpha_i(x - x_{i-1}) + \beta_i(x_i - x), \qquad x \in [x_{i-1}, x_i], \tag{7.13}$$

donde α_i y β_i son constantes de integración. Igualando s_2 a f en los nodos x_{i-1} queda:

$$f(x_{i-1}) = \frac{1}{6}\sigma_{i-1}h_i^2 + h_i\beta_i, \qquad f(x_i) = \frac{1}{6}\sigma_i h_i^2 + h_i\alpha_i. \tag{7.14}$$

Expresando α_i y β_i a partir de la expresión anterior, insertándolas en (7.13) y aprovechando la continuidad de s_2' en los nodos internos (es decir que $s_2'(x_i-) = s_2'(x_i+)$, para $i = 1, 2, \ldots, m - 1$) se tiene que:

$$h_i\sigma_{i-1} + 2\left(h_{i+1} + h_i\right)\sigma_i + h_{i+1}\sigma_{i+1} = 6\left(\frac{f(x_{i+1}) - f(x_i)}{h_{i+1}} - \frac{f(x_i) - f(x_{i-1})}{h_i}\right) \tag{7.15}$$

para $i = 1, 2, \ldots, m - 1$, lo que en conjunto con:

$$\sigma_0 = \sigma_m = 0,$$

constituye un sistema de ecuaciones lineales para σ_i. La matriz del sistema es tridiagonal y nosingular. Al resolver este sistema lineal se obtienen los coeficientes σ_i, $i = 0, 1, \ldots, m$ y por lo tanto todos los coeficientes α_i, β_i, $i = 1, 2, \ldots, m - 1$ planteados en (7.14).

Si es necesario identificar una cota para el error de interpolación, se utiliza el siguiente resultado:

Teorema 21. *Sea $f \in C^4[a, b]$, y sea s el spline cúbico que interpola f a través de los nodos definidos en K. Entonces, una cota para el error que se comete en la interpolación es:*

$$\|f - s\|_\infty \leq \frac{1}{24}h^4\|f^{iv}\|_\infty,$$

donde f^{iv} es la derivada cuarta de f con respecto a x, $h = \max_i\{h_i\} = \max_i\{x_i - x_{i-1}\}$ y $\|\cdot\|$ denota la norma infinito en el intervalo $[a, b]$.

Ejemplo 45. *Sea* $f(x) = e^{-3x}$, *definida en* $[0, 1]$. *Se quiere interpolar la función utilizando cuatro puntos equiespaciados, entonces a partir de (7.15) se plantea el sistema de ecuaciones:*

$$\begin{aligned} h_1\sigma_0 + 2\left(h_2 + h_1\right)\sigma_1 + h_2\sigma_2 &= 6\left(\frac{f(x_2)-f(x_1)}{h_2} - \frac{f(x_1)-f(x_0)}{h_1}\right) \\ h_2\sigma_1 + 2\left(h_3 + h_2\right)\sigma_2 + h_3\sigma_3 &= 6\left(\frac{f(x_3)-f(x_2)}{h_3} - \frac{f(x_2)-f(x_1)}{h_2}\right) \end{aligned}$$

que reemplazando los valores de las constantes con los datos de la tabla 7.4 lleva al sistema de ecuaciones:

$$\begin{aligned} 0{,}3333\sigma_0 + 1{,}333\sigma_1 + 0{,}3334\sigma_2 &= 7{,}188 \\ 0{,}3334\sigma_1 + 1{,}333\sigma_2 + 0{,}3333\sigma_3 &= 2{,}646 \end{aligned}$$

y al ser resuelto utilizando el método de Gauss con pivoteo parcial, da como resultado $\sigma_1 = 5{,}223$ *y* $\sigma_2 = 0{,}6788$. *Recordar que* σ_0 *y* σ_3 *toman el valor cero por definición. Calculando* α_i *y* β_i *para* $i = 1, 2, 3$ *a partir de (7.14) se obtiene la expresión del spline*

i	x_i	$f(x_i)$	h_i
0	0	1	-
1	0,3333	0,3679	0,3333
2	0,6667	0,1353	0,3334
3	1	0,04979	0,3334

Tabla 7.4: Datos para la interpolación por *splines* cúbicos de $f(x)$, del ejemplo 45.

cúbico que interpola a $f(x)$*:*

$$s_2 = \begin{cases} s_{21}(x) & si \quad 0 \le x \le 0{,}3333 \\ s_{22}(x) & si \quad 0{,}3333 \le x \le 0{,}6667 \\ s_{23}(x) & si \quad 0{,}6667 \le x \le 1 \end{cases}$$

donde:

$$\begin{aligned} s_{21}(x) &= 0\left(x_1 - x\right)^3 + 2{,}612\left(x - x_0\right)^3 + 0{,}8137\left(x - x_0\right) + 3\left(x_1 - x\right), \\ s_{22}(x) &= 2{,}611\left(x_2 - x\right)^3 + 0{,}3393\left(x - x_1\right)^3 + 0{,}3681\left(x - x_1\right) + 0{,}8133\left(x_2 - x\right), \\ s_{23}(x) &= 0{,}3394\left(x_3 - x\right)^3 + 0\left(x - x_2\right)^3 + 0{,}1494\left(x - x_2\right) + 0{,}3682\left(x_3 - x\right). \end{aligned}$$

En la figura 7.7 es posible ver los nodos de interpolación (círculos de color negro), la función a interpolar (línea punteada azul) y el spline cúbico interpolante (línea continua de color rojo).

Nota. *Tener en cuenta la forma cuasi simétrica de los coeficientes cúbicos y lineales del spline antes calculados. Es una buena forma de verificar la exactitud de los coeficientes obtenidos.*

Ejercicio 24. *Verificar la continuidad de las derivadas primeras y segundas del spline utilizado en el ejercicio anterior.*

Comandos de EMT. *Los comandos para generar* splines *cúbicos es:*

- `spline(x:vector numérico, y:vector numérico)`, *donde* ***x*** *e* ***y*** *son los datos a interpolar. La salida es un vector con los valores de la derivada segunda en los puntos dados. El* spline *generado es el* spline *natural.*

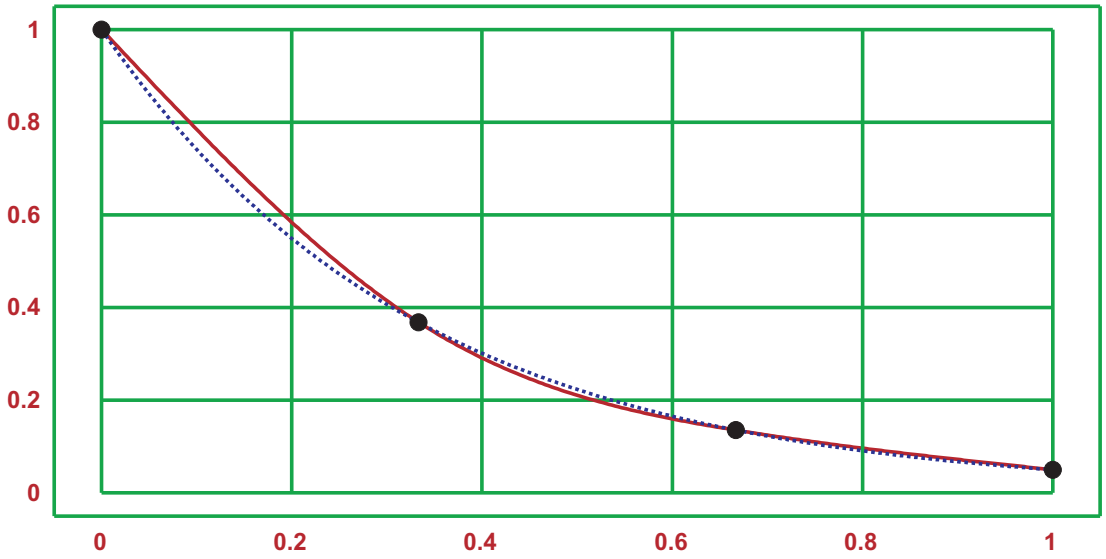

Figura 7.7: *Spline* cúbico que interpola a $f(x)$ del ejemplo 45.

- `splineval(t:número, x:vector numérico, y:vector numérico, h:vector numérico)`, *donde* $\boldsymbol{t}$ *es el valor de abscisa donde se calculará el* spline*;* $\boldsymbol{x}$ *e* $\boldsymbol{y}$ *son los datos a interpolar,* $\boldsymbol{h}$ *es el vector que contiene las derivadas segundas en los puntos dados. La salida es el valor de ordenada del* spline *en la abscisa dada.*

Ejemplo en EMT 16. *Generar y graficar el* spline *cúbico que interpola los datos:*

$$x = [1; 1{,}35; 1{,}7; 2{,}05; 2{,}4; 2{,}75]^T$$
$$y = [1{,}40887; 1{,}70009; -2{,}04741; 2{,}21126; 2{,}98832; -1{,}58884]^T$$

```
>xp=[1,1.35,1.7,2.05,2.4,2.75];
>yp=[1.40887,1.70009,-2.04741,2.21126,2.98832,-1.58884];
>s=spline(xp,yp)
 [0,  -83.1559,  134.809,  -63.9399,  -49.5769,  0]
>plot2d("splineval(x,xp,yp,s)",xmin=1,xmax=2.75,color=blue);
>plot2d(xp,yp,points=1,add=1,style="o#");
```

7.2.3. *Splines* de Hermite

En la subsección previa se consideró $f \in C^2[a,b]$ y se exigió que s pertenezca a $C^2[a,b]$. Ahora, el foco se centrará en la suavidad y exactitud del *spline* interpolador con respecto a la función a interpolar, exigiendo sólo que $s \in C^1[a,b]$.

Sea $K = x_0, x_1, \ldots, x_m$ un conjunto de nodos en el intervalo $[a,b]$ con $a = x_0 < x_1 < \ldots < x_m = b$ y $m \geq 2$. Se define el ***spline* cúbico de Hermite** como una función $s \in C^1[a,b]$ tal que

- $s(x_i) = f(x_i), \quad s'(x_i) = f'(x_i), \quad i = 0, 1, \ldots, m.$
- s es un polinomio cúbico sobre $[x_{i-1}, x_i], \quad i = 1, 2, \ldots, m.$

Escribiendo el polinomio cúbico s dentro del intervalo $[x_{i-1}, x_i]$ como:

$$s(x) = c_0 + c_1(x - x_{i-1}) + c_2(x - x_{i-1})^2 + c_3(x - x_{i-1})^3, \quad x \in [x_{i-1}, x_i], \qquad (7.16)$$

donde:

$$c_0 = f(x_{i-1}), \qquad c_1 = f'(x_{i-1}),$$

$$c_2 = 3\frac{f(x_i) - f(x_{i-1})}{h_i^2} - \frac{f'(x_i) + 2f'(x_{i-1})}{h_i},$$
$$c_3 = \frac{f'(x_i) + f'(x_{i-1})}{h_i^2} - 2\frac{f(x_i) - f(x_{i-1})}{h_i^3}.$$

Se debe tener en cuenta que el *spline* de Hermite tiene sólo derivada continua en los nodos y por lo tanto no es un *spline* interpolador cúbico en el sentido de la definición original.

A diferencia de los *splines* cúbicos, los coeficientes de un *spline* de Hermite en cada subintervalo se pueden calcular explícitamente sin necesidad de resolver un sistema de ecuaciones.

A fin de realizar un análisis (mínimo) del error de interpolación, se utiliza el siguiente resultado.

Teorema 22. *Sea $f \in C^4[a,b]$, y sea s el spline de Hermite cúbico que interpola f a través de los nodos definidos en K. Entonces, una cota para el error que se comete en la interpolación es:*

$$\|f - s\|_\infty \leq \frac{1}{384} h^4 \|f^{iv}\|_\infty,$$

donde f^{iv} es la derivada cuarta de f con respecto a x, $h = \max_i\{h_i\} = \max_i\{x_i - x_{i-1}\}$ y $\|\cdot\|$ denota la norma infinito en el intervalo $[a,b]$.

Ejemplo 46. *Sea $f(x) = e^{-3x}$, definida en $[0,1]$. Se quiere interpolar la función utilizando cuatro puntos equiespaciados, también es necesario conocer el valor de la derivada primera en los nodos a utilizar. Estos datos se resumen en la tabla 7.5*

i	x_i	$f(x_i)$	$f'(x_i)$	h_i
0	0	1	-3	-
1	0,3333	0,3679	-1,104	0,3333
2	0,6667	0,1353	-0,4060	0,3334
3	1	0,04979	-0,1494	0,3333

Tabla 7.5: Datos para la interpolación por *splines* de Hermite de $f(x)$ en el ejemplo 46.

Entonces, los coeficientes se muestran en la tabla 7.6. y el spline de Hermite es:

$$s = \begin{cases} s_1(x) & si \quad 0 \leq x \leq 0{,}3333 \\ s_2(x) & si \quad 0{,}3333 \leq x \leq 0{,}6667 \\ s_3(x) & si \quad 0{,}6667 \leq x \leq 1 \end{cases}$$

donde:

$s_1(x) = 1 - 3(x-0) + 4{,}244(x-0)^2 - 2{,}800(x-0)^3,$
$s_2(x) = 0{,}3679 - 1{,}104(x-0{,}3333) + 1{,}563(x-0{,}3333)^2 - 1{,}032(x-0{,}3333)^3,$
$s_3(x) = 0{,}1353 - 0{,}4060(x-0{,}6667) + 0{,}5753(x-0{,}6667)^2 - 0{,}3807(x-0{,}6667)^3.$

	c_0	c_1	c_2	c_3
s_1	1	-3	4,244	-2,800
s_2	0,3679	-1,104	1,563	-1,032
s_3	0,1353	-0,4060	0,5753	-0,3807

Tabla 7.6: Esquema de coeficientes para el *spline* de Hermite del ejemplo 46.

En la figura 7.8 se puede ver los nodos de interpolación (círculos de color negro), la función a interpolar (línea punteada de color azul) y el spline de Hermite (línea continua de color rojo).

Figura 7.8: *Spline* de Hermite que interpola a $f(x)$ del ejemplo 46.

Nota. *Tener en cuenta que, en forma general, los coeficientes c_i mantienen el mismo signo para cada coeficiente, por más que sean de dominios diferentes.*

Ejercicio 25. *Verificar la continuidad de las derivadas primeras y segundas del spline de Hermite utilizado en el ejercicio anterior.*

Comandos de EMT. *Los comandos para generar* splines *de Hermite son:*

- `hermiteinterp(x:vector numérico, y:vector numérica)`, *donde* $\boldsymbol{x}$ *es el vector de abscisas, con cada valor duplicado;* $\boldsymbol{y}$ *es un vector con los datos de las ordenadas y la derivada primera, en orden y alternados. La salida es un vector con los coeficientes finales de la tabla de diferencias divididas.*

- `interpval(x:vector numérico, d:vector numérico, t:número)`, *donde* $\boldsymbol{x}$ *es el vector de abscisas con los valores duplicados;* $\boldsymbol{d}$ *es el vector de coeficientes finales en la tabla de diferencias divididas;* $\boldsymbol{t}$ *es el valor de abscisa donde se calculará el* spline *de Hermite.*

Ejemplo en EMT 17. *Generar y graficar el* spline *de Hermite que interpole los valores de* $\arctan(x)$ *entre* -3 *y* 3.

```
>x=linspace(-3,3,6); y=arctan(x); dy=diff("arctan(x)",x);
>X=multdup(x,2); Y=zeros(1,2*length(y));
>for i=1 to length(y), Y[2*i-1]=y[i]; Y[2*i]=dy[i]; end
>d=hermiteinterp(X,Y)
 [-1.24905,  0.1,  0.0418971,  0.0162059,  0.00780896,  0.00226429,
 -0.00228808,  -0.00209971,  0.00162668,  -0.000348638,  -0.000118149,
 0.000150501,  -6.03254e-005,  2.01085e-005]
>plot2d("interpval(X,d,x)",-3,3,color=blue);
>plot2d(x,y,points=1,style="o#",add=1);
```

7.3. Ejercicios

1. Construir los siguientes algoritmos en PC:

   a) **Polinomio de Lagrange**. Entrada: matriz con los datos necesarios para interpolar. Salida: coeficientes del polinomio. Opcional: graficar los puntos y el polinomio interpolador.

   b) **Spline de Hermite**. Idénticas condiciones que las utilizadas para el polinomio de Lagrange.

2. La tabla 7.7 proporciona información sobre la cantidad de pasajeros que suben a un colectivo de acuerdo al horario matutino.

hora	8	9	10	11	12	13	14
pasajeros	41	35	21	9	11	17	32

Tabla 7.7

   a) Estimar la cantidad de pasajeros a las 10.30hs, por medio de interpolación lineal.

   b) Repertir el inciso anterior utilizando interpolación cuadrática y cúbica.

3. La solubilidad del cloro amónico en el agua toma, a distintas temperaturas (medidas en grados Celsius), los valores dados en la tabla 7.8.

   a) Elegir los 4 puntos centrales para definir un polinomio interpolador de grado 3.

   b) Con los mismos puntos que el inciso anterior, definir el *spline* de Hermite.

   c) Calcular con los resultados obtenidos en los incisos *a*) y *b*) la solubilidad para las temperaturas $25°C$, $35°C$ y $45°C$.

4. Dada la función $f(x) = \cos^{10}(x)$, definida en el intervalo $[-1; 1]$:

   a) Calcular el *spline* cúbico que interpola esta función, tomando tres puntos.

   b) Calcular el *spline* de Hermite, con los mismos puntos del inciso anterior.

   c) Graficar ambos polinomios y la función f. Estimar el máximo error cometido.

5. Elegir 5 puntos de la función $f(x) = \sin(x + 1) - 1$ dentro del intervalo $[0; 7]$. Construir:

   a) El polinomio interpolador de grado 4.

   b) El *spline* de Hermite.

6. Dado el conjunto de abscisas $\mathbf{x} = [1; 2; 3; 4; 5]$, calcular el vector $\mathbf{y}$ utilizando el polinomio $P(x) = -x^3 + 5x^2 - 2x + 1$, para luego:

temperatura	10	20	30	40	50	60
solubilidad	33	37	42	46	52	55

Tabla 7.8

x	1.08	1.13	1.20	1.27	1.31
$f(x)$	1.302	1.386	1.509	1.217	1.284

Tabla 7.9

x	-2	-1	0	1	2
$P(x)$	7	3	1	0	3

Tabla 7.10

a) Interpolar $(\mathbf{x};\mathbf{y})$ por medio de un polinomio de grado 4.

b) Crear el *spline* de Hermite que pasa por los puntos de $(\mathbf{x};\mathbf{y})$.

c) Estimar el error máximo cometido con ambas aproximaciones.

7. Utilizando la interpolación de Neville, estimar el valor de solubilidad para la temperatura $15°C$ del ejercicio 3, utilizando:

   a) Los tres datos más cercanos.

   b) Los cuatro datos más cercanos.

8. Por medio de la interpolación de Neville, calcular $P(3{,}5)$ del polinomio definido en el ejercicio 6.

9. Evaluar $f(1{,}4)$ a través de polinomios interpolantes de grados 1, 2 y 3 utilizando los valores de la tabla 7.9.

10. Interpolar la función $f(x) = |x|$ dentro del intervalo $[-3;5]$ utilizando 6 puntos no necesariamente equiespaciados. Graficar ambas funciones en el mismo sistema de ejes.

11. Interpolar la función $f(x) = x$ dentro del intervalo $[-3;5]$ utilizando 6 puntos no necesariamente equiespaciados, a través de un *spline*. Graficar ambas funciones en el mismo sistema de ejes.

12. $[EMT]$ Interpolar la función $f(x) = x$ dentro del intervalo $[-3;5]$ utilizando 6 puntos no necesariamente equiespaciados. Representar ambas funciones en el mismo sistema de ejes.

13. Si fuera necesario aproximar una función dentro de un intervalo para calcular su longitud, ¿qué es mejor? ¿Un *spline* de Hermite? ¿Un *spline* cúbico? ¿Por qué?

14. Si fuera necesario aproximar una función dentro de un intervalo para calcular extremos relativos, ¿qué es mejor? ¿Un *spline* de Hermite? ¿Un *spline* cúbico? ¿Por qué?

15. $[EMT]$ Es posible generar curvas paramétricas a través de interpoladores. Para ello se generan dos polinomios, dependientes de una única variable, que toman cada secuencia de puntos por separado. Generar un rectángulo con lados paralelos a los ejes coordenados a través de 9 puntos, donde el primero y el último punto deben coincidir. Reconstruirlo a través de dos splines cúbicos.

16. La tabla 7.10 muestra valores de un polinomio de segundo grado. Hay exactamente un error en la segunda fila, identificarlo.

x	-2	-1	0	1	2	3
y	1	4	11	16	13	-4

Tabla 7.11

17. Elegir los primeros tres nodos de la tabla 7.10 y:

    a) Generar el polinomio interpolador por definición.

    b) Rehacer nuevamente el polinomio, pero ahora realizar un desplazamiento del origen al punto medio de las abscisas de interpolación.

    c) Calcular el número de condición de las matrices del sistema de los incisos anteriores.

18. Los polinomios obtenidos por diferencias divididas se ven afectados por las simplificaciones aritméticas. Mostrar que, aplicando redondeo en una aritmética de 5 dígitos, el polinomio que interpola los puntos $(6000; \frac{1}{3})$ y $(6001; -\frac{2}{3})$ no pasa por los puntos dados si se lo escribe de la forma $P(x) = mx + b$.

19. ¿Cuál de los siguientes polinomios no es una función cardinal del mismo conjunto de datos? Una vez identificado el incorrecto, calcularlo nuevamente.

$$P_1(x) = \frac{1}{2240}x^5 - \frac{19}{2240}x^4 + \frac{19}{320}x^3 - \frac{421}{2240}x^2 + \frac{293}{1120}x - \frac{1}{8}$$
$$P_2(x) = -\frac{1}{360}x^5 + \frac{7}{120}x^4 - \frac{157}{360}x^3 + \frac{173}{120}x^2 - \frac{371}{180}x + 1$$
$$P_3(x) = \frac{1}{96}x^5 - \frac{23}{96}x^4 + \frac{63}{32}x^3 - \frac{673}{96}x^2 + \frac{505}{48}x - \frac{21}{4}$$
$$P_4(x) = -\frac{1}{90}x^5 + \frac{4}{15}x^4 - \frac{104}{45}x^3 + \frac{131}{15}x^2 - \frac{1231}{90}x + 6$$
$$P_5(x) = \frac{1}{210}x^5 - \frac{13}{105}x^4 + \frac{6}{5}x^3 - \frac{557}{105}x^2 + \frac{2147}{210}x - 6$$
$$P_6(x) = -\frac{1}{576}x^5 + \frac{3}{64}x^4 - \frac{277}{576}x^3 + \frac{149}{64}x^2 - \frac{1517}{288}x + \frac{35}{8}$$

20. Verificar que los polinomios:

$$p(x) = 5x^3 - 27x^2 + 45x - 21$$
$$q(x) = x^4 - 5x^3 + 8x^2 - 5x + 3$$

    interpolan los datos $\mathbf{x} = [1; 2; 3; 4]$, $\mathbf{y} = [2; 1; 6; 47]$. Explicar por qué no viola la condición de unicidad del polinomio interpolador.

21. El polinomio $p(x) = x^4 - x^3 + x^2 - x + 1$ interpola los datos $\mathbf{x} = [-2; -1; 0; 1; 2; 3]$, $\mathbf{y} = [31; 5; 1; 1; 11; 61]$. Con mínimo esfuerzo, encontrar un polinomio $q(x)$ que interpole los datos $\mathbf{x} = [-2; -1; 0; 1; 2; 3]$, $\mathbf{y} = [31; 5; 1; 1; 11; 30]$.

22. Se sospecha que los datos de la tabla 7.11 provienen de un polinomio cúbico. ¿Es posible identificarlo? Justificar.

23. Identificar el polinomio de grado mínimo que interpola los valores de la tabla 7.12. Idear alguna estrategia que permita hacer un desarrollo mínimo.

24. [*EMT*] La interpolación polinómica es muy sensible a los datos con que se construye el interpolador. Verificar esto generando el polinomio $p(x)$ que pasa por los datos de los vectores $\mathbf{x} = [0; 1; 2; 3; 4; 5; 6]$, $\mathbf{y} = [0; 1; 2{,}001; 3; 4; 5; 6]$. ¿Son lógicos los valores de $p(14)$ y $p(20)$?

x	1,73	1,82	2,61	5,22	8,26
y	0	0	7,8	0	0

Tabla 7.12

25. [*EMT*] Los nodos equiespaciados dentro del intervalo $[a;b]$ en un interpolador polinómico generan el *efecto Runge*. Dada una función, es posible atenuar esto utilizando los **nodos de Chebyshev**. Se generan $n+1$ nodos en el intervalo $[-1;1]$ a través de:

$$x_i = \cos\left[\left(\frac{2i+1}{2n+2}\right)\pi\right],$$

para $0 \leq i \leq n$. Luego se transforma linealmente el intervalo $[-1;1]$ en $[a;b]$ por medio de:

$$x_i = \frac{1}{2}(a+b) + \frac{1}{2}(b-a)\cos\left[\left(\frac{2i+1}{2n+2}\right)\pi\right],$$

para $0 \leq i \leq n$. Interpolar la función de Runge $R(x) = \dfrac{1}{1+x^2}$, con 9 nodos equiespaciados en el intervalo $[-5;5]$ y luego interpolar, en el mismo intervalo, con los nodos de Chebyshev. Graficar ambos interpoladores.

Bibliografía

- *A theoretical introduction to numerical analysis**, V. RYABEN'KII y S. TSYNKOV, Cap.2
- *Análisis numérico - Primer curso*, Hernán GONZÁLEZ, Cap.4
- *Análisis numérico - Un enfoque práctico*, M. MARON y R. LÓPEZ, Cap.6
- *Análisis numérico con aplicaciones*, C. GERALD y P. WHEATLEY, Cap.3
- *Fundamental numerical methods for electrical engineering**, Stanislaw ROSLONIEC, Cap.4
- *Numerical methods*, G. DAHLQUIST y A. BJÖRK, Cap.7

8 Ajuste de Datos

Mínimos cuadrados es una clase general de métodos para ajustar datos (generalmente obtenidos a través de mediciones) a una función que, teóricamente, modela los datos tabulados. Es decir que los datos son como los observados en la tabla 8.1 para el caso bidimensional y la función que modela los datos pertenece a una clase de funciones $\mathcal{F}$. El objetivo es encontrar la *mejor* $f \in \mathcal{F}$ que ajuste los datos a $y = f(x)$. Usualmente, la clase de funciones $\mathcal{F}$ estará determinada por un número pequeño de parámetros, mucho más pequeño, que la cantidad de datos obtenidos por medición. Resta definir el concepto de *mejor* función de ajuste y encontrar métodos para identificar los parámetros que la constituyen.

8.1. Mínimos cuadrados

Primero deben considerarse dos vectores de longitud $n+1$. Esto es:

$$\mathbf{x} = [x_0, x_1, \ldots, x_n]^T, \qquad \mathbf{y} = [y_0, y_1, \ldots, y_n]^T.$$

El **error**, o **residuo**, de una función dada con respecto a los datos es el vector $\mathbf{e} = f(\mathbf{x}) - \mathbf{y}$. Esto es:

$$\mathbf{e} = [e_0, e_1, \ldots, e_n]^T, \quad e_i = f(x_i) - y_i.$$

El objetivo es encontrar $f \in \mathcal{F}$ de forma tal que $\mathbf{e}$ sea razonablemente pequeño. Para medir un vector se utilizan las normas vectoriales. Si bien se puede medir el error con cualquiera de ellas, se recomienda utilizar $\|\cdot\|_2$.

Definición 20. *La **mejor aproximante** en el sentido de los mínimos cuadrados con respecto a un conjunto de datos,* $\mathbf{x}$, $\mathbf{y}$, *tomada de una clase de funciones* $\mathcal{F}$, *es la función* $f^* \in \mathcal{F}$ *que minimiza la norma 2 del vector de error. Esto es, si* f^* *es la mejor aproximante, entonces:*

$$\|f^*(\mathbf{x}) - \mathbf{y}\|_2 = \min_{f \in \mathcal{F}} \|f(\mathbf{x}) - \mathbf{y}\|_2$$

Se asume que este mínimo es único. Este método es llamado el método ordinario de mínimos cuadrados puesto que siempre se considera que no hay error en las mediciones de los datos de $\mathbf{x}$.

x	x_0	x_1	$\ldots$	x_n
y	y_0	y_1	$\ldots$	y_n

Tabla 8.1: Datos para ajustar con una función f, caso bidimensional.

8.2. Ajuste polinomial por mínimos cuadrados

El caso más simple implica ajustar datos por medio de recta. Para ello se asume que:

$$\mathcal{F} = \{f(x) = ax + b, \quad a, b \in \mathbb{R}\}.$$

Entonces se debe encontrar $f(x) = ax + b$ tal que minimice:

$$\|\mathbf{e}\|_2 = \left(\sum_{i=0}^{n} [f(x_i) - y_i]^2\right)^{1/2}.$$

Sin embargo, se puede minimizar el cuadrado del error, entonces:

$$\|\mathbf{e}\|_2^2 = \sum_{i=0}^{n} [ax_i + b - y_i]^2.$$

Para determinar a y b se debe derivar al error con respecto a cada una de las variables e igualar las expresiones a cero:

$$\frac{\partial \mathbf{e}}{\partial a} = \sum_{i=0}^{n} 2\,[ax_i + b - y_i]\,x_i = 0, \qquad \frac{\partial \mathbf{e}}{\partial b} = \sum_{i=0}^{n} 2\,[ax_i + b - y_i] = 0 \tag{8.1}$$

A fin de resolver (8.1), es necesario aplicar ciertas propiedades de sumatorias, con lo que queda:

$$\begin{aligned} a\sum_{i=0}^{n} x_i^2 + b\sum_{i=0}^{n} x_i &= \sum_{i=0}^{n} x_i y_i \\ a\sum_{i=0}^{n} x_i + b\sum_{i=0}^{n} 1 &= \sum_{i=0}^{n} y_i, \end{aligned} \tag{8.2}$$

lo que es un sistema de ecuaciones lineales (ecuaciones normales) y se puede resolver por cualquiera de los métodos vistos.

Ejemplo 47. *Los datos de la tabla 8.2, obtenidos por medición, deben ser ajustados a una función lineal. Para determinar los coeficientes a y b que logren el mejor ajuste es necesario crear el sistema de ecuaciones normales. Utilizando la expresión de las ecuaciones normales, (8.2), se obtiene el sistema:*

$$\begin{aligned} 252{,}98a + 47b &= 8886{,}3 \\ 47a + 9b &= 1602{,}7, \end{aligned}$$

que al ser resuelto por el método de Gauss da como solución:

$$a = 68{,}561; \qquad b = -179{,}96.$$

El error cuadrático asociado al ajuste es:

$$\begin{aligned} \|\mathbf{e}\|_2^2 &= \|a\mathbf{x} + b - \mathbf{y}\|_2^2 \\ &= \|68{,}561\mathbf{x} - 179{,}96 - \mathbf{y}\|_2^2 \\ &= \sum_{i=0}^{n} [68{,}561x_i - 179{,}96 - y_i]^2 \\ &= 194{,}88 \end{aligned}$$

El gráfico de los datos de la tabla 8.2 (círculos negros) y el ajuste lineal (línea continua azul) se aprecia en la figura 8.1.

x	4,0	4,2	4,5	4,7	5,1	5,5	5,9	6,3	6,8
y	100,75	112,07	127,59	137,91	165,67	190,08	221,67	254,98	292,02

Tabla 8.2: Datos para ajustar con una función lineal, planteados para el ejemplo 47.

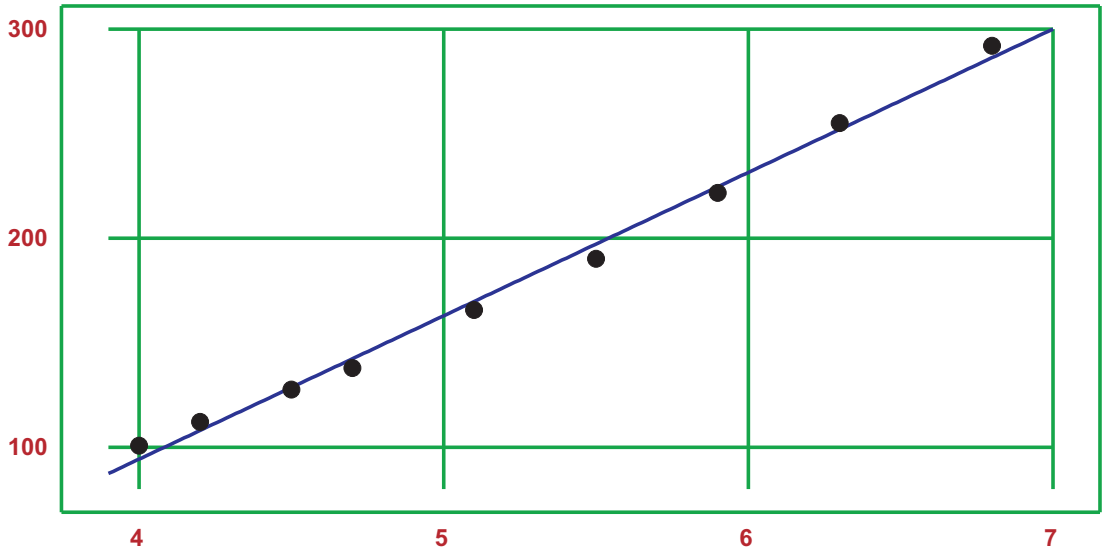

Figura 8.1: Ajuste lineal a los datos de la tabla 8.2 del ejemplo 47.

Ejercicio 26. *Calcular el número de condición de la matriz de coeficientes del sistema normal del ejemplo anterior.*

Si se observa que los datos pueden tomar una tendencia distinta de la que ofrece una recta, tal vez sea bueno verificar si su tendencia no es similar a la que modela una parábola o polinomio cuadrático. Sea entonces:

$$\mathcal{F} = \left\{ f(x) = ax^2 + bx + c, \qquad a, b, c \in \mathbb{R} \right\}.$$

Entonces se debe encontrar $f(x) = ax^2 + bx + c$ tal que minimice:

$$\|\mathbf{e}\|_2 = \left(\sum_{i=0}^{n} \left[f(x_i) - y_i \right]^2 \right)^{1/2}.$$

Realizando un procedimiento similar al desarrollado anteriormente, se generan las ecuaciones normales minimizando la expresión del cuadrado del error en norma dos, con respecto a cada variable involucrada. En este caso, cada uno de los coeficientes del polinomio cuadrático:

$$\frac{\partial \mathbf{e}}{\partial a} = \sum_{i=0}^{n} 2 \left[ax_i^2 + bx_i + c - y_i \right] x_i^2 = 0$$
$$\frac{\partial \mathbf{e}}{\partial b} = \sum_{i=0}^{n} 2 \left[ax_i^2 + bx_i + c - y_i \right] x_i = 0$$
$$\frac{\partial \mathbf{e}}{\partial c} = \sum_{i=0}^{n} 2 \left[ax_i^2 + bx_i + c - y_i \right] = 0,$$

con lo que se obtiene un sistema de ecuaciones lineales que puede ser resuelto por cualquiera de los métodos ya vistos:

$$\begin{aligned} a\sum_{i=0}^{n} x_i^4 + b\sum_{i=0}^{n} x_i^3 + c\sum_{i=0}^{n} x_i^2 &= \sum_{i=0}^{n} y_i x_i^2 \\ a\sum_{i=0}^{n} x_i^3 + b\sum_{i=0}^{n} x_i^2 + c\sum_{i=0}^{n} x_i &= \sum_{i=0}^{n} y_i x_i \\ a\sum_{i=0}^{n} x_i^2 + b\sum_{i=0}^{n} x_i + c\sum_{i=0}^{n} 1 &= \sum_{i=0}^{n} y_i. \end{aligned}$$

Ejemplo 48. *Observando el gráfico 8.1, se nota que los puntos extremos están por encima de la recta de ajuste, mientras que los puntos centrales se ubican por debajo. Tal vez se obtenga un mejor ajuste si se utiliza una función cuadrática en lugar de una recta. Para determinar los coeficientes a, b y c que logren el mejor ajuste es necesario crear el sistema de ecuaciones normales. Utilizando la expresión de las ecuaciones normales, (8.2), se obtiene el sistema:*

$$\begin{aligned} 7981{,}9a + 1401{,}9b + 252{,}98c &= 50617 \\ 1401{,}9a + 252{,}98b + 46{,}999c &= 8886{,}3 \\ 252{,}98a + 46{,}999b + 9c &= 1602{,}7 \end{aligned}$$

que al ser resuelto por el método de Gauss da como solución:

$$a = 6{,}2041; \qquad b = 2{,}0135; \qquad c = -6{,}8245.$$

El error cuadrático asociado al ajuste es:

$$\begin{aligned} \|\mathbf{e}\|_2^2 &= \|a\mathbf{x}^2 + b\mathbf{x} + c - \mathbf{y}\|_2^2 \\ &= \|6{,}2041\mathbf{x}^2 + 2{,}0135\mathbf{x} - 6{,}8245 - \mathbf{y}\|_2^2 \\ &= \sum_{i=0}^{n} [6{,}2041x_i^2 + 2{,}0135x_i - 6{,}8245 - y_i]^2 \\ &= 20{,}116 \end{aligned}$$

Ejercicio 27. *Calcular el número de condición de la matriz de coeficientes del sistema normal del ajuste cuadrático antes realizado.*

Comandos de EMT. *El comando para ajustar datos a través de polinomios es:*

- `polyfit(x:vector, y:vector, n:natural, w:vector)`, *donde* $\boldsymbol{x}$, $\boldsymbol{y}$ *es la nube de datos para ajustar;* $\boldsymbol{n}$ *es el grado del polinomio de ajuste;* $\boldsymbol{w}$ *es un parámetro opcional para agregarle pesos a los nodos. La salida es un vector con los coeficientes del polinomio de ajuste, ordenados de la menor potencia a la mayor.*

Ejemplo en EMT 18. *Ajustar los datos:*

$$\begin{aligned} x &= [0{,}7726; 1{,}973; 2{,}405; 2{,}637; 3{,}869] \\ y &= [1{,}337; 1{,}524; 2{,}004; 2{,}072; 2{,}733] \end{aligned}$$

a través de un polinomio de grado 2.

```
>X=[0.7726, 1.973, 2.405, 2.637, 3.869];
>Y=[1.337, 1.524, 2.004, 2.072, 2.733];
>polyfit(X,Y,2)
 [1.16307,  0.126529,  0.0739135]
```

8.2.1. Índice de determinación

En el caso de ajuste polinómico, a medida que aumenta el grado del polinomio disminuye el error cuadrático del ajuste de un conjunto dado de puntos. De hecho, dados $n+1$ puntos, el mejor ajuste se consigue con el polinomio interpolador de grado n, cuyo error cuadrático es cero. Sin embargo, a menudo es necesario un polinomio de grado reducido que ajuste una gran cantidad de puntos.

El error cuadrático no es un buen indicador de la aproximación polinómica, pues depende del número de puntos y de las unidades de medida empleadas. Como alternativa se define el **índice de determinación**:

$$I = \frac{\sum_{i=0}^{n} \left(f\left(x_i\right) - \overline{y}\right)^2}{\sum_{i=0}^{n} \left(y_i - \overline{y}\right)^2},$$

donde $\overline{y}$ es el promedio de las ordenadas y_i.

El índice de determinación es una cantidad adimensional entre 0 y 1. Cuanto más próximo esté a 1, mejor es el ajuste.

Ejemplo 49. *El índice de determinación para los dos ajuste polinómicos de los ejemplos 47 y 48 es el siguiente:*

$$\begin{aligned} I_1 &= \frac{\sum_{i=0}^{n} \left(f\left(x_i\right) - \overline{y}\right)^2}{\sum_{i=0}^{n} \left(y_i - \overline{y}\right)^2} \\ &= \frac{\sum_{i=0}^{n} \left(68{,}561x_i - 179{,}96 - 178{,}08\right)^2}{\sum_{i=0}^{n} \left(y_i - 178{,}08\right)^2} \\ &= \frac{35421}{35594} = 0{,}99513 \end{aligned}$$

$$\begin{aligned} I_2 &= \frac{\sum_{i=0}^{n} \left(f\left(x_i\right) - \overline{y}\right)^2}{\sum_{i=0}^{n} \left(y_i - \overline{y}\right)^2} \\ &= \frac{\sum_{i=0}^{n} \left(6{,}2041x_i^2 + 2{,}0135x_i - 6{,}8245 - 178{,}08\right)^2}{\sum_{i=0}^{n} \left(y_i - 178{,}08\right)^2} \\ &= \frac{35573}{35594} = 0{,}99941, \end{aligned}$$

con lo que se muestra que ambos ajustes son buenos, independientemente del error cuadrático asociado.

8.3. Ajuste discreto por mínimos cuadrados

Utilizando las ideas de la sección anterior, es sencillo ajustar cualquier conjunto de datos bidimensional por medio de una función arbitraria. Las opciones clásicas son:

- Polinomios de grado n, aunque generalmente se utilizan grados pequeños.

x	4,0	4,2	4,5	4,7	5,1	5,5	5,9	6,3	6,8
$\ln(y)$	4.6126	4.7191	4.8488	4.9266	5.11	5.2474	5.4012	5.5412	5.6768

Tabla 8.3: Datos preparados para ser ajustados a $y = be^{ax}$.

- Trigonométricas, combinaciones lineales de $\sin(ax)$, $\cos(bx)$ y $\tan(cx)$.
- Exponenciales, aunque puede dar errores altos debido a que es necesario linealizar la expresión de la función de ajuste. En el ejemplo 50 se aplica linealización.
- Combinaciones de todos los tipos de funciones antes mencionadas.

Es necesario tener en cuenta que cuando las funciones de ajuste dependen de varios coeficientes, el sistema lineal a resolver es mal condicionado. Esto aumenta de acuerdo al tamaño de la matriz normal.

Ejemplo 50. *Se desea ajustar los datos de la tabla 8.2 con una función exponencial del tipo $f(x) = be^{xa}$. Como no es posible de la expresión de la norma del error cuadrático obtener un sistema lineal, es necesario linealizarlo. Entonces, como se desea que $f(x) \sim y$:*

$$y = be^{ax}$$
$$\ln(y) = \ln(b) + ax$$

lo que implica recalcular la tabla de datos. La tabla 8.3 muestra los datos preparados especialmente para este ejemplo.

Ahora, la expresión del cuadrado de la norma 2 del error es:

$$\|\mathbf{e}\|_2^2 = \sum_{i=0}^{8} (ax_i + \ln(b) - \ln(y_i))^2,$$

las ecuaciones normales son:

$$a\sum_{i=0}^{8} x_i^2 + \ln(b)\sum_{i=0}^{8} x_i = \sum_{i=0}^{8} \ln(y_i)x_i$$
$$a\sum_{i=0}^{8} x_i + \ln(b)\sum_{i=0}^{8} 1 = \sum_{i=0}^{8} \ln(y_i),$$

y cuando se resuelve el sistema se obtienen los coeficientes:

$$a = 0{,}38339, \qquad \ln(b) = 3{,}1182 \to b = 22{,}605.$$

El número de condición de la matriz normal es $1{,}01 \times 10^3$. El gráfico de los datos (círculos negros) y su función de ajuste (línea continua azul) se observa en la figura 8.2. El error asociado al ajuste se calcula sobre la definición del problema y no sobre la linealización. Entonces:

$$\|\mathbf{e}\|_2^2 = \sum_{i=0}^{8} (be^{ax_i} - y_i)^2 \quad = 303{,}64,$$

lo que es un error bastante grande por más que el ajuste se vea correcto. En este caso, la diferencia entre $f(x_8)$ y el valor de y_8 es lo que aporta la mayor cantidad de error. En el caso de tener puntos muy alejados del ajuste, se los considera outliers *y hay diversas técnicas para lidiar con estos datos especiales.*

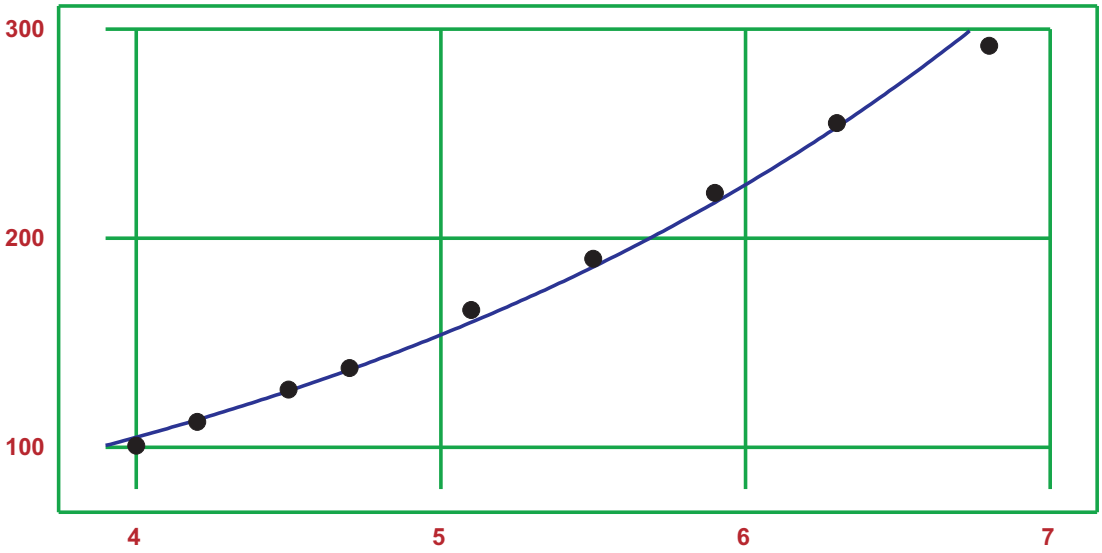

Figura 8.2: Ajuste exponencial ($f(x) = be^{ax}$) a los datos de la tabla 8.2.

$\ln(x)$	1,3863	1,4351	1,5041	1,5476	1,629	1,7047	1,775	1,8405	1,917
$\ln(y)$	4,6126	4,7191	4,8488	4,9266	5,11	5,2474	5,4012	5,5412	5,677

Tabla 8.4: Datos preparados para ser ajustados a $y = bx^a$.

Ejemplo 51. *Se desea ajustar los datos de la tabla 8.2 con una función exponencial del tipo $f(x) = bx^a$. Como no es posible de la expresión de la norma del error cuadrático obtener un sistema lineal, es necesario linealizarlo. Entonces, como se desea que $f(x) \sim y$:*

$$y = bx^a$$
$$\ln(y) = \ln(b) + a\ln(x)$$

lo que implica recalcular la tabla de datos. La tabla 8.4 muestra los datos preparados especialmente para este ejemplo.

Ahora, la expresión del cuadrado de la norma 2 del error es:

$$\|\mathbf{e}\|_2^2 = \sum_{i=0}^{8} (a\ln(x_i) + \ln(b) - \ln(y_i))^2,$$

las ecuaciones normales son:

$$a\sum_{i=0}^{8} \ln^2(x_i) + \ln(b)\sum_{i=0}^{8} \ln(x_i) = \sum_{i=0}^{8} \ln(y_i)\ln(x_i)$$
$$a\sum_{i=0}^{8} \ln(x_i) + \ln(b)\sum_{i=0}^{8} 1 = \sum_{i=0}^{8} \ln(y_i),$$

y cuando se resuelve el sistema se obtienen los coeficientes:

$$a = 2{,}0106, \qquad \ln(b) = 1{,}8278 \rightarrow b = 6{,}2202.$$

El número de condición de la matriz normal es $6{,}55 \times 10^1$, lo que supone un sistema bien condicionado. El gráfico de los datos (círculos negros) y su función de ajuste (línea continua azul) se observa en la figura 8.3. El error asociado al ajuste se calcula sobre

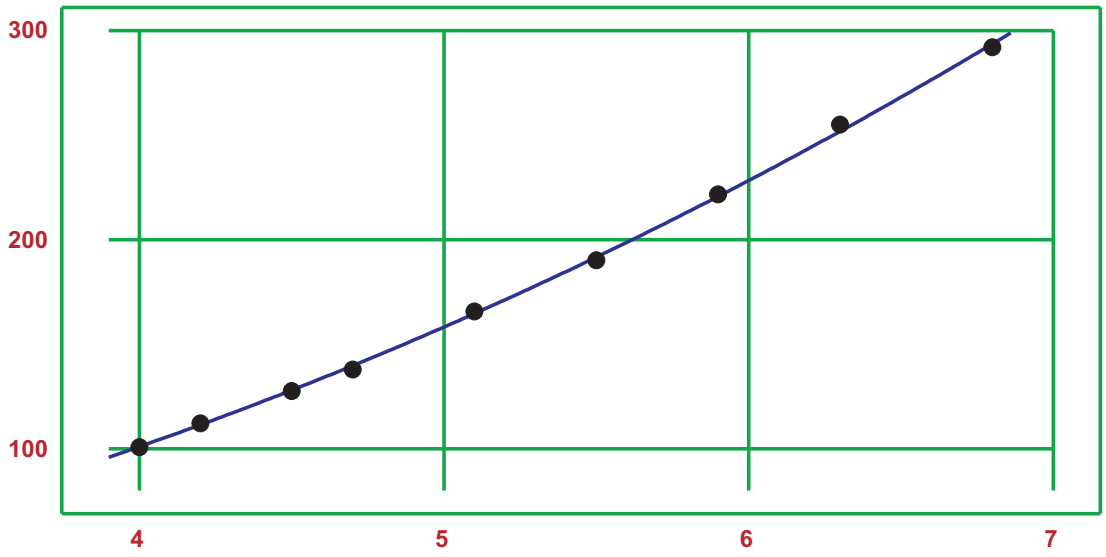

Figura 8.3: Ajuste exponencial ($f(x) = bx^a$) a los datos de la tabla 8.4.

la definición del problema y no sobre la linealización. Entonces:

$$\|\mathbf{e}\|_2^2 = \sum_{i=0}^{8} (bx_i^a - y_i)^2 \quad = 20{,}611,$$

lo que es un error mucho más aceptable que el del ejemplo anterior.

Ejercicio 28. *Ajustar los datos de la tabla 8.2 a las siguientes funciones:*

- $f(x) = ax^2 + bx + c$,
- $g(x) = ax^2 + b\cos(x)$,
- $h(x) = ax + be^x$,

y luego calcular el error cometido por el ajuste.

8.3.1. Linealización de funciones habituales

En la tabla 8.5 se muestran las funciones linealizadas más comunes para ajustar datos. Tener en cuenta que luego de la identificación de las constantes, es necesario volver a transformar los valores hacia la función original.

8.4. Ajuste funcional por mínimos cuadrados

Sea $y(x) \in C[a, b]$ y se intenta construir un polinomio de grado n que juste lo mejor posible a dicha función. En la sección anterior, se planteó el error de ajuste como:

$$\|\mathbf{e}\|_2^2 = \sum_{i=0}^{n} [f(x_i) - y_i]^2 ,$$

ahora no se puede calcular dicho vector, ya que la función y no es discreta sino continua. Sin embargo, se utiliza la misma idea que antes y se define el error como si fuera una sumatoria de términos infinitos, es decir una integral definida. Entonces el error asociado al proceso es:

$$E^2 = \int_a^b [P_n(x) - y(x)]^2 \, \mathrm{d}x.$$

Función a ajustar	**Función linealizada**
$y = \frac{a}{x} + b$	$y = a\frac{1}{x} + b \quad \rightarrow \quad y = a\overline{x} + b$
$y = \frac{b}{x+a}$	$\frac{1}{y} = \frac{1}{b}x + \frac{a}{b} \quad \rightarrow \quad \overline{y} = \overline{a}x + \overline{b}$
$y = ab^x$	$\ln(y) = [\ln(b)]x + ln(a) \quad \rightarrow \quad \overline{y} = \overline{a}x + \overline{b}$
$y = be^{ax}$	$\ln(y) = ax + \ln(b) \quad \rightarrow \quad \overline{y} = ax + \overline{b}$
$y = c - be^{-ax}$	$\ln(c - y) = -ax + \ln(b) \quad \rightarrow \quad \overline{y} = \overline{a}x + \overline{b}$
$y = ax^b$	$\ln(y) = b[\ln(x)] + \ln(a) \quad \rightarrow \quad \overline{y} = \overline{ax} + \overline{b}$
$y = axe^{bx}$	$\ln(y) - \ln(x) = bx + \ln(a) \quad \rightarrow \quad \overline{y} = \overline{a}x + \overline{b}$
$y = \frac{c}{1+be^{ax}}$	$\ln\left(\frac{c}{y} - 1\right) = ax + \ln(b) \quad \rightarrow \quad \overline{y} = ax + \overline{b}$

Tabla 8.5

La idea entonces se reduce a encontrar un polinomio tal que E^2 sea mínimo. Para ello se construye el polinomio en forma genérica:

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = \sum_{i=0}^{n} a_i x^i,$$

con lo que:

$$E^2 = \int_a^b \left[\sum_{i=0}^{n} a_i x^i - y(x)\right]^2 \mathrm{d}x$$

Como $E = E(a_0, a_1, \ldots, a_n)$, si se quiere reducir E al mínimo debe cumplirse que:

$$\frac{\partial E}{\partial a_i} = 0, \qquad i = 0, 1, \ldots, n.$$

Por lo tanto:

$$\begin{aligned} E^2 &= \int_a^b \left[\sum_{i=0}^{n} a_i x^i - y(x)\right]^2 \mathrm{d}x \\ &= \int_a^b \left[\sum_{i=0}^{n} a_i x^i\right]^2 \mathrm{d}x - 2\int_a^b y(x)\sum_{i=0}^{n} a_i x^i \,\mathrm{d}x + \int_a^b [y(x)]^2 \,\mathrm{d}x \end{aligned}$$

Derivando e igualando a cero:

$$\begin{aligned} \frac{\partial E^2}{\partial a_j} &= \int_a^b 2\left[\sum_{i=0}^{n} a_i x^i\right] x^j \,\mathrm{d}x - 2\int_a^b y(x)x^j \,\mathrm{d}x + 0 = 0 \\ &= 2\sum_{i=0}^{n} a_i \int_a^b x^{i+j} \,\mathrm{d}x - 2\int_a^b y(x)x^j \,\mathrm{d}x = 0, \end{aligned}$$

surgen, para cada j, las ecuaciones normales:

$$\sum_{i=0}^{n} a_i \int_a^b x^{i+j} \,\mathrm{d}x = \int_a^b y(x)x^j \,\mathrm{d}x,$$

o expandiendo la sumatoria:

$$\begin{aligned}
a_0\int_a^b x^0\,\mathrm{d}x + a_1\int_a^b x^1\,\mathrm{d}x + a_2\int_a^b x^2\,\mathrm{d}x + \cdots + a_n\int_a^b x^n\,\mathrm{d}x &= \int_a^b y(x)x^0\,\mathrm{d}x\\
a_0\int_a^b x^1\,\mathrm{d}x + a_1\int_a^b x^2\,\mathrm{d}x + a_2\int_a^b x^3\,\mathrm{d}x + \cdots + a_n\int_a^b x^{n+1}\,\mathrm{d}x &= \int_a^b y(x)x^1\,\mathrm{d}x\\
\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad &= \quad\vdots\\
a_0\int_a^b x^n\,\mathrm{d}x + a_1\int_a^b x^{n+1}\,\mathrm{d}x + a_2\int_a^b x^{n+2}\,\mathrm{d}x + \cdots + a_n\int_a^b x^{2n}\,\mathrm{d}x &= \int_a^b y(x)x^n\,\mathrm{d}x
\end{aligned}$$

con lo que debe resolverse un sistema lineal de tamaño n. Puede demostrarse que las $n+1$ ecuaciones normales tiene solución única si $f \in [a,b]$.

Ejemplo 52. *Se desea aproximar la función $y(x) = e^x$, definida en el intervalo $[0,3]$, por medio de un polinomio lineal. Entonces, la expresión del error cuadrático es:*

$$E^2 = \int_0^3 [a_0 + a_1 x - e^x]^2\,\mathrm{d}x,$$

de donde surgen las ecuaciones normales:

$$\begin{aligned}
a_0\int_0^3 1\,\mathrm{d}x + a_1\int_0^3 x\,\mathrm{d}x &= \int_0^3 e^x\,\mathrm{d}x\\
a_1\int_0^3 x\,\mathrm{d}x + a_2\int_0^3 x^2\,\mathrm{d}x &= \int_0^3 xe^x\,\mathrm{d}x,
\end{aligned}$$

que al resolverse las integrales se obtiene la solución del ejemplo. En este caso:

$$a_0 = -1{,}993; \qquad a_1 = 5{,}571.$$

El error asociado a esta aproximación es $E^2 = 9{,}87$. En la figura 8.4 se muestra la función $y(x) = e^x$, en línea punteada azul, y el polinomio aproximante, en línea continua roja.

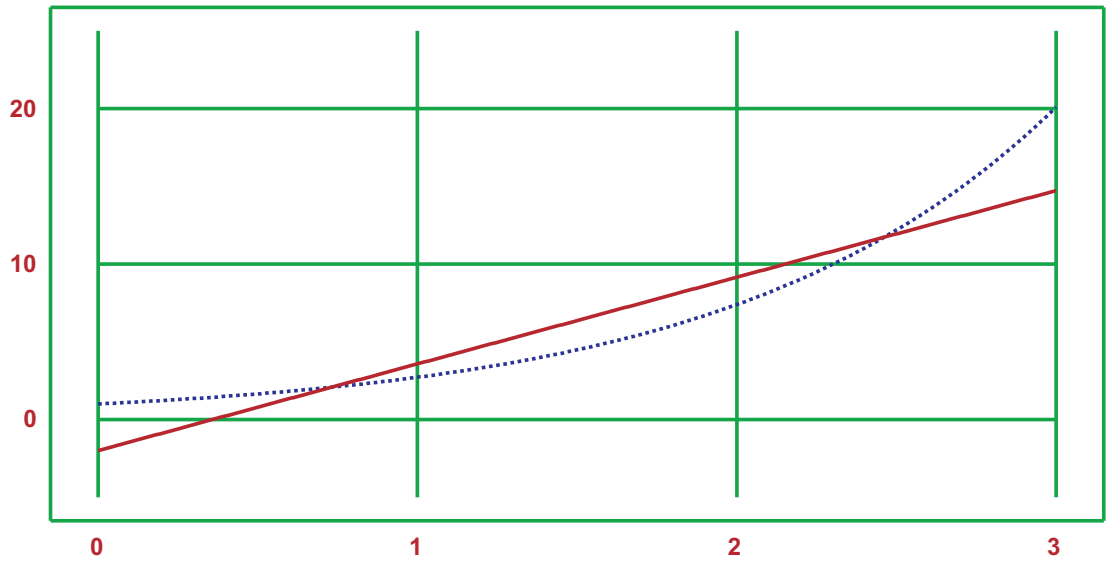

Figura 8.4: Ajuste lineal continuo a a función $f(x) = e^x$.

Ejemplo 53. *Se desea aproximar la función $y(x) = e^x$, definida en el intervalo $[0, 3]$, por medio de un polinomio cuadrático. Entonces, la expresión del error cuadrático es:*

$$E^2 = \int_0^3 \left[a_0 + a_1 x + a_2 x^2 - e^x\right]^2 \, \mathrm{d}x,$$

de donde surgen las ecuaciones normales:

$$a_0 \int_0^3 1 \, \mathrm{d}x + a_1 \int_0^3 x \, \mathrm{d}x + a_2 \int_0^3 x^2 \, \mathrm{d}x = \int_0^3 e^x \, \mathrm{d}x$$
$$a_0 \int_0^3 x \, \mathrm{d}x + a_1 \int_0^3 x^2 \, \mathrm{d}x + a_2 \int_0^3 x^3 \, \mathrm{d}x = \int_0^3 x e^x \, \mathrm{d}x$$
$$a_0 \int_0^3 x^2 \, \mathrm{d}x + a_1 \int_0^3 x^3 \, \mathrm{d}x + a_2 \int_0^3 x^4 \, \mathrm{d}x = \int_0^3 x^2 e^x \, \mathrm{d}x$$

que al resolverse las integrales se obtiene la solución del ejemplo. En este caso:

$$a_0 = 1{,}957; \qquad a_1 = -2{,}329; \qquad a_2 = 2{,}633.$$

El error asociado a esta aproximación es $E^2 = 5{,}78 \times 10^{-1}$. En la figura 8.5 se muestra la función $y(x) = e^x$, en línea punteada azul, y el polinomio aproximante, en línea continua roja.

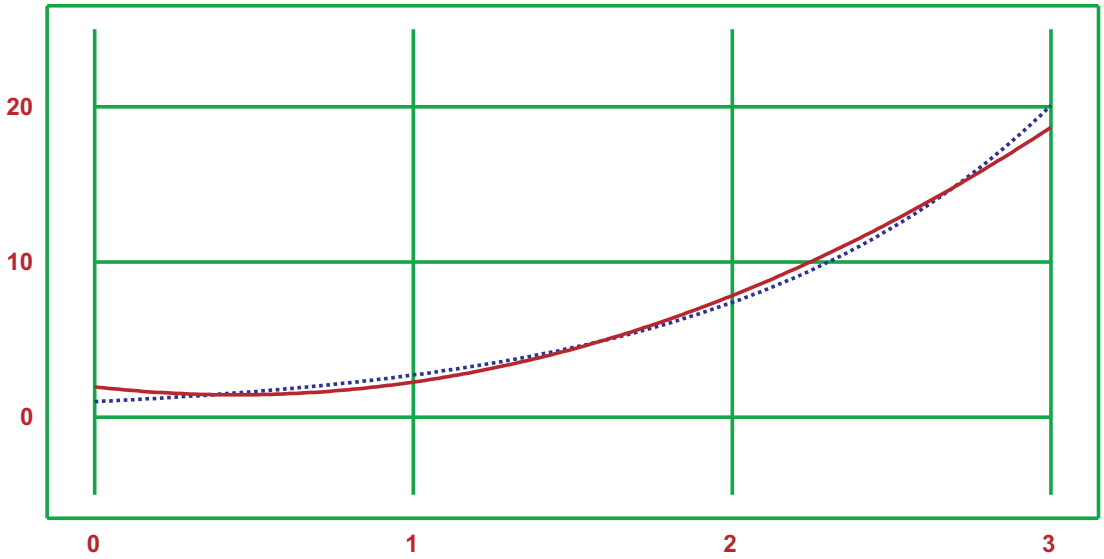

Figura 8.5: Ajuste cuadrático continuo a a función $f(x) = e^x$.

Ejemplo 54. *Se desea aproximar la función $y(x) = e^x$, definida en el intervalo $[0, 3]$, por medio de un polinomio cúbico. Entonces, la expresión del error cuadrático es:*

$$E^2 = \int_0^3 \left[a_0 + a_1 x + a_2 x^2 + a_3 x^3 - e^x\right]^2 \, \mathrm{d}x,$$

de donde surgen las ecuaciones normales:

$$a_0 \int_0^3 1 \, \mathrm{d}x + a_1 \int_0^3 x \, \mathrm{d}x + a_2 \int_0^3 x^2 \, \mathrm{d}x + a_3 \int_0^3 x^3 \, \mathrm{d}x = \int_0^3 e^x \, \mathrm{d}x$$
$$a_0 \int_0^3 x \, \mathrm{d}x + a_1 \int_0^3 x^2 \, \mathrm{d}x + a_2 \int_0^3 x^3 \, \mathrm{d}x + a_3 \int_0^3 x^4 \, \mathrm{d}x = \int_0^3 x e^x \, \mathrm{d}x$$
$$a_0 \int_0^3 x^2 \, \mathrm{d}x + a_1 \int_0^3 x^3 \, \mathrm{d}x + a_2 \int_0^3 x^4 \, \mathrm{d}x + a_3 \int_0^3 x^5 \, \mathrm{d}x = \int_0^3 x^2 e^x \, \mathrm{d}x$$
$$a_0 \int_0^3 x^3 \, \mathrm{d}x + a_1 \int_0^3 x^4 \, \mathrm{d}x + a_2 \int_0^3 x^5 \, \mathrm{d}x + a_3 \int_0^3 x^6 \, \mathrm{d}x = \int_0^3 x^3 e^x \, \mathrm{d}x$$

que al resolverse las integrales se obtiene la solución del ejemplo. En este caso:

$$a_0 = 0{,}8596; \qquad a_1 = 2{,}059; \qquad a_2 = -1{,}024; \qquad a_3 = 0{,}8126$$

El error asociado a esta aproximación es $E^2 = 2{,}05 \times 10^{-2}$. *En la figura 8.6 se muestra la función* $y(x) = e^x$, *en línea punteada azul, y el polinomio aproximante, en línea continua roja.*

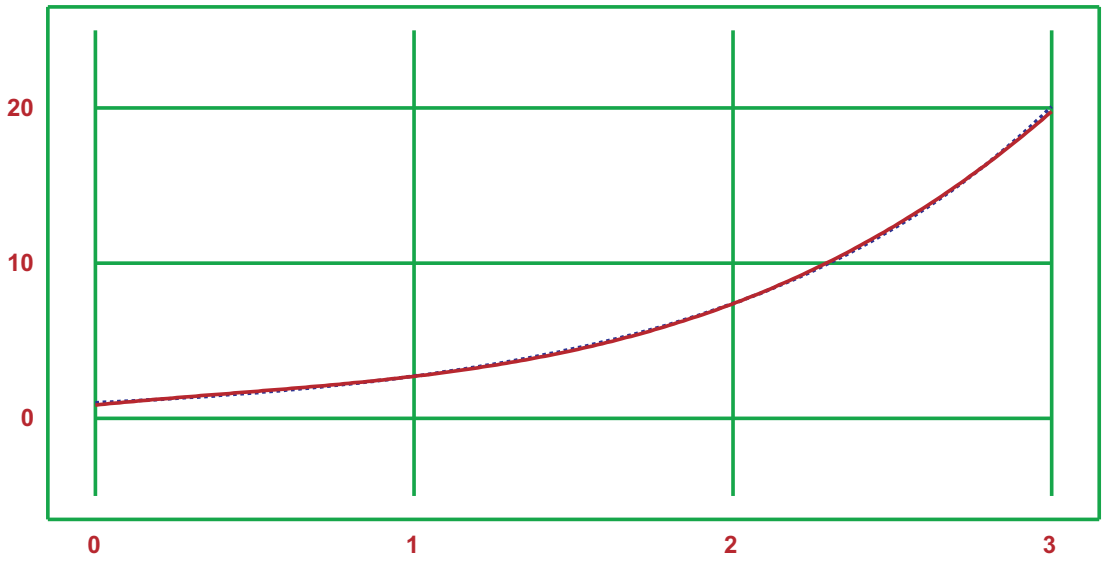

Figura 8.6: Ajuste lineal cúbico a a función $f(x) = e^x$.

Una de las ventajas principales de este método es que si se quiere aumentar el grado del polinomio que ajusta a una función, no es necesario calcular todos los coeficientes de nuevo, sino que se calculan dos integrales más: una para los coeficientes a_i y otra para los términos libres. La única complicación que surge es la resolución del sistema lineal, cada vez de mayor tamaño.

Nota. *Si bien a medida de que aumenta el grado del polinomio, el ajuste a una función es cada vez mejor, también es cierto que aumenta el grado del sistema lineal a resolver y el número de condición de la matriz de coeficientes. De los ejemplos anteriores:*

- *El número de condición de la matriz de coeficientes para el polinomio lineal es* $1{,}93 \times 10^1$;
- *El número de condición de la matriz de coeficientes para el polinomio cuadrático es* $6{,}01 \times 10^2$;
- *El número de condición de la matriz de coeficientes para el polinomio cúbico es* $2{,}86 \times 10^4$.

Ejercicio 29. *Sea la función* $y(x) = \sin(x)+2$, *definida en el intervalo* $[-1; 3]$. *Sabiendo que:*

$$\int x \sin(x)\, \mathrm{d}x = \sin(x) - x\cos(x) + \mathcal{C}$$

y

$$\int x^2 \sin(x)\, \mathrm{d}x = 2x\sin(x) + (2 - x^2)\cos(x) + \mathcal{C},$$

construir los polinomios de grado 1 y 2 que aproximen a $y(x)$. *Aproximar el valor del error cuadrático asociado, con alguna rutina computacional.*

8.5. Aproximación *Minimax*

La aproximación *minimax*, también denominada **mejor aproximación**, es una aproximación polinomial de norma mínima. Dos normas funcionales de uso habitual son la **norma infinito** ó **norma del máximo** y la **norma dos**.

Definición 21. *El conjunto $C[a,b]$ de funciones reales f, definidas y continuas en el intervalo $[a,b]$, es un espacio lineal normado con la norma:*

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

denominada ***norma infinito****.*

Definición 22. *El conjunto $C[a,b]$ de funciones reales f, definidas y continuas en el intervalo $[a,b]$, es un espacio lineal normado con la norma:*

$$\|f\|_2 = \left(\int_a^b w(x)\,|f(x)|^2 dx\right)^{1/2},$$

denominada ***norma dos****, donde $w(x)$ es una* ***función de peso****, continua, positiva e integrable en (a,b).*

Ambas normas se encuentran relacionadas, a pesar de tener expresiones muy diferentes. El siguiente teorema las relaciona.

Teorema 23. *Si la función de peso w es definida, continua, positiva e integrable en el intervalo (a,b), entonces se cumple que, para cualquier función $f \in C[a,b]$:*

$$\|f\|_2 \leq W\,\|f\|_\infty,$$

donde:

$$W = \left[\int_a^b w(x)\,dx\right]^{1/2}.$$

Además, dados dos números positivos ε (sin importar cuán pequeño sea) y M (sin importar cuán grande sea), existe una función $f \in C[a,b]$ tal que:

$$\|f\|_2 < \varepsilon, \|f\|_\infty > M.$$

De acuerdo al primer axioma de las normas, es posible pensar que $f \in C[a,b]$ puede ser bien aproximada por un polinomio p en $[a,b]$ si $\|f-p\|$ es pequeño, sin importar la norma con la que se esté calculando. El cálculo de polinomios mínimos en la norma dos ya fue realizado[1], ahora se centrará la atención en el análisis de la norma infinito, a partir del siguente teorema.

Teorema 24 (de Aproximación de Weierstrass)**.** *Suponiendo que f es una función real, definida y continua en un intervalo cerrado $[a,b]$, entonces dado cualquier $\varepsilon > 0$ existe un polinomio p tal que:*

$$\|f-p\|_\infty \leq \varepsilon.$$

De acuerdo al teorema 24, cualquier función f definida en $C[a,b]$ puede ser aproximada con error arbitrario a partir del conjunto de todos los polinomios. Si en vez de utilizar el conjunto de todos los polinomios se restringe la elección al conjunto $\mathcal{P}_n$

[1] con $w(x) = 1$

de polinomios de grado menor o igual a n, entonces no se cumple que, para cualquier $f \in C[a,b]$ y para cualquier $\varepsilon > 0$, existe $p_n \in \mathcal{P}_n$ tal que:

$$\|f - p\|_\infty \leq \varepsilon.$$

Entonces es relevante preguntarse cuán bien una función dada $f \in C[a,b]$ puede ser aproximada por polinomios de grado fijo $n \geq 0$. Por lo tanto, la definición de norma debe adaptarse a la siguiente expresión:

$$\|f - p_n^*\|_\infty = \inf_{p_n \in \mathcal{P}_n} \|f - p_n\|_\infty ,$$

donde el polinomio p_n^* se denomina **polinomio de mejor aproximación** de grado n a la función f en la norma infinito. Como la idea es que este polinomio minimice el máximo valor absoluto del error $f(x) - p_n(x)$ sobre $[a,b]$, se lo conoce como **polinomio *minimax***.

Lamentablemente los teoremas vistos no muestran la forma de construcción del polinomio *minimax*, sólo aseguran la existencia. Lo ideal sería que el error $f - p_n^*$ esté distribuido de manera uniforme sobre $[a,b]$, a fin de asegurar una convergencia óptima en todo el intervalo. El teorema de oscilación establece las condiciones finales con las que se definirá un polinomio *minimax*.

Teorema 25 (de Oscilación)**.** *Sea $f \in C[a,b]$. Entonces un polinomio $p_n^* \in \mathcal{P}_n$ es* minimax *para f en $[a,b]$ si y sólo si existe una secuencia de $n+2$ puntos x_i para $i = 0, 1, \ldots, n, n+1$ de forma tal que para:*

$$a \leq x_0 < \ldots < x_{n+1} \leq b;$$

se cumpla que:

$$|f(x_i) - p_n^*(x_i)| = \|f - p_n^*\|_\infty ,$$

con $i = 0, 1, \ldots, n, n+1$ y:

$$f(x_i) - p_n^*(x_i) = -\left[f(x_{i+1}) - p_n^*(x_{i+1})\right],$$

con $i = 0, 1, \ldots, n$.

8.5.1. Algoritmo de Remez

Este algoritmo genera una sucesión de polinomios $p_n^{(k)}$ que converge al polinomio *minimax* $p_n^* \in \mathcal{P}_n$, donde la sucesión de conjuntos $X^{(k)}$ converge a X^* sobre el cuál $f - p_n^*$ es equioscilante:

1. Escoger un conjunto $X = \{x_0, x_1, x_2, \ldots, x_n, x_{n+1}\}$.
2. Resolver el sistema de ecuaciones lineales:

$$f(x_i) - p_n^{(k)}(x_i) = (-1)^i E$$

3. Para generar $X^{(k+1)}$, deben tomarse aquellos valores de x donde el error *minimax* sea máximo en valor absoluto, incluidos ambos extremos de intervalo. En el caso de que sobre un punto, debe dejarse en el conjunto aquel extremo de intervalo que muestre mayor error y mantenga la alternancia de signos. El otro extremo debe ser eliminado de la lista.

Este algoritmo finaliza cuando la convergencia de $X^{(k)}$ se logra ó bien cuando el error obtenido en la resolución del sistema lineal alcanza un valor de tolerancia previamente establecido.

Ejemplo 55. *Debe conseguirse la aproximación* minimax *de grado uno a la función* $f(x) = \cos(x) + \sqrt{x}$ *en* $[0{,}5; 3]$. *Para ello se aplicará el algoritmo de Remez. Para* $X^{(0)} = \{0{,}5; 1{,}75; 3\}$, *debe resolverse el sistema (luego de acomodar convenientemente los términos):*

$$\begin{aligned} f(0{,}5) - (a(0{,}5) + b) &= E \\ f(1{,}75) - (a(1{,}75) + b) &= -E \\ f(3) - (a(3) + b) &= E \end{aligned}$$

cuya solución es $E = 0{,}0093721$, $a = -0{,}33705$ *y* $b = 1{,}7438$. *Los extremos de:* $f(x) - (-0{,}33705x + 1{,}7438)$ *son:* $0{,}99442$ *y* $2{,}4230$.

De los extremos del intervalo, se elige $x = 3$ *puesto que la alternancia de signos se cumple de acuerdo a:*

$$\begin{aligned} f(x) - (0{,}33705x + 1{,}7438)|_{x=0{,}5} &> 0 \\ f(x) - (0{,}33705x + 1{,}7438)|_{x=0{,}99442} &> 0 \\ f(x) - (0{,}33705x + 1{,}7438)|_{x=2{,}4230} &< 0 \\ f(x) - (0{,}33705x + 1{,}7438)|_{x=3} &> 0 \end{aligned}$$

Ahora $X^{(1)} = \{0{,}99442; 2{,}4230; 3\}$, *entonces luego de acomodar los términos en forma conveniente, debe resolverse el sistema:*

$$\begin{aligned} f(0{,}99442) - (a(0{,}99442) + b) &= E \\ f(2{,}4230) - (a(2{,}4230) + b) &= -E \\ f(3) - (a(3) + b) &= E \end{aligned}$$

cuya solución es $E = 0{,}084194$, $a = -0{,}39895$ *y* $b = 1{,}8547$. *Los extremos de* $f(x) - (-0{,}39895x + 1{,}8547)$ *son:* $1{,}0772$ *y* $2{,}3281$.

De los extremos del intervalo, se elige $x = 3$ *con el mismo criterio que en la iteración anterior.*

Siguiendo de forma similar a las dos primeras iteraciones se obtiene:

$$X^{(2)} = \{1{,}0791; 2{,}3260; 3\}$$

y

$$X^{(3)} = \{1{,}0791; 2{,}3260; 3\},$$

con lo que el algoritmo termina y la aproximación minimax *de grado uno es:*

$$p_1^* = -0{,}40026x + 1{,}8567.$$

En las figuras 8.7 y 8.8 se muestran la función $f(x)$ *(en línea punteada de color azul) y su correspondiente polinomio* minimax *(línea continua de color rojo) y el error entre función y polinomio* minimax *respectivamente.*

Ejemplo 56. *Se mostrará la convergencia del polinomio* minimax *de tercer grado a la función* $f(x) = \cos(x) + \sqrt{x}$ *en* $[0{,}5; 3]$. *Para ello se realizarán 5 iteraciones del algoritmo de Remez. Sea* $X^{(0)} = \{0{,}5; 1{,}125; 1{,}75; 2{,}375; 3\}$, *entonces:*

$$\begin{aligned} X^{(1)} &= \{0{,}5;\ 0{,}87248;\ 1{,}8073;\ 2{,}6718;\ 3\} \\ X^{(2)} &= \{0{,}5;\ 0{,}88988;\ 1{,}8535;\ 2{,}6811;\ 3\} \\ X^{(3)} &= \{0{,}5;\ 0{,}89000;\ 1{,}8504;\ 2{,}6792;\ 3\} \\ X^{(4)} &= \{0{,}5;\ 0{,}88997;\ 1{,}8499;\ 2{,}6810;\ 3\} \\ X^{(5)} &= \{0{,}5;\ 0{,}88861;\ 1{,}8496;\ 2{,}6802;\ 3\}. \end{aligned}$$

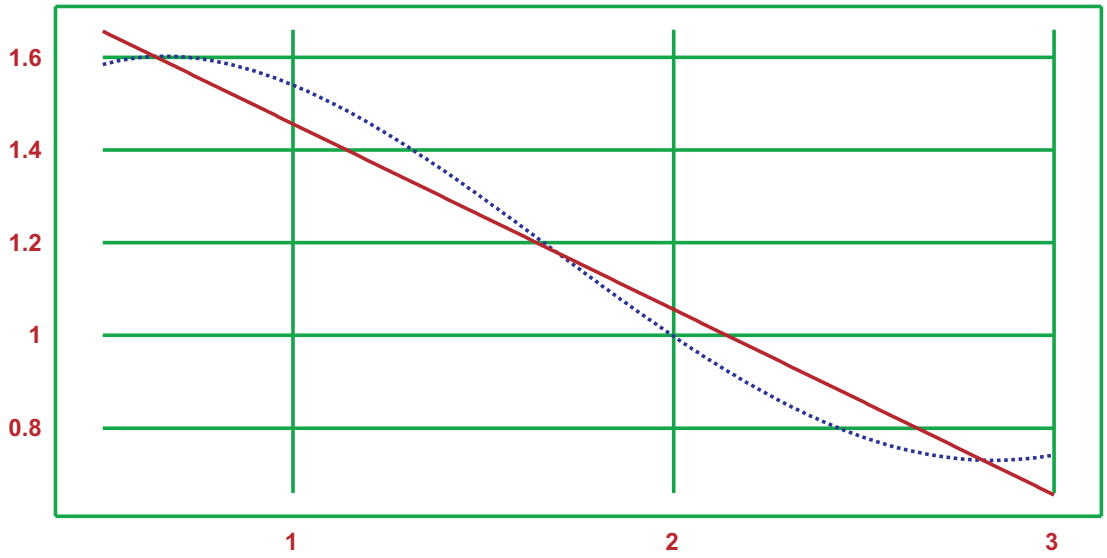

Figura 8.7: Gráfica de $f(x) = \cos(x) + \sqrt{x}$ y su polinomio *minimax*.

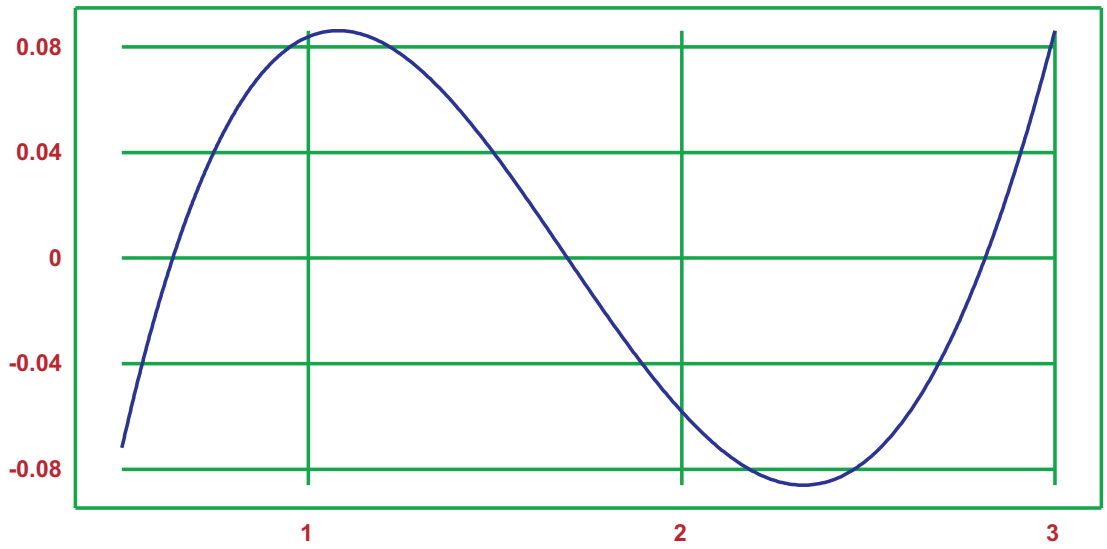

Figura 8.8: Error entre $f(x) = \cos(x) + \sqrt{x}$ y su polinomio *minimax* de grado uno.

La sucesión de polinomios minimax *es la siguiente:*

$$
\begin{aligned}
p_3^{(0)}(x) &= 0{,}17041x^3 - 0{,}88268x^2 + 0{,}90045x + 1{,}3272\\
p_3^{(1)}(x) &= 0{,}16892x^3 - 0{,}87469x^2 + 0{,}90848x + 1{,}3329\\
p_3^{(2)}(x) &= 0{,}16904x^3 - 0{,}87526x^2 + 0{,}90920x + 1{,}3326\\
p_3^{(3)}(x) &= 0{,}16900x^3 - 0{,}87511x^2 + 0{,}90903x + 1{,}3327\\
p_3^{(4)}(x) &= 0{,}16905x^3 - 0{,}87538x^2 + 0{,}90950x + 1{,}3325\\
p_3^{(5)}(x) &= 0{,}16906x^3 - 0{,}87541x^2 + 0{,}90945x + 1{,}3325.
\end{aligned}
$$

Sin embargo, con una partición de 10000 puntos en $[0{,}5; 3]$*, el error* minimax *es:*

$$
\begin{aligned}
E^{(0)} &= 7{,}30 \times 10^{-3}\\
E^{(1)} &= 5{,}20 \times 10^{-3}\\
E^{(2)} &= 5{,}20 \times 10^{-3}\\
E^{(3)} &= 5{,}20 \times 10^{-3}\\
E^{(4)} &= 5{,}10 \times 10^{-3}\\
E^{(5)} &= 5{,}24 \times 10^{-3},
\end{aligned}
$$

lo que sugiere que tal vez debería haberse finalizado el algoritmo en la cuarta iteración.

8.6. Ejercicios

1. Construir un algoritmo en PC que permita ajustar un conjunto de datos por medio de un polinomio de grado n.

2. Dada la tabla de valores 8.6, generada por medio de $f(x) = a\sin(x) + b\cos(x)$, determinar los valores de a y b. ¿Es necesario utilizar todos los datos?

x	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
$f(x)$	3,16	3,01	2,73	2,47	2,13	1,82	1,52	1,21	0,76	0,43	0,03

Tabla 8.6

3. Ajustar la tabla de datos 8.7 a una función de la forma:

$$f(x) = \frac{1}{1 + e^{-ax}}.$$

   Si no es posible, justificar por qué.

x	-8	-7	-6	-5	-4	-3	-2	-1
$f(x)$	0,0150	0,0338	0,0468	0,0712	0,1152	0,1850	0,2716	0,3775

Tabla 8.7

4. La densidad relativa ρ del aire se midió a diferentes altitudes, generando la tabla 8.8. Utilizando un ajuste cuadrático, determinar la densidad relativa cuando la altura es de 10.5km.

h (km)	0	1,525	3,050	4,575	6,10	7,625	9,150
ρ	1	0,8617	0,7385	0,6292	0,5328	0,4481	0,3741

Tabla 8.8

x	0,5	1,0	1,5	2,0	2,5
$f(x)$	0,541	0,398	0,232	0,106	0,052

Tabla 8.9

5. Ajustar los datos de la tabla 8.9 a la función $f(x) = axe^{bx}$. Calcular el error cuadrático total cometido.

6. ¿Será posible encontrar algún ajuste polinomial que presente menos error para los datos del inciso anterior?

7. Encontrar los coeficientes necesarios para ajustar la función $f(x) = e^x$ por medio de $g(x) = a_0x + a_1\cos(x)$, dentro del intervalo $[0, 2]$.

8. Sugerir funciones que, agregadas al resultado anterior, mejoren notablemente el ajuste. Resolver nuevamente.

9. $[EMT]$ Ajustar la función $f(x) = \sin(x)$ por medio de un polinomio de grado 3 dentro del intervalo $[0; 2\pi]$. Aumentar el grado del polinomio de ajuste a grado 5, ¿mejora el resultado?

10. $[EMT]$ Repetir el ejercicio del inciso anterior, pero ahora utilizar un polinomio *minimax* de grado 3. Compararlo con el polinomio de ajuste de grado 5.

11. El ajuste polinomial es fácilmente extensible a datos de más dos variables. Verificar que los datos de la tabla 8.10 se ajustan a un modelo del tipo $z = a + bx - cy$, resolviendo el sistema lineal asociado al ajuste cuadrático.

x	0	2	2,5	1	4	7
y	0	1	2	3	6	2
z	5	10	9	0	3	27

Tabla 8.10

12. En Barcelona, la tarifa del taxi se compone de una parte fija (bajada de bandera), otra parte proporcional a la distancia recorrida y otra proporcional al tiempo de espera. La tabla 8.11 recoge los datos de distintos viajes. Es decir que:

$$CV = a + bD + cT,$$

donde D es la distancia y T la espera.

   a) Determinar el valor de las constantes a, b y c a partir de los datos de la tabla.

   b) Estimar el precio de un viaje de 5km con 18min de espera.

Distancia (km)	8,5	2,0	8,4	6,8	8,3	7,1	3,0	1,9	3,0
Tiempo (min)	10,5	13,4	0,4	7,6	10,1	8,6	3,8	13,6	10,8
Costo de Viaje (euros)	10,5	6,6	7,7	8,5	10,2	9,0	4,7	6,6	6,6

Tabla 8.11

13. Es posible demostrar que si $f \in C^2[a,b]$ con $f''(x) > 0$ para $a \leq x \leq b$, entonces $q_1^*(x) = a_0 + a_1 x$ es la aproximación lineal *minimax* a $f(x)$ en $[a,b]$, donde:

$$a_1 = \frac{f(b) - f(a)}{b-a}; \qquad a_0 = \frac{f(a)+f(c)}{2} - \left(\frac{a+c}{2}\right)\left[\frac{f(b)-f(a)}{b-a}\right]$$

y c es la única solución de:

$$f'(c) = \frac{f(b)-f(a)}{b-a}$$

Verificar este enunciado con alguna función que cumpla con la condición pedida.

14. Identificar el valor de α que minimiza:

$$\int_0^1 |e^x - \alpha|\, \mathrm{d}x.$$

¿Cuál es el mínimo? Esta es una forma simple de ilustrar otra forma de medir el error de una aproximación. Este caso representa la aproximación de la constante α a la función e^x en el intervalo $[0;1]$.

15. $[EMT]$ Para $f(x) = e^x$ en $[-1;1]$ crear el polinomio de Taylor de grado cuatro, $p_4(x)$, expandido alrededor de $x = 0$. Ahora, identificar el polinomio *minimax* de grado tres $p_3^*(x)$ que aproxima a $p_4(x)$. Graficar los errores $e^x - p_3^*(x)$ y $e^x - p_3(x)$. Este proceso de reducir en un grado el polinomio de Taylor por medio de un polinomio *minimax* se denomina **economización** ó ***telescoping*** y se aplica varias veces para reducir un polinomio Taylor de grado alto hacia una aproximación de grado mucho menor.

16. Crear el conjunto de datos $y_i = 0{,}354x_i + 0{,}28 + (-1)^i 0{,}5$ a partir de $\mathbf{x} = [1;2;3;4;5;6;7]$, donde $i = 0, 1, \ldots, 6$. ¿Cuál es la recta de ajuste de mínimos cuadrados?

17. De acuerdo al conjunto de datos $y_i = 0{,}354x_i^2 + 0{,}28x_i - 2{,}5 + (-1)^i 0{,}5$ donde $\mathbf{x} = [1;2;3;4;5;6]$, ajustarlo con un polinomio cuadrático. ¿Se llega a la misma conclusión que la obtenida en el ejercicio anterior?

18. Determinar los parámetros a y b para que la fórmula:

$$y = \left(\frac{a+\sqrt{x}}{b\sqrt{x}}\right)^2$$

ajuste los datos de la tabla 8.12.

19. La concentración de la bacteria *E. Coli* en una pileta es monitoreada luego de una tormenta, arrojando los datos de la tabla 8.13. El tiempo es medido en horas a partir de la finalización de la tormenta y la unidad CFU significa *Colony Forming Unit* o Unidad de Formación de Colonias. Ajustar los datos a un modelo de la forma:

$$y = ae^{bx}$$

y estimar:

x	0,5	1	2	3	4
y	10,4	5,8	3,3	2,5	2

Tabla 8.12

a) La concentración al finalizar la tormenta, $t = 0$.

b) El tiempo en el cual la concentración alcanzará 200CFU/100mL.

t (horas)	4	8	12	16	20	24
c (CFU/100mL)	1590	1320	1000	900	650	560

Tabla 8.13

20. Repetir el ejercicio 4 pero ahora normalizar los datos. Es decir que se minimizará un modelo de la forma:

$$\|\mathbf{e}\|_2 = \left(\sum_{i=0}^{n} [f(x_i - \overline{x}) - (y_i - \overline{y})]^2 \right)^{1/2},$$

donde $\overline{x}$ e $\overline{y}$ son los promedios de los datos tabulados. ¿Mejora el error cuadrático? ¿Mejora el índice de determinación?

21. Dados los datos de la tabla 8.14:

a) Ajustarlos a un modelo de la forma $y = ax + b$.

b) Invertir los roles de las variables y ajustarlos a un modelo de la forma $x = c + dy$.

c) Comparar los resultados de los incisos anteriores en base a los errores cuadráticos y al índice de determinación.

x	0,0	0,1	0,2	0,3	0,4	0,5	0,7
y	1,9	2,8	2,6	2,3	2,1	2,1	1,6

Tabla 8.14

22. Usando como base al conjunto de datos de la tabla 8.15, generar el ajuste por mínimos cuadrados de $(\mathbf{x}; \mathbf{y} + \delta\mathbf{y})$ a un polinomio cuadrático, donde:

a) $\delta\mathbf{y} = [0; 0; 0; 1{,}1; 0]$.

b) $\delta\mathbf{y} = [0; 1{,}1; 0; 0; 0]$.

c) Sabiendo que los datos de la tabla 8.15 provienen de la relación $y = -x^2 + 3$, ¿qué conclusión se puede sacar con respecto a las perturbaciones en los datos?

23. Encontrar una función de la forma $y = e^{cx}$ que ajuste los datos: $\mathbf{x} = [0; 1]$; $\mathbf{y} = [a; b]$.

24. [*EMT*] Considerar el ajuste lineal de la forma $y = a + b(x - c)$ a los datos de la tabla 8.16.

x	1,791	2,418	3,469	4,898	5,388
y	-0,2076	-2,846	-9,033	-20,99	-26,03

Tabla 8.15

a) Mostrar que, con el desplazamiento $c = 1{,}000$ y una aritmética de 4 dígitos, no se resuelve correctamente el problema.

b) Resolver correctamente el ajuste, eligiendo un valor de c que no haga que la matriz del sistema de ecuaciones normales sea mal condicionada.

x	1	3	4	6	7
y	-2,1	-0,9	-0,6	0,6	0,9

Tabla 8.16

25. Demostrar que la recta de ajuste lineal pasa exactamente por el punto promedio de la nube de datos.

Bibliografía

- *An introduction to numerical analysis*, Kendall ATKINSON, Cap.4
- *Análisis numérico*, R. BURDEN y J. FAIRES, Cap.8
- *Métodos numéricos con MATLAB*, J. MATHEWS y K. FINK, Cap.5
- *Numerical calculations and algorithms*, R. BECKETT y J. HURT, Cap.8

9 Derivación Numérica

La derivación numérica surge al presentarse el siguiente problema: dada una función $y = f(x)$, se desea obtener una de sus derivadas en el punto $x = x_k$. El término *dada* se refiere a que se posee un algoritmo con el que se calculan los puntos de la función, o se tiene un conjunto discreto de datos (x_i, y_i), $i = 1, 2, \ldots, n$. En dicho caso, sólo se tiene acceso a un número finito de pares (x, y) con los cuales se computará la derivada.

Uno de los métodos básicos consiste en aproximar la función localmente por un polinomio y luego diferenciarlo. Otra herramienta efectiva es utilizar una serie de Taylor alrededor del punto x_k. Este segundo método tiene una ventaja: se puede calcular una cota para el error.

La derivación numérica no es un proceso particularmente preciso. Se sufren los errores de redondeo, de truncamiento y los errores que son inherentes a la interpolación. Por esta razón, la derivada de una función nunca puede ser calculada con la misma exactitud que cuando se utiliza la definición de derivada como límite del cociente incremental.

9.1. Aproximación de derivadas por diferencias finitas

La derivación por medio de diferencias finitas se basa en la expansión de series de Taylor (hacia adelante y hacia atrás), dadas como:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x) + \frac{h^4}{4!}f^{(4)}(x) + \cdots \tag{9.1}$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2!}f''(x) - \frac{h^3}{3!}f'''(x) + \frac{h^4}{4!}f^{(4)}(x) - \cdots \tag{9.2}$$

$$f(x+2h) = f(x) + 2hf'(x) + \frac{(2h)^2}{2!}f''(x) + \frac{(2h)^3}{3!}f'''(x) + \frac{(2h)^4}{4!}f^{(4)}(x) + \cdots \tag{9.3}$$

$$f(x-2h) = f(x) - 2hf'(x) + \frac{(2h)^2}{2!}f''(x) - \frac{(2h)^3}{3!}f'''(x) + \frac{(2h)^4}{4!}f^{(4)}(x) - \cdots \tag{9.4}$$

También se puede plantear la suma y resta de las series:

$$f(x+h)+f(x-h) = 2f(x)+h^2f''(x)+\frac{h^4}{12}f^{(4)}(x)+\cdots \tag{9.5}$$

$$f(x+h)-f(x-h) = 2hf'(x)+\frac{h^3}{3}f'''(x)+\cdots \tag{9.6}$$

$$f(x+2h)+f(x-2h) = 2f(x)+4h^2f''(x)+\frac{4h^4}{3}f^{(4)}(x)+\cdots \tag{9.7}$$

$$f(x+2h)-f(x-2h) = 4hf'(x)+\frac{8h^3}{3}f'''(x)+\cdots \tag{9.8}$$

Es importante notar que las sumas contienen solamente derivadas de orden par, mientras que las restas contiene las derivadas de orden impar. Las ecuaciones (9.2)–(9.8) pueden ser vistas como ecuaciones simultáneas que deben ser resueltas para obtener las derivadas de $f(x)$. El número de ecuaciones involucradas y la cantidad de términos de cada ecuación dependen del orden de la derivada y el grado de precisión deseados.

9.1.1. Primera aproximación por diferencias centrales

La solución de la ecuación (9.6) para $f'(x)$ es:

$$f'(x)=\frac{f(x+h)-f(x-h)}{2h}-\frac{h^2}{6}f'''(x)-\cdots$$

Manteniendo sólo el primer término del lado derecho de la igualdad anterior resulta:

$$f'(x)=\frac{f(x+h)-f(x-h)}{2h}+\mathcal{O}(h^2), \tag{9.9}$$

lo que se denomina **primera aproximación por diferencias centrales**. El término $\mathcal{O}(h^2)$ indica que el error de truncamiento se comporta como h^2.

A partir de la ecuación (9.5) se obtiene:

$$f''(x)=\frac{f(x+h)-2f(x)+f(x-h)}{h^2}+\frac{h^2}{12}f^{(4)}(x)+\cdots$$

o mejor aún:

$$f''(x)=\frac{f(x+h)-2f(x)+f(x-h)}{h^2}+\mathcal{O}(h^2).$$

Al igual que en el caso de la aproximación de $f'(x)$, el error para la aproximación de $f''(x)$ es del mismo orden que h^2.

Las aproximaciones por diferencias centrales para derivadas de orden superior se obtienen a partir de las ecuaciones (9.2)–(9.8) de manera similar. Por ejemplo, eliminando $f'(x)$ de las ecuaciones (9.6) y (9.8), y resolviendo para $f'''(x)$ se sigue que:

$$f'''(x)=\frac{f(x+2h)-2f(x+h)+2f(x-h)-f(x-2h)}{2h^3}+\mathcal{O}(h^2).$$

La aproximación:

$$f^{(4)}(x)=\frac{f(x+2h)-4f(x+h)+6f(x)-4f(x-h)+f(x-2h)}{h^4}+\mathcal{O}(h^2),$$

se consigue a través de las ecuaciones (9.5) y (9.7) luego de eliminar $f''(x)$. La tabla 9.1 resume estos resultados.

	$f(x-2h)$	$f(x-h)$	$f(x)$	$f(x+h)$	$f(x+2h)$
$2hf'(x)$		-1	0	1	
$h^2f''(x)$		1	-2	1	
$2h^3f'''(x)$	-1	2	0	-2	1
$h^4f^{(4)}(x)$	1	-4	6	-4	1

Tabla 9.1: Coeficientes de la aproximación por diferencias centrales.

	$f(x)$	$f(x+h)$	$f(x+2h)$	$f(x+3h)$	$f(x+4h)$
$hf'(x)$	-1	1			
$h^2f''(x)$	1	-2	1		
$h^3f'''(x)$	-1	3	-3	1	
$h^4f^{(4)}(x)$	1	-4	6	-4	1

Tabla 9.2: Coeficientes de la primera aproximación por diferencias hacia adelante.

9.1.2. Primera aproximación por diferencias no centrales

No siempre es posible construir aproximaciones de derivada por diferencias centrales. Por ejemplo, si se considera la situación donde la función está dada en n puntos, $x_1, x_2, \ldots, x_n$, no se puede calcular la derivada en los puntos extremos. Esto es debido a que se necesita información de ambos lados del punto sobre el cual se aproximará la derivada. Si sólo se tiene información a derecha o a izquierda se utilizarán las aproximaciones *hacia adelante* y *hacia atrás*.

La aproximación de derivadas por diferencias no centrales se obtienen a partir de las ecuaciones (9.2)–(9.8). Resolviendo la ecuación (9.2) para $f'(x)$ se obtiene:

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{h}{2}f''(x) - \frac{h^2}{6}f'''(x) - \frac{h^3}{4!}f^{(4)}(x) - \cdots$$

Mantentiendo sólo el primer término del lado derecho, se consigue lo que se denomina **primera aproximación por diferencia hacia adelante**:

$$f'(x) = \frac{f(x+h) - f(x)}{h} + \mathcal{O}(h)$$

Por medio de un procedimiento similar, se logra la **primera aproximación por diferencia hacia atrás**:

$$f'(x) = \frac{f(x) - f(x-h)}{h} + \mathcal{O}(h)$$

Un punto a tener en cuenta es que el error en este caso es del orden de h, $\mathcal{O}(h)$, mientras que en las diferencias centrales el error es de orden h^2, $\mathcal{O}(h^2)$.

Es posible deducir las aproximaciones para derivadas de mayor orden de la misma manera. Por ejemplo, de las ecuaciones (9.2) y (9.4):

$$f''(x) = \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2} + \mathcal{O}(h)$$

Las aproximaciones a las derivadas tercera y cuarta pueden ser originadas por un procedimiento similar. Este resultado se muestra en las tablas 9.2 y 9.3.

	$f(x-4h)$	$f(x-3h)$	$f(x-2h)$	$f(x-h)$	$f(x)$
$hf'(x)$				-1	1
$h^2f''(x)$			1	-2	1
$h^3f'''(x)$		-1	3	-3	1
$h^4f^{(4)}(x)$	1	-4	6	-4	1

Tabla 9.3: Coeficientes de la primera aproximación por diferencias hacia atrás.

	$f(x)$	$f(x+h)$	$f(x+2h)$	$f(x+3h)$	$f(x+4h)$	$f(x+5h)$
$2hf'(x)$	-3	4	-1			
$h^2f''(x)$	2	-5	4	-1		
$2h^3f'''(x)$	-5	18	-24	14	-3	
$h^4f^{(4)}(x)$	3	-14	26	-24	11	-2

Tabla 9.4: Coeficientes de la segunda aproximación por diferencias hacia adelante.

9.1.3. Segunda aproximación por diferencias no centrales

Las aproximaciones finitas cuyo error es de orden h, no son muy populares, ya que se prefiere utilizar expresiones cuyo error sea de orden h^2. Para obtener fórmulas de aproximación no central, deben utilizarse una mayor cantidad de términos de la serie de Taylor. Como ejemplo, se derivará la expresión para $f'(x)$. Partiendo de las ecuaciones (9.2) y (9.4):

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{(4)}(x) + \cdots$$
$$f(x+2h) = f(x) + 2hf'(x) + 2h^2f''(x) + \frac{4h^3}{3}f'''(x) + \frac{2h^4}{3}f^{(4)}(x) + \cdots$$

Ahora, se eliminará $f''(x)$ al multiplicar la primera ecuación por 4 y restando de la segunda ecuación. El resultado es:

$$f(x+2h) - 4f(x+h) = -3f(x) - 2hf'(x) + \frac{2h^2}{3}f'''(x) + \cdots$$

Así:

$$f'(x) = \frac{-f(x+2h) + 4f(x+h) - 3f(x)}{2h} + \frac{h^2}{3}f'''(x) + \cdots$$

o mejor aún:

$$f'(x) = \frac{-f(x+2h) + 4f(x+h) - 3f(x)}{2h} + \mathcal{O}(h^2)$$

La ecuación anterior se denomina **segunda aproximación por diferencias hacia adelante**.

La deducción de las fórmulas de diferencias finitas para derivadas de orden superior, requieren términos adicionales en la serie de Taylor. Así, la aproximación por diferencias hacia adelante para $f''(x)$ utiliza expansiones de $f(x+h)$, $f(x+2h)$ y $f(x+3h)$; la aproximación para $f'''(x)$ depende de $f(x+h)$, $f(x+2h)$, $f(x+3h)$ y $f(x+4h)$, a modo de ejemplo. Los resultados para ambas aproximaciones (hacia adelante y hacia atrás), se resumen en las tablas 9.4 y 9.5.

Comandos de EMT. *El comando para derivar numéricamente una función es:*

	$f(x-5h)$	$f(x-4h)$	$f(x-3h)$	$f(x-2h)$	$f(x-h)$	$f(x)$
$2hf'(x)$				1	-4	3
$h^2f''(x)$			-1	4	-5	2
$2h^3f'''(x)$		3	-14	24	-18	5
$h^4f^{(4)}(x)$	-2	11	-24	26	-14	3

Tabla 9.5: Coeficientes de la segunda aproximación por diferencias hacia atrás.

- `diff(f$:string, x:número, n:natural, h:número)`, *donde* ***f$*** *es la función de* x *a derivar, expresada como* string*;* ***x*** *es la abscisa en la cual se derivará;* ***n*** *es el orden de la derivada, debe ser menor que 6;* ***h*** *es un parámetro optativo y permite seleccionar el incremento a utilizar en las fórmulas.*

Ejemplo en EMT 19. *Derivar numéricamente, en orden 1 y 2, la función:*

$$f(x) = \cos\left(e^{x/6+1} + 2\right)$$

en $x_=$2,5.

```
>function f(x):=cos(exp(x/6+1)+2)
>diff("f(x)",2.5)
 0.109373763622
>diff("f(x)",2.5,2)
 -0.448030253124
```

9.2. Extrapolación de Richardson

De la ecuación (9.6), se puede conseguir la expresión:

$$f'(x) = \frac{1}{2h}\left[f(x+h) - f(x-h)\right] + a_2h^2 + a_4h^4 + \ldots, \tag{9.10}$$

donde las constantes $a_2, a_4, \ldots$ dependen de f y de x. Cuando esta información está disponible sobre un proceso numérico, entonces es posible aplicar la **extrapolación de Richardson** para conseguir mayor precisión. Es decir que la extrapolación de Richardson se utiliza para generar resultados de gran precisión con el uso de fórmulas de bajo orden. Puede ser aplicada cada vez que se conozca la expresión del error de una técnica de aproximación, dependiente de un parámetro (habitualmente la longitud de paso h).

Para el caso particular de la derivada primera, se define la función:

$$\phi(h) = \frac{1}{2h}\left[f(x+h) - f(x-h)\right], \tag{9.11}$$

donde f y x están bien definidos. Se puede ver que $\phi(x)$ es una aproximación a $f'(x)$ con un error de $\mathcal{O}(h^2)$. Siguiendo las ideas del cálculo diferencial, se tratará de computar $\lim_{h\to 0}\phi(h)$ para conseguir una aproximación de mejor calidad que una simple aproximación discreta. Sin embargo, no es posible utilizar $h = 0$ en la ecuación (9.11). Pero, si se computa $\phi(h)$ para algún h, puede computarse $\phi(h/2)$. A partir de la ecuación (9.10) se tiene:

$$\begin{aligned}\phi(h) &= f'(x) - a_2h^2 - a_4h^4 - a_6h^6 - \ldots\\ \phi\left(\frac{h}{2}\right) &= f'(x) - a_2\left(\frac{h}{2}\right)^2 - a_4\left(\frac{h}{2}\right)^4 - a_6\left(\frac{h}{2}\right)^6 - \ldots.\end{aligned}$$

Pero el término dominante del error puede ser eliminado fácilmente: restando la primera expresión de la segunda expresión multiplicada por 4. Esto resulta en:

$$\phi(h) - 4\phi\left(\frac{h}{2}\right) = -3f'(x) - \frac{3}{4}a_4h^4 - \frac{15}{16}a_6h^6 - \dots.$$

Dividiendo toda la expresión por -3 y reacomodando términos:

$$\phi\left(\frac{h}{2}\right) + \frac{1}{3}\left[\phi\left(\frac{h}{2}\right) - \phi(h)\right] = f'(x) + \frac{1}{4}a_4h^4 + \frac{5}{16}a_6h^6 + \dots.$$

El resultado anterior es muy importante, puesto que, al agregar $\frac{1}{3}[\phi(h/2) - \phi(h)]$ a $\phi(h/2)$, se mejoró la aproximación inicial a la derivada primera hasta un orden de $\mathcal{O}(h^4)$. Como h es pequeño, es una mejora de muy buena calidad. Pero este proceso puede, teóricamente, repetirse. Definiendo:

$$\Phi(h) = \frac{4}{3}\phi\left(\frac{h}{2}\right) - \frac{1}{3}\phi(h),$$

y aplicando la misma idea que en las expresiones anteriores:

$$\Phi(h) = f'(x) + b_4h^4 + b_6h^6 + \dots$$
$$\Phi\left(\frac{h}{2}\right) = f'(x) + b_4\left(\frac{h}{2}\right)^4 + b_6\left(\frac{h}{2}\right)^6 + \dots.$$

Combinando convenientemente ambas expresiones para eliminar el término de orden superior del error:

$$\Phi(h) - 16\Phi\left(\frac{h}{2}\right) = -15f'(x) + \frac{3}{4}b_6h^6 + \dots.$$

Por lo tanto, se obtiene:

$$\Phi\left(\frac{h}{2}\right) + \frac{1}{15}\left[\Phi\left(\frac{h}{2}\right) - \Phi(h)\right] = f'(h) - \frac{1}{20}b_6h^6 + \dots,$$

logrando una mejora de calidad. El error de truncamiento para esta expresión de la derivada es del orden de $\mathcal{O}(h^6)$. Este proceso, llamado **extrapolación de Richardson** puede ser aplicado en secuencia para ir eliminando los términos de orden superior en la expresión del error.

En forma general, para un cálculo F cuyo valor depende de h (un parámetro pequeño) y se conoce la expresión completa del error, se tiene que para $h_1 > h_2$ donde $h_2 = h_1/k$:

$$\phi(h_1) = F - a_\alpha h_1^\alpha - a_\beta h_1^\beta - a_\gamma h_1^\gamma - \dots$$
$$\phi(h_2) = F - a_\alpha h_2^\alpha - a_\beta h_2^\beta - a_\gamma h_2^\gamma - \dots,$$

pero $\phi(h_2)$ puede ser escrita como:

$$\phi\left(\frac{h_1}{k}\right) = F - a_\alpha\left(\frac{h_1}{k}\right)^\alpha - a_\beta\left(\frac{h_1}{k}\right)^\beta - a_\gamma\left(\frac{h_1}{k}\right)^\gamma - \dots,$$

o mejor aún:

$$k^\alpha\phi\left(\frac{h_1}{k}\right) = k^\alpha F - a_\alpha\left(h_1\right)^\alpha - a_\beta k^\alpha\left(\frac{h_1}{k}\right)^\beta - a_\gamma k^\alpha\left(\frac{h_1}{k}\right)^\gamma - \dots,$$

con lo que restando $k^\alpha \phi\left(\frac{h_1}{k}\right) - \phi(h_1)$ se obtiene:

$$k^\alpha \phi\left(\frac{h_1}{k}\right) - \phi(h_1) = \left(k^\alpha - 1\right) F - b_\beta \left(\frac{h_1}{k}\right)^\beta - b_\gamma \left(\frac{h_1}{k}\right)^\gamma - \dots$$

$$\frac{k^\alpha \phi\left(\frac{h_1}{k}\right) - \phi(h_1)}{k^\alpha - 1} = F - c_\beta \left(\frac{h_1}{k}\right)^\beta - c_\gamma \left(\frac{h_1}{k}\right)^\gamma - \dots .$$

Es decir que se obtiene una aproximación a F de orden $\mathcal{O}\left((h_1/k)\right)^\beta$.

Ejemplo 57. *Se desea calcular* $f'(2{,}8)$ *de* $f(x) = \cos(x)e^x$. *Para ello se elige trabajar con PF(10,8,2), la primera aproximación por diferencias hacia adelante y* $h_1 = 0{,}2$, $h_2 = 0{,}1$. *Entonces* $k = 2$ *y:*

$$\begin{aligned}
\phi(h_1) &= f'(2{,}8) + \mathcal{O}(h_1) \\
&= \frac{f(2{,}8 + 0{,}2) - f(2{,}8)}{0{,}2} + \mathcal{O}(h_1) \\
&= \frac{-19{,}884531 - (-15{,}494514)}{0{,}2} + \mathcal{O}(h_1) \\
&= -21{,}950085 + \mathcal{O}(h_1)
\end{aligned}$$

y

$$\begin{aligned}
\phi(h_2) &= f'(2{,}8) + \mathcal{O}(h_2) \\
&= \frac{f(2{,}8 + 0{,}1) - f(2{,}8)}{0{,}1} + \mathcal{O}(h_2) \\
&= \frac{-17{,}646335 - (-15{,}494514)}{0{,}1} + \mathcal{O}(h_2) \\
&= -21{,}518210 + \mathcal{O}(h_2).
\end{aligned}$$

Ahora:

$$\begin{aligned}
\frac{k^\alpha \phi\left(\frac{h_1}{k}\right) - \phi(h_1)}{k^\alpha - 1} &= f'(2{,}8) + \mathcal{O}\left(h_2^2\right) \\
\frac{2^1 \phi(h_2) - \phi(h_1)}{2^1 - 1} &= f'(2{,}8) + \mathcal{O}\left(h_2^2\right) \\
-21{,}086335 &= f'(2{,}8) + \mathcal{O}\left(h_2^2\right)
\end{aligned}$$

lo que mejora notablemente las aproximaciones obtenidas con $\phi(h_1)$ *y* $\phi(h_2)$.

Ejemplo 58. *A los datos del ejemplo anterior se agrega* $h_3 = 0{,}06$. *Entonces:*

$$\begin{aligned}
\phi(h_3) &= f'(2{,}8) + \mathcal{O}\left(h_3\right) \\
&= \frac{f(2{,}8 + 0{,}06) - f(2{,}8)}{0{,}06} + \mathcal{O}\left(h_3\right) \\
&= \frac{-16{,}773789 - (-15{,}494514)}{0{,}06} + \mathcal{O}\left(h_3\right) \\
&= -21{,}321250 + \mathcal{O}\left(h_3\right),
\end{aligned}$$

con lo que se puede operar a partir de $\phi(h_2)$ *y* $\phi(h_3)$*, con* $k = 5/3$*:*

$$\begin{aligned}
\frac{k^{\alpha}\phi\left(\frac{h_2}{k}\right) - \phi(h_2)}{k^{\alpha} - 1} &= f'(2{,}8) + \mathcal{O}\left(h_3^2\right) \\
\frac{\left(\frac{5}{3}\right)^{1}\phi(h_3) - \phi(h_2)}{\left(\frac{5}{3}\right)^{1} - 1} &= f'(2{,}8) + \mathcal{O}\left(h_3^2\right) \\
-21{,}025810 &= f'(2{,}8) + \mathcal{O}\left(h_3^2\right)
\end{aligned}$$

Sin embargo, se pueden volver a extrapolar las aproximaciones de órdenes $\mathcal{O}\left(h_2^2\right)$ *y* $\mathcal{O}\left(h_3^2\right)$*, tomando nuevamente* $k = 5/3$*:*

$$\begin{aligned}
\frac{k^{\beta}\phi\left(\left[\frac{h_2}{k}\right]^2\right) - \phi(h_2^2)}{k^{\alpha} - 1} &= f'(2{,}8) + \mathcal{O}\left(h_3^3\right) \\
\frac{\left(\frac{5}{3}\right)^{2}\phi\left(h_3^2\right) - \phi(h_2^2)}{\left(\frac{5}{3}\right)^{2} - 1} &= f'(2{,}8) + \mathcal{O}\left(h_3^3\right) \\
-20{,}991764 &= f'(2{,}8) + \mathcal{O}\left(h_3^3\right)
\end{aligned}$$

con lo que $f'(2{,}8) = -20{,}991764$*.*

Nota. *En los ejemplos anteriores, se desarrolla el orden del error dependiendo de* h_1*,* h_2 *y* h_3*. Sólo se hizo con fines ilustrativos ya que el orden depende de* h *y sus potencias.*

Nota. *Si bien es un proceso que puede repetirse indefinidamente, debe tenerse cuidado con la aritmética utilizada, ya que puede agregar más error al cálculo debido al error de redondeo.*

Ejercicio 30. *Rehacer el ejemplo anterior, pero esta vez utilizar la primera aproximación por diferencias centrales.*

9.2.1. Expresiones de derivadas por extrapolación

La extrapolación no sólo permite mejorar la precisión de cálculo sobre magnitudes puntuales, sino que también es posible mejorar las fórmulas dadas de derivación por medio de transformaciones sencillas.

Utilizando la ecuación (9.9) es posible definir:

$$\begin{aligned}
\phi(2h) &= \frac{f(x+2h) - f(x-2h)}{4h} + \mathcal{O}(h^2) \\
\phi(2h) &= f'(x) + \mathcal{O}(h^2),
\end{aligned}$$

y

$$\begin{aligned}
\phi(h) &= \frac{f(x+h) - f(x-h)}{2h} + \mathcal{O}(h^2) \\
\phi(h) &= f'(x) + \mathcal{O}(h^2),
\end{aligned}$$

entonces, aplicando extrapolación con $k = 2h/h = 2$:

$$\begin{aligned}
\frac{2^2\phi(h) - \phi(2h)}{2^2 - 1} &= \frac{4\left[\frac{f(x+h) - f(x-h)}{2h}\right] - \left[\frac{f(x+2h) - f(x-2h)}{4h}\right]}{4 - 1} + \mathcal{O}(h^4) \\
&= \frac{8f(x+h) - 8f(x-h) - f(x+2h) + f(x-2h)}{12h} + \mathcal{O}(h^4) \\
&= f'(x) + \mathcal{O}(h^4),
\end{aligned}$$

con lo que se consigue una aproximación a $f'(x)$ del orden de $\mathcal{O}(h^4)$.

9.3. Errores en las aproximaciones finitas

Hasta ahora las fórmulas de derivación por cálculos discretos sólo dependen del paso elegido: h, por lo que parece simple reducir h hasta un valor mínimo y conseguir aproximaciones de calidad. Sin embargo, esta consideración es sólo una parte del error involucrado en el desarrollo, ya que se tiene en cuenta el error de truncamiento de la serie de Taylor pero no el que se genera por el redondeo aritmético.

Con el fin de analizar en detalle la aproximación de la derivada primera, se supondrá que las siguientes cantidades:

$$f(x-2h), \quad f(x-h), \quad f(x), \quad f(x+h), \quad f(x+2h),$$

contienen el error de redondeo aritmético:

$$y_{-2} = f(x-2h) + e_{-2}, \quad y_{-1} = f(x-h) + e_{-1}, \quad y_0 = f(x) + e_0,$$

$$y_1 = f(x+h) + e_1, \quad y_2 = f(x+2h) + e_2,$$

donde las magnitudes de los errores de redondeo e_i son más pequeños que cierto número positivo ε. Entonces, el error total de la aproximación por diferencia hacia adelante es:

$$\begin{aligned} D_f(x,h) &= \frac{y_1 - y_0}{h} \\ &= \frac{f(x+h) + e_1 - f(x) - e_0}{h} \\ &= f'(x) + \frac{e_1 - e_0}{h} + \frac{K_1}{2}h, \end{aligned}$$

entonces:

$$\left|D_f(x,h) - f'(x)\right| \leq \left|\frac{e_1 - e_0}{h}\right| + \frac{|K_1|}{2}h \leq \frac{2\varepsilon}{h} + \frac{|K_1|}{2}h,$$

con $K_1 = f''(x)$. En la segunda parte de la inecuación anterior se presenta el error: la primera expresión corresponde al redondeo aritmético (inversamente proporcional a h) y la segunda corresponde al truncamiento de la serie de Taylor (directamente proporcional a h). Es posible minimizar la expresión del error por medio de un cálculo sencillo:

$$\frac{d}{dh}\left(\frac{2\varepsilon}{h} + \frac{|K_1|}{2}h\right) = -\frac{2\varepsilon}{h^2} + \frac{|K_1|}{2} = 0, \qquad h_{op} = 2\sqrt{\frac{\varepsilon}{|K_1|}}.$$

De forma similar, el error total en la aproximación por diferencias centrales es:

$$\begin{aligned} D_c(x,h) &= \frac{y_1 - y_{-1}}{2h} \\ &= \frac{f(x+h) + e_1 - f(x-h) - e_{-1}}{2h} \\ &= f'(x) + \frac{e_1 - e_{-1}}{2h} + \frac{K_2}{6}h^2, \end{aligned}$$

con lo que:

$$\left|D_c(x,h) - f'(x)\right| \leq \left|\frac{e_1 - e_{-1}}{2h}\right| + \frac{|K_2|}{6}h^2 \leq \frac{2\varepsilon}{2h} + \frac{|K_2|}{6}h^2,$$

con $K_2 = f'''(x)$. Al minimizar la expresión del error, se consigue el valor óptimo de h:

$$\frac{d}{dh}\left(\frac{\varepsilon}{h} + \frac{|K_2|}{6}h^2\right) = -\frac{\varepsilon}{h^2} + \frac{|K_2|}{3}h = 0, \qquad h_{op} = \sqrt[3]{\frac{3\varepsilon}{|K_2|}}.$$

h	PF(10,6,2)	PF(10,8,2)
0,64	0,380610	0,38060911
0,32	0,371035	0,37102939
0,16	0,368711	0,36866484
0,08	0,368281	0,36807656
0,04	0,36875	0,36783125
0,02	0,37	0,3679
0,01	0,38	0,3679
0,005	0,40	0,3676
0,0025	0,48	0,3680
0,00125	1,28	0,3712

Tabla 9.6: Cálculo de la derivada segunda de $f(x) = e^{-x}$ en $x = 1$, para diferentes valores de h y dos tipos de aritmética.

Ejercicio 31. *Deducir la expresión completa del error para el caso de la aproximación de la derivada primera por diferencias hacia atrás.*

Por lo visto hasta ahora, a medida que el paso h decrece el error de redondeo puede incrementarse, mientras que el de truncamiento decae. Este es el **dilema de la longitud de paso**. Si bien se plantean expresiones para calcular el valor óptimo de h, son poco realistas y de valor sólo teórico, puesto que habitualmente se carece de información sobre $K_1, K_2, \ldots$, es decir sobre las derivadas de orden superior. Además, hay que notar que h_{op} minimiza no el error real, sino su límite superior.

Ejemplo 59. *Se desea calcular la derivada segunda de $f(x) = e^{-x}$ en $x = 1$ a partir de la fórmula de diferencias centrales. Para ello se utilizará una aritmética de 6 y 8 dígitos de precisión, utilizando diferentes valores de h. Los resultados se muestran en la tabla 9.6. El valor exacto es $f''(1) = e^{-1} = 0{,}36787944$. En los cálculos con PF(10,6,2) y de acuerdo a la tabla 9.6, el valor óptimo para h es 0,08, dando un resultado exacto hasta el tercer dígito significativo. Tres de los dígitos significativos se perdieron debido a una combinación de truncamiento y error de redondeo. Por encima del h óptimo, el error dominante es que surge por truncamiento, por debajo prima el error debido al redondeo. En los cálculos con PF(10,8,2), el mejor resultado obtenido en la tabla 9.6 es exacto hasta el cuarto dígito significativo. Como al aumentar la aritmética, decrece el error de redondeo, el h óptimo es más pequeño que en PF(10,6,2).*

9.4. Ejercicios

1. Construir los siguientes algoritmos en PC:

   a) **Primera aproximación por diferencias centrales**. Entrada: Tabla de valores cuyo dato central sea el valor sobre el cual se calcula la derivada. Salida: Aproximación de derivada primera, segunda y tercera.

   b) **Extrapolación de Richardson**. Entrada: Tabla de valores con los datos para extrapolar; orden del error utilizado. Salida: Tabla completa de extrapolación.

2. Utilizando la primera aproximación por diferencias centrales, calcular la derivada de las funciones $f_i(x)$, en los puntos indicados. Utilizar $h = 0{,}5$; $h = 0{,}2$; $h = 0{,}15$ y $h = 0{,}05$:

a) $f_1(x) = \log(x)$, definida en $[2;6]$, $x_0 = 4$.
b) $f_2(x) = \sin(x) + \cos(x)$, definida en $[1;3]$, $x_0 = 2$.

3. Aproximar el valor de la derivada primera por aproximación no central en las funciones e abscisas indicadas. Utilizar los valores de h dados en el ejercicio anterior:

   a) $f_3(x) = e^{-2x} + 4x$, definida en $[0;4]$, $x_0 = 4$.
   b) $f_4(x) = \ln(x)$, definida en $[1;5]$, $x_0 = 1$.

4. Para las funciones $f_i(x)$, $i = 1, \ldots, 4$ de los dos ejercicios anteriores, calcular la derivada segunda en las abscisas indicadas.

5. Dada una tabla de valores, ¿es útil construir un *spline* cúbico para obtener la derivada en uno de los puntos de la tabla? ¿Y el *spline* de Hermite?

6. Dadas las tablas 9.7 a 9.10 como datos, calcular la derivada primera en las abscisas indicadas.

x	0.7000	0.9000	1.1000	1.3000	1.5000	1.7000	1.9000
$f(x)$	2.6445	2.3148	1.9108	1.4525	0.9622	0.4635	-0.0199

Tabla 9.7: Datos tabulados de $f(x) = 3\cos(x) + x/2$. Calcular $f'(1{,}3)$.

7. Calcular $f''(x_0)$ para las tablas del inciso anterior.

8. Mejorar las aproximaciones obtenidas en los incisos 2, 3 y 6 utilizando extrapolación de Richardson.

9. A partir de la extrapolación de Richardson, generar una mejor fórmula para la expresión de la derivada segunda utilizando como base la fórmula de la primera aproximación por diferencias centrales.

10. *EMT*, de acuerdo a la aritmética de IEEE, utiliza un valor de ϵ_M del orden de 10^{-16}. Comprobarlo con el siguiente código:

```
>i=1:14;
>x=(sin(pi/3.2+10^(-i))-sin(pi/3.2))/10^(-i);
>y=x[i]-cos(pi/3.2);
>x'|y'
```

    que muestra los diferentes valores aproximados de la derivada primera de $\sin(x)$ en $x_0 = \pi/3{,}2$ por diferencia hacia adelante. Suponiendo que, para este caso, $K_1 = 0{,}8$: ¿qué valor aproximado toma ϵ_M?

11. [*EMT*] Para la función $f(x) = \frac{\cos(x)}{x}$, crear particiones de 10, 20, 30 y 40 elementos en el intervalo $[1;7]$. Graficar las derivadas primera y segunda calculadas con los algoritmos de diferencias centrales sobre los datos de las particiones. Además, en cada caso, graficar la derivada correcta.

x	-1.4000	-0.8000	-0.2000	0.4000	1.0000	1.6000
$f(x)$	17.3200	5.6800	-0.9200	-2.4800	1.0000	9.5200

Tabla 9.8: Datos tabulados de $g(x) = -4x + 7x^2 - 2$. Calcular $f'(-0{,}8)$.

Tiempo (s)	0	4	8	12	16	20
Altura (km)	0	0,84	3,53	8,41	15,97	27,00
Velocidad (km/s)						
Aceleración (km/s^2)						

Tabla 9.11

18. Los datos telemétricos de un cohete en ascenso vertical se muestran en la tabla 9.11. Completar los datos faltantes y estimar aproximadamente en qué momento ocurre la segunda etapa del lanzamiento.

19. [EMT] Considerar la fórmula de aproximación:

$$f'(x) \approx \frac{3}{2h^3}\int_{-h}^{h} tf\,(x+t)\,\mathrm{d}t.$$

Establecer el orden del error en forma similar a lo realizado en el ejercicio 16 con alguna función conveniente. Esta es la **derivada generalizada de Lanczos**.

20. Cierto cálculo requiere una fórmula de aproximación de $f'(x)+f''(x)$. El esquema:

$$\left(\frac{2+h}{2h^2}\right)f(x+h)-\left(\frac{2}{h^2}f(x)\right)+\left(\frac{2-h}{2h^2}\right)f(x-h),$$

a) ¿Sirve para este propósito?

b) ¿Qué orden de error tiene?

c) Mejorar el orden del error, si es posible, con otro esquema similar.

21. ¿Es posible derivar la fórmula:

$$f'(x) \approx \frac{4f(x+h)-3f(x)-f(x-2h)}{6h}$$

utilizando el teorema de Taylor? ¿Qué orden de convergencia se logra con esta aproximación?

22. Deducir la fórmula:

$$f''(x) \approx \frac{2}{h^2}\left[\frac{f(x_0)}{(1+\alpha)}-\frac{f(x_1)}{\alpha}+\frac{f(x_2)}{\alpha(1+\alpha)}\right],$$

que está preparada para $x_0 < x_1 < x_2$ donde la partición no es equiespaciada, es decir que $x_1 - x_0 = h$ y $x_2 - x_1 = \alpha h$. Para ello calcular los coeficientes A, B y C de la expresión:

$$f''(x) \approx Af(x_0)+Bf(x_1)+Cf(x_2)$$

resolviendo el sistema que se genera al suponerla exacta para los tres polinomios 1, $x - x_1$ y $(x - x_1)^2$ y por lo tanto exacta para todos los polinomios de grado menor o igual que 2. Este es el método de **coeficientes indeterminados**.

23. Calcular la derivada de $f(x) = -2x^2 + 3x + 1$ en $x = -2$ utilizando la fórmula del ejercicio 19. ¿Qué error se obtiene?

24. Calcular la derivada de $f(x) = x^3 + 2x - 1$ en $x = -2$ utilizando la fórmula del ejercicio 19. ¿Qué error se obtiene?

25. Para las aproximaciones de $f'(x)$ dadas a continuación, calcular el orden de convergencia con respecto a h:

$a)$ $\dfrac{-11f(x)+18f(x+h)-9f(x+2h)+2f(x+3h)}{6h}$

$b)$ $\dfrac{f(x-2h)-6f(x-h)+3f(x)+2f(x+h)}{6h}$

$c)$ $\dfrac{-f(x-2h)-12f(x)+16f(x+h)-3f(x+2h)}{12h}$

Bibliografía

- *Análisis numérico - Un enfoque práctico*, M. MARON y R. LÓPEZ, Cap.7
- *Fundamental numerical methods for electrical engineering**, Stanislaw ROSLONIEC, Cap.6
- *Métodos numéricos con MATLAB*, J. MATHEWS y K. FINK, Cap.6

10 Integración Numérica

En este capítulo se tratará de resolver el problema de la integral definida de $f(x)$, dada en un intervalo cerrado $[a, b]$, es decir:

$$\int_a^b f(x)\,\mathrm{d}x.$$

Esta integral puede ser calculada fácilmente siempre que el integrando sea una función integrable $f(x)$, acotada y continua sobre el intervalo y cuando la función primitiva $F(x)$ sea conocida y cumpla que $F'(x) = f(x)$. En este caso especial, sólo se debe aplicar la regla de Barrow:

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a).$$

Sin embargo, en otros casos es imposible (o excesivamente costoso) determinar la función primitiva $F(x)$. Por ejemplo, si se tiene una función discreta $y_i = f(x_i)$, para $i = 0, 1, 2, \ldots, n$ no se puede crear una primitiva discreta. Este es el momento en que los métodos numéricos aparecen, para aproximar el valor de una integral definida.

Los métodos numéricos para calcular integrales definidas se separarán en tres grupos, sólo con la finalidad de organizar mejor el texto:

- **Integrandos expansibles por series**, aquí se reemplazará el integrando $f(x)$, por la expansión de una serie de funciones elementales que se pueden integrar fácilmente. El permitir varios términos de la serie (funciones elementales) permite ajustar la precisión del cálculo final.
- **Métodos de paso finito**, este grupo contiene a aquellos métodos finitos en los cuales se subdivide el intervalo de integración $[a, b]$ en n partes para luego aproximar cada una de las partes en forma geométrica.
- **Métodos de cuadratura**, son los métodos también conocidos como *Gaussianos*. Es decir que se reemplazará el integrando por polinomios interpoladores y/o ortogonales (Legendre, Jacobi o Chebyshev). La principal complicación es determinar los nodos a utilizar y los valores de los pesos aplicados a cada función.

10.1. Integrandos expansibles por series

En este caso, el integrando $f(x)$ definido sobre un intervalo $[a, b]$ puede ser representado por una serie de funciones elementales, fácilmente integrables en el sentido de los

métodos analíticos. Generalmente, las expansiones más comunes son las *series de Taylor* y las *series de Fourier*. El análisis del error es muy sencillo, ya que la componente principal es aportada por error de truncamiento.

Ejemplo 60. *En análisis espectral de imágenes, es necesario calcular la siguiente integral:*

$$Si(x) = \int_0^x \frac{\sin(t)}{t}\,\mathrm{d}t, \tag{10.1}$$

que no tiene primitiva analítica. Sin embargo, es útil conocer la serie:

$$\frac{\sin(t)}{t} = \frac{1}{t}\left[t - \frac{t^3}{3!} + \frac{t^5}{5!} - \frac{t^7}{7!} + \ldots + (-1)^n \frac{t^{2n+1}}{(2x+1)!} + \ldots\right]$$

que es convergente en todo valor de t. Luego de reemplazar el integrando de (10.1) por su expansión en serie, se obtiene

$$\begin{aligned} Si(x) &= x - \frac{x^3}{3!3} + \frac{x^5}{5!5} - \frac{x^7}{7!7} + \ldots + (-1)^n \frac{x^{2n+1}}{(2n+1)!(2n+1)} + \ldots \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!(2n+1)} \end{aligned}$$

Esta expansión en series permite calcular la función $Si(x)$ con la precisión que se desee. Por ejemplo, para $x = 2$ y $n = 6$,

$$\begin{aligned} Si(2) &= 2 - \frac{8}{18} + \frac{32}{600} - \frac{128}{35280} + \frac{512}{3265920} - \frac{2048}{439084800} + \sum_{n=6}^{\infty} (-1)^n \frac{2^{2n+1}}{(2n+1)!(2n+1)} \\ &\approx 1{,}605412 \end{aligned}$$

El séptimo término de la serie, no considerado en los cálculos anteriores, toma el valor $8{,}192/(8{,}905 \times 10^7)$, que es mucho más pequeño que 10^{-6}.

Ejercicio 32. *Utilizando* series de Taylor*, aproximar la integral definida*

$$\frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\,\mathrm{d}t$$

y acotar su error, utilizando $x = 2$ y $n = 4$.

10.2. Métodos de paso finito

En este caso, dado un integrando $f(x)$ definido en un intervalo cerrado $[a, b]$, se puede dividir el intervalo en n segmentos diferentes (subintervalos) $\Delta x_i = x_i - x_{i-1}$, para $i = 1, 2, \ldots, n$. El soporte teórico para los métodos que se presentarán se basa en el teorema por el cual se define la integral definida, la *suma de Riemann*. Su versión finita es:

$$S = S_1 + S_2 + S_3 + \ldots + S_n = \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} f(\xi_i)\Delta x_i$$

donde ξ_i es un valor arbitrario de la variable x, tomado en cada subintervalo i, es decir, $x_{i-1} \leq \xi_i \leq x_i$.

En la figura 10.1 se muestra una función arbitraria $f(x)$ y una partición finita (n partes) del intervalo de integración.

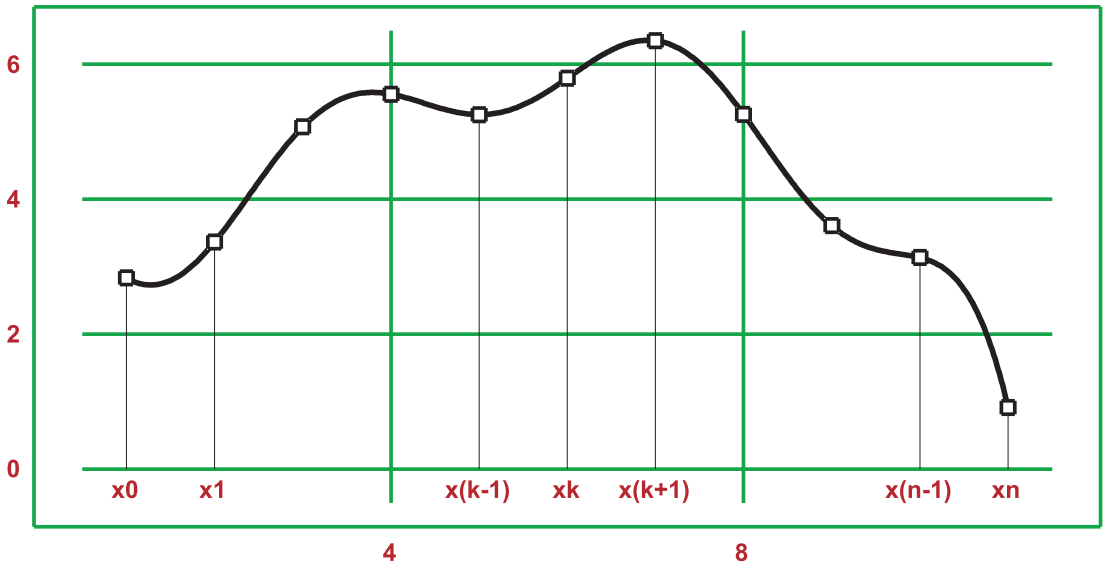

Figura 10.1: Partición del intervalo $[a, b]$ en n segmentos.

10.2.1. Método del rectángulo

Es el más simple de los métodos finitos y también el más intuitivo. Consiste en dividir el intervalo de integración $[a, b]$ en n subintervalos utilizando una partición de $n+1$ puntos. Existen tres formas de aproximar la integral definida sumando el área de rectángulos, siempre dependiendo de la elección de la altura del rectángulo, ya que la base es siempre $h_i = x_i - x_{i-1}$, para $i = 1, 2, \ldots, n$:

- La altura del *i-ésimo* rectángulo es $f(x_{i-1})$, para $i = 1, 2, \ldots, n$. En este caso:

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n} (x_i - x_{i-1}) f(x_{i-1}) + E(h).$$

  Un esquema se observa en la figura 10.2.

- La altura del *i-ésimo* rectángulo es $f(x_i)$, para $i = 1, 2, \ldots, n$. Esto lleva a:

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n} (x_i - x_{i-1}) f(x_i) + E(h).$$

  Este esquema se muestra en la figura 10.3.

- La altura del *i-ésimo* rectángulo se calcula como $f(\frac{x_{i-1}+x_i}{2})$, para $i = 1, 2, \ldots, n$. Entonces:

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n} (x_i - x_{i-1}) f\left(\frac{x_{i-1} + x_i}{2}\right) + E(h).$$

  El esquema correspondiente es la figura 10.4.

El límite común de las tres expresiones anteriores se da cuando máx$\{x_i - x_{i-1}\} \to 0$ y es igual al valor de la integral resuelta por métodos algebraicos, es decir el valor exacto. Si se quiere acotar el error debe tenerse en cuenta que:

$$\frac{\mathrm{d}^n F}{\mathrm{d}x^n} = \frac{\mathrm{d}^{n-1} f}{\mathrm{d}x^{n-1}}, \tag{10.2}$$

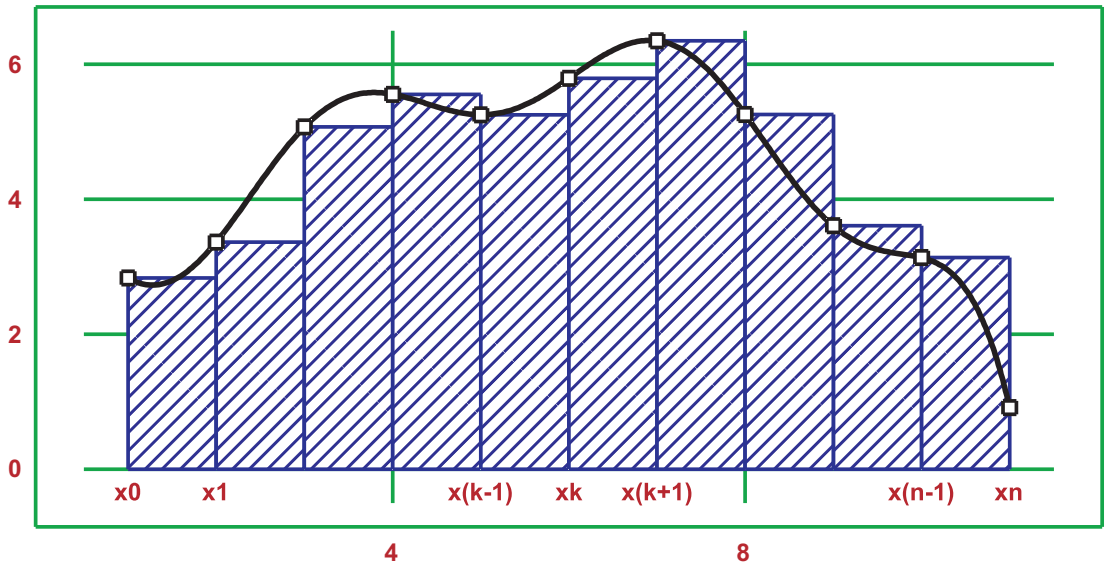

Figura 10.2: Integración por aproximación de rectángulos, altura $f(x_{i-1})$.

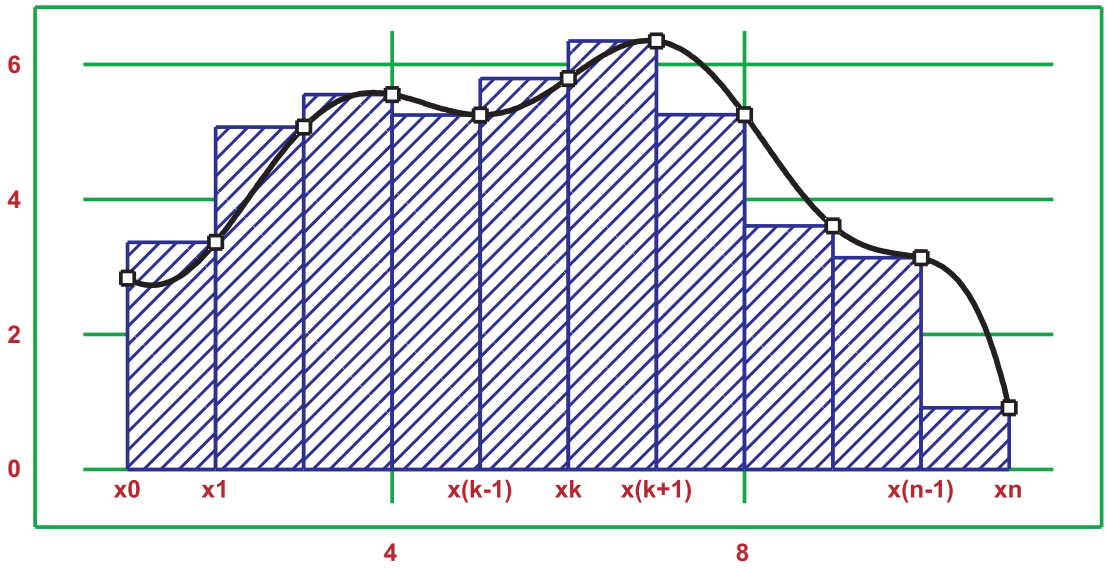

Figura 10.3: Integración por aproximación de rectángulos, altura $f(x_i)$.

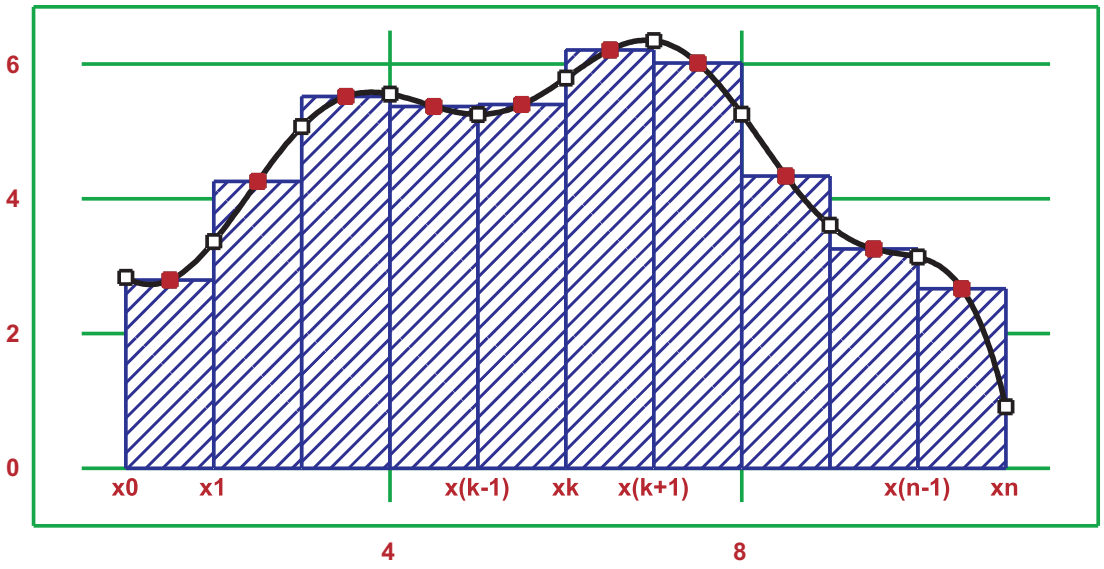

Figura 10.4: Integración por aproximación de rectángulos, altura $f\left(\frac{x_{i-1}+x_i}{2}\right)$.

ya que $F(x)$ es primitiva de $f(x)$. Además, como F y f son funciones continuas y derivables, se las puede expandir por medio de una serie de Taylor:

$$\begin{aligned} F(x+h) &= F(x) + hF'(x) + \tfrac{h^2}{2!}F''(x) + \tfrac{h^3}{3!}F'''(x) + \mathcal{O}(h^4) \\ f(x+h) &= f(x) + hf'(x) + \tfrac{h^2}{2!}f''(x) + \tfrac{h^3}{3!}f'''(x) + \mathcal{O}(h^4), \end{aligned} \tag{10.3}$$

esto permite acotar el error cometido por truncado. Partiendo de la aproximación central, y analizando el área del intervalo $[x_i; x_i + h_i]$:

$$\begin{aligned} E(h_i) &= |I_A - I_R| \\ &= \left| \int_{x_i}^{x_i+h_i} f(x)\,\mathrm{d}x - hf\left(\frac{x_i + (x_i+h_i)}{2}\right)\right| \\ &= \left| F(x_i+h_i) - F(x_i) - h_i f(x_i + \frac{h_i}{2})\right|. \end{aligned}$$

Ahora, aplicando (10.3) y (10.2), queda la siguiente expresión para el error:

$$E(h_i) = \frac{h_i^3}{24}f''(x_i) + \mathcal{O}(h_i^4),$$

con lo que el error cometido por truncamiento es del orden de h^3. Si se quiere acotar el error para todo el proceso de integración, sólo debe sumarse cada error:

$$E(h) = \sum_{i=1}^{n} \frac{h_i^3}{24}f''(x_i) \le \frac{nh^3}{24}f''(\xi), \qquad \xi \in [x_0, x_n].$$

La última expresión sólo es válida si la partición realizada sobre el intervalo de integración es regular.

Ejercicio 33. *Acotar el error para los otros dos métodos de integración finita basada en rectángulos.*

Ejemplo 61. *Utilizando el método del rectángulo (partición central), se desea aproximar el valor de*

$$\int_0^3 [x\cos(x) + 4]\,\mathrm{d}x$$

x_i	0,25	0,75	1,25	1,75	2,25	2,75
$f(x_i)$	4,2422	4,5488	4,3942	3,6881	2,5866	1,4582

Tabla 10.1: Partición regular sobre el intervalo $[0, 3]$. Método del rectángulo.

con una partición regular de 6 rectángulos. Como $n = 6$*, entonces* $h = \frac{3-0}{6} = 0{,}5$*. Para ello deben identificarse los puntos medios de cada subintervalo de la partición generada. Esos datos se muestran en la tabla 10.1 y el gráfico correspondiente es el que se ve en la figura 10.5.*

Figura 10.5: Integración por aproximación de rectángulos, para $f(x) = x\cos(x) + 4$ en $[0, 3]$.

Entonces:

$$\begin{aligned}\int_0^3 [x\cos(x) + 4]\,\mathrm{d}x &= h\sum_{i=1}^{n} f(x_i) + E(h)\\ &= 10{,}459 + E(h),\end{aligned}$$

donde $E(h)$ *depende de la cota de la derivada segunda de* $f(x)$*. Una cota para la función* $f''(x) = -x\cos(x) - 2\sin(x)$ *es* 4*. Por lo tanto:*

$$E(h) = \frac{6 \cdot 0{,}5^3}{24}4 = 0{,}125,$$

lo que se confirma al integrar algebraicamente la función deseada.

10.2.2. Método del trapecio

La esencia del método del trapecio se muestra en la figura 10.6. En este caso, el área bajo la curva se aproxima por medio de trapecios de base $h_i = x_i - x_{i-1}$. Es decir que:

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n} (x_i - x_{i-1})\left(\frac{f(x_{i-1}) + f(x_i)}{2}\right) + E(h)$$

Al igual que en el método de los rectángulos, se acota el error de truncamiento por medio de la expansión en series de Taylor. Para analizar el error cometido en el

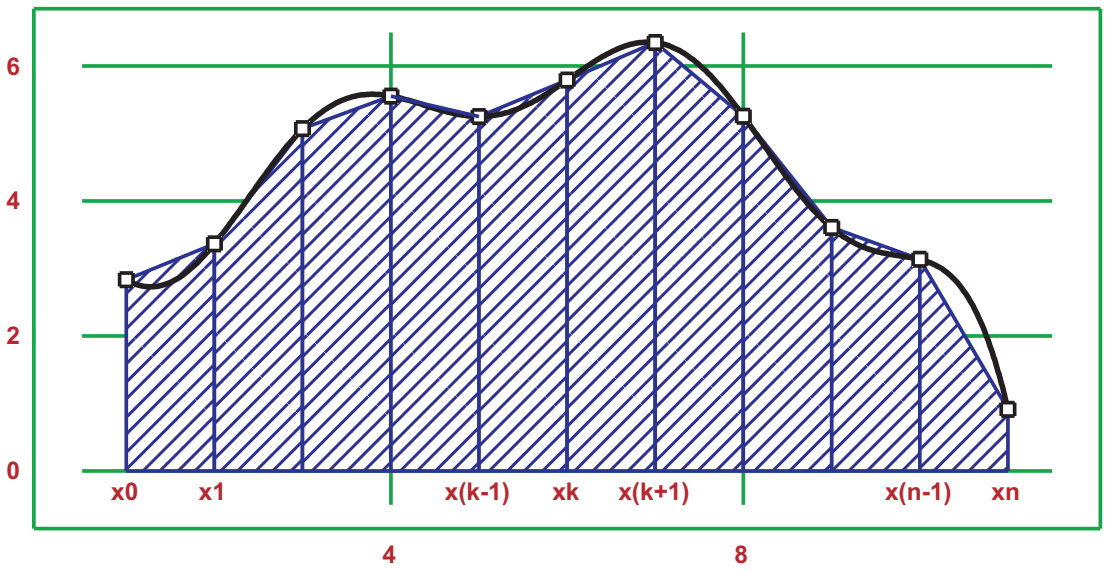

Figura 10.6: Integración por aproximación de trapecios.

intervalo $[x_i, x_i + h_i]$:

$$\begin{aligned} E(h_i) &= |I_A - I_T| \\ &= \left| \int_{x_i}^{x_i+h_i} f(x)\,\mathrm{d}x - \frac{h}{2}\left[f(x_i) + f(x_i+h_i)\right] \right| \\ &= \left| F(x_i+h_i) - F(x_i) - \frac{h}{2}\left[f(x_i) + f(x_i+h_i)\right] \right| \end{aligned}$$

Aplicando (10.3) y (10.2), se obtiene una expresión más simple del error:

$$E(h_i) = \frac{h_i^3}{12} f''(x_i) + \mathcal{O}(h_i^4),$$

lo que muestra que el error debido al truncamiento es del orden de h^3. Para acotar correctamente el error, debe sumarse sobre cada uno de los intervalos involucrados:

$$E(h) = \sum_{i=1}^{n} \frac{h_i^3}{12} f''(x_i) \leq \frac{nh^3}{12} f''(\xi), \qquad \xi \in [x_0, x_n].$$

Al igual que en el método de los rectángulos, la expresión última es válida sólo si la partición generada es regular.

Ejemplo 62. *Utilizando el método del trapecio, se desea aproximar el valor de:*

$$\int_0^3 \left[x\cos(x) + 4\right]\,\mathrm{d}x$$

con una partición regular de 6 trapecios. Como $n = 6$, *entonces* $h = \frac{3-0}{6} = 0{,}5$. *El gráfico correspondiente es el que se ve en la figura 10.7.*

Entonces:

$$\begin{aligned} \int_0^3 \left[x\cos(x) + 4\right]\,\mathrm{d}x &= \frac{h}{2}\sum_{i=1}^{n}\left[f(x_i) + f(x_i+h)\right] + E(h) \\ &= 10{,}383 + E(h), \end{aligned}$$

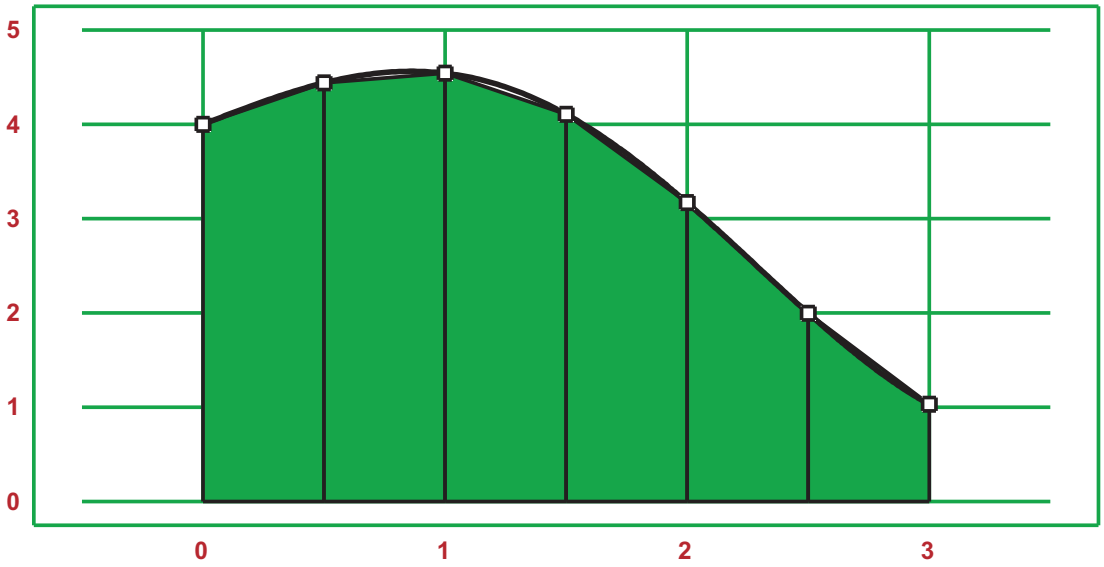

Figura 10.7: Integración por aproximación de trapecios, para $f(x) = x\cos(x)+4$ en $[0,3]$.

donde $E(h)$ depende de la cota de la derivada segunda de $f(x)$. Una cota para la función $f''(x) = -x\cos(x) - 2\sin(x)$ es 4. Por lo tanto:

$$E(h) = \frac{6 \cdot 0{,}5^3}{12} 4 = 0{,}25,$$

lo que se confirma al integrar algebraicamente la función deseada.

Nota. *Contrariamente a lo que se supone al observar los gráficos, el método de integración por rectángulos con partición central tiene menos error por truncamiento que el método de integración por trapecios.*

Ejemplo 63. *Se desea resolver nuevamente el ejercicio anterior. Esta vez utilizando particiones de 2 y 6 trapecios para poder aplicar extrapolación de Richardson. Ahora:*

$$\begin{aligned} \phi(h_1) &= A_{2T} = 9{,}932 + \mathcal{O}(h^3); \qquad h_1 = 1{,}5 \\ \phi(h_2) &= A_{6T} = 10{,}383 + \mathcal{O}(h^3); \qquad h_2 = 0{,}5 \end{aligned}$$

con lo que $k = 3$ y $\alpha = 3$. Además es posible extrapolar extrapolar estos dos valores y obtener un refinamiento superior en el sentido del error de truncamiento:

$$\begin{aligned} \frac{k^\alpha \phi(h_2) - \phi(h_1)}{k^\alpha - 1} &= \int_0^3 [x\cos(x) + 4]\ \mathrm{d}x + \mathcal{O}(h_2^4) \\ \frac{3^3 10{,}383 - 9{,}932}{3^3 - 1} &= \int_0^3 [x\cos(x) + 4]\ \mathrm{d}x + \mathcal{O}(h^4) \\ 10{,}400 &= \int_0^3 [x\cos(x) + 4]\ \mathrm{d}x + \mathcal{O}(h^4) \end{aligned}$$

10.2.3. Método de Simpson

El método del trapecio para aproximar integrales definidas se puede considerar un método de integración basado en el concepto de interpolación, ya que para cada subintervalo se construye un polinomio lineal con el que se aproxima el área bajo la curva de dicho subintervalo. Si se pretende *mejorar* la precisión, es decir reducir el error de truncamiento, es útil considerar polinomios cuadráticos en vez de lineales.

Si se desea integrar la función $f(x)$ en el intervalo $[a, b]$, se debe generar una partición $[x_0, x_1, \ldots, x_n]$ con n par y cada 3 puntos de la partición se puede crear un polinomio interpolador cuadrático. Ahora:

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=1}^{n/2} \int_{x_{2i-2}}^{x_{2i}} P_{2i-1}(x)\,\mathrm{d}x + E(h). \tag{10.4}$$

Para construir los polinomios $P_i(x)$ de grado 2 en forma eficiente, es importante centrarlos en el punto medio de cada subintervalo, simplificando de este modo el desarrollo. Entonces, para identificar el polinomio que pasa por los puntos $(x_{k-1}, f(x_{k-1}))$, $(x_k, f(x_k))$ y $(x_{k+1}, f(x_{k+1}))$ es necesario hallar los coeficientes a_k, b_k y c_k tales que:

$$P_k(x) = a_k(x - x_k)^2 + b_k(x - x_k) + c_k.$$

Reemplazando los tres puntos a interpolar, se obtiene un sistema de ecuaciones lineales:

$$\begin{aligned} a_k(x_{k-1} - x_k)^2 + b_k(x_{k-1} - x_k) + c_k &= f(x_{k-1}) \\ a_k(x_k - x_k)^2 + b_k(x_k - x_k) + c_k &= f(x_k) \\ a_k(x_{k+1} - x_k)^2 + b_k(x_{k+1} - x_k) + c_k &= f(x_{k+1}) \end{aligned}$$

que se resuelve de manera sencilla si se considera que $h = x_k - x_{k-1} = x_{k+1} - x_k$. Entonces:

$$a_k = \frac{f(x_{k+1}) - 2f(x_k) + f(x_{k-1})}{2h^2}, \qquad b_k = \frac{f(x_{k+1}) - f(x_{k-1})}{2h}, \qquad c_k = f(x_k).$$

Por lo tanto, para calcular el área bajo la curva:

$$\begin{aligned} \int_{x_{k-1}}^{x_{k+1}} f(x)\,\mathrm{d}x &= \int_{x_{k-1}}^{x_{k+1}} P_k(x)\,\mathrm{d}x + E(h) \\ &= a_k \int_{x_{k-1}}^{x_{k+1}} [x - x_k]^2\,\mathrm{d}x + b_k \int_{x_{k-1}}^{x_{k+1}} [x - x_k]\,\mathrm{d}x + c_k \int_{x_{k-1}}^{x_{k+1}} \mathrm{d}x + E(h) \\ &= a_k \left.\frac{(x - x_k)^3}{3}\right|_{x_{k-1}}^{x_{k+1}} + b_k \left.\frac{(x - x_k)^2}{2}\right|_{x_{k-1}}^{x_{k+1}} + c_k\, x|_{x_{k-1}}^{x_{k+1}} + E(h) \\ &= a_k \left[\frac{h^3}{3} + \frac{h^3}{3}\right] + b_k \left[\frac{h^2}{2} + \frac{h^2}{2}\right] + c_k 2h + E(h) \\ &= \frac{f(x_{k+1}) - 2f(x_k) + f(x_{k-1})}{2h^2} \left[\frac{2h^3}{3}\right] + f(x_k)2h + E(h) \\ &= \frac{h}{3}\left[f(x_{k+1}) + 4f(x_k) + f(x_{k-1})\right] + E(h). \end{aligned}$$

Aplicando el desarrollo anterior a la fórmula (10.4) se obtiene la **fórmula compuesta de Simpson**:

$$\begin{aligned} \int_a^b f(x)\,\mathrm{d}x &= \int_{x_0}^{x_2} P_1(x)\,\mathrm{d}x + \int_{x_2}^{x_4} P_3(x)\,\mathrm{d}x + \cdots + \int_{x_{n-2}}^{x_n} P_5(x)\,\mathrm{d}x + E(h) \\ &= \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)\right] + E(h) \\ &= \frac{h}{3}\left[f(x_0) + 2\sum_{i=1}^{n/2-1} f(x_{2i}) + 4\sum_{i=1}^{n/2} f(x_{2i-1}) + f(x_n)\right] + E(h), \end{aligned}$$

donde n es par y $n \geq 2$.

Se puede demostrar que una cota para el error asociado a la aproximación por medio de la fórmula compuesta de Simpson es:

$$E(h) = \frac{nh^4}{180} f^{(4)}(\xi), \qquad \xi \in [x_0, x_n].$$

Ejemplo 64. *Utilizando el método de Simpson, se desea aproximar el valor de*

$$\int_0^3 [x\cos(x) + 4]\ \mathrm{d}x$$

con una partición regular de 7 puntos. Como $n = 6$, entonces se construirán 3 polinomios interpoladores. El gráfico de la función, y los polinomios interpoladores que determinan el área se muestra como figura 10.8

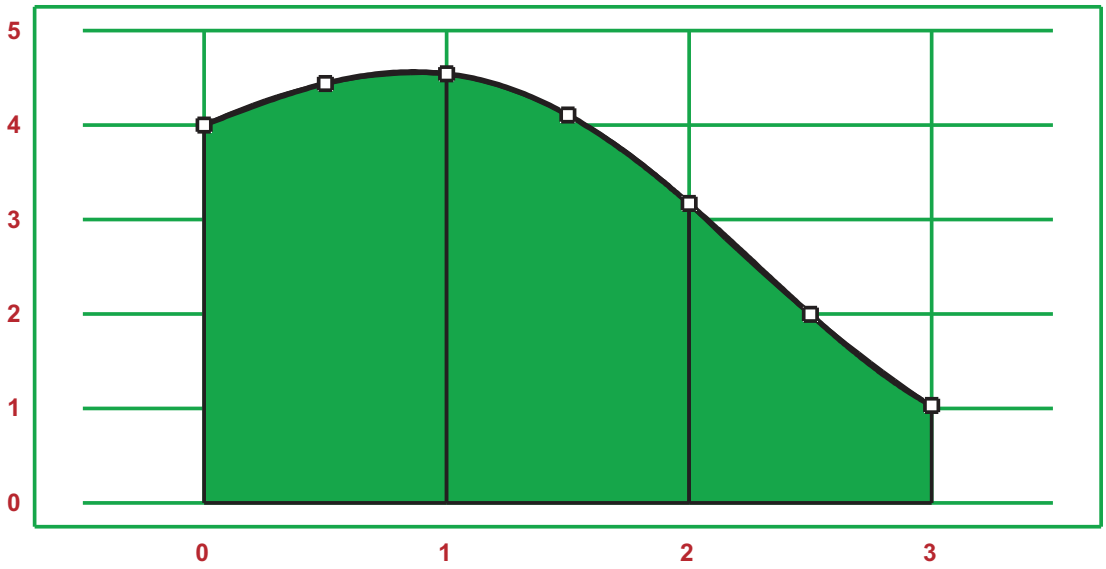

Figura 10.8: Integración por aproximación de Simpson, para $f(x) = x\cos(x) + 4$ en $[0, 3]$.

Dichos polinomios son:

$$\begin{aligned} P_1(x) &= -0{,}6746x^2 + 1{,}215x + 4{,}000 \\ P_3(x) &= -1{,}008x^2 + 1{,}651x + 3{,}898 \\ P_5(x) &= 0{,}4069x^2 - 4{,}173x + 9{,}885, \end{aligned}$$

por lo que:

$$\begin{aligned} \int_0^3 [x\cos(x) + 4]\ \mathrm{d}x &= \int_0^1 P_1\,\mathrm{d}x + \int_1^2 P_3\,\mathrm{d}x + \int_2^3 P_5\,\mathrm{d}x + E(h) \\ &= 4{,}3826 + 4{,}0225 + 2{,}0295 + E(h) \\ &= 10{,}435 + E(h), \end{aligned}$$

donde $E(h)$ puede calcularse como:

$$E(h) = \frac{6 \cdot 0{,}5^4}{180} 4{,}5 = 0{,}009375$$

Ejercicio 34. *Resolver el ejemplo anterior utilizando 4 elementos en la partición. Mejorar el cálculo de la partición de 6 puntos utilizando extrapolación.*

Comandos de EMT. *El comando para integrar numéricamente es:*

- `integrate(f$:string, a:número, b:número, eps, steps, method, fast)`, *donde* ***f$*** *es la función de* x *a integrar, expresada como* string*;* ***a*** *y* ***b*** *forman el intervalo de integración;* ***eps*** *es un parámetro optativo y se utiliza dentro de la rutina como tolerancia en el caso de aplicar refinamientos ;* ***steps*** *es un parámetro optativo e indica la partición a utilizar, el valor por defecto es 10;* ***method*** *es un parámetro optativo y permite elegir entre 4 diferentes métodos, la opción por defecto es el método adaptativo de Gauss;* ***fast*** *es un parámetro optativo e indica si se aplica el método de Gauss con sólo una partición.*

Ejemplo en EMT 20. *Integrar numéricamente la función* $f(x) = \cos(x) + \sin(x)$ *en* $[0; 5]$.

```
>integrate("cos(x)+sin(x)",0,5)
 -0.242586460126
```

10.3. Métodos de cuadratura

En este caso, se tratará de generar funciones que interpolen a $f(x)$ en:

$$\int_a^b f\,(x)\,\mathrm{d}x$$

con el fin de integrar el interpolador obtenido en vez de la $f(x)$ original. Sin embargo, debido al efecto Runge no es muy recomendable generar interpoladores polinómicos de grado elevado. Con el fin de resolver el problema planteado se utilizan las **fórmulas de cuadratura**, aproximando la integral mediante una suma ponderada de valores de la función en ciertos puntos del intervalo $[a, b]$ llamados **nodos** y utilizando coeficientes llamados **pesos**. De esta forma es posible, bajo ciertas condiciones, aproximar la integral deseada mediante:

$$\int_a^b f\,(x)\,\mathrm{d}x \approx p_0 f(x_0) + p_1 f(x_1) + \ldots + p_n f(x_n)$$

Ahora el problema es distinto: es necesario elegir convenientemente los nodos, en el caso en que sea posible, y los pesos de manera que el error de integración sea lo más pequeño posible.

10.3.1. Cuadratura de Newton-Cotes

Estos métodos ya fueron vistos como **métodos de paso finito**. El análisis es un poco más complejo que el ya desarrollado, pero es importante comprender el concepto de cuadratura partir de métodos simples y cuyo desarrollo se conoce en forma completa. Para las fórmulas planteadas no se calculará el error, ya conocido a partir de las expresiones de expansión de Taylor.

Fórmula de un punto

Se desea que la fórmula:

$$\int_a^b f\,(x)\,\mathrm{d}x \approx p_0 f(x_0)$$

sea exacta para polinomios de grado 0. Entonces x_0 es el punto medio del intervalo $[a, b]$ y:

$$\int_a^b 1\mathrm{d}x = b - a \approx p_0 1,$$

entonces $p_0 = b - a$ con lo que la fórmula de cuadratura de un punto es:

$$\int_a^b f(x)\,\mathrm{d}x \approx (b-a) f(x_0 + h/2). \tag{10.5}$$

Si se divide un intervalo de integración en partes más pequeñas y se aplica la fórmula 10.5 se obtiene:

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x \\ &\approx h\left[f(x_0 + h/2) + f(x_1 + h/2) + \ldots + f(x_{n-1} + h/2)\right]\end{aligned}$$

Fórmula de dos puntos

Se desea que la fórmula:

$$\int_a^b f(x)\,\mathrm{d}x \approx p_0 f(x_0) + p_1 f(x_1)$$

sea exacta para polinomios de grado 0 y de grado 1. Entonces se eligen los dos puntos $x_0 = a$, $x_1 = b$ y:

$$\begin{aligned}\int_a^b 1\mathrm{d}x &= p_0 1 + p_1 1 \\ \int_a^b x\mathrm{d}x &= p_0 a + p_1 b,\end{aligned}$$

con lo que se genera el sistema de ecuaciones lineales:

$$\begin{aligned}b - a &= p_0 + p_1 \\ \frac{b^2}{2} - \frac{a^2}{2} &= p_0 a + p_1 b,\end{aligned}$$

donde, después de algunas operaciones algebraicas, se llega a la solución:

$$p_0 = p_1 = \frac{b-a}{2}.$$

Entonces la fórmula de cuadratura de dos puntos es:

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{b-a}{2}\left(f(a) + f(b)\right). \tag{10.6}$$

Si se divide un intervalo de integración en partes más pequeñas y se aplica la fórmula 10.6 se obtiene:

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x \\ &\approx \frac{h}{2}\left[f(x_0) + 2f(x_1) + 2f(x_2) + \ldots + 2f(x_{n-1}) + f(x_n)\right]\end{aligned}$$

Ejercicio 35. *Desarrollar la fórmula de cuadratura de tres puntos, cuya expresión final es equivalente al método de Simpson.*

Fórmulas de orden superior

Las fórmulas de 1, 2 y 3 puntos ofrecen la misma solución al problema de integración que los métodos de paso finito. En realidad son interpolaciones de grado 0, 1 y 2 para una función $f(x)$ en un intervalo $[a, b]$. Son casos particulares de las fórmulas de Newton-Cotes, que permiten aproximar integrales por interpolación sin la necesidad de generar el polinomio interpolador que aproxima al integrando.

Sea:

$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b P_n(x)\,\mathrm{d}x,$$

entonces por medio de la fórmula de Lagrange:

$$\begin{aligned}\int_a^b P_n(x)\,\mathrm{d}x &= \int_a^b \sum_{j=0}^{n} l_{j,n}(x)\, f(x_j)\,\mathrm{d}x \\ &= \sum_{j=0}^{n} w_{j,n} f(x_j),\end{aligned}$$

donde:

$$w_{j,n} = \int_a^b l_{j,n}(x)\,\mathrm{d}x,$$

con $j = 0, 1, \ldots, n$. Los pesos w ya han sido calculados[1] para los casos donde $n = 0$, $n = 1$ y $n = 2$. Ahora se calculará w_0 para $n = 3$, los demás pesos (independiente del grado del interpolante) se desarrollan en forma similar.

Para aproximar la integral:

$$\int_a^b f(x)\mathrm{d}x$$

se descompondrá el intervalo de integración en 4 puntos equiespaciados. Entonces:

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x &\approx \int_a^b P_3(x)\mathrm{d}x \\ &\approx w_0 f(x_0) + w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3).\end{aligned}$$

Resolviendo analíticamente:

$$w_0 = \int_a^b l_0(x)\mathrm{d}x = \int_{x_0}^{x_3} \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)}\mathrm{d}x,$$

ahora se introduce el cambio de variables: $x = x_0 + \mu h$, con $0 \leq \mu \leq 3$. Por lo tanto $\mathrm{d}x = h\mathrm{d}\mu$ y además:

$$\begin{aligned}x = x_0 &\Rightarrow \mu h = 0 \\ x = x_3 &\Rightarrow \mu h = 3 \\ xi - xj &= (i - j)h.\end{aligned}$$

Aplicando los cambios sugeridos:

$$\begin{aligned}w_0 &= \int_0^3 \frac{[(\mu - 1)h][(\mu - 2)h][(\mu - 3)h]}{(-h)(-2h)(-3h)} h\mathrm{d}\mu \\ &= -\frac{h}{6}\int_0^3 (\mu - 1)(\mu - 2)(\mu - 3)\,\mathrm{d}\mu \\ &= \frac{3h}{8}\end{aligned}$$

[1] a través de métodos más simples

De manera similar a lo ya mostrado, se tiene que:

$$w_1 = \frac{9h}{8}; \qquad w_2 = \frac{9h}{8}; \qquad w_3 = \frac{3h}{8}.$$

Por lo tanto:

$$\int_a^b f(x)\mathrm{d}x \approx \frac{3h}{8}\left[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right]$$

Con el fin de estimar el error en la cuadratura de Newton-Cotes, se tendrá en cuenta el siguiente teorema.

Teorema 26. ***(a)*** *Para n par, y asumiendo que $f(x)$ es $n+2$ veces diferenciable en $[a,b]$, entonces el error cometido por la cuadratura de Newton-Cotes es:*

$$\left|\int_a^b f(x)\,dx - \int_a^b P_n(x)\,dx\right| = C_n h^{n+3} f^{(n+2)}(\xi),$$

con $\xi \in [a,b]$ y:

$$C_n = \frac{1}{(n+2)!}\int_0^n \mu^2\,(\mu-1)\,(\mu-2)\ldots(\mu-n)\,d\mu$$

(b) *Para n impar, y asumiendo que $f(x)$ es $n+1$ veces diferenciable en $[a,b]$, entonces el error cometido por la cuadratura de Newton-Cotes es:*

$$\left|\int_a^b f(x)\,dx - \int_a^b P_n(x)\,dx\right| = C_n h^{n+2} f^{(n+1)}(\xi),$$

con $\xi \in [a,b]$ y:

$$C_n = \frac{1}{(n+1)!}\int_0^n \mu\,(\mu-1)\,(\mu-2)\ldots(\mu-n)\,d\mu$$

Nota. *La cota de error planteada por el teorema 26 para el método de Simpson (o Newton-Cotes de tres puntos) difiere con el mostrado anteriormente. Esto es así porque el teorema 26 brinda una cota para el* ***error global****, mientras que la expresión ya desarrollada muestra una cota para el error de la fórmula compuesta de Simpson.*

Ejemplo 65. *Utilizando la cuadratura de Newton-Cotes y 4 puntos equiespaciados, se desea aproximar el valor de:*

$$\int_0^3 \left[x\cos(x) + 4\right] dx.$$

Entonces $n = 3$, $x_i = i$ y el polinomio que interpola los datos es:

$$P_3 = 0{,}1913x^3 - 1{,}530x^2 + 1{,}879x + 4{,}000,$$

con lo que se obtiene:

$$\begin{aligned}\int_0^3 \left[x\cos(x) + 4\right] dx &\approx \int_0^3 P_3(x)\,dx \\ &\approx 10{,}557.\end{aligned}$$

Sin embargo, es posible aproximar la integral sin calcular la expresión del polinomio interpolador, sólo a partir de la cuadratura de 4 puntos:

$$\begin{aligned}\int_0^3 \left[x\cos(x) + 4\right] dx &\approx \frac{3h}{8}\left[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right] \\ &\approx \frac{3}{8}\left[f(0) + 3f(1) + 3f(2) + f(3)\right] \\ &\approx \frac{3}{8}\left[4 + 3(4{,}5403) + 3(3{,}1677) + 1{,}0300\right] \\ &\approx 10{,}557\end{aligned}$$

Ejercicio 36. *Estimar el error cometido en la aproximación del ejemplo 65.*

Sólo con fines didácticos, se mostrarán las fórmulas de aproximación para $n = 4, 5, 6$:

$$n = 4 \Rightarrow \int_a^b f(x)\mathrm{d}x \approx \frac{2h}{45}\left[7f(a) + 32f(a+h) + 12f\left(\frac{a+b}{2}\right) + 32f(b-h) + 7f(b)\right]$$

$$n = 5 \Rightarrow \int_a^b f(x)\mathrm{d}x \approx \frac{5h}{288}[19f(a) + 75f(a+h) + 50f(a+2h) + \\ +50f(b-2h) + 75f(b-h) + 19f(b)]$$

$$n = 6 \Rightarrow \int_a^b f(x)\mathrm{d}x \approx \frac{h}{140}\left[41f(a) + 216f(a+h) + 27f(a+2h) + 272f\left(\frac{a+b}{2}\right) + \right. \\ \left. +27f(b-2h) + 216f(b-h) + 41f(b)\right]$$

Ejercicio 37. *Acotar los errores que se cometen al aplicar las fórmulas de cuadratura para $n = 4, 5, 6$.*

La cuadratura de Newton-Cotes, al estar basada en polinomios interpoladores, es sensible al efecto Runge.

Ejemplo 66. *Se quiere aproximar el valor de:*

$$I = \int_{-4}^{4} \frac{1}{1+x^2}\,dx,$$

cuyo valor exacto es $2\arctan^{-1}(4) \approx 2{,}6516$. *Para ello se utilizará la fórmula de cuadratura de Newton-Cotes, integrando con una cantidad impar de nodos. Los resultados se resumen en la tabla 10.2, que muestra un comportamiento oscilatorio de mayor amplitud a medida que aumenta la cantidad de nodos de integración.*

n	I
2	5,4902
4	2,2776
6	3,3288
8	1,9411
10	3,5956

Tabla 10.2: Resolución de integral por medio de la fórmula de Newton-Cotes.

10.3.2. Cuadratura de Gauss

La cuadratura de Newton-Cotes ofrece una solución simple al problema de obtener una aproximación de integrales definidas. Pero es recomendable subdividir el intervalo de integración con el fin de evitar el efecto Runge, ya que su aproximación se basa en polinomios interpoladores. La cuadratura gaussiana ofrece fórmulas de buena calidad con un esquema simple y de bajo costo computacional.

Teorema 27 (Cuadratura Gaussiana)**.** *Sea q un polinomio no trivial de grado $n+1$ de forma tal que:*

$$\int_a^b x^k q(x)\,dx = 0,$$

con $0 \leq k \leq n$. Sean $x_0, x_1, \ldots, x_n$ las raíces de q. Entonces la fórmula:

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} A_i f(x_i),$$

donde:

$$A_i = \int_a^b l_i(x)dx,$$

con los x_i antes mencionados como nodos, es exacta para todos los polinomios de grado máximo $2n+1$. Más aún, los nodos se ubican dentro del intervalo abierto (a, b).

Demostración. Sea f un polinomio de grado menor ó igual que $2n+1$. Dividiendo f por q, se obtiene un cociente p y un resto r, ambos de grado menor ó igual a n. Entonces:

$$f = pq + r.$$

Por hipótesis, $\int_a^b q(x)p(x)\mathrm{d}x = 0$. Es más, como cada x_i es una raíz de q, se tiene que:

$$f(x_i) = p(x_i)q(x_i) + r(x_i) = r(x_i).$$

Finalmente, como r tiene grado menor ó igual a n, entonces el teorema asegura que se calculará $\int_a^b r(x)\mathrm{d}x$ en forma exacta y:

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x &= \int_a^b p(x)q(x)\mathrm{d}x + \int_a^b r(x)\mathrm{d}x \\ &= \int_a^b r(x)\mathrm{d}x \\ &= \sum_{i=0}^{n} A_i r(x_i) \\ &= \sum_{i=0}^{n} A_i f(x_i)\end{aligned}$$

□

Lo que plantea el teorema anterior es que, con nodos arbitrarios (generalmente equiespaciados), la integración de cuadratura utilizando el polinomio de Lagrange será exacta para polinomios de grado menor ó igual a n, pero con los nodos gaussianos, la cuadratura será exacta para polinomios de grado menor ó igual a $2n+1$.

Las cuadraturas que surgen a partir del teorema 27 son llamadas **Cuadraturas Gaussianas** ó **Fórmulas de Cuadratura de Gauss-Legendre**. Existen fórmulas diferentes para cada intervalo $[a, b]$ y cada valor de n. También es posible generar fórmulas gaussianas para dar valores aproximados de integrales especiales como:

$$\int_0^\infty f(x)e^{-x}\mathrm{d}x; \qquad \int_{-1}^1 f(x)\sqrt{1-x^2}\mathrm{d}x; \qquad \int_{-\infty}^\infty f(x)e^{-x^2}\mathrm{d}x;$$

por citar algunos ejemplos comunes.

La derivación de cuadraturas no es muy difícil de resolver si se requiere una aproximación de:

$$\int_{-1}^1 f(x)\mathrm{d}x,$$

utilizando una cantidad definida de pesos en la fórmula. Para ello es necesario escribir el polinomio $q(x)$ y obtener sus raíces. Luego las raíces son los puntos en los que se evaluará la función y resta calcular los pesos resolviendo un sistema de ecuaciones lineales.

Fórmula de dos puntos

En este caso, deben calcularse dos pesos y dos puntos para generar la fórmula de cuadratura. Para ello se utiliza un polinomio de grado dos:

$$q(x) = a_0 + a_1 x + a_2 x^2,$$

que debe cumplir las condiciones:

$$\int_{-1}^{1} q(x)\mathrm{d}x = \int_{-1}^{1} xq(x)\mathrm{d}x = 0.$$

Como:

$$\int_{-1}^{1} a_0 + a_1 x + a_2 x^2 \mathrm{d}x = \frac{12a_0 + 4a_2}{6} = 0$$
$$\int_{-1}^{1} x\left(a_0 + a_1 x + a_2 x^2\right)\mathrm{d}x = \frac{2a_1}{3} = 0,$$

entonces, una de las soluciones posibles es $a_0 = 1$; $a_1 = 0$ y $a_2 = -3$. Por lo tanto:

$$q(x) = 1 - 3x^2$$

y sus raíces son $-\sqrt{\frac{1}{3}}$ y $\sqrt{\frac{1}{3}}$. Esos valores son los nodos gaussianos de la cuadratura buscada. Sólo resta calcular los pesos A_i de forma tal que la aproximación:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx A_0 f\left(-\sqrt{\frac{1}{3}}\right) + A_1 f\left(\sqrt{\frac{1}{3}}\right),$$

sea exacta para $f(x)$ de la forma $ax + b$. Para obtener los pesos A_0 y A_1 se utiliza el método de coeficientes indeterminados[2]. Ya que la integración es un proceso lineal, entonces puede separarse la integral en dos integrales:

$$\int_{-1}^{1} \mathrm{d}x = 2 = A_0 + A_1$$
$$\int_{-1}^{1} x\mathrm{d}x = 0 = -\sqrt{\frac{1}{3}}A_0 + \sqrt{\frac{1}{3}}A_1,$$

de donde surge el sistema de ecuaciones lineales equivalente al generado por definición:

$$A_0 + A_1 = 2$$
$$-A_0 + A_1 = 0.$$

La solución del sistema anterior es $A_0 = A_1 = 1$, de lo que sigue que la fórmula de cuadratura es:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx f\left(-\sqrt{\frac{1}{3}}\right) + f\left(\sqrt{\frac{1}{3}}\right), \tag{10.7}$$

y permite integrar en forma exacta todos los polinomios de grado menor ó igual a 3.

Ejemplo 67. *Es posible calcular la integral de* $f(x) = 1 + 2x + 3x^2 + 4x^3$, *entre* -1 *y* 1 *a través de la cuadratura gaussiana de dos puntos. Para ello se aplica la fórmula (10.7):*

$$\begin{aligned}\int_{-1}^{1} 1 + 2x + 3x^2 + 4x^3 dx &\approx f\left(-\sqrt{\frac{1}{3}}\right) + f\left(\sqrt{\frac{1}{3}}\right)\\ &\approx 2 - \frac{10}{\sqrt{3^3}} + \frac{10}{\sqrt{3^3}} + 2\\ &\approx 4.\end{aligned}$$

[2] utilizado en los ejercicios de derivación numérica

En este caso se obtuvo el valor correcto, recordando que la fórmula de cuadratura gaussiana de dos puntos es exacta para polinomios de grado menor ó igual a 3.

Fórmula de tres puntos

Para este caso, son necesarios tres pesos en la fórmula. Como el grado de q es 3, entonces debe tener la forma:

$$q(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3.$$

Las condiciones que debe satisfacer son:

$$\int_{-1}^{1} q(x)\mathrm{d}x = \int_{-1}^{1} xq(x)\mathrm{d}x = \int_{-1}^{1} x^2 q(x)\mathrm{d}x = 0.$$

Si se eligen $a_0 = a_2 = 0$, entonces $q(x) = a_1 x + a_3 x^3$ y:

$$\int_{-1}^{1} q(x)\mathrm{d}x = \int_{-1}^{1} x^2 q(x)\mathrm{d}x = 0,$$

puesto que la integral de una función impar sobre un intervalo simétrico equivale a 0. Para determinar a_1 y a_3 debe cumplirse la condición:

$$\begin{aligned} 0 &= \int_{-1}^{1} x\left(a_1 x + a_3 x^3\right)\mathrm{d}x \\ &= \left.\frac{3a_3 x^5 + 5a_1 x^3}{15}\right|_{-1}^{1} \\ &= \frac{2\left(3a_3 + 5a_1\right)}{15}, \end{aligned}$$

con lo que, una de las soluciones posibles es $a_1 = -3$ y $a_3 = 5$. Entonces:

$$q(x) = 5x^3 - 3x,$$

y sus raíces son $-\sqrt{3/5}$; 0 y $\sqrt{3/5}$. Esos valores son los nodos gaussianos para la cuadratura buscada. Deben entonces seleccionarse A_0, A_1 y A_2 en la fórmula:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx A_0 f\left(-\sqrt{\frac{3}{5}}\right) + A_1 f(0) + A_2 f\left(\sqrt{\frac{3}{5}}\right),$$

de forma tal que la aproximación sea exacta para $f(x)$ de la forma $ax^2 + bx + c$. Como la integración es un proceso lineal, entonces es posible calcular por separado los valores de integración de 1, x y x^2:

$$\begin{aligned} \int_{-1}^{1} \mathrm{d}x &= 2 = A_0 + A_1 + A_2 \\ \int_{-1}^{1} x\mathrm{d}x &= 0 = -\sqrt{\frac{3}{5}}A_0 + \sqrt{\frac{3}{5}}A_2 \\ \int_{-1}^{1} x^2\mathrm{d}x &= \frac{2}{3} = \frac{3}{5}A_0 + \frac{3}{5}A_2 \end{aligned}$$

Resolviendo el sistema anterior, se obtienen los pesos: $A_0 = A_2 = \frac{5}{9}$ y $A_1 = \frac{8}{9}$. Por lo tanto, la fórmula de cuadratura es:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right), \tag{10.8}$$

y permite integrar en forma exacta todos los polinomios de grado menor ó igual a 5.

Ejemplo 68. *Es posible calcular la integral de* $f(x) = 1 + 2x + 3x^2 + 4x^3 + 5x^4 + 6x^5$, *entre* -1 *y* 1 *a través de la cuadratura gaussiana de tres puntos. Para ello se aplica la fórmula (10.8):*

$$\int_{-1}^{1} 1 + 2x + 3x^2 + 4x^3 + 5x^4 + 6x^5 dx \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right) \approx 6.$$

En este caso se obtuvo el valor correcto, recordando que la fórmula de cuadratura gaussiana de tres puntos es exacta para polinomios de grado menor ó igual a 5.

Fórmulas de orden superior

Si bien la deducción de las fórmulas de dos y tres puntos resultó simple, es posible sistematizar la generación de fórmulas de orden superior a través de los **polinomios de Legendre**. Las raíces de los polinomios de Legendre serán utilizadas como nodos de la cuadratura gaussiana para el intervalo $[-1; 1]$. La principal ventaja de los polinomios de Legendre es que son ortogonales, además de que se pueden obtener por medio de una expresión recursiva:

$$\begin{aligned} q_0(x) &= 1 \\ q_1(x) &= x \\ q_2(x) &= \frac{3}{2}x^2 - \frac{1}{2} \\ &\vdots \\ q_n(x) &= \left(\frac{2n-1}{n}\right) x q_{n-1}(x) - \left(\frac{n-1}{n}\right) q_{n-2}(x) \end{aligned} \tag{10.9}$$

Ejemplo 69. *Para obtener el polinomio* $q(x)$, *necesario para la fórmula de cuadratura de cuatro puntos, no hace falta establecer condiciones sobre el polinomio sino que es posible generarlo a través del esquema recursivo (10.9), economizando gran parte del trabajo:*

$$\begin{aligned} q_3(x) &= \frac{5}{3}x q_2(x) - \frac{2}{3} q_1(x) \\ &= \frac{5}{3}x\left(\frac{3}{2}x^2 - \frac{1}{2}\right) - \frac{2}{3}x \\ &= \frac{5}{2}x^3 - \frac{3}{2}x; \\ q_4(x) &= \frac{7}{4}x q_3(x) - \frac{3}{4} q_2(x) \\ &= \frac{7}{4}x\left(\frac{5}{2}x^3 - \frac{3}{2}x\right) - \frac{3}{4}\left(\frac{3}{2}x^2 - \frac{1}{2}\right) \\ &= \frac{35}{8}x^4 - \frac{15}{4}x^2 + \frac{3}{8}. \end{aligned}$$

Siendo $q_4(x)$ *una expresión bicuadrática cuyas raíces (nodos de la cuadratura gaussiana) son:*

$$r_1 = -\sqrt{\tfrac{15+2\sqrt{30}}{35}} \approx -0{,}861136311594; \quad r_2 = \sqrt{\tfrac{15+2\sqrt{30}}{35}} \approx 0{,}861136311594$$
$$r_3 = -\sqrt{\tfrac{15-2\sqrt{30}}{35}} \approx -0{,}339981043585; \quad r_4 = \sqrt{\tfrac{15-2\sqrt{30}}{35}} \approx 0{,}339981043585$$

Ejercicio 38. *Determinar los coeficientes* A_i *para las fórmulas de cuadratura de cuatro puntos.*

Extensión del intervalo de integración

Hasta ahora se generaron las fórmulas de la cuadratura gaussiana para el intervalo $[-1;1]$. Por medio de la transformación lineal:

$$x(t) = \frac{b-a}{2}t + \frac{b+a}{2}; \qquad \mathrm{d}x = \frac{b-a}{2}\mathrm{d}t,$$

puede rescribirse la integral:

$$\int_a^b f(x)\mathrm{d}x$$

como:

$$\int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{b+a}{2}\right)\frac{b-a}{2}\mathrm{d}t.$$

Entonces la fórmula de cuadratura es:

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x &= \int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{b+a}{2}\right)\frac{b-a}{2}\mathrm{d}t \\ &= \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{b+a}{2}\right)\mathrm{d}t \\ &\approx \frac{b-a}{2}\sum_{i=0}^{n-1} A_i f\left(\frac{b-a}{2}t_i + \frac{b+a}{2}\right)\end{aligned}$$

Ejemplo 70. *Utilizando la cuadratura gaussiana de 3 puntos, se desea aproximar el valor de:*

$$\int_0^3 [x\cos(x) + 4]\,\mathrm{d}x$$

Entonces $a = 0$ *y* $b = 3$*, con lo que:*

$$x(t) = \frac{3}{2}t + \frac{3}{2}$$

y

$$t_0 \approx -0{,}77459666; \qquad t_1 \approx 0; \qquad t_2 \approx 0{,}77459666.$$

De lo anterior se sigue que:

$$\begin{aligned}\int_0^3 [x\cos(x) + 4]\,\mathrm{d}x &= \int_{-1}^{1} f\left(\frac{3-0}{2}t + \frac{3+0}{2}\right)\frac{3-0}{2}\mathrm{d}t \\ &\approx \frac{3}{2}\left[\frac{5}{9}f(0{,}33810501) + \frac{8}{9}f(1{,}5) + \frac{5}{9}f(2{,}6618949)\right] \\ &\approx 10{,}439\end{aligned}$$

Ejercicio 39. *Integrar la función de Runge, en el intervalo* $[-4;4]$*, con las fórmulas de tres y cuatro puntos de la cuadratura gaussiana.*

10.3.3. Cuadratura de Chebyshev

Para construir la cuadratura de Gauss-Laguerre fue necesario, para una cantidad de nodos dados, identificar los nodos x_i y sus pesos asociados, planteando el trabajo de construcción de las fórmulas en dos grandes partes. En la cuadratura de Chebyshev se considera que el valor de todos los pesos, para una cantidad determinada de nodos, es constante. Sin embargo, la ubicación de los nodos queda sujeta a la resolución de un sistema de ecuaciones no lineales, con lo que a medida que aumenta la cantidad de nodos se torna muy complicado (y a partir de cierto grado, imposible) resolver el sistema no lineal asociado.

Fórmula de dos puntos

Se desea que la fórmula:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \sum_{i=0}^{n-1} A_i f(x_i),$$

sea exacta para polinomios de grado menor ó igual a 2 y que los A_i sean todos iguales. Entonces, recordando que la integración es una operación lineal, obtenemos el sistema asociado:

$$\begin{aligned} \int_{-1}^{1} \mathrm{d}x = 2 &= \sum_{i=0}^{n-1} A_i \\ \int_{-1}^{1} x\mathrm{d}x = 0 &= \sum_{i=0}^{n-1} A_i x_i \\ \int_{-1}^{1} x^2\mathrm{d}x = \frac{2}{3} &= \sum_{i=0}^{n-1} A_i x_i^2 \end{aligned} \tag{10.10}$$

Observando el sistema, se nota que la primera sumatoria del conjunto de ecuaciones (10.10) puede resolverse en forma independiente, ya que todos los pesos de esta cuadratura para $n = 2$ toman el mismo valor:

$$2 = \sum_{i=0}^{n-1} A = nA$$

por lo que:

$$\frac{2}{n} = A. \tag{10.11}$$

Utilizando (10.11) en el sistema (10.10) y sabiendo que $n = 2$, disminuye la complejidad del sistema:

$$\begin{aligned} 0 &= \frac{2}{n}\sum_{i=0}^{n-1} x_i = x_0 + x_1, \\ \frac{2}{3} &= \frac{2}{n}\sum_{i=0}^{n-1} x_i^2 = x_0^2 + x_1^2. \end{aligned}$$

Resolviendo numéricamente se obtiene:

$$x_0 \approx -0{,}57735026; \qquad x_1 \approx 0{,}57735026.$$

Entonces la fórmula de cuadratura de dos puntos es:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx f(-0{,}57735026) + f(0{,}57735026),$$

y esta fórmula coincide con la cuadratura de dos puntos de Gauss-Laguerre.

Fórmula de tres puntos

Se desea que la fórmula:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \sum_{i=0}^{n-1} A_i f(x_i),$$

sea exacta para polinomios de grado menor ó igual a 3 y que los A_i sean todos iguales. Sabiendo que la integración es un proceso lineal, es posible separar la integral propuesta en cuatro integrales distintas, cada una de ellas con un polinomio mínimo trivial:

$$\begin{aligned}
\int_{-1}^{1} \mathrm{d}x = 2 &= \sum_{i=0}^{n-1} A_i \\
\int_{-1}^{1} x\mathrm{d}x = 0 &= \sum_{i=0}^{n-1} A_i x_i \\
\int_{-1}^{1} x^2\mathrm{d}x = \frac{2}{3} &= \sum_{i=0}^{n-1} A_i x_i^2 \\
\int_{-1}^{1} x^3\mathrm{d}x = 0 &= \sum_{i=0}^{n-1} A_i x_i^3
\end{aligned} \tag{10.12}$$

En este caso, al igual que en el desarrollo de la fórmula de dos puntos, la primera ecuación del conjunto (10.12) puede resolverse en forma independiente al sistema y se cumple la igualdad (10.11).

Entonces, debe resolverse el sistema no lineal:

$$\begin{aligned}
0 &= \frac{2}{n}\sum_{i=0}^{n-1} x_i = \frac{2}{3}\left(x_0 + x_1 + x_2\right), \\
\frac{2}{3} &= \frac{2}{n}\sum_{i=0}^{n-1} x_i^2 = \frac{2}{3}\left(x_0^2 + x_1^2 + x_2^2\right), \\
0 &= \frac{2}{n}\sum_{i=0}^{n-1} x_i^3 = \frac{2}{3}\left(x_0^3 + x_1^3 + x_2^3\right),
\end{aligned}$$

cuya solución es:

$$x_0 \approx -0{,}70710678; \qquad x_1 = 0; \qquad x_2 \approx 0{,}70710678.$$

Entonces la fórmula de cuadratura es:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{2}{3}\left[f(-0{,}70710678) + f(0) + f(0{,}70710678)\right].$$

Ejemplo 71. *Es posible calcular la integral de* $f(x) = 1 + 2x + 3x^2 + 4x^3$*, entre* -1 *y* 1 *a través de la cuadratura de Chebyshev de tres puntos. Para ello se aplica la fórmula 10.3.3:*

$$\begin{aligned}
\int_{-1}^{1} \left[1 + 2x + 3x^2 + 4x^3\right]\mathrm{d}x &\approx \frac{2}{3}\left[f(-0{,}70710678) + f(0) + f(0{,}70710678)\right] \\
&\approx 3{,}9999999.
\end{aligned}$$

El valor exacto de la integral es 4. En este caso, si bien la fórmula es exacta para polinomios de grado 3 e inferior, en lugar del valor exacto se obtiene una aproximación muy buena debido a la aritmética utilizada.

Fórmulas de orden superior

Contrario al método de cuadratura de Gauss, no es posible construir un esquema recursivo para generar polinomios cuyas raíces sean los nodos a utilizar en la suma

finita de la aproximación a la integral. Pero sí es posible crear un esquema general para plantear en forma simple el sistema no lineal cuya solución son los nodos de integración:

$$\frac{n\left[1+(-1)^k\right]}{2(k+1)} = \sum_{i=0}^{n-1} x_i^k, \tag{10.13}$$

donde n es la cantidad de nodos a utilizar y k se corresponde con el número de ecuación a resolver.

Ejemplo 72. *Con el fin de construir la fórmula de Chebyshev de cuatro puntos, es necesario plantear y resolver el sistema no lineal asociado. Utilizando la relación (10.13) queda el sistema:*

$$\begin{aligned} 0 &= x_0 + x_1 + x_2 + x_3 \\ \frac{4}{3} &= x_0^2 + x_1^2 + x_2^2 + x_3^2 \\ 0 &= x_0^3 + x_1^3 + x_2^3 + x_3^3 \\ \frac{4}{5} &= x_0^4 + x_1^4 + x_2^4 + x_3^4, \end{aligned}$$

cuya solución numérica es:

$$x_0 \approx -0{,}79465447; \quad x_1 \approx -0{,}18759247; \quad x_2 \approx 0{,}18759247; \quad x_3 \approx 0{,}79465447,$$

por lo que la fórmula de cuadratura es:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{1}{2}\left[f(-0{,}79465447) + f(-0{,}18759247) + f(0{,}18759247) + f(0{,}79465447)\right]$$

Nota. *La gran desventaja de la cuadratura de Chebyshev es que el sistema no lineal asociado al cálculo de coeficientes, no posee soluciones reales para $n = 8$ y $n \geq 10$.*

Extensión del intervalo de integración

La relación necesaria para extender esta cuadratura desde el intervalo $[-1;1]$ a un intervalo $[a;b]$ es muy similar al que se planteó para la cuadratura de Gauss, en la página 188, por lo que no es necesario realizar nuevamente el análisis.

Ejemplo 73. *Utilizando la cuadratura de Chebyshev de 3 puntos, se desea aproximar el valor de:*

$$\int_0^3 \left[x\cos(x) + 4\right]\mathrm{d}x$$

Entonces $a = 0$ y $b = 3$, con lo que:

$$x(t) = \frac{3}{2}t + \frac{3}{2}$$

y

$$t_0 \approx -0{,}70710678; \quad t_1 \approx 0; \quad t_2 \approx 0{,}70710678.$$

De lo anterior se sigue que:

$$\begin{aligned} \int_0^3 \left[x\cos(x) + 4\right]\mathrm{d}x &= \int_{-1}^{1} f\left(\frac{3-0}{2}t + \frac{3+0}{2}\right)\frac{3-0}{2}\mathrm{d}t \\ &\approx \frac{3}{2}\frac{2}{3}\left[f(0{,}43933983) + f(1{,}5) + f(2{,}5606601)\right] \\ &\approx 10{,}363 \end{aligned}$$

10.4. Ejercicios

1. Construir los siguientes algoritmos en PC:

   a) **Integración por rectángulos**. Entrada: una tabla de valores con los puntos de la función a integrar; extremos de integración finita; cantidad de subdivisiones del intervalo de integración. Salida: resultado de la integración numérica. Opcional: gráfico del área calculada.

   b) **Integración por trapecios**. Idénticas condiciones que la integración por rectágulos.

   c) **Integración por Simpson**. Idénticas condiciones que la integración por rectángulos.

2. Aproximar las siguientes integrales utilizando expansión por series de Taylor. Comparar los resultados de la aproximación por 2, 3 y 4 términos con el resultado de la rutina de integración de *Euler Math Toolbox*:

   a) $\int_2^4 2\cos(x)\mathrm{d}x$

   b) $\int_3^7 \sqrt{x+1}\mathrm{d}x$

   c) $\int_{-2}^0 \frac{x+1}{3x^2+5}\mathrm{d}x$

3. Calcular utilizando la aproximación por rectángulos, en todos los casos usar una partición mínima de 3 rectángulos:

   a) $\int_2^3 \sqrt{1+\cos^2(x)}\mathrm{d}x$

   b) $\int_0^\pi \sin(x)\mathrm{d}x$

   c) $\int_0^{10} \frac{\cos(x)}{\sqrt{x}}\mathrm{d}x$

4. Rehacer el ejercicio 3 pero esta vez utilizar el método del trapecio, con la misma cantidad de particiones que las usadas con el método del rectángulo. Comparar el resultado de ambos métodos.

5. Resolver utilizando el método de Simpson:

   a) $\int_{-3}^0 x\sin(x)\mathrm{d}x$

   b) $\int_0^1 \frac{1+e^x}{1+2e^x}\mathrm{d}x$

   c) $\int_0^1 x^2e^x\mathrm{d}x$

6. Acotar el error cometido en las integraciones de los incisos anteriores, de acuerdo a las fórmulas de error planteado.

7. Calcular las integrales a través de las cuadraturas de dos puntos:

   a) $\int_{-2}^0 (x-2)(x+1)\mathrm{d}x$

   b) $\int_{-3}^0 (9+2x)^{-1}\mathrm{d}x$

   c) $\int_{-2}^2 \frac{4}{9+2x^2}\mathrm{d}x$

8. Repetir el ejercicio anterior, pero ahora utilizar fórmulas de cuadratura de tres puntos.

9. [*EMT*] El **método de Romberg** propone integrar a través del método del trapecio una cantidad finita de veces y luego aplicar extrapolación de Richardson. Con $h = 0{,}6$, $h = 0{,}3$ y $h = 0{,}1$ aplicadas al método de Romberg, aproximar los valores de:

   a) $\int_0^{1,2} \sqrt{x+7}\mathrm{d}x$

   b) $\int_3^{3,6} e^x \cos(x)\mathrm{d}x$

   Estimar el orden del error de truncamiento cometido.

10. Construir una fórmula de cuadratura de la forma:

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \alpha f\left(-\frac{1}{2}\right) + \beta f(0) + \gamma f\left(\frac{1}{2}\right),$$

    que sea exacta para polinomios de grado menor ó igual a 2.

11. Derivar una fórmula de la forma:

$$\int_a^b f(x)\mathrm{d}x \approx w_0 f(a) + w_1 f(b) + w_2 f'(a) + w_3 f'(b),$$

    exacta para polinomios del grado más alto posible.

12. La regla de integración:

$$\int_0^{3h} f(x)\mathrm{d}x \approx \frac{3h}{8}\left[f(0) + 3f(h) + 3f(2h) + f(3h)\right]$$

    es exacta para polinomios de grado menor ó igual a n. Determinar el valor máximo de n para el cual la afirmación anterior es cierta.

13. Considerar la integral:

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}}\mathrm{d}x.$$

    Debido a que contiene singularidades en los extremos de integración, no es posible utilizar fórmulas cerradas de integración[3]. Aplicar las fórmulas de cuadratura de Chebyshev de dos y tres puntos y comparar con el valor exacto de la integral: π.

14. ¿Qué es mejor? ¿Integrar una función dada en forma de tabla tabla utilizando la fórmula de trapecios, la fórmula de rectángulos no centrales, ó la fórmula de rectángulos centrales? Comprobar la deducción con los datos de la tabla 10.3, que contiene puntos de la función $f(x) = x\sin(x-1) - \cos(x)$ y:

$$\begin{aligned}\int_2^5 f(x)\mathrm{d}x &= -\sin(5) + \sin(4) - 5\cos(4) + \sin(2) - \sin(1) + 2\cos(1)\\ &\approx 4{,}6187709.\end{aligned}$$

x_i	2	2,5	3	3,5	4	4,5	5
$f(x_i)$	2,09908	3,29488	3,71788	3,0311	1,21812	-1,36772	-4,06767

Tabla 10.3

[3] son las fórmulas que utilizan a los extremos de integración como nodos

15. Dada la tabla 10.4, determinar la distancia recorrida a partir de los datos:

    a) Utilizar la regla de los trapecios.

    b) Ajustar los datos a través de un polinomio cúbico completo, integrar la expresión cúbica para determinar la distancia.

t	1	2	3,25	4,5	6	7	8	8,5	9	10
v	5	6	5,5	7	8,5	8	6	7	7	5

Tabla 10.4

16. Los métodos adaptativos de cuadratura utilizan un conjunto de nodos propio a cada problema de integración a resolver. Dado el gráfico de:

$$f(x) = \frac{1}{x} + 4\cos(4x)$$

en el intervalo $(0; 2)$:

    a) Sugerir una partición de 10 puntos para calcular la integral pedida con el mínimo error posible.

    b) Calcular la integral con la partición del inciso anterior y el método del trapecio.

    c) Comprobar la precisión obtenida resolviendo la integral con *EMT*.

17. Utilizando la expresión del error asintótico[4] de la regla de los trapecios, estimar la cantidad de nodos a utilizar para evaluar las integrales pedidas con el error solicitado:

    a) $\int_1^3 \ln(x)\mathrm{d}x; \qquad E(h) \leq 10^{-3}$

    b) $\int_0^2 \frac{e^x - e^{-x}}{2}\mathrm{d}x; \qquad E(h) \leq 10^{-5}$

    c) $\int_0^2 e^{-x^2}\mathrm{d}x; \qquad E(h) \leq 10^{-8}$

18. El error de truncamiento $E_f(h)$, para los métodos de Simpsons y de los trapecios con una f determinada, tiene propiedades que simplifican su estimación:

    a) Mostrar que $E_{f+g}(h) = E_f(h) + E_g(h)$, para todas las funciones continuas $f(x)$ y $g(x)$.

    b) Mostrar que $E_{cf}(h) = cE_f(h)$, para todas las funciones continuas $f(x)$ y $c \in \mathbb{R}$.

19. [*EMT*] Para los métodos de paso finito, es posible estimar la tasa de convergencia p a partir de tres evaluaciones de la misma integral:

$$\frac{I_{2n} - I_n}{I_{4n} - I_{2n}} \approx 2^p,$$

donde n es la cantidad de nodos involucrados en la integración. Utilizando:

$$\int_0^4 \frac{x^2 + 3}{x + 1}\mathrm{d}x,$$

calcular las tasas de convergencia de:

[4] también llamado error de truncamiento

a) Método de Simpson.

b) Método del rectángulo central.

20. [*EMT*] Dada la integral:

$$I = \int_0^1 \sin\left(\sqrt{x}\right) \mathrm{d}x$$

a) Estimar su valor numérico utilizando el método de Simpson utilizando $n = 2, 4, 8, \ldots, 128$. Calcular su tasa de convergencia.

b) Repetir el inciso anterior, pero utilizar el cambio de variables $x = t^2$, que lleva a la integral:

$$I = 2\int_0^1 t\sin(t)\mathrm{d}t$$

21. A partir de la resta de las expansiones de Taylor $f(x+h)$ y $f(x-h)$, es posible crear una fórmula de aproximación de integrales definidas basada en derivadas.

a) Construir una fórmula que dependa de la derivada, hasta el orden cinco, de la función a integrar.

b) Acotar su error, si es posible.

c) Probar el desempeño de la fórmula creada aproximando el valor de:

$$\int_1^2 e^x + 4\cos(x)\mathrm{d}x$$

22. Crear una cuadratura de dos puntos:

$$\int_0^1 f(x)\mathrm{d}x \approx \alpha f(0) + \beta f(1),$$

tal que devuelva valores exactos para $f(x) = 1$ y $f(x) = x^2$. ¿Qué pasa cuando $f(x) = x$?

23. ¿Qué es mejor? ¿Aplicar en el intervalo $[a, b]$ una cuadratura de tres puntos o aplicar cuadratura de dos puntos en $[a, c]$ y luego en $[c, b]$ siendo c un punto interior al intervalo? Justificar la respuesta.

24. En el ejercicio 2 se propone reemplazar la función a integrar por una serie de Taylor e integrarla analíticamente. En cambio, en el ejercicio 21 se propone utilizar dos expansiones y evaluarlas utilizando los extremos de integración. ¿Es el mismo método? Aplicar ambos para aproximar:

$$\int_0^5 \cos(x)\mathrm{d}x$$

y comparar resultados.

25. Utilizando integración numérica, verificar ó refutar el valor de las siguientes integrales:

a) $\int_0^1 \sqrt{x^3}\mathrm{d}x = \frac{2}{5}$

b) $\int_0^1 \dfrac{1}{1+10x^2}\mathrm{d}x = \frac{1}{2}$

c) $\int_0^1 25e^{-25x}\mathrm{d}x = 1$

Bibliografía

- *Análisis numérico*, R. BURDEN y J. FAIRES, Cap.4
- *Análisis numérico con aplicaciones*, C. GERALD y P. WHEATLEY, Cap.5
- *Fundamental numerical methods for electrical engineering**, Stanislaw ROSLONIEC, Cap.5
- *Métodos numéricos con MATLAB*, J. MATHEWS y K. FINK, Cap.7
- *Numerical calculations and algorithms*, R. BECKETT y J. HURT, Cap.5

11 Resolución Numérica de EDO

Una ecuación diferencial es una ecuación que involucra una ó más derivadas de una función desconocida. Si todas las derivadas son tomadas con respecto a una sola variable independiente, se la denomina **ecuación diferencial ordinaria**, con lo que una **ecuación diferencial parcial** (ó en derivadas parciales) es denominada así cuando existen derivadas con respecto a más de una variable.

Una ecuación diferencial (ordinaria ó parcial) tiene **orden** p si p es la máxima cantidad de derivadas, con respecto a cualquier variable involucrada.

La forma general de una ecuación diferencial ordinaria de primer orden es:

$$y' = f(x, y), \tag{11.1}$$

donde $y' = \frac{\mathrm{d}y}{\mathrm{d}x}$ y $f(x, y)$ es una función dada. La solución de esta ecuación contiene una constante arbitraria (constante de integración). Para identificar el valor de esta constante, se debe conocer un punto por el cual pase la curva solución, es decir debe conocerse $y(a) = \alpha$.

Una ecuación diferencial de orden n:

$$y^{(n)} = f(x, y, y', y'', \ldots, y^{(n-1)}), \tag{11.2}$$

se puede transformar en n ecuaciones diferenciales de primer orden. Utilizando la notación:

$$y_1 = y, \quad y_2 = y', \quad y_3 = y'', \quad \ldots \quad y_n = y^{(n-1)}, \tag{11.3}$$

las ecuaciones equivalentes de primer orden son:

$$y_1' = y_2, \quad y_2' = y_3, \quad y_3' = y_4, \quad \ldots \quad y_n' = f(x, y_1, y_2, \ldots, y_n). \tag{11.4}$$

Para resolverlas es necesario conocer n condiciones auxiliares. Si esas condiciones son especificadas en el mismo valor de x, si dice que es un **problema de valor inicial**. Entonces las condiciones auxiliares, llamadas **condiciones iniciales**, tienen la forma:

$$y_1(a) = \alpha_1, \qquad y_2(a) = \alpha_2, \qquad y_3(a) = \alpha_3, \qquad \ldots \qquad y_n(a) = \alpha_n. \tag{11.5}$$

Si los valores de y_i son especificados en diferentes valores de x, el problema es denominado **problema de valor de contorno**.

Ejemplo 74. *La ecuación diferencial:*

$$y'' - y, \quad y(0) = 1, \quad y'(0) = 0$$

es un problema de valor inicial ya que las dos condiciones impuestas sobre la solución están dadas sobre el mismo punto, $x = 0$. Sin embargo:

$$y'' = -y, \quad y(0) = 1, \quad y(\pi) = 0$$

es un problema de valor de contorno ya que las dos condiciones estás especificadas en diferentes valores de x.

11.1. Métodos Básicos

11.1.1. Campos direccionales e isóclinas

Dada la ecuación diferencial ordinaria de primer orden del tipo (11.1), donde f es una función real y continua, $f : D \to \mathbb{R}$, $D \subset \mathbb{R}^2$, puede ser considerada como una fórmula para calcular la pendiente y' en cada punto (x, y) pertenecientes a la región D. El segmento de línea que pasa por (x, y) con pendiente $f(x, y)$, indica la dirección de la tangente al gráfico de la solución en el punto antes mencionado. El conjunto de todos los segmentos tangentes es denominado **campo direccional** de la ecuación diferencial ordinaria.

Para ciertas curvas de D, el campo direccional de una ecuación diferencial ordinaria tiene pendiente constante. Dichas curvas son denominadas **isóclinas** y se identifican haciendo que la pendiente sea constante y arbitraria:

$$f(x, y) = c,$$

donde c será la pendiente asociada a dicha isóclina.

Ejemplo 75. *Sea la ecuación diferencial:*

$$xy' = 2y,$$

que puede ser reescrita como:

$$y' = 2\frac{y}{x} = c,$$

donde c representa la pendiente constante que se puede identificar en cada punto $(x, y) \in D$. Entonces, despejando queda:

$$y(x, c) = \frac{c}{2}x.$$

Si se toma $c = 1$, la recta $y = \frac{x}{2}$ contiene, a lo largo de su curva, segmentos de inclinación (pendiente) igual a 1. Al elegir $c = -0{,}3$ la recta $y = -0{,}15x$ contiene segmentos de pendiente igual a $-0{,}3$. En la figura 11.1 se muestran el campo direccional y algunas de las isóclinas (soluciones particulares de la ecuación diferencial) asociadas a él.

Al resolver en forma exacta, la solución es $y(x) = ax^2$, lo que coincide con el trazado de isóclinas: una familia de parábolas cuyo eje central coincide con $x = 0$.

Ejemplo 76. *La ecuación diferencial $4y' + y = 0$ debe ser resuelta gráficamente. Para ello se construye el campo direccional y algunas isóclinas asociadas a él. Entonces*

$$y(x, c) = -4c.$$

Graficando estas líneas paralelas al eje x cuyos segmentos de pendiente son variables, es que se obtiene el gráfico mostrado en la figura 11.2.

Ejercicio 40. *Resolver, en forma gráfica, las ecuaciones diferenciales:*

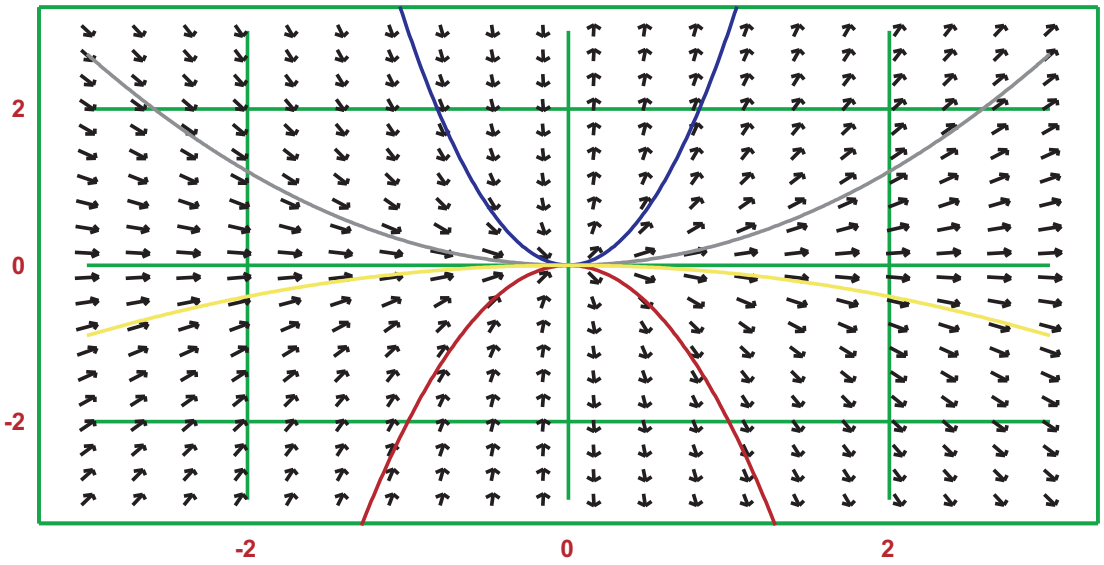

Figura 11.1: Campo direccional de $xy' = 2y$ y algunas isóclinas.

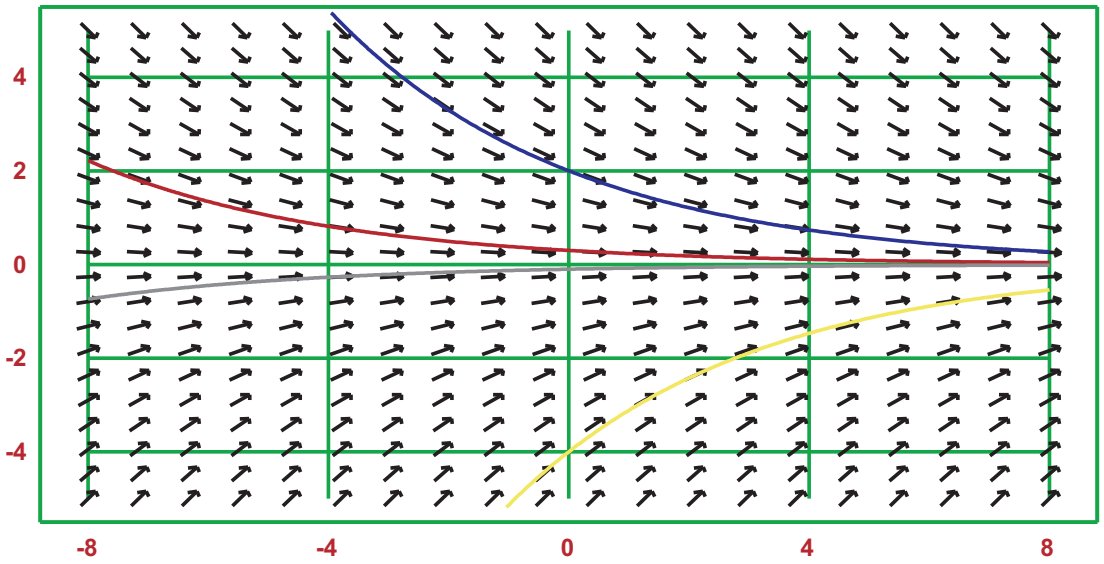

Figura 11.2: Campo direccional de $4y' + y = 0$ y algunas isóclinas.

- $y' = x^2 + y^2$
- $y' = x^2$

Estimar alguna solución particular en el plano $x, y \in [-3, 3]$.

Comandos de EMT. *El comando para representar gráficamente las isóclinas de una ecuación diferencial es:*

- `vectorfield(f$:string, x1:número, x2:número, y1:número, y2:número, nx:número, ny:número)`, *donde* ***f$*** *es la función de* x *e* y *que representa* y', *expresada como* string; ***x1***, ***x2*** *es el dominio para el eje* x; ***y1***, ***y2*** *es el dominio para el eje* y; ***nx***, ***ny*** *son parámetros optativos y permiten seleccionar la partición a utilizar en la gráfica, valor por defecto: 20.*

Ejemplo en EMT 21. *Graficar las isóclinas de la ecuación diferencial* $y' \sin(y) = \cos(x) + 1$, *en* $[-7, 7] \times [-4, 4]$.

```
>vectorfield("(cos(x)+1)/(sin(y))",-7,7,-4,4);
```

La salida se muestra en la figura 11.3.

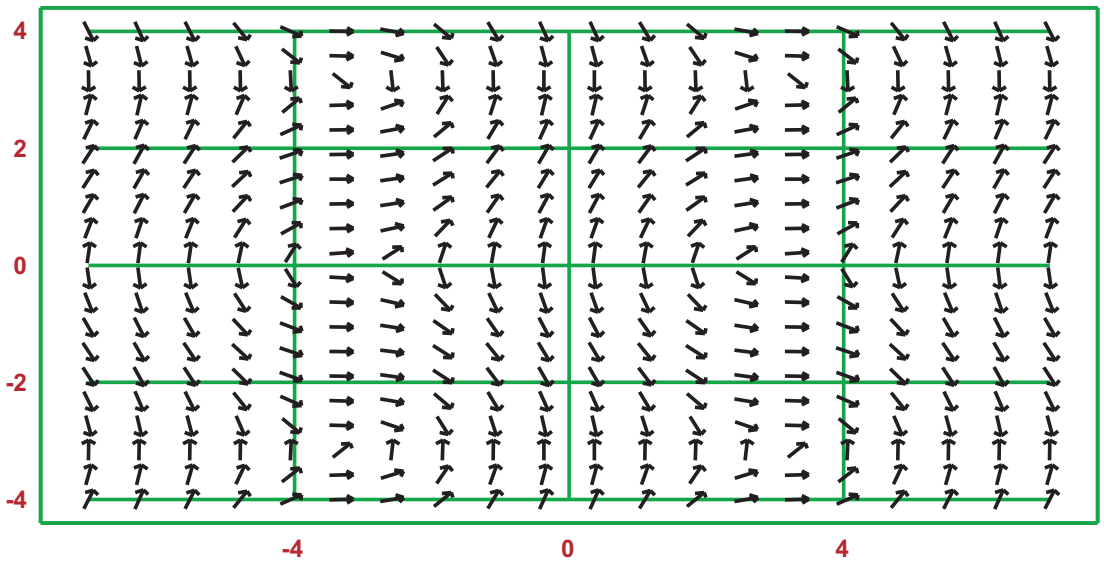

Figura 11.3: Campo direccional de $y' \sin(y) = cos(x) + 1$.

11.1.2. Métodos de Euler y Crank-Nicolson

Si se considera la ecuación diferencial (11.1), se observa que el valor de la derivada es fundamental para resolver el problema propuesto, lo que implica conocer los valores de $y(x)$ para evaluar correctamente f. Es necesario crear un esquema iterativo con el fin de aproximar la derivada y la función $f(x, y)$ a través de pequeños incrementos. Si esto es posible, existirán y se calcularán cotas para el error de truncamiento asociado.

Método de Euler Explícito

Surge al aproximar la derivada de la ecuación (11.1) a través del esquema de derivada hacia adelante. Es decir que:

$$y'_n = f(x_n, y_n)$$
$$\frac{y_{n+1} - y_n}{h} + \mathcal{O}(h) \approx f(x_n, y_n).$$

Luego de algunas operaciones algebraicas se obtiene la expresión conocida como **método de Euler explícito**:

$$y_{n+1} \approx y_n + hf(x_n, y_n) + \mathcal{O}(h),$$

tal que su **error local** de truncamiento es $\mathcal{O}(h^2)$, pero su **error global** de truncamiento es $\mathcal{O}(h)$. La disminución en el orden del error de truncamiento se debe a que cada paso se inicia con el error que se propaga del paso anterior.

Ejemplo 77. *Se desea resolver el problema de valor inicial* $y' = y - x$, $y(0) = -2$, *con el fin de obtener el valor de* $y(1)$. *Para ello se aplicará el método de Euler explícito y se generará un esquema iterativo:*

$$\begin{aligned} y_{n+1} &\approx y_n + hf(x_n, y_n) \\ &\approx y_n + h(y_n - x_n), \end{aligned}$$

cuya evolución, para $h = 0{,}1$, *se muestra en la tabla 11.1. En la figura 11.4 se muestra la aproximación discreta (círculos de color negro y línea continua de color rojo) y la solución exacta (línea punteada de color azul).*

i	x_i	$\tilde{y}_i$	y_i
0	0,0	-2,00000	-2,00000
1	0,1	-2,20000	-2,21551
2	0,2	-2,43000	-2,46420
3	0,3	-2,69300	-2,74957
4	0,4	-2,99230	-3,07547
5	0,5	-3,33153	-3,44616
6	0,6	-3,71468	-3,86635
7	0,7	-4,14614	-4,34125
8	0,8	-4,63075	-4,87662
9	0,9	-5,17382	-5,47880
10	1,0	-5,78120	-6,15484

Tabla 11.1: Resolución del ejemplo 77 a través del método de Euler explícito.

Método de Euler Implícito

Si en el esquema iterativo anterior, generado con el fin de resolver numéricamente la ecuación (11.1) se utilizó la aproximación de la derivada hacia adelante, es posible ahora generar un nuevo esquema utilizando la aproximación de la derivada hacia atrás:

$$\begin{aligned} y'_n &= f(x_n, y_n) \\ \frac{y_n - y_{n-1}}{h} + \mathcal{O}(h) &\approx f(x_n, y_n). \end{aligned}$$

Luego de aplicar operaciones algebraicas y de aumentar en uno cada índice de la expresión iterativa se obtiene la expresión conocida como **método de Euler implícito**:

$$y_{n+1} \approx y_n + hf(x_{n+1}, y_{n+1}) + \mathcal{O}(h),$$

tal que su **error local** de truncamiento es $\mathcal{O}(h^2)$, pero su **error global** de truncamiento es $\mathcal{O}(h)$. La disminución en el orden del error de truncamiento se debe a que cada paso se inicia con el error que se propaga del paso anterior.

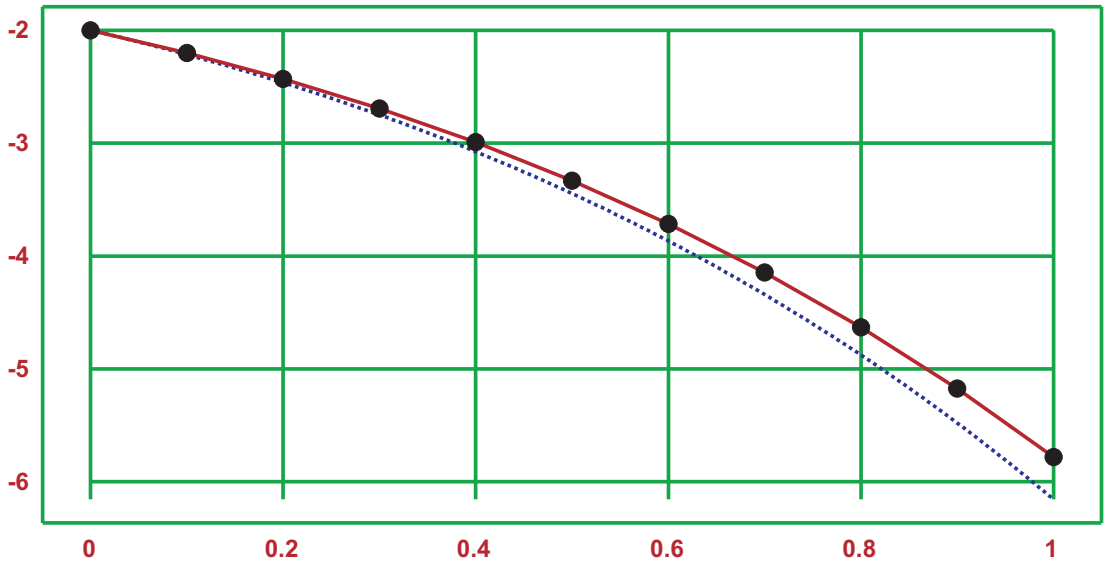

Figura 11.4: Solución numérica (método de Euler explícito) y exacta de $y' = y - x$.

Ejemplo 78. *Se desea resolver nuevamente el problema del ejemplo 77. Esta vez se aplicará el método de Euler implícito y se generará un esquema iterativo:*

$$\begin{aligned} y_{n+1} &\approx y_n + hf(x_{n+1}, y_{n+1}) \\ &\approx y_n + h(y_{n+1} - x_{n+1}) \\ &\approx \frac{y_n - hx_{n+1}}{1-h} \end{aligned}$$

cuya evolución, para $h = 0{,}1$, se muestra en la tabla 11.2. En la figura 11.5 se muestra la aproximación discreta (círculos de color negro y línea continua de color rojo) y la solución exacta (línea punteada de color azul).

i	x_i	$\tilde{y}_i$	y_i
0	0,0	-2,00000	-2,00000
1	0,1	-2,23333	-2,21551
2	0,2	-2,50370	-2,46420
3	0,3	-2,81522	-2,74957
4	0,4	-3,17246	-3,07547
5	0,5	-3,58051	-3,44616
6	0,6	-4,04501	-3,86635
7	0,7	-4,57223	-4,34125
8	0,8	-5,16914	-4,87662
9	0,9	-5,84348	-5,47880
10	1,0	-6,60386	-6,15484

Tabla 11.2: Resolución del ejemplo 78 a través del método de Euler implícito.

Método de Crank-Nicolson

En el desarrollo de los métodos de Euler, se utilizó $f(x_n, y_n)$ y $f(x_{n+1}, y_{n+1})$ para los métodos explícito e implícito respectivamente. Como en uno de ellos se cometió error por exceso y en el otro error por defecto, es lógico pensar que evaluando f en el punto medio

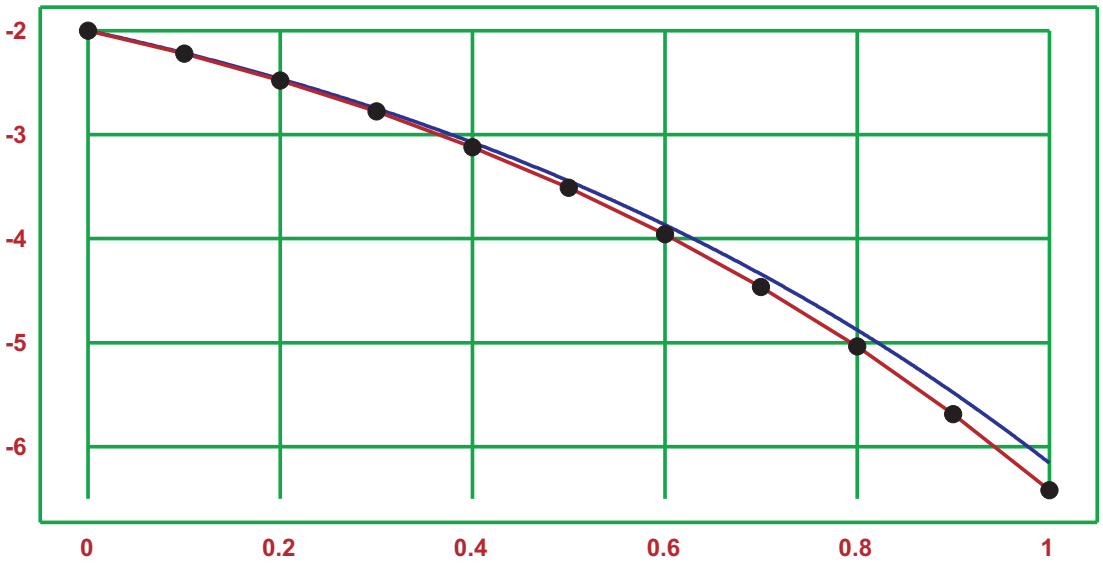

Figura 11.5: Solución numérica (método de Euler implícito) y exacta de $y' = y - x$.

entre (x_n, y_n) y (x_{n+1}, y_{n+1}) se obtenga una aproximación con un error más pequeño que las dos anteriores. En este desarrollo se basa el **método de Crank-Nicolson**:

$$y_{n+1} \approx y_n + hf\left(\frac{x_n + x_{n+1}}{2}, \frac{y_n + y_{n+1}}{2}\right)$$

Ejemplo 79. *Se desea resolver nuevamente el problema del ejemplo 77. Esta vez se aplicará el método de Crank-Nicolson y se generará un esquema iterativo:*

$$\begin{aligned} y_{n+1} &\approx y_n + hf((x_n + x_{n+1})/2, (y_n + y_{n+1})/2) \\ &\approx y_n + h((y_n + y_{n+1})/2 - (x_n + x_{n+1})/2) \\ &\approx \frac{y_n + \frac{h}{2}(y_n - x_n - x_{n+1})}{1 - \frac{h}{2}} \end{aligned}$$

cuya evolución, para $h = 0{,}1$, se muestra en la tabla 11.3. En la figura 11.6 se muestra la aproximación discreta (círculos de color negro y línea continua de color rojo) y la solución exacta (línea punteada de color azul). En este caso, la diferencia entre el valor exacto y la aproximación es muy pequeña.

Ejercicio 41. *Estimar el orden de convergencia del método de Crank-Nicolson.*

Los tres métodos presentados son convergentes, es decir que cuando $h \to 0$ se logra que $E(h) \to 0$. En la figura 11.7 se muestra el error absoluto que surge de la diferencia entre la secuencia de valores aproximados y los valores exactos. En color azul y con línea punteada se grafica el error con respecto al método de Euler explícito; en color rojo y con línea discontinua, el error con respecto al método de Euler implícito; y en color negro con línea alternada, el error con respecto al método de Crank-Nicolson.

No siempre es posible encontrar una expresión iterativa para el desarrollo de los métodos de Euler implícito y Crank-Nicolson, y por más que sea posible es a veces es un proceso tedioso. Por lo tanto es necesario ajustar los métodos a los datos y a las aproximaciones disponibles.

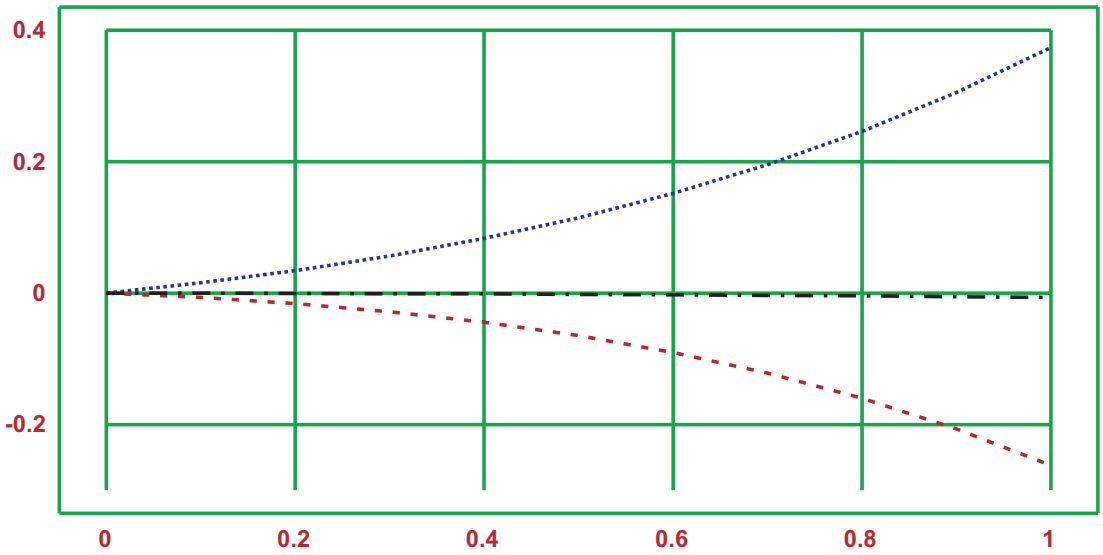

Figura 11.7: Errores absolutos de los métodos de Euler y Crank-Nicolson para $y' = y - x$.

esquema iterativo (también denominado Euler implícito ó ***esquema de Euler Predictor Corrector****). Se observa que la diferencia de estimaciones entre ambos métodos es despreciable.*

i	x_i	$\tilde{y}_i$	y_i
0	0,0	1,00000	1,00000
1	0,1	1,09900	1,09965
2	0,2	1,19483	1,19721
3	0,3	1,28516	1,29046
4	0,4	1,36767	1,37712
5	0,5	1,44016	1,45499
6	0,6	1,50064	1,52196
7	0,7	1,54746	1,57617
8	0,8	1,57934	1,61607
9	0,9	1,59545	1,64049
10	1,0	1,59545	1,64872

Tabla 11.4: Resolución del ejemplo 80 a través del esquema de Euler Predictor Corrector.

Ejercicio 42. *Resolver nuevamente el ejercicio 80 pero esta vez generar un esquema predictor corrector basado en el método de Crank-Nicolson.*

11.1.3. Métodos Runge-Kutta

El objetivo de los métodos Runge-Kutta es eliminar la necesidad de repetir derivadas de las ecuaciones diferenciales que surgen de las aproximaciones por series de Taylor. Como dicha diferenciación no está presente en la integración del método de Taylor de primer orden:

$$y_{n+1} = y_n + hf(x_n, y_n), \tag{11.6}$$

puede ser considerado como el **método de primer orden de Runge-Kutta**. También es denominado *método de Euler explícito*. Debido al gran error de truncamiento, rara

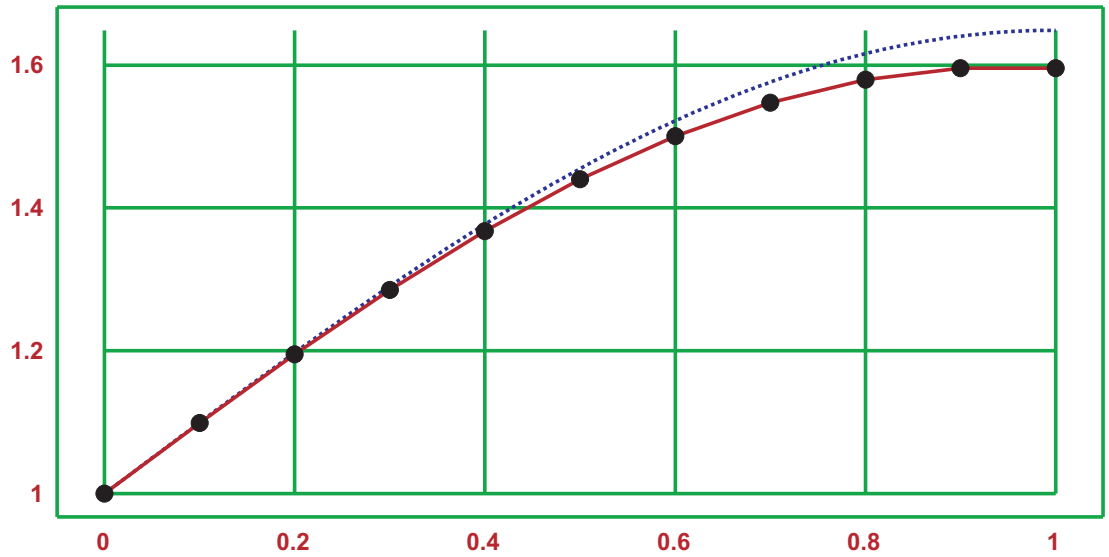

Figura 11.8: Solución numérica (esquema de Euler Predictor Corrector) y exacta de $y' = y - xy$.

vez es utilizado en la práctica, sólo se aplica con fines pedagógicos. Si se analiza la interpretación gráfica de la fórmula de Euler, plasmado en la figura 11.10, se observa que el error cometido es realmente excesivo. Analizando la fórmula (11.6), la diferencia entre y_{n+1} y y_n puede calcularse por integración:

$$y_{n+1} - y_n = \int_{x_n}^{x_{n+1}} y' \, \mathrm{d}x = \int_{x_n}^{x_{n+1}} f(x,y) \, \mathrm{d}x,$$

donde dependiendo del método de integración numérica elegido, surgen los diferentes métodos Runge-Kutta.

Método de Euler

El método de Euler, o RK1, surge de despejar $y_{n+1} = y(x_{n+1}) = y(x_n + h)$ en la expresión de Taylor. Sin embargo, se llega a la misma expresión si se integra $\int_{x_n}^{x_{n+1}} f(x,y) \, \mathrm{d}x$ numéricamente por el método del rectángulo[1]:

$$\begin{aligned} y_{n+1} - y_n &= \int_{x_n}^{x_{n+1}} f(x,y) \, \mathrm{d}x \\ &\approx [x_{n+1} - x_n] \, f(x_n, y_n) \\ &= h f(x_n, y_n), \end{aligned}$$

con lo que :

$$y_{n+1} \approx y_n + h f(x_n, y_n) \tag{11.7}$$

Método de Heun

El método de Heun, o RK2, surge al integrar $\int_{x_n}^{x_{n+1}} f(x,y) \, \mathrm{d}x$ numéricamente por el método del trapecio. La principal diferencia con Euler implícito es que no se despeja y_{n+1} sino que se aproxima su valor por medio del método de Euler, es decir que éste

[1] primera aproximación de la integral por rectángulos

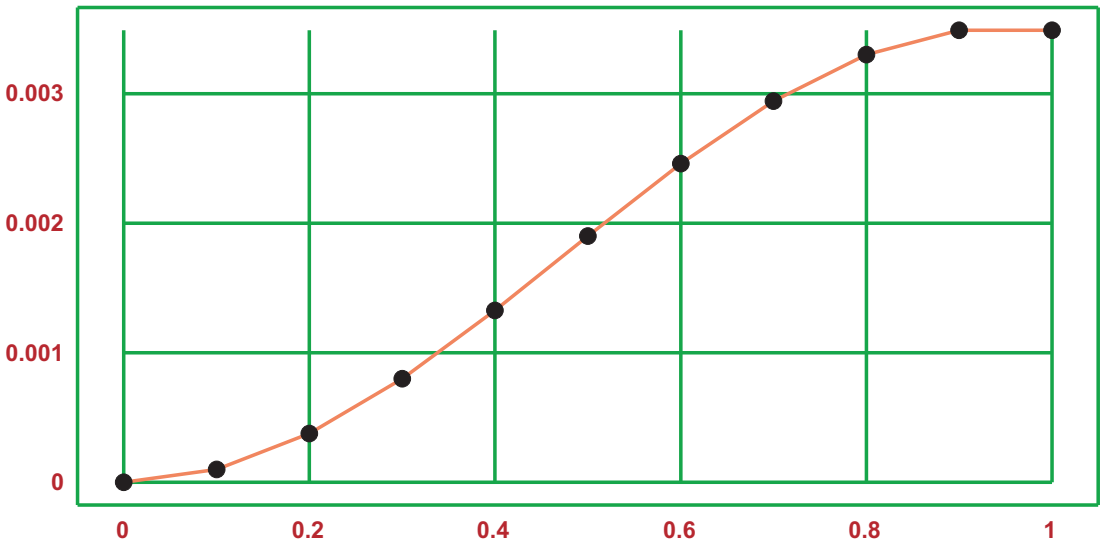

Figura 11.9: Diferencia entre el método de Euler implícito y el esquema de Euler Precictor Corrector para $y' = y - xy$.

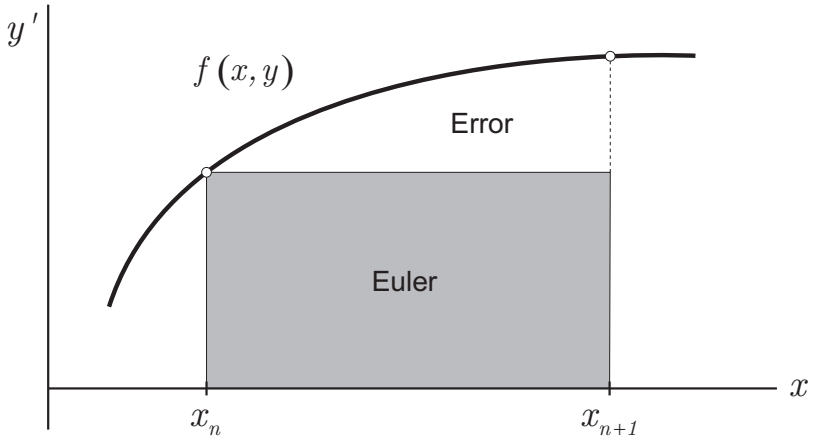

Figura 11.10: Área bajo la curva $y' = f(x, y)$, entre x_n y x_{n+1}.

también es un esquema del tipo predictor corrector:

$$\begin{aligned} y_{n+1} - y_n &= \int_{x_n}^{x_{n+1}} f(x,y)\,\mathrm{d}x \\ &\approx [x_{n+1} - x_n]\,\frac{f(x_n,y_n) + f(x_{n+1},y_{n+1})}{2} \\ &= h\frac{f(x_n,y_n) + f(x_{n+1},y_{n+1})}{2}, \end{aligned}$$

con lo que:

$$y_{n+1} \approx y_n + \frac{h}{2}\left[f(x_n,y_n) + f(x_{n+1},\tilde{y_{n+1}})\right], \tag{11.8}$$

donde:

$$\tilde{y_{n+1}} = y_n + hf(x_n,y_n).$$

Ejemplo 81. *Se desea resolver la ecuación diferencial del ejemplo 80, utilizando RK2. Para ello debe construirse la función de iteración que aproximará los valores sucesivos*

a partir de $y(0)$ *hasta* $y(1)$. *La función de iteración es:*

$$\begin{aligned} y_{n+1} &\approx y_n + \frac{h}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1}^p)\right] \\ &= y_n + \frac{h}{2}\left[y_n - x_n y_n + \tilde{y_{n+1}} - x_{n+1}\tilde{y_{n+1}}\right], \end{aligned}$$

donde:

$$\begin{aligned} \tilde{y_{n+1}} &\approx y_n + hf(x_n, y_n) \\ &\approx y_n + h(y_n - x_n y_n). \end{aligned}$$

Primero se calculará la predicción y luego se hará la corrección. Los valores de las iteraciones se muestran en la tabla 11.5. En la figura 11.11 se muestra la aproximación discreta (círculos de color negro y línea continua de color rojo) y la solución exacta (línea punteada de color azul).

i	x_i	$pred(\tilde{y}_i)$	$correc(\tilde{y}_i)$	y_i
0	0,0	1,00000	1,00000	1,00000
1	0,1	1,10000	1,09950	1,09965
2	0,2	1,19845	1,19692	1,19721
3	0,3	1,29267	1,29004	1,29046
4	0,4	1,38034	1,37660	1,37712
5	0,5	1,45919	1,45437	1,45499
6	0,6	1,52709	1,52128	1,52196
7	0,7	1,58213	1,57543	1,57617
8	0,8	1,62270	1,61529	1,61607
9	0,9	1,64760	1,63968	1,64049
10	1,0	1,65608	1,64788	1,64872

Tabla 11.5: Resolución del ejemplo 80 a través del método de Heun.

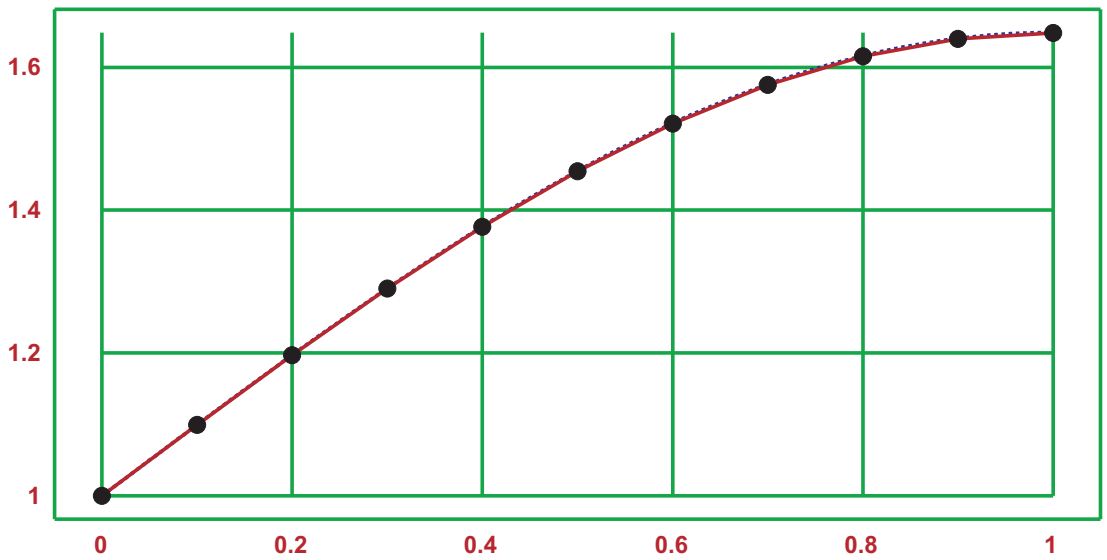

Figura 11.11: Solución numérica (Método de Heun) y exacta de $y' = y - xy$.

Comandos de EMT. *El comando para resolver una EDO a través del método de Heun es:*

- `heun(f$:string, t:vector, y0:número)`, *donde* ***f$*** *es la función de* x *e* y *que representa* y'*, expresada como* string*;* ***t*** *es el dominio discreto sobre el que se resolverá la EDO (nodos);* ***y0*** *es el valor inicial de la EDO. La salida es un vector con las aproximaciones en cada nodo.*

Ejemplo en EMT 22. *Resolver, a través del método de Heun, la EDO* $y' = (x+y)^2$ *con la condición inicial* $y(-2) = 1$ *para estimar el valor de* $y(-1)$*. Utilizar 15 pasos.*

```
>heun("(x+y)^2",linspace(-2,-1,15),1)
 [1,  1.05851,  1.1032,  1.13709,  1.16245,  1.18106,  1.19431,
 1.20334,  1.20912,  1.21244,  1.21403,  1.21455,  1.21459,  1.21477,
 1.21568,  1.21795]
```

Método RK4

De manera similar a los anteriores, se consigue desarrollar el método más conocido e implementado en rutinas numéricas: RK4. La integral $\int_{x_n}^{x_{n+1}} f(x,y)\,\mathrm{d}x$, debe ser resuelta ahora utilizando el método de Simpson[2]. Por lo tanto:

$$\begin{aligned} y_{n+1} - y_n &= \int_{x_n}^{x_{n+1}} f(x,y)\,\mathrm{d}x \\ &\approx \frac{[x_{n+1}-x_n]/2}{3}\left[f(x_n,y_n) + 4f\left(x_n+\frac{h}{2}, y_n+\frac{h}{2}\right) + f(x_{n+1},y_{n+1})\right], \end{aligned}$$

donde $k_1 = f(x_n, y_n)$. Se puede estimar $y_n + \frac{h}{2}$ por Euler:

$$y_n + \frac{h}{2} \approx y_n + k_1\frac{h}{2}$$

y $k_2 = f(x_n + \frac{h}{2}, y_n + k_1\frac{h}{2})$. Ahora se puede aproximar $y_n + \frac{h}{2}$ por Euler, pero con pendiente k_2. Entonces

$$y_n + \frac{h}{2} \approx y_n + k_2\frac{h}{2},$$

de donde $k_3 = f(x_n + \frac{h}{2}, y_n + k_2\frac{h}{2})$. Por último, se estima y_{n+1} por Euler con pendiente k_3 y se evalúa f una vez más, $k_4 = f(x_{n+1}, y_n + k_3 h)$. Sustituyendo el término que tiene factor 4 de la fórmula de Simpson por dos sumandos, cada uno con factor 2, se logra el algoritmo de Runge-Kutta.

$$y_{n+1} = y_n + \frac{h}{6}\left[k_1 + 2k_2 + 2k_3 + k_4\right].$$

Ejemplo 82. *Se desea resolver la ecuación diferencial del ejemplo 80, utilizando RK4. Para ello debe construirse la función de iteración que aproximará los valores sucesivos a partir de* $y(0)$ *hasta* $y(1)$*. La función de iteración es:*

$$y_{n+1} = y_n + \frac{h}{6}\left[k_1 + 2k_2 + 2k_3 + k_4\right]$$

donde:

$$\begin{aligned} k_1 &= f\left(x_n, y_n\right) \\ k_2 &= f\left(x_n + \frac{h}{2}, y_n + k_1\frac{h}{2}\right) \\ k_3 &= f\left(x_n + \frac{h}{2}, y_n + k_2\frac{h}{2}\right) \\ k_4 &= f\left(x_{n+1}, y_n + k_3 h\right) \end{aligned}$$

[2] con una pequeña corrección para aumentar más la precisión

Los valores de las iteraciones se muestran en la tabla 11.6. En la figura 11.12 se muestra la aproximación discreta (círculos de color negro y línea continua de color rojo) y la solución exacta (línea punteada de color azul). En este caso, la aproximación es casi exacta.

i	x_i	k_1	k_2	k_3	k_4	$\tilde{y}_i$	y_i
0	0,0	-	-	-	-	1,00000	1,00000
1	0,1	1,00000	0,99750	0,99738	0,98976	1,09966	1,09965
2	0,2	0,98969	0,97677	0,97622	0,95783	1,19722	1,19721
3	0,3	0,95777	0,93383	0,93293	0,90336	1,29046	1,29046
4	0,4	0,90332	0,86816	0,86702	0,82630	1,37713	1,37712
5	0,5	0,82628	0,78014	0,77887	0,72751	1,45499	1,45499
6	0,6	0,72750	0,67112	0,66985	0,60879	1,52196	1,52196
7	0,7	0,60879	0,54334	0,54220	0,47285	1,57617	1,57617
8	0,8	0,47285	0,39995	0,39904	0,32322	1,61607	1,61607
9	0,9	0,32322	0,24484	0,24425	0,16405	1,64050	1,64049
10	1,0	0,16405	0,08244	0,08223	0,00000	1,64872	1,64872

Tabla 11.6: Resolución del ejemplo 80 a través del método de RK4.

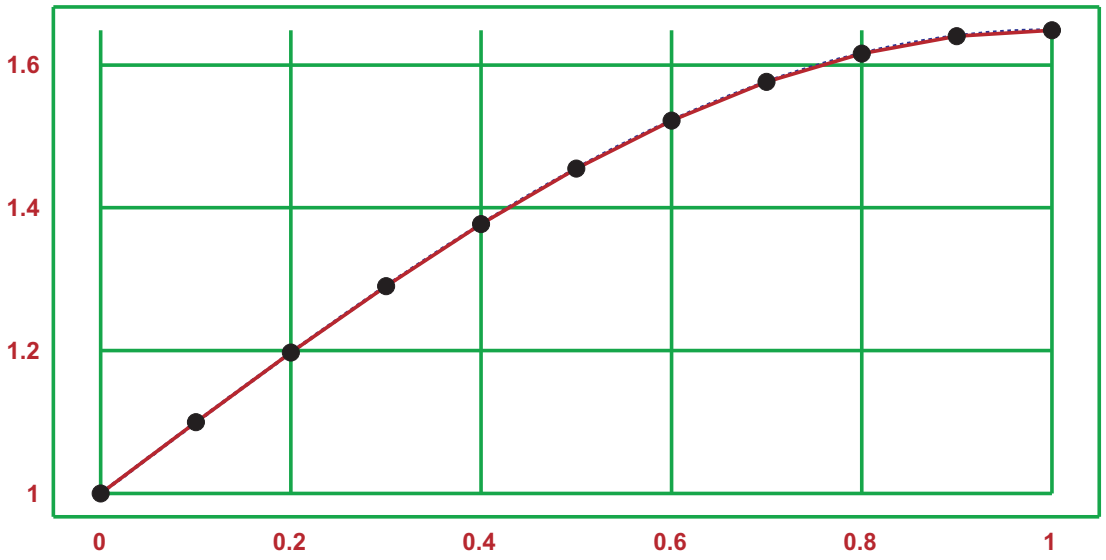

Figura 11.12: Solución numérica (Método RK4) y exacta de $y' = y - xy$.

Ejercicio 43. *Determinar el orden de convergencia de los métodos Runge-Kutta, basándose en el orden de convergencia de los métodos de integración. Comprobar numéricamente con los ejercicios desarrollados.*

Comandos de EMT. *El comando para resolver una EDO a través del método RK4 es:*

- `runge(f$:string, t:vector, y0:número)`, *donde* ***f$*** *es la función de* x *e* y *que representa* y', *expresada como* string; ***t*** *es el dominio discreto sobre el que se resolverá la EDO (nodos);* ***y0*** *es el valor inicial de la EDO. La salida es un vector con las aproximaciones en cada nodo.*

Ejemplo en EMT 23. *Resolver, a través del método RK4, la EDO* $y' = (x+y)^2$ *con la condición inicial* $y(-2) = 1$ *para estimar el valor de* $y(-1)$. *Utilizar 15 pasos.*

```
>runge("(x+y)^2",linspace(-2,-1,15),1)
 [1,  1.05851,  1.1032,  1.13709,  1.16245,  1.18106,  1.19431,
 1.20335,  1.20912,  1.21245,  1.21404,  1.21455,  1.2146,  1.21478,
 1.21569,  1.21796]
```

11.2. Resolución por Derivación

Dada una ecuación diferencial ordinaria es posible resolverla utilizando el concepto de derivación. La primera familia de métodos surge de la conocida expansión por series de Taylor, aprovechando la función dy/dx definida en el problema original. La segunda familia de métodos es obtenida al derivar el polinomio interpolante que pasa a través de los puntos anteriores al actual de la ecuación diferencial a resolver.

11.2.1. Series de Taylor

Si se considera la expansión por serie de Taylor de $y(x+h)$:

$$y(x+h) = y(x) + hy'(x) + \frac{h^2 y''(x)}{2!} + \frac{h^3 y'''(x)}{3!} + \dots,$$

entonces es posible construir una fórmula iterativa de aproximación para resolver el problema (11.1):

$$y_{n+1} = y_n + hf(x_n, y_n) + \frac{h^2}{2!}\frac{\mathrm{d}f(x_n,y_n)}{\mathrm{d}x} + \frac{h^3}{3!}\frac{\mathrm{d}^2 f(x_n,y_n)}{\mathrm{d}x^2} + \dots,$$

donde:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x} = f^{(I)}$$
$$\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} = \frac{\mathrm{d}f^{(I)}}{\mathrm{d}x} = \frac{\partial f^{(I)}}{\partial x} + \frac{\partial f^{(I)}}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x} = f^{(II)},$$

y así sucesivamente. Es importante notar que en todos los términos de derivada aparece $\frac{\mathrm{d}y}{\mathrm{d}x}$, cuya expresión es conocida.

Este esquema iterativo permite deducir, nuevamente, el método de Euler como la expansión de Taylor de orden 1. Con órdenes superiores provee una solución eficiente al problema de resolver un ecuación diferencial ordinaria con valor inicial, aunque no siempre simple de calcular, teniendo en cuenta que el proceso de derivación de $f(x,y)$ se torna tedioso[3] con el aumento del orden de expansión. Una de las mejoras introducidas al método de Taylor son los métodos de Runge-Kutta, aunque también es posible mantener el esquema iterativo de expansión de Taylor aproximando en cada caso $\frac{\mathrm{d}^{(n)}f}{\mathrm{d}x^n}$ a través de diferencias finitas.

Ejemplo 83. *Se desea resolver la ecuación diferencial* $y' = \cos(x+y) - 3y$ *con la condición inicial* $y(0) = 2$*, para obtener el valor de* $y(6)$*. Para solucionar este problema a través de expansiones de Taylor de orden 2, es necesario obtener* $f'(x,y)$*:*

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}x} &= \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x} = f^{(I)}(x,y)\\ &= -\sin(x+y) + [-\sin(x+y) - 3]\,[\cos(x+y) - 3y]\\ &= -\sin(x+y)\,[1 + \cos(x+y) - 3y] - 3\cos(x+y) + 9y\end{aligned}$$

entonces:

$$y(x+h) \approx y(x) + hf(x,y) + \frac{h^2}{2!}f^{(I)}(x,y) + \mathcal{O}(h^3),$$

[3] y es imposible para una función dada en forma de tabla

con lo que puede generarse el esquema iterativo:

$$y_{n+1} \approx y_n + hf(x_n, y_n) + \frac{h^2}{2!} f^{(I)}(x_n, y_n) + \mathcal{O}(h^3). \tag{11.9}$$

Para $h = 0{,}4$, los valores de las iteraciones se muestran en la tabla 11.7 y la figura 11.13 muestra la solución exacta *(en línea punteada y de color azul) y la aproximación de Taylor de orden 2 (círculos negros y línea continua en color rojo).*

i	x_i	$\tilde{y}_i$	y_i
0	0,0	2,000000	2,000000
1	0,4	1,367410	0,665208
2	0,8	0,938520	0,313526
3	1,2	0,617694	0,165885
4	1,6	0,367233	0,057333
5	2,0	0,165180	-0,043520
6	2,4	-0,000025	-0,137897
7	2,8	-0,132189	-0,220344
8	3,2	-0,229610	-0,283212
9	3,6	-0,286646	-0,317009
10	4,0	-0,294742	-0,310339
11	4,4	-0,244107	-0,251447
12	4,8	-0,129656	-0,134856
13	5,2	0,033513	0,024117
14	5,6	0,194031	0,179689
15	6,0	0,291024	0,280303

Tabla 11.7: Resolución del ejemplo 83 a través de la expansión de Taylor de orden 2.

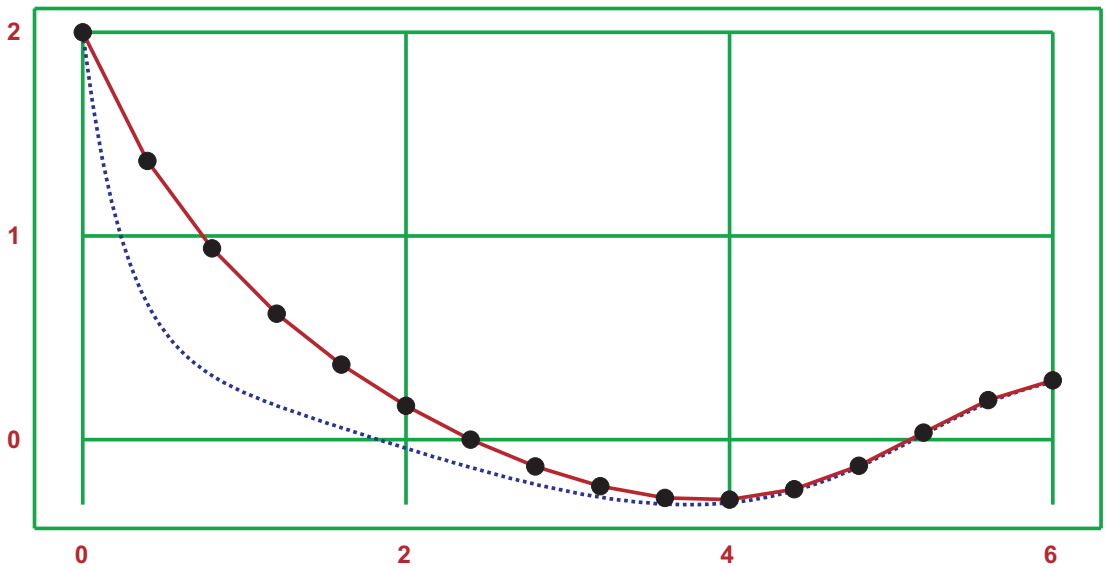

Figura 11.13: Solución numérica (expansión de Taylor de orden 2) y exacta de $y' = \cos(x + y) - 3y$.

Nota. *Por ser $y' = \cos(x + y) - 3y$ una ecuación diferencial que no puede resolverse por métodos algebraicos, cuando se nombra la solución* exacta *quiere decir una solución numérica calculada con métodos de gran potencia y un incremento h pequeño.*

Ejercicio 44. *Resolver nuevamente el problema planteado en el ejemplo anterior, pero esta vez utilizar la aproximación de derivada hacia adelante:*

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}x} &= \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x} \\ &\approx \frac{f(x+h,y)-f(x,y)}{h} + \frac{f(x,y+h)-f(x,y)}{h} f(x,y).\end{aligned}$$

Luego, graficar la diferencia entre el valor de $\tilde{y_{n+1}}$ utilizando la derivada de $f(x,y)$ calculada en forma exacta y el valor de $\tilde{y_{n+1}}$ con la aproximación de derivada de $f(x,y)$.

Ejemplo 84. *Se desea resolver nuevamente el ejemplo 83, pero ahora utilizando la expansión de Taylor de orden 3. Debe para ello calcularse $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}$:*

$$\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} = \frac{\mathrm{d}f^{(I)}}{\mathrm{d}x} = \frac{\partial f^{(I)}}{\partial x} + \frac{\partial f^{(I)}}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x} = f^{(II)}(x,y),$$

donde:

$$\begin{aligned}\frac{\partial f^{(I)}}{\partial x} &= -\cos(x+y)\left[1+\cos(x+y)-3y\right] + \sin^2(x+y) + 3\sin(x+y) \\ \frac{\partial f^{(I)}}{\partial y} &= -\cos(x+y)\left[1+\cos(x+y)-3y\right] - \sin(x+y)\left[-\sin(x+y)-3\right] \\ \frac{\mathrm{d}y}{\mathrm{d}x} &= \cos(x+y) - 3y.\end{aligned}$$

Entonces el esquema iterativo a utilizar es:

$$y_{n+1} = y_n + hf(x_n,y_n) + \frac{h^2}{2!}f^{(I)}(x_n,y_n) + \frac{h^3}{3!}f^{(II)}(x_n,y_n) + \mathcal{O}(h^4). \tag{11.10}$$

Para $h = 0{,}4$, los valores de las iteraciones se muestran en la tabla 11.8 y la figura 11.14 muestra la solución exacta *(en línea punteada y de color azul) y la aproximación de Taylor de orden 2 (círculos negros y línea continua en color rojo).*

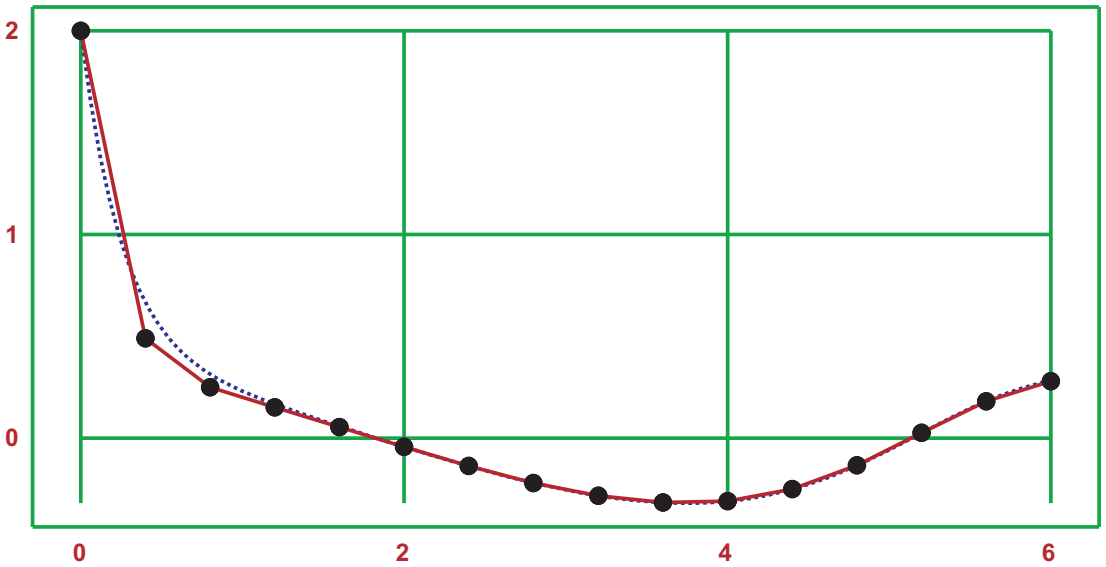

Figura 11.14: Solución numérica (expansión de Taylor de orden 3) y exacta de $y' = \cos(x+y) - 3y$.

i	x_i	$\tilde{y}_i$	y_i
0	0,0	2,000000	2,000000
1	0,4	0,489613	0,665208
2	0,8	0,248800	0,313526
3	1,2	0,150215	0,165885
4	1,6	0,054069	0,057333
5	2,0	-0,044172	-0,043520
6	2,4	-0,138050	-0,137897
7	2,8	-0,220420	-0,220344
8	3,2	-0,283297	-0,283212
9	3,6	-0,317110	-0,317009
10	4,0	-0,310384	-0,310339
11	4,4	-0,251172	-0,251447
12	4,8	-0,133717	-0,134856
13	5,2	0,026129	0,024117
14	5,6	0,180395	0,179689
15	6,0	0,278530	0,280303

Tabla 11.8: Resolución del ejemplo 84 a través de la expansión de Taylor de orden 3.

Ejercicio 45. *Repetir el ejercicio 77, pero esta vez aplicar la expansión de orden 4 de Taylor. ¿Mejora el resultado obtenido al aplicar el método de Crank-Nicolson?*

Ejercicio 46. *Los esquemas iterativos (11.9) y (11.10) son desarrollados hasta el orden del error de truncamiento asociado. ¿Este error es local o global? ¿Por qué?*

11.2.2. Fórmulas de Diferenciación hacia Atrás

En la literatura, son conocidos como los métodos *BDF*, sigla de *Backward Differentiation Formula* y es posible definir esta familia sólo hasta la fórmula de 6 pasos, ya que para 7 pasos (ó más), la fórmula es completamente inestable o bien su región de convergencia estable es muy pequeña[4]. Todos los métodos obtenidos de esta manera son esquemas implícitos y no son autoiniciables, con excepción de *BDF1* siempre y cuando se aplique a una ecuación diferencial ordinaria lineal. Al momento de generar los valores necesarios para la interpolación, se recomienda utilizar algún método de alto orden ó, en el peor de los casos, un método del mismo orden que la aproximación de *BDF* que se esté por utilizar.

Fórmula de un paso

A partir de los nodos (x_n, y_n) y (x_{n+1}, y_{n+1}), es posible construir el polinomio interpolante de grado uno que pasa a través de ellos:

$$P(x) = \frac{x - x_{n+1}}{x_n - x_{n+1}} y_n + \frac{x - x_n}{x_{n+1} - x_n} y_{n+1}.$$

Derivando $P(x)$ con respecto a x y reemplazando $x_n = x_{n+1} - h$:

$$P'(x) = \frac{y_{n+1} - y_n}{h},$$

[4] la explicación de este tema está fuera de los alcances de esta asignatura

Pero si $P(x)$ interpola los nodos de $y(x)$, entonces $P'(x)$ debe interpolar nodos de $y'(x) = f(x, y)$, en particular permite aproximar $f(x_{n+1}, y_{n+1})$:

$$\begin{aligned} P'(x_{n+1}) &= f(x_{n+1}, y_{n+1}) \\ \frac{y_{n+1} - y_n}{h} &= f(x_{n+1}, y_{n+1}). \end{aligned}$$

De la igualdad anterior es posible obtener la *Fórmula de diferenciación hacia atrás de un paso*, más conocida como *BDF1*:

$$y_{n+1} - y_n = hf(x_{n+1}, y_{n+1}). \tag{11.11}$$

Fórmula de dos pasos

A partir de los nodos (x_{n-1}, y_{n-1}); (x_n, y_n) y (x_{n+1}, y_{n+1}), es posible construir el polinomio interpolante de grado dos que pasa a través de ellos:

$$\begin{aligned} P(x) = \frac{(x - x_n)(x - x_{n+1})}{(x_{n-1} - x_n)(x_{n-1} - x_{n+1})} y_{n-1} + \frac{(x - x_{n-1})(x - x_{n+1})}{(x_n - x_{n-1})(x_n - x_{n+1})} y_n + \\ + \frac{(x - x_{n-1})(x - x_n)}{(x_{n+1} - x_{n-1})(x_{n+1} - x_n)} y_{n+1}. \end{aligned}$$

Derivando $P(x)$ con respecto a x y reemplazando $x_{n-1} = x_{n+1} - 2h$ y $x_n = x_{n+1} - h$:

$$P'(x) = \frac{y_{n+1}(-2x_{n+1} + 2x + 3h) + y_n(4x_{n+1} - 4x - 4h) + y_{n-1}(-2x_{n+1} + 2x + h)}{2h^2},$$

Pero si $P(x)$ interpola los nodos de $y(x)$, entonces $P'(x)$ debe interpolar nodos de $y'(x) = f(x, y)$, en particular permite aproximar $f(x_{n+1}, y_{n+1})$:

$$\begin{aligned} P'(x_{n+1}) &= f(x_{n+1}, y_{n+1}) \\ \frac{3y_{n+1} - 4y_n + y_{n-1}}{2h} &= f(x_{n+1}, y_{n+1}). \end{aligned}$$

De la igualdad anterior es posible obtener la *Fórmula de diferenciación hacia atrás de dos pasos*, más conocida como *BDF2*:

$$3y_{n+1} - 4y_n + y_{n-1} = 2hf(x_{n+1}, y_{n+1}). \tag{11.12}$$

Ejemplo 85. *Se desea resolver la ecuación diferencial $y' = x(1 - y)$, con la condición inicial $y(0) = 0$, para estimar el valor de $y(3)$. Se utilizará el esquema iterativo* BDF *de dos pasos, (11.12), con un conjunto de 11 nodos. Entonces $h = 0{,}3$; $y_0 = 0$ y además:*

$$y_1 = 0{,}044002518.$$

En la tabla 11.9 se muestran los valores de las iteraciones y la solución exacta. En la figura 11.15 se muestran la solución exacta (línea punteada de color azul) y la solución numérica (nodos negros y línea continua de color rojo).

i	x_i	$\tilde{y}_i$	y_i
0	0,0	0,00000000	0,00000000
1	0,3	0,04400251	0,04400251
2	0,6	0,15952680	0,16472978
3	0,9	0,32036856	0,33302318
4	1,2	0,49514716	0,51324774
5	1,5	0,65646668	0,67534753
6	1,8	0,78694107	0,80210130
7	2,1	0,88058629	0,88974947
8	2,4	0,94040632	0,94386523
9	2,7	0,97425087	0,97387859
10	3,0	0,99095774	0,98889100

Tabla 11.9: Resolución del ejemplo 85 a través del método *BDF* de dos pasos.

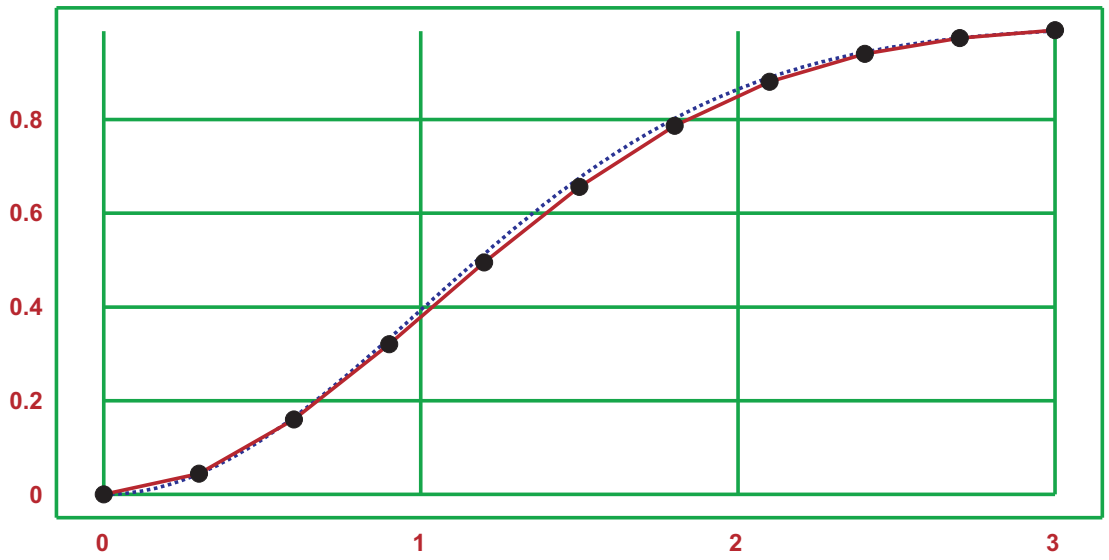

Figura 11.15: Solución numérica (método *BDF* de dos pasos) y exacta de $y' = x(1-y)$.

Fórmula de tres pasos

A partir de los nodos (x_{n-2}, y_{n-2}); (x_{n-1}, y_{n-1}); (x_n, y_n) y (x_{n+1}, y_{n+1}), es posible construir el polinomio interpolante de grado tres que pasa a través de ellos:

$$\begin{aligned} P(x) = & \frac{(x-x_{n-1})(x-x_n)(x-x_{n+1})}{(x_{n-2}-x_{n-1})(x_{n-2}-x_n)(x_{n-2}-x_{n+1})} y_{n-2} + \\ & + \frac{(x-x_{n-2})(x-x_n)(x-x_{n+1})}{(x_{n-1}-x_{n-2})(x_{n-1}-x_n)(x_{n-1}-x_{n+1})} y_{n-1} + \\ & + \frac{(x-x_{n-2})(x-x_{n-1})(x-x_{n+1})}{(x_n-x_{n-2})(x_n-x_{n-1})(x_n-x_{n+1})} y_n + \\ & + \frac{(x-x_{n-2})(x-x_{n-1})(x-x_n)}{(x_{n+1}-x_{n-2})(x_{n+1}-x_{n-1})(x_{n+1}-x_n)} y_{n+1} \end{aligned}$$

Derivando $P(x)$ con respecto a x y reemplazando $x_{n-2} = x_{n+1} - 3h$; $x_{n-1} = x_{n+1} - 2h$ y $x_n = x_{n+1} - h$, para luego evaluar en x_{n+1} se obtiene:

$$\begin{aligned} P'(x_{n+1}) &= f(x_{n+1}, y_{n+1}) \\ \frac{11y_{n+1} - 18y_n + 9y_{n-1} - 2y_{n-2}}{6h} &= f(x_{n+1}, y_{n+1}) \end{aligned}$$

De la igualdad anterior es posible obtener la *Fórmula de diferenciación hacia atrás de tres pasos*, más conocida como *BDF3*:

$$11y_{n+1} - 18y_n + 9y_{n-1} - 2y_{n-2} = 6hf(x_{n+1}, y_{n+1}). \tag{11.13}$$

Ejemplo 86. *Se desea resolver nuevamente el ejercicio 85. Se utilizará el esquema iterativo* BDF *de tres pasos, (11.13), con un conjunto de 11 nodos. Entonces* $h = 0{,}3$*;* $y_0 = 0$ *y además:*

$$y_1 = 0{,}044002518$$
$$y_2 = 0{,}16472978.$$

En la tabla 11.10 se muestran los valores de las iteraciones y la solución exacta. En la figura 11.16 se muestran la solución exacta (línea punteada de color azul) y la solución numérica (nodos negros y línea continua de color rojo).

i	x_i	$\tilde{y}_i$	y_i
0	0,0	0,00000000	0,00000000
1	0,3	0,04400251	0,04400251
2	0,6	0,16472978	0,16472978
3	0,9	0,33194243	0,33302318
4	1,2	0,51218850	0,51324774
5	1,5	0,67601246	0,67534753
6	1,8	0,80494471	0,80210130
7	2,1	0,89372595	0,88974947
8	2,4	0,94742752	0,94386523
9	2,7	0,97604358	0,97387859
10	3,0	0,98959687	0,98889100

Tabla 11.10: Resolución del ejemplo 86 a través del método *BDF* de tres pasos.

Ejercicio 47. *Estimar el orden del error de truncamiento de los métodos* BDF*, teniendo en cuenta el error de interpolación.*

11.3. Resolución por Integración Numérica

La ecuación diferencial (11.1) puede, por medio de una aproximación hacia adelante de la derivada primera de $y(x)$, escribirse como:

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(x, y)\mathrm{d}x.$$

Los métodos de Adams surgen al generar un polinomio interpolador para los datos de $f(x, y)$ y luego integrarlo. De acuerdo a los puntos involucrados en la interpolación, se denominan métodos de Adams-Bashforth ó métodos de Adams-Moulton.

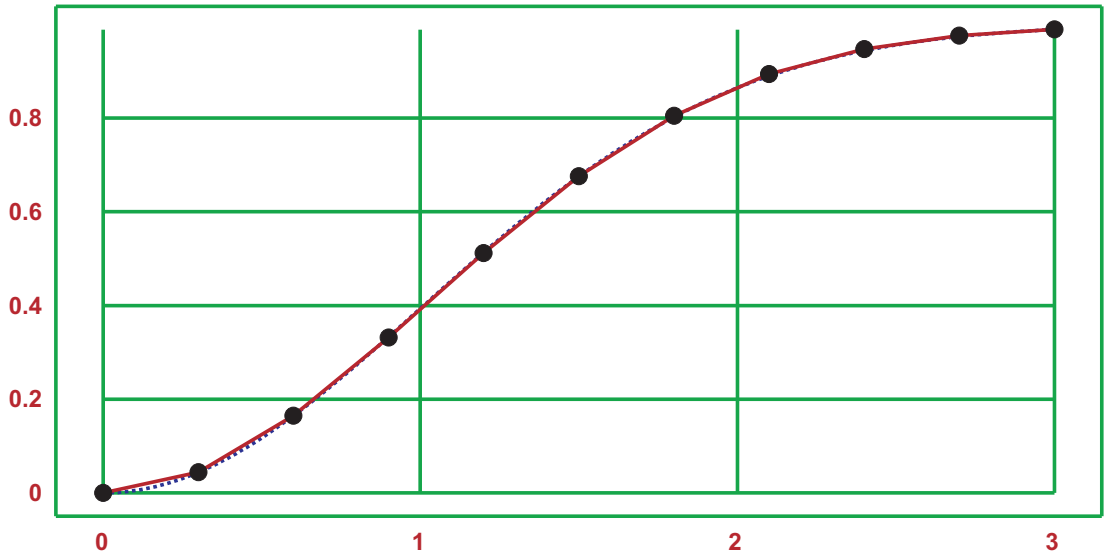

Figura 11.16: Solución numérica (método *BDF* de tres pasos) y exacta de $y' = x(1-y)$.

11.3.1. Métodos de Adams-Bashforth

Los métodos de Adams-Bashforth surgen al considerar como último punto de interpolación a $(x_n, f(x_n, y_n))$. Salvo el método de un paso, también conocido como método de Euler, ninguno de ellos es autoiniciable y al momento de generar los puntos necesarios para la interpolación, se recomienda utilizar algún método de alto orden ó, en el peor de los casos, un método del mismo orden que la aproximación de Adams-Bashforth.

Método de dos pasos

Surge al generar el interpolante que pasa por los puntos $(x_{n-1}, f(x_{n-1}, y_{n-1}))$ y $(x_n, f(x_n, y_n))$, para luego integrar entre x_n y x_{n+1}.

Entonces:

$$\begin{aligned} y_{n+1} &= y_n + \int_{x_n}^{x_{n+1}} f(x,y)\mathrm{d}x \\ &\approx y_n + \int_{x_n}^{x_{n+1}} P(x)\mathrm{d}x, \end{aligned}$$

donde, para el método de dos pasos:

$$\begin{aligned} P(x) &= \frac{(x-x_n)}{x_{n-1}-x_n} f_{n-1} + \frac{(x-x_{n-1})}{x_n - x_{n-1}} f_n \\ &= \frac{(x_n - x)}{h} f_{n-1} + \frac{(x-x_{n-1})}{h} f_n \\ &= \frac{1}{h}\left[(x_{n-1} + h - x) f_{n-1} + (x - x_{n-1}) f_n\right], \end{aligned}$$

por lo tanto, la integral buscada es:

$$\begin{aligned} \int_{x_n}^{x_{n+1}} P(x)\mathrm{d}x &= \frac{1}{h}\frac{h^2}{2}\left[3f_n - f_{n-1}\right] \\ &= \frac{h}{2}\left[3f_n - f_{n-1}\right], \end{aligned}$$

donde $f_n = f(x_n, y_n)$ y $f_{n-1} = f(x_{n-1}, y_{n-1})$. Entonces el **método de dos pasos de Adams-Bashforth** tiene la expresión recursiva:

$$y_{n+1} = y_n + \frac{h}{2}\left[3f(x_n, y_n) - f(x_{n-1}, y_{n-1})\right]. \tag{11.14}$$

Ejemplo 87. *Se desea resolver la ecuación diferencial* $y' + y = \cos(x)$, *con la condición inicial* $y(0) = -1$, *para estimar el valor de* $y(3)$. *Se utilizará el esquema iterativo de Adams-Bashforth de dos pasos, (11.14), con un conjunto de 11 nodos. Entonces* $h = 0{,}3$; $y_0 = -1$ *y además:*

$$y_1 = -0{,}48579898.$$

En la tabla 11.11 se muestran los valores de las iteraciones y la solución exacta. En la figura 11.17 se muestran la solución exacta (línea punteada de color azul) y la solución numérica (nodos negros y línea continua de color rojo).

i	x_i	$\tilde{y}_i$	y_i
0	0,0	-1,00000000	-1,00000000
1	0,3	-0,48579898	-0,48579898
2	0,6	-0,13728802	-0,12822841
3	0,9	0,07972229	0,09261395
4	1,2	0,17917820	0,19540710
5	1,5	0,18032585	0,19942085
6	1,8	0,10353403	0,12537444
7	2,1	-0,02885893	-0,00450301
8	2,4	-0,19344274	-0,16704220
9	2,7	-0,36682260	-0,33915440
10	3,0	-0,52699225	-0,49911685

Tabla 11.11: Resolución del ejemplo 87 a través del método de Adams-Bashforth de dos pasos.

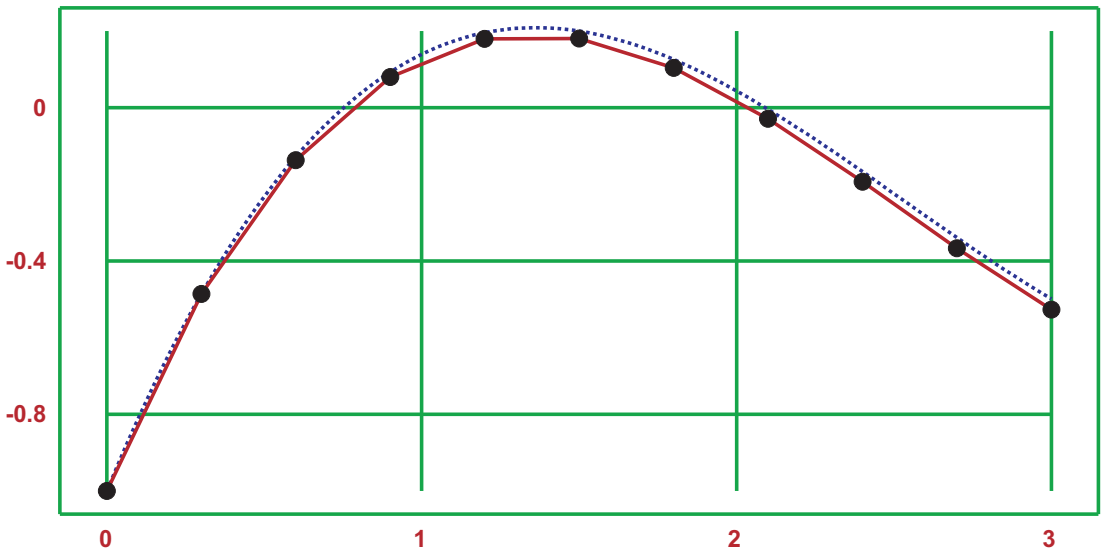

Figura 11.17: Solución numérica (método de Adams-Bashforth de dos pasos) y exacta de $y' + y = \cos(x)$.

Nota. *En el ejemplo 87 y los siguientes, se dan los valores exactos necesarios para comenzar a aplicar el método que se está mostrando. La idea de estos ejemplos es ver cuánto mejora la aproximación con el aumento del método (y por consiguiente del orden) utilizado. Más adelante se analizará la sensibilidad de los métodos de Adams con respecto a los valores iniciales necesarios.*

Método de tres pasos

Surge al generar el interpolante que pasa por los puntos $(x_{n-2}, f(x_{n-2}, y_{n-2}))$; $(x_{n-1}, f(x_{n-1}, y_{n-1}))$ y $(x_n, f(x_n, y_n))$, para luego integrar entre x_n y x_{n+1}.

Entonces:

$$\begin{aligned} y_{n+1} &= y_n + \int_{x_n}^{x_{n+1}} f(x,y)\mathrm{d}x \\ &\approx y_n + \int_{x_n}^{x_{n+1}} P(x)\mathrm{d}x, \end{aligned}$$

donde, para el método de tres pasos:

$$\begin{aligned} P(x) = &\frac{(x-x_{n-1})}{(x_{n-2}-x_{n-1})}\frac{(x-x_n)}{(x_{n-2}-x_n)}f_{n-2} + \frac{(x-x_{n-2})}{(x_{n-1}-x_{n-2})}\frac{(x-x_n)}{(x_{n-1}-x_n)}f_{n-1} + \\ &+ \frac{(x-x_{n-2})}{(x_n-x_{n-2})}\frac{(x-x_{n-1})}{(x_n-x_{n-1})}f_n \\ = &\frac{1}{2h^2}\left[(x-x_{n-1})(x-x_n)f_{n-2} - 2(x-x_{n-2})(x-x_n)f_{n-1} + \right. \\ &\left. + (x-x_{n-2})(x-x_{n-1})f_n\right], \end{aligned}$$

por lo tanto, la integral buscada es:

$$\begin{aligned} \int_{x_n}^{x_{n+1}} P(x)\mathrm{d}x &= \frac{1}{2h^2}\frac{h^3}{6}\left[23f_n - 16f_{n-1} + 5f_{n-2}\right] \\ &= \frac{h}{12}\left[23f_n - 16f_{n-1} + 5f_{n-2}\right] \end{aligned}$$

donde $f_n = f(x_n, y_n)$; $f_{n-1} = f(x_{n-1}, y_{n-1})$ y $f_{n-2} = f(x_{n-2}, y_{n-2})$. Entonces el **método de tres pasos de Adams-Bashforth** tiene la expresión recursiva:

$$y_{n+1} = y_n + \frac{h}{12}\left[23f(x_n, y_n) - 16f(x_{n-1}, y_{n-1}) + 5f(x_{n-2}, y_{n-2})\right]. \tag{11.15}$$

Ejemplo 88. *Se desea resolver nuevamente el ejercicio 87. Ahora se utilizará el esquema iterativo de Adams-Bashforth de tres pasos, (11.15), nuevamente con un conjunto de 11 nodos. Entonces* $h = 0{,}15$*;* $y_0 = -1$ *y además:*

$$\begin{aligned} y_1 &= -0{,}48579898 \\ y_2 &= -0{,}12822841. \end{aligned}$$

En la tabla 11.12 se muestran los valores de las iteraciones y la solución exacta. En la figura 11.18 se muestran la solución exacta (línea punteada de color azul) y la solución numérica (nodos negros y línea continua de color rojo).

Método de cuatro pasos

Surge al generar el interpolante que pasa por los puntos $(x_{n-3}, f(x_{n-3}, y_{n-3}))$; $(x_{n-2}, f(x_{n-2}, y_{n-2}))$; $(x_{n-1}, f(x_{n-1}, y_{n-1}))$ y $(x_n, f(x_n, y_n))$, para luego integrar entre x_n y x_{n+1}.

i	x_i	$\tilde{y}_i$	y_i
0	0,0	-1,00000000	-1,00000000
1	0,3	-0,48579898	-0,48579898
2	0,6	-0,12822841	-0,12822841
3	0,9	0,09361672	0,09261395
4	1,2	0,19592916	0,19540710
5	1,5	0,19962380	0,19942085
6	1,8	0,12494173	0,12537444
7	2,1	-0,00518276	-0,00450301
8	2,4	-0,16774248	-0,16704220
9	2,7	-0,33944458	-0,33915440
10	3,0	-0,49870285	-0,49911685

Tabla 11.12: Resolución del ejemplo 88 a través del método de Adams-Bashforth de tres pasos.

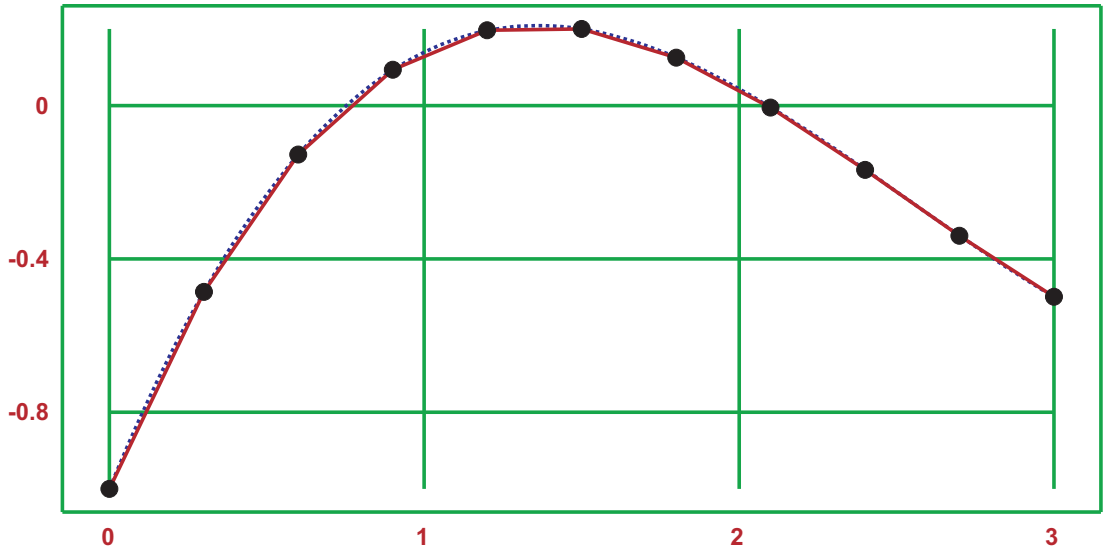

Figura 11.18: Solución numérica (método de Adams-Bashforth de tres pasos) y exacta de $y' + y = \cos(x)$.

Entonces:

$$\begin{aligned} y_{n+1} &= y_n + \int_{x_n}^{x_{n+1}} f(x,y)\mathrm{d}x \\ &\approx y_n + \int_{x_n}^{x_{n+1}} P(x)\mathrm{d}x. \end{aligned}$$

Para el método de cuatro pasos se sigue un esquema similar a los antes desarrollados, generando el polinomio interpolador[5] y luego resolviendo la integral:

$$\begin{aligned} \int_{x_n}^{x_{n+1}} P(x)\mathrm{d}x &= \frac{1}{6h^3}\frac{h^4}{6}\left[55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}\right] \\ &= \frac{h}{24}\left[55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}\right], \end{aligned}$$

donde $f_n = f(x_n, y_n)$; $f_{n-1} = f(x_{n-1}, y_{n-1})$; $f_{n-2} = f(x_{n-2}, y_{n-2})$ y por último $f_{n-3} = f(x_{n-3}, y_{n-3})$. Entonces el **método de cuatro pasos de Adams-Bashforth**

[5] en este caso de grado 3

tiene la expresión recursiva:

$$y_{n+1} = y_n + \frac{h}{24}\left[55f(x_n, y_n) - 59f(x_{n-1}, y_{n-1}) + 37f(x_{n-2}, y_{n-2}) - 9f(x_{n-3}, y_{n-3})\right]. \tag{11.16}$$

Ejemplo 89. *Se desea resolver nuevamente el ejercicio 87. Ahora se utilizará el esquema iterativo de Adams-Bashforth de cuatro pasos, (11.16), nuevamente con un conjunto de 11 nodos. Entonces* $h = 0{,}15$*;* $y_0 = -1$ *y además:*

$$y_1 = -0{,}48579898$$
$$y_2 = -0{,}12822841$$
$$y_3 = 0{,}09261395.$$

En la tabla 11.13 se muestran los valores de las iteraciones y la solución exacta. En la figura 11.19 se muestran la solución exacta (línea punteada de color azul) y la solución numérica (nodos negros y línea continua de color rojo).

i	x_i	$\tilde{y}_i$	y_i
0	0,0	-1,00000000	-1,00000000
1	0,3	-0,48579898	-0,48579898
2	0,6	-0,12822841	-0,12822841
3	0,9	0,09261395	0,09261395
4	1,2	0,19457040	0,19540710
5	1,5	0,19868526	0,19942085
6	1,8	0,12436250	0,12537444
7	2,1	-0,00488686	-0,00450301
8	2,4	-0,16738201	-0,16704220
9	2,7	-0,33874959	-0,33915440
10	3,0	-0,49870534	-0,49911685

Tabla 11.13: Resolución del ejemplo 89 a través del método de Adams-Bashforth de cuatro pasos.

Error de truncamiento

Se hará un breve análisis del error de truncamiento de los métodos de Adams-Bashforth, basándose en la expansión de Taylor. En particular se expandirá la expresión del método de dos pasos, (11.14), omitiendo el resto de las fórmulas puesto que el análisis es similar.

Si, de acuerdo a (11.14):

$$y_{n+1} = y_n + \frac{h}{2}\left[3f_n - f_{n-1}\right],$$

entonces puede expandirse y_{n+1} alrededor de y_n por medio de un polinomio de Taylor:

$$y_{n+1} = y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(x_n) + \frac{h^3}{6}y'''(x_n) + \mathcal{O}(h^4). \tag{11.17}$$

Teniendo en cuenta que $f_n = y'(x_n)$, también puede expandirse f_{n-1} a través de un polinomio de Taylor: $f_{n-1} = y'(x_{n-1}) = y'(x_n - h)$:

$$y'(x_{n-1}) = y'(x_n) - hy''(x_n) + \frac{h^2}{2}y'''(x_n) - \frac{h^3}{6}y^{(iv)}(x_n) + \mathcal{O}(h^4). \tag{11.18}$$

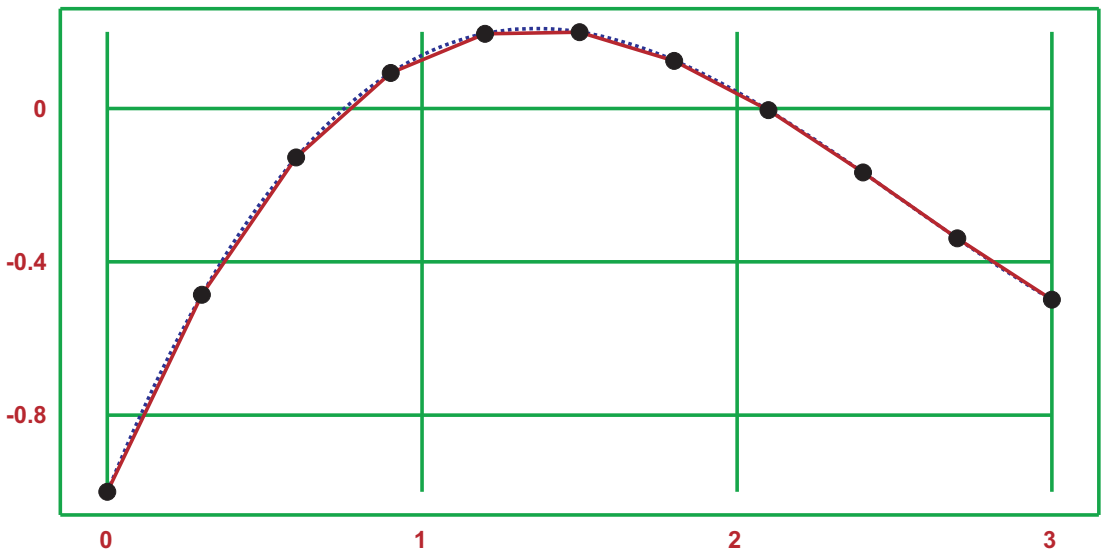

Figura 11.19: Solución numérica (método de Adams-Bashforth de cuatro pasos) y exacta de $y' + y = \cos(x)$.

Reemplazando (11.17) y (11.18) en la expresión de error:

$$\begin{aligned} E(h) &= y_{n+1} - y_n - \frac{h}{2}\left[3f_n - f_{n-1}\right] \\ &= \left(\frac{h^3}{6}y'''(x_n) + \mathcal{O}(h^4)\right) + \left(\frac{h^3}{4}y'''(x_n) + \mathcal{O}(h^4)\right) \\ &= \frac{5}{12}h^3 y'''(x_n) + \mathcal{O}(h^4), \end{aligned} \tag{11.19}$$

el error de truncamiento local es del orden de h^3. Pero, para métodos *cero-estables*[6], el orden de truncamiento global es siempre un orden menor que el orden de truncamiento local. Por lo tanto, el método de Adams-Bashforth de dos pasos tiene un error *global* de truncamiento del orden de h^2.

De manera similar a lo mostrado anteriormente, el error de truncamiento global para el método de Adams-Bashforth de n pasos es $\mathcal{O}(h^n)$.

Ejercicio 48. *Mostrar que el error de truncamiento global para el método de Adams-Bashforth de tres pasos es del orden de h^3.*

Dependencia de los valores iniciales

En los ejemplos 87 a 89, se utilizaron los valores exactos necesarios para dar inicio a las iteraciones de Adams-Bashforth, sólo con fines didácticos. En problemas reales, es imposible saber los valores iniciales por lo que deben aproximarse con algún método de orden similar. Se sugiere utilizar métodos iniciadores cuyo orden sea, como mínimo, similar al que se usará en el esquema iterativo. La calidad de la solución obtenida por iteración depende fuertemente de la aproximación de los valores iniciales.

Ejemplo 90. *Se desea resolver nuevamente la ecuación diferencial del ejemplo 88 con el método de Adams-Bashforth de tres pasos, pero ahora utilizando diferentes valores de h y con los valores de y_2 e y_3 calculados de dos maneras, a fin de comprobar el orden del error de truncamiento cometido. Se espera que el error de truncamiento sea del orden de h^3, pero esto se logra al utilizar los valores exactos (ó un método autoiniciable de*

[6] la explicación de este tema está fuera de los alcances de esta asignatura

orden h^3). Si los dos valores necesarios para iniciar el método de Adams-Bashforth se calculan por medio del método de Euler explícito, cuyo error de truncamiento global es del orden de h, la calidad de la solución se degrada. En la tabla 11.14 se muestran los valores obtenidos utilizando valores iniciales exactos, $\tilde{y}$; utilizando valores aproximados por el método de Euler explícito, $\overline{y}$; y sus correspondientes convergencia de los errores de truncamiento.

Como la cantidad de pasos se duplicó en cada caso, se espera que los cocientes R_i de los métodos cuyo error de truncamiento sea del orden de h^3 converjan alrededor de 8. Sin embargo, la convergencia de los cocientes R_i al utilizar el método de Euler explícito fue alrededor de 4.

pasos	$\tilde{\mathbf{y}}(\mathbf{3})$	R_i	$\overline{\mathbf{y}}(\mathbf{3})$	R_i
10	-0,49870285	-	-0,48574818	-
20	-0,49899868	3,5034	-0,49635820	4,8461
40	-0,49909743	6,0843	-0,49849060	4,4051
80	-0,49911412	7,1173	-0,49896844	4,2198
160	-0,49911649	7,5764	-0,49908079	4,1153
320	-0,49911680	7,7925	-0,49910796	4,0592
640	-0,49911684	7,8973	-0,49911464	4,0300

Tabla 11.14

Ejercicio 49. *Verificar la convergencia de los cocientes R_i cuando, para calcular los valores iniciales necesarios, se utiliza algún método cuyo error de truncamiento sea del orden de h^3.*

11.3.2. Métodos de Adams-Moulton

La idea subyacente detrás de los métodos de Adams-Moulton es la misma que la utilizada en Adams-Bashforth: construir un polinomio interpolante para luego resolver la integral asociada entre x_n y x_{n+1}. La diferencia radica en que, en los métodos de Adams-Moulton el primer nodo de integración está dado por $(x_{n+1}, f(x_{n+1}, y_{n+1}))$. De esta manera, siempre es necesario contar con un método predictor de, al menos, el mismo orden que el método corrector a utilizar.

Método de dos pasos

Surge al generar el polinomio interpolante que pasa por los puntos $(x_n, f(x_n, y_n))$ y $(x_{n+1}, f(x_{n+1}, y_{n+1}))$, para luego integrar entre x_n y x_{n+1}.

Entonces:

$$\begin{aligned} y_{n+1} &= y_n + \int_{x_n}^{x_{n+1}} f(x,y)\mathrm{d}x \\ &\approx y_n + \int_{x_n}^{x_{n+1}} P(x)\mathrm{d}x, \end{aligned}$$

donde, para el método de dos pasos:

$$\begin{aligned} P(x) &= \frac{(x-x_{n+1})}{x_n - x_{n+1}} f_n + \frac{(x-x_n)}{x_{n+1}-x_n} f_{n+1} \\ &= \frac{(x_{n+1}-x)}{h} f_n + \frac{(x-x_n)}{h} f_{n+1} \\ &= \frac{1}{h}\left[(x_n + h - x) f_n + (x - x_n) f_{n+1}\right], \end{aligned}$$

por lo tanto, la integral buscada es:

$$\begin{aligned}\int_{x_n}^{x_{n+1}} P(x)\mathrm{d}x &= \frac{1}{h}\frac{h^2}{2}\left[f_{n+1}+f_n\right]\\ &= \frac{h}{2}\left[f_{n+1}+f_n\right],\end{aligned}$$

donde $f_{n+1}=f(x_{n+1},y_{n+1})$ y $f_n=f(x_n,y_n)$. Entonces el **método de dos pasos de Adams-Moulton** tiene la expresión recursiva:

$$y_{n+1}=y_n+\frac{h}{2}\left[f(x_{n+1},y_{n+1})+f(x_n,y_n)\right], \tag{11.20}$$

y es la misma expresión que la utilizada en el método de Heun.

Método de tres pasos

Surge al generar el polinomio interpolante de grado dos que pasa por los puntos $(x_{n-1},f(x_{n-1},y_{n-1}))$; $(x_n,f(x_n,y_n))$ y $(x_{n+1},f(x_{n+1},y_{n+1}))$, para luego integrar entre x_n y x_{n+1}.

Entonces:

$$\begin{aligned}y_{n+1} &= y_n+\int_{x_n}^{x_{n+1}} f(x,y)\mathrm{d}x\\ &\approx y_n+\int_{x_n}^{x_{n+1}} P(x)\mathrm{d}x,\end{aligned}$$

donde, para el método de tres pasos:

$$\begin{aligned}P(x) &= \frac{(x-x_n)}{(x_{n-1}-x_n)}\frac{(x-x_{n+1})}{(x_{n-1}-x_{n+1})}f_{n-1}+\frac{(x-x_{n-1})}{(x_n-x_{n-1})}\frac{(x-x_{n+1})}{(x_n-x_{n+1})}f_n+\\ &\qquad+\frac{(x-x_{n-1})}{(x_{n+1}-x_{n-1})}\frac{(x-x_n)}{(x_{n+1}-x_n)}f_{n+1}\\ &= \frac{1}{2h^2}\left[(x-x_n)(x-x_{n+1})f_{n-1}-2(x-x_{n-1})(x-x_{n+1})f_n+\right.\\ &\qquad\left.+(x-x_{n-1})(x-x_n)f_{n+1}\right],\end{aligned}$$

por lo tanto, la integral buscada es:

$$\begin{aligned}\int_{x_n}^{x_{n+1}} P(x)\mathrm{d}x &= \frac{1}{2h^2}\frac{h^3}{6}\left[5f_{n+1}+8f_n-f_{n+1}\right]\\ &= \frac{h}{12}\left[5f_{n+1}+8f_n-f_{n-1}\right],\end{aligned}$$

donde $f_{n+1}=f(x_{n+1},y_{n+1})$; $f_n=f(x_n,y_n)$ y $f_{n-1}=f(x_{n-1},y_{n-1})$. Entonces el **método de tres pasos de Adams-Moulton** tiene la expresión recursiva:

$$y_{n+1}=y_n+\frac{h}{12}\left[5f(x_{n+1},y_{n+1})+8f(x_n,y_n)-f(x_{n-1},y_{n-1})\right]. \tag{11.21}$$

Ejemplo 91. *Se desea resolver nuevamente el ejemplo 87, pero ahora utilizando el método de Adams-Bashforth de tres pasos como predictor y Adams-Moulton de tres pasos como corrector. Al igual que en el ejemplo antes mencionado,* $h=0{,}3$*;* $y_0=-1$ *y además:*

$$\begin{aligned}y_1 &= -0{,}48579898\\ y_2 &= -0{,}12822841.\end{aligned}$$

En la tabla 11.15 se muestran los valores de las iteraciones y la solución exacta. En la figura 11.20 se muestran la solución exacta (línea punteada de color azul) y la solución numérica (nodos negros y línea continua de color rojo).

i	x_i	$\tilde{AB3}_i$	$\tilde{AM3}_i$	y_i
0	0,0	-1,00000000	-1,00000000	-1,00000000
1	0,3	-0,48579898	-0,48579898	-0,48579898
2	0,6	-0,12822841	-0,12822841	-0,12822841
3	0,9	0,09361671	0,09245516	0,09261395
4	1,2	0,19543550	0,19531230	0,19540710
5	1,5	0,19889702	0,19947254	0,19942085
6	1,8	0,12477589	0,12555209	0,12537444
7	2,1	-0,00490674	-0,00427277	-0,00450301
8	2,4	-0,16709268	-0,16685621	-0,16704220
9	2,7	-0,33878022	-0,33911087	-0,33915440
10	3,0	-0,49832027	-0,49929871	-0,49911685

Tabla 11.15: Resolución del ejemplo 91 a través del método predictor-corrector de Adams-Bashforth-Moulton de tres pasos.

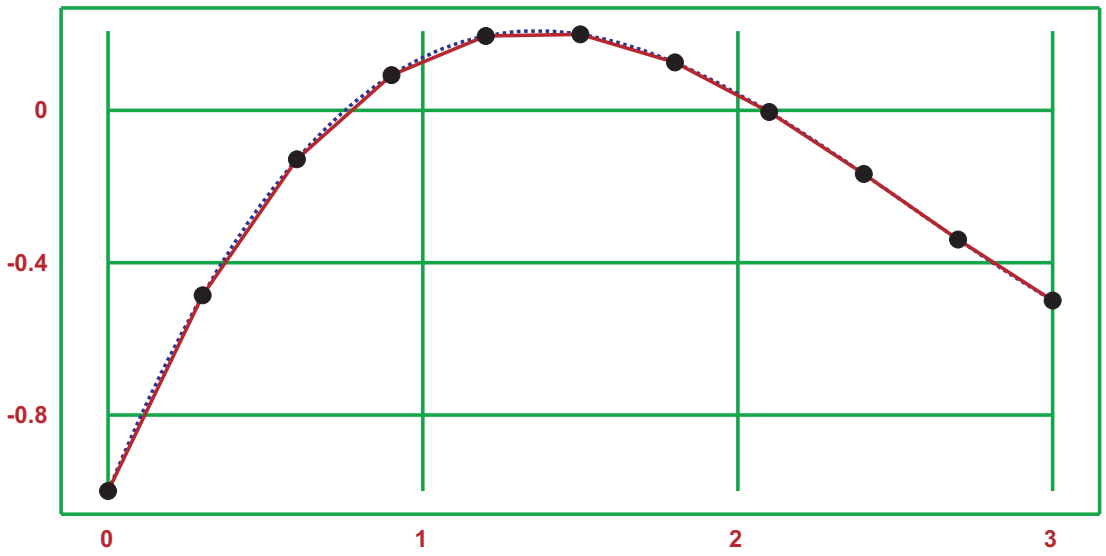

Figura 11.20: Solución numérica (método de Adams-Bashforth-Moulton de tres pasos) y exacta de $y' + y = \cos(x)$.

Error de truncamiento

Se hará un breve análisis del error de truncamiento de los métodos de Adams-Moulton, similar al ya realizado para Adams-Bashforth, basándose en la expansión de Taylor. En particular se expandirá la expresión del método de tres pasos, (11.21), omitiendo el resto de las fórmulas puesto que el análisis es similar.

De acuerdo a la fórmula (11.21):

$$y_{n+1} = y_n + \frac{h}{12}\left[5f_{n+1} + 8f_n - f_{n-1}\right],$$

entonces es posible expandir, por medio de la fórmula de Taylor las expresiones de:

$$\begin{aligned}
y_{n+1} &= y_n + hy'_n + \frac{h^2}{2}y''_n + \frac{h^3}{6}y'''_n + \frac{h^4}{24}y_n^{(iv)} + \mathcal{O}(h^5)\\
f_{n+1} &= y'_n + hy''_n + \frac{h^2}{2}y'''_n + \frac{h^3}{6}y_n^{(iv)} + \frac{h^4}{24}y_n^{(v)} + \mathcal{O}(h^5)\\
f_{n-1} &= y'_n - hy''_n + \frac{h^2}{2}y'''_n - \frac{h^3}{6}y_n^{(iv)} + \frac{h^4}{24}y_n^{(v)} + \mathcal{O}(h^5).
\end{aligned}$$

Reemplazando las tres expresiones anteriores en la expresión de error y simplificando se obtiene la expresión del error local:

$$\begin{aligned} E(h) &= y_{n+1} - y_n - \frac{h}{12}\left[5f_{n+1} + 8f_n - f_{n-1}\right] \\ &= -\frac{h^4}{24}y_n^{(iv)} + \mathcal{O}(h^5). \end{aligned}$$

De la misma manera que en los métodos de Adams-Bashforth, en los métodos de Adams-Moulton el error global decrece un orden de magnitud con respecto al error local, pues se trata de un método cero-estable. Por lo tanto, el error de truncamiento global para el método de Adams-Moulton de n pasos es $\mathcal{O}(h^n)$.

Ejercicio 50. *Calcular la expresión del error para método de Adams-Moulton de dos pasos, también conocido como método de Heun.*

11.4. Ecuaciones Diferenciales Rígidas

Las ecuaciones diferenciales rígidas, ó como son más conocidas en la literatura: ecuaciones diferenciales *stiff* presentan algunos problemas al momento de ser resueltas con los métodos numéricos tradicionales. Una EDO *stiff* se caracteriza por una etapa en la cual existe un crecimiento ó decrecimiento de gran pendiente, para luego estabilizarse alrededor de alguna función llamada envolvente. Si bien al resolverse analíticamente, el término de gran crecimiento (decrecimiento) se vuelve despreciable, es fundamental considerarlo durante la primera parte del intervalo de resolución al aplicar un método numérico. El orden de magnitud del error global de truncamiento en un paso determinado depende de la precisión del o los pasos anteriores. Es por esto que no todos los métodos vistos pueden resolver correctamente, dentro de un margen de error estimado, los problemas *stiff*.

Las ecuaciones diferenciales *stiff* surgen como solución de una gran cantidad de aplicaciones ingenieriles, incluyendo modelado de reacciones químicas, solución numérica de ecuaciones diferenciales en derivadas parciales parabólicas e hiperbólicas, teoría de control y modelado de circuitos eléctricos.

Ejemplo 92. *Se desea resolver la ecuación diferencial ordinaria $y' = -20y$ con la condición inicial $y(0) = 1$ a fin de estimar el valor de $y(1)$. Para ello se utilizará una partición de 11 nodos equiespaciados y los métodos: Euler explícito y Euler implícito (*BDF1*), cuyos errores globales son del orden h; Crank-Nicolson, Heun (AM2) y Taylor de orden 2, cuyos errores globales son del orden h^2. En todos los casos, se utilizará $h = 0{,}1$. Como la ecuación diferencial a resolver es lineal, entonces los esquemas implícitos planteados son autoiniciables de acuerdo a las fórmulas iterativas dadas a continuación:*

- *Euler explícito:* $y_{n+1} = y_n(1 - 20h)$.
- *Euler implícito:* $y_{n+1} = \dfrac{y_n}{1 + 20h}$.
- *Crank-Nicolson:* $y_{n+1} = y_n \dfrac{1 - 10h}{1 + 10h}$.
- *Heun:* $y_{n+1} = y_n \dfrac{1 - 10h}{1 + 10h}$.
- *Taylor de orden 2:* $y_{n+1} = y_n(1 + 400h - 4000h^2)$

Los resultados se muestran en la tabla 11.16. El único método que converge, desde el punto de vista de la estabilidad de las iteraciones, es el método de Euler implícito. En las figuras 11.21 y 11.22 se muestran las iteraciones del método de Euler Implícito.

i	xi	Euler Exp.	Euler Imp.	CN/Heun	Taylor	Exacto
0	0,0	1,0000E+00	1,00000E+00	1,0000E+00	1,0000E+00	1,00000E+00
1	0,1	-1,0000E+00	3,33333E-01	0,0000E+00	1,0000E+00	1,35335E-01
2	0,2	1,0000E+00	1,11111E-01	0,0000E+00	1,0000E+00	1,83156E-02
3	0,3	-1,0000E+00	3,70370E-02	0,0000E+00	1,0000E+00	2,47875E-03
4	0,4	1,0000E+00	1,23457E-02	0,0000E+00	1,0000E+00	3,35463E-04
5	0,5	-1,0000E+00	4,11523E-03	0,0000E+00	1,0000E+00	4,53999E-05
6	0,6	1,0000E+00	1,37174E-03	0,0000E+00	1,0000E+00	6,14421E-06
7	0,7	-1,0000E+00	4,57247E-04	0,0000E+00	1,0000E+00	8,31529E-07
8	0,8	1,0000E+00	1,52416E-04	0,0000E+00	1,0000E+00	1,12535E-07
9	0,9	-1,0000E+00	5,08053E-05	0,0000E+00	1,0000E+00	1,52300E-08
10	1,0	1,0000E+00	1,69351E-05	0,0000E+00	1,0000E+00	2,06115E-09

Tabla 11.16

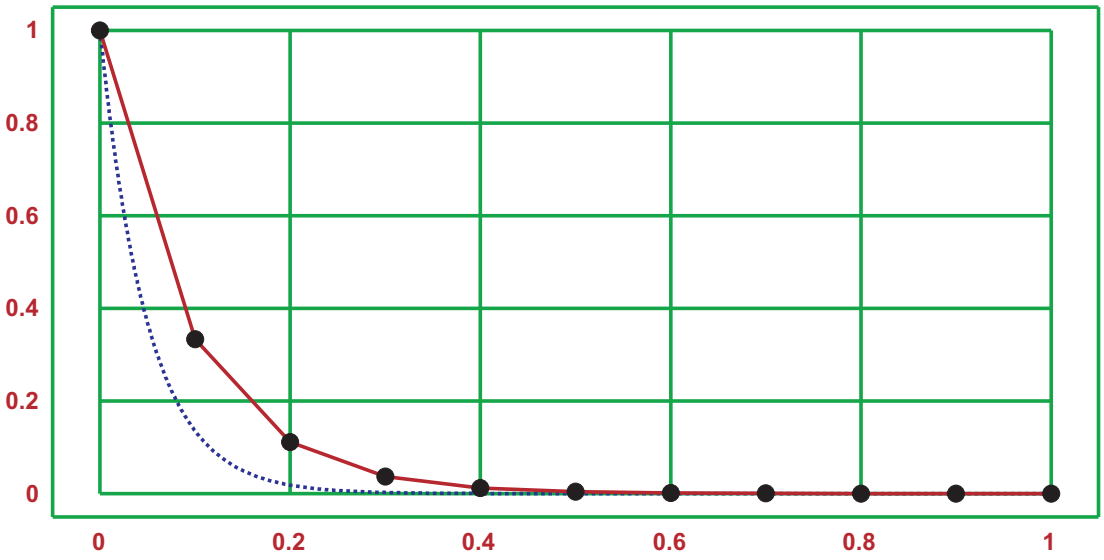

Figura 11.21: Solución numérica (método de Euler Implícito) y exacta de $y' = -20y$.

En el ejemplo anterior todos los métodos empleados, salvo Euler Implícito, fallaron al intentar resolver una ecuación diferencial sencilla con pocas iteraciones. Sin embargo, si se utiliza un incremento h lo suficientemente pequeño, es posible resolver exitosamente la ecuación diferencial propuesta con cualquiera de los métodos utilizados. La familia de métodos *BDF* es recomendada para resolver ecuaciones *stiff*.

No siempre los métodos que no convergen a la solución se estancan alrededor de un valor. Generalmente tienden a crecer en forma indefinida ó bien oscilar alrededor de la solución, como se muestra en el ejemplo siguiente.

Ejemplo 93. *Se desea resolver la ecuación diferencial ordinaria* $y' = -50\,(y - \cos(x)) - \sin(x)$ *con la condición inicial* $y(0) = 0$ *a fin de estimar el valor de* $y(2)$. *Para ello se utilizará una partición de 11 nodos equiespaciados y los*

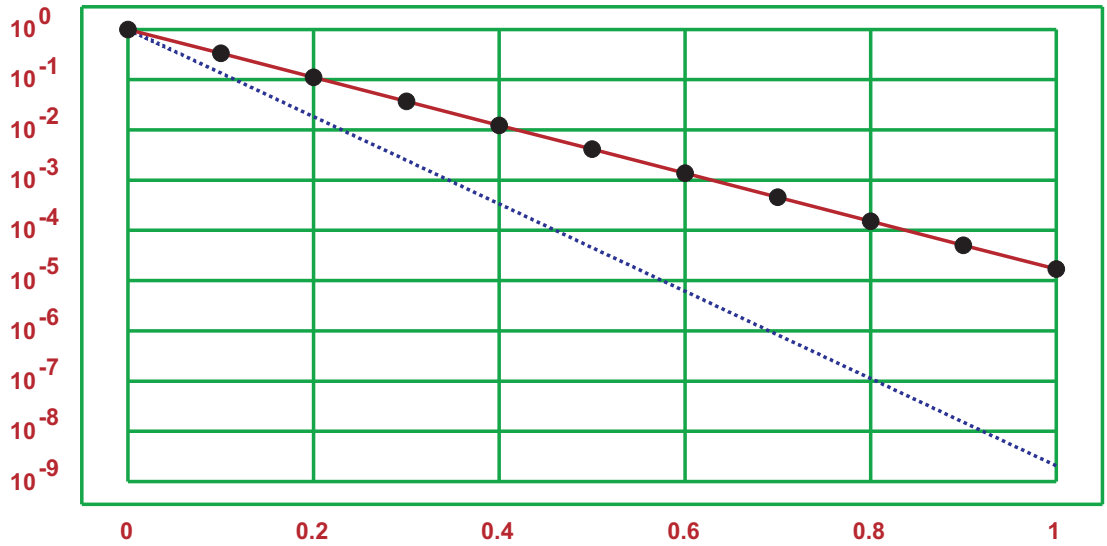

Figura 11.22: Solución numérica (método de Euler Implícito) y exacta de $y' = -20y$, eje vertical en escala logarítmica.

*métodos: Euler explícito y Euler implícito (*BDF1*), cuyos errores globales son del orden h; Crank-Nicolson, Heun (AM2) y Taylor de orden 2, cuyos errores globales son del orden h^2. En todos los casos, se utilizará $h = 0{,}2$. Como la ecuación diferencial a resolver es lineal, entonces los esquemas implícitos planteados son autoiniciables de acuerdo a las fórmulas iterativas dadas a continuación:*

- *Euler explícito:* $y_{n+1} = y_n - h\,[50(y_n - \cos(x_n)) + \sin(x_n)]$.
- *Euler implícito:* $y_{n+1} = \dfrac{y_n + h\,[50\cos(x_{n+1}) - \sin(x_{n+1})]}{1 + 50h}$.
- *Crank-Nicolson:* $y_{n+1} = \dfrac{y_n + h\left[-25y_n + 50\cos\left(\frac{x_n+x_{n+1}}{2}\right) - \sin\left(\frac{x_n+x_{n+1}}{2}\right)\right]}{1 + 25h}$.
- *Heun:* $y_{n+1} = \dfrac{y_n + \frac{h}{2}\,[50(\cos(x_{n+1}) - y_n + \cos(x_n)) - \sin(x_{n+1}) - \sin(x_n)]}{1 + 25h}$.
- *Taylor de orden 2:* $y_{n+1} = y_n + hf_n - \frac{h^2}{2}\,[50\sin(x_n) + \cos(x_n) + 50f_n]$, *con* $f_n = -50\,(y_n - \cos(x_n)) - \sin(x_n)$.

Los resultados obtenidos con la aplicación de los métodos mencionados, a excepción de Euler explícito (divergente y oscilante) y Taylor de orden 2 (divergente hacia menos infinito), se muestran en la tabla 11.17. Nuevamente la convergencia monótona la logra el método de Euler implícito, pero Crank-Nicolson y Heun convergen en forma oscilante con una pequeña diferencia entre ellos. Las figuras 11.23 y 11.24 muestran la convergencia de los métodos Euler implícito y Heun respectivamente.

Ejercicio 51. *Repetir el ejemplo anterior, pero ahora utilizar el método BDF3 calculando los valores necesarios para iniciar las iteraciones a través del método de Euler Implícito, con 4 pasos por nodo.*

Comandos de EMT. *El comando para resolver una EDO* stiff *es:*

- `ode(f$:string, t:vector, y0:número)`, *donde* ***f$*** *es la función de x e y que representa y', expresada como* string*;* ***t*** *es el dominio discreto sobre el que se resolverá la EDO (nodos);* ***y0*** *es el valor inicial de la EDO. La salida es un vector con las aproximaciones en cada nodo.*

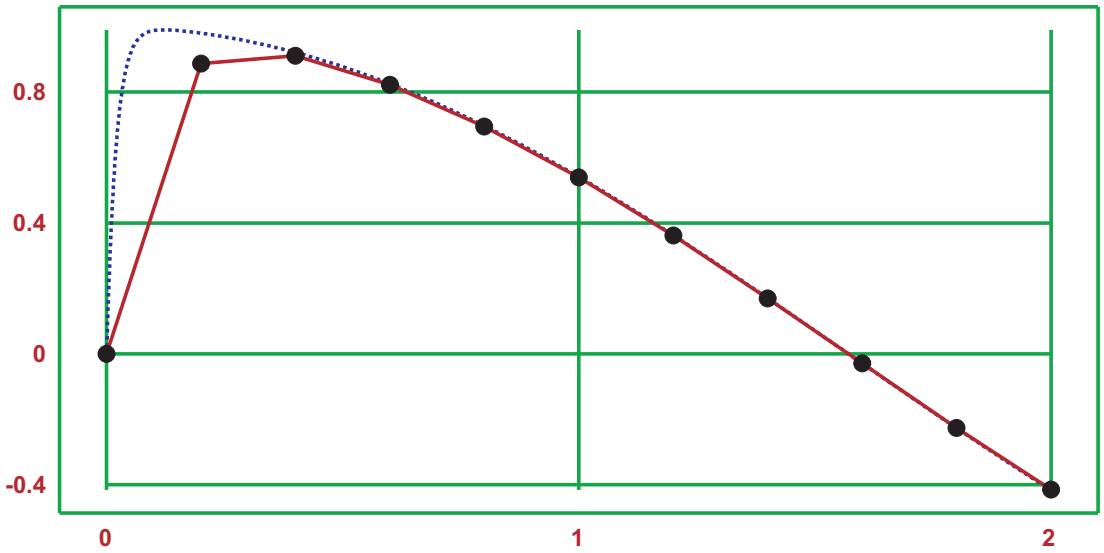

Figura 11.23: Solución numérica (método de Euler Implícito) y exacta de $y' = -50\,(y - \cos(x)) - \sin(x)$.

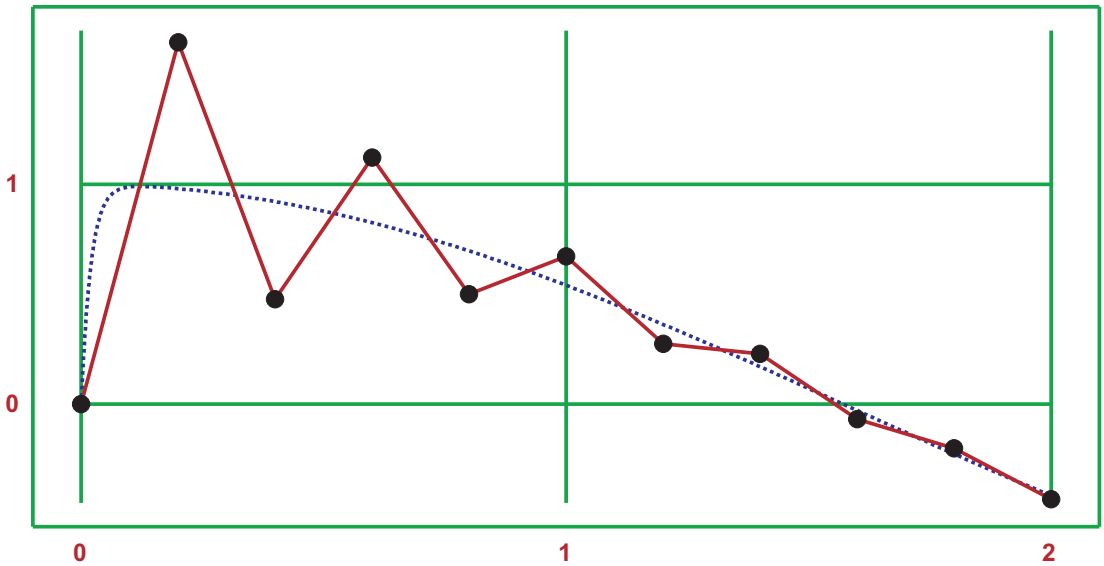

Figura 11.24: Solución numérica (método de Heun) y exacta de $y' = -50\,(y - \cos(x)) - \sin(x)$.

i	xi	Euler Imp.	CN	Heun	Exacto
0	0,0	0,00000E+00	0,00000E+00	0,00000E+00	0,00000E+00
1	0,2	8,87357E-01	1,65501E+00	1,64674E+00	9,80021E-01
2	0,4	9,10917E-01	4,79035E-01	4,76642E-01	9,21061E-01
3	0,6	8,22849E-01	1,12730E+00	1,12167E+00	8,25336E-01
4	0,8	6,95131E-01	5,01730E-01	4,99223E-01	6,96707E-01
5	1,0	5,39078E-01	6,75419E-01	6,72045E-01	5,40302E-01
6	1,2	3,61477E-01	2,76007E-01	2,74628E-01	3,62358E-01
7	1,4	1,69460E-01	2,29708E-01	2,28560E-01	1,69967E-01
8	1,6	-2,93136E-02	-6,84932E-02	-6,81510E-02	-2,91995E-02
9	1,8	-2,26919E-01	-2,02134E-01	-2,01124E-01	-2,27202E-01
10	2,0	-4,15477E-01	-4,35603E-01	-4,33427E-01	-4,16147E-01

Tabla 11.17

Ejemplo en EMT 24. *Resolver, a través del método para ecuaciones* stiff, $y' = 30-5y$ *con la condición inicial* $y(0) = 1$ *para estimar el valor de* $y(5)$. *Utilizar 10 pasos.*

```
>longformat
>ode("30-5*y",linspace(0,5,10),1)
 [1,  5.58957500688,  5.96631026504,  5.99723457817,  5.99977300034,
 5.99998136673,  5.99999847049,  5.99999987445,  5.99999998969,
 5.99999999915,  5.99999999993]
```

11.5. Ejercicios

1. Construir los siguientes algoritmos en PC:

   a) **Método de Crank-Nicolson**. Entrada: $f(x, y)$; $(x_0; y_0)$; h; x_n. Salida: el valor numérico de $y(x_n)$. Opcional: graficar los puntos intermedios hasta llegar a la solución.

   b) **Método de Euler - RK1**. Idénticas condiciones que las definidas para el método de Crank-Nicolson.

   c) **Método de Runge-Kutta - RK4**. Idénticas condiciones que las definidas para el método de Crank-Nicolson.

   d) **Método predictor-corrector de Adams-Bashforth-Moulton de tres pasos**. Idénticas condiciones que las definidas para el método de Crank-Nicolson.

2. Para las ecuaciones diferenciales dadas a continuación, graficar sus isóclinas y deducir a qué tiende la solución cuando $x \to \infty$.

   a) $y' = \sin(y + x^2)$, en $[-4, 5]$.

   b) $y' = y^2 - x$, en $[-2, 10]$.

3. [*EMT*] En la familia de métodos predictor-corrector se sigue un esquema básico de 1 predicción y 1 corrección por cada paso de iteración realizada. Iterar sobre la corrección, ¿logra la convergencia a la solución exacta? ¿O sólo es una aproximación un poco mejor que la obtenida con una única corrección realizada por paso? Verificar la respuesta dada con la ecuación diferencial $y' + \cos(x) = y$, $y(0) = -1$, y el método predictor-corrector de Euler para aproximar $y(1)$ con $h = 0{,}2$.

4. Resolver los siguientes problemas mediante el método de Euler explícito con los valores de h indicados. Calcular las estimaciones del error y aplicar, si es posible, extrapolación de Richardson:

   a) $y' = 1 - y$; $y(0) = 0$; $y(1) =?$. $\quad h = 0{,}5$; $h = 0{,}25$.
   b) $y' = y^2 + 2x - x^4$; $y(0) = 0$; $y(0{,}2) =?$. $\quad h = 0{,}2$; $h = 0{,}1$.
   c) $y' = y + e^x + xy$; $y(1) = 2$; $y(1{,}03) =?$. $\quad h = 0{,}01$; $h = 0{,}005$.

5. $[EMT]$ Dada la ecuación diferencial $y' = x\cos(x - y)$, con la condición inicial $y(0) = 0$, resolverla para calcular $y(1)$ utilizando el método de Euler implícito, pero en vez de predecir el valor por medio de otro método resolver la ecuación no lineal planteada en cada paso. Operar con aritmética de 5 dígitos, truncamiento y un mínimo de 5 pasos.

6. El problema de valor inicial $y' = y^{1/3}$, $y(0) = 0$ tiene dos soluciones: $y_1 = 0$ e $y_2 = \left(\frac{2}{3}x\right)^{3/2}$, para $x \geq 0$. Si se aplica sobre ambos el método de Series de Taylor, ¿qué sucede?

7. $[EMT]$ Considerar la ecuación diferencial $y' = y$. Si la condición inicial es $y(0) = c$, entonces la solución es $y = ce^x$. Si un error de redondeo ε ocurre en la PC al ingresar el valor inicial c, ¿qué efecto tiene sobre la solución cuando $x = 10$? ¿Y cuando $x = 20$?

8. Repetir el ejercicio anterior, con las mismas condiciones planteadas, pero con la ecuación diferencial $y' = -y$.

9. Suponer que una ecuación diferencial se resuelve numéricamente en el intervalo $[a, b]$ y el error local de truncamiento es ch^p. Mostrar que, si todos los errores de truncamiento tienen el mismo signo (el peor caso posible), entonces el error global de truncamiento es $(b - a)ch^{p-1}$, donde $h = \frac{b-a}{n}$.

10. Considerar la ecuación diferencial ordinaria $y' = (xy)^3 - \left(\frac{x}{y}\right)^2$, con la condición inicial $y(1) = 1$. Realizar un paso con $h = 0{,}1$ y:

    a) El método de Series de Taylor de orden 2.
    b) El método RK2.

    Comparar los resultados obtenidos.

11. $[EMT]$ El método RK3 responde a la ecuación iterativa:

$$y_{n+1} = y_n + \frac{1}{9}(2K_1 + 3K_2 + 4K_3),$$

    donde:

$$K_1 = hf(x_n, y_n)$$
$$K_2 = hf\left(x_n + \frac{h}{2}, y_n + \frac{K_1}{2}\right)$$
$$K_3 = hf\left(x_n + \frac{3h}{4}, y_n + \frac{3K_2}{4}\right)$$

    ¿Cuál es el orden del error de truncamiento?

12. $[EMT]$ Resolver la ecuación diferencial $y' = 10y + 11x - 5x^2 - 1$ con la condición inicial $y(0) = 0$, el método RK4 y $h = 2^{-8}$. Calcular la solución exacta y graficar ambas soluciones en el intervalo $[0, 3]$. Repetir el ejercicio pero ahora con la condición inicial a $y(0) = \varepsilon$, con ε pequeño. ¿Qué ocurre?

13. [EMT] Considerar la ecuación diferencial:

$$y' = \begin{cases} y + x, & x \in [-1, 0] \\ y - x, & x \in [0, 1] \end{cases},$$

con la condición inicial $y(-1) = 1$. Resolverla en el intervalo $[-1, 1]$ y RK4 utilizando $h = 0{,}1$. Resolver nuevamente pero ahora utilizar $h = 0{,}09$. ¿Qué solución es más precisa? ¿Por qué?

14. Resolver las siguientes ecuaciones diferenciales para aproximar $y(3)$, con el mismo incremento h para todos los casos, la condición inicial $y(2) = 1$ y los métodos: Heun, AB2 y AM2:

    a) $y' = \sin(x) - e^y$
    b) $y' = 2\cos(x) - 3y + x$
    c) $y' = \sin(y)(x^2 + \cos(x))$

15. Mostrar que el método RK3, descrito en el inciso 11, plantea el mismo esquema iterativo que el método de Series de Taylor con la ecuación diferencial $y' = x + y$. En general, esto se cumple para cualquier ecuación diferencial, con la expansión hasta h^3.

16. [EMT] Determinar con qué valores iniciales la ecuación diferencial $y' = \dfrac{y}{1 + x^2}$ diverge cuando $x \to \infty$.

17. Determinar el valor numérico de:

$$\int_2^3 s\cos(s)\mathrm{d}s,$$

de tres maneras: integrando numéricamente, resolviendo una ecuación diferencial ordinaria y calculando la integral en forma exacta. Comparar los resultados.

18. Dada la ecuación diferencial $y' = \sin(y)$ con el valor inicial $y(0) = 1$, utilizar el método RK4 para obtener una aproximación de $y(0{,}5)$ con 2 y 4 pasos. Mejorar la estimación utilizando el método de extrapolación de Richardson.

19. Mostrar que el método de Heun falla al tratar de aproximar $y(2)$ para la ecuación diferencial $y' = 3y^{1/3}$ con la condición inicial $y(0) = 0$. ¿Es posible resolver con algún método diferente este problema?

20. ¿Es el método de Heun estable al tratar de calcular $y(1)$ para la ecuación diferencial $y' = 2xy^2$ con la condición inicial $y(0) = 1$? ¿Por qué?

21. Estimar el tamaño de paso necesario para lograr la convergencia de la ecuación diferencial $y' = -\lambda y$, con $\lambda > 0$ y la condición inicial $y(0) = 1$ en el intervalo $[0, 2]$.

22. [EMT] Es posible resolver una ecuación diferencial de orden superior generando un sistema de ecuaciones no lineales y utilizando los esquemas de resolución para ecuaciones de orden 1. Aproximar numéricamente $y(1)$ para la ecuación diferencial $y'' - 2y' - 3y = \cos(x)$ con las condiciones iniciales $y(0) = 1$; $y'(0) = -1$. Utilizar el método de Euler explícito y, al menos, 20 pasos.

23. [EMT] Repetir el ejercicio anterior, con idénticas condiciones de trabajo, pero ahora utilizando el método Crank-Nicolson. Para la predicción de valores, utilizar el método de Euler explícito. ¿Es posible aplicar Crank-Nicolson como un método autoiniciable?

24. [EMT] Dado el sistema de ecuaciones diferenciales:

$$\frac{dx}{dt} = x - xy$$
$$\frac{dy}{dt} = -y + xy,$$

con las condiciones iniciales $x(0) = 4$; $y(0) = 1$, resolverlo con el método predictor-corrector de Euler explícito y Crank-Nicolson, utilizando $t \in [0, 20]$ y:

a) 100 pasos.

b) 1000 pasos.

Graficar x vs y en los 2 casos planteados.

25. [EMT] Repetir el ejercicio anterior, pero ahora utilizar:

a) $x(0) = 2$; $y(0) = 1$

b) $x(0) = 4{,}5$; $y(0) = 1{,}5$,

c) $x(0) = 7$; $y(0) = 2$,

con 1000 pasos. La solución gráfica de este tipo de sistemas se asocia con los denominados *ciclos límite*.

Bibliografía

- *An introduction to numerical analysis*, Kendall ATKINSON, Cap.6
- *Análisis numérico*, R. BURDEN y J. FAIRES, Cap.5
- *Análisis numérico - Primer curso*, Hernán GONZÁLEZ, Cap.6
- *Análisis numérico - Un enfoque práctico*, M. MARON y R. LÓPEZ, Cap.8
- *Análisis numérico con aplicaciones*, C. GERALD y P. WHEATLEY, Cap.6
- *Fundamental numerical methods for electrical engineering**, Stanislaw ROSLONIEC, Cap.7
- *Numerical methods*, G. DAHLQUIST y A. BJÖRK, Cap.8
- *Numerical calculations and algorithms*, R. BECKETT y J. HURT, Cap.6

Solución de los ejercicios de número impar

Capítulo 1 - Conceptos Básicos del Cálculo Numérico

Ejercicio 3:

Para el caso planteado, $p = \frac{a+b+c}{2}$ y $a \approx b+c$ con lo que $p \approx a$. Entonces $p-a \approx 0$ lo que implica una cancelación catastrófica. Por lo tanto, el área dará cero.

Ejercicio 5:

Salvo en el primer ejemplo, los otros tres no dan exactamente cero. Los resultados son (en orden de ejecución): $-\epsilon_M$; $\epsilon_M/4$ y $\epsilon_M/8$.

Ejercicio 7:

a) Antes de anidar 9 *flops* y luego de anidar 6 *flops*.

b) Antes de anidar 8 *flops* y luego de anidar 7 *flops*.

c) Antes de anidar 9 *flops* y luego de anidar 7 *flops*.

Ejercicio 9:

$37191_{10} = 4441036_6 = 110507_8 = 25\text{A}40_{11} = \text{B}046_{15}$.

Ejercicio 11:

a) $e_A = 3{,}20 \times 10^{-2}$

b) $e_A = 1{,}51 \times 10^{-2}$

c) $e_A = 3{,}40 \times 10^{-3}$

d) $e_A = 5{,}00 \times 10^{-4}$

Ejercicio 13:

El código correspondiente se encuentra en la sección de Códigos para *Euler Math Toolbox*.

Ejercicio 15:

La segunda expresión es estable, la primera no.

Ejercicio 17:

La primera expresión es más sensible al error que la segunda, aunque ambas se pueden considerar bien condicionadas alrededor de $x_0 = 2$.

Ejercicio 19:

Si bien en todo el intervalo tiene buena precisión, la peor se consigue en los alrededores de $\pi/4$.

Ejercicio 21:

Con la fórmula original, $f(0{,}0001) = 0$ puesto que el numerador de la fracción de anula. Pero utilizando:

$$\begin{aligned} f(x) &= \frac{e^x - 1}{x} \\ &\approx \frac{1 + x + \frac{x^2}{2} - 1}{x} \\ &\approx \frac{x + \frac{x^2}{2}}{x} \end{aligned}$$

se obtiene un valor que coincide con el límite de la función: $f(0{,}0001) = 1$.

Ejercicio 23:

1. El desbordamiento de memoria ocurre por *overflow*.
2. En este caso no, ya que el *overflow* ocurre al calcular 300^{125}.
3.
```
function Ej23TP1(L,k)
f=1;
for i=1 to k
    f=f*L/(i*exp(1));
end
return {f}
endfunction
```
4. El resultado obtenido al ejecutar `Ej23TP1(300,125)` es $1{,}19831 \times 10^{46}$.

Ejercicio 25:

Sea $f(r) = 0$. Asumiendo que r es una raíz simple, se sabe que $f'(r) \neq 0$. Entonces es posible generar una función $F(x) = f(x) + \varepsilon g(x)$ y aproximarla con una serie de Taylor en cercanías de r, donde estará la nueva raíz:

$$\begin{aligned} F(r+h) &= 0 \\ &= f(r+h) + \varepsilon g(r+h) \\ &\approx \left[f(r) + hf'(r) + \frac{1}{2}h^2 f''(\xi) \right] + \varepsilon \left[g(r) + hg'(r) + \frac{1}{2}h^2 g''(\eta) \right], \end{aligned}$$

tomando como negligentes los términos en h^2 y aplicando que $f(r) = 0$:

$$h \approx -\varepsilon \frac{g(r)}{f'(r) + \varepsilon g'(r)} \approx -\varepsilon \frac{g(r)}{f'(r).}$$

Para el ejercicio planteado, $h \approx -\varepsilon \dfrac{g(20)}{f'(20)} = -\varepsilon \dfrac{20^{20}}{19!}$, lo que es un incremento muy grande.

Capítulo 2 - Resolución de Ecuaciones No Lineales

Ejercicio 3:

La raíz por Bisección es $x_B = -1{,}99805$, la solución exacta es $x = -2$.

Ejercicio 5:

La raíz por Newton Raphson es $x_{NR} = 4{,}7300$, en 3 iteraciones. La raíz por Secante es $x_S = 4{,}7300$, en 4 iteraciones.

Ejercicio 7:

Con Newton-Raphson se obtiene una convergencia superior a lineal. No se puede determinar con precisión debido a que son pocas las iteraciones. Con Bisección la convergencia es lineal, de tasa 0,5.

Ejercicio 9:

a) `>plot2d("8x-cos(x)-2x^2",xmin=-1,xmax=5);`

b) Por Regula Falsi, la primera raíz es $x_{RF} = 0{,}128077$, en 4 iteraciones; la segunda raíz es $x_{RF} = 4{,}07322$, en 6 iteraciones. Por Bisección, la primera raíz es $x_B = 0{,}128074$, en 16 iteraciones; la segunda raíz es $x_B = 4{,}07321$, en 16 iteraciones. El intervalo utilizado para la primera raíz fue $[0; 1]$ y para la segunda $[4; 5]$.

c) Ambos esquemas son útiles para el cálculo de la primera raíz:

- $\phi_1(x) = \dfrac{2x^2 + \cos(x)}{8}$
- $\phi_2(x) = \dfrac{\cos(x)}{8 - 2x}$

d) En los cuatro casos analizados, la convergencia es lineal y se calculó sobre la primera raíz:

- Bisección: $CL = 0{,}5$
- Regula Falsi: $CL = 0{,}02463$
- Punto Fijo - $\phi_1(x)$: $CL = 0{,}04804$, $x_0 = 0$
- Punto Fijo - $\phi_2(x)$: $CL = 0{,}01658$, $x_0 = 0$

Ejercicio 11:

Los intervalos quedan determinados por la aritmética utilizada, porque se observa un comportamiento fractal.

Ejercicio 13:

Una de las raíces complejas de la función es $-1{,}00286e - 011 + 1i$, obtenida en 6 iteraciones.

Ejercicio 15:

La raíz obtenida por Bisección y $\varepsilon = 1 \times 10^{-3}$ es $x_B = 6{,}00010$. Luego de cambiar el coeficiente no se encuentra raíz en el intervalo indicado.

Ejercicio 17:

La función a utilizar es $f(x) = x^3 - R$, es convergente para cualquier número positivo y con cualquier semilla.

Ejercicio 19:

a) La raíz por Bisección es $x_B = 0{,}999984741211$, convergiendo en 30 iteraciones y con la condición de corte: $\varepsilon = 1 \times 10^{-4}$.

b) La raíz por Regula Falsi, en 500 iteraciones y $\varepsilon = 1 \times 10^{-8}$ es $x_{RF} = 1{,}54928409254$.

c) La raíz por Secante, en 10 iteraciones y $\varepsilon = 1 \times 10^{-3}$ es $x_S = 0{,}245121540175$.

d) La raíz por Secante, en 15 iteraciones y $\varepsilon = 1 \times 10^{-5}$ es $x_S = -0{,}04523352526$.

e) Falla el algoritmo, pero la salida obtenida por Regula Falsi, en 3 iteraciones y $\varepsilon = 1 \times 10^{-3}$ es $x_{RF} = 3{,}99998776364$. No es raíz, falló el algoritmo.

Ejercicio 21:

No se puede establecer claramente el orden de convergencia. Se puede asegurar que es superior a lineal.

Ejercicio 23:

La función dada es asintótica a x, por lo tanto nunca se verifica que $x = f(x)$. Ésto no contradice al Teorema de Punto Fijo puesto que no se cumple la condición de función autocontenida.

Ejercicio 25:

a) Los puntos fijos de la ecuación logística son $x = 0$ y $x = 1 - \frac{1}{a}$.

b) Como $\phi'(x) = a(1 - 2x)$: $x = 0$ es atractor si $a \in (0; 1)$ y $x = 1 - \frac{1}{a}$ es atractor si $a \in (1; 3)$.

c) Las iteraciones converge cuadráticamente cuando la derivada del mapeo se anula, esto es $a = 0$ para $x = 0$ y $a = 2$ para $x = \frac{1}{2}$.

d) `A=[0:0.5:4]'; plot2d("A*x*(1-x)",xmin=0,xmax=1);`

Capítulo 3 - S.E.L. - Métodos Directos

Ejercicio 3:

Luego de aplicar tres transformaciones, cada una involucrando dos operaciones entre filas: $F_2^* = F_2 + (-1)F_1$, $F_3^* = F_3 + (-1)F_1$; $F_1^* = F_1 + (-1)F_2$, $F_3^* = F_3 + (-2)F_2$; $F_1^* = F_1 + (1)F_3$, $F_2^* = F_2 + (-2)F_3$; se obtiene la inversa:

$$\mathbf{A} = \begin{bmatrix} 3 & -3 & 1 \\ -3 & 5 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$

Ejercicio 5:

La inversa de la matriz $\mathbf{A}$ es:

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & -\delta \\ 0 & 1 \end{bmatrix},$$

con lo que las normas son $\|\mathbf{A}\|_\infty = \max\{1+|\delta|, 1\} = 1 + |\delta|$ y $\left\|\mathbf{A}^{-1}\right\|_\infty = \max\{1 + |-\delta|, 1\} = 1+|\delta|$. El producto de ambas normas es $(1+|\delta|)^2$. Si δ es pequeño, $\mathbf{A}$ es bien condicionada. Si δ es grande, $\mathbf{A}$ es mal condicionada.

Ejercicio 7:

a) $\|\mathbf{A}\|_\infty = 0{,}7469$; $\left\|\mathbf{A}^{-1}\right\|_\infty = 3{,}757$; $\mathcal{K}_\infty(\mathbf{A}) = 2{,}806$;

$$\mathbf{A}^{-1} = \begin{bmatrix} -1{,}014 & 2{,}326 \\ 2{,}583 & -1{,}174 \end{bmatrix}$$

b) $\|\mathbf{B}\|_\infty = 1{,}523$; $\left\|\mathbf{B}^{-1}\right\|_\infty = 3{,}828$; $\mathcal{K}_\infty(\mathbf{B}) = 5{,}830$;

$$\mathbf{B}^{-1} = \begin{bmatrix} 1{,}4 & -0{,}3724 \\ -1{,}828 & 2 \end{bmatrix}$$

c) $\|\mathbf{C}\|_\infty = 0{,}3025$; $\left\|\mathbf{C}^{-1}\right\|_\infty = 4{,}376$; $\mathcal{K}_\infty(\mathbf{C}) = 1{,}323$;

$$\mathbf{C}^{-1} = \begin{bmatrix} 4{,}726 & -0{,}1001 \\ -0{,}05006 & 3{,}346 \end{bmatrix}$$

d) $\|\mathbf{D}\|_\infty = 1{,}179$; $\left\|\mathbf{D}^{-1}\right\|_\infty = 12580$; $\mathcal{K}_\infty(\mathbf{D}) = 14840$;

$$\mathbf{D}^{-1} = \begin{bmatrix} 136{,}1 & -5739 \\ -289{,}1 & 12300 \end{bmatrix}$$

Ejercicio 9:

$\|\mathbf{r}_1\|_\infty = 7{,}50 \times 10^{-4}$; $\|\mathbf{r}_2\|_\infty = 2{,}54 \times 10^{-4}$.

Ejercicio 11:

En todos los ítems se utilizó descomposición de Doolittle.

a)

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 1{,}459 & 1 & 0 \\ 1{,}944 & 0{,}5435 & 1 \end{bmatrix}; \quad \mathbf{U} = \begin{bmatrix} 0{,}3959 & 0{,}6252 & 0{,}2368 \\ 0 & -0{,}7657 & -0{,}007099 \\ 0 & 0 & 0{,}4454 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} 1{,}023 & 2{,}773 & -1{,}31 \\ 1{,}927 & -1{,}293 & -0{,}02081 \\ -2{,}581 & -1{,}22 & 2{,}245 \end{bmatrix}$$

Solución de los ejercicios de número impar

b)

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0{,}9119 & 1 & 0 \\ 0{,}8172 & -0{,}2318 & 1 \end{bmatrix}; \quad \mathbf{U} = \begin{bmatrix} 0{,}9061 & 0{,}7827 & 0{,}8737 \\ 0 & -0{,}6684 & -0{,}5963 \\ 0 & 0 & 0{,}1411 \end{bmatrix}$$

$$\mathbf{B}^{-1} = \begin{bmatrix} 1{,}335 & 0{,}9739 & -1{,}372 \\ 7{,}863 & -2{,}961 & -6{,}322 \\ -7{,}285 & 1{,}642 & 7{,}087 \end{bmatrix}$$

c)

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 1{,}224 & 1 & 0 \\ 2{,}686 & -0{,}6035 & 1 \end{bmatrix}; \quad \mathbf{U} = \begin{bmatrix} 0{,}3517 & 0{,}3487 & 0{,}0223 \\ 0 & 0{,}5242 & 0{,}7987 \\ 0 & 0 & 1{,}276 \end{bmatrix}$$

$$\mathbf{C}^{-1} = \begin{bmatrix} 1{,}273 & -1{,}206 & 1{,}134 \\ 1{,}755 & 1{,}186 & -1{,}194 \\ -2{,}683 & 0{,}4729 & 0{,}7836 \end{bmatrix}$$

Ejercicio 13:

La matriz $\mathbf{A}$ es definida no negativa. La matriz $\mathbf{B}$ es definida positiva.

Ejercicio 15:

Suponiendo posible la descomposición LU de Doolittle:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 \\ L_{21} & 1 \end{bmatrix}; \quad \mathbf{U} = \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix};$$

$$\mathbf{A} = \mathbf{LU} = \begin{bmatrix} U_{11} & U_{12} \\ L_{21}U_{11} & L_{21}U_{12} + U_{22} \end{bmatrix},$$

pero es imposible que $U_{11} = 0$ y $L_{21}U_{11} = 1$.

Ejercicio 17:

Sí, debido a que los autovalores de $\mathbf{A}$, $\lambda_1, \lambda_2, \ldots, \lambda_n$, son necesariamente positivos y reales. Por lo tanto los autovalores de $\mathbf{A}^{-1}$ también son positivos y reales: $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_n}$.

Ejercicio 19:

Independientemente de la norma matricial utilizada:

$$\mathcal{K}(\alpha\mathbf{A}) = \|\alpha\mathbf{A}\| \left\|(\alpha\mathbf{A})^{-1}\right\| = |\alpha| \, \|\mathbf{A}\| \left|\frac{1}{\alpha}\right| \left\|\mathbf{A}^{-1}\right\| = \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| = \mathcal{K}(\mathbf{A})$$

Ejercicio 21:

Operando con 3 dígitos de mantisa, $\mathbf{x} = [0{,}999; -99{,}9]^T$.

a) $\mathbf{x} = [1{,}14; -85{,}7]^T$.

b) $\mathbf{x} = [-8{,}65 \times 10^{15}; -8{,}67 \times 10^{17}]^T$.

c) $\mathbf{x} = [-1{,}15; -325]^T$.

Ejercicio 23:

$a_0 = 10$; $a_1 = 34$; $a_2 = -9$; $a_3 = 0$.

Ejercicio 25:

$a > 0$; $c > 0$; no hay condiciones para b.

Capítulo 4 - S.E.L. - Métodos Iterativos

Ejercicio 3:

La matriz $\mathbf{A}$ se generó con `A=CondMat(10^4,2)` y se operó con una mantisa de 4 dígitos:

$$\mathbf{A} = \begin{bmatrix} 0{,}5952 & 0{,}7472 \\ 0{,}2382 & 0{,}2990 \end{bmatrix}$$

La solución obtenida por Gauss:

$$\mathbf{x} = \begin{bmatrix} 73730 \\ -58740 \end{bmatrix}$$

Resolviendo por Refinamiento Iterativo, en sólo un paso:

$$\mathbf{x} = \begin{bmatrix} 73750 \\ -58750 \end{bmatrix}$$

Ejercicio 5:

Con el método de Jacobi, una aritmética de 3 dígitos y truncamiento, el sistema no converge a la solución.

Ejercicio 9:

Los axiomas del elemento nulo y la linealidad se cumplen. No se cumple al axioma de la desigualdad triangular, por ejemplo con este par de matrices:

$$\mathbf{A} = \begin{bmatrix} 0{,}0469 & 0{,}0469 \\ 0{,}979 & 0{,}457 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0{,}513 & 0{,}941 \\ 0{,}253 & 0{,}226 \end{bmatrix}$$

Ejercicio 11:

La solución por Gauss es $\mathbf{x} = [-1{,}25; 10; 0]^T$. El sistema planteado no converge al utilizar Gauss Seidel.

Ejercicio 13:

La iteración de Jacobi es:

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} + \mathbf{D}_A^{-1}\left(\mathbf{b} - \mathbf{A}\mathbf{x}^{(n)}\right),$$

ahora, $\mathbf{D}_{AB} = \mathbf{B}\mathbf{D}_A$, entonces:

$$\begin{aligned} \mathbf{x}^{(n+1)} &= \mathbf{x}^{(n)} + (\mathbf{B}\mathbf{D}_A)^{-1}\left(\mathbf{B}\mathbf{b} - \mathbf{B}\mathbf{A}\mathbf{x}^{(n)}\right) \\ &= \mathbf{x}^{(n)} + \mathbf{D_A}^{-1}\mathbf{B}^{-1}\mathbf{B}\left(\mathbf{b} - \mathbf{A}\mathbf{x}^{(n)}\right) \\ &= \mathbf{x}^{(n)} + \mathbf{D_A}^{-1}\left(\mathbf{b} - \mathbf{A}\mathbf{x}^{(n)}\right) \end{aligned}$$

Ejercicio 15:

Este sistema fue descrito por Wilkinson en un artículo de 1961 sobre estabilidad numérica en operaciones matriciales.

a) $\rho(\mathbf{P}) = 1{,}4565$

b) El sistema converge a $\mathbf{x} = [0{,}39472; 0{,}62470]^T$, que es la solución, pero después de una gran cantidad de iteraciones. Este fenómeno es frecuente debido a la mala condición de la matriz de iteración, que además es casi singular, y entonces genera perturbaciones en la aritmética de punto flotante. Se converge a la solución del sistema en forma excesivamente lenta.

Ejercicio 17:

Se desea calcular el valor de $\frac{1}{r}$, entonces sea $f(x) = \frac{1}{x} - r$. El esquema iterativo de Newton Raphson es $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$, por lo que:

$$\begin{aligned} x_{n+1} &= x_n - \frac{\frac{1}{x_n} - r}{-\frac{1}{x_n^2}} \\ &= x_n + x_n(1 - rx_n) \\ &= x_n(2 - rx_n) \end{aligned}$$

Este esquema es convergente siempre que $\rho(\mathbf{I} - \mathbf{AB}^{(0)}) < 1$.

Ejercicio 19:

Para ambos sistemas, se utilizó $\mathbf{x}^{(0)} = [2; -1]^T$.

a) $\omega_{opt} = 1{,}4$, logra convergencia en 19 iteraciones. Analizando en forma más profunda, $\omega_{opt} = 1{,}3168$, convergiendo en 14 iteraciones.

b) $\omega_{opt} = 1$, logra convergencia en 9 iteraciones. Analizando en forma más profunda, $\omega_{opt} = 1{,}0366$ o sino $\omega_{opt} = 1{,}0540$, convergiendo con cualquiera de ambos parámetros en 6 iteraciones.

Ejercicio 21:

A y **b** se componen a través de los siguientes bloques:

$$\mathbf{A}_1 = \begin{bmatrix} 0{,}925 & 0{,}650 \\ 0{,}464 & 0{,}900 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 0{,}833 & 0{,}172 \\ 0{,}859 & 0{,}00712 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0{,}484 & 0{,}715 \\ 0{,}563 & 0{,}544 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{bmatrix} 0{,}895 \\ 0{,}813 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 0{,}559 \\ 0{,}896 \end{bmatrix}$$

En 87 iteraciones se obtiene $\mathbf{x} = [5{,}77; 7{,}21]^T$, $\mathbf{y} = [-7{,}23; -7{,}89]^T$. Con los métodos de Jacobi y Gauss Seidel ambos iterados divergen, ya que $\rho(\mathbf{B}_J) = 12{,}25$ y $\rho(\mathbf{B}_{GS}) = 26{,}04$.

Ejercicio 23:

Sea $\mathbf{x}$ un autovector de $\mathbf{A}$ asociado a λ tal que $\|\mathbf{x}\| = 1$ y $\rho(\mathbf{A}) = |\lambda|$. Entonces:

$$\rho(\mathbf{A}) = |\lambda| = |\lambda| \, \|\mathbf{x}\| = \|\lambda \mathbf{x}\| = \|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\| \, \|\mathbf{x}\| = \|\mathbf{A}\|$$

Ejercicio 25:

El sistema de ecuaciones:

$$\begin{aligned} 0{,}6232x_1 + 0{,}04384x_2 + 0{,}6298x_3 &= 0{,}9502 \\ 0{,}3633x_1 + 0{,}5284x_2 + 0{,}7228x_3 &= 0{,}08094 \\ 0{,}2564x_1 + 0{,}2823x_2 + 0{,}7924x_3 &= 0{,}9269 \end{aligned}$$

cumple con las condiciones pedidas y es convergente si se aplica el método de Gauss Seidel.

Capítulo 5 - Autovalores

Ejercicio 3:

a) $P_A(\lambda) = \lambda^3 - 14\lambda^2 + 87\lambda - 354$; $P(8{,}6751) = P(2{,}6624 + 5{,}8067i) = P(2{,}6624 - 5{,}8067i) = 0$

b) $P_B(\lambda) = \lambda^3 - 18\lambda^2 + 137\lambda - 36$; $P(8{,}8638 + 7{,}3215i) = P(8{,}8638 - 7{,}3215i) = P(0{,}27237) = 0$

c) $P_C(\lambda) = \lambda^3 - \lambda^2 - 43\lambda + 16$; $P(6{,}8977) = P(0{,}37009) = P(-6{,}2678) = 0$

Ejercicio 5:

a) $\lambda_3 = -0{,}071588$

b) $\lambda_{2,3} = -0{,}29817 \pm 0{,}099980i$

c) $\lambda_3 = 0{,}17682$

Ejercicio 7:

Sea $\mathbf{A}$ no singular, entonces existe $\mathbf{A}^{-1}$ y, por propiedades de determinantes:

$$\begin{aligned} |\mathbf{I} - \mathbf{AB} - \lambda\mathbf{I}| &= \left|\mathbf{A}^{-1}\right| |\mathbf{I} - \mathbf{AB} - \lambda\mathbf{I}| \, |\mathbf{A}| \\ &= \left|\mathbf{A}^{-1} (\mathbf{I} - \mathbf{AB} - \lambda\mathbf{I}) \, \mathbf{A}\right| \\ &= \left|\left(\mathbf{A}^{-1} - \mathbf{B} - \lambda\mathbf{A}^{-1}\right) \mathbf{A}\right| \\ &= |\mathbf{I} - \mathbf{BA} - \lambda\mathbf{I}| \, . \end{aligned}$$

Como ambos polinomios característicos son iguales, entonces ambas matrices tienen los mismos autovalores.

Ejercicio 9:

Como $\mathbf{L}$ es triangular superior y su diagonal principal contiene sólo unos, entonces $|\mathbf{L}| = 1$ y por lo tanto existe $\mathbf{L}^{-1}$:

$$\begin{aligned} |\mathbf{A} - \lambda\mathbf{I}| &= |\mathbf{LU} - \lambda\mathbf{I}| \\ &= \left|\mathbf{L}^{-1}\right| |\mathbf{LU} - \lambda\mathbf{I}| \, |\mathbf{L}| \\ &= |\mathbf{UL} - \lambda\mathbf{I}| \, . \end{aligned}$$

Como ambos polinomios característicos son iguales, entonces ambas matrices tienen los mismos autovalores.

Ejercicio 11:

Si la matriz $\mathbf{A}$ es de orden n, entonces la cantidad de discos de Gerschgorin en n. Ahora, cada disco debe contener un autovalor, ya que no pueden existir regiones de Gerschgorin disjuntas vacías. Esto excluye la posibilidad de autovalores complejos, ya que en los polinomios reales si $a + bi$ es raíz, $a - bi$ también lo es. ahora, si $\lambda = a + bi \in D_k$ entonces $\overline{\lambda} = a - bi \in D_k$ lo cual no es compatible con el segundo teorema de Gerschgorin. Por lo tanto, $\mathbf{A}$ sólo tiene autovalores reales.

Ejercicio 13:

Aplicando el segundo teorema de Gerschgorin:

a) $\mathcal{M}_1 = D_1 \cup D_2 \cup D_3$ y $\mathcal{M}_2 = D_4$, por lo tanto puede haber como máximo dos autovalores complejos (conjugados) en $\mathcal{M}_1$.

b) $\mathcal{M}_1 = D_1$ y $\mathcal{M}_2 = D_2 \cup D_3 \cup D_4$, por lo tanto puede haber como máximo dos autovalores complejos (conjugados) en $\mathcal{M}_2$.

Ejercicio 15:

Esto ocurre porque $\mathbf{x} = [4; -5; 1]^T$ es uno de los autovectores asociados a $\lambda = -3$, cualquier otro vector que no sea una compresión o expansión de $\mathbf{x}$ convergirá a $\lambda = 4$, siempre dentro de los límites de precisión numérica de la computadora utilizada.

Ejercicio 17:

Deben reemplazarse dos líneas de código, programando en *EMT*:

- `q=z/norm(z)` reemplazar por `q=z/norm(z,0)`
- `L=q'.A.q` reemplazar por `L=q'.A.q/(q'.q)`

Ejercicio 19:

Si $\mathbf{A} = \mathbf{PDP}^{-1}$ y $\mathbf{A}^n = \mathbf{PD}^n\mathbf{P}^{-1}$, entonces:

$$\begin{aligned}\mathbf{A}^{n+1} &= \mathbf{A}^n\mathbf{A} \\ &= \mathbf{PD}^n\mathbf{P}^{-1}\mathbf{PDP}^{-1} \\ &= \mathbf{PD}^n\mathbf{DP}^{-1} \\ &= \mathbf{PD}^{n+1}\mathbf{P}^{-1}\end{aligned}$$

Ejercicio 21:

a) El código está disponible en el Anexo: Códigos para *Euler Math Toolbox*.

b) Con la matriz $\mathbf{A}$ dada, se obtiene la siguiente salida:

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 4 & -1 & 0 & 0 \\ -69 & -7 & 1 & 0 \\ 53 & 80 & 6 & -1 \end{bmatrix}$$

Ejercicio 23:

a) Calculando sólo los valores de $P_i(\alpha)$:

- $P_0(4) = 1$; $P_1(4) = 1$; $P_2(4) = -4$; $P_3(4) = -1$; $P_4(4) = 15$, por lo tanto existen dos autovalores menores que 4.
- $P_0(2) = 1$; $P_1(2) = 3$; $P_2(2) = 2$; $P_3(2) = 1$; $P_4(2) = -5$, por lo tanto existe un autovalor menor que 2.

b) $P(\lambda) = \lambda^4 - 18\lambda^3 + 112\lambda^2 - 278\lambda + 231$.

c) Luego de 1 iteración, la raíz de $P(\lambda)$ es $\lambda = 3$.

Ejercicio 25:

```
>A=[1,2,3;-2,4,5;-3,-5,4];
```

a)
```
>x=[1;1;1]; k=300; L=zeros(1,k);
>for i=1 to k, x=A.x/norm(A.x); L[i]=x'.A.x; end
>plot2d(L);
```

b)
```
>x=[1;1;1]; k=1.5E5; L=zeros(1,k);
>for i=1 to k, x=A.x/norm(A.x); L[i]=x'.A.x; end
>mean(L)
 3.53922351004
```

c) En ambos casos, se utiliza para graficar y calcular el promedio, las 200 iteraciones siguientes a las indicadas en el enunciado.

```
>B=inv(A-13.817*eye(3));
>x=[1;1;1]; for i=1 to 5E5, x=B.x/norm(B.x); end
>L=zeros(1,200);
>for i=1 to 200, x=B.x/norm(B.x); L[i]=x'.A.x; end
>plot2d(L); mean(L)
 1.92172939569
>B=inv(A-13.818*eye(3));
>x=[1;1;1]; for i=1 to 5E5, x=B.x/norm(B.x); end
>L=zeros(1,200);
>for i=1 to 200, x=B.x/norm(B.x); L[i]=x'.A.x; end
>plot2d(L); mean(L)
 3.41401005706
```

Esto ocurre porque, si $\mu \leq 13{,}817$, es menor la distancia del desplazamiento al autovalor real, por lo que el algoritmo converge muy lentamente. En cambio si $\mu \geq 13{,}818$, la distancia es menor con respecto a ambos autovalores complejos, por lo que el algoritmo oscilará sin lograr convergencia.

Capítulo 6 - Sistemas de Ecuaciones No Lineales

Ejercicio 3:

a)
```
>plot2d("x^2-2x-y+0.5",level=0,a=-4,b=4,c=-2,d=2)
>plot2d("x^2+4y^2-4",level=0,add=1)
```

b) Utilizando $\mathbf{x}^{(0)} = [5; 3{,}5]^T$ y una aritmética con truncamiento de 8 dígitos, converge a $[1{,}9006767; 0{,}31121856]^T$ en 8 iteraciones.

c) Con $\alpha \approx 2{,}12356$ se obtiene estabilidad de los cocientes $\Delta x_{k+2}/\Delta x_{k+1}$ en 0,801133.

Ejercicio 5:

Se logra convergencia a $[0{,}63889691; 0{,}76929235]^T$ después de 3 iteraciones.

Ejercicio 7:

Independiente de la semilla elegida, el sistema *a*) converge en sólo en $[1; 0]^T$, mientras que el sistema *b*) converge sólo en $[0; 1]^T$.

Ejercicio 9:

a)
```
>a=-1; b=2; c=4;
>plot2d("x^2+y^2-(c-a^3/b^2)*x+2*b*y-b^2-c*a^3/b^2",level=0,
  a=-2,b=6,c=-6,d=2);
>plot2d("x*y",level=a^3/b,add=1);
```

b) La ecuación de circunferencia queda reducida a un solo punto. El sistema de ecuaciones tiene por solución: $[1; -1]^T$, lo que coincide con la solución real de la ecuación cúbica asociada: $x = 1$;

Ejercicio 11:

a)
```
>plot2d("x^2-3y+cos(x*y)",level=0,a=-4,b=0,c=0,d=4);
>plot2d("x+y-sin(x)*sin(y)",level=0,add=1);
```

b) Converge a $[-3{,}2053306; 3{,}2015160]^T$ luego de 30 iteraciones, con un valor de tolerancia de $\varepsilon = 1 \times 10^{-5}$.

c) La variable y converge en 29 iteraciones, mientras que x converge en 30 iteraciones.

Ejercicio 13:

a) Una posible región $\mathcal{R}$ es $[-2; 2] \times [-2; 2]$.

b)
```
>plot2d("x+x^3+2*x*y+x*y^2",level=0.5,a=-2,b=2,c=-2,d=2);
>plot2d("y+y*x^2-2*x*y^2+y^3",level=0.5,add=1);
```

c) Con un error de tolerancia de 1×10^{-5} y aritmética de 8 dígitos con truncamiento, se logra la convergencia a [0,3072664; 0,48473223] en 11 iteraciones.

Ejercicio 15:

a) Puede minimizarse el cuadrado de la expresión de la distancia:

$$D(x, u) = (x - u)^2 + (x^2 + 3 * x + 5 - \cos(u))^2,$$

donde las funciones que componen el Gradiente:

$$g_1(x, u) = \partial D(x, u)/\partial x$$
$$g_2(x, u) = \partial D(x, u)/\partial u$$

permiten generar el Jacobiano.

b) Utilizando como semilla a $[0;0]^T$, se logra la convergencia luego de 6 iteraciones a $[-1{,}2815405; -0{,}45217055]$. Se utilizó una aritmética de 8 dígitos con truncamiento.

Ejercicio 17:

Utilizando el método de Newton Rapshon, un error de tolerancia de 1×10^{-5} y:

- $[0{,}8;0]^T$ como semilla, se obtiene la convergencia a $[0{,}88159259; 0{,}21359471]^T$ en 4 iteraciones.
- $[1{,}4;3]^T$ como semilla, se obtiene la convergencia a $[1{,}3294021; 3{,}0618225]^T$ en 5 iteraciones.

Ejercicio 19:

a) Partiendo de la semilla $[2;2]^T$, se logra la convergencia en 6 iteraciones. Los resultados intermedios se muestran en la tabla 11.18, donde se verifica que $x_k + y_k = 1$.

b) Partiendo de la semilla $[2;2;2]^T$, se logra la convergencia en 6 iteraciones. Los resultados intermedios se muestran en la tabla 11.19, donde se verifica que $3x_k + 4y_k + z_k = -2$ y $-x_k + y_k - z_k = -1$.

k	x_k	y_k
0	2,0000000E+00	2,0000000E+00
1	2,3127091E+00	-1,3127091E+00
2	1,3567677E+00	-3,5676775E-01
3	1,0464386E+00	-4,6438667E-02
4	1,0008879E+00	-8,8791974E-04
5	1,0000003E+00	-3,2842950E-07
6	1,0000000E+00	0,0000000E+00

Tabla 11.18

k	x_k	y_k	z_k
0	2,0000000E+00	2,0000000E+00	2,0000000E+00
1	1,7597972E+00	-1,3039188E+00	-2,0637161E+00
2	5,0623552E-01	-8,0249420E-01	-3,0872973E-01
3	-1,3387528E-01	-5,4644988E-01	5,8742539E-01
4	-1,4557778E-01	-5,4176888E-01	6,0380889E-01
5	-1,4559041E-01	-5,4176383E-01	6,0382657E-01
6	-1,4559041E-01	-5,4176383E-01	6,0382657E-01

Tabla 11.19

Ejercicio 21:

a) $\|\mathbf{J_F}(\mathbf{x})\| = x^2 - y + 1$.

b) $\mathbf{x}^{(0)} = [x_0; y_0]^T$; $\mathbf{x}^{(1)} = \dfrac{1}{x_0^2 - y_0 + 1}[x_0^3 + 1 - x_0 y_0; x_0(x_0^3 + 2 - x_0 - x_0 y_0)]^T$. No se verifica que, al sumarle 1 al cuadrado de la primera componente, el resultado sea igual a la segunda componente.

Ejercicio 23:

Cada función se considera evaluada en x_k, y_k:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - J_F^{-1}\left(\mathbf{x}^{(0)}\right) F\left(\mathbf{x}^{(0)}\right)$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - \frac{1}{f_x g_y - f_y g_x} \begin{bmatrix} g_y & -f_y \\ -g_x & f_x \end{bmatrix} \begin{bmatrix} f \\ g \end{bmatrix}$$

$$= \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - \frac{1}{f_x g_y - f_y g_x} \begin{bmatrix} f g_y - f_y g \\ f_x g - f g_x \end{bmatrix}$$

Entonces, separando los vectores en sus componentes iterativas:

$$x_1 = x_0 - \frac{f g_y - f_y g}{f_x g_y - f_y g_x}; \qquad y_1 = y_0 - \frac{f_x g - f g_x}{f_x g_y - f_y g_x}$$

Ejercicio 25:

a)
```
>plot2d("x^2+y^2-25",level=0,a=-10,b=10,c=-6,d=6);
>plot2d("x^2-y^2-2",level=0,add=1);
```

b) La región simétrica para las raíces del primer y segundo cuadrante, así como para las raíces del tercer y cuarto cuadrante, son los puntos de la forma $(x; 0)$. La región simétrica para las raíces del segundo y tercer cuadrante, como para las raíces del primer y cuarto cuadrante son los puntos de la forma $(0; y)$. Sin embargo, no se puede elegir ningún punto de ambas regiones como semilla, puesto que $\mathbf{J_F}(\mathbf{x}) = -8xy$.

c) Cualquier punto de $\mathbb{R}^2$ que no esté en las regiones descritas en el inciso anterior converge a alguna de las 4 soluciones identificadas en el gráfico.

Capítulo 7 - Interpolación

Ejercicio 3:

a) $p(x) = 0{,}0004999x^3 - 0{,}05x^2 + 2{,}0499x + 12$.

b) No es posible construir un *spline* de Hermite por no tener información de la derivada.

c) $p(25) = 39{,}808$; $p(35) = 43{,}929$; $p(45) = 48{,}548$.

Ejercicio 5:

Utilizando la partición equiespaciada del intervalo dado:

a) $L(x) = -0{,}024115x^4 + 0{,}35281x^3 - 1{,}4821x^2 + 1{,}3795x - 0{,}15852$

b) $s_1(x) = 0{,}046203x^3 - 0{,}53974x^2 + 0{,}54030x - 0{,}15852$
$s_2(x) = 0{,}13657x^3 - 0{,}87167x^2 + 0{,}87176x - 0{,}20639$
$s_3(x) = -0{,}094891x^3 + 1{,}5912x^2 - 7{,}8621x + 10{,}115$
$s_4(x) = -0{,}10274x^3 + 1{,}5609x^2 - 6{,}8936x + 7{,}0051$

Ejercicio 7:

Para ambos incisos se utiliza la tabla 11.20:

x	$k=0$	$k=1$	$k=2$	$k=3$
10,000	33,000	35,000	34,875	34,750
20,000	37,000	34,500	34,125	-
30,000	42,000	36,000	-	-
40,000	46,000	-	-	-

Tabla 11.20

a) Solubilidad a $15°C$: 34.875

b) Solubilidad a $15°C$: 34.750

Ejercicio 9:

No es posible interpolar $f(1{,}4)$ con los valores dados, ya que esa abscisa no está dentro del intervalo de interpolación. Es necesario utilizar alguna técnica de extrapolación en este caso.

Ejercicio 11:

```
>x=[-3,  -1.501,  -0.8344,  1.912,  2.558,  4]
 [-3,  -1.501,  -0.8344,  1.912,  2.558,  4]
>S=Spline(x,x,5)
 5.13e-005         1         0         0
 1.678e-005        1         0         0
 7.872e-006  0.99999         0         0
 -3.38e-005        1         0         0
 -3.62e-005        1         0         0
>plot2d(x,x,color=blue);
>plot2d("polyval(S[1,:],x)",x[1],x[2],add=1,color=red,style=".-");
>plot2d("polyval(S[2,:],x)",x[2],x[3],add=1,color=red,style=".-");
>plot2d("polyval(S[3,:],x)",x[3],x[4],add=1,color=red,style=".-");
>plot2d("polyval(S[4,:],x)",x[4],x[5],add=1,color=red,style=".-");
>plot2d("polyval(S[5,:],x)",x[5],x[6],add=1,color=red,style=".-");
```

Ejercicio 13:

Es mejor utilizar un *spline* de Hermite puesto que tiene menos error con respecto a la curva original. Si bien no se asegura la continuidad de la derivada segunda, sí se asegura que $s \in C^1$.

Ejercicio 15:

Se utilizará el rectángulo que tiene por vértices a: $[0;0]$, $[2;0]$, $[2;2]$ y $[0;2]$. Además, se agregan los puntos intermedios de cada segmento: $[1;0]$, $[2;1]$, $[1;2]$ y $[0;1]$.

```
>t=1:9; x=[0,0,0,1,2,2,2,1,0]; y=[2,1,0,0,0,1,2,2,2];
>tx=spline(t,x);
>ty=spline(t,y);
>X=splineval(linspace(1,9,100),t,x,tx);
>Y=splineval(linspace(1,9,100),t,y,ty);
>plot2d(X,Y,color=blue,a=-1,b=3,c=-1,d=3);
>plot2d(x,y,points=1,style="o#",add=1);
```

Ejercicio 17:

a) $p(x) = -0{,}066305x^3 + 0{,}97145x^2 - 3{,}9931x + 4{,}2096$

b) $p(x) = -0{,}066305(x-5{,}4747)^3 - 0{,}11756(x-5{,}4747)^2 + 0{,}68170(x-5{,}4747) + 0{,}58528$

c) Para el sistema original: $\mathcal{K}_\infty = 1{,}39 \times 10^4$. Para el sistema desplazado: $\mathcal{K}_\infty = 8{,}14 \times 10^1$.

Ejercicio 19:

Graficando $P_1(x)$ se sabe que las raíces son $[1;2;4;5;7]$ y que $P(9) = 1$, con lo que ya se identificaron los nodos de interpolación. Quien no cumple con estos nodos es $P_4(x)$, la versión correcta es:

$$P_4(x) = -\frac{1}{90}x^5 + \frac{4}{15}x^4 - \frac{104}{45}x^3 + \frac{131}{15}x^2 - \frac{1231}{90}x + 7$$

Ejercicio 21:

Utilizando la construcción del polinomio de Lagrange:

$$\begin{aligned} q(x) &= p(x) - 61l_6(x) + 30l_6(x) \\ &= p(x) - 31l_6(x) \\ &= x^4 - x^3 + x^2 - x + 1 - 31\frac{(x-2)(x-1)x(x+1)(x+2)}{120} \end{aligned}$$

Ejercicio 23:

El polinomio mínimo que cumple con las condiciones del enunciado es $p(x) = 7{,}8l_3(x)$, puesto que $l_3(x)$ vale 0 en todos los nodos, salvo en $x = 2{,}61$ que vale 1.

Ejercicio 25:

```
>x=linspace(-5,5,8); y=1/(1+x^2);
>plot2d(x,y,points=1,style="o#",color=blue);
>d=interp(x,y); p=polytrans(x,d);
>plot2d("polyval(p,x)",xmin=-5,xmax=5,add=1,color=blue);
>i=0:8; x2=0.5*(5-5)+0.5*(-5-5)*cos((2*i+1)/(2*8+2)*pi);
>y2=1/(1+x2^2);
```

```
>plot2d(x2,y2,points=1,style="o#",color=red,add=1);
>d2=interp(x2,y2); p2=polytrans(x2,d2);
>plot2d("polyval(p2,x)",add=1,color=red);
```

Capítulo 8 - Ajuste de Datos

Ejercicio 3:

Debe linealizarse la función de acuerdo al siguiente esquema:

$$\ln\left(\frac{1}{y}-1\right) = -ax$$
$$Y = Ax,$$

por lo que el valor de la constante a determinar es: $a = -0{,}5046$.

Ejercicio 5:

Debe linealizarse la función sugerida de acuerdo al siguiente esquema:

$$\ln(y) - \ln(x) = \ln(a) + bx$$
$$Y = A + bx,$$

donde las constantes originales toman los valores: $a = 3{,}06$; $b = -2{,}01$, cometiendo el error: $\|\mathbf{e}\|_2^2 = 5{,}72 \times 10^{-4}$.

Ejercicio 7:

Los coeficientes de mejor ajuste son: $a_0 = 3{,}1347$ y $a_1 = 0{,}074451$.

Ejercicio 9:

Los coeficientes de mejor ajuste para un polinomio de grado tres son: $a_0 = -0{,}084833$; $a_1 = 1{,}6918$; $a_2 = -0{,}79628$ y $a_3 = 0{,}084637$, dados en orden creciente. El polinomio de grado cinco que ajusta los datos dados tiene por coeficientes a: $a_0 = -0{,}055778$; $a_1 = 1{,}093$; $a_2 = 0{,}1067$; $a_3 = -0{,}3485$; $a_4 = 0{,}082131$ y $a_5 = -0{,}0053744$. Sin embargo, no es lo ideal pues el número de condición de la matriz de coeficientes es del orden de 10^9 y los coeficientes de orden superior son pequeños.

Ejercicio 11:

Los valores de las constantes del modelo elegido es: $a = 5$; $b = 4$ y $c = 3$.

Ejercicio 13:

El cálculo analítico de la recta *minimax* de la función $f(x) = e^x$ en $[-1; 1]$ es:

$$y^* = 1{,}175201x + 1{,}264279,$$

mientras que, al aplicar el algoritmo de Remez al conjunto inicial $X^{(0)} = \{-1; 0; 1\}$ se obtiene:

$$X^{(1)} = \{-1;\ 0{,}24019;\ 1\}$$
$$X^{(2)} = \{-1;\ 0{,}23568;\ 1\}$$
$$X^{(3)} = \{-1;\ 0{,}23570;\ 1\}$$
$$X^{(4)} = \{-1;\ 0{,}23570;\ 1\}$$

Con lo que el polinomio *minimax* de grado uno es: $y^* = 1{,}2658 + 1{,}1751x$.

Ejercicio 15:

$p_4(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}$
$p_3^* = 0{,}997076 + 1{,}0119x + 0{,}541665x^2 + 0{,}154756x^3$

Ejercicio 17:

El ajuste de segundo grado es $p_2(x) = -2{,}8 + 0{,}36571x + 0{,}354x^2$. Es posible recuperar en forma simple el valor del coeficiente principal, debido a que la cantidad de puntos es par, al igual que el grado del polinomio interpolador.

Ejercicio 19:

El modelo que ajusta los datos es: $c = 1976e^{-0{,}05314t}$.

a) $c(0) = 1976$.

b) $c(43{,}10) = 200$.

Ejercicio 21:

a) El ajuste lineal de los datos es: $y_1 = 2{,}5065 - 0{,}97540x$.

b) El ajuste lineal de los datos intercambiados es: $y_2 = 1{,}0622 - 0{,}33999x$.

c) $\|\mathbf{e}_1\|_2^2 = 0{,}66836$; $\|\mathbf{e}_2\|_2^2 = 0{,}23297$; $I_1 = 0{,}33163$; $I_2 = 0{,}33161$.

Ejercicio 23:

El ajuste pasa siempre por la segunda coordenada y su expresión es: $y = e^{\ln(b)x}$. La primera coordenada no afecta el resultado final.

Ejercicio 25:

De la expresión de mínimos cuadrados:

$$\sum_{i=0}^{n} [ax_i + b - y_i] = 0$$

$$a\sum_{i=0}^{n} x_i + bn = \sum_{i=0}^{n} y_i$$

$$\frac{a}{n}\sum_{i=0}^{n} x_i + \frac{b}{n}n = \frac{1}{n}\sum_{i=0}^{n} y_i$$

$$a\overline{x} + b = \overline{y}$$

Capítulo 9 - Derivación Numérica

Ejercicio 3:

a) Utilizando la primera aproximación por diferencia hacia atrás, se logran los siguientes valores, de acuerdo al valor de h: 3,99884; 3,99917; 3,99921 y 3,99929.

b) Utilizando la primera aproximación por diferencia hacia adelante, se logran los siguientes valores, de acuerdo al valor de h: 0,810930; 0,911607; 0,931746 y 0,975803.

Ejercicio 5:

Sí, el *spline* cúbico aproxima bastante bien la curvatura de la función, por lo que dará una aproximación buena a la derivada. El *spline* de Hermite ofrece una mejor aproximación que el *spline* cúbico, pero no se lo puede construir porque su desarrollo depende de los valores de las derivadas en los nodos.

Ejercicio 7:

a) Utilizando la primera aproximación por diferencias centrales, se aproxima $f''(1{,}3) = -0{,}80000$.

b) Utilizando la primera aproximación por diferencias centrales, se aproxima $f''(-{,}8) = 14{,}000$.

c) Utilizando la primera aproximación por diferencias hacia atrás, se aproxima $f''(4) = 9{,}8686$.

d) Utilizando la primera aproximación por diferencias hacia adelante, se aproxima $f''(-1) = 2{,}3600$.

Ejercicio 9:

Considerando los desarrollos anteriores y $k = 2$:

$$\begin{aligned} f''(x) &\approx \frac{2^2\phi(h) - \phi(2h)}{2^2 - 1} \\ &\approx \frac{-f(x-2h) + 16f(x-h) - 30f(x) + 16f(x+h) - f(x+2h)}{12h^2} + \mathcal{O}(h^4) \end{aligned}$$

Ejercicio 11:

Cambiando el valor de n dentro del código se obtienen los distintos gráficos pedidos:

```
>function f(x)::=cos(x)/x
>n=10;
>x=linspace(1,7,n-1); d1y=zeros(1,n); d2y=zeros(1,n); h=x[2]-x[1];
>for i=2 to n-1, d1y[i]=(f(x[i+1])-f(x[i-1]))/(2*h); end
>plot2d("&diff(f(x),x)",1,7);
>plot2d(x[2:length(x)-1],d1y[2:length(x)-1],add=1,color=red);
>
>for i=2 to n-1, d2y[i]=(f(x[i+1])-2*f(x[i])+f(x[i-1]))/h^2; end
>plot2d("&diff(f(x),x,2)",1,7);
>plot2d(x[2:length(x)-1],d2y[2:length(x)-1],add=1,color=blue);
```

Ejercicio 13:

a) El ajuste del conjunto de datos cuyas abscisas son: [0,9; 1,1; 1,3; 1,5], es $-0{,}539375x^2 - 0{,}96355x + 3{,}62$ y la derivada del ajuste en 1,3 es $-2{,}36592$.

b) El ajuste del conjunto de datos cuyas abscisas son: $[-1{,}4; -0{,}8; -0{,}2; 0{,}4]$, es $7x^2 - 4x - 2$ y la derivada del ajuste en $-0{,}8$ es $-15{,}2$.

c) El ajuste del conjunto de datos cuyas abscisas son: $[3; 3{,}3333; 3{,}6667; 4]$, es $4{,}23629x^2 - 21{,}0572x + 30{,}082$ y la derivada del ajuste en 4 es 12,8331.

d) El ajuste del conjunto de datos cuyas abscisas son: $[-1; -0{,}85; -0{,}7; -0{,}55]$, es $1{,}10333x^2 + 3{,}95303x + 0{,}416955$ y la derivada del ajuste en -1 es 1,74637.

Ejercicio 15:

a) $\dfrac{\partial f}{\partial y} \approx -39{,}0000$

b) $\dfrac{\partial^2 f}{\partial x^2} \approx 6{,}00000$

c) $\dfrac{\partial^2 f}{\partial x \partial y} \approx 2{,}99999$

Ejercicio 17:

En este caso se estabiliza alrededor de 4. Esto ocurre porque $h_i = h_{i-1}/2$ y el orden de convergencia de la fórmula es $\mathcal{O}(h^2)$:

$$\begin{aligned} R_i &= 3{,}999000 \\ &= 3{,}999750 \\ &= 3{,}999937 \\ &= 3{,}999986 \\ &= 4{,}000036 \end{aligned}$$

Ejercicio 19:

Con la función $f(x) = \cos(x)$, la abscisa $x = 2$ y tomando valores para h del vector $[0{,}1; 0{,}05; 0{,}025; 0{,}0125; 0{,}00625]$, se obtienen los siguientes cocientes:

$$\begin{aligned} R_i &= 3{,}99892 \\ &= 3{,}99973 \\ &= 3{,}99994 \\ &= 4{,}00163, \end{aligned}$$

con lo que el orden de convergencia es $\mathcal{O}(h^2)$.

Ejercicio 21:

Sí es posible, la expresión surge al restar las expansiones de $4f(x+h)$ y $f(x-2h)$. Su orden de convergencia es $\mathcal{O}(h^3)$.

Ejercicio 23:

El error para este tipo de polinomios es cero.

$$\begin{aligned} f'(x) &\approx \frac{2\left(8h^3 - 9h^2\right) + 12h^3 - 2\left(-8h^3 - 9h^2\right)}{4h^3} \\ &\approx 11 \end{aligned}$$

Ejercicio 25:

Las tres fórmulas de aproximación propuesta son del orden de $\mathcal{O}(h^3)$.

Capítulo 10 - Integración Numérica

Ejercicio 3:

En todos los casos se utilizó una aritmética de 5 dígitos y truncamiento.

a) Con 7 nodos y el método de rectángulos centrales: $I \approx 1{,}2695$.

b) Con 5 nodos y el método de rectángulos hacia atrás: $I \approx 1{,}8961$.

c) Con 26 nodos y el método de rectángulos hacia adelante: $I \approx 0{,}12239$.

Ejercicio 5:

En todos los casos se utilizó una aritmética de 5 dígitos y truncamiento.

a) $I \approx 3{,}2040$.

b) $I \approx 0{,}61828$.

c) $I \approx 0{,}72781$.

Ejercicio 7:

Los ejercicios se resolvieron con la cuadratura gaussiana y una aritmética de 6 dígitos, con lo que: $t_0 = -0{,}577350$ y $t_1 = -t_0$.

a) $I \approx 0{,}666670$.

b) $I \approx 0{,}545452$.

c) $I \approx 1{,}53175$.

Ejercicio 9:

Ambos incisos se calcularon con una aritmética de 7 dígitos y truncamiento.

a) $I \approx 3{,}307253 + \mathcal{O}(h^5)$

b) $I \approx -15{,}97944 + \mathcal{O}(h^5)$

Ejercicio 11:

Los coeficientes son:

$$w_0 = \frac{b-a}{2}; \qquad w_1 = w0; \qquad w_2 = \frac{a^2 - 2ab + b^2}{12}; \qquad w_3 = -w2;$$

y la fórmula de cuadratura es exacta para polinomios de grado tres ó menor.

Ejercicio 13:

Ambas aproximaciones no son buenas. Con dos puntos, $I \approx 2{,}449489$. Con tres puntos, $I \approx 2{,}552284$.

Ejercicio 15:

a) $I \approx 60{,}125$

b) El polinomio de ajuste es: $4{,}85066 + 0{,}0602885x + 0{,}175315x^2 - 0{,}0180009x^3$. Entonces, $I \approx 60{,}022$.

Ejercicio 17:

a) Una cota para la derivada segunda es 1, entonces $n \geq 26$.

b) Una cota para la derivada segunda es 4, entonces $n \geq 517$.

c) Una cota para la derivada segunda es 2, entonces $n \geq 11548$.

Ejercicio 19:

a) La convergencia del método de Simpson es de orden h^4, $p \approx 4$.

b) La convergencia del método de rectángulos centrales es de orden h^3, $p \approx 3$.

Ejercicio 21:

a) $\int_{x-h}^{x+h} f(x)\mathrm{d}x \approx 2h\left(f(x) + \dfrac{h^2 f''(x)}{3!} + \dfrac{h^4 f^{(4)}(x)}{5!}\right)$

b) De acuerdo al error de truncamiento estimado por la serie de Taylor: $E(h) \approx \dfrac{2h^7 f^{(6)}}{7!}$

c) $I \approx 4{,}94208$ y $E(h) \approx 3{,}10 \times 10^{-5}$.

Ejercicio 23:

En la mayoría de los casos se cumple que se logra una mejor aproximación por fórmula de tres puntos en vez de aplicar dos veces la fórmula de dos puntos. Esto se debe a que el error de truncamiento es, al menos, un orden de magnitud más grande en la fórmula de tres puntos.

Ejercicio 25:

La única aproximación que no es cierta, corresponde al segundo inciso.

Capítulo 11 - Resolución Numérica de EDO

Ejercicio 3:

No se logra convergencia a la solución exacta, sólo mejora la obtenida por el esquema predictor-corrector. Es equivalente a, en cada paso, resolver una ecuación no lineal asociada a la corrección. No es posible disminuir la cota de error de truncamiento del método utilizado, sea cual fuere la elección.

Ejercicio 5:

Las valores obtenidos por iteración son:

$$[0;\quad 0{,}039485;\quad 0{,}11628;\quad 0{,}22807;\quad 0{,}37375;\quad 0{,}5542]$$

Ejercicio 7:

Todas los métodos numéricos amplificarán el error de ingreso con el paso de las iteraciones. Por lo tanto, para un valor de h no muy pequeño, es muy probable se obtenga un *overflow* en $x = 10$, sino de seguro ocurrirá para $x = 20$.

Ejercicio 9:

Si $E(h) = \sum_{i=0}^{n} c_i h^p$, donde c_i depende de cotas impuestas sobre las derivadas de la función a resolver, entonces:

$$\begin{aligned} E(h) &= \sum_{i=0}^{n} c_i h^p \\ &\approx nch^p \\ &= hnch^{p-1} \\ &= \frac{b-a}{n} nch^{p-1} \\ &= (b-a)ch^{p-1} \end{aligned}$$

Ejercicio 11:

El método RK3 presenta un error de truncamiento global del orden de h^3.

Ejercicio 13:

Es más precisa la solución calculada con $h = 0{,}1$ ya que es posible estimar *directamente* el valor de $y(1) \approx 3{,}902080$. Al utilizar $h = 0{,}09$ se logra una estimación del valor de $y(0{,}98) \approx 3{,}849800$ siendo necesario algún método de interpolación para estimar $y(1)$.

Ejercicio 15:

Ambas expresiones, para la ecuación diferencial dada, pueden escribirse en forma iterativa como:

$$y_{n+1} = y_n \left(1 + h + \frac{h^2}{2} + \frac{h^3}{6}\right) + x_n \left(h + \frac{h^2}{2} + \frac{h^3}{6}\right) + \frac{h^2}{2} + \frac{h^3}{6}$$

Ejercicio 17:

Resolviendo de acuerdo a lo solicitado:

- Integración numérica, con el método del trapecio y 5 pasos: $-1{,}96634325387$.
- Resolviendo la EDO $y' = x\cos(x)$, con $y(2) = 0$ y el método de Heun con 5 pasos: $-1{,}96908030727$.
- Integración exacta: $3\sin(3) + \cos(3) - 2\sin(2) - \cos(2) \approx -1{,}96908048952$.

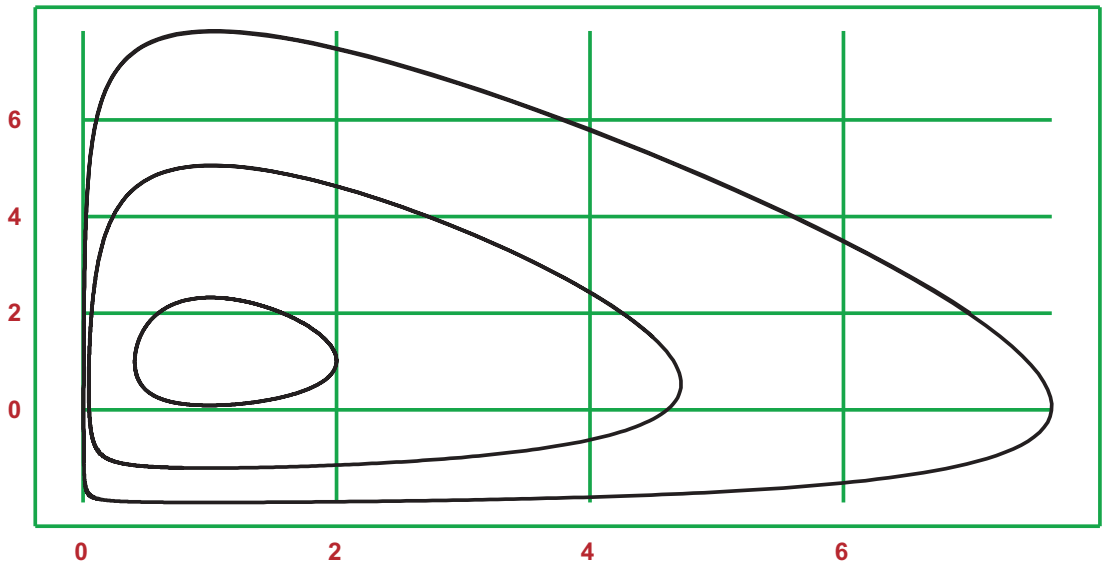

Figura 11.25: Ciclos límite del sistema del ejercicio 25.

Ejercicio 19:

Como $f(x_n, y_n) = f(x_{n+1}, y_{n+1})$, entonces todos los valores de y_{n+1} son iguales a cero. Sin embargo, aproximando $y(h) = (3h)^{(3/2)}$ al resolver la ecuación no lineal asociada, es posible utilizar cualquier método iterativo y $y(2) \approx 8$ lo que concuerda con la solución exacta.

Ejercicio 21:

Al utilizar el método de Euler explícito: $y_{n+1} = y_n(1-\lambda h)$, con lo que $|1-\lambda h| < 1$ para evitar crecimiento en las iteraciones. Entonces $h < \frac{2}{\lambda}$. El análisis de otros métodos es similar.

Ejercicio 23:

Utilizando los esquemas iterativos:

$$\begin{aligned}
\tilde{y_{n+1}} &= y_n + hf(z_n) \\
\tilde{z_{n+1}} &= z_n + hg(x_n, y_n, z_n) \\
y_{n+1} &= y_n + hf\left(\frac{z_n + \tilde{z_{n+1}}}{2}\right) \\
z_{n+1} &= z_n + hg\left(\frac{x_n + x_{n+1}}{2}, \frac{y_n + \tilde{y_{n+1}}}{2}, \frac{z_n + \tilde{z_{n+1}}}{2}\right),
\end{aligned}$$

donde $f(z) = z$ y $g(x, y, z) = 2z + 3y + \cos(x)$ con 20 pasos se obtiene $y(1) \approx 1{,}712878886$.

Ejercicio 25:

El gráfico con las soluciones se muestra como figura 11.25. A medida que decrecen los valores de las condiciones iniciales, se reduce la longitud del ciclo límite, siempre conservando el mismo centro.

Códigos para *Euler Math Toolbox*

Condición de función

El valor de condición de una función f en $x = a$ permite estimar su sensibilidad a las operaciones en un entorno de a. Gráficamente se observa que, en aquellos valores donde una función es mal condicionada, el valor de:

$$k = \left| \frac{xf'(x)}{f(x)} \right|$$

crece con gran pendiente.

```
function CondFunc(f$:string, a:real, b:real, n:natural=100)
## Condición de función
## * Parámetros de entrada:
##      f$: función de x, del tipo string
##      a: inicio del intervalo
##      b: fin del intervalo
##      n: cantidad de elementos en que se particionará el intervalo
## * Parámetro de salida:
##      vector con los números de condición de la función f en [a,b]
if a==b then
    X = a;
else
    X = linspace(a,b,n);
endif
y = zeros(1,length(X));
for i=1 to length(X)
    y[i] = abs(X[i]*diff(f$,X[i])/f$(X[i],args()));
end
return {y}
endfunction
```

Descomposición LU de Doolittle

La descomposición LU de Doolittle genera, para una matriz cuadrada $\mathbf{A}$, dos matrices $\mathbf{L}$, triangular inferior, y $\mathbf{U}$, triangular superior, tales que $\mathbf{A} = \mathbf{LU}$ y además $L_{ii} = 1$, $i = 1, 2, \ldots, n$.

```
function Lu(A)
## Descomposición LU [Doolittle]
## * Parámetro de entrada:
##      A: matriz cuadrada
## * Parámetros de salida:
```

```
##      L: matriz triangular inferior
##      U: matriz triangular superior
U=A;
m=size(U)[1]; n=size(U)[2]; L=eye(m);
for i=2 to m;
    if U[i,i-1]<>0 then;
        for j=i to n;
            L[j, i-1] = U[j,i-1]/U[i-1,i-1];
            U[j,:] = U[j,:] - U[i-1,:].(U[j,i-1]/U[i-1,i-1]);
        end;
    endif;
end;
return {L,U}
endfunction
```

Matriz de Pascal

La matriz de Pascal de orden n es una matriz simétrica y definida positiva con valores enteros, tomados del triángulo de Pascal:

$$P_{ij} = \frac{(i+j-2)!}{(i-1)!(j-1)!},$$

donde $i = 1, 2, \ldots, n$. Su inversa tiene valores enteros.

```
function Pascal(n:natural)
## Matriz de Pascal
## * Parámetro de entrada:
##      n: tamaño de la matriz a generar
## * Parámetro de salida:
##      matriz de Pascal de orden n
A=zeros(n,n);
for i=1 to n;
    for j=1 to n;
        A[i,j]=(i+j-2)!/((i-1)!*(j-1)!);
    end
end
return {A}
endfunction
```

Matriz de Hilbert

La matriz de Hilbert **H** es el ejemplo clásico de matriz mal condicionada. Los elementos se definen como:

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

donde $i, j = 1, 2, \ldots, n$.

```
function Hilbert(n:natural, d:natural)
## Matriz de Hilbert
## * Parámetros de entrada:
##      n: tamaño de la matriz a generar
##      d: cantidad de decimales en que se truncará cada elemento
```

```
## * Parámetro de salida:
##      matriz de Hilbert de orden n, con d decimales de mantisa
A=zeros(n,n);
for i=1 to n;
    for j=1 to n;
        A[i,j]=Trunc(1/(i+j-1),d);
    end
end
return {A}
endfunction
```

Método de la Potencia

A través del método de la potencia, y dados una matriz cuadrada y un vector semilla, es posible encontrar por medio de una iteración sencilla el autovalor dominante de la matriz.


```
function MetPot(A:real, q:column, n:natural, tol:positive)
## Método de la potencia
## * Parámetros de entrada:
##      A: matriz cuadrada
##      q: vector columna, semilla de las iteraciones
##      n: cantidad máxima de iteraciones
##      tol: tolerancia utilizada como stop del algoritmo
## * Parámetro de salida:
##      valor de la última iteración realizada
if size(A)[1]==size(A)[2] then
    dim=size(A)[1];
    i=1;
    error=1;
    repeat while i<=n and error>tol;
        L1=q'.A.q;
        z=A.q;
        q=z/norm(z);
        L=q'.A.q;
        error=abs(L-L1);
        i=i+1;
    end;
    if i==n+1 then
        "Algoritmo terminado por límite de iteraciones"
    endif
    if error<tol then
        "Algoritmo terminado por convergencia"
    endif
else
    "La matriz ingresada no es cuadrada."
endif
return {L}
endfunction
```

Método de Regula-Falsi

Dados una función f, un intervalo $[a;b]$ tal que $f(a)f(b) < 0$, una cantidad de iteraciones y una condición de *stop*, el algoritmo de Regula-Falsi:

$$x_{n+1} = \frac{x_{n'}f(x_n) - x_n f(x_{n'})}{f(x_n) - f(x_{n'})},$$

converge a la raíz de f dentro del intervalo bajo las condiciones iniciales dadas.

```
function RFalsi(f$:string, xa:real, xb:real, iter:natural, tol:
                                                  positive, d:natural)
## Método de Regula-Falsi
## * Parámetros de entrada:
##      f$: función de x, del tipo string
##      xa: inicio del intervalo
##      xb: fin del intervalo
##      iter: cantidad máxima de iteraciones a realizar
##      tol: tolerancia utilizada como stop del algoritmo
##      d: cantidad de decimales en que se truncará cada elemento
## * Parámetro de salida:
##      raíz aproximada de f en el intervalo [a,b]
if f$(xa,args())*f$(xb,args())>0 then
   "No se puede aplicar regula falsi, f(a) y f(b) tienen el mismo
                                                              signo"
else
i=1; error=1; y=zeros(1,iter); A=y; B=y;
repeat while i<iter && error>tol
    A[i]=xa; B[i]=xb;
    x = Trunc((xb*f$(xa,args())-xa*f$(xb,args()))/(f$(xa,args())-
                                                  f$(xb,args())),d);
    y[i]=x;
    if f$(xa,args())*f$(x,args())>0 then
        xa = x;
    else
        xb = x;
    endif
    if i>1 then
        error = abs(y[i-1]-y[i]);
    endif
    i = i+1;
end
return {x,y[1:i-1],A[1:i-1],B[1:i-1]}
endif
endfunction
```

Método de Sturn

Dada una matriz $\mathbf{A}$ de orden n, simétrica y tridiagonal, el algoritmo de Sturn genera $n+1$ polinomios P_i de grado creciente donde P_n es el polinomio característico de $\mathbf{A}$.

```
function Sturn(A)
## Método de Sturn
```

```
## * Parámetro de entrada:
##      A: matriz cuadrada, tridiagonal y simétrica
## * Parámetro de salida:
##      matriz con los coeficientes de los polinomios de la secuencia
                                                                de Sturn
n = length(A);
S = eye(n+1);
S[2,1:2] = [A[1,1],-1];
for i=3 to n+1
    Pol = polymult([A[i-1,i-1],-1], S[i-1,:]);
    Pol = polyadd(Pol, -(A[i-1,i-2])^2*S[i-2,:]);
    S[i,:] = Pol[1:n+1];
end
return {S}
endfunction
```

Norma matricial

La norma matricial es una función de valor real sobre el conjunto de matrices de orden n. A partir de ella es posible calcular el número de condición de una matriz.

```
function Norm(A:real, p=2)
## Normas matriciales
## * Parámetros de entrada:
##      A: matriz cuadrada
##      p=0 [norma infinito]: máxima suma absoluta de las filas de
                                                        la matriz
##      p=1 [norma uno]: máxima suma absoluta de las columnas de la
                                                        matriz
##      p=2 [norma dos]: utilizada por defecto
## * Parámetro de salida:
##      norma p de la matriz A
if p==1 then;
    return {max((sum(abs(A')))')}
elseif p==0 then;
    return {max(sum(abs(A))')}
elseif p==2 then;
    return {sqrt(max(abs(eigen(A'.A))))}
else
    "Los valores que acepta el parámetro p son: 0, 1 o 2"
endif
endfunction
```

Número de condición de matriz

El número de condición de una matriz permite saber si, al utilizarla con alguna aritmética restringida, será sensible a las operaciones tales como resolución de sistemas de ecuaciones lineales ó cálculo de inversa, por citar dos ejemplos.

```
function Cond(A:real, k=2)
## Número de condición
## * Parámetros de entrada:
```

```
##      A: matriz cuadrada
##      k: tipo de norma matricial a utilizar
## * Parámetro de salida:
##      número de condición, asociado a la norma k, de la matriz A
return {Norm(A,k)*Norm(inv(A),k)}
endfunction
```

Truncamiento a n dígitos

Truncamiento de los valores ingresados a una mantisa de n dígitos, útil para simular procesadores aritméticos limitados en potencia de almacenamiento.

```
function map Trunc(x:real, n:natural)
## Truncamiento de números reales
## * Parámetros de entrada:
##      x: elemento real o vector de reales
##      n: cantidad de decimales de mantisa
## * Parámetro de salida:
##      valor o valores de x truncados a n decimales
s=sign(x);
x=s*x;
j=0;
if x<10^n && x<>0 then
    repeat while x<10^n
        x=x*10;
        j=j+1;
    end
    x=10^(-j+1)*floor(x/10);
else
    repeat while x>10^n
        x=x/10;
        j=j+1;
    end
x=10^(j)*floor(x);
endif
return {s*x}
endfunction
```

Sebastián Hernández

Nací en Puerto San Julián en 1980 y a los pocos meses mi familia se trasladó a Río Gallegos, lugar en el que resido desde entonces. Poseo dos títulos, de grado (Prof. en Matemática) y de posgrado (Mg. en Informática y Sistemas), ambos obtenidos en UNPA. Mi área de docencia es Métodos Numéricos y Computación y actualmente estoy en proceso de inscripción para el Doctorado en la Universidad Nacional de San Luis.

Datos del autor

Agradecimientos

Nada de esto habría sido posible sin el apoyo, primero de mis padres y hermana (Rodrigo, Inés y Daniela) y luego de mi mujer (Paola).

Desde Nov 2016 estoy en la carrera más complicada y satisfactoria, soy el papá de Tomás..