PROBLEMAS DE COMPUTACIÓN NUMÉRICA

Fernando Sánchez Lasheras

Esperanza García Gonzalo

Problemas de Computación Numérica

Fernando Sánchez Lasheras
Departamento de Matemáticas. Universidad de Oviedo

Esperanza García Gonzalo
Departamento de Matemáticas. Universidad de Oviedo

A mi mujer Isabel, con todo mi amor.
Fernando

"La linea orizzontale
Ci spinge verso la materia
Quella verticale verso lo spirito"
Battiato, F. (2009). Inneres Auge. En Inneres Auge. Il tutto è più della somma delle sue parti [CD]. Milán, Italia: Universal Music.

Prólogo

El presente libro es fruto de la experiencia adquirida en la docencia del cálculo numérico. Gran parte de esta experiencia proviene del tiempo que llevamos impartiendo la asignatura de Computación Numérica en el grado en Ingeniería Informática del Software de la Escuela de Ingeniería Informática de la Universidad de Oviedo.

El objetivo de esta obra es ofrecer a los alumnos problemas relativos a los contenidos de la asignatura de Computación Numérica. Se trata de problemas que presentan una doble finalidad, por una parte, la de servir para afianzar los conocimientos teóricos y, por otra, permitir un primer contacto con la práctica del cálculo numérico, tan importante para la ciencia y la técnica contemporáneas.

La mayoría de los problemas que se recogen en este libro han sido propuestos en exámenes de la asignatura o se han inspirado, de una u otra forma, en los mismos. Si bien todos ellos han sido creados por nosotros, no podemos negar la influencia que en estas propuestas han tenido tanto la bibliografía clásica de análisis numérico como las ideas de otros compañeros del Departamento de Matemáticas de la Universidad de Oviedo con los que hemos compartido docencia y proyectos a lo largo de los años. Por tanto, siendo plenamente conscientes de esta deuda, no cabe más que hacer nuestra la frase que se atribuye a Juan de Salisbury: "Nos sumus sicut nanus positus super humeros gigantis".

Oviedo, octubre de 2019

Los autores

Índice

1. Aritmética finita y análisis de error **1**
- 1.1. Almacenamiento de enteros . . . 1
- 1.2. Almacenamiento de números en punto flotante . . . 6
- 1.3. Dígitos significativos . . . 32

2. Raíces de ecuaciones no lineales **35**
- 2.1. El método de bisección . . . 35
- 2.2. El método de *Régula-Falsi* . . . 44
- 2.3. El método de Newton . . . 48
- 2.4. El método de la secante . . . 53
- 2.5. El método de punto fijo . . . 56

3. Aproximación de funciones **67**
- 3.1. Interpolación de Lagrange . . . 67
- 3.2. Interpolación a trozos . . . 76
- 3.3. Aproximación lineal . . . 88
- 3.4. Aproximación con funciones polinómicas . . . 97
- 3.5. Aproximación con polinomios ortogonales . . . 102
- 3.6. Aproximación con funciones trigonométricas (Fourier) . . . 108

4. Derivación e integración numérica **111**
- 4.1. Derivación numérica: funciones de una variable . . . 111
- 4.2. Derivación numérica: funciones de dos variables . . . 120
- 4.3. Integración numérica: fórmulas de Newton-Cotes . . . 125
- 4.4. Integración numérica: fórmulas gaussianas . . . 140

5. Sistemas de ecuaciones lineales **147**
- 5.1. Métodos directos . . . 147
- 5.2. Métodos iterativos . . . 162

6. Optimización **177**
- 6.1. Condiciones necesarias y suficientes de óptimo . . . 177
- 6.2. El método de Newton . . . 179
- 6.3. El método del gradiente . . . 188
- 6.4. Optimización con restriciones . . . 191
- 6.5. Problemas lineales con restricciones lineales . . . 204

Tema 1

Aritmética finita y análisis de error

1.1. Almacenamiento de enteros

Problema 1.1:

Disponemos de 12 bits para almacenar números enteros en binario:

(a) Si lo usamos para representar sólo números positivos, ¿cómo representaríamos el número 22 dado en base 10 en este sistema de 12 bits?

(b) Si lo usamos para representar enteros con signo con representación sesgada, ¿cuál es el mayor entero que podemos representar? ¿Y el mínimo?

(a) Dividimos 22 entre 2 y almacenamos el cociente y el resto. Repetimos el proceso con los sucesivos cocientes obtenidos hasta que el cociente sea 1.

Cociente	22	11	5	2	1
Resto	0	1	1	0	↙

Para construir el número en binario, empezamos con el 1 del cociente y seguimos con los ceros y unos del resto y el número será 10110. Rellenando a la izquierda con ceros tenemos

Solución:

$$\boxed{(22)_{10} = (000000010110)_2}$$

(b) El número de enteros con signo que podemos representar con m bits sería

$$2^m = 2^{12} = 4096$$

y como empezaríamos con el 0 acabaríamos con 4095. Pero los enteros representados serían los anteriores menos el sesgo

$$\text{sesgo} = 2^{m-1} = 2^{11} = 2048,$$

es decir, el rango de números a representar sería

$$[0 - 2048, \ldots, 4095 - 2048] = [-2048, \ldots, 2047]$$

y por lo tanto el valor mínimo y máximo son

Solución:

$$\boxed{min = -2048, \qquad max = 2047}$$

Problema 1.2:

Si disponemos de 8 bits para almacenar números enteros en binario:

(a) Si lo usamos para representar sólo números positivos ¿cómo representaríamos 18 en base 10 en este sistema de 8 bits?

(b) Si lo usamos para representar enteros con signo con representación sesgada ¿cuál es el mayor entero que podemos representar? ¿Y el mínimo?

(a) Dividimos 18 entre 2 y almacenamos el cociente y el resto. Repetimos el proceso con los sucesivos cocientes obtenidos hasta que el cociente sea 1.

Cociente	18	9	4	2	1
Resto	0	1	0	0	↙

Para construir el número en binario, empezamos con el 1 del cociente y seguimos con los ceros y unos del resto y el número será 10010. Rellenando a la izquierda con ceros tenemos

Solución:

$$\boxed{(18)_{10} = (00010010)_2}$$

(b) El número de enteros con signo que podemos representar con m bits serían

$$2^m = 2^8 = 256$$

y como empiezan en 0 acabarían en 255. Pero los enteros representados serían los anteriores menos el sesgo

$$\text{sesgo} = 2^{m-1} = 2^7 = 128,$$

es decir, el rango de números a representar sería

$$[0 - 128, \ldots, 255 - 128] = [-128, \ldots, 127]$$

y por lo tanto el valor mínimo y máximo son

Solución:

$$\boxed{min = -128, \qquad max = 127}$$

Problema 1.3:

En una máquina controlada numéricamente, los enteros no negativos necesitan ser almacenados en un espacio de memoria:

(a) ¿Cuál es el número mínimo de bits necesarios para representar todos los enteros entre 0 y 1300?

(b) ¿Cuál es el máximo entero no negativo que se podría representar con estos bits?

(c) Para el mismo número de bits, si fueran enteros con signo y se usara la representación sesgada, ¿cuál sería el mayor positivo representable?

(d) ¿Cómo se representaría entonces -1000?

(a) Con m bits representamos 2^m enteros. Como empezamos representado el 0 podemos representar enteros positivos dentro del rango $[0, (2^m - 1)_{10}]$. Si tomamos $m = 10$

$$1300 \not< (2^m - 1) = 1024 - 1 = 1023$$

que no nos vale. Pero si tomamos $m = 11$

$$1300 < (2^m - 1) = 2048 - 1 = 2047$$

Solución:

$$\boxed{\text{11 bits}}$$

(b) Con m bits representamos enteros en $[0, (2^m - 1)_{10}]$. Por lo tanto

$$[0, (2^m - 1)_{10}] = [0, (2^{11} - 1)_{10}] = [0, 2047],$$

Solución:

$$\boxed{2047}$$

(c) El número de enteros con signo que podemos representar con m bits serían los mismos que el número de enteros sin signo pero los enteros representados serían los anteriores menos el sesgo

$$\text{sesgo} = 2^{m-1} = 2^{10} = 1024,$$

es decir, el rango de números a representar sería

$$[0 - 1024, 2047 - 1024] = [-1024, 1023]$$

y por lo tanto el valor máximo es

Solución:

$\boxed{1023}$

(d) Para obtener el número a almacenar en binario tendríamos que empezar sumándole el sesgo

$$-1000 + \text{sesgo} = -1000 + 1024 = 24$$

Y ahora lo convertimos a binario dividiendo sucesivamente entre dos y nos quedamos con el último cociente, que tiene que ser 1, y todos los restos empezando por el último obtenido.

Cociente	24	12	6	3	1
Resto	0	0	0	1	↙

Y el número en binario es 11000. Completando con ceros a la izquierda

Solución:

$\boxed{(00000011000)_2}$

Problema 1.4:

Un ingeniero que trabaja en el Ministerio de Defensa está escribiendo un programa que transforma números reales no negativos al formato entero. Para evitar problemas de overflow:

(a) ¿Cuál es el máximo entero no negativo que se puede representar con un entero de 12 bits?

(b) Y si fueran enteros con signo y se usara la representación sesgada, ¿cuál sería el mayor positivo representable?

(c) ¿Cómo se representaría entonces -2000?

(a) Con m bits representamos 2^m enteros. Como empezamos representado el 0 podemos representar enteros positivos dentro del rango $[0, (2^m - 1)_{10}]$. Si tomamos $m = 12$

$$2^m - 1 = 2^{12} - 1 = 4096 - 1 = 4095$$

Solución:

$\boxed{4095}$

(b) El número de enteros con signo que podemos representar con m bits serían los mismos que el número de enteros sin signo. Pero los enteros representados serían los anteriores menos el sesgo

$$\text{sesgo} = 2^{m-1} = 2^{11} = 2048,$$

es decir, el rango de números a representar sería

$$[0 - 2048, 4095 - 2048] = [-2048, 2047]$$

y por lo tanto el valor máximo es

Solución:

$\boxed{2047}$

(c) Para representar el número, tendríamos que empezar sumándole el sesgo

$$-2000 + \text{sesgo} = -2000 + 2048 = 48$$

Convertimos a binario dividiendo sucesivamente entre dos y nos quedamos con el último cociente, que tiene que ser 1, y todos los restos empezando por el último obtenido:

Cociente	48	24	12	6	3	1
Resto	0	0	0	0	1	↙

y el número en binario es 110000. Rellenando con ceros a la izquierda

Solución:

$\boxed{(000000110000)_2}$

Problema 1.5:

Representamos un entero utilizando 7 bits:

(a) ¿Cuál es el máximo entero no negativo que se puede representar?

(b) Y si fueran enteros con signo y usáramos la representación sesgada, ¿cuál sería el mayor positivo representable?

(a) Con m bits representamos enteros en el rango $[0, (2^m - 1)_{10}]$. Por lo tanto

$$[0, (2^m - 1)_{10}] = [0, (2^7 - 1)_{10}] = [0, 127],$$

y el máximo entero no negativo es

Solución:

$\boxed{127}$

(b) Los enteros con signo que podemos representar con m bits serían los mismos que en el caso anterior menos el sesgo

$$sesgo = 2^{m-1} = 2^6 = 64,$$

es decir

$$[0 - 64, 127 - 64] = [-64, 63]$$

y por lo tanto el valor máximo es

Solución:

$\boxed{63}$

1.2. Almacenamiento de números en punto flotante

Problema 1.6:

Si el número binario

signo	1 bit	0
exponente	8 bits	10001111
mantisa	23 bits	11101100000000000000000

sigue la norma IEEE 754 para representación en punto flotante con precisión simple, calcular su representación en base 10.

(a) EXPONENTE

Teniendo en cuenta la posición de los unos dentro del exponente, el valor nominal del exponente es $2^7 + 2^3 + 2^2 + 2^1 + 2^0 = 143$. Si el número de bits del exponente es $m = 8$, entonces el sesgo es

$$\text{sesgo} = 2^{m-1} - 1 = 127$$

y el exponente es

$$142 - \text{sesgo} = 143 - 127 = 16$$

(b) MANTISA

Teniendo en cuenta el bit escondido, la mantisa es (**1**),111011 que en base 10 es

$$\mathbf{1} + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-5} + 2^{-6} = 1{,}921875$$

(c) NÚMERO

Como el bit del signo es 0, el número es positivo. Teniendo en cuenta la mantisa y el exponente, el número en base 10 es

$$+(\mathbf{1} + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-5} + 2^{-6}) \times 2^{16} = +125952$$

y el número representado en base 10 es

Solución:

$$\boxed{125952}$$

Problema 1.7:

¿Cómo se almacenaría en precisión simple según la norma IEEE 754 el número $-107,90625$?

En precisión simple tenemos 32 bits (4 bytes) en total: 1 bit para el signo, 8 bits para el exponente y 23 para la mantisa.

(a) MANTISA

Para convertir a binario la parte fraccionaria, la multiplicamos por 2, le quitamos la parte entera, que será nuestro dígito binario y repetimos el proceso.

Parte fraccionaria :	0,90625	0,8125	0,625	0,25	0,5	0
Parte entera :		1	1	1	0	1

y tomamos los dígitos de la parte entera en el orden que se generan

$$0{,}11101$$

Para convertir a binario la parte entera 118 dividimos primero el número y luego los sucesivos cocientes de forma reiterada por 2 y guardamos los restos

cociente :	107	53	26	13	6	3	1
resto :	1	1	0	1	0	1	↙

empezando por el último cociente, que tiene que ser 1, seguimos con los restos en orden inverso y la parte entera de la mantisa es

$$1101011$$

Juntamos la parte entera y la fraccionaria en binario y para almacenar según la norma IEEE, normalizamos

$$1101011{,}11101 \rightarrow (\mathbf{1}),10101111101 \times 2^6$$

y no hace falta almacenar el primer uno que será el bit escondido. Rellenamos con ceros a la derecha hasta 23 bits y la mantisa se almacena como

$$10101111101000000000000$$

(b) EXPONENTE

Si el número de bits del exponente es $m = 8$, entonces el sesgo $= 2^{m-1} - 1 = 127$. El exponente sesgado será

$$\text{exponente} + \text{sesgo} = 6 + 127 = 133 = 2^7 + 2^2 + 2^0,$$

que en binario es 100000101.

(c) NÚMERO

Como el número es negativo, el signo es 1

signo	exponente	mantisa
1	10000101	10101111101000000000000

Problema 1.8:

Si utilizamos precisión simple según la norma IEEE 754:

(a) ¿Cómo se almacenaría el número $-34{,}875$?

(b) ¿Cual es el ϵ de máquina?

(c) ¿Y el mayor entero positivo que se puede almacenar de forma exacta de forma que todos los enteros menores se pueden almacenar de forma exacta también?

En precisión simple tenemos 32 bits (4 bytes) en total:

- 1 bit para el signo,
- 8 para el exponente,
- 23 para la mantisa.

(a) 1. MANTISA

Primero la parte entera: dividimos de forma reiterada por 2 primero el número y luego los sucesivos cocientes y guardamos los cocientes y los restos:

Cociente :	34	17	8	4	2	1
Resto :	0	1	0	0	0	↙

y tomamos los dígitos en orden inverso, empezando por el último cociente, que es necesariamente 1:

$$100010$$

Comprobamos que es correcto:

$$1 \times 2^5 + 0 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 0 \times 2^0 = 34$$

Para convertir a binario la parte fraccionaria, multiplicamos por 2, y le quitamos la parte entera, que será nuestro dígito binario y repetimos el proceso.

Parte fraccionaria :	0,875	0,75	0,5
Parte entera :	1	1	1

y tomamos los dígitos:

$$0{,}111$$

Comprobamos que es correcto:

$$1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} = 0{,}875$$

Juntando la parte entera y la fraccionaria, tenemos el número completo. Para almacenarlo según la norma IEEE, primero lo normalizamos:

$$100010{,}111 = \mathbf{1}{,}00010111 \times 2^5$$

No hace falta almacenar el primer uno (técnica del bit escondido). Rellenamos a la derecha con ceros, hasta 23 bits. Por lo tanto la mantisa se almacena como

$$00010111000000000000000$$

2. EXPONENTE
   Como hay $m = 8$ bits para el exponente el sesgo es

$$\text{sesgo} = 2^{m-1} - 1 = 2^{8-1} - 1 = 127.$$

   El exponente en base 10 es $127 + 5 = 132$. Si lo pasamos a base 2:

$$\begin{array}{rcccccccc} \text{Cociente}: & 132 & 66 & 33 & 16 & 8 & 4 & 2 & 1 \\ \text{Resto}: & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \swarrow \end{array}$$

   y comenzado por el último cociente, uno, y siguiendo con los restos en orden inverso, el exponente es

$$10000100$$

   Comprobamos que está bien

$$2^2 + 2^7 = 132$$

3. NÚMERO
   Como el número es negativo, el bit del signo sería 1

signo	exponente	mantisa
1	10000100	00010111000000000000000

(b) El ϵ de máquina es el número más pequeño que se le puede sumar a 1. Como 1 normalizado en este estándar es

$$1.\overbrace{00000000000000000000000}^{23} \times 2^0$$

el número más pequeño que se le puede sumar es

$$0.\overbrace{00000000000000000000001}^{23} \times 2^0 = 2^{-23} = 0{,}000000119209$$

Solución:

$$\boxed{2^{-23} = 0{,}000000119209}$$

(c) Como disponemos de 23 bits más el bit escondido (en negrita), hasta el número

$$\mathbf{1}\overbrace{11111111111111111111111}^{23}$$

que normalizado es

$$\mathbf{1}.\overbrace{11111111111111111111111}^{23} \times 2^{23}$$

podemos almacenar todos los dígitos de un número entero y por tanto representarlo de forma exacta. El número siguiente es

$$\mathbf{1}\overbrace{11111111111111111111111}^{23} + 1 = \mathbf{1}\overbrace{00000000000000000000000}^{23}0$$

que normalizado es

$$\mathbf{1}.\overbrace{00000000000000000000000}^{23} \times 2^{24} = 1677216$$

y aunque no hemos podido almacenar su último dígito, como es un cero, no cometemos error. Sin embargo, el número entero siguiente es

$$\mathbf{1}\overbrace{00000000000000000000000}^{23}1$$

que normalizado es, otra vez

$$\mathbf{1}.\overbrace{00000000000000000000000}^{23} \times 2^{24} = 1677216$$

porque no podemos almacenar el último dígito, que en este caso es un 1, y redondeamos al número anterior, por lo tanto estamos cometiendo un error. Así que

Solución:

$$\boxed{2^{24} = 1677216}$$

Problema 1.9:

Redondear al par más cercano los siguientes números en base 2:

(a) Si la precisión es 4, $n_1 = 0{,}110010$, $n_2 = 1{,}111100$, $n_3 = 1{,}010110$, $n_4 = 1{,}010010$, $n_5 = 1,111111$, $n_6 = 1,111001$, $n_7 = 1,010100$, $n_8 = 1,010101$, $n_9 = 1,010001$ y $n_{10} = 1,011100$.

(b) Si la precisión es 3, $n_1 = 1,111000$, $n_2 = 1,101000$, $n_3 = 1,111001$, $n_4 = 1{,}110001$.

(a) $n_1 = 0{,}110010 \longrightarrow 0{,}1100$, puesto que n_1 es equidistante del truncado 0,1100 y el siguiente 0,1101 (con cuatro dígitos significativos) y redondea al par más cercano (el que acaba en cero).

$n_2 = 1,111100 \longrightarrow 10,00$ puesto que n_2 es equidistante del truncado $1,111$ y el siguiente 10,00 (con cuatro dígitos significativos) y redondea al par más cercano (el que acaba en cero).

$n_3 = 1,010110 \longrightarrow 1,011$ puesto n_3 truncado es $1,010$ y el siguiente, con cuatro dígitos significativos, es 1,011 y redondea al más cercano que es el segundo.

$n_4 = 1,010010 \longrightarrow 1,010$ puesto n_4 truncado es $1,010$ y el siguiente, con cuatro dígitos significativos, es $1,011$ y redondea al más cercano, que es el primero.

$n_5 = 1,111111 \longrightarrow 10,00$ puesto n_5 truncado es $1,111$ y el siguiente, con cuatro dígitos significativos, es $10,00$ y redondea al más cercano que es el segundo.

$n_6 = 1,111001 \longrightarrow 1,111$ puesto n_6 truncado es $1,111$ y el siguiente, con cuatro

dígitos significativos, es $10,00$ y redondea al más cercano que es el primero.

$n_7 = 1,010100 \longrightarrow 1,010$, puesto que n_7 es equidistante del truncado $1,010$ y el siguiente $1,011$ (con cuatro dígitos significativos) y redondea al par más cercano (el que acaba en cero).

$n_8 = 1,010101 \longrightarrow$ 1,011 puesto n_8 truncado es $1,010$ y el siguiente, con cuatro dígitos significativos, es 1,011 y redondea al más cercano que es el segundo.

$n_9 = 1,010001 \longrightarrow 1,010$ puesto n_9 truncado es $1,010$ y el siguiente, con cuatro dígitos significativos, es $1,011$ y redondea al más cercano que es el primero.

$n_{10} = 1,011100 \longrightarrow 1,100$ puesto que n_{10} es equidistante del truncado $1,011$ y el siguiente $1,100$ (con cuatro dígitos significativos) y redondea al par más cercano (el que acaba en cero).

(b) $n_1 = 1,111000 \longrightarrow 10,0$ puesto que n_1 es equidistante del truncado $1,11$ y el siguiente 10,0 (con tres dígitos significativos) y redondea al par más cercano (el que acaba en cero).

$n_2 = 1,101000 \longrightarrow 1,10$, puesto que n_2 es equidistante del truncado $1,10$ y el siguiente $1,11$ (con tres dígitos significativos) y redondea al par más cercano (el que acaba en cero).

$n_3 = 1,111001 \longrightarrow$ 10,0 puesto n_3 truncado es $1,11$ y el siguiente, con tres dígitos significativos, es 10,0 y redondea al más cercano que es el segundo.

$n_4 =$ 1,110001 $\longrightarrow 1,11$ puesto n_4 truncado es $1,11$ y el siguiente, con tres dígitos significativos, es $10,0$ y redondea al más cercano que es el primero.

Problema 1.10:

Si el número

signo	1 bit	`1`
exponente	5 bits	`10001`
mantisa	10 bits	`0110100000`

sigue la norma IEEE 754 para representación en punto flotante con 16 bits llamada de media precisión:

(a) Calcular su representación en base 10.

(b) ¿Cual sería el ϵ de máquina expresado en base 10?

(c) ¿Cuál es el mayor entero positivo que se puede almacenar de forma exacta de manera que todos los enteros positivos anteriores se pueden representar de forma exacta también? Escribirlo en decimal y en binario.

(d) ¿Cuál sería la representación de 0, $+\infty$, $-\infty$? Da un ejemplo de representación de `NaN`.

(e) Representar 0,2 en media precisión. Dar el error absoluto en base 10.

(a) El número de bits del exponente es $m = 5$, y por tanto el sesgo es

$$2^{m-1} - 1 = 2^4 - 1 = 15.$$

El valor nominal del exponente es

$$(10001)_2 \to 2^4 + 2^0 = 17,$$

y si tenemos en cuenta el sesgo el exponente es $17 - \text{sesgo} = 17 - 15 = 2$. Para calcular la mantisa hemos de tener en cuenta el bit escondido

$$\mathbf{1},01101 \times 2^2 = \left(1 + 2^{-2} + 2^{-3} + 2^{-5}\right) \times 2^2 = 5{,}625.$$

Finalmente, como el bit del signo es 1, el número es negativo.

Solución:

$$\boxed{-5{,}625}$$

(b) El número de dígitos de la mantisa es 10. El número 1 se representa

$$1 \;=\; 1{,}0000000000 \times 2^0$$

y el siguiente número representable es

$$1 + \epsilon \;=\; 1{,}0000000001 \times 2^0$$

Por lo que

$$\epsilon \;=\; 0{,}0000000001 \times 2^0 \;=\; 2^{-10}$$

Y en base 10,

Solución:

$$\boxed{9{,}77 \times 10^{-4}}$$

(c) Como disponemos de 10 bits más el bit escondido (en negrita), hasta el número

$$\mathbf{1}\overbrace{1111111111}^{10}$$

que normalizado es

$$\mathbf{1}.\overbrace{1111111111}^{10} \times 2^{10}$$

podemos almacenar todos los dígitos de un número entero y por tanto representarlo de forma exacta. El número siguiente es

$$\mathbf{1}\overbrace{1111111111}^{10} + 1 = \mathbf{1}\overbrace{0000000000}^{10}0$$

que normalizado es

$$\mathbf{1}.\overbrace{0000000000}^{10} \times 2^{11}$$

y aunque no hemos podido almacenar su último dígito, como es un cero, no cometemos error. Sin embargo, el número entero siguiente es

$$\mathbf{1}\overbrace{0000000000}^{10}1$$

que normalizado es, otra vez

$$\mathbf{1}.\overbrace{0000000000}^{10}\times 2^{11} = 2048$$

porque no podemos almacenar el último dígito, que en este caso es un 1, y redondeamos al número anterior, por lo tanto estamos cometiendo un error. El resultado es

Solución:

$$\boxed{2^{11} = (2048)_{10} = (100000000000)_2}$$

(d) Podemos representar cero

	signo	exponente	mantisa
cero	0	00000	0000000000

	signo	exponente	mantisa
$+\infty$	0	11111	0000000000

	signo	exponente	mantisa
$-\infty$	1	11111	0000000000

	signo	exponente	mantisa
NaN	0	11111	0001000000

(e) MANTISA: Convertimos a binario:

Parte fraccionaria	0,2	0,4	0,8	0,6	0,2	$\cdots$
Parte entera	0	0	1	1	0	$\cdots$

Y tenemos que $(0{,}2)_{10} = (0{,}0011\ 0011\ 0011\ldots)_2$, que normalizado es

$$(0{,}2)_{10} = \mathbf{1}.\overbrace{100\ 1100\ 110}^{10}0\ 1100\ldots \times 2^{-3}$$

y redondeado al par más cercano (está más cerca del truncado que del número siguiente en este formato)

$$(0{,}2)_{10} \approx \mathbf{1}.\overbrace{100\ 1100\ 110}^{10}\times 2^{-3}$$

y ya tenemos la mantisa si quitamos el bit escondido (en negrita).

EXPONENTE: El sesgo es 15, y $e = -3 + sesgo = -3 + 15 = 12$. Pasándolo a binario

Cociente	12	6	3	1
Resto	0	0	1	↙

Y el exponente es, 1100. Completaríamos a 5 bits con un cero a la izquierda.

SIGNO: Como es positivo, el bit del signo es cero.

signo	exponente	mantisa
0	01100	1001100110

El número redondeado es

$$x^* = \left(1 + 2^{-1} + 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9}\right) \times 2^{-3} \approx 0{,}199951$$

Solución:

$$\boxed{\text{error} = |0{,}2 - x^*| \approx 4{,}9 \times 10^{-5}}$$

Problema 1.11:

Una máquina almacena números en punto flotante en base 2 en 8 bits, siguiendo un criterio similar al de la norma IEEE 754. El primer bit se usa para el signo del número, los cinco siguientes para el exponente sesgado y los últimos dos bits para la mantisa.

(a) Calcular los exponentes máximo y mínimo y dar su valor en base 10.

(b) Escribir todos los números desnormalizados positivos en este sistema, en binario (siguiendo la norma) y en decimal.

(c) Calcular en base 10 el valor del número representado en este formato como:

signo	exponente	mantisa
1	11110	00

(a) El número de exponentes que podemos representar con m bits sería

$$2^m = 2^5 = 32$$

que son, en binario

$$[00000, 00001, 00010, \ldots, 11101, 11110, 11111]$$

Si no tenemos en cuenta el signo, el valor nominal en base 10 es

$$[0, 1, \ldots, 30, 31]$$

y teniendo en cuenta que el primero y el último están reservados

$$[R, 1, \ldots, 30, R]$$

Pero los enteros representados serían los anteriores menos el sesgo, donde

$$\text{sesgo} = 2^{m-1} - 1 = 2^{5-1} - 1 = 2^4 - 1 = 15$$

es decir, el rango de números a representar sería

$$[R, 1 - 15, \ldots, 30 - 15, R] = [R, -14, \ldots, 15, R]$$

por lo tanto los exponentes mínimo y máximo en base 10 son

Solución:

$$\boxed{e_{min} = -14 \text{ y } e_{max} = 15}$$

(b) Los números desnormalizados se caracterizan por el valor nominal de su exponente, que es cero y

- El valor asignado al exponente es el mínimo.
- En la mantisa, el bit escondido es cero.

Con este estándar los números desnormalizados son

Num(bin)		Num(dec)
0 00000 01	$0{,}01 \times 2^{-14}$	$1{,}53 \times 10^{-5}$
0 00000 10	$0{,}10 \times 2^{-14}$	$3{,}05 \times 10^{-5}$
0 00000 11	$0{,}11 \times 2^{-14}$	$4{,}56 \times 10^{-5}$

El valor mínimo se corresponde con el valor

$$0{,}01 \times 2^{-14} \longrightarrow 2^{-2} \times 2^{-14} = 2^{-16} \approx 1{,}53 \times 10^{-5}$$

El valor intermedio se corresponde con el valor

$$0{,}10 \times 2^{-14} \longrightarrow 2^{-1} \times 2^{-14} = 2^{-16} \approx 3{,}05 \times 10^{-5}$$

El máximo se corresponde con el valor

$$0{,}11 \times 2^{-14} \longrightarrow \left(2^{-1} + 2^{-2}\right) \times 2^{-14} \approx 4{,}56 \times 10^{-5}$$

(c) El número representado en este estandar como

signo	exponente	mantisa
1	11110	00

tiene exponente máximo (el número siguiente, 11111, está reservado), mantisa mínima y signo negativo. Por lo tanto, el número es

$$-1{,}00 \times 2^{15}$$

que expresado en decimal es

$$-2^{15} = -32768$$

Solución:

$$\boxed{-32768}$$

Problema 1.12:

Una máquina almacena números en punto flotante en base 2 en 10 bits, siguiendo un criterio similar al de la norma IEEE 754. El primer bit se usa para el signo del número, los seis siguientes para el exponente sesgado y los últimos tres bits para la mantisa.

(a) Calcular los exponentes máximo y mínimo y dar su valor en base 10.

(b) Escribir todos los números desnormalizados positivos en este sistema en binario (siguiendo la norma). Dar el valor en decimal de los números máximo y mínimo desnormalizados.

(c) Calcular en base 10 el valor del número representado en este formato:

signo	exponente	mantisa
1	111110	000

(a) El número de exponentes que podemos representar con m bits sería

$$2^m = 2^6 = 64$$

que son, en binario

$$[000000, 000001, 000010, \ldots, 111101, 111110, 111111]$$

Si no tenemos en cuenta el signo, el valor nominal es

$$[0, 1, \ldots, 62, 63]$$

y teniendo en cuenta que el primero y el último están reservados

$$[R, 1, \ldots, 62, R]$$

Pero los enteros representados serían los anteriores menos el sesgo, siendo

$$\text{sesgo} = 2^{m-1} - 1 = 2^{6-1} - 1 = 2^5 - 1 = 31,$$

es decir, el rango de números a representar sería

$$[R, 1 - 31, \ldots, 62 - 31, R] = [R, -30, \ldots, 31, R]$$

Por lo tanto, los exponentes máximo y mínimo, para esta norma, en base 10 son

Solución:

$$\boxed{e_{min} = -30 \text{ y } e_{max} = 31}$$

(b) Los números desnormalizados se caracterizan por el valor nominal de su exponente, que es cero y

- El valor asignado al exponente es el mínimo.
- En la mantisa, el bit escondido es cero.

Con esta norma los números desnormalizados son

Num(bin)		Num(dec)
0 000000 001	$0{,}001 \times 2^{-30}$	$1{,}16 \times 10^{-10}$
0 000000 010	$0{,}010 \times 2^{-30}$	
0 000000 011	$0{,}011 \times 2^{-30}$	
0 000000 100	$0{,}100 \times 2^{-30}$	
0 000000 101	$0{,}101 \times 2^{-30}$	
0 000000 110	$0{,}110 \times 2^{-30}$	
0 000000 111	$0{,}111 \times 2^{-30}$	$8{,}15 \times 10^{-10}$

El mínimo se corresponde con el valor

$$0{,}001 \times 2^{-30} \longrightarrow 2^{-3} \times 2^{-30} = 2^{-33} \approx 1{,}16 \times 10^{-10}$$

El máximo se corresponde con el valor

$$0{,}111 \times 2^{-30} \longrightarrow \left(2^{-1} + 2^{-2} + 2^{-3}\right) \times 2^{-30} \approx 8{,}15 \times 10^{-10}$$

Solución:

Mínimo número desnormalizado: $1{,}16 \times 10^{-10}$
Máximo número desnormalizado: $8{,}15 \times 10^{-10}$

(c) El número representado en este estandar como

signo	exponente	mantisa
1	111110	000

tiene exponente máximo (el número siguiente, 111111, está reservado), mantisa mínima y signo negativo. Por lo tanto, el número es

$$-1{,}000 \times 2^{31}$$

que expresado en decimal es

$$-2^{31} = -2147483648$$

Solución:

$$\boxed{-2147483648}$$

Problema 1.13:

Una máquina almacena números en punto flotante en base 2 en 19 bits, siguiendo un criterio similar al de la norma IEEE 754. El primer bit se usa para el signo del número, los once siguientes para el exponente sesgado y los últimos siete bits para la mantisa.

(a) ¿Cual es el valor (en decimal) de los exponentes máximo y mínimo en este formato?

(b) Convertir a decimal el número almacenado en este formato

signo	exponente	mantisa
1	0 00000 00000	1000 000

(c) Calcular el número 18,7 en este formato. Redondear al par más cercano.

(d) Si solo dispusiéramos de 5 bits para la mantisa, en lugar de 7, como en el caso anterior (exponente y signo, con el mismo número de bits que el caso anterior) ¿cómo lo representaríamos si redondeamos al par más cercano?

(a) El número de exponentes que podemos representar con m bits sería

$$2^m = 2^{11} = 2048$$

Si no tenemos en cuenta el signo los valores nominales serían

$$[0, 1, \ldots, 2046, 2047]$$

y teniendo en cuenta que el primero y el último están reservados

$$[R, 1, \ldots, 2046, R]$$

Pero los enteros representados serían los anteriores menos el sesgo, donde

$$\text{sesgo} = 2^{m-1} - 1 = 2^{11-1} - 1 = 2^{10} - 1 = 1023,$$

es decir, el rango de números a representar sería

$$[R, 1 - 1023, \ldots, 2046 - 1023, R] = [R, -1022, \ldots, 1023, R]$$

por lo tanto

Solución:

$$\boxed{e_{min} = -1022 \text{ y } e_{max} = 1023}$$

(b) Dado el número

signo	exponente	mantisa
1	0 00000 00000	1000 000

como todos los elementos del exponente son cero, el número es desnormalizado y se cumple que el bit escondido va a ser cero y el exponente será el mínimo. Por lo tanto el número es

$$-0{,}1000000 \times 2^{-1022} \rightarrow -2^{-1} \times 2^{-1022} = -2^{-1023} = -1{,}11 \times 10^{-308}$$

y por lo tanto

Solución:

$$\boxed{-1{,}11 \times 10^{-308}}$$

(c) Calcular el número 18,7 en este formato.

MANTISA:
Dividiendo el número sucesivamente por dos y almacenando el cociente y el resto de cada división

Cociente	18	9	4	2	1
Resto	0	1	0	0	↙

Empezando por el último cociente y siguiendo con los restos, la parte entera es

$$(18)_{10} = (10010)_2$$

La parte fracionaria se obtiene multiplicando en cada paso estapor dos y separándola en entera y fraccionaria y volviendo a multiplicar sólo la parte fraccionaria por dos. La representación vendrá dada por los dígitos correspondientes a las partes enteras en el orden en el que se generan

Parte fraccionaria	0,7	0,4	**0,8**	**0,6**	**0,2**	**0,4**	0,8	0,6	0,2	0,4	$\cdots$
Parte entera		1	**0**	**1**	**1**	**0**	0	1	1	0	$\cdots$

y el número es periódico

$$(0{,}7)_{10} = (0{,}1\ \mathbf{0110}\ 0110\ \mathbf{0110}\ \ldots)_2$$

Y el número completo es

$$(18{,}7)_{10} = (10010{,}1\ \mathbf{0110}\ 0110\ \mathbf{0110}\ \ldots)_2$$

Si lo normalizamos

$$\mathbf{1}{,}0010\ 1\ 01\mathbf{10}\ \mathbf{0110}\ \ldots \times 2^4$$

Ahora, en negrita, el bit escondido y los números que no caben en la mantisa pero que tendremos que tener en cuenta al redondearla. Cómo el número está más próximo al valor siguiente representable que al truncado, redondeamos al siguiente, y el número a almacenar es

$$\mathbf{1}{,}0010\ 1\ 10 \times 2^4$$

donde la mantisa a almacenar sería 0010110, sin el bit escondido, y el exponente es 4.

EXPONENTE:
Como sesgo = 1023 representaremos el exponente por el número binario que se corresponde con

$$4 + 1023 = 1027 = 1024 + 3 = 1024 + 2 + 1 = 2^{10} + 2^1 + 2^0 \longrightarrow (1\,00000\,00011)_2$$

También podemos obtenerlo

Cociente	1027	513	256	128	64	32	16	8	4	2	1
Resto	1	1	0	0	0	0	0	0	0	0	↙

y el exponente en binario es $(1\,00000\,00011)_2$ y, teniendo en cuenta que el número es positivo (bit del signo cero), la representación completa será

signo	exponente	mantisa
0	1 00000 00011	0010 110

(d) El número es el mismo pero ahora tenemos dos bits menos para la mantisa

$$(18{,}7)_{10} = (10010{,}1\ 0110\ 0110\ 0110\ \ldots)_2 \longrightarrow \mathbf{1}{,}0010\ 1\ \mathbf{0110\ 0110}\ \ldots \times 2^4$$

En negrita, el bit escondido y los números que no caben en la mantisa pero que tendremos que tener en cuenta al redondearla. Cómo el número está más próximo al valor truncado que al siguiente, redondeamos al truncado, y el número a almacenar es

$$\mathbf{1}{,}0010\,1 \times 2^4$$

Como el exponente no cambia respecto al anterior, el número en este formato es

signo	exponente	mantisa
0	1 00000 00011	0010 1

Problema 1.14:

Una máquina almacena números en punto flotante en base 2 en 16 bits, siguiendo un criterio similar al de la norma IEEE 754. El primer bit se usa para el signo del número, los seis siguientes para el exponente sesgado y los últimos nueve bits para la mantisa.

(a) Calcular los exponentes máximo y mínimo en base 10.

(b) Calcular el valor mínimo y máximo de los números positivos (excluido el cero) desnormalizados. Expresarlos en binario y en decimal. ¿Qué precisión tendría cada uno?

(c) ¿Cuántos números positivos desnormalizados se pueden representar con este formato?

(d) Calcular el número 17,6 en este formato. Redondear al par más cercano. Calcular el error relativo cometido al almacenarlo.

(a) El número de exponentes que podemos representar con m bits sería

$$2^m = 2^6 = 64$$

Si no tenemos en cuenta el signo, su valor nominal es

$$[0, 1, \ldots, 62, 63]$$

y teniendo en cuenta que el primero y el último están reservados

$$[R, 1, \ldots, 62, R]$$

Pero los enteros representados serían los anteriores menos el sesgo

$$\text{sesgo} = 2^{m-1} - 1 = 2^{6-1} - 1 = 2^5 - 1 = 31,$$

es decir, el rango de números a representar sería

$$[R, 1 - 31, \ldots, 62 - 31, R] = [R, -30, \ldots, 31, R]$$

por lo tanto

Solución:

$$\boxed{e_{min} = -30 \text{ y } e_{max} = 31}$$

(b) Los números desnormalizados se reconocen porque el exponente es el número binario reservado `000 000`. Entoces el bit escondido es cero y se asume que el exponente tiene el valor mínimo: en este caso, -30.

- El valor mínimo desnormalizado tiene mantisa mínima (distinta de cero) y se representa en binario

signo	exponente	mantisa
0	000 000	000 000 001

que se corresponde con

$$0{,}000\,000\,001 \times 2^{-30} \longrightarrow 2^{-9} \times 2^{-30} = 2^{-39} \approx 1{,}8 \times 10^{-12}$$

Solución:

$$\boxed{\text{Mínimo número desnormalizado: } 1{,}8 \times 10^{-12}}$$

- El valor máximo tiene mantisa máxima. Se representa en binario

signo	exponente	mantisa
0	000 000	111 111 111

que, teniendo en cuenta el bit escondido, el número máximo es

$$0{,}111\,111\,111 \times 2^{-30}$$

que expresado en decimal es

$$\left(0 + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} + 2^{-5} + 2^{-6} + 2^{-7} + 2^{-8} + 2^{-9}\right) \times 2^{-30} \approx 9{,}3 \times 10^{-10}$$

Solución:

Máximo número desnormalizado: $9{,}3 \times 10^{-10}$

La precisión del número mínimo es 1 porque es el número de dígitos almacenados, ya que los ceros a la izquierda no cuentan. Análogamente la precisión del máximo es 9

(c) El número de mantisas posibles es $2^9 = 512$ porque el bit escondido es siempre 0, no hay dos opciones. Y sólo hay un exponente. Si quitamos la mantisa correspondiente a la representación del cero (mantisa todo ceros) tenemos 511

Solución:

511 números desnormalizados

(d) Calcular el número 17,6 en este formato.

MANTISA:
Dividiendo el número y los cocientes sucesivamente por dos y almacenando el cociente y el resto de cada división

Cociente	17	8	4	2	1
Resto	1	0	0	0	↙

La parte entera se obtiene empezando por el último cociente (uno) y siguiendo con los restos en orden inverso

$$(17)_{10} = (10001)_2$$

La parte fracionaria se convierte a binario multiplicando en cada paso por dos, separándola en entera y fraccionaria y volviendo a multiplicar sólo la parte fraccionaria por dos.

Fraccionaria	0,6	**0,2**	**0,4**	**0,8**	**0,6**	0,2	0,4	0,8	0,6	$\cdots$
Entera		**1**	**0**	**0**	**1**	1	0	0	1	$\cdots$

Guardamos la parte entera en el orden en que la vamos generando. En este caso se repiten cuatro dígitos y por lo tanto es periódica

$$(0{,}6)_{10} = (0.\mathbf{1001}\ 1001\ \mathbf{1001}\ 1001\ \ldots)_2$$

Y el número completo es

$$(17{,}6)_{10} = (10001{,}1001\ 1001\ \ldots)_2$$

Normalizando

$$\mathbf{1}{,}0001\ 1001\ 1\mathbf{001}\ \ldots \times 2^4$$

Ahora, en negrita, el bit escondido y los números que no caben en la mantisa pero tendremos que tener en cuenta al redondearla. Cómo el número está más próximo al valor del truncado que al siguiente, redondeamos al truncado, y el número a almacenar es

$$\mathbf{1}{,}0001\ 1001\ 1 \times 2^4$$

La mantisa a almacenar es 000110011 y el exponente es 4.

EXPONENTE:
Como el sesgo = 31 el exponente sesgado es

$$4 + 31 = 35$$

Cociente	35	17	8	4	2	1
Resto	1	1	0	0	0	↙

que en binario es $(100011)_2$ y como el número es positivo, el bit del signo es cero y la representación completa será

signo	exponente	mantisa
0	100011	000 110 011

Problema 1.15:

Una máquina almacena números en punto flotante en base 2 en 18 bits, siguiendo un criterio similar al de la norma IEEE 754. El primer bit se usa para el signo del número, los doce siguientes para el exponente sesgado y los últimos cinco bits para la mantisa.

(a) Calcular los exponentes máximo y mínimo.

(b) Calcular el valor mínimo y máximo normalizados. Expresarlos en binario y en decimal (dejarlos indicados como potencias de 2). ¿Qué precisión tendría cada uno?

(a) El número de exponentes que podemos representar con m bits sería

$$2^m = 2^{12} = 4096$$

Si no tenemos en cuenta el signo van desde

$$[0, 1, \ldots, 4094, 4095]$$

y teniendo en cuenta que el primero y el último están reservados

$$[R, 1, \ldots, 4094, R]$$

Pero los enteros representados serían los anteriores menos el sesgo $= 2^{m-1} - 1 = 2^{12-1} - 1 = 2^{11} - 1 = 2047$, es decir, el rango de números a representar sería

$$[R, 1 - 2047, \ldots, 4094 - 2047, R] = [R, -2046, \ldots, 2047, R]$$

por lo tanto

Solución:

$$\boxed{e_{min} = -2046 \text{ y } e_{max} = 2047}$$

(b) Los números normalizados tienen 1 como bit escondido.

- El valor mínimo tiene mantisa mínima (todo ceros) y exponente mínimo (uno como valor nominal en binario, porque cero es un valor reservado) y se representa, en binario

signo	exponente	mantisa
0	000000000001	00000

que, teniendo en cuenta el bit escondido, se corresponde con

$$1{,}00000 \times 2^{-2046} \longrightarrow 2^{-2046}$$

Solución:

$$\boxed{\text{Mínimo número normalizado: } 2^{-2046}}$$

- El valor máximo tiene mantisa máxima (todo unos) y exponente máximo (el anterior al último, todo unos, que está reservado). Se representa en binario

signo	exponente	mantisa
0	111111111110	11111

que se corresponde con

$$1{,}11111 \times 2^{2047} \longrightarrow \left(1 + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} + 2^{-5}\right) \times 2^{2047} = 1{,}96875 \times 2^{2047}$$

Solución:

$$\boxed{\text{Máximo número normalizado: } 1{,}96875 \times 2^{2047}}$$

La precisión, tanto del numero mínimo como del máximo es el número de dígitos de la mantisa, que incluyen el bit escondido. Por lo tanto la precisión es $5 + 1 = 6$

Solución:

$$\boxed{p_{min} = 6 \text{ y } p_{max} = 6}$$

Problema 1.16:

Una máquina almacena números en punto flotante en base 2 en 15 bits siguiendo un criterio similar al de la norma IEEE 754. El primer bit se usa para el signo del número, los ocho siguientes para el exponente sesgado y los últimos seis bits para la mantisa.

(a) Calcular los exponentes máximo y mínimo.

(b) ¿Cual sería el ϵ de máquina expresado en base 10?

(c) ¿Cuál es el mayor entero que se puede almacenar de forma exacta?

(d) Calcular el número 271 en este formato. Redondear al par más cercano ¿Qué error absoluto cometemos al hacer esta operación?

(a) El número de exponentes que podemos representar con m bits sería

$$2^m = 2^8 = 256$$

Si no tenemos en cuenta el signo, van desde

$$[0, 1, \ldots, 254, 255]$$

y teniendo en cuenta que el primero y el último están reservados

$$[R, 1, \ldots, 254, R]$$

Pero los enteros representados serían los anteriores menos el sesgo, donde

$$\text{sesgo} = 2^{m-1} - 1 = 2^{8-1} - 1 = 2^7 - 1 = 127,$$

es decir, el rango de números a representar sería

$$[R, 1 - 127, \ldots, 254 - 127, R] = [R, -126, \ldots, 127, R]$$

por lo tanto

Solución:

$$\boxed{e_{min} = -126 \text{ y } e_{max} = 127}$$

(b) El ϵ de máquina es el menor número que se le puede sumar a 1 usando esta norma. Como el número de dígitos de la mantisa es 6, el número 1 se representa

$$1 \;=\; 1{,}000000 \times 2^0$$

y el siguiente número representable es

$$1 + \epsilon \;=\; 1{,}000001 \times 2^0$$

Por lo que

$$\epsilon \;=\; 0{,}000001 \times 2^0 \;=\; 2^{-6}$$

Y en base 10,

Solución:

$$\boxed{\epsilon = 0{,}015625}$$

(c) Como disponemos de 6 bits más el bit escondido (en negrita), hasta el número

$$\mathbf{1}\overbrace{111111}^{6}$$

que normalizado es

$$\mathbf{1}.\overbrace{111111}^{6} \times 2^6$$

podemos almacenar todos los dígitos de un número entero y por tanto representarlo de forma exacta. El número siguiente es

$$\mathbf{1}\overbrace{111111}^{6}+1=\mathbf{1}\overbrace{000000}^{6}0$$

que normalizado es

$$\mathbf{1}.\overbrace{000000}^{6}\times 2^7$$

y aunque no hemos podido almacenar su último dígito, como es un cero, no cometemos error. Sin embargo, el número entero siguiente es

$$\mathbf{1}\overbrace{000000}^{6}1$$

que normalizado es, otra vez

$$\mathbf{1}.\overbrace{000000}^{6}\times 2^7=128$$

porque no podemos almacenar el último dígito, que en este caso es un 1, y redondeamos al número anterior y, por lo tanto, estamos cometiendo un error.

Solución:

$$\boxed{2^7=(128)_{10}=(10000000)_2}$$

(d) Calcular el número 271 en este formato.

MANTISA: Dividimos el número y luego los cocientes por 2. Guardamos los cocientes y los restos.

Cociente	271	135	67	33	16	8	4	2	1
Resto		1	1	1	1	0	0	0	0 ↙

Para formar el número, empezamos por el último cociente, uno, y seguimos con los restos en orden inverso.

$$(271)_{10}=(100001111)_2$$

Normalizamos

$$\mathbf{1},00001\mathbf{11}\times 2^8$$

En negrita, el bit escondido, que no se almacena y los dígitos extra, que no se almacenan, pero hay que tener en cuenta al redondear el número. El número truncado es **1**,000011 y el siguiente número representable con esta norma es **1**,000011 + 0,000001 = **1**,000100. Como el número completo está más cerca de este que de el truncado los dígitos de la mantisa a almacenar son 000100 y el exponente es 8.

EXPONENTE: Dado que el sesgo es igual a 127 representaremos

$$8+127=135$$

Cociente	135	67	33	16	8	4	2	1
Resto		1	1	1	0	0	0	0 ↙

y el exponente en binario es $(10000111)_2$ y la representación completa será

signo	exponente	mantisa
0	10000111	000100

Hemos representado 271 con

$$1{,}000100 \times 2^8$$

por lo que el error es

$$\left|271 - (1 + 2^{-4}) \times 2^8\right| = |271 - 272| = 1$$

Problema 1.17:

Una máquina almacena números en punto flotante en 9 bits. El primer bit se usa para el signo del número, los cuatro siguientes para el exponente sesgado y los últimos cuatro bits para la magnitud de la mantisa. Si se sigue un criterio similar al de la norma IEEE 754:

(a) Calcular el número $(100110110)_2$ en base 10.

(b) ¿Cual sería el ϵ de máquina expresado en base 10?

(c) ¿Cuál es el mayor entero que se puede almacenar de forma exacta?

(d) ¿Cuales son el menor y el mayor real positivo que se almacena en forma normalizada? ¿Cómo se almacenarían en binario?

(e) ¿Cuántos números normalizados se pueden representar con este sistema?

(a) El número a convertir a decimal es

signo	exponente	mantisa
1	0011	0110

EXPONENTE

Como el número de bits es $m = 4$, el sesgo es $2^{m-1} - 1 = 2^3 - 1 = 7$. El exponente sesgado es $e - \text{sesgo} = e - 7$, siendo e el valor nominal del exponente. Como

$$(0011)_2 \to e = 2^1 + 2^0 = 3,$$

y el exponente sesgado es $3 - 7 = -4$.

MANTISA

Teniendo en cuenta el bit escondido

$$\mathbf{1}{,}0110 \times 2^{-4} = \left(1 + 2^{-2} + 2^{-3}\right) \times 2^{-4} = 0{,}0859375$$

Y teniendo en cuenta el signo, el número es

Solución:

$$\boxed{-0{,}0859375}$$

(b) El ϵ de máquina es la distancia que existe entre el número 1 y el siguiente número representable en ese sistema. El ϵ de máquina es una cota superior del error relativo que cometemos al almacenar cualquier número con este sistema. Como el número de dígitos de la mantisa es 4, el número 1 se representa

$$1 \;=\; 1{,}0000 \times 2^0$$

y el siguiente número representable es

$$1+\epsilon \;=\; 1{,}0001 \times 2^0$$

Por lo que ϵ es la diferencia entre ambos

$$\epsilon \;=\; 0{,}0001 \times 2^0 = 2^{-4}$$

Y en base 10 el ϵ de máquina es

Solución:

$$\boxed{2^{-4} = 0{,}0625}$$

(c) Como disponemos de 4 bits más el bit escondido (en negrita), hasta el número

$$\mathbf{1}1111$$

que normalizado es

$$\mathbf{1},1111 \times 2^4$$

podemos almacenar todos los dígitos de un número entero y por tanto representarlo de forma exacta. El número siguiente es

$$\mathbf{1}1111 + 1 = \mathbf{1}00000$$

que normalizado es

$$\mathbf{1},0000 \times 2^5$$

y aunque no hemos podido almacenar su último dígito, como es un cero, no cometemos error. Sin embargo, el número entero siguiente es

$$\mathbf{1}00001$$

que normalizado es, otra vez

$$\mathbf{1},0000 \times 2^5 = 32$$

porque no podemos almacenar el último dígito, que en este caso es un 1, y redondeamos al número anterior y, por lo tanto, estamos cometiendo un error.

Solución:

$$\boxed{2^5 = (32)_{10} = (100000)_2}$$

(d) Tenemos cuatro bits para los exponentes, es decir, $2^4 = 16$ valores distintos por lo que los exponentes posibles son

$$[0000, 0001, 0010, 0011, \ldots, 1101, 1110, 1111]$$

Pero el primer valor y el último están reservados.

$$[R, 0001, 0010, 0011, \ldots, 1101, 1110, R]$$

y su valor nominal es

$$[R, 1, 2, 3, \ldots, 13, 14, R]$$

Y si tenemos en cuenta el sesgo = 7, los valores que puede tomar son

$$[R, 1-7, 2-7, 3-7, \ldots, 13-7, 14-7, R] = [R, -6, -5, -4, \ldots, 6, 7, R]$$

con lo que el exponente mínimo es -6 en decimal y 0001 en binari y el exponente máximo es 7 y 1110 en binario. Por otra parte la mantisa mínima sería 0000 y la máxima 1111. También habría que tener en cuenta el bit escondido.

- El menor real positivo normalizado es, en binario

signo	exponente	mantisa
0	0001	0000

es decir

$$1{,}0000 \times 2^{-6}$$

que en decimal será

$$2^{-6} = 0{,}015625$$

- El mayor real positivo normalizado es, en binario

signo	exponente	mantisa
0	1110	1111

es decir

$$1{,}1111 \times 2^{7}$$

que en decimal será

$$(1 + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4}) \times 2^7 = 248$$

(e) Con 4 bits tenemos $2^m = 2^4 = 16$ posibles exponentes. Pero le tenemos que quitar el primer valor 0000 y el último 1111 que están reservados. Por lo tanto tenemos $16 - 2 = 14$ posibles exponentes.
Por otra parte, tenemos 4 bits para la mantisa, lo que quiere decir $2^4 = 16$ posibles mantisas.
Por lo tanto 14 exponentes $\times$ 16 mantisas = 224 números positivos normalizados distintos.

Problema 1.18:

Sea el conjunto de números en punto flotante en base 2, con precisión 4 y $e_{max} = 7$ y que sigue normas análogas a la IEEE 754 pero con distinto número de bits para el exponente y la mantisa. Calcular:

(a) Número de bits del exponente y la mantisa.

(b) El valor mínimo y máximo normalizados en binario y en decimal.

(c) El valor mínimo y máximo desnormalizados en binario y en decimal.

(a) Como seguimos la norma IEEE 754, el valor máximo del exponente coincide con el sesgo, por lo que

$$\text{sesgo} = 2^{m-1} - 1 = 7 \rightarrow 2^{m-1} = 8 \Rightarrow m - 1 = 3 \rightarrow m = 4.$$

Y si tenemos en cuenta que la precisión es 4 y uno de los bits es el bit escondido, el número de bits de la mantisa que almacenamos es tres y la representación sería:

signo	exponente	mantisa
s	$e_1e_2e_3e_4$	$m_1m_2m_3$

Solución:

bits exponente = 4, bits mantisa = 3

(b) La representación binaria del número más grande normalizado es

signo	exponente	mantisa
0	1110	111

que tendría el exponente máximo (1111 está reservado, así que tomamos el número anterior, 1110) y mantisa máxima (111)

$$(1{,}111) \times 2^7 \rightarrow \left(1 + 2^{-1} + 2^{-2} + 2^{-3}\right) \times 2^7 = 240.$$

La representación binaria del número más pequeño normalizado es

signo	exponente	mantisa
0	0001	000

que tendría exponente mínimo (0000 está reservado, así que tomamos el número siguiente, 0001) y mantisa mínima (000). Como el valor nominal del exponente es 1, tenemos que restarle el sesgo y entonces, el valor del exponente sería $1 - \text{sesgo} = 1 - 7 = -6$ y por lo tanto

$$(1{,}000) \times 2^{-6} \rightarrow 2^{-6} = 0{,}015625$$

Solución:

$$\boxed{\text{Max} = 240,\ \text{min} = 0{,}015625}$$

(c) Si el número es desnormalizado el exponente es 0000 pero se le atribuye el valor del menor exponente posible, es decir, -6. Además, el bit escondido es cero. La representación binaria del número más grande desnormalizado es

signo	exponente	mantisa
0	0000	111

que, de acuerdo con lo dicho para el exponente y el bit escondido, sería

$$(0{,}111) \times 2^{-6} \rightarrow \left(2^{-1} + 2^{-2} + 2^{-3}\right) 2^{-6} = 0{,}013671875.$$

La representación binaria del número más pequeño normalizado es

signo	exponente	mantisa
0	0000	001

por lo tanto

$$(0{,}001) \times 2^{-6} \rightarrow 2^{-3} \times 2^{-6} = 2^{-9} = 0{,}001953125$$

Solución:

$$\boxed{\text{Max} = 0{,}013671875,\ \text{min} = 0{,}001953125}$$

1.3. Dígitos significativos

Problema 1.19:

El error relativo aproximado al final de una iteración para calcular la raíz de una ecuación es 0,09 %. ¿Cuál es el mayor número de cifras significativas que podemos dar por buenas en la solución?

Decimos que x^* aproxima a x con p dígitos significativos si p es el mayor entero no negativo tal que el error relativo satisface

$$E_r = \left| \frac{x - x^*}{x} \right| \leq 5 \times 10^{-p}$$

Se tiene que, en este caso, el error relativo viene dado por

$$E_r = \frac{0{,}09}{100} = 0{,}0009 = 9 \times 10^{-4} = 0{,}9 \times 10^{-3} \leq 5 \times 10^{-3}$$

pero

$$E_r = 9 \times 10^{-4} \leq 5 \times 10^{-4}$$

y por tanto tenemos

Solución:

3 dígitos significativos

Problema 1.20:

¿Con cuantos dígitos significativos aproxima $x^* = 1000$ a $x = 999{,}99$? Entonces, ¿cómo deberíamos escribir x^*?

Se tiene que el error relativo

$$E_r = \frac{|x - x^*|}{|x|} = \frac{|999{,}99 - 1000|}{999{,}99} = 1 \times 10^{-5} \leq 5 \times 10^{-5},$$

y se tiene que

$$E_r = 1 \times 10^{-5} = 10 \times 10^{-6} \leq 5 \times 10^{-6},$$

Por lo tanto

Solución:

5 dígitos significativos y se escribiría $x^* = 1000{,}0$

Problema 1.21:

¿Con cuantos dígitos significativos aproxima $x_1^* = 0{,}27351$ a $x_1 = 0{,}2736$? ¿Y $x_2^* = 1$ a $x_2 = 0{,}9999$? Entonces, ¿cómo deberíamos escribir x_2^*?

En el primer caso, el error relativo al representar el exacto por el aproximado, viene dado por

$$E_r = \frac{0{,}2736 - 0{,}27351}{0{,}2736} \approx 3{,}3 \times 10^{-4} \leq 5 \times 10^{-4},$$

Y por lo tanto tenemos

Solución:

4 dígitos significativos

En el segundo caso, el error relativo viene dado por

$$E_r = \frac{0{,}9999 - 1}{0{,}9999} = 1 \times 10^{-4} \leq 5 \times 10^{-4},$$

Por lo tanto

Solución:

4 dígitos significativos y se escribiría $x_2^* = 1000{,}0$

Tema 2

Raíces de ecuaciones no lineales

2.1. El método de bisección

Problema 2.1:

Sea la ecuación

$$x^2 e^{1-x} - 2 = 0$$

(a) Demostrar que en $[-1{,}5, -0{,}5]$ existe una única raíz.

(b) ¿Se puede calcular por el método de bisección partiendo de dicho intervalo?

(c) Aproximar la raíz haciendo tres iteraciones con el método de bisección en dicho intervalo.

(d) Dar una cota del error cometido al calcular esta raíz.

(a) Sea $f(x) = x^2 e^{1-x} - 2$. Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[-1{,}5, -0{,}5]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es el producto de un polinomio por una exponencial y sumada a una función constante, todas ellas funciones continuas.
2. f tiene distinto signo en los extremos del intervalo:

$$f(-1{,}5) = 25{,}4 \quad \text{y} \quad \text{f}(-0{,}5) = -0{,}87$$

3. f es estrictamente creciente o decreciente en $[-1{,}5, -0{,}5]$. Es decir $f' > 0$ o $f' < 0$ en $(-1{,}5, -0{,}5)$:

$$f'(x) = 2x\, e^{1-x} + x^2 e^{1-x}(-1) = e^{1-x}(2x - x^2) = e^{1-x} x(2 - x).$$

Teniendo en cuenta el signo de los factores en el intervalo $(-1{,}5, -0{,}5)$. Se tiene

$$f'(x) = e^{1-x} x(2 - x) = (+)(-)(+) < 0$$

Y $f'(x) < 0$ en $(-1{,}5, -0{,}5)$.

(b) Si, porque se cumplen las condiciones necesarias, que son las condiciones 1. y 2. de la pregunta anterior.

(c) Iteramos. Para ello calculamos el punto medio del intervalo $[a, b]$ usando la fórmula

$$c = \frac{a+b}{2}.$$

En cada iteración escogemos entre $[a, c]$ y $[c, b]$ el nuevo intervalo de forma que se mantenga la condición de que el signo de la función es distinto en los extremos (en negrita).

k	a	$c = (a+b)/2$	b	$f(a)$	$f(c)$	$f(b)$	*cota de error*
1	$-1{,}5$	$(-1{,}5 + (-0{,}5))/2 = -1$	$-0{,}5$	25,4	**5,4**	**−0,87**	0,5
2	-1	$(-1 + (-0{,}5))/2 = -0{,}75$	$-0{,}5$	5,4	**1,23**	**−0,87**	0,25
3	$-0{,}75$	$(-0{,}75 + (-0{,}5))/2 = -0{,}625$	$-0{,}5$	1,23		$-0{,}87$	0,125

Y podemos dar como raíz aproximada $-0{,}625$ (la solución exacta es $-0{,}6269$)

(d) La cota de error es

$$\frac{b_0 - a_0}{2^n} = \frac{-0{,}5 - (-1{,}5)}{2^3} = \frac{1}{8} = 0{,}125$$

Problema 2.2:

Sea la ecuación:

$$\ln(3+x) + 3x^2 + x - 3 = 0$$

(a) Demostrar que en $[0, 2]$ existe una única raíz.

(b) Aproximar la raíz haciendo tres iteraciones por el método de bisección tomando como intervalo de partida $[0, 2]$.

(c) Dar una cota del error cometido al calcular esta raíz.

(a) Las condiciones (suficientes, no necesarias) que ha de cumplir

$$f(x) = \ln(3+x) + 3x^2 + x - 3$$

en $[0, 2]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es la suma de funciones continuas, el polinomio lo es siempre y $\ln(3+x)$ es continua en todo su dominio.
2. f tiene distinto signo en los extremos del intervalo: $f(0) \approx -1{,}9$ y $f(2) \approx 12{,}6$
3. $f' > 0$ o $f' < 0$ en $(0, 2)$:

$$f'(x) = 6x + 1 + \frac{1}{3+x} = \frac{6x^2 + 19x + 4}{3+x}.$$

Como $3+x$ es positivo en $(0,2)$ hace falta estudiar el signo del numerador. Si calculamos las raíces del numerador ($x_1 = -2{,}94$ y $x_2 = -0{,}23$) y factorizamos:

$$6x^2 + 19x + 4 = 6(x + 2{,}94)(x + 0{,}23)$$

Como $(x + 2{,}94)$ es siempre positivo en $(0,2)$ y $(x + 0{,}23)$ también

$$6x^2 + 19x + 4 = 6(x + 2{,}94)(x + 0{,}23) = (+)(+)(+) > 0$$

en $(0,2)$. Y recordamos que el denominador también era positivo.
Por lo tanto $f' > 0$ en $(0,2)$.

(b) Iteramos. Para ello calculamos el punto medio del intervalo $[a,b]$ usando la fórmula

$$c = \frac{a+b}{2}.$$

En cada iteración escogemos entre $[a,c]$ y $[c,b]$ el nuevo intervalo de forma que se mantenga la condición de que el signo de la función es distinto en los extremos (en negrita).

k	a	$c = (1+b)/2$	b	*f(a)*	*f(c)*	*f(b)*	*cota de error*
1	0	1	2	**−1,9**	**2,38**	12,6	$1-0=1$
2	0	0,5	1	−1,9	**−0,5**	**2,38**	$1-0{,}5=0{,}5$
3	0,5	0,75	1	−0,5		2,38	0,25

Y podemos dar como raíz aproximada 0,75 (la raíz verdadera es 0,60797).

(c) La cota de error es la longitud del último intervalo, que es $[0{,}5, 0{,}75]$ o $[0{,}75, 1]$ ambos de longitud 0,25. O podemos usar la fórmula

$$\frac{b_0 - a_0}{2^n} = \frac{2-0}{2^3} = 0{,}25$$

Problema 2.3:

Sea la función

$$h(x) = \left(x^3 - 1\right) e^{2x}$$

(a) Demostrar que en $[0{,}5, 1{,}5]$ existe un único extremo de h.

(b) Aproximar el extremo haciendo tres iteraciones con el método de bisección en dicho intervalo.

(c) Dar una cota del error cometido al calcular este extremo.

(a) Sea $h(x) = (x^3 - 1)\, e^{2x}$. La condición necesaria de extremo es $h'(x) = 0$. Por lo que, teniendo en cuenta que

$$h'(x) = 3x^2\, e^{2x} + \left(x^3 - 1\right) e^{2x} 2 = \left(2x^3 + 3x^2 - 2\right) e^{2x}$$

para que $h'(x) = 0$ habrá de ser $2x^3 + 3x^2 - 2 = 0$ porque e^{2x} es siempre positivo. Por lo tanto estamos buscando las raíces de la función

$$f(x) = 2x^3 + 3x^2 - 2$$

Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[0{,}5, 1{,}5]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es un polinomio, que es una función continua.
2. f tiene distinto signo en los extremos del intervalo:
$$f(0{,}5) = -1 \quad \text{y} \quad \text{f}(1{,}5) = 11{,}5$$
3. f es estrictamente creciente o decreciente en $(0{,}5, 1{,}5)$. Es decir $f' > 0$ o $f' < 0$ en $(0{,}5, 1{,}5)$:
$$f'(x) = 6x^2 + 6x = 6x(1 + x) = (+)(+) > 0.$$
para todos los $x \in (0{,}5, 1{,}5)$ del intervalo.Y $f'(x) > 0$ en $(0{,}5, 1{,}5)$.

(b) Iteramos. Para ello calculamos el punto medio del intervalo $[a, b]$ usando la fórmula

$$c = \frac{a+b}{2}.$$

En cada iteración escogemos entre $[a, c]$ y $[c, b]$ el nuevo intervalo de forma que se mantenga la condición de que el signo de la función es distinto en los extremos (en negrita).

k	a	$c = (a+b)/2$	b	*f(a)*	*f(c)*	*f(b)*	*cota de error*
1	0,5	$(0{,}5 + 1{,}5)/2 = 1$	1.5	$\mathbf{-1}$	**3**	11,5	0,5
2	0,5	$(0{,}5 + 1)/2 = 0{,}75$	1	$\mathbf{-1}$	**0,53**	3	0,25
3	0,5	$(0{,}5 + 0{,}75)/2 = 0{,}625$	0,75	-1		0,53	0,125

Y podemos dar como raíz aproximada 0,625 (la solución exacta es 0,6777)

(c) La cota de error es

$$\frac{b_0 - a_0}{2^n} = \frac{1{,}5 - 0{,}5}{2^3} = \frac{1}{8} = 0{,}125$$

Problema 2.4:

Sea la función

$$h(x) = x^2 + \frac{2x}{1+x}.$$

(a) Demostrar que esta función tiene un único extremo en $[-2, -1{,}5]$.

(b) ¿Se puede calcular por el método de bisección partiendo de dicho intervalo?

(c) Aproximar el extremo haciendo tres iteraciones.

(d) Dar una cota del error cometido al calcularlo.

(e) ¿Cuántos pasos se necesitarían para aproximar el extremo con un error menor que 10^{-8} ?

(a) Para que el punto α será un extremo de h es necesario que $h'(\alpha) = 0$. Como

$$h'(x) = 2x + \frac{2(1+x) - 2x(1)}{(1+x)^2} = \frac{2x(1+x)^2 + 2(1+x) - 2x(1)}{(1+x)^2} =$$

$$= \frac{2x(1+2x+x^2) + (2+2x) - 2x}{(1+x)^2} = \frac{(2x+4x^2+2x^3) + (2+2x) - 2x}{(1+x)^2} =$$

$$= \frac{2+2x+4x^2+2x^3}{(1+x)^2}$$

Es decir

$$h'(x) = \frac{2(1+x+2x^2+x^3)}{(1+x)^2}$$

Como $(1+x)^2 > 0$ en $[-2,-1{,}5]$, si tomamos $f(x) = 1+x+2x^2+x^3$ entonces

$$h \quad \text{tiene un extremo en } [-2,-1{,}5] \Longleftrightarrow f \quad \text{tiene una raíz en } [-2,-1{,}5].$$

O lo que es lo mismo

$$h'(t) = 0 \quad \text{en } [-2,-1{,}5] \Longleftrightarrow f(t) = 0 \quad \text{en } [-2,-1{,}5].$$

Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[-2,-1{,}5]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es un polinomio.
2. f tiene distinto signo en los extremos del intervalo:

$$f(-2) = -1 \quad \text{y} \quad \text{f}(-1{,}5) = 0{,}6$$

3. f es estrictamente creciente o decreciente en $[-2,-1{,}5]$. Es decir, $f' > 0$ o $f' < 0$ en $(-2,-1{,}5)$:

$$f'(x) = 1 + 4x + 3x^2.$$

Calculemos los ceros de este polinomio.

$$3x^2 + 4x + 1 = 0 \quad \Longrightarrow \quad x = \frac{-2 \pm \sqrt{4-3}}{3} = \frac{-2 \pm 1}{3}$$

$$x_1 = \frac{-3}{3} = -1 \qquad x_2 = \frac{-1}{3} = -0{,}33$$

y podemos factorizar

$$f'(x) = 1 + 4x + 3x^2 = 3(x+1)(x+0{,}33)$$

Teniendo en cuenta el signo de los factores en el intervalo $(-2,-1{,}5)$:

$$f'(x) = 3(x+1)(x+0{,}33) = (+)(-)(-) > 0$$

Y $f'(x) > 0$ en $(-2,-1{,}5)$.

(b) Si, porque se cumplen las condiciones necesarias, que son las condiciones 1. y 2. de la pregunta anterior.

(c) Iteramos. Para ello calculamos el punto medio del intervalo $[a, b]$ usando la fórmula

$$c = \frac{a+b}{2}.$$

En cada iteración escogemos entre $[a, c]$ y $[c, b]$ el nuevo intervalo de forma que se mantenga la condición de que el signo de la función es distinto en los extremos (en negrita).

k	a	$c = (a+b)/2$	b	$f(a)$	$f(c)$	$f(b)$	*cota de error*
1	-2	$-1{,}75$	$-1{,}5$	$\mathbf{-1{,}0}$	$\mathbf{0{,}016}$	0,63	0,25
2	-2	$-1{,}875$	$-1{,}75$	$-1{,}0$	$\mathbf{-0{,}44}$	$\mathbf{0{,}016}$	0,125
3	$-1{,}875$	$-1{,}8125$	$-1{,}75$	$-0{,}44$		0,016	0.0625

Y podemos dar como raíz aproximada $-1{,}8125$ (la solución exacta es $-1{,}75488$)

(d) La cota de error es

$$\frac{b_0 - a_0}{2^n} = \frac{-1{,}5 - (-2)}{2^3} = \frac{0{,}5}{8} = 0{,}0625$$

(e) El número de iteraciones n necesarias para que el error E sea menor que una tolerancia tol al aplicar el método de Bisección a un intervalo $[a, b]$ (donde se cumplen las condiciones) es

$$E < \frac{b-a}{2^n}$$

Si hacemos $\dfrac{b-a}{2^n} < tol$ se cumple que $E < tol$

$$\frac{b-a}{2^n} < tol \quad \Longrightarrow \quad \frac{b-a}{2^n} < tol \quad \Longrightarrow \quad \frac{b-a}{tol} < 2^n$$

Y como si aplicamos logaritmos (función creciente) a los dos miembros de una desigualdad, la desigualdad se mantiene

$$\ln \frac{b-a}{tol} < \ln 2^n \quad \Longrightarrow \quad \ln \frac{b-a}{tol} < n \ln 2$$

Si dividimos los dos términos de la desigualdad por $\ln 2$, como es positivo, no cambia el sentido de la desigualdad

$$\ln \frac{b-a}{tol} < \ln 2^n \quad \Longrightarrow \quad \frac{1}{\ln 2} \ln \frac{b-a}{tol} < n$$

Teniendo en cuenta que $tol = 10^{-8}$ y $b - a = -1{,}5 - (-2) = 0{,}5$

$$\frac{1}{\ln 2} \ln \frac{0{,}5}{10^{-8}} < n \quad \Longrightarrow \quad 25{,}6 < n$$

y para garantizar que el error es menor que 10^{-8} podemos tomar $n = 26$.

Problema 2.5:

Sea la función

$$h(x) = \frac{x-1}{x^3+1}.$$

(a) Demostrar que esta función tiene un único extremo en $[1,2{,}5]$.

(b) ¿Se puede calcular por el método de bisección partiendo de dicho intervalo?

(c) Aproximar el extremo haciendo tres iteraciones.

(d) Dar una cota del error cometido al calcularlo.

(e) ¿Cuántos pasos se necesitarían para aproximar el extremo con un error menor que 10^{-8} ?

(a) Para que el punto α sea un extremo de h es necesario que $h'(\alpha) = 0$. Como

$$h'(t) = \frac{(x^3+1)-(x-1)3x^2}{(x^3+1)^2} = \frac{x^3+1-3x^3+3x^2}{(x^3+1)^2} = \frac{1+3x^2-2x^3}{(x^3+1)^2}$$

Como $(x^3+1)^2 > 0$ en $[1,2{,}5]$, y si tomamos $f(x) = 1+3x^2-2x^3$ entonces

$$h \quad \text{tiene un extremo en } [1,2{,}5] \Longleftrightarrow f \quad \text{tiene una raíz en } [1,2{,}5].$$

O lo que es lo mismo

$$h'(t) = 0 \quad \text{en } [1,2{,}5] \Longleftrightarrow f(t) = 0 \quad \text{en } [1,2{,}5].$$

Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[-1{,}5, -0{,}5]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es un polinomio.
2. f tiene distinto signo en los extremos del intervalo:

$$f(1) = 2 \quad \text{y} \quad \text{f}(2{,}5) = -11{,}5$$

3. f es estrictamente creciente o decreciente en $[1,2{,}5]$. Es decir, $f' > 0$ o $f' < 0$ en $(1,2{,}5)$:

$$f'(x) = 6x - 6x^2 = -6(x-1)x$$

Teniendo en cuenta el signo de los factores en el intervalo $(1,2{,}5)$:

$$f'(x) = -6(x-1)x = (-)(+)(+) < 0$$

Y $f'(x) < 0$ en $(1,2{,}5)$.

(b) Si, porque se cumplen las condiciones necesarias, que son las condiciones 1 y 2 de la pregunta anterior.

(c) Iteramos. Para ello calculamos el punto medio del intervalo $[a, b]$ usando la fórmula

$$c = \frac{a+b}{2}.$$

En cada iteración escogemos entre $[a, c]$ y $[c, b]$ el nuevo intervalo de forma que se mantenga la condición de que el signo de la función es distinto en los extremos (en negrita).

k	*a*	*c*	*b*	*f(a)*	*f(c)*	*f(b)*	*cota de error*
1	1	$(1+2{,}5)/2 = 1{,}75$	2,5	**2**	**−0,53**	−11,5	0,75
2	1	$(1+1{,}75)/2 = 1{,}375$	1,75	2	**1,47**	**−0,53**	0,375
3	1,375	$(1{,}375+1{,}75)/2 = 1{,}5625$	1,75	1,47		−0,53	0,1875

Y podemos dar como raíz aproximada 1,5625 (la solución exacta es 1,67765)

(d) La cota de error es

$$\frac{b_0 - a_0}{2^n} = \frac{2{,}5-1}{2^3} = \frac{1{,}5}{8} = 0{,}1875$$

(e) El número de iteraciones n necesarias para que el error E sea menor que una tolerancia *tol* al aplicar el método de Bisección a un intervalo $[a, b]$ (donde se cumplen las condiciones) es

$$E < \frac{b-a}{2^n}$$

Si hacemos $\frac{b-a}{2^n} < tol$ se cumple que $E < tol$

$$\frac{b-a}{2^n} < tol \quad \Longrightarrow \quad \frac{b-a}{2^n} < tol \quad \Longrightarrow \quad \frac{b-a}{tol} < 2^n$$

Dado que si aplicamos logaritmos (función creciente) a los dos miembros de una desigualdad, la desigualdad se mantiene

$$\ln \frac{b-a}{tol} < \ln 2^n \quad \Longrightarrow \quad \ln \frac{b-a}{tol} < n \ln 2$$

Si dividimos los dos términos de la desigualdad por $\ln 2$, como es positivo, no cambia el sentido de la misma

$$\ln \frac{b-a}{tol} < \ln 2^n \quad \Longrightarrow \quad \frac{1}{\ln 2} \ln \frac{b-a}{tol} < n$$

Teniendo en cuenta que $tol = 10^{-8}$ y $b - a = 2{,}5 - 1 = 1{,}5$

$$\frac{1}{\ln 2} \ln \frac{1{,}5}{10^{-8}} < n \quad \Longrightarrow \quad 27{,}16 < n$$

y para garantizar que el error es menor que 10^{-8} podemos tomar $n = 28$.

Problema 2.6:

Si partimos del intervalo $[a_0, b_0]$, el error absoluto del método de bisección e_a en el paso n, está acotado por la expresión

$$e_a < \frac{b_0 - a_0}{2^n}.$$

Calcular un número suficiente de iteraciones para que el error absoluto e_a sea menor que una tolerancia dada, tol.

Dado que

$$e_a < \frac{b_0 - a_0}{2^n},$$

una condición suficiente para que le error absoluto sea menor que tol es que

$$\frac{b_0 - a_0}{2^n} < tol.$$

Como tanto tol como 2^n son cantidades positivas y teniendo en cuenta que

$$a < b \quad \text{y} \quad c > 0 \Longrightarrow ac < bc$$

se tiene

$$\frac{b_0 - a_0}{2^n} < tol \Longrightarrow \frac{b_0 - a_0}{tol} < 2^n \tag{2.1}$$

Como $f(x) = \log(x)$ es una función estrictamente creciente se tiene que si

$$0 < x < y \Longrightarrow \log(x) < \log(y)$$

y a partir de (2.1)

$$\log\left(\frac{b_0 - a_0}{tol}\right) < \log\left(2^n\right)$$

y teniendo en cuenta las propiedades de los logaritmos

$$\log\left(\frac{b_0 - a_0}{tol}\right) < n \log 2.$$

Como $\log 2 > 0$, si dividimos los dos miembros de la desigualdad por $\log 2$ el sentido de la desigualdad no cambia y

$$\frac{1}{\log 2} \log\left(\frac{b_0 - a_0}{tol}\right) < n$$

y si $E(x)$ es la función parte entera de x la solución es

$$n = E\left[\frac{1}{\log 2} \log\left(\frac{b_0 - a_0}{tol}\right)\right] + 1.$$

2.2. El método de *Régula-Falsi*

Problema 2.7:

Sea la ecuación

$$x^3 - 3x + 3 = 0$$

(a) Demostrar que en $[-3, -2]$ existe una única raíz.

(b) ¿Se puede calcular por el método de *régula-falsi* partiendo de dicho intervalo?

(c) Aproximar la raíz haciendo dos iteraciones con el método de *régula-falsi* en dicho intervalo.

(d) Dar una cota del error cometido al calcular esta raíz. ¿Es una buena cota?¿Por qué?

(a) Sea

$$f(x) = x^3 - 3x + 3$$

Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[-3, -2]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es un polinomio, que es una función continua.
2. f tiene distinto signo en los extremos del intervalo:

$$f(-3) = -15 \quad \text{y} \quad \text{f}(-2) = 1$$

3. f es estrictamente creciente o decreciente en $[0, 1]$. Es decir $f' > 0$ o $f' < 0$ en $(0, 1)$:

$$f'(x) = 3x^2 - 3.$$

Teniendo en cuenta que

$$3x^2 - 3 = 0 \Longrightarrow x_{1,2} = \pm\sqrt{1} \Longrightarrow x_1 = -1. \quad x_2 = 1.$$

podemos factorizar f

$$f'(x) = 3\,(x-1)\,(x+1) = (+)\,(-)\,(-) > 0$$

para todos los $x \in (-3, -2)$ del intervalo.Y $f'(x) > 0$ en $(-3, -2)$.

(b) Si, porque se cumplen las condiciones necesarias, que son las condiciones 1. y 2. de la pregunta anterior.

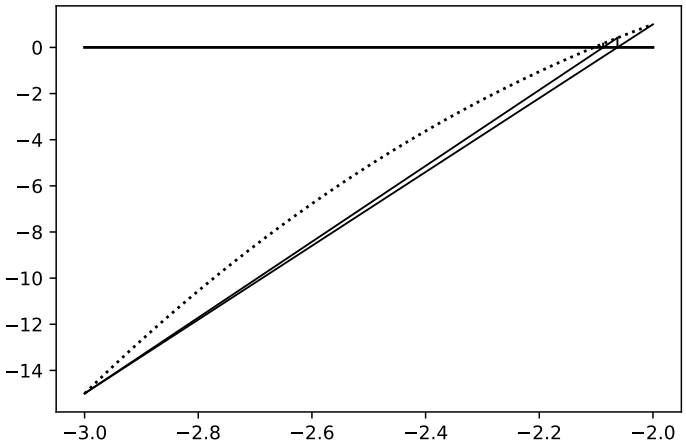
(c) El método de *régula-falsi* calcula un nuevo punto en cada iteración con la fórmula

$$c = b - f(b)\frac{b-a}{f(b)-f(a)} = \frac{bf(b) - bf(a) - bf(b) + af(b)}{f(b)-f(a)} = \frac{a\,f(b) - b\,f(a)}{f(b)-f(a)}$$

y selecciona el intervalo siguiente de forma que los signos de los extremos sean distintos

k	a	$c = \frac{a\,f(b) - b\,f(a)}{f(b) - f(a)}$	b	$f(a)$	$f(c)$	$f(b)$	$cota\ de\ error = b - a$
1	-3	$-2{,}0625$	-2	$\mathbf{-15}$	$\mathbf{0{,}41}$	1	1
2	-3	$-2{,}0877$	$-2{,}0625$	-15		0,41	0,9375

Y podemos dar como raíz aproximada $-2{,}0877$ (la solución exacta es $-2{,}1038$)



(d) La cota de error es la longitud del intervalo que contiene la raíz

$$b_2 - a_2 = 0{,}9375$$

Régula-falsi no suele dar buenas cotas de error porque la longitud del intervalo que contiene la raíz en muchos casos no se reduce significativamente con las iteraciones. En este caso el error absoluto es $E_a = -2{,}0877 - (-2{,}1038) \approx 0{,}02$ que es mucho menor que la cota dada.

Problema 2.8:

Sea la ecuación

$$x^3 - 4x + 1 = 0$$

(a) Demostrar que en $[0, 1]$ existe una única raíz.

(b) ¿Se puede calcular por el método de *régula-falsi* partiendo de dicho intervalo?

(c) Aproximar la raíz haciendo dos iteraciones con el método de *régula-falsi* en dicho intervalo.

(d) Dar una cota del error cometido al calcular esta raíz. ¿Es una buena cota?¿Por qué?

(a) Sea

$$f(x) = x^3 - 4x + 1$$

Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[0, 1]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es un polinomio, que es una función continua.
2. f tiene distinto signo en los extremos del intervalo:

$$f(0) = 1 \quad \text{y} \quad \text{f}(1) = -2$$

3. f es estrictamente creciente o decreciente en $[0, 1]$. Es decir $f' > 0$ o $f' < 0$ en $(0, 1)$:

$$f'(x) = 3x^2 - 4.$$

Teniendo en cuenta que

$$3x^2 - 4 = 0 \Longrightarrow x_{1,2} = \pm\sqrt{\frac{4}{3}} \Longrightarrow x_1 = -1{,}15 \quad x_2 = 1{,}15$$

podemos factorizar f

$$f'(x) = 3\,(x - 1{,}15)\,(x + 1{,}15) = (+)\,(-)\,(+) < 0$$

para todos los $x \in (0, 1)$ del intervalo.Y $f'(x) < 0$ en $(0, 1)$.

(b) Si, porque se cumplen las condiciones necesarias, que son las condiciones 1. y 2. de la pregunta anterior.

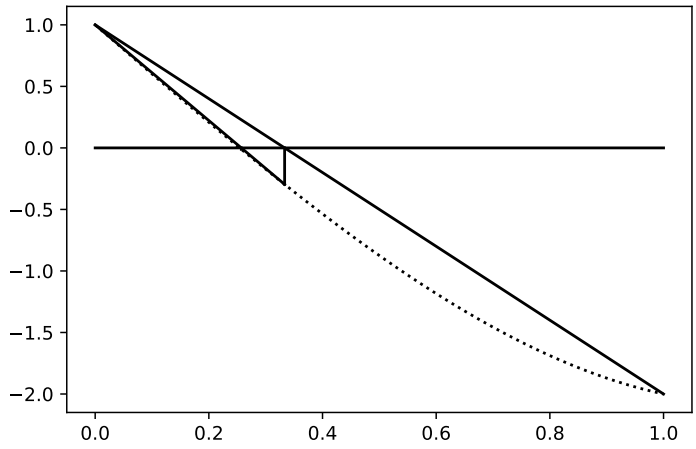
(c) El método de *régula-falsi* calcula un nuevo punto en cada iteración con la fórmula

$$c = b - f(b)\frac{b - a}{f(b) - f(a)} = \frac{bf(b) - bf(a) - bf(b) + af(b)}{f(b) - f(a)} = \frac{a\,f(b) - b\,f(a)}{f(b) - f(a)}$$

y selecciona el intervalo siguiente de forma que los signos de los extremos sean distintos

k	a	$c = \frac{a\,f(b) - b\,f(a)}{f(b) - f(a)}$	b	$f(a)$	$f(c)$	$f(b)$	$cota\ de\ error = b - a$
1	0	0,3333	1	**1**	**−0,296**	−2	1
2	0	0,2571	0,3333	1		−0,296	0,3333

Y podemos dar como raíz aproximada 0,2571 (la solución exacta es 0,2541)



(d) La cota de error es la longitud del intervalo que contiene la raíz

$$b_2 - a_2 = 0{,}3333$$

El métodod de *régula-falsi* no suele dar buenas cotas de error porque la longitud del intervalo que contiene la raíz en muchos casos no se reduce significativamente con las iteraciones. En este caso el error absoluto es $E_a = 0{,}2571 - 0{,}2541 = 0{,}003$ que es mucho menor que la cota dada.

2.3. El método de Newton

Problema 2.9:

Aproximar utilizando el método de Newton

$$(a)\ r = \sqrt[4]{60} \quad (b)\ r = \sqrt[3]{75}.$$

Utilizar como punto inicial $x_0 = 3$ y $x_0 = 4$ respectivamente. Realizar dos iteraciones. Redondear a 5 cifras decimales en cada paso.

(a) Buscamos un valor de x que verifique

$$x = \sqrt[4]{60} \quad \Rightarrow \quad x^4 = 60 \quad \Rightarrow \quad x^4 - 60 = 0$$

Por lo tanto, si tomamos la función $f(x) = x^4 - 60$ y calculamos su raíz positiva, habremos solucionado nuestro problema. Como $f'(x) = 4x^3$, la función de iteración de Newton será

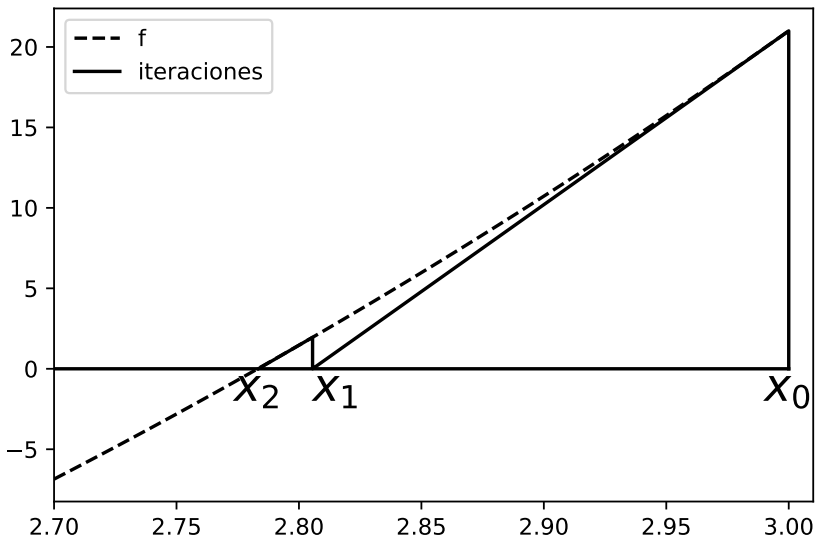
$$x_{k+1} = x_k - \frac{x_k^4 - 60}{4x_k^3}$$

Si $x_0 = 3$

$$x_1 = x_0 - \frac{x_0^4 - 60}{4x_0^3} = 3 - \frac{3^4 - 60}{4(3^3)} = 3 - \frac{81 - 60}{4(27)} = 3 - \frac{21}{108} = 3 - 0{,}19444 = 2{,}80556$$

Si $x_1 = 2{,}80556$

$$x_2 = x_1 - \frac{x_1^4 - 60}{4x_1^3} = 2{,}80556 - \frac{2{,}80556^4 - 60}{4(2{,}80556^3)} = 2{,}80556 - 0{,}02214 = 2{,}78342$$



(b) Buscamos un valor de x que verifique

$$x = \sqrt[3]{75} \quad \Rightarrow \quad x^3 = 75 \quad \Rightarrow \quad x^3 - 75 = 0$$

Por lo tanto, si tomamos la función $f(x) = x^3 - 75$ y calculamos su raíza positiva, habremos solucionado el problema. Teniendo en cuenta $f'(x) = 3x^2$, la función de iteración de Newton será

$$x_{k+1} = x_k - \frac{x_k^3 - 75}{3x_k^2}$$

Si $x_0 = 4$

$$x_1 = x_0 - \frac{x_0^3 - 75}{3x_0^2} = 4 - \frac{4^3 - 75}{3(4^2)} = 4 - \frac{64 - 75}{3(16)} = 4 - \frac{11}{48} = 4 - \frac{11}{48} = 4 + 0{,}22917 = 4{,}22917$$

Si $x_1 = 4{,}22917$

$$x_2 = x_1 - \frac{x_1^3 - 75}{3x_1^2} = 4{,}22917 - \frac{4{,}22917^3 - 75}{3(4{,}22917^2)} = 4{,}21720$$

Problema 2.10:

Sea la función

$$h(t) = \left(t^3 - 4t^2\right) + \ln(3 + t)$$

(a) Demostrar que esta función tiene un extremo relativo en $[0, 1]$.

(b) Aproximar el extremo utilizando el Método de Newton. Utilizar como punto inicial $t_0 = 0$ y realizar 2 iteraciones.

(c) Si la solución exacta es $\alpha = 0{,}0417484$ obtener el error absoluto y relativo.

(a) Para que el punto α sea un extremo de h es necesario que $h'(\alpha) = 0$. Como

$$h'(t) = 3t^2 - 8t + \frac{1}{3 + t} = \frac{3t^3 + t^2 - 24t + 1}{3 + t}$$

y como si tomamos $f(x) = 3t^3 + t^2 - 24t + 1$ entonces

$$h \quad \text{tiene un extremo en } [0, 1] \iff f \quad \text{tiene una raíz en } [0, 1].$$

O lo que es lo mismo

$$h'(t) = 0 \quad \text{en } [0, 1] \iff f(t) = 0 \quad \text{en } [0, 1].$$

Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[0, 1]$ para que exista una única raíz en el intervalo son

1. f continua: f es continua porque es un polinomio.
2. f tiene distinto signo en los extremos del intervalo: $f(0) = 1$ y $f(1) = -19$

3. $f' > 0$ o $f' < 0$ en $[0,1]$:

$$f'(t) = 9t^2 + 2t - 24.$$

Si calculamos las raíces y factorizamos:

$$f'(t) = 9t^2 + 2t - 24 = 9(t + 1{,}75)(t - 1{,}53)$$

Como $(t + 1{,}75)$ es siempre positivo en $[0,1]$ y $(t - 1{,}53)$ siempre es negativo $f'(t) < 0$ en $[0,1]$.

(b) Si iteramos por Newton, usaremos la fórmula

$$t_{k+1} = t_k - \frac{f(t_k)}{f'(t_k)}$$

Como $f'(t) = 9t^2 + 2t - 243$

$$t_{k+1} = t_k - \frac{3t_k^3 + t_k^2 - 24t_k + 1}{9t_k^2 + 2t_k - 24}$$

Si $t_0 = 0$ entonces $t_1 = 0{,}0416667$

Y tenemos $t_1 = -1{,}2$. Entonces $t_2 = 0{,}0417484$

Y los errores son

$E_a = |t_2 - \alpha| = |0{,}0417484 - 0{,}0417484| \approx 3{,}8 \times 10^{-10}$

$E_r = \dfrac{E_a}{|\alpha|} = \dfrac{3{,}84232 \times 10^{-10}}{0{,}0417484} \approx 9{,}2 \times 10^{-9}$

Problema 2.11:

Sea la función

$$h(t) = \left(t^3 + t - 1\right) e^{-t}.$$

(a) Demostrar que esta función tiene un único extremo en $[2,3]$.

(b) Aproximar el extremo utilizando el método de Newton. Utilizar como punto inicial $t_0 = 3$ y realizar 2 iteraciones.

(a) Para que el punto α será un extremo de h es necesario que $h'(\alpha) = 0$. Como

$$h'(t) = e^{-t}\left(1 + 3t^2\right) - e^{-t}\left(t^3 + t - 1\right) = -e^{-t}\left(t^3 - 3t^2 + t - 2\right)$$

Como $e^{-t} > 0$, si tomamos $f(x) = t^3 - 3t^2 + t - 2$ entonces

$$h \quad \text{tiene un extremo en } [2,3] \iff f \quad \text{tiene una raíz en } [2,3].$$

O lo que es lo mismo

$$h'(t) = 0 \quad \text{en } [2,3] \iff f(t) = 0 \quad \text{en } [2,3].$$

Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[2,3]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es un polinomio.
2. f tiene distinto signo en los extremos del intervalo: $f(2) = -4$ y $f(3) = 1$
3. f es estrictamente creciente $f' > 0$ o decreciente $f' < 0$ en $(2, 3)$:

$$f'(t) = 3t^2 - 6t + 1.$$

Calculamos las raíces de este polinomio de segundo grado

$$t_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{6 \pm \sqrt{6^2 - 4(3)(1)}}{2(3)} \quad t_1 = 0{,}18 \quad t_2 = 1{,}81$$

Si factorizamos:

$$f'(t) = 3t^2 - 6t + 1 = 3(t - t_1)(t - t_2) = 3(t - 0{,}18)(t - 1{,}81) > 0$$

porque los tres factores son siempre positivos para $t \in (2, 3)$.

(b) La sucesión se define

$$t_{k+1} = t_k - \frac{f(t_k)}{f'(t_k)} \qquad t_0 = 3$$

o lo que es lo mismo

$$t_{k+1} = t_k - \frac{t_k^3 - 3t_k^2 + t_k - 2}{3t_k^2 - 6t_k + 1} \qquad t_0 = 3.$$

Por lo tanto

$$t_1 = t_0 - \frac{t_0^3 - 3t_0^2 + t_0 - 2}{3t_0^2 - 6t_0 + 1} = 3 - \frac{3^3 - 3(3^2) + 3 - 1}{3(3^2) - 6(3) + 1} = 3 - \frac{1}{10} = 2{,}9$$

$$t_2 = t_1 - \frac{t_1^3 - 3t_1^2 + t_1 - 2}{3t_1^2 - 6t_1 + 1} = 2{,}9 - \frac{2{,}9^3 - 3(2{,}9^2) + 2{,}9 - 1}{3(2{,}9^2) - 6(2{,}9) + 1} = 2{,}8933$$

(La raíz es $\alpha = 2{,}89329$)

Problema 2.12:

Sea la función

$$h(t) = t + \ln|t| - \arctan t.$$

(a) Demostrar que esta función tiene un único extremo relativo en $[-2, -1]$.

(b) Aproximar el extremo utilizando el método de Newton. Utilizar como punto inicial $t_0 = -1$ y realizar 2 iteraciones.

(a) Para que el punto α será un extremo de h es necesario que $h'(\alpha) = 0$. Como

$$h'(t) = 1 + \frac{1}{t} - \frac{1}{1 + t^2} = \frac{1 + t^2 + t^3}{t(1 + t^2)}$$

Teniendo en cuenta que el denominador no se anula en el intervalo $[-2,-1]$, si tomamos $f(x) = 1 + t^2 + t^3$ entonces

$$h \quad \text{tiene un extremo en } [-2,-1] \Longleftrightarrow f \quad \text{tiene una raíz en } [-2,-1].$$

O lo que es lo mismo

$$h'(t) = 0 \quad \text{en } [-2,-1] \Longleftrightarrow f(t) = 0 \quad \text{en } [-2,-1].$$

Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[-2,-1]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es un polinomio.
2. f tiene distinto signo en los extremos del intervalo: $f(-2) = -3$ y $f(-1) = 1$
3. $f' > 0$ o $f' < 0$ en $(-2,-1)$:

$$f'(t) = 2t + 3t^2.$$

Si factorizamos:

$$f'(t) = 2t + 3t^2 = t(2+3t) = 3t\left(\frac{2}{3} + t\right)$$

Como t es siempre negativo en $(-2,-1)$ y $\left(\frac{2}{3} + t\right)$ también es negativo y entonces $f'(t) = (+)(-)(-) > 0$ en $(-2,-1)$.

(b) La sucesión se define

$$t_{k+1} = t_k - \frac{f(t_k)}{f'(t_k)} \qquad t_0 = -1$$

o lo que es lo mismo

$$t_{k+1} = t_k - \frac{1 + t_k^2 + t_k^3}{2t_k + 3t_k^2} \qquad t_0 = -1.$$

Por lo tanto

$$t_1 = t_0 - \frac{1 + t_0^2 + t_0^3}{2t_0 + 3t_0^2} = (-1) - \frac{1 + (-1)^2 + (-1)^3}{2(-1) + 3(-1)^2} = -1 - \frac{1+1-1}{-2+3} = -2$$

$$t_2 = t_1 - \frac{1 + t_1^2 + t_1^3}{2t_1 + 3t_1^2} = (-2) - \frac{1 + (-2)^2 + (-2)^3}{2(-2) + 3(-2)^2} = -2 - \frac{1+4-8}{-4+12} = -1{,}625$$

2.4. El método de la secante

Problema 2.13:

Aproximar utilizando el método de la secante $r = \sqrt[4]{3}$. Utilizar como puntos iniciales $x_0 = 1$ y $x_1 = 2$. Realizar 2 iteraciones. Si $r = 1{,}3161$, calcular el error absoluto y relativo de la aproximación.

Vamos a plantear el problema como el cálculo de una raíz que resolveremos usando el método de la secante.

Si tenemos $x = \sqrt[4]{3} \Rightarrow x^4 = 3 \Rightarrow x^4 - 3 = 0$ y la función que usaremos para obtener sus raíces será

$$f(x) = x^4 - 3.$$

Si iteramos usando el método de la Secante usaremos la fórmula

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

Por lo tanto

$$x_{k+1} = x_k - \frac{(x_k^4 - 3)(x_k - x_{k-1})}{(x_k^4 - 3) - (x_{k-1}^4 - 3)}$$

o lo que es lo mismo

$$x_{k+1} = x_k - \frac{(x_k^4 - 3)(x_k - x_{k-1})}{(x_k^4 - x_{k-1}^4)}$$

Si $x_0 = 1$ y $x_1 = 2$

$$x_2 = 2 - \frac{(2^4 - 3)(2 - 1)}{(2^4 - 1^4)} = 2 - \frac{(16 - 3)(1)}{(16 - 1)} = 2 - \frac{13}{15}$$

Si $x_1 = 2$ y $x_2 = 1{,}1333$

$$x_3 = 1{,}133 - \frac{(1{,}133^4 - 3)(1{,}133 - 2)}{(1{,}133^4 - 2^4)} = 1{,}2149$$

Los errores absoluto y relativo son

$$e_a = |x - x^*| = |1{,}3161 - 1{,}2149| = 0{,}1012 \qquad e_r = \frac{e_a}{\sqrt[4]{3}} = \frac{0{,}1012}{1{,}3161} \approx 0{,}08$$

Problema 2.14:

Aproximar utilizando el método de la secante $r = \sqrt[3]{2}$. Utilizar como puntos iniciales $x_0 = 1$ y $x_1 = 2$. Realizar 2 iteraciones y calcular el residual. Si $r = 1{,}2599$, calcular el error absoluto de la aproximación.

Resolvamos el problema como el cálculo de una raíz que calcularemos usando el método de la secante. Si tenemos $x = \sqrt[3]{2} \Rightarrow x^3 = 2 \Rightarrow x^3 - 2 = 0$ y nuestra función puede ser

$$f(x) = x^3 - 2.$$

Si iteramos usando el método de la Secante usaremos la fórmula

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

Por lo tanto

$$x_{k+1} = x_k - \frac{(x_k^3 - 2)(x_k - x_{k-1})}{(x_k^3 - 2) - (x_{k-1}^3 - 2)}$$

o lo que es lo mismo

$$x_{k+1} = x_k - \frac{(x_k^3 - 2)(x_k - x_{k-1})}{(x_k^3 - x_{k-1}^3)}$$

Si $x_0 = 1$ y $x_1 = 2$ entonces $x_2 = 1{,}1429$
Si $x_1 = 2$ y $x_2 = 1{,}1429$ entonces $x_3 = 1{,}2097$
El residual será

$$|f(x_3)| = \left|x_3^3 - 2\right| = 0{,}2298.$$

Los errores absoluto y relativo son

$$e_a = |x - x^*| = |1{,}2599 - 1{,}2097| = 0{,}0502 \qquad e_r = \frac{e_a}{\sqrt[3]{2}} = \frac{0{,}0502}{1{,}2599} \approx 0{,}04$$

Problema 2.15:

Sea la función

$$h(x) = \frac{x^3}{3} - 2\ln(2 + x)$$

(a) Demostrar que en $[0, 1]$ existe un único extremo de h.

(b) Aproximar el extremo haciendo dos iteraciones con el método de Newton con $x_0 = 1$.

(c) Aproximar el extremo haciendo dos iteración con el método de la secante con $x_0 = 0$ y $x_1 = 1$.

(a) Sea $h(x) = x^3/3 - 2\ln(2+x) = 0$. La condición necesaria de extremo es $h'(x) = 0$. Por lo que, teniendo en cuenta que

$$h'(x) = x^2 - \frac{2}{2+x} = \frac{x^3 + 2x^2 - 2}{2+x}$$

para que $h'(x) = 0$ habrá de ser $x^3 + 2x^2 - 2 = 0$ y $2 + x \neq 0$. Cómo $x \neq -2$ en el intervalo $[0, 1]$, estamos buscando las raíces de la función

$$f(x) = x^3 + 2x^2 - 2$$

Las condiciones (suficientes, no necesarias) que ha de cumplir f en $[0, 1]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es un polinomio, que es una función continua.
2. f tiene distinto signo en los extremos del intervalo:

$$f(0) = -2 \quad \text{y} \quad \text{f}(1) = 1$$

3. f es estrictamente creciente o decreciente en $(0, 1)$. Es decir $f' > 0$ o $f' < 0$ en $(0, 1)$:
$$f'(x) = 3x^2 + 4x = x(3x + 4) = (+)(+) > 0.$$
para todos los $x \in (0, 1)$ del intervalo.Y $f'(x) > 0$ en $(0, 1)$.

(b) Hagamos dos iteraciones usando el método de Newton usando $x_0 = 1$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = x_0 - \frac{x_0^3 + 2x_0^2 - 2}{3x_0^2 + 4x_0} = 1 - \frac{1+2-2}{3+4} = 1 - \frac{1}{7} = 0{,}8571$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = x_1 - \frac{x_1^3 + 2x_1^2 - 2}{3x_1^2 + 4x_1} = 0{,}8571 - \frac{0{,}8571^3 + 2(0{,}8571)^2 - 2}{3(0{,}8571)^2 + 4(0{,}8571)} = 0{,}8395$$

(c) Hagamos dos iteraciones usando el método de la secante

$$x_0 = 0 \quad x_1 = 1 \quad f(x_0) = -2 \quad f(x_1) = 1$$

$$x_2 = x_1 - f(x_1)\frac{x_1 - x_0}{f(x_1) - f(x_0)} = 1 - 1\frac{1-0}{1-(-2)} = 1 - \frac{1}{3} = 0{,}6667$$

$$x_1 = 1 \quad x_2 = 0{,}6667 \quad f(x_1) = 1 \quad f(x_2) = -0{,}8148$$

$$x_3 = x_2 - f(x_2)\frac{x_2 - x_1}{f(x_2) - f(x_1)} = 0{,}6667 - (-0{,}8148)\frac{0{,}6667 - 1}{-0{,}8148 - 1} = 0{,}8163$$

(El extremo es $\alpha = 0{,}8393$)

2.5. El método de punto fijo

Problema 2.16:

Sean las funciones

$$f(x) = x - \cos\frac{x}{2} \qquad g_1(x) = \cos\frac{x}{2} \qquad g_2(x) = 2x - \cos\frac{x}{2}$$

(a) Demostrar que la ecuación $f(x) = 0$ tiene la misma raíz que $g_i(x) = x$ con $i = 1, 2$.

(b) Enunciar el teorema de la aplicación contractiva.

(c) Demostrar analíticamente que la función g_1 cumple las condiciones del teorema de la aplicación contractiva el intervalo con $[0, 2]$.

(d) Demostrar analíticamente que la función g_2 no cumple las condiciones del teorema de la aplicación contractiva el intervalo con $[0, 2]$.

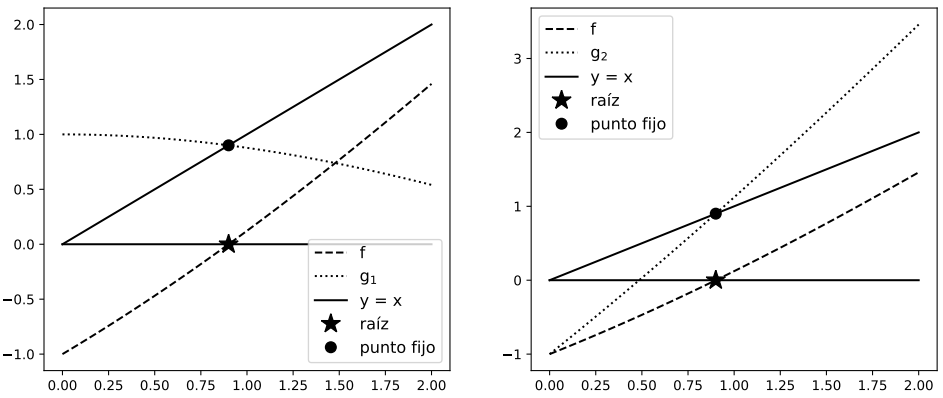
(e) Hacer dos iteraciones con g_1 tomando como punto inicial $x_0 = 0$

(a) Demostremos que la ecuación $f(x) = 0$ tiene la misma raíz que $g_i(x) = x$ con $i = 1, 2$.

$$g_1(x) = \cos\frac{x}{2} \iff x = \cos\frac{x}{2} \iff x - \cos\frac{x}{2} = 0 \iff f(x) = x - \cos\frac{x}{2}$$

$$g_2(x) = 2x - \cos\frac{x}{2} \iff x = 2x - \cos\frac{x}{2} \iff 0 = x - \cos\frac{x}{2} \iff f(x) = x - \cos\frac{x}{2}$$

Y hemos demostrado que son ecuaciones equivalentes. Gráficamente:



(b) El Teorema de la aplicación contractiva dice: sea g una función derivable definida en el intervalo $[a, b] \subset \mathbb{R}$ y $x_0 \in [a, b]$ un punto del intervalo. Si

1. $x \in [a, b] \quad \Rightarrow \quad g(x) \in [a, b]$
2. $|g'(x)| \leq k < 1$ para todo $x \in [a, b]$, $k \in \mathbb{R}$

Entonces g tiene un único punto fijo $\alpha \in [a,b]$, y la sucesión x_n definida como $x_{i+1} = f(x_i)$ que tiene como punto inicial x_0 converge a α con orden al menos lineal.

(c) Empecemos por la **condición 2**

$$|g_1'(x)| = \left|-\frac{1}{2}\text{sen}\frac{x}{2}\right| = \frac{1}{2}\left|\text{sen}\frac{x}{2}\right|$$

Como la función seno está siempre comprendida entre -1 y 1, en valor absoluto es siempre menor que 1 y

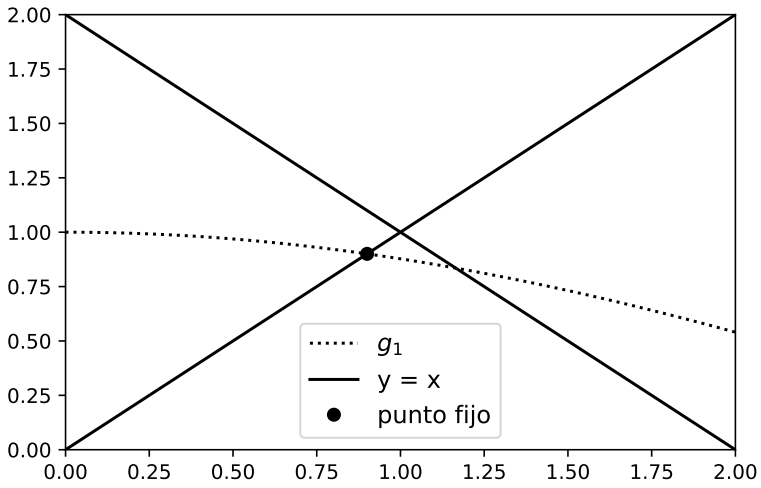
$$|g_1'(x)| = \left|-\frac{1}{2}\text{sen}\frac{x}{2}\right| = \frac{1}{2}\left|\text{sen}\frac{x}{2}\right| \leq \frac{1}{2} \times 1 = 0{,}5 \quad \text{en} \quad [0,2]$$

y se cumple la segunda condición 2.

Veamos la **condición 1**. Como la función coseno $\cos\dfrac{x}{2}$ es positiva en $[0,2]$ y está siempre comprendida entre -1 y 1.

$$x \in [0,2] \;\Rightarrow\; 0 \leq \cos\frac{x}{2} \leq 1 \;\Rightarrow\; g(x) \in [0,1] \subset [0,2]$$

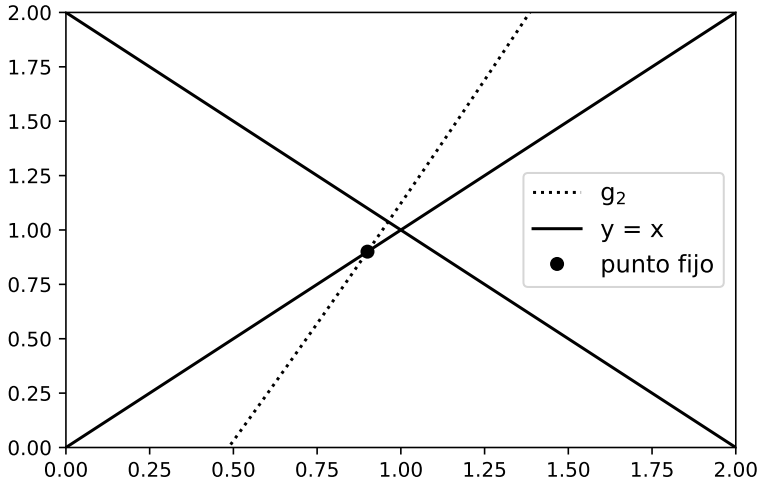
y se cumple también la primera condición. También podemos ver gráficamente que se cumplen las condiciones 1 y 2, porque toda la gráfica de g_1 está dentro de cuadrado $[0,2] \times [0,2]$ y la pendiente de la función es menor que la de las bisectrices del cuadrado, que son 1 y -1, respectivamente:



(d) Empecemos por la **condición 2**.

$$|g_2'(x)| = 2 + \frac{1}{2}\text{sen}\frac{x}{2} \geq 2 \quad \text{en} \quad [0,2]$$

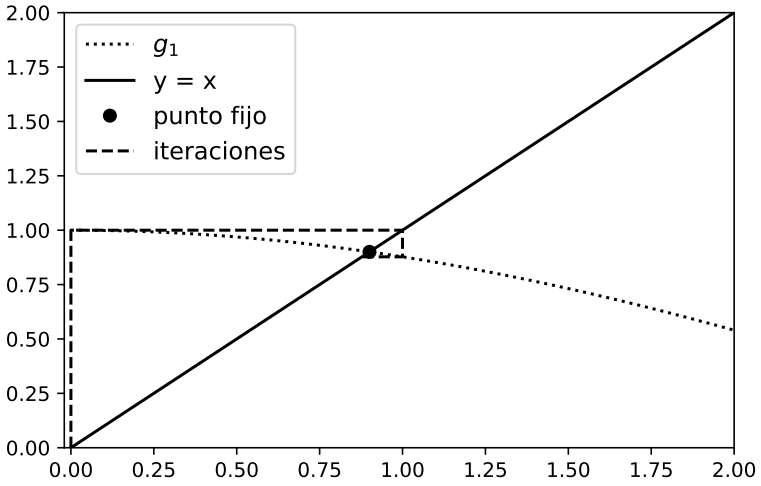
puesto que en $[0,2]$ se verifica que $\text{sen}\dfrac{x}{2}$ es positivo. Y no se cumple la segunda condición, por lo que no cumple las condiciones del teorema de la aplicación contractiva. También podemos ver gráficamente que no se cumplen las condiciones 1 y 2, porque parte de la gráfica de g_2 está fuera de cuadrado $[0,2] \times [0,2]$ y la pendiente de la función es mayor que la de la bisectriz del cuadrado de pendiente 1:



(e) Hagamos iteraciones con la función g_1 partiendo de $x_0 = 0$

k	$x_{k+1} = g_1(x_k)$
0	$x_0 = 0$
1	$x_1 = g_1(x_0) = g_1(0) = \cos(0/2) = 1$
2	$x_2 = g_1(x_1) = g_1(1) = \cos(1/2) = 0{,}8776$

En la siguiente figura vemos gráficamente como el método de punto fijo converge utilizando como función de iteración la función g_1



Problema 2.17:

Sean las funciones

$$f(x) = x - \cos(x) \qquad g_1(x) = \cos(x) \qquad g_2(x) = 2x - \cos(x)$$

(a) Demostrar que la ecuación $f(x) = 0$ tiene la misma raíz que $g_i(x) = x$ con $i = 1, 2$.

(b) Enunciar el teorema de la aplicación contractiva.

(c) Demostrar analíticamente que la función g_1 cumple las condiciones del teorema de la aplicación contractiva el intervalo con $[0, 1]$.

(d) Demostrar analíticamente que la función g_2 no cumple las condiciones del teorema de la aplicación contractiva el intervalo con $[0, 1]$.

(e) Hacer dos iteraciones con g_1 tomando como punto inicial $x_0 = 0$

(a) Demostremos que la ecuación $f(x) = 0$ tiene la misma raíz que $g_i(x) = x$ con $i = 1, 2$.

$$g_1(x) = \cos(x) \Longleftrightarrow x = \cos(x) \Longleftrightarrow x - \cos(x) = 0 \Longleftrightarrow f(x) = x - \cos(x)$$

$$g_2(x) = 2x - \cos(x) \Longleftrightarrow x = 2x - \cos(x) \Longleftrightarrow 0 = x - \cos(x) \Longleftrightarrow f(x) = x - \cos(x)$$

(b) El Teorema de la aplicación contractiva dice: sea g una función derivable definida en el intervalo $[a, b] \subset \mathbb{R}$ y $x_0 \in [a, b]$ un punto del intervalo. Si

1. $x \in [a, b] \quad \Rightarrow \quad g(x) \in [a, b]$
2. $|g'(x)| \leq k < 1$ para todo $x \in [a, b]$ con $k \in \mathbb{R}$

Entonces g tiene un único punto fijo $\alpha \in [a, b]$, y la sucesión x_n definida como $x_{i+1} = f(x_i)$ que tiene como punto inicial x_0 converge a α con orden al menos lineal.

(c) Empecemos por la **condición 2**

$$|g_1'(x)| = |-\sin(x)| = \sin(x)$$

Como $1 < \dfrac{\pi}{2} = 1{,}57$ y como sen(x) es creciente en $\left[0, \dfrac{\pi}{2}\right]$ se verifica que

$$|g_1'(x)| = |-\text{sen}(x)| = \text{sen}(x) \leq \sin(1) = 0{,}84 \quad \text{en} \quad [0, 1]$$

y se cumple la segunda condición 2.

Veamos la **condición 1**. Como el $g_1(x) = \cos(x)$ es decreciente en $[0, 1]$ tendrá su máximo en 0 y su mínimo en 1.

$$x \in [0, 1] \;\Rightarrow\; g_1(1) \leq g(x) \leq g_1(0) \;\Rightarrow\; 0{,}54 \leq g(x) \leq 1 \;\Rightarrow\; g(x) \in [0{,}54, 1] \subset [0, 1]$$

y se cumple también la primera condición.

(d) Empecemos por la **condición 2**.

$$|g_2'(x)| = 2 + \text{sen}(x) \geq 2 \quad \text{en} \quad [0,1]$$

puesto que en $[0,1]$ se verifica que $\text{sen}(x)$ es positivo. Y no se cumple la segunda condición, por lo que no cumple las condiciones del teorema de la aplicación contractiva.

(e) Hacemos dos iteraciones con g_1 tomando como punto inicial $x_0 = 0$

k	$x_{k+1} = g_1(x_k)$
0	$x_0 = 0$
1	$x_1 = g_1(x_0) = g_1(0) = \cos 0 = 1$
2	$x_2 = g_1(x_1) = g_1(1) = \cos 1 = 0{,}54030$

Problema 2.18:

Demostrar que, si tenemos la ecuación

$$f(x) = x^3 - 5 + \ln(1 + x^2) = 0$$

y consideramos sus raíces en el intervalo $[1, 2]$:

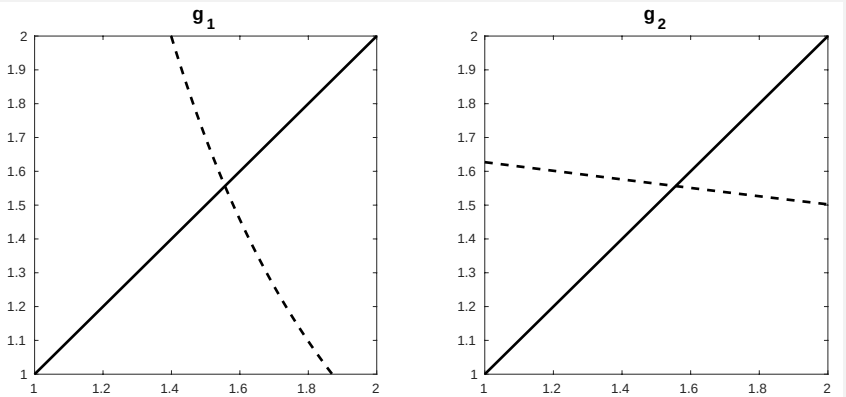
(a) La ecuación $f(x) = 0$ tiene la misma raíz que $g_i(x) = x$ con $i = 1, 2$ siendo

$$g_1(x) = \frac{5 - \ln(1 + x^2)}{x^2} \quad g_2(x) = \sqrt[3]{5 - \ln(1 + x^2)}$$

(b) Enunciar las condiciones del teorema de la aplicación contractiva.

(c) Utilizándolo y teniendo en cuenta las gráficas, escoger, una de las dos funciones para aproximar la solución por el método de iteración de punto fijo comenzando en $x_0 = 1$.

(d) Realizar dos iteraciones con la función escogida.



(a) Queremos demostrar que $f(x) = 0 \iff g_i(x) = x$.
Para la primera función de iteración

$$g_1(x) = x \iff \frac{5 - \ln(1+x^2)}{x^2} = x \iff 5 - \ln(1+x^2) = x^3 \iff$$

$$\iff x^3 - 5 + \ln(1+x^2) = 0 \iff f(x) = 0.$$

Y para la segunda

$$g_2(x) = x \iff \sqrt[3]{5 - \ln(1+x^2)} = x \iff 5 - \ln(1+x^2) = x^3 \iff$$

$$\iff x^3 - 5 + \ln(1+x^2) = 0 \iff f(x) = 0.$$

(b) Sea g una función definida en el intervalo $[a, b] \subset \mathbb{R}$ y $x_0 \in [a, b]$ una aproximación inicial de la iteración de punto fijo dada por $x_{k+1} = g(x_k)$, con $k \geq 0$.

1. $g(x) \in [a, b]$ para todo $x \in [a, b]$,
2. g es diferenciable en $[a, b]$ y existe una constante $k < 1$ tal que $|g'(x)| \leq k$ para todo $x \in [a, b]$.

(c) La función g_1 no cumple las condiciones: no cumple la condición 1 porque la gráfica de g_1 no está contenida totalmente en el intervalo $[1, 2]$. No está entre las rectas $y = 1$ y $y = 2$. Y no cumple la condición 2 porque la gráfica tiene una pendiente mayor en valor absoluto que la recta $y = -x$ que tiene pendiente menos uno.

La función g_2 sí cumple las condiciones: cumple la condición 1 porque la gráfica de g_2 está totalmente contenida en el intervalo $[1, 2]$. Está entre las rectas $y = 1$ y $y = 2$. Y cumple la condición 2 porque la gráfica tiene una pendiente menor en valor absoluto que la recta $y = -x$ que tiene pendiente menos uno.

Por lo tanto escogeríamos la función g_2 para aproximar la solución por el método de iteración de punto fijo.

(d) $x_1 = g_2(x_0) = \sqrt[3]{5 - \ln(1 + x_0^2)} = \sqrt[3]{5 - \ln(1+1)} = 1{,}627$
$x_2 = g_2(x_1) = \sqrt[3]{5 - \ln(1 + x_1^2)} = \sqrt[3]{5 - \ln(1+1{,}627)} = 1{,}54752$
(El punto fijo es $\alpha = 1{,}55634$)

Problema 2.19:

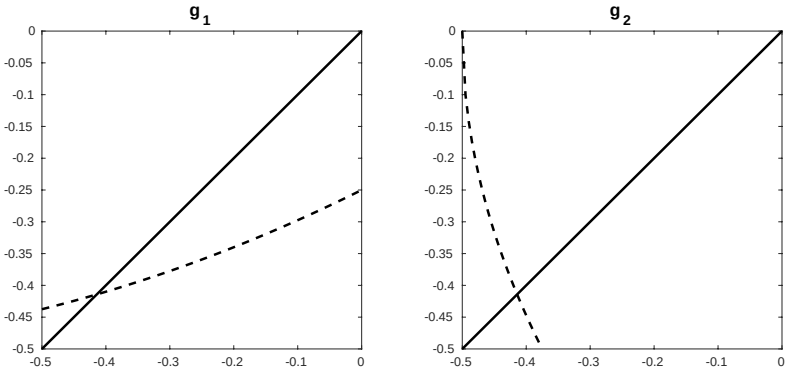
Demostrar que si tenemos la ecuación $f(x) = x^2 - 2x - 1$ y consideramos sus raíces en el intervalo $[-0{,}5, 0]$:

(a) La ecuación $f(x) = 0$ tiene la misma raíz que $g_i(x) = x$ con $i = 1, 2$ siendo

$$g_1(x) = \frac{x^2 + 2x - 1}{4} \quad g_2(x) = -\sqrt{2x + 1}$$

(b) Enunciar las condiciones del teorema de la aplicación contractiva.

(c) Utilizándolo y teniendo en cuenta las gráficas, escoger, una de las dos funciones para aproximar la solución por el método de iteración de punto fijo comenzando en $x_0 = 0$.



(a) Queremos demostrar que $f(x) = 0 \iff g_i(x) = x$.

Para la primera función de iteración

$$g_1(x) = x \iff \frac{x^2 + 2x - 1}{4} = x \iff x^2 + 2x - 1 = 4x \iff$$

$$\iff x^2 - 2x - 1 = 0 \iff f(x) = 0.$$

Y para la segunda

$$g_2(x) = x \iff -\sqrt{2x + 1} = x \iff 2x + 1 = x^2 \iff x^2 - 2x - 1 = 0 \iff f(x) = 0.$$

(b) Sea g una función definida en el intervalo $[a, b] \subset \mathbb{R}$ y $x_0 \in [a, b]$ una aproximación inicial de la iteración de punto fijo dada por $x_{k+1} = g(x_k)$, con $k \geq 0$.

1. $g(x) \in [a, b]$ para todo $x \in [a, b]$,
2. g es diferenciable en $[a, b]$ y existe una constante $k < 1$ tal que $g'(x) \leq k$ para todo $x \in [a, b]$.

(c) La función g_1 sí cumple las condiciones: cumple la condición 1 porque la gráfica de g_1 está totalmente contenida en el intervalo $[-0{,}5, 0]$, entre las rectas $y = -0{,}5$ y

$y = 0$. Y cumple la condición 2 porque la gráfica tiene una pendiente menor que la recta $y = x$ que tiene pendiente uno.
La función g_2 no cumple las condiciones: no cumple la condición 1 porque la gráfica de g_2 no está totalmente contenida en el intervalo $[-0{,}5, 0]$, entre las rectas $y = -0{,}5$ y $y = 0$. Y no cumple la condición 2 porque la gráfica tiene una pendiente mayor en valor absoluto que la recta $y = x$ que tiene pendiente uno.
Por lo tanto escogeríamos la función g_1 para aproximar la solución por el método de iteración de punto fijo.

Problema 2.20:

Sea la ecuación:

$$\frac{1}{x} + x^2 - 2 = 0$$

(a) Demostrar que en $[-2, -1]$ existe una única raíz.

(b) Aproximar la raíz haciendo tres iteraciones utilizando el método de bisección.

(c) Dar una cota del error cometido al calcular esta raíz.

(d) Dada la función

$$f(x) = \frac{1}{x} + x^2 - 2,$$

demostrar que la raíz de $f(x) = 0$ y el punto fijo de $g(x)$ en $[-2, -1]$ es la misma, siendo

$$g(x) = -\sqrt{\frac{2x-1}{x}}.$$

(e) Utilizando g, aproximar la raíz de f. Hacer 3 iteraciones partiendo de $x_0 = -2$.

(a) Las condiciones (suficientes, no necesarias) que ha de cumplir $f(x) = \frac{1}{x} + x^2 - 2$ en $[-2, -1]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es la suma de funciones continuas, el polinomio lo es siempre y $\frac{1}{x}$ es continua en $[-2, -1]$ porque no se anula su denominador.
2. f tiene distinto signo en los extremos del intervalo: $f(-2) = 1{,}5$ y $f(-1) = -2$.
3. $f' > 0$ o $f' < 0$ en $(-2, -1)$:

$$f'(x) = -\frac{1}{x^2} + 2x = \frac{2x^3 - 1}{x^2}.$$

Como x es negativo en $(-2, -1)$, $2x^3$ también y por lo tanto $2x^3 - 1$ es negativo en el intervalo. Como x^2 es siempre positivo $f' < 0$ en $(-2, -1)$.

(b) Hacemos tres iteraciones por bisección. Para ello calculamos el punto medio del intervalo $[a, b]$ usando la fórmula

$$c = \frac{a+b}{2}.$$

En cada iteración escogemos entre $[a, c]$ y $[c, b]$ el nuevo intervalo de forma que se mantenga la condición de que el signo de la función es distinto en los extremos (en negrita).

k	a	$c = (a+b)/2$	b	$f(a)$	$f(m)$	$f(b)$	*cota de error*
1	-2	$-1{,}5$	-1	$\mathbf{1{,}5}$	$\mathbf{-0{,}42}$	-2	0,5
2	-2	$-1{,}75$	$-1{,}5$	1,5	$\mathbf{0{,}49}$	$\mathbf{-0{,}42}$	0,25
3	$-1{,}75$	$-1{,}625$	$-1{,}5$	0,49		$-0{,}42$	0,125

Y podemos dar como raíz aproximada $-1{,}625$ (la raíz verdadera es $-1{,}61803$).

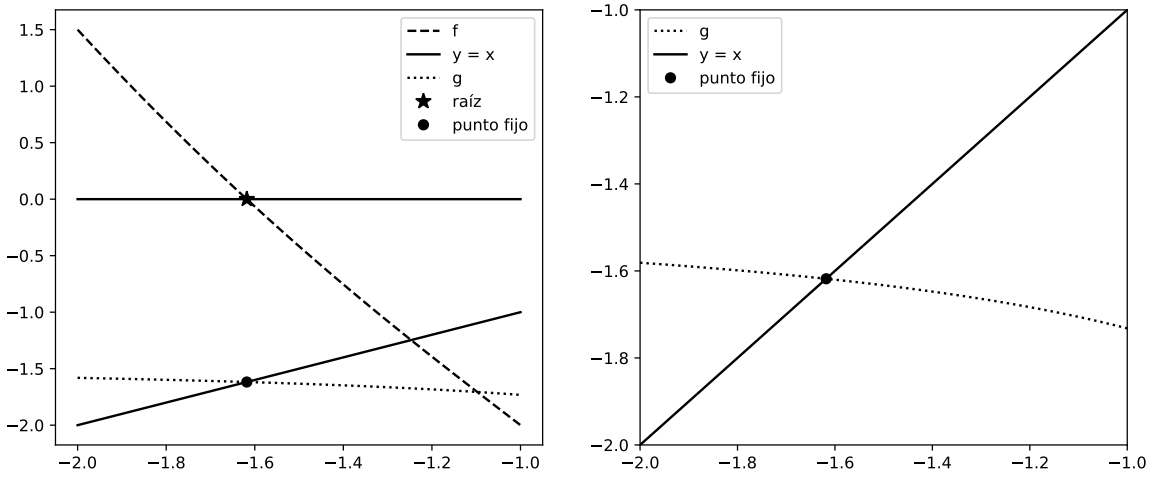
(c) La cota de error es

$$\frac{b_0 - a_0}{2^n} = \frac{-1-(-2)}{2^3} = \frac{1}{8} = 0{,}125$$

(d) Demostremos que la ecuación $x = g(x)$ es equivalente a $f(x) = 0$.

$$f(x) = 0 \Rightarrow \frac{1}{x} + x^2 - 2 = 0 \Rightarrow x^2 = 2 - \frac{1}{x} \Rightarrow x = -\sqrt{\frac{2x-1}{x}}$$

como la raíz que estamos buscando es negativa (está en $[-2, -1]$) tomamos la raíz cuadrada negativa. Si llamamos $g(x) = -\sqrt{\dfrac{2x-1}{x}}$ equivale a $x = g(x)$. Podemos ver, gráficamente, en la siguiente figura, que el punto fijo de g coincide con la raíz de f. Y que la función g cumple las condiciones para ser utilizada como función de iteración para el método de punto fijo utilizando como punto inicial cualquier punto del intervalo $[-2, -1]$.



(e) Hagamos 3 iteraciones

k	$x_{k+1} = g(x_k)$
0	$x_0 = -2$
1	$x_1 = g(x_0) = g(-2) = -1{,}58114$
2	$x_2 = g(x_1) = g(-1{,}58114) = -1{,}62248$
3	$x_3 = g(x_2) = g(-1{,}62248) = -1{,}61751$

Problema 2.21:

Sea la ecuación

$$te^{-\frac{t}{5}} - 1 = 0 \quad (1)$$

(a) Demostrar que en $[1, 2]$ existe una única raíz.

(b) ¿Se puede calcular por el método de bisección partiendo de dicho intervalo?

(c) Realizar tres iteraciones con el método de bisección.

(d) Dar una cota del error cometido al calcular la raíz con tres iteraciones.

(e) Demostrar que la ecuación (1) tiene la misma raíz que $g_i(t) = t$ con $i = 1, 2$ siendo

$$g_1(t) = e^{\frac{t}{5}} \quad g_2(t) = 5 \ln t$$

(f) Enunciar las 2 condiciones del Teorema de la Aplicación Contractiva.

(g) Utilizándolo, escoger una de las dos funciones para aproximar la solución por el método de iteración de punto fijo comenzando en $t_0 = 1$.

(a) Las condiciones (suficientes, no necesarias) que ha de cumplir $f(t) = te^{-\frac{t}{5}} - 1$ en $[1, 2]$ para que exista una única raíz en el intervalo son:

1. f continua: f es continua porque es la suma, producto y composición de funciones continuas. El polinomio lo es siempre y la función exponencial también es continua en todo su dominio.
2. f tiene distinto signo en los extremos del intervalo: $f(1) = -0{,}18$ y $f(2) = 0{,}34$
3. $f' > 0$ o $f' < 0$ en $(1, 2)$:

$$f'(t) = e^{-\frac{t}{5}} - \frac{1}{5}te^{-\frac{t}{5}} = -\frac{1}{5}e^{-\frac{t}{5}}(t - 5).$$

Como $-\frac{1}{5}$ es negativo, $e^{-\frac{t}{5}}$ es positivo y $(t - 5) < 0$ el producto es negativo en $(1, 2)$. Por lo tanto $f' = (-)(+)(-) < 0$ en $(1, 2)$.

(b) Sí, porque se cumplen las condiciones necesarias, que son las condiciones 1 y 2 de la pregunta anterior.

(c) Hagamos tres iteraciones por bisección. Para ello calculamos el punto medio del intervalo $[a, b]$ usando la fórmula

$$c = \frac{a + b}{2}.$$

En cada iteración escogemos entre $[a, c]$ y $[c, b]$ el nuevo intervalo de forma que se mantenga la condición de que el signo de la función es distinto en los extremos (en negrita).

k	a	$c=(a+b)/2$	b	$f(a)$	$f(c)$	$f(b)$	*cota de error*
1	1	1,5	2	**−0,18**	**0,11**	0,34	0,5
2	1	1,25	1,5	−0,18	**−0,026**	**0,11**	0,25
3	1,25	1,375	1,5	−0,026		0,11	0,125

Y podemos dar como raíz aproximada 1,375 (la raíz exacta es 1,29586).

(d) La raíz está en el intervalo [1,25, 1,375] por lo que el máximo error será 1,375−1,25 = 0,125, que podemos dar como cota de error.

(e) Podemos transformar la ecuación

$$f(t)=0 \Longleftrightarrow te^{-\frac{t}{5}}-1=0 \Longleftrightarrow te^{-\frac{t}{5}}=1 \Longleftrightarrow t=e^{\frac{t}{5}} \Longleftrightarrow g_1(t)=t \text{ con } g_1(t)=e^{\frac{t}{5}}.$$

Lo mismo con

$$f(t)=0 \Longleftrightarrow t=e^{\frac{t}{5}} \Longleftrightarrow \ln t=\frac{t}{5} \Longleftrightarrow 5\ln t=t \Longleftrightarrow g_2(t)=t \text{ con } g_2(x)=5\ln t.$$

(f) Sea g una función definida en el intervalo $[a,b]\subset\mathbb{R}$ y $t_0\in[a,b]$ una aproximación inicial de la iteración de punto fijo dada por $t_{k+1}=g(t_k)$, con $k\geq 0$.

1. $g(t)\in[a,b]$ para todo $t\in[a,b]$,
2. g es diferenciable en $[a,b]$ y existe una constante $k<1$ tal que $g'(t)\leq k$ para todo $t\in[a,b]$.

(g) Estudiemos las funciones de iteración.

1. Función g_1: las funciones

$$g_1(t)=e^{\frac{t}{5}} \quad \text{y} \quad g_1'(t)=\frac{1}{5}e^{\frac{t}{5}}$$

son crecientes y para $t\in[1,2]$ el mínimo está en 1 y el máximo en 2

$$g_1(1)<g_1(t)<g_1(2) \Longleftrightarrow 1{,}22<g_1(t)<1{,}49 \Rightarrow g_1(t)\in[1,2]$$

y se cumple la condición 1.

$$g_1'(1)<g_1'(t)<g_1'(2) \Longleftrightarrow 0{,}24<g_1'(t)<0{,}3 \Rightarrow |g_1'(t)|<1$$

y se cumple la condición 2.

2. Función g_2: la función es decreciente y para $t\in[1,2]$ tiene su máximo en 1 y su mínimo en 2

$$|g_2'(t)|=\frac{5}{t}\geq g_2'(2)=\frac{5}{2}>1 \Longleftrightarrow |g_2'(t)|>1$$

y no se cumple la condición 2.

Así que para aproximar el punto fijo escogeríamos la función g_1.

Tema 3

Aproximación de funciones

3.1. Interpolación de Lagrange

Problema 3.1:

Dados los puntos $x_0 = 0$, $x_1 = 1$ y $x_2 = 3$ y la función $f(x) = x^3$:

(a) Calcular los polinomios fundamentales de Lagrange y dibujarlos.

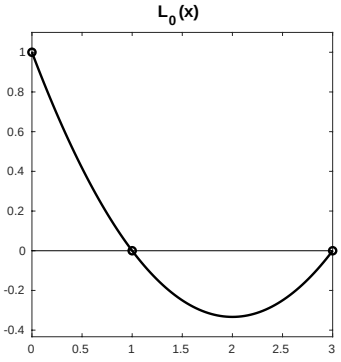
(b) Calcular el polinomio interpolante por el método de Lagrange.

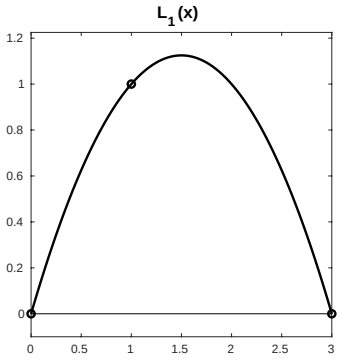
(a) Los polinomios fundamentales de Lagrange correspondientes a los nodos dados son

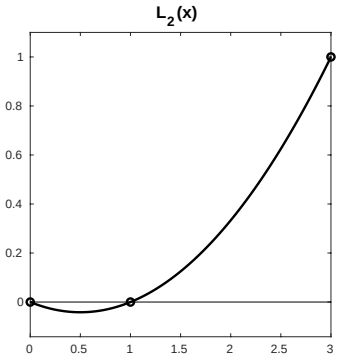
$$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-1)(x-3)}{(0-1)(0-3)} = \frac{1}{3}(x-1)(x-3)$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-0)(x-3)}{(1-0)(1-3)} = -\frac{1}{2}x(x-3)$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-0)(x-1)}{(3-0)(3-1)} = \frac{1}{6}x(x-1)$$







(b) El polinomio interpolante de Lagrange viene dado por

$$P_2(x) = f(x_0)L_0(x) + f(x_1)L_1(x) + f(x_2)L_2(x)$$

$$P_2(x) = 0 \cdot L_0(x) + 1 \cdot L_1(x) + 27 \cdot L_2(x)$$

$$P_2(x) = -\frac{1}{2}x\,(x-3) + 27\frac{1}{6}x\,(x-1) = x\left(-\frac{1}{2}x + \frac{3}{2} + \frac{9}{2}x - \frac{9}{2}\right)$$

$$P_2(x) = x(4x-3)$$

Problema 3.2:

Dados los puntos $x_0 = -1$, $x_1 = 0$ y $x_2 = 1$ y la función $f(x) = x^5 - x^4$:

(a) Calcular los polinomios fundamentales de Lagrange y dibujarlos.

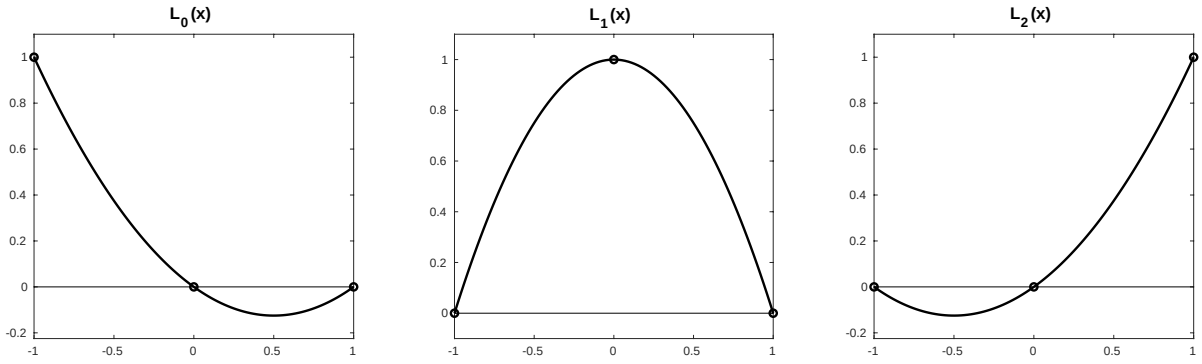
(b) Calcular el polinomio interpolante por el método de Lagrange.

(a) Los polinomios fundamentales de Lagrange correspondientes a los nodos dados son

$$L_0(x) = \frac{(x-x_1)\,(x-x_2)}{(x_0-x_1)\,(x_0-x_2)} = \frac{(x-0)\,(x-1)}{(-1-0)\,(-1-1)} = \frac{1}{2}x\,(x-1)$$

$$L_1(x) = \frac{(x-x_0)\,(x-x_2)}{(x_1-x_0)\,(x_1-x_2)} = \frac{(x-(-1))\,(x-1)}{(0-(-1))\,(0-1)} = -\,(x+1)\,(x-1)$$

$$L_2(x) = \frac{(x-x_0)\,(x-x_1)}{(x_2-x_0)\,(x_2-x_1)} = \frac{(x-(-1))\,(x-0)}{(1-(-1))\,(1-0)} = \frac{1}{2}\,(x+1)\,x$$



(b) El polinomio interpolante de Lagrange viene dado por

$$P_2(x) = f(x_0)L_0(x) + f(x_1)L_1(x) + f(x_2)L_2(x)$$

$$P_2(x) = (-2) \cdot L_0(x) + 0 \cdot L_1(x) + 0 \cdot L_2(x)$$

$$P_2(x) = -\,2 \cdot L_0(x) = -\frac{2}{2}x\,(x-1)$$

$$P_2(x) = x - x^2$$

Problema 3.3:

Dados los puntos

k	0	1	2	3
x_k	-1	0	1	2
y_k	3	0	-1	0

(a) Calcular los polinomios fundamentales de Lagrange y dibujarlos.

(b) Calcular el polinomio interpolante por el método de Lagrange.

(c) Construir la tabla de diferencias divididas de Newton.

(d) Usando la tabla anterior, calcular el polinomio interpolante.

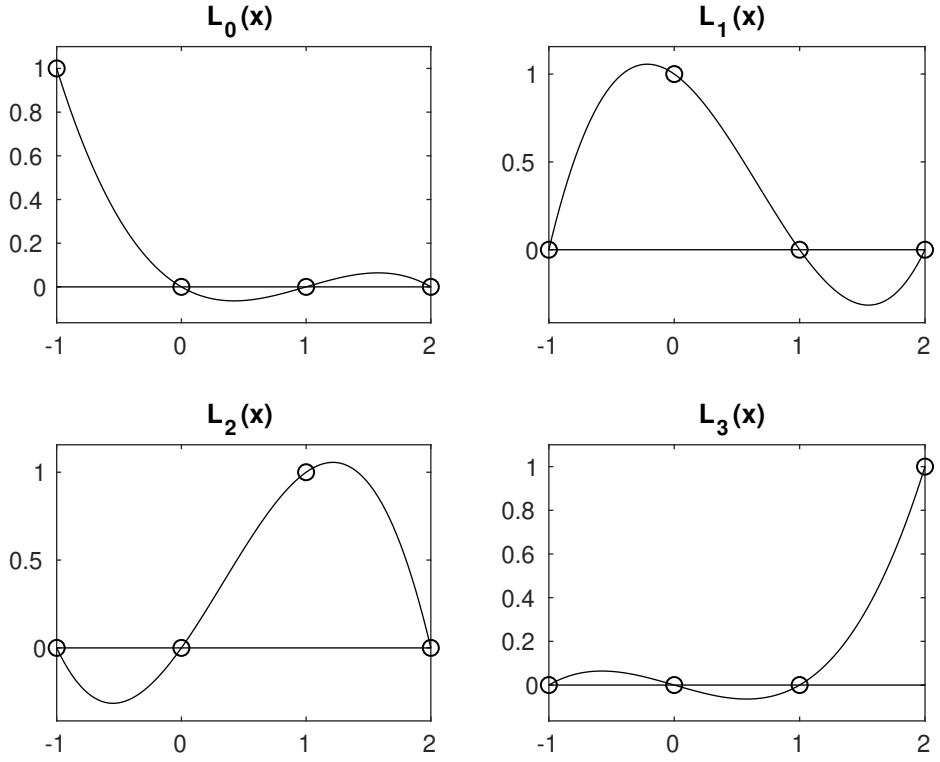
(a)

$$L_0(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = \frac{(x-0)(x-1)(x-2)}{(-1-0)(-1-1)(-1-2)} = -\frac{1}{6}x(x-1)(x-2)$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{(x-(-1))(x-1)(x-2)}{(0-(-1))(0-1)(0-2)} = \frac{1}{2}(x+1)(x-1)(x-2)$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = \frac{(x-(-1))(x-0)(x-2)}{(1-(-1))(1-0)(1-2)} = -\frac{1}{2}(x+1)x(x-2)$$

$$L_3(x) = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{(x-(-1))(x-0)(x-1)}{(2-(-1))(2-0)(2-1)} = \frac{1}{6}(x+1)x(x-3)$$



(b)

$$P_3(x) = f(x_1)L_1(x) + f(x_2)L_2(x) + f(x_3)L_3(x) + f(x_4)L_4(x)$$

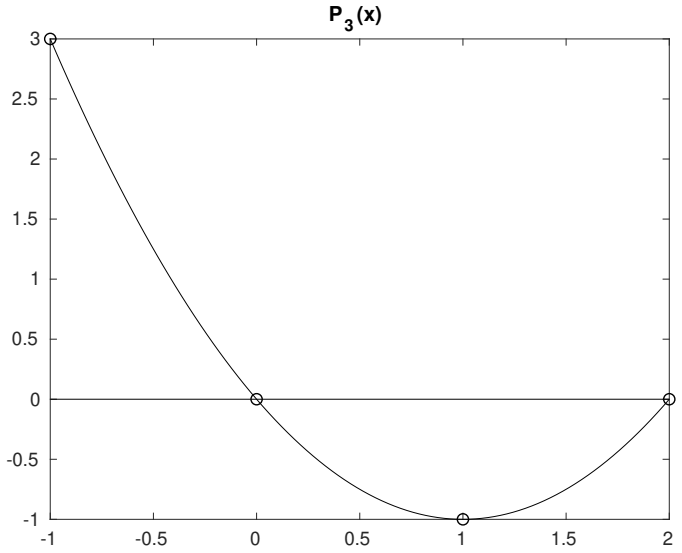
$$P_3(x) = (3) \cdot L_1(x) + (0) \cdot L_2(x) + (-1) \cdot L_3(x) + (2) \cdot L_4(x)$$

$$P_3(x) = \frac{1}{2}(x-2)x(x+1) - \frac{1}{2}(x-2)(x-1)x$$

$$P_3(x) = (x-2)x\left(\frac{1-x}{2} + \frac{x+1}{2}\right)$$

$$P_3(x) = (x-2)x(1)$$

$$P_3(x) = x^2 - 2x$$



(c) En la tabla de diferencias divididas de Newton los numeradores se forman a partir de la columna anterior y los denominadores a partir de la primera columna. En este caso

x	$f(x)$			
-1	3			
		$\frac{0-3}{0-(-1)} = -3$		
0	0		$\frac{-1-(-3)}{1-(-1)} = 1$	
		$\frac{-1-0}{1-0} = -1$		$\frac{1-1}{2-(-1)} = 0$
1	-1		$\frac{1-(-1)}{2-0} = 1$	
		$\frac{0-(-1)}{2-1} = 1$		
2	0			

Para construir el polinomio interpolante necesitamos el primer elemento de cada

columna

$$f[x_0] = 3, \quad f[x_0, x_1] = -3, \quad f[x_0, x_1, x_2] = 1, \quad f[x_0, x_1, x_2, x_3] = 0$$

(d) Construímos el polinomio interpolante en la forma de Newton

$$\begin{aligned} P_3(x) =& f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \\ &+ f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2) \end{aligned}$$

que es

$$\begin{aligned} P_3(x) =& 3 + (-3)(x - (-1)) + (1)(x - (-1))(x - 0) + \\ &+ (0)(x - (-1))(x - 0)(x - 1) \end{aligned}$$

Por lo tanto

$$P_3(x) = 3 - 3(x + 1) + (x + 1)x = 3 - 3x - 3 + x^2 + x$$

Que es

$$P_3(x) = x^2 - 2x$$

que coincide con el obtenido con los polinomios fundamentales de Lagrange, como era de esperar.

Problema 3.4:

Tenemos los siguientes datos del movimiento de un cuerpo

t(s)	0	2	4	6	8	10
v(m/s)	0	6	20	42	72	110

(a) ¿Cuál será la velocidad a los 6,2 s? Usar interpolación polinomial lineal con el polinomio interpolante en la forma de Newton.

(b) ¿Y la aceleración? Utilizar derivación numérica con $h = 0{,}2$.

(a) Para interpolar con un polinomio de grado uno, necesitamos dos puntos. Los t más próximos a 6,2 son 6 y 8. Además, el intervalo $[6, 8]$ contiene a 6,2.
Para usar la notación habitual $x = t$ y $f(x) = v(t)$. Construimos la tabla de diferencias divididas con $x_0 = 6$ y $x_1 = 8$:

x	$f(x)$	
6	42	
		$\frac{72 - 42}{8 - 6} = 15$
8	72	

Y tenemos

$$f[x_0] = 42, \quad f[x_0, x_1] = 15$$

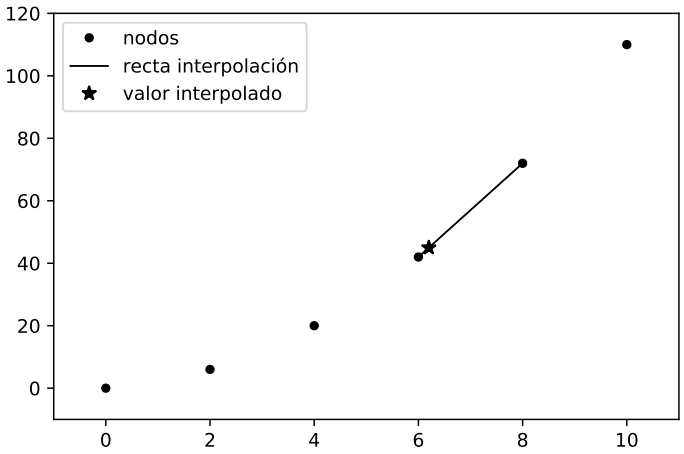
Construímos el polinomio interpolante en la forma de Newton

$$P_2(x) = f[x_0] + f[x_0, x_1](x - x_0)$$
$$P_2(x) = 42 + 15(x - 6)$$

y entonces

$$v(6{,}2) \approx P_1(6{,}2) = 42 + 15(6{,}2 - 6) = 45\text{m/s}$$

$$v(6{,}2) \approx 45\text{m/s}$$



(b) Podemos aproximar la derivada de una función usando la fórmula centrada:

$$f'(t) \approx \frac{1}{2h}(f(t+h) - f(t-h))$$

Si $f = P_1$, $t = 6{,}2$ y $h = 0{,}2$

$$a(6{,}2) \approx \frac{1}{2h}(P_1(6{,}2 + 0{,}2) - P_1(6{,}2 - 0{,}2)) = \frac{1}{2(0{,}2)}(48 - 42) = 15\,\text{m/s}^2$$

Problema 3.5:

Tenemos los siguientes datos del movimiento de un cuerpo

t(s)	0	1	2	3	4	5
v(m/s)	0	14	32	48	56	50

(a) ¿Cuál será la velocidad a los 4,5 s? Usar interpolación polinomial cuadrática con el polinomio interpolante en la forma de Newton.

(b) ¿Y la aceleración? Utilizar derivación numérica con $h = 0{,}2$.

(a) Para interpolar con un polinomio de grado dos, necesitamos tres puntos. Los t más próximos a 4,5 son 3, 4 y 5. Además el intervalo $[3, 5]$ contiene a 4,5.

Para usar la notación habitual $x = t$ y $f(x) = v(t)$. Construimos la tabla de diferencias divididas con $x_0 = 3$, $x_1 = 4$ y $x_2 = 5$:

$$\begin{array}{cc} x & f(x) \\ 3 & 48 \\ & \frac{56-48}{4-3} = 8 \\ 4 & 56 \qquad\qquad \frac{(-6)-8}{5-3} = -7 \\ & \frac{50-56}{5-4} = -6 \\ 5 & 50 \end{array}$$

Y tenemos

$$f\left[x_0\right] = 48, \quad f\left[x_0, x_1\right] = 8, \quad f\left[x_0, x_1, x_2\right] = -7$$

Construímos el polinomio interpolante en la forma de Newton

$$\begin{aligned} P_2\left(x\right) =& f\left[x_0\right] + f\left[x_0, x_1\right]\left(x - x_0\right) + f\left[x_0, x_1, x_2\right]\left(x - x_0\right)\left(x - x_1\right) \\ P_2\left(x\right) =& 48 + 8\left(x - 3\right) + \left(-7\right)\left(x - 3\right)\left(x - 4\right) \end{aligned}$$

y entonces

$$v(4{,}5) \approx P_2\left(4{,}5\right) = 48 + 8\left(4{,}5 - 3\right) + \left(-7\right)\left(4{,}5 - 3\right)\left(4{,}5 - 4\right) = 54{,}75\,\mathrm{m/s}$$

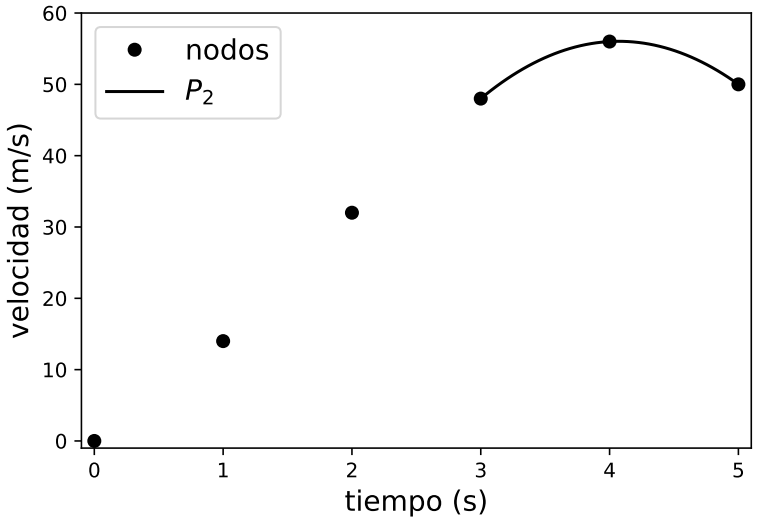
(b) Podemos aproximar la derivada de una función usando la fórmula centrada:

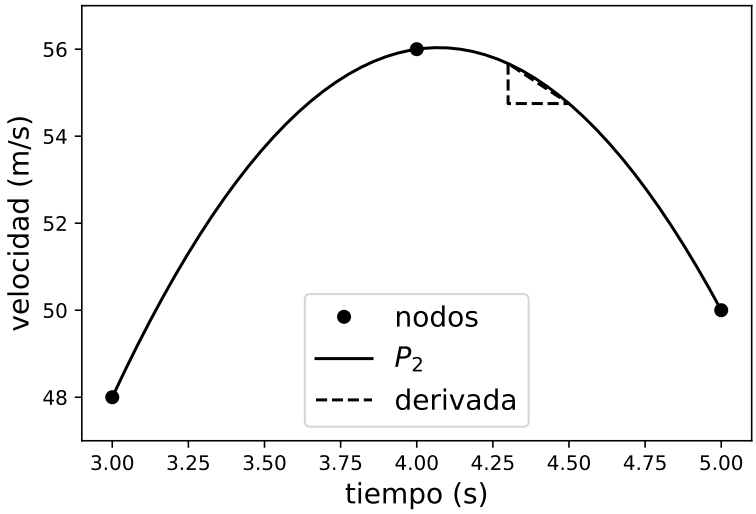
$$f'(t) \approx \frac{1}{2h}(f(t+h) - f(t-h))$$

Si $f = P_2$, $t_0 = 4{,}5$ y $h = 0{,}2$

$$a(t_0) \approx \frac{1}{2h}(P_2(t_0+h) - P_2(t_0-h))$$

$$a(4{,}5) \approx \frac{1}{2(0{,}2)}(P_2(4{,}5+0{,}2) - P_2(4{,}5+0{,}2)) = \frac{1}{2(0{,}2)}(53{,}27 - 55{,}67) = -6\,\mathrm{m/s^2}$$





Problema 3.6:

Tenemos los siguientes datos del movimiento de un cuerpo

t(s)	0	1	2	3	4	5
v(m/s)	0	14	32	48	56	50

(a) ¿Cuál será la velocidad a los 1,5 s? Usar interpolación polinomial de grado 3 con el polinomio interpolante en la forma de Newton.

(b) ¿Y la aceleración? Utilizar derivación numérica con $h = 0{,}1$.

(a) Para interpolar con un polinomio de grado tres, necesitamos cuatro puntos. Los t más próximos a 1,5 son 0, 1, 2 y 3. Además el intervalo $[3, 5]$ contiene 1,5.

Para usar la notación habitual $x = t$ y $f(x) = v(t)$. Construimos la tabla de diferencias divididas con $x_0 = 3$, $x_1 = 4$ y $x_2 = 5$:

$$\begin{array}{cc} x & f(x) \\ 0 & 0 \\ & \frac{14-0}{1-0} = 14 \\ 1 & 14 \qquad\qquad \frac{18-14}{2-0} = 2 \\ & \frac{32-14}{2-1} = 18 \qquad\qquad \frac{-1-2}{3-0} = -1 \\ 2 & 32 \qquad\qquad \frac{16-18}{3-1} = -1 \\ & \frac{48-32}{3-2} = 16 \\ 3 & 48 \end{array}$$

Y tenemos

$$f[x_0] = 0, \quad f[x_0, x_1] = 14, \quad f[x_0, x_1, x_2] = 2, \quad f[x_0, x_1, x_2, x_3] = -1$$

Construímos el polinomio interpolante en la forma de Newton

$$P_3(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) +$$

$$+f\left[x_0, x_1, x_2, x_3\right](x-x_0)(x-x_1)(x-x_2)$$

$$P_3(x) = 0 + 14(x-0) + 2(x-0)(x-1) + (-1)(x-0)(x-1)(x-2)$$

y entonces

$$v(1{,}5) \approx P_3(1{,}5) = 0 + 14(1{,}5) + 2(1{,}5)(0{,}5) + (-1)(1{,}5)(0{,}5)(-0{,}5) = 22{,}875\,\mathrm{m/s}$$

$$v(1{,}5) \approx 22{,}875\,\mathrm{m/s}$$

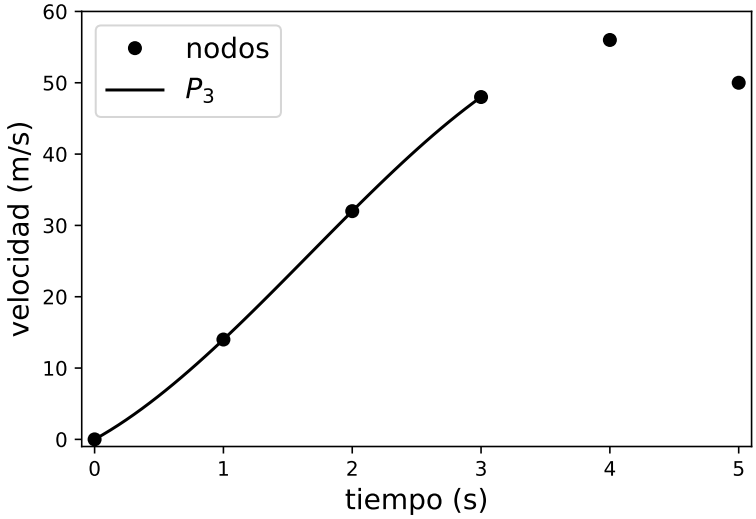
(b) Podemos aproximar la derivada de una función usando la fórmula centrada:

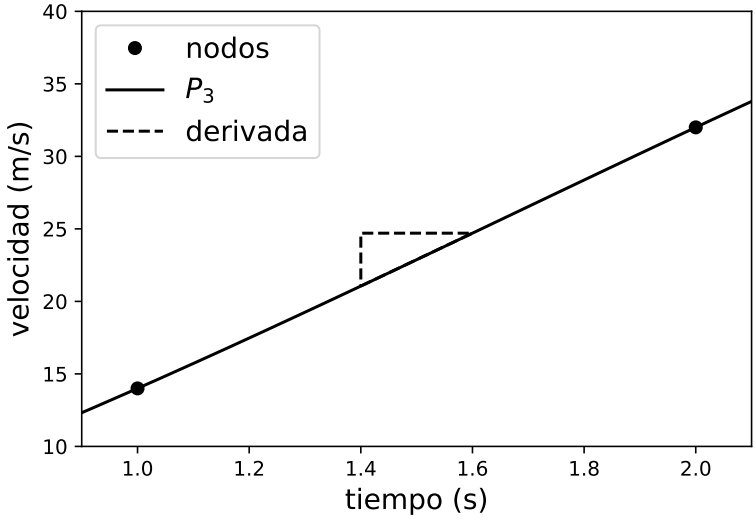
$$f'(t) \approx \frac{1}{2h}(f(t+h) - f(t-h))$$

Si $f = P_3$, $t_0 = 1{,}5$ y $h = 0{,}1$

$$a(4{,}5) \approx \frac{1}{2h}(P_3(t_0+h) - P_3(t_0+h)) =$$

$$= \frac{1}{2(0{,}1)}(P_3(1{,}5+0{,}1) - P_3(1{,}5-0{,}1)) = \frac{1}{2(0{,}1)}(24{,}7 - 21{,}1) = 18\,\mathrm{m/s^2}$$





3.2. Interpolación a trozos

Problema 3.7:

De una tabla de logaritmos neperianos, extraemos

x	1	2	3	4	5
$\ln x$	0,0000	0,6931	1,0986	1,3863	1,6094

Usando interpolación lineal a trozos, estimar el valor de ln(3,4). Para interpolar, utilizar el método de Newton. Redondear a cuatro cifras decimales.

Para calcular el valor aproximado por interpolación lineal en 3,4 usaremos los dos valores más cercanos que son 3 y 4 con sus correspondientes logaritmos. Construimos la tabla de diferencias divididas

$$\begin{array}{ll} x & f(x) \\ 3 & 1{,}0986 \\ & \qquad \dfrac{1{,}3863 - 1{,}0986}{4-3} = 0{,}2877 \\ 4 & 1{,}3863 \end{array}$$

Ya tenemos los coeficientes del polinomio

$$f[x_0] = 1{,}0986, \quad f[x_0, x_1] = 0{,}2877.$$

Construímos el polinomio interpolante en la forma de Newton

$$\begin{aligned} P_1(x) &= f[x_0] + f[x_0, x_1](x - x_0) \\ P_1(x) &= 1{,}0986 + 0{,}2877(x-3) \\ P_1(3{,}4) &= 1{,}0986 + 0{,}2877(3{,}4-3) = 1{,}0986 + 0{,}2877(0{,}4) = 1{,}2137 \end{aligned}$$

Problema 3.8:

Consideremos la función $f(x) = x\ln x - x$ y su polinomio interpolador con los nodos x_0 y x_1.

(a) Demostrar que el error cometido al aproximar $f(x)$ mediante tal polinomio en cualquier punto de $[x_0, x_1]$ está acotado por

$$\frac{(x_1 - x_0)^2}{8x_0}.$$

(b) Construir el polinomio interpolante, utilizando el método de Newton para $x_0 = 1$ y $x_1 = 2$ y dar una cota del error cometido.

(a) El error de interpolación viene dado por

$$E(x) = f(x) - P_n(x) = f^{(n+1)}(c)\frac{(x-x_0)\dots(x-x_n)}{(n+1)!},$$

donde las x_i son los puntos de interpolación, c un punto del intervalo de interpolación, f la función a interpolar y P_n el polinomio de interpolación obtenido con los puntos de interpolación. En este caso, como tenemos dos nodos de interpolación, la interpolación es lineal y el error es

$$|E(x)| = |f(x) - P_1(x)| = \left|f^{(2)}(c)\right| \frac{|(x-x_0)(x-x_1)|}{2!}.$$

Por una parte, suponiendo $x_0 < x_1$ tenemos

$$f'(x) = \ln x + x\frac{1}{x} - 1 = \ln x \quad f^{(2)}(x) = \frac{1}{x}$$

Como $c \in [x_0, x_1]$, se tiene que $x_0 < c$. Y por lo tanto $\dfrac{1}{c} < \dfrac{1}{x_0}$. Y entonces tenemos

$$\left|f^{(2)}(c)\right| = \frac{1}{c} < \frac{1}{x_0}$$

Por otra parte

$$g(x) = |(x-x_0)(x-x_1)| = (x-x_0)(x_1-x)$$

y

$$g'(x) = (x_1-x) + (x-x_0)(-1) = -2x + x_0 + x_1 = 0$$

con lo que el extremo está en

$$x = \frac{x_0+x_1}{2}.$$

Y como $g''(x) = -2$ en este punto tenemos un máximo. El valor de la función g en este máximo es

$$g\left(\frac{x_0+x_1}{2}\right) = \left(\frac{x_0+x_1}{2} - x_0\right)\left(x_1 - \frac{x_0+x_1}{2}\right) = \left(\frac{x_1-x_0}{2}\right)\left(\frac{x_1-x_0}{2}\right) = \frac{1}{4}(x_1-x_0)^2.$$

Por lo tanto se verifica que

$$|E(x)| = |f(x) - P_1(x)| = \left|f^{(2)}(c)\right| \frac{|(x-x_0)(x-x_1)|}{2!} < \frac{1}{x_0}\frac{1}{4}(x_1-x_0)^2\frac{1}{2!} = \frac{(x_1-x_0)^2}{8x_0}.$$

Es decir

$$|E(x)| < \frac{(x_1-x_0)^2}{8x_0}. \tag{3.1}$$

(b) Para construir el polinomio

x	$f(x)$	
1	-1	
		$\dfrac{2\ln 2 - 2 - (-1)}{2-1} = 2\ln 2 - 1$
2	$2\ln 2 - 2$	

Ya tenemos los coeficientes del polinomio

$$f[x_0] = -1, \quad f[x_0, x_1] = 2\ln 2 - 1.$$

Construímos el polinomio interpolante en la forma de Newton

$$\begin{aligned} P_1(x) =& f[x_0] + f[x_0, x_1](x - x_0) \\ P_1(x) =& -1 + (2\ln 2 - 1)(x - 1) \end{aligned}$$

Y, teniendo en cuenta (3.1) la cota del error es

$$\frac{(x_1 - x_0)^2}{8x_0} = \frac{(1-0)^2}{8(1)} = \frac{1}{8} = 0{,}125$$

Problema 3.9:

Consideremos la función $f(x) = \dfrac{1}{x}$ y su polinomio interpolador con los nodos x_0 y x_1.

(a) Demostrar que el error cometido al aproximar $f(x)$ mediante tal polinomio en cualquier punto de $[x_0, x_1]$ está acotado por $\dfrac{(x_1 - x_0)^2}{4x_0^3}$.

(b) Construir el polinomio interpolante, utilizando el método de Newton para $x_0 = 1$ y $x_1 = 2$ y dar una cota del error cometido.

(a) El error de interpolación viene dado por

$$E(x) = f(x) - P_n(x) = f^{(n+1)}(c)\frac{(x - x_0)\ldots(x - x_n)}{(n+1)!},$$

donde las x_i son los puntos de interpolación, c un punto del intervalo de interpolación, f la función a interpolar y P_n el polinomio de interpolación obtenido con los puntos de interpolación. En este caso, como tenemos dos nodos de interpolación, la interpolación es lineal y el error es

$$|E(x)| = |f(x) - P_1(x)| = \left|f^{(2)}(c)\right| \frac{|(x - x_0)(x - x_1)|}{2!}.$$

Por una parte, suponiendo $x_0 < x_1$, y como $c \in (x_0, x_1)$ entonces $x_0 < c$ y tenemos

$$\left|f^{(2)}(c)\right| = \frac{2}{c^3} < \frac{2}{x_0^3}$$

Por otra parte

$$g(x) = |(x - x_0)(x - x_1)| = (x - x_0)(x_1 - x)$$

y

$$g'(x) = (x_1 - x) + (x - x_0)(-1) = -2x + x_0 + x_1 = 0$$

con lo que el extremo está en

$$x = \frac{x_0 + x_1}{2}.$$

Y como $g''(x) = -2$ en este punto tenemos un máximo. El valor de la función g en este máximo es

$$g\left(\frac{x_0+x_1}{2}\right) = \left(\frac{x_0+x_1}{2} - x_0\right)\left(x_1 - \frac{x_0+x_1}{2}\right) = \left(\frac{x_1-x_0}{2}\right)\left(\frac{x_1-x_0}{2}\right) = \frac{1}{4}(x_1-x_0)^2.$$

Por lo tanto se verifica que

$$|E(x)| = |f(x) - P_1(x)| = \left|f^{(2)}(c)\right| \frac{|(x-x_0)(x-x_1)|}{2!} < \frac{2}{x_0^3}\frac{1}{4}(x_1-x_0)^2\frac{1}{2!} = \frac{(x_1-x_0)^2}{4x_0^3}$$

y

$$|E(x)| < \frac{(x_1-x_0)^2}{4x_0^3} \tag{3.2}$$

(b) Para construir el polinomio

x	$f(x)$	
1	1	
		$\frac{1/2-1}{2-1} = -1/2$
2	1/2	

Ya tenemos los coeficientes del polinomio

$$f[x_0] = 1, \quad f[x_0, x_1] = -\frac{1}{2}.$$

Construímos el polinomio interpolante en la forma de Newton

$$\begin{aligned} P_1(x) &= f[x_0] + f[x_0, x_1](x - x_0) \\ P_1(x) &= 1 + \left(-\frac{1}{2}\right)(x-1) \\ P_1(x) &= \frac{3-x}{2} \end{aligned}$$

Y, teniendo en cuenta (3.2) la cota del error es

$$\frac{(x_1-x_0)^2}{4x_0^3} = \frac{(2-1)^2}{4(1)^3} = \frac{1}{4} = 0{,}25$$

Problema 3.10:

Consideremos la función $f(x) = \ln(x)$ y su polinomio interpolador con los nodos x_0 y x_1.

(a) Demostrar que el error cometido al aproximar $f(x)$ mediante tal polinomio en cualquier punto de $[x_0, x_1]$ está acotado por $\dfrac{(x_1 - x_0)^2}{8x_0^2}$.

(b) Construir el polinomio interpolante, utilizando el método de Newton para $x_0 = 1$ y $x_1 = 2$ y dar una cota del error cometido.

(c) Construir el polinomio interpolante, utilizando el método de Newton para $x_0 = 2$ y $x_1 = 4$ y dar una cota del error cometido.

(d) Se desea tabular $f(x)$ para ser capaces de obtener, utilizando interpolación lineal en dos puntos consecutivos, cualquier valor de $f(x)$ con un error menor que 10^{-2}. Calcular el número de subintervalos necesarios, considerando los puntos igualmente espaciados, cuando $[x_0, x_n] = [1, 100]$.

(a) El error de interpolación viene dado por

$$E(x) = f(x) - P_n(x) = f^{(n+1)}(c)\frac{(x - x_0)\dots(x - x_n)}{(n+1)!},$$

donde las x_i son los puntos de interpolación, c un punto del intervalo de interpolación, f la función a interpolar y P_n el polinomio de interpolación obtenido con los puntos de interpolación. En este caso, como tenemos dos nodos de interpolación, la interpolación es lineal y el error es

$$|E(x)| = |f(x) - P_1(x)| = \left|f^{(2)}(c)\right| \frac{|(x - x_0)(x - x_1)|}{2!}.$$

Por una parte, suponiendo $x_0 < x_1$, y como $c \in (x_0, x_1)$ se verifica que $x_0 < c$ y tenemos

$$\left|f^{(2)}(c)\right| = \frac{1}{c^2} < \frac{1}{x_0^2}$$

Por otra parte

$$g(x) = |(x - x_0)(x - x_1)| = (x - x_0)(x_1 - x)$$

y

$$g'(x) = (x_1 - x) + (x - x_0)(-1) = -2x + x_0 + x_1 = 0$$

con lo que el extremo está en

$$x = \frac{x_0 + x_1}{2}.$$

Y como $g''(x) = -2$ en este punto tenemos un máximo. El valor de la función g en este máximo es

$$g\left(\frac{x_0 + x_1}{2}\right) = \left(\frac{x_0 + x_1}{2} - x_0\right)\left(x_1 - \frac{x_0 + x_1}{2}\right) = \left(\frac{x_1 - x_0}{2}\right)\left(\frac{x_1 - x_0}{2}\right) = \frac{1}{4}(x_1 - x_0)^2.$$

Por lo tanto se verifica que

$$|E(x)| = |f(x) - P_1(x)| = \left|f^{(2)}(c)\right| \frac{|(x-x_0)(x-x_1)|}{2!} < \frac{1}{x_0^2}\frac{1}{4}(x_1-x_0)^2\frac{1}{2!} = \frac{(x_1-x_0)^2}{8x_0^2}.$$

Es decir,

$$|E(x)| < \frac{(x_1-x_0)^2}{8x_0^2} \tag{3.3}$$

(b) Para construir el polinomio

x	$f(x)$	
1	0	
		$\frac{\ln 2 - 0}{2-1} = \ln 2$
2	$\ln 2$	

Ya tenemos los coeficientes del polinomio

$$f[x_0] = 0, \quad f[x_0, x_1] = \ln 2.$$

Construímos el polinomio interpolante en la forma de Newton

$$\begin{aligned} P_1(x) &= f[x_0] + f[x_0, x_1](x - x_0) \\ P_1(x) &= 0 + \ln 2\,(x-1) \\ P_1(x) &= \ln 2\,(x-1) \end{aligned}$$

Y, teniendo en cuenta (3.3) la cota del error es

$$\frac{(x_1-x_0)^2}{8x_0^2} = \frac{(2-1)^2}{8(1)^2} = \frac{1}{8} = 0{,}125$$

(c) Para construir el polinomio

x	$f(x)$	
2	$\ln 2$	
		$\frac{\ln 4 - \ln 2}{4-2} = \frac{1}{2}(\ln 4 - \ln 2) = \frac{1}{2}\ln\frac{4}{2} = \frac{\ln 2}{2}$
3	$\ln 3$	

Ya tenemos

$$f[x_0] = \ln 2, \quad f[x_0, x_1] = \frac{\ln 2}{2}.$$

Construímos el polinomio interpolante en la forma de Newton

$$\begin{aligned} P_1(x) &= f[x_0] + f[x_0, x_1](x - x_0) \\ P_1(x) &= \ln 2 + \frac{\ln 2}{2}(x-2) \\ P_1(x) &= \frac{\ln 2}{2}x \end{aligned}$$

Y la cota del error es

$$\frac{(x_1-x_0)^2}{8x_0^2} = \frac{(3-2)^2}{8(2)^2} = \frac{1}{32} = 0{,}031$$

(d) Si dividimos el intervalo $[1, 100]$ en n intervalos iguales de longitud h tendremos que

$$h = \frac{100 - 1}{n} = \frac{99}{n} \tag{3.4}$$

Los extremos de los intervalos serán

$$x_i = 1 + i\,h \quad i = 0, \ldots, n.$$

Y los intervalos serán de la forma

$$[x_i, x_{i+1}] \quad i = 0, \ldots, n-1.$$

Teniendo en cuenta (3.3) para cada uno de estos intervalo la cota de error será

$$|E(x)| < \frac{(x_{i+1} - x_i)^2}{8x_i^2}.$$

Como $h = x_{i+1} - x_i$ es la longitud de todos los intervalos

$$|E(x)| < \frac{h^2}{8x_i^2}.$$

Y por la fórmula (3.4)

$$|E(x)| < \frac{(99/n)^2}{8x_i^2}.$$

Por otra parte, dado que todos los x_i están en el intervalo $[1, 100]$ el menor valor que puede tomar x_i es 1 y podemos decir

$$|E(x)| < \frac{(99/n)^2}{8x_i^2} < \frac{(99/n)^2}{8(1)^2} = \frac{99^2}{8n^2}.$$

Si hacemos

$$\frac{99^2}{8n^2} < 10^{-2} \tag{3.5}$$

entonces

$$|E(x)| < 10^{-2}$$

que es lo que estamos buscando. Veamos cuantos intervalos necesitamos como mínimo para que se cumpla (3.5).

$$\frac{99^2}{8n^2} < 10^{-2} \iff \frac{99^2}{(8)(10^{-2})} < n^2 \iff \sqrt{\frac{99^2 10^2}{8}} < n \iff 350{,}02 < n$$

y si dividimos el intervalo $[1, 100]$ en $n = 351$ subintervalos iguales y realizamos interpolación lineal en cada uno de ellos podemos garantizar que el error de interpolación $|E(x)| < 10^{-2}$.

Problema 3.11:

Dados los nodos:

x	-1	0	1
y	0	1	0

Si escribimos el spline natural que los ajusta como

$$s(x) = \begin{cases} s_1(x) = ax^3 + bx^2 + cx + d & \text{si } x \in [-1, 0] \\ s_2(x) = ex^3 + fx^2 + gx + h & \text{si } x \in [0, 1] \end{cases}$$

plantear las ecuaciones del sistema cuya solución son los coeficientes a, b, c, d, e, f, g y h. Especificar las condiciones aplicadas.

Se tienen que cumplir las siguientes condiciones:

(a) La curva ha de pasar por los tres puntos. Por lo tanto:

$$s_1(-1) = 0,\ s_1(0) = 1,\ s_2(0) = 1,\ s_2(1) = 0.$$

(b) Han de coincidir las derivadas primera y segunda en los puntos intermedios:

$$s_1'(0) = s_2'(0),\ s_1''(0) = s_2''(0).$$

(c) Y como hay 8 incógnitas y de momento solo tenemos 6 condiciones (ecuaciones), imponemos dos más en los extremos. Como el spline es natural las condiciones adicionales son:

$$s_1''(-1) = 0,\ s_2''(1) = 0.$$

Calculamos

$$s'(x) = \begin{cases} s_1'(x) = 3ax^2 + 2bx + c & \text{si } x \in [-1, 0] \\ s_2'(x) = 3ex^2 + 2fx + g & \text{si } x \in [0, 1] \end{cases}$$

$$s''(x) = \begin{cases} s_1''(x) = 6ax + 2b & \text{si } x \in [-1, 0] \\ s_2''(x) = 6ex + 2f & \text{si } x \in [0, 1] \end{cases}$$

y las ecuaciones son:

1. $s_1(-1) = 0 \Longrightarrow -a + b - c + d = 0$
2. $s_1(0) = 2 \Longrightarrow d = 1$
3. $s_2(0) = 2 \Longrightarrow h = 1$
4. $s_2(1) = 0 \Longrightarrow e + f + g + h = 0$
5. $s_1'(0) = s_2'(0) \Longrightarrow c = g$
6. $s_1''(0) = s_2''(0) \Longrightarrow 2b = 2f$

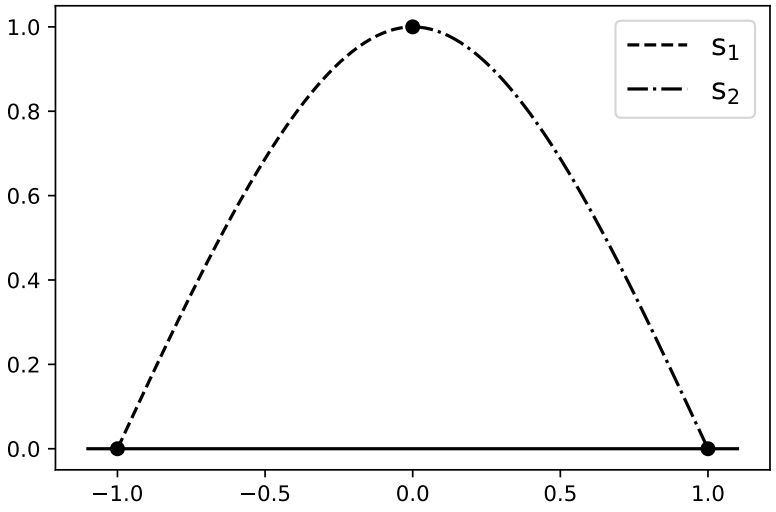
7. $s_1''(-1) = 0 \Longrightarrow -6a + 2b = 0$
8. $s_2''(1) = 0 \Longrightarrow 6e + 2f = 0$

Y tenemos un sistema lineal de ocho ecuaciones lineales para calcular ocho incógnitas, que expresado matricialmente es

$$\begin{pmatrix} -1 & 1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 \\ -6 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6 & 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

La solución de este sistema es $a = -1/2$, $b = -3/2$, $c = 0$, $d = 1$, $e = 1/2$, $f = -3/2$, $g = 0$ y $h = 1$. Por lo tanto la spline cúbica puede definirse

$$s(x) = \begin{cases} s_1(x) = -\dfrac{1}{2}x^3 - \dfrac{3}{2}x^2 + 1 & \text{si } x \in [-1, 0] \\ s_2(x) = +\dfrac{1}{2}x^3 - \dfrac{3}{2}x^2 + 1 & \text{si } x \in [0, 1] \end{cases}$$



Problema 3.12:

Dados los nodos:

x	0	1	2	3
y	1	0	-1	1/2

Si escribimos el spline natural que los ajusta como

$$s(x) = \begin{cases} s_1(x) = a\,x^3 + b\,x^2 + c\,x + d & \text{si } x \in [0,1] \\ s_2(x) = e\,(x-1)^3 + f\,(x-1)^2 + g\,(x-1) + h & \text{si } x \in [1,2] \\ s_3(x) = i\,(x-2)^3 + j\,(x-2)^2 + k\,(x-2) + l & \text{si } x \in [2,3] \end{cases}$$

plantear las ecuaciones del sistema cuya solución son los coeficientes a, b, c, d, ... , k y l. Especificar las condiciones aplicadas.

Se tienen que cumplir las siguientes condiciones:

(a) La curva ha de pasar por los tres puntos. Por lo tanto:

$$s_1(0) = 1,\ s_1(1) = 0,\ s_2(1) = 0,\ s_2(2) = -1,\ s_3(2) = -1,\ s_2(3) = 1/2.$$

(b) Han de coincidir las derivadas primera y segunda en los puntos intermedios:

$$s_1'(1) = s_2'(1),\ s_2'(2) = s_3'(2),\ s_1''(1) = s_2''(1),\ s_2''(2) = s_3''(2)$$

(c) Y como hay 12 incógnitas y de momento sólo tenemos 10 condiciones (ecuaciones) imponemos dos más en los extremos. Como el spline es natural las condiciones adicionales son:

$$s_1''(0) = 0,\ s_3''(3) = 0.$$

Calculamos

$$s'(x) = \begin{cases} s_1'(x) = 3a\,x^2 + 2b\,x + c & \text{si } x \in [0,1] \\ s_2'(x) = 3e\,(x-1)^2 + 2f\,(x-1) + g & \text{si } x \in [1,2] \\ s_3'(x) = 3i\,(x-2)^2 + 2j\,(x-2) + k & \text{si } x \in [2,3] \end{cases}$$

$$s''(x) = \begin{cases} s_1''(x) = 6a\,x + 2b & \text{si } x \in [0,1] \\ s_2''(x) = 6e\,(x-1) + 2f & \text{si } x \in [1,2] \\ s_3''(x) = 6i\,(x-2) + 2j & \text{si } x \in [2,3] \end{cases}$$

y las ecuaciones son:

1. $s_1(0) = 1 \Longrightarrow d = 1$
2. $s_1(1) = 0 \Longrightarrow a + b + c + d = 0$
3. $s_2(1) = 0 \Longrightarrow h = 0$
4. $s_2(2) = -1 \Longrightarrow e + f + g + h = -1$

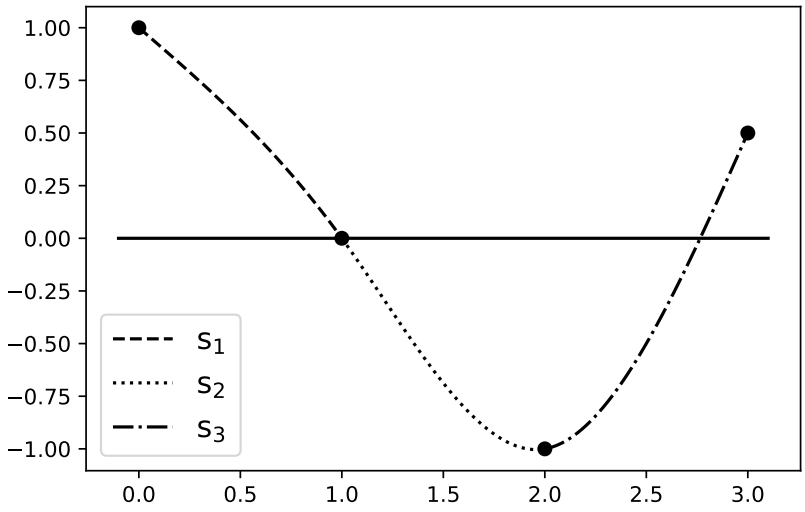
5. $s_3(2) = -1 \Longrightarrow l = -1$
6. $s_3(3) = 1/2 \Longrightarrow i + j + k + l = 1/2$
7. $s_1'(1) = s_2'(1) \Longrightarrow 3a + 2b + c = g$
8. $s_2'(2) = s_3'(2) \Longrightarrow 3e + 2f + g = k$
9. $s_1''(1) = s_2''(1) \Longrightarrow 6a + 2b = 2f$
10. $s_2''(2) = s_3''(2) \Longrightarrow 6e + 2f = 2j$
11. $s_1''(0) = 0 \Longrightarrow 2b = 0$
12. $s_3''(3) = 0 \Longrightarrow 6i + 2j = 0$

Y tenemos un sistema lineal de 12 ecuaciones para calcular 12 incógnitas, que expresado en forma matricial es

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 3 & 2 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 2 & 1 & 0 & 0 & 0 & -1 & 0 \\ 6 & 2 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6 & 2 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ i \\ j \\ k \\ l \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \\ -1 \\ 1/2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

La solución de este sistema es $a = -1/6$, $b = 0$, $c = -5/6$, $d = 1$, $e = 5/6$, $f = -1/2$, $g = -4/3$, $h = 0$, $i = -2/3$, $j = 2$, $k = -1/6$ y $l = -1$. Por lo tanto la spline cúbica puede definirse

$$s(x) = \begin{cases} s_1(x) = -\dfrac{1}{6}\,x^3 - \dfrac{5}{6}\,x + 1 & \text{si } x \in [0,1] \\ s_2(x) = \dfrac{5}{6}\,(x-1)^3 - \dfrac{1}{2}\,(x-1)^2 - \dfrac{4}{3}\,(x-1) & \text{si } x \in [1,2] \\ s_3(x) = -\dfrac{2}{3}\,(x-2)^3 + 2\,(x-2)^2 + \dfrac{1}{6}\,(x-2) - 1 & \text{si } x \in [2,3] \end{cases}$$


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0.0
0.5
1.0
1.5
2.0
2.5
3.0
s_1
s_2
s_3

3.3. Aproximación lineal

Problema 3.13:

Dada la tabla de valores

x	-2	-1	0	1
y	$-0{,}5$	$0{,}5$	1	$2{,}1$

calcular la recta de regresión $P_1(x) = a_0 + a_1 x$ que minimiza los errores cuadráticos. Deducir la fórmula empleada para calcular esta recta a partir de los errores cuadráticos.

La recta será de la forma:

$$P_1(x) = a_0 + a_1 x$$

Queremos calcular la recta que minimiza la suma de los errores cuadráticos. Si d_k es la distancia del punto y_k a su estimación con la recta $P_1(x_k)$

$$E(a_0, a_1) = \sum_{k=1}^{4} d_k^2 = \sum_{k=1}^{4} (P_1(x_k) - y_k)^2 = \sum_{k=1}^{4} (a_0 + a_1 x_k - y_k)^2.$$

Para hallar el error mínimo, calculamos las derivadas parciales respecto a las dos variables y las igualamos a cero:

$$\begin{aligned} \frac{\partial E}{\partial a_0} &= \sum_{k=1}^{4} 2(a_0 + a_1 x_k - y_k) = 0 \\ \frac{\partial E}{\partial a_0} &= \sum_{k=1}^{4} 2(a_0 + a_1 x_k - y_k) x_k = \sum_{k=1}^{4} 2(a_0 x_k + a_1 x_k^2 - x_k y_k) = 0 \end{aligned}$$

Que equivale a

$$\begin{array}{rcrcl} a_0 \sum_{k=1}^{4} 1 & + & a_1 \sum_{k=1}^{4} x_k & = & \sum_{k=1}^{4} y_k \\ a_0 \sum_{k=1}^{4} x_k & + & a_1 \sum_{k=1}^{4} x_k^2 & = & \sum_{k=1}^{4} x_k y_k \end{array}$$

Sistema, que expresado matricialmente es:

$$\begin{pmatrix} \sum_{k=1}^{4} 1 & \sum_{k=1}^{4} x_k \\ \sum_{k=1}^{4} x_k & \sum_{k=1}^{4} x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{4} y_k \\ \sum_{k=1}^{4} x_k y_k \end{pmatrix}$$

El mismo razonamiento, pero utilizando la forma expandida: queremos minimizar la suma de los errores cuadráticos que es

$$\begin{aligned} E(a_0, a_1) &= d_1^2 + d_2^2 + d_3^2 + d_4^2 = \\ &= (P_1(-2) - (-0{,}5))^2 + (P_1(-1) - 0{,}5)^2 + (P_1(0) - 1)^2 + (P_1(1) - 2{,}1)^2 = \\ &= (a_0 + a_1(-2) - (-0{,}5))^2 + (a_0 + a_1(-1) - 0{,}5)^2 + \\ &\quad + (a_0 + a_1(0) - 1)^2 + (a_0 + a_1(1) - 2{,}1)^2 \end{aligned}$$

Derivando respecto a las variables a_0 y a_1 e igualando a cero:

$$\begin{aligned}
\frac{\partial E}{\partial a_0} = 0 \qquad & 2(a_0 + a_1(-2) - (-0{,}5)) + 2(a_0 + a_1(-1) - 0{,}5) + \\
& +2(a_0 + a_1(0) - 1) + 2(a_0 + a_1(1) - 2{,}1) = 0 \\
\frac{\partial E}{\partial a_1} = 0 \qquad & 2(a_0 + a_1(-2) - (-0{,}5))(-2) + 2(a_0 + a_1(-1) - 0{,}5)(-1) + \\
& +2(a_0 + a_1(0) - 1)(0) + 2(a_0 + a_1(1) - 2{,}1)(1) = 0
\end{aligned}$$

Y sacando a_0 y a_1 factor común

$$\begin{aligned}
a_0(1+1+1+1) + a_1((-2)+(-1)+0+1) &= (-0{,}5) + 0{,}5 + 1 + 2{,}1 \\
a_0((-2)+(-1)+0+1) + a_1((-2)^2+(-1)^2+0^2+1^2) &= (-2)(-0{,}5) + (-1)0{,}5 + (0)1 + 1(2{,}1)
\end{aligned}$$

Que es

$$\begin{array}{rcrcl} 4a_0 & - & 2a_1 & = & 3{,}1 \\ -2a_0 & + & 6a_1 & = & 2{,}6 \end{array}$$

Resolvemos el sistema por Gauss: la segunda ecuación $e_2 \to e_2 - (-2/4)e_1$

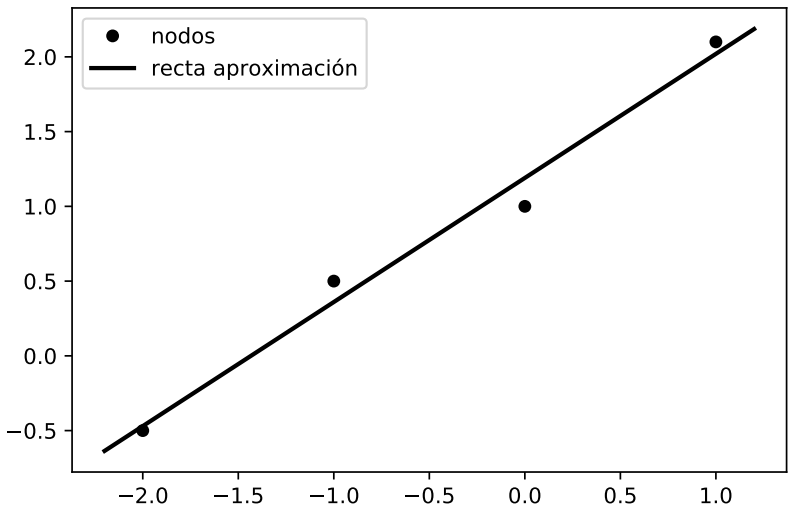
$$\begin{array}{rcrcl} 4a_0 & - & 2a_1 & = & 3{,}1 \\ & & 5a_1 & = & 4{,}15 \end{array}$$

y por sutitución reversiva

$$\begin{array}{rcccl} a_1 & = & 4{,}15/5 & = & 0{,}83 \\ a_0 & = & (3{,}1 + 2a_1)/4 & = & 1{,}19 \end{array}$$

Y la recta de regresión mínimo cuadrática es

$$P_1(x) = 1{,}19 + 0{,}83x$$



Problema 3.14:

Un dispositivo montado en una guía recta se mueve a velocidad constante. Si tomamos la posición y el tiempo inicial como posición y velocidad cero y a partir de ellas medimos la distancia y el tiempo transcurrido, construímos la tabla:

tiempo (min)	0	1	2	3	4	5
posición (m)	0	1,3	2,6	4	5,2	6,6

Calcular la recta que pasa por el origen y que minimiza los errores cuadráticos. Deducir la fórmula empleada para calcular la recta a partir de los errores cuadráticos. Calcular la velocidad del dispositivo que se deduce de la recta.

Una recta que pasa por el origen será de la forma:

$$P(x) = mx$$

Queremos calcular la recta que minimiza la suma de los errores cuadráticos:

$$E(m) = \sum_{k=1}^{5} d_k^2 = \sum_{k=1}^{5} (P(x_k) - y_k)^2 = \sum_{k=1}^{5} (m \times x_k - y_k)^2.$$

Para hallar el error mínimo derivamos:

$$\begin{aligned} E'(m) &= \sum_{k=1}^{5} 2(m \times x_k - y_k)x_k = \sum_{k=1}^{5} 2(m \times x_k^2 - y_k x_k) = 0 \\ & m \sum_{k=1}^{5} x_k^2 - \sum_{k=1}^{5} y_k x_k = 0 \end{aligned}$$

Y por lo tanto:

$$m = \frac{\sum_{k=1}^{4} y_k x_k}{\sum_{k=1}^{4} x_k^2}$$

O lo que es lo mismo **con notación expandida**, la suma de errores cuadráticos es:

$$\begin{aligned} E(m) &= d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 = \\ &= (P(1) - 1{,}3)^2 + (P(2) - 2{,}6)^2 + (P(3) - 4)^2 + (P(4) - 5{,}2)^2 + (P(5) - 6{,}6)^2 = \\ &= (m \times 1 - 1{,}3)^2 + (m \times 2 - 2{,}6)^2 + (m \times 3 - 4)^2 + (m \times 4 - 5{,}2)^2 + (m \times 5 - 6{,}6)^2 \end{aligned}$$

Y la derivada de esta función es

$$\begin{aligned} E'(m) &= 2(m \times 1 - 1{,}3) \times 1 + 2(m \times 2 - 2,6) \times 2 + 2(m \times 3 - 4) \times 3 + \\ &\quad +2(m \times 4 - 5{,}2) \times 4 + 2(m \times 5 - 6{,}6) \times 5 = \\ &= 2m(1 \times 1 + 2 \times 2 + 3 \times 3 + 4 \times 4 + 5 \times 5) \\ &\quad -2(1 \times 1{,}3 + 2 \times 2{,}6 + 3 \times 4 + 4 \times 5{,}2 + 5 \times 6{,}6) = 0 \end{aligned}$$

Y por lo tanto:

$$m = \frac{1 \times 1{,}3 + 2 \times 2{,}6 + 3 \times 4 + 4 \times 5{,}2 + 5 \times 6{,}6}{1 \times 1 + 2 \times 2 + 3 \times 3 + 4 \times 4 + 5 \times 5} = \frac{72{,}3}{55} = 1{,}31$$

Y la recta de aproximación es

$$P(x) = 1{,}31x$$

y como x es tiempo y $P(x)$ posición, m es la velocidad constante y es 1,31 m/min.

Problema 3.15:

Ajustar la curva $Q(t) = \dfrac{a}{t+b}$, calculando los valores a y b utilizando el criterio de los mínimos cuadrados, linealizando previamente la función, para las tablas de valores

(a)

t	1	3	5
Q	0,7	0,4	0,3

(b)

t	1	3	5
Q	0,3	0,5	0,7

Vamos a linealizar la función a ajustar. Para un punto cualquiera, debería cumplirse, aproximadamente

$$Q_k = \frac{a}{t_k+b} \quad \Longrightarrow \quad \frac{1}{Q_k} = \frac{t_k+b}{a} \quad \Longrightarrow \quad \frac{1}{Q_k} = \frac{1}{a}t_k + \frac{b}{a} \quad \Longrightarrow \quad \frac{1}{Q_k} = \frac{b}{a} + \frac{1}{a}t_k$$

Y si llamamos

$$y_k = \frac{1}{Q_k}, \quad x_k = t_k, \quad a_0 = \frac{b}{a}, \quad a_1 = \frac{1}{a}$$

tenemos

$$Q_k = \frac{a}{t_k+b} \quad \Longrightarrow \quad y_k = a_0 + a_1\, x_k$$

el problema es ahora ajustar una recta de regresión mínimo cuadrática

$$P_1(x) = a_0 + a_1\, x$$

a los datos transformados (x_k, y_k), $\quad k = 1, 2, 3$ con el sistema

$$\begin{pmatrix} \sum_{k=1}^3 1 & \sum_{k=1}^3 x_k \\ \sum_{k=1}^3 x_k & \sum_{k=1}^3 x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^3 y_k \\ \sum_{k=1}^3 x_k y_k \end{pmatrix}$$

(a)

	$x_k = t_k$	Q_k	$y_k = \dfrac{1}{Q_k}$	x_k^2	$x_k\, y_k$
	1	0,7	1,43	1	1,43
	3	0,4	2,50	9	7,50
	5	0,3	3,34	25	16,67
$\sum$	9		7,26	35	25,60

Sustituyendo los datos y operando

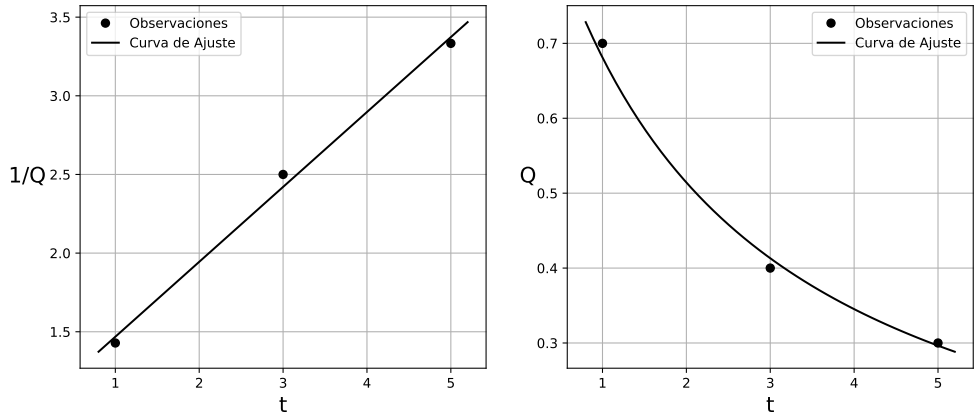
$$\begin{array}{rcrcl} 3a_0 & + & 9a_1 & = & 7{,}26 \\ 9a_0 & + & 35a_1 & = & 25{,}60 \end{array}$$

Y tenemos $a_0 = 0{,}99 \quad a_1 = 0{,}48$. Como

$$a_1 = \frac{1}{a} \Longrightarrow a = \frac{1}{a_1} = 2{,}1 \quad a_0 = \frac{b}{a} \Longrightarrow b = a_0 \times a = 2{,}08$$

y

$$Q\,(t) = \frac{2{,}1}{t + 2{,}08}$$



(b)

	$x_k = t_k$	Q_k	$y_k = 1/Q_k$	x_k^2	$x_k\, y_k$
	1	0,3	3,33	1	3,33
	3	0,5	2,00	9	6,00
	5	0,7	1,43	25	7,14
$\sum$	9		6,76	35	16,48

Sustituyendo los datos y operando

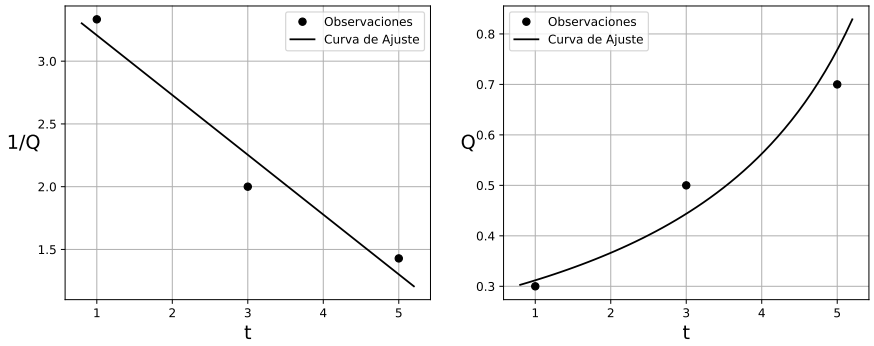
$$\begin{array}{rcrcl} 3a_0 & + & 9a_1 & = & 6{,}76 \\ 9a_0 & + & 35a_1 & = & 16{,}47 \end{array}$$

Y la solución de este sistema es $a_0 = 3{,}68 \quad a_1 = -0{,}47$. Como

$$a_1 = \frac{1}{a} \Longrightarrow a = \frac{1}{a_1} = -2{,}1 \quad a_0 = \frac{b}{a} \Longrightarrow b = a_0 \times a = -7{,}73$$

y

$$Q\,(t) = \frac{-2{,}1}{t - 7{,}73}$$



Problema 3.16:

Dados los siguientes datos

t	1	2	3	4	5
Q	0,1	0,6	3	6,4	7

se quiere encontrar una curva que ajuste estos datos usando una función de la forma

$$Q(t) = \frac{7{,}2}{1 + b\,e^{\,c\,t}}$$

Calcular los valores b y c utilizando el criterio de los mínimos cuadrados.

Necesitamos, primero, linealizar la función a ajustar. Para un punto cualquiera, debería cumplirse, aproximadamente

$$Q_k = \frac{7{,}2}{1 + b\,e^{\,c\,t_k}} \implies \frac{1}{Q_k} = \frac{1 + b\,e^{\,c\,t_k}}{7{,}2} \implies \frac{7{,}2}{Q_k} = 1 + b\,e^{\,c\,t_k} \implies$$

$$\implies \frac{7{,}2}{Q_k} - 1 = b\,e^{\,c\,t_k} \implies \ln\left(\frac{7{,}2}{Q_k} - 1\right) = \ln\left(b\,e^{\,c\,t_k}\right) \implies$$

$$\implies \ln\frac{7{,}2 - Q_k}{Q_k} = \ln b + \ln e^{\,c\,t_k} \implies \ln\frac{7{,}2 - Q_k}{Q_k} = \ln b + c\,t_k \ln e \implies$$

$$\ln\frac{7{,}2 - Q_k}{Q_k} = \ln b + c\,t_k$$

Y si llamamos

$$y_k = \ln\frac{7{,}2 - Q_k}{Q_k}, \quad x_k = t_k, \quad a_0 = \ln b, \quad a_1 = c$$

tenemos

$$\ln\frac{7{,}2 - Q_k}{Q_k} = \ln b + c\,t_k \implies y_k = a_0 + a_1\,x_k$$

el problema es ahora ajustar una recta de regresión mínimo cuadrática

$$P_1(x) = a_0 + a_1\,x$$

a los datos transformados (x_k, y_k), $k = 1, \ldots, 5$ con el sistema del ejercicio anterior.

$$\begin{pmatrix} \sum_{k=1}^5 1 & \sum_{k=1}^5 x_k \\ \sum_{k=1}^5 x_k & \sum_{k=1}^5 x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 y_k \\ \sum_{k=1}^5 x_k y_k \end{pmatrix}$$

Calculamos los valores correspondientes

	$x_k = t_k$	Q_k	$y_k = \ln((7{,}2 - Q_k)/Q_k)$	x_k^2	$x_k\,y_k$
	1	0,1	4,263	1	4,263
	2	0,6	2,398	4	4,796
	3	3,0	0,336	9	1,009
	4	6,4	−2,079	16	−8,318
	5	7,0	−3,555	25	−17,777
$\sum$	15		1,362	55	−16,027

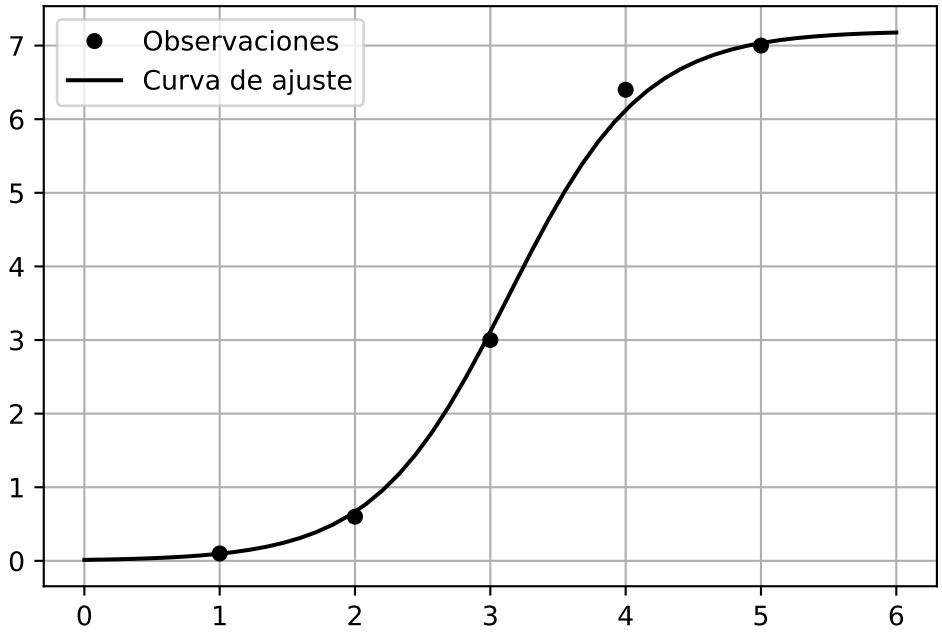
Sustituyendo los datos y operando

$$\begin{array}{rcrcr} 5\,a_0 & + & 15\,a_1 & = & 1{,}362 \\ 15\,a_0 & + & 55\,a_1 & = & -16{,}027 \end{array}$$

Y la solución de este sistema es $a_0 = 6{,}306 \quad a_1 = -2{,}011$. Y tenemos

$$a_0 = \ln b \quad a_1 = c \quad \Longrightarrow b = e^{a_0} \approx 548 \quad c = a_1 \approx -2$$

$$Q\,(t) = \frac{7{,}2}{1 + 548\,e^{-2\,t}}$$



Problema 3.17:

Una elipse, en coordenadas polares (α, r), con un foco en el origen y el otro foco en la parte no negativa del eje X tiene la forma

$$r = \frac{p}{1 - \varepsilon \cos \alpha}$$

con $0 \leq \varepsilon < 1$ y $p > 0$. Si han medido las coordenadas aproximadas de varios puntos de una elipse y tenemos que

α	0^o	15^o	30^o	45^o	60^o
r	15	14	13	11	10

aproximar el valor de p y ε por mínimos cuadrados, linealizando previamente la ecuación.

Para un punto cualquiera, debería cumplirse, aproximadamente

$$r_k = \frac{p}{1 - \varepsilon \cos\left(\alpha_k\right)} \quad \Longrightarrow \quad \frac{1}{r_k} = \frac{1 - \varepsilon \cos\left(\alpha_k\right)}{p} \quad \Longrightarrow \quad \frac{1}{r_k} = \frac{1}{p} - \frac{\varepsilon}{p} \cos\left(\alpha_k\right)$$

Y si llamamos

$$y_k = \frac{1}{r_k}, \quad x_k = \cos(\alpha_k), \quad a_0 = \frac{1}{p}, \quad a_1 = -\frac{\varepsilon}{p}$$

tenemos

$$\frac{1}{r_k} = \frac{1}{p} - \frac{\varepsilon}{p}\cos(\alpha_k) \quad \Longrightarrow \quad y_k = a_0 + a_1\, x_k$$

el problema consiste ahora en ajustar una recta de regresión mínimo cuadrática

$$P_1(x) = a_0 + a_1\, x$$

a los datos transformados $(x_k, y_k), \quad k = 1, \ldots, 5$

$$\begin{pmatrix} \sum_{k=1}^5 1 & \sum_{k=1}^5 x_k \\ \sum_{k=1}^5 x_k & \sum_{k=1}^5 x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 y_k \\ \sum_{k=1}^5 x_k y_k \end{pmatrix}$$

Calculamos los valores correspondientes

	r_k	α_k	$y_k = 1/r_k$	$x_k = \cos\,\alpha_k$	x_k^2	$x_k\, y_k$
	15	0^o	0,0667	1,000	1,000	0,0667
	14	15^o	0,0714	0,966	0,933	0,0690
	13	30^o	0,0769	0,866	0,750	0,0666
	11	45^o	0,0909	0,707	0,500	0,0643
	10	60^o	0,1000	0,500	0,250	0,0500
$\sum$			0,4059	4,039	3,433	0,3166

Sustituyendo los datos y operando

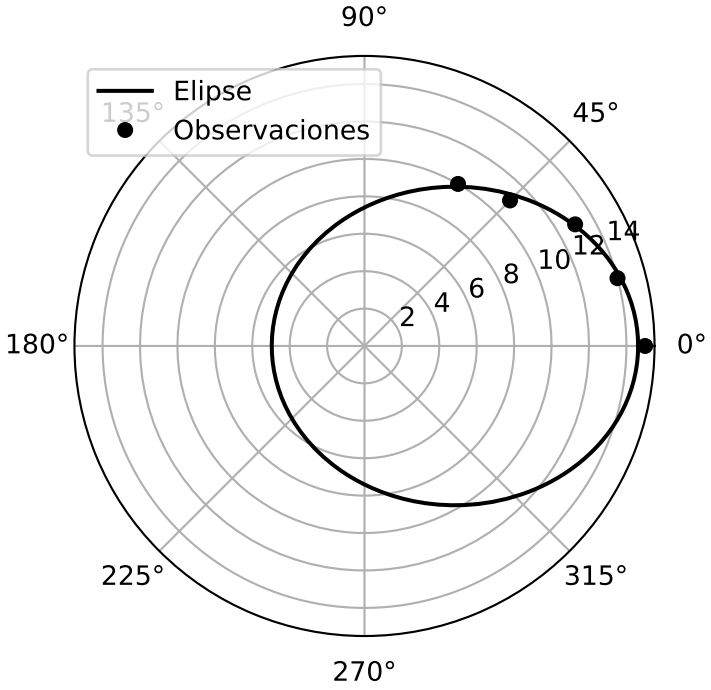
$$\begin{array}{rcrcl} 5\, a_0 & + & 4{,}039\, a_1 & = & 0{,}4059 \\ 4{,}039\, a_0 & + & 3{,}433\, a_1 & = & 0{,}3166 \end{array}$$

Y la solución de este sistema es $a_0 = 0{,}135 \quad a_1 = -0{,}0667$. Como

$$a_0 = \frac{1}{p} \quad a_1 = -\frac{\varepsilon}{p} \quad \Longrightarrow p = \frac{1}{a_0} \quad \varepsilon = -a_1 \times p$$

y $p = 7{,}4$, $\varepsilon = 0{,}5$ y la ecuación es

$$r = \frac{7{,}4}{1 - 0{,}5\cos(\alpha)}$$


90°
45°
135°
180°
0°
225°
315°
270°
Elipse
Observaciones
2
4
6
8
10
12
14

3.4. Aproximación con funciones polinómicas

Problema 3.18:

Dados cinco nodos equiespaciados en el intervalo $[-1, 1]$ y la función

$$f(x) = \cos x,$$

calcular los coeficientes del polinomio

$$P(x) = a_0 + a_2 x^2 + a_4 x^4$$

que aproxima la función en los nodos utilizando el criterio de mínimos cuadrados. Utilizar la base de polinomios

$$\{P_0, P_2, P_4\} = \left\{1, x^2, x^4\right\}$$

La solución al problema planteado sería la solución del sistema

$$\begin{pmatrix} \langle P_0, P_0\rangle & \langle P_0, P_2\rangle & \langle P_0, P_4\rangle \\ \langle P_2, P_0\rangle & \langle P_2, P_2\rangle & \langle P_2, P_4\rangle \\ \langle P_4, P_0\rangle & \langle P_4, P_2\rangle & \langle P_4, P_4\rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \\ a_4 \end{pmatrix} = \begin{pmatrix} \langle P_0, f(x)\rangle \\ \langle P_2, f(x)\rangle \\ \langle P_4, f(x)\rangle \end{pmatrix}$$

siendo el producto escalar

$$\langle g(x), h(x)\rangle = \sum_{k=1}^{5} g(x_k)\, h(x_k).$$

Por lo tanto

$$\begin{pmatrix} \sum_{k=1}^5 1\times 1 & \sum_{k=1}^5 1\times x_k^2 & \sum_{k=1}^5 1\times x_k^4 \\ \sum_{k=1}^5 x_k^2\times 1 & \sum_{k=1}^5 x_k^2\times x_k^2 & \sum_{k=1}^5 x_k^2\times x_k^4 \\ \sum_{k=1}^5 x_k^4\times 1 & \sum_{k=1}^5 x_k^4\times x_k^2 & \sum_{k=1}^5 x_k^4\times x_k^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \\ a_4 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 1\times f(x_k) \\ \sum_{k=1}^5 x_k^2\times f(x_k) \\ \sum_{k=1}^5 x_k^4\times f(x_k) \end{pmatrix}$$

que equivale a

$$\begin{pmatrix} \sum_{k=1}^5 1 & \sum_{k=1}^5 x_k^2 & \sum_{k=1}^5 x_k^4 \\ \sum_{k=1}^5 x_k^2 & \sum_{k=1}^5 x_k^4 & \sum_{k=1}^5 x_k^6 \\ \sum_{k=1}^5 x_k^4 & \sum_{k=1}^5 x_k^6 & \sum_{k=1}^5 x_k^8 \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \\ a_4 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 1\times\cos(x_k) \\ \sum_{k=1}^5 x_k^2\times\cos(x_k) \\ \sum_{k=1}^5 x_k^4\times\cos(x_k) \end{pmatrix}$$

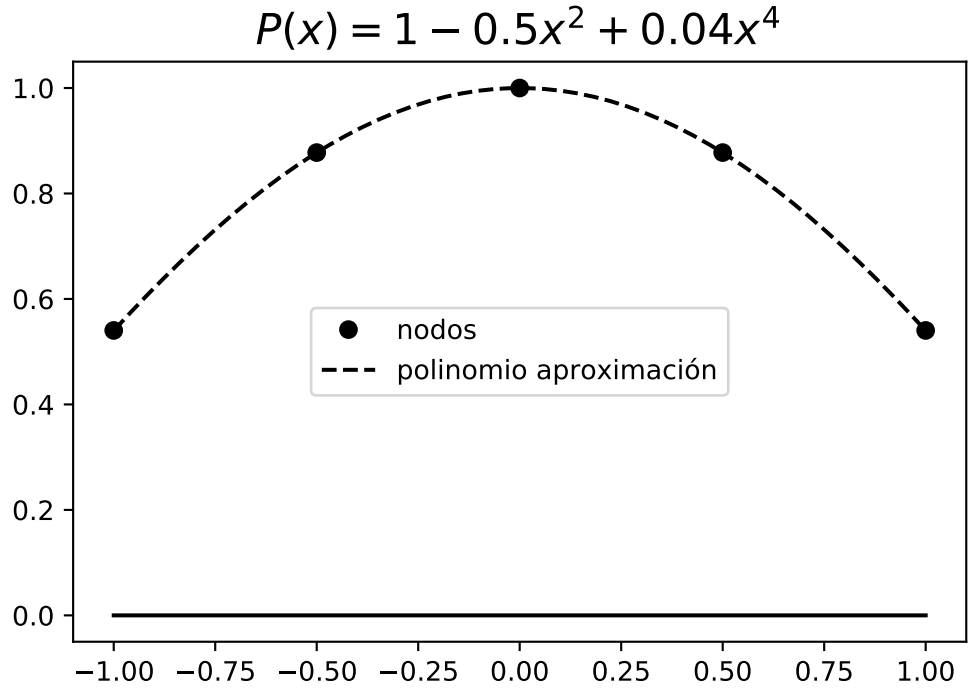
Tomando $x_1 = -1$, $x_2 = -0{,}5$, $x_3 = 0$,$x_4 = 0{,}5$ y $x_5 = 1$,

	x_k	x_k^2	x_k^4	x_k^6	x_k^8	$\cos(x_k)$	$x_k^2\cos(x_k)$	$x_k^4\cos(x_k)$
	-1	1,0000	1,0000	1,0000	1,0000	0,5403	0,5403	0,5403
	$-0{,}5$	0,2500	0,0625	0,0156	0,0039	0,8776	0,2194	0,0548
	-0	0,0000	0,0000	0,0000	0,0000	1,0000	0,0000	0,0000
	$+0{,}5$	0,2500	0,0625	0,0156	0,0039	0,8776	0,2194	0,0548
	$+1$	1,0000	0,1000	1,0000	1,0000	0,5403	0,5403	0,5403
$\sum$		2,5000	2,1250	2,0313	2,0078	3,8358	1,5194	1,1903

Sustituyendo los datos y operando

$$\begin{pmatrix} 5 & 2{,}5000 & 2{,}1250 \\ 2{,}5000 & 2{,}1250 & 2{,}0313 \\ 2{,}1250 & 2{,}0313 & 2{,}0078 \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \\ a_4 \end{pmatrix} = \begin{pmatrix} 3{,}8358 \\ 1{,}5194 \\ 1{,}1903 \end{pmatrix}$$

Y la solución del sistema es $a_0 = 1 \quad a_2 = -0{,}5 \quad a_4 = 0{,}04$.



Problema 3.19:

Dados siete nodos equiespaciados en el intervalo $[-1, 1]$ y dada la función

$$f(x) = \tan x$$

calcular los coeficientes del polinomio de la forma

$$P(x) = a_1 x + a_3 x^3$$

que aproxima la función en los nodos. Utilizar la base de polinomios

$$\{P_1, P_3\} = \left\{x, x^3\right\}$$

La solución al problema planteado sería la solución del sistema

$$\begin{pmatrix} \langle P_1, P_1 \rangle & \langle P_1, P_3 \rangle \\ \langle P_3, P_1 \rangle & \langle P_3, P_3 \rangle \end{pmatrix} \begin{pmatrix} a_1 \\ a_3 \end{pmatrix} = \begin{pmatrix} \langle P_1, f(x) \rangle \\ \langle P_3, f(x) \rangle \end{pmatrix}$$

siendo el producto escalar

$$\langle g(x), h(x) \rangle = \sum_{k=1}^{7} g(x_k)\, h(x_k).$$

Por lo tanto

$$\begin{pmatrix} \sum_{k=1}^{7} x_k \times x_k & \sum_{k=1}^{7} x_k \times x_k^3 \\ \sum_{k=1}^{7} x_k^3 \times x_k & \sum_{k=1}^{7} x_k^3 \times x_k^3 \end{pmatrix} \begin{pmatrix} a_1 \\ a_3 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{7} x_k \times f(x_k) \\ \sum_{k=1}^{7} x_k^3 \times f(x_k) \end{pmatrix},$$

que equivale a

$$\begin{pmatrix} \sum_{k=1}^{7} x_k^2 & \sum_{k=1}^{7} x_k^4 \\ \sum_{k=1}^{7} x_k^4 & \sum_{k=1}^{7} x_k^6 \end{pmatrix} \begin{pmatrix} a_1 \\ a_3 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{7} x_k \times \tan x_k \\ \sum_{k=1}^{7} x_k^3 \times \tan x_k \end{pmatrix}.$$

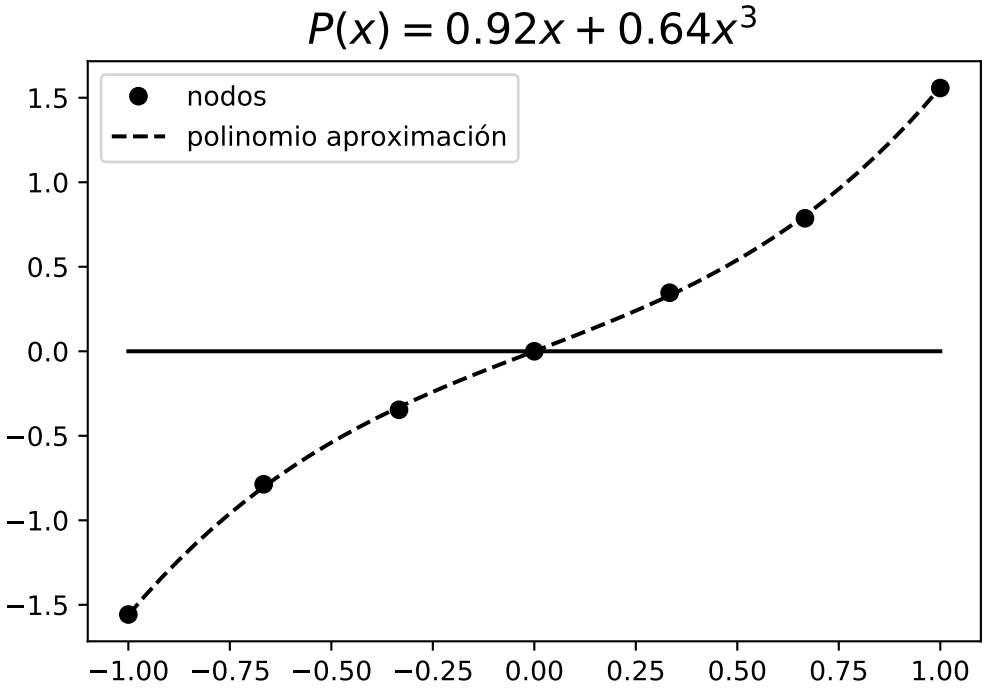
Tomando los x_k, equiespaciados en $[-1, 1]$ tenemos

	x_k	x_k^2	x_k^3	x_k^4	x_k^6	$x_k \tan(x_k)$	$x_k^3 \tan(x_k)$
	−1,0000	1,0000	−1,0000	1,0000	1,0000	1,5574	1,5574
	−0,6667	0,4444	−0,2963	0,1975	0,0878	0,5246	0,2332
	−0,3333	0,1111	−0,0370	0,0123	0,0014	0,1154	0,0128
	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	0,3333	0,1111	0,0370	0,0123	0,0014	0,1154	0,0128
	0,6667	0,4444	0,2963	0,1975	0,0878	0,5246	0,2332
	1,0000	1,0000	1,0000	1,0000	1,0000	1,5574	1,5574
$\sum$		3,1111		2,4198	2,1783	4,3948	3,6067

Sustituyendo los datos y operando

$$\begin{pmatrix} 3{,}1111 & 2{,}4198 \\ 2{,}4198 & 2{,}1783 \end{pmatrix} \begin{pmatrix} a_1 \\ a_3 \end{pmatrix} = \begin{pmatrix} 4{,}3948 \\ 3{,}6067 \end{pmatrix}$$

Y la solución del sistema es $a_1 = 0{,}92 \quad a_3 = 0{,}64$.



Problema 3.20:

Dada la función $f(x) = x^4$ en $[-1, 1]$ calcular la parábola que aproxima de forma continua a la función utilizando la base de polinomios $\{1, x^2\}$.

Para una base de polinomios $\{P_0, P_2\} = \{1, x^2\}$ el polinomio de aproximación es de la forma $P(x) = a_0 P_0(x) + a_2 P_2(x)$ donde los coeficientes a_0y a_2 son la solución del sistema lineal:

$$\begin{pmatrix} \langle P_0, P_0\rangle & \langle P_0, P_2\rangle \\ \langle P_2, P_0\rangle & \langle P_2, P_2\rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f(x)\rangle \\ \langle P_2, f(x)\rangle \end{pmatrix}.$$

En este caso, como aproximamos en un intervalo, el producto escalar podría ser

$$\langle g(x), h(x)\rangle = \int_{-1}^{1} g(x)h(x)dx.$$

Por lo tanto

$$\begin{pmatrix} \int_{-1}^{1} P_0P_0dx & \int_{-1}^{1} P_0P_2dx \\ \int_{-1}^{1} P_2P_0dx & \int_{-1}^{1} P_2P_2dx \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \end{pmatrix} = \begin{pmatrix} \int_{-1}^{1} P_0f(x)dx \\ \int_{-1}^{1} P_2f(x)dx \end{pmatrix},$$

es decir

$$\begin{pmatrix} \int_{-1}^{1} 1dx & \int_{-1}^{1} x^2dx \\ \int_{-1}^{1} x^2dx & \int_{-1}^{1} x^4dx \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \end{pmatrix} = \begin{pmatrix} \int_{-1}^{1} f(x)dx \\ \int_{-1}^{1} x^2f(x)dx \end{pmatrix},$$

que es

$$\begin{pmatrix} 2 & 2/3 \\ 2/3 & 2/5 \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2/5 \\ 2/7 \end{pmatrix}.$$

Tenemos dos ecuaciones

$$\begin{cases} 2a_0 + \frac{2}{3}a_2 = \frac{2}{5} \\ \frac{2}{3}a_0 + \frac{2}{5}a_2 = \frac{2}{7} \end{cases}.$$

Despejando a_0 en la primera ecuación

$$a_0 = \frac{1}{2}\left(\frac{2}{5} - \frac{2}{3}a_2\right)$$

Y sustituyendo en la segunda ecuación:

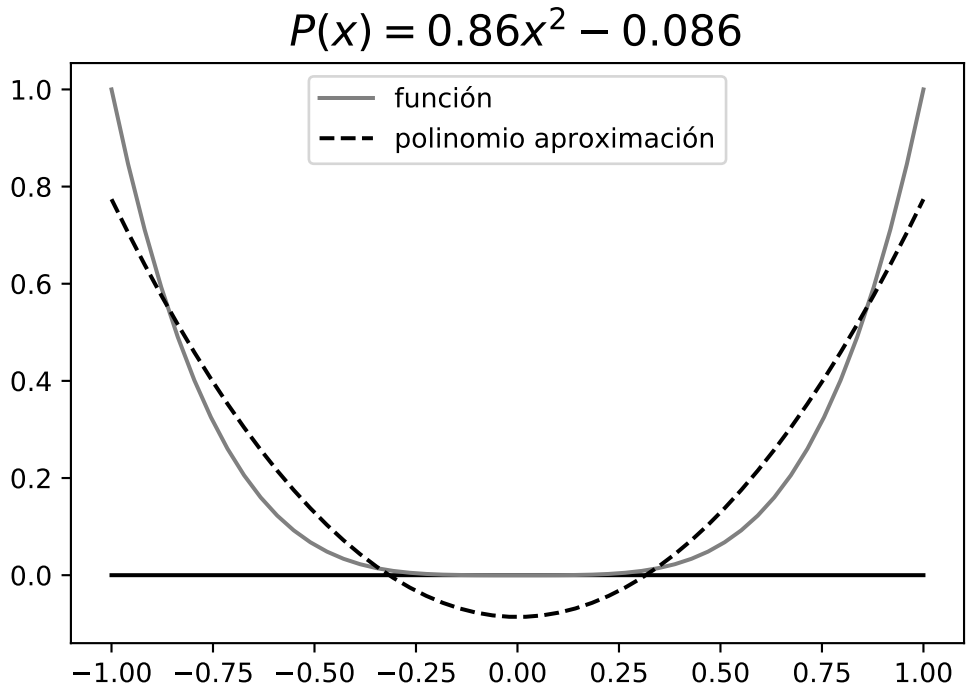
$$\frac{2}{3}\left(\frac{1}{5} - \frac{1}{3}a_2\right) + \frac{2}{5}a_2 = \frac{2}{7} \Longrightarrow \frac{2}{15} + \frac{8}{45}a_2 = \frac{2}{7} \Longrightarrow a_2 = \frac{6}{7} \approx 0{,}86.$$

Entonces

$$a_0 = \frac{1}{2}\left(\frac{2}{5} - \frac{2}{3}a_2\right) = \frac{1}{2}\left(\frac{2}{5} - \frac{2}{3}\left(\frac{6}{7}\right)\right) = -\frac{3}{35} \approx -0{,}086.$$

Y el polinomio que aproxima la función f en el intervalo $[-1, 1]$

$$P(x) = a_0P_0(x) + a_2P_2(x) = -\frac{3}{35} \times 1 + \frac{6}{7} \times x^2 \approx -0{,}086 + 0{,}86x^2.$$


$P(x) = 0.86x^2 - 0.086$
función
polinomio aproximación
1.0
0.8
0.6
0.4
0.2
0.0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

3.5. Aproximación con polinomios ortogonales

Problema 3.21:

Dada la función $f(x) = \ln x$ en $[1, 2]$, calcular la parábola que aproxima de forma continua esta función en este intervalo utilizando la base de polinomios ortonormales

$$B = \{P_0(x), P_1(x), P_2(x)\} = \left\{1, \sqrt{12}\left(x - \frac{3}{2}\right), \sqrt{180}\left(x^2 - 3x + \frac{13}{6}\right)\right\}.$$

Tener en cuenta que

$$\int_1^2 \ln x\, dx = 0{,}3863, \quad \int_1^2 x \ln x\, dx = 0{,}6363, \quad \int_1^2 x^2 \ln x\, dx = 1{,}0706$$

Para una base de polinomios $\{P_0, P_1, P_2\}$ el polinomio de aproximación es de la forma

$$P(x) = a_0 P_0(x) + a_1 P_1(x) + a_2 P_2(x)$$

donde los coeficientes a_0, a_1 y a_2 son la solución del sistema lineal:

$$\begin{pmatrix} \langle P_0, P_0\rangle & \langle P_0, P_1\rangle & \langle P_0, P_2\rangle \\ \langle P_1, P_0\rangle & \langle P_1, P_1\rangle & \langle P_1, P_2\rangle \\ \langle P_2, P_0\rangle & \langle P_2, P_1\rangle & \langle P_2, P_2\rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f\rangle \\ \langle P_1, f\rangle \\ \langle P_2, f\rangle \end{pmatrix}.$$

Si los polinomios son ortonormales para el producto escalar utilizado se tiene que

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f\rangle \\ \langle P_1, f\rangle \\ \langle P_2, f\rangle \end{pmatrix},$$

y entonces

$$a_0 = \langle P_0, f\rangle \quad a_1 = \langle P_1, f\rangle \quad a_2 = \langle P_2, f\rangle.$$

En este caso el producto escalar es

$$\langle g, h\rangle = \int_1^2 g(x)\, h(x)\, dx.$$

Si hacemos los cálculos:

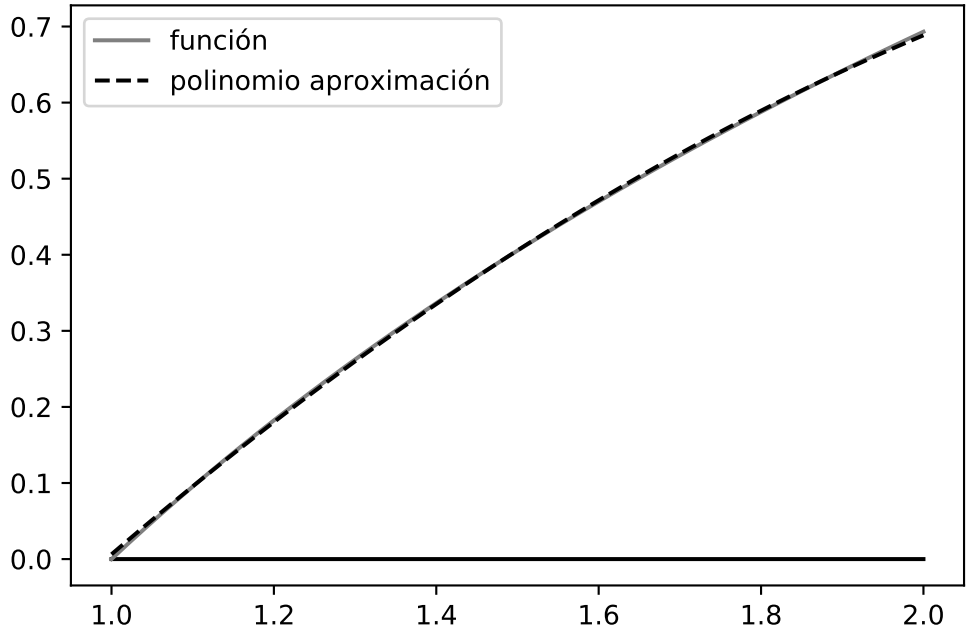
$$a_0 = \langle P_0, f(x)\rangle = \int_1^2 P_0(x)\, f(x)\, dx = \int_1^2 \ln x\, dx = 0,3863$$

$$\begin{aligned} a_1 = \langle P_1, f(x)\rangle = \int_1^2 P_1(x)\, f(x)\, dx = \int_1^2 \sqrt{12}\left(x - \frac{3}{2}\right) \ln x\, dx = \\ = \sqrt{12}\left(\int_1^2 x \ln x\, dx - 1{,}5 \int_1^2 \ln x\, dx\right) = \\ = \sqrt{12}(0{,}6363 - 1{,}5 \times 0{,}3863) = 0{,}1969 \end{aligned}$$

$$a_2 = \langle P_2, f(x)\rangle = \int_1^2 P_2(x)\, f(x)\, dx = \sqrt{180}\int_1^2 (x^2 - 3x + 2{,}17)\, \ln x\, dx =$$
$$= \sqrt{180}\left(\int x^2\, \ln x\, dx - 3\int_1^2 x\, \ln x\, dx + 2{,}17\int_1^2 \ln x\, dx\right) =$$
$$= \sqrt{180}\,(1{,}0706 - 3\times 0{,}6363 + 2{,}17\times 0{,}3863) = -0{,}0174$$

Entonces

$$\begin{aligned} P(x) &= a_0P_0(x) + a_1P_1(x) + a_2P_2(x) = \\ &= 0,3863 + 0{,}1969\sqrt{12}(x-1{,}5) - 0{,}0174\sqrt{180}(x^2 - 3x + 2{,}1667) = \\ &= 0,3863 + 0{,}6822\,(x-1{,}5) - 0{,}2335\,(x^2 - 3x + 2{,}1667) \end{aligned}$$



Problema 3.22:

Dada la función $f(x) = e^{-x}$ en $[0, 2]$, calcular la parábola que aproxima de forma continua esta función en este intervalo, utilizando la base de polinomios ortonormales

$$B = \{P_0(x), P_1(x), P_2(x)\} = \left\{\frac{\sqrt{2}}{2}, \frac{\sqrt{6}}{2}(x-1), \frac{\sqrt{10}}{4}(3x^2 - 6x + 2)\right\}.$$

$$\int_0^2 e^{-x}dx = 0{,}8646, \quad \int_0^2 x\, e^{-x}dx = 0{,}5940, \quad \int_0^2 x^2e^{-x}dx = 0{,}6466$$

Para una base de polinomios $\{P_0, P_1, P_2\}$ el polinomio de aproximación es de la forma

$$P(x) = a_0P_0(x) + a_1P_1(x) + a_2P_2(x)$$

donde los coeficientes a_0, a_1 y a_2 son la solución del sistema lineal:

$$\begin{pmatrix} \langle P_0, P_0\rangle & \langle P_0, P_1\rangle & \langle P_0, P_2\rangle \\ \langle P_1, P_0\rangle & \langle P_1, P_1\rangle & \langle P_1, P_2\rangle \\ \langle P_2, P_0\rangle & \langle P_2, P_1\rangle & \langle P_2, P_2\rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f\rangle \\ \langle P_1, f\rangle \\ \langle P_2, f\rangle \end{pmatrix}.$$

Si los polinomios son ortonormales para el producto escalar utilizado se tiene que

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f\rangle \\ \langle P_1, f\rangle \\ \langle P_2, f\rangle \end{pmatrix},$$

y entonces

$$a_0 = \langle P_0, f\rangle \quad a_1 = \langle P_1, f\rangle \quad a_2 = \langle P_2, f\rangle.$$

En este caso el producto escalar es

$$\langle g, h\rangle = \int_0^2 g(x)\, h(x)\, dx.$$

Si hacemos los cálculos:

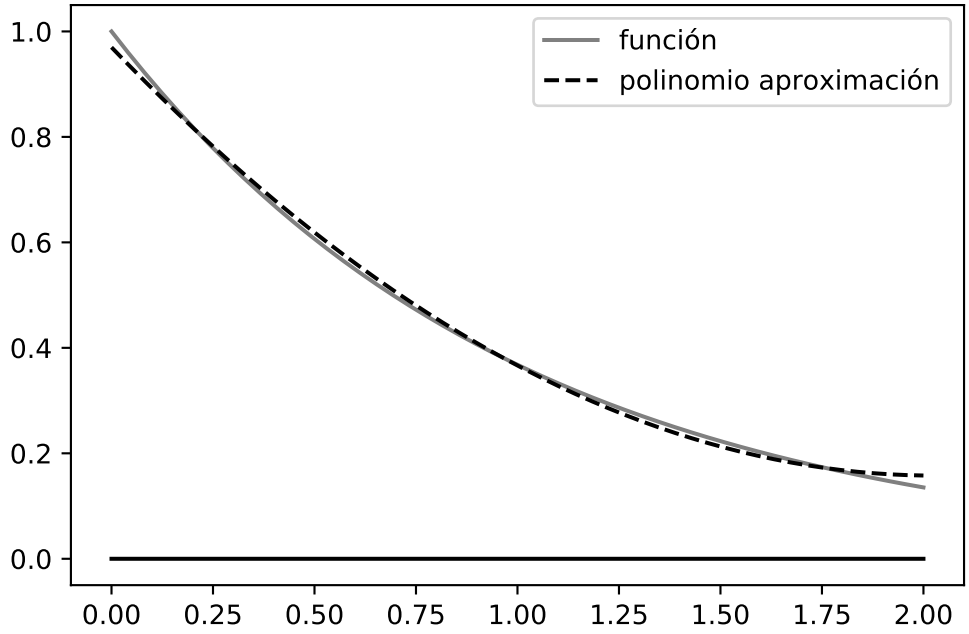
$$a_0 = \langle P_0, f(x)\rangle = \int_0^2 P_0(x)\, f(x)\, dx = \frac{\sqrt{2}}{2}\int_0^2 e^{-x}dx = \\ = \frac{\sqrt{2}}{2} 0{,}8647 = 0{,}6114$$

$$a_1 = \langle P_1, f(x)\rangle = \int_0^2 P_1(x)\, f(x)\, dx = \int_0^2 \frac{\sqrt{6}}{2}(x-1)\, e^{-x}dx = \\ = \frac{\sqrt{6}}{2}\left(\int_0^2 x\, e^{-x}dx - \int_0^2 e^{-x}dx\right) = \\ = \frac{\sqrt{6}}{2}(0{,}5940 - 0{,}8647) = -0{,}3315$$

$$a_2 = \langle P_2, f(x)\rangle = \int_0^2 P_2(x)\, f(x)\, dx = \frac{\sqrt{10}}{4}\int_0^2 (3x^2 - 6x + 2)\, e^{-x}dx = \\ = \frac{\sqrt{10}}{4}\left(3\int_0^2 x\, e^{-x}dx - 6\int_0^2 x\, e^{-x}dx + 2\int_0^2 e^{-x}dx\right) = \\ = \frac{\sqrt{10}}{4}(3 \times 0{,}6466 - 6 \times 0{,}5940 + 2 \times 0{,}8647) = 0{,}08325$$

Entonces

$$\begin{aligned} P(x) &= a_0 P_0(x) + a_1 P_1(x) + a_2 P_2(x) = \\ &= 0{,}6114\frac{\sqrt{2}}{2} - 0{,}3315\frac{\sqrt{6}}{2}(x-1) + 0{,}08325\frac{\sqrt{10}}{4}(3x^2 - 6x + 2) = \\ &= 0{,}4323 - 0{,}4060\,(x-1) + 0{,}0658\,(3x^2 - 6x + 2) \end{aligned}$$



Problema 3.23:

Dada la función $f(x) = x^4 + x^2$ y dados los puntos $x_1 = -2$, $x_2 = -1$, $x_3 = 0$, $x_4 = 1$ y $x_5 = 2$, calcular la parábola que aproxima a estos nodos de la función utilizando la base de polinomios ortogonales $\{P_0, P_1, P_2\} = \{1, x, x^2 - 2\}$.

Para una base de polinomios $\{P_0, P_1, P_2\}$ el polinomio de aproximación es de la forma $P(x) = a_0 P_0(x) + a_1 P_1(x) + a_2 P_2(x)$ donde los coeficientes a_0, a_1 y a_2 son la solución del sistema lineal:

$$\begin{pmatrix} \langle P_0, P_0\rangle & \langle P_0, P_1\rangle & \langle P_0, P_2\rangle \\ \langle P_1, P_0\rangle & \langle P_1, P_1\rangle & \langle P_1, P_2\rangle \\ \langle P_2, P_0\rangle & \langle P_2, P_1\rangle & \langle P_2, P_2\rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f(x)\rangle \\ \langle P_1, f(x)\rangle \\ \langle P_2, f(x)\rangle \end{pmatrix}.$$

Si los polinomios son ortogonales para el producto escalar utilizado se tiene que

$$\begin{pmatrix} \langle P_0, P_0\rangle & 0 & 0 \\ 0 & \langle P_1, P_1\rangle & 0 \\ 0 & 0 & \langle P_2, P_2\rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f(x)\rangle \\ \langle P_1, f(x)\rangle \\ \langle P_2, f(x)\rangle \end{pmatrix},$$

y entonces

$$a_0 = \frac{\langle P_0, f(x)\rangle}{\langle P_0, P_0\rangle} \quad a_1 = \frac{\langle P_1, f(x)\rangle}{\langle P_1, P_1\rangle} \quad a_2 = \frac{\langle P_2, f(x)\rangle}{\langle P_2, P_2\rangle}.$$

En este caso el producto escalar es

$$\langle g(x), h(x)\rangle = \sum_{k=1}^{5} g(x_k)\, h(x_k).$$

Si hacemos los cálculos:

$$\langle P_0, P_0\rangle = \sum_{k=1}^{5} P_0(x_k)\, P_0(x_k) = \sum_{k=1}^{5} 1 \times 1 = \sum_{k=1}^{5} 1$$
$$\langle P_1, P_1\rangle = \sum_{k=1}^{5} P_1(x_k)\, P_1(x_k) = \sum_{k=1}^{5} x_k \times x_k = \sum_{k=1}^{5} x_k^2$$
$$\langle P_2, P_2\rangle = \sum_{k=1}^{5} P_2(x_k)\, P_2(x_k) = \sum_{k=1}^{5} \left(x_k^2 - 2\right)^2$$
$$\langle P_0, f(x)\rangle = \sum_{k=1}^{5} P_0(x_k)\, f(x_k) = \sum_{k=1}^{5} 1 \times (x_k^4 + x_k^2) = \sum_{k=1}^{5} x_k^4 + x_k^2$$
$$\langle P_1, f(x)\rangle = \sum_{k=1}^{5} P_1(x_k)\, f(x_k) = \sum_{k=1}^{5} x_k \times (x_k^4 + x_k^2)$$
$$\langle P_2, f(x)\rangle = \sum_{k=1}^{5} P_2(x_k)\, f(x_k) = \sum_{k=1}^{5} \left(x_k^2 - 2\right) \times (x_k^4 + x_k^2)$$

Construyamos una tabla para calcular estos valores

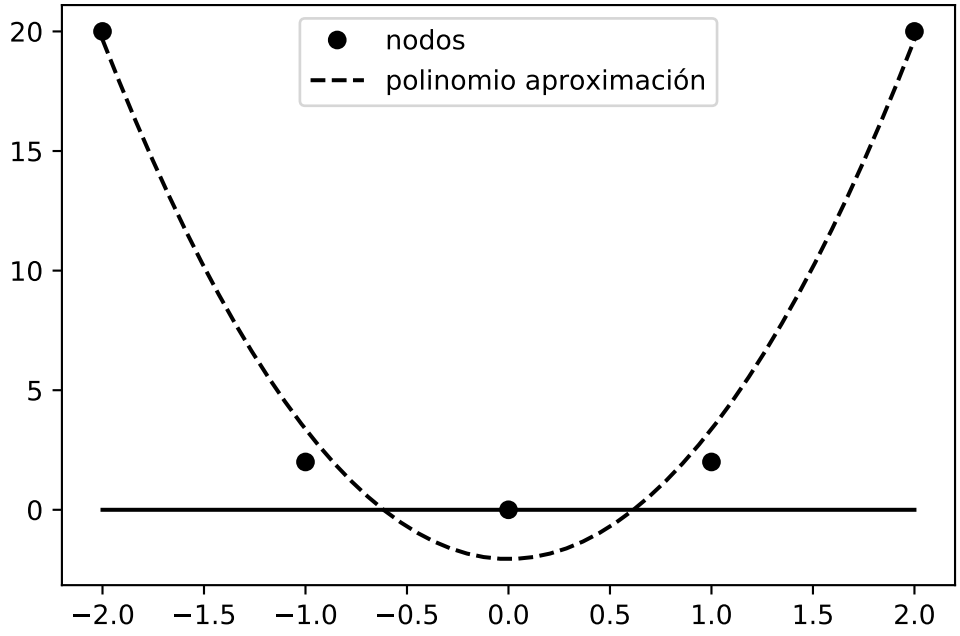
	1	x_k^2	$(x_k^2-2)^2$	$x_k^4+x_k^2$	$x_k(x_k^4+x_k^2)$	$(x_k^2-2)(x_k^4+x_k^2)$
	1	4	4	20	-40	40
	1	1	1	2	-2	-2
	1	0	4	0	0	0
	1	1	1	2	2	-2
	1	4	4	20	40	40
$\sum$	5	10	14	44	0	76

$$a_0 = \frac{\langle P_0, f(x)\rangle}{\langle P_0, P_0\rangle} = \frac{\sum_{k=1}^{5}(x_k^4 - x_k^2)}{\sum_{k=1}^{5} 1} = \frac{44}{5}$$
$$a_1 = \frac{\langle P_1, f(x)\rangle}{\langle P_1, P_1\rangle} = \frac{\sum_{k=1}^{5}(x_k^4 - x_k^2)}{\sum_{k=1}^{5} x_k^2} = \frac{0}{10} = 0$$
$$a_2 = \frac{\langle P_2, f(x)\rangle}{\langle P_2, P_2\rangle} = \frac{\sum_{k=1}^{5}(x_k^2-2)(x_k^4 + x_k^2)}{\sum_{k=1}^{5}\left(x_k^2-2\right)^2} = \frac{76}{14}$$

y el polinomio es

$$P(x) = a_0 P_0(x) + a_1 P_1(x) + a_2 P_2(x) = \frac{44}{5} \times P_0(x) + 0 \times P_1(x) + \frac{76}{14} \times P_2(x)$$

$$P(x) = \frac{44}{5} + \frac{76}{14}(x^2 - 2) = \frac{2}{35}(95x^2 - 36)$$


nodos
polinomio aproximación
20
15
10
5
0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

3.6. Aproximación con funciones trigonométricas (Fourier)

Problema 3.24:

Calcular el desarrollo de Fourier de la función f de periodo π que en el intervalo $[0, \pi]$ se define como $f(x) = x(\pi - x)$ usando la base ortogonal $\{1, \cos 2x, \text{sen}2x\}$

La base es $\{1, \cos 2x, \text{sen}2x\}$ y la función f está definida en $[0, \pi]$ por lo que el producto escalar es

$$\langle f(x), g(x)\rangle = \int_0^\pi f(x)g(x)dx$$

La función aproximada será de la forma

$$P(x) = \frac{a_0}{2}\cdot 1 + a_1 \cdot \cos 2x + b_1 \cdot \text{sen}2x$$

Obtenemos los coeficientes a_0, a_1 y b_1 como solución del sistema lineal

$$\begin{pmatrix} \langle 1,1\rangle & \langle 1,\cos 2x\rangle & \langle 1,\text{sen}2x\rangle \\ \langle \cos 2x,1\rangle & \langle \cos 2x,\cos 2x\rangle & \langle \cos 2x,\text{sen}2x\rangle \\ \langle \text{sen}2x,1\rangle & \langle \text{sen}2x,\cos 2x\rangle & \langle \text{sen}2x,\text{sen}2x\rangle \end{pmatrix} \begin{pmatrix} \frac{a_0}{2} \\ a_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} \langle 1, f(x)\rangle \\ \langle \cos 2x, f(x)\rangle \\ \langle \text{sen}2x, f(x)\rangle \end{pmatrix}$$

Como la base es ortogonal

$$\begin{pmatrix} \langle 1,1\rangle & 0 & 0 \\ 0 & \langle \cos 2x,\cos 2x\rangle & 0 \\ 0 & 0 & \langle \text{sen}2x,\text{sen}2x\rangle \end{pmatrix} \begin{pmatrix} \frac{a_0}{2} \\ a_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} \langle 1, f(x)\rangle \\ \langle \cos 2x, f(x)\rangle \\ \langle \text{sen}2x, f(x)\rangle \end{pmatrix}$$

y por lo tanto

$$\frac{a_0}{2} = \frac{\langle 1, f(x)\rangle}{\langle 1,1\rangle} \quad a_1 = \frac{\langle \cos 2x, f(x)\rangle}{\langle \cos 2x, \cos 2x\rangle} \quad b_1 = \frac{\langle \text{sen}2x, f(x)\rangle}{\langle \text{sen}2x, \sin x\rangle}$$

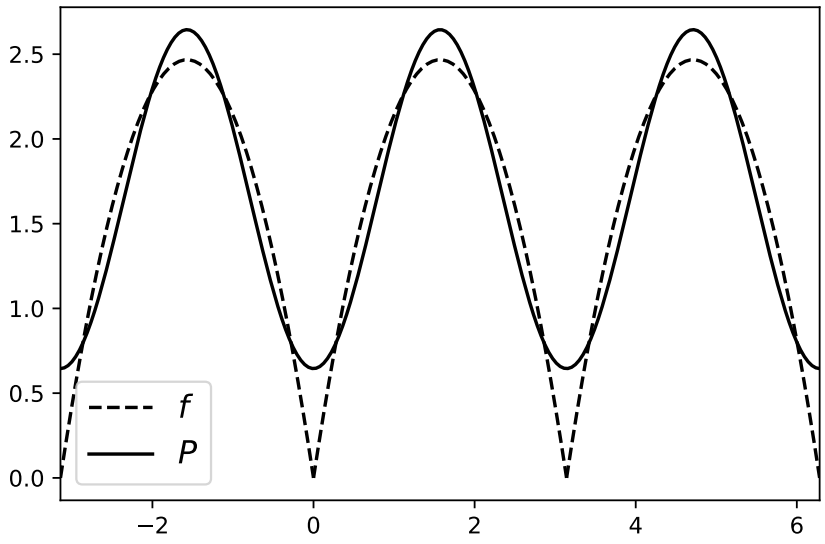
$$\frac{a_0}{2} = \frac{\int_0^\pi 1\cdot f(x)dx}{\int_0^\pi 1\cdot 1 dx} = \frac{\pi^3/6}{\pi} = \frac{\pi^2}{6} \qquad a_1 = \frac{\int_0^\pi f(x)\,\cos 2x\,dx}{\int_0^\pi (\cos 2x)^2\,dx} = \frac{-\pi/2}{\pi/2} = -1$$

$$b_1 = \frac{\int_0^\pi f(x)\,\text{sen}2x\,dx}{\int_0^\pi (\text{sen}2x)^2\,dx} = \frac{0}{\pi/2} = 0$$

Y por lo tanto

$$P(x) = \frac{\pi^2}{6}\cdot 1 + (-1)\cdot \cos 2x + (0)\cdot \text{sen}2x$$

$$P(x) = \frac{\pi^2}{6} - \cos 2x$$


2.5
2.0
1.5
1.0
0.5
0.0
−2
0
2
4
6
f
P

Tema 4

Derivación e integración numérica

4.1. Derivación numérica: funciones de una variable

Problema 4.1:

Dada la función $f(x) = \text{arcsen}(x)$ y la tabla de valores

x	0,5	0,6	0,7
$y = \text{arcsen}(x)$	0,5236	0,6435	0,7554

(a) Calcular $f'(0{,}6)$ usando fórmulas de aproximación de la derivada de dos puntos progresiva, regresiva y centrada con $h = 0{,}1$. Si el valor exacto con cuatro cifras significativas es $f'(0{,}6) = 1{,}250$, calcular el error absoluto para cada una de las fórmulas.

(b) Estudiar el orden de las fórmulas del apartado anterior.

(a) Vamos a utilizar tres fórmulas aproximadas de derivación. Si $h > 0$

1. **Fórmula progresiva**

$$f'(a) \approx \frac{f(a+h) - f(a)}{h}$$

Si $a = 0{,}6$ y $h = 0{,}1$

$$f'(2{,}6) \approx \frac{f(0{,}6 + 0{,}1) - f(0{,}6)}{0{,}1} = \frac{f(0{,}7) - f(0{,}6}{0{,}1} = \frac{0{,}7554 - 0{,}6435}{0{,}1} = 1{,}119$$

El error absoluto es

$$E_a = |1{,}250 - 1{,}119| = 0{,}131$$

2. **Fórmula regresiva**

$$f'(a) \approx \frac{f(a) - f(a-h)}{h}$$

Si $a = 2{,}6$ y $h = 0{,}1$

$$f'(2{,}6) \approx \frac{f(2{,}6) - f(2{,}6 - 0{,}1)}{0{,}1} = \frac{f(2{,}6) - f(2{,}5)}{0{,}1} = \frac{0{,}6435 - 0{,}5236}{0{,}1} = 1{,}199$$

El error absoluto es

$$E_a = |1{,}250 - 1{,}199| = 0{,}051$$

3. **Fórmula centrada**

$$f'(a) \approx \frac{f(a+h) - f(a-h)}{2h}$$

Si $a = 2{,}6$ y $h = 0{,}1$

$$f'(2{,}6) \approx \frac{f(2{,}6+0{,}1) - f(2{,}6-0{,}1)}{2(0{,}1)} = \frac{f(2{,}7) - f(2{,}5}{0{,}2} = \frac{0{,}7754 - 0{,}5236}{0{,}2} = 1{,}259$$

El error absoluto es

$$E_a = |1{,}250 - 1{,}259| = 0{,}009$$

(b) Veamos el orden de las respectivas fórmulas.

1. **Fórmula progresiva**

$$f'(a) \approx \frac{f(a+h) - f(a)}{h}$$

La fórmula de Taylor se puede escribir

$$f(x) = f(x_0) + f'(x_0)\frac{x - x_0}{1!} + f''(c)\frac{(x - x_0)^2}{2!} \quad c \in (x, x_0) \text{ o } c \in (x_0, x)$$

Si $x_0 = a$ y $x = a + h$ entonces $h = x - x_0$ y podemos reescribir la fórmula como

$$f(a+h) = f(a) + f'(a)\frac{h}{1!} + f''(c)\frac{h^2}{2!} \quad c \in (a+h, a) \text{ o } c \in (a, a+h)$$

Si tenemos en cuenta la fórmula progresiva, restando $f(a)$ a los dos miembros

$$f(a+h) - f(a) = f'(a)\frac{h}{1!} + f''(a)\frac{h^2}{2!} + f'''(c)\frac{h^3}{3!}$$

y dividiendo por h

$$\frac{f(a+h) - f(a)}{h} = f'(a) + f''(c)\frac{h}{2}$$

Es decir, el error de la fórmula es de la forma

$$E_h = -f''(c)\frac{h}{2} = K\,h$$

y como el exponente de la h nos da el orden de la fórmula, esta es una fórmula de orden 1.

2. **Fórmula regresiva**

$$f'(a) \approx \frac{f(a) - f(a-h)}{h}$$

La fórmula de Taylor se puede escribir

$$f(x) = f(x_0) + f'(x_0)\frac{x - x_0}{1!} + f''(c)\frac{(x - x_0)^2}{2!} \quad c \in (x, x_0) \text{ o } c \in (x_0, x)$$

Si $x_0 = a$ y $x = a - h$ entonces $-h = x - x_0$ y podemos reescribir la fórmula como

$$f(a-h) = f(a) + f'(a)\frac{-h}{1!} + f''(c)\frac{(-h)^2}{2!} \quad c \in (a+h, a) \text{ o } c \in (a, a+h)$$

Si tenemos en cuenta la fórmula progresiva, restando $f(a)$ a los dos miembros

$$f(a-h) - f(a) = -f'(a)\frac{h}{1!} + f''(c)\frac{h^2}{2!}$$

y dividiendo por $-h$

$$\frac{f(a-h) - f(a)}{-h} = f'(a) - f''(c)\frac{h}{2}$$

que es

$$\frac{f(a) - f(a-h)}{h} = f'(a) - f''(c)\frac{h}{2}$$

Es decir, el error de la fórmula es de la forma

$$E_h = f''(c)\frac{h}{2} = K\,h$$

y como el exponente de la h nos da el orden de la fórmula, esta es una fórmula de orden 1.

3. **Fórmula centrada**

$$f'(a) \approx \frac{f(a+h) - f(a-h)}{2h}$$

La fórmula de Taylor se puede escribir

$$f(x) = f(x_0) + f'(x_0)\frac{x - x_0}{1!} + f''(x_0)\frac{(x-x_0)^2}{2!} + f'''(c)\frac{(x-x_0)^3}{3!}$$

con

$$c \in (x, x_0) \text{ o } c \in (x_0, x)$$

Si $x_0 = a$ y $x = a + h$ entonces $h = x - x_0$ y podemos reescribir la fórmula como

$$f(a+h) = f(a) + f'(a)\frac{h}{1!} + f''(a)\frac{h^2}{2!} + f'''(c_1)\frac{h^3}{3!}$$

Si $x_0 = a$ y $x = a - h$ entonces $-h = x - x_0$ y podemos reescribir la fórmula como

$$f(a-h) = f(a) + f'(a)\frac{-h}{1!} + f''(a)\frac{(-h)^2}{2!} + f'''(c_2)\frac{(-h)^3}{3!}$$

Y $-f(a-h)$ es

$$-f(a-h) = -f(a) + f'(a)\frac{h}{1!} - f''(a)\frac{h^2}{2!} + f'''(c_2)\frac{h^3}{3!}$$

Si tenemos en cuenta la fórmula centrada, sumando $f(a+h) + (-f(a-h))$ a los dos miembros y teniendo en cuenta, que si f''' es continua, por el teorema del valor intermedio, podemos encontrar un valor c_3 en (c_1, c_2) tal que

$$f'''(c_3) = \frac{f'''(c_1) + f'''(c_2)}{2}$$

se tiene

$$f(a+h) - f(a-h) = 2f'(a)h + 2f'''(c_3)\frac{h^3}{6}$$

y dividiendo por $2h$

$$\frac{f(a+h) - f(a-h)}{2h} = f'(a) + f'''(c_3)\frac{h^2}{6}$$

Es decir, el error de la fórmula es de la forma

$$E_h = -f'''(c)\frac{h^2}{6} = K\,h^2$$

y como el exponente de la h nos da el orden de la fórmula, esta es una fórmula de orden 2.

Problema 4.2:

Supongamos que tenemos tres puntos (x_0, y_0), (x_1, y_1) y (x_2, y_2) de una función f, de forma que $x_1 = x_0 + h$ y $x_2 = x_0 + 2h$ con $0 < h < 1$.

(a) Construir una fórmula numérica que aproxime la derivada primera, sea regresiva, utilice estos tres puntos y sea de orden 2.

(b) Demostrar que esta fórmula regresiva es de orden 2.

(c) Calcular $f'(1{,}3)$ usando esta fórmula, con $f(x) = \ln x$ y $h = 0{,}1$.

x	1,1	1,2	1,3	1,4	1,5
$\ln x$	0,09531	0,1823	0,2624	0,3365	0,4055

(d) Calcular $f'(2{,}5)$ usando esta fórmula, con $f(x) = \sqrt{x}$ y $h = 0{,}1$.

x	2,3	2,4	2,5	2,6	2,7
$y = \sqrt{x}$	1,52	1,55	1,58	1,61	1,64

(e) Construir una fórmula numérica de diferenciación progresiva, que utilice estos tres puntos y que sea de orden 2.

(f) Demostrar que esta fórmula progresiva es de orden 2.

(g) Calcular $f'(1{,}3)$ usando esta última fórmula, con $f(x) = \ln x$ y $h = 0{,}1$.

(h) Construir una fórmula que utilizando estos tres puntos aproxime la derivada segunda. Utilizando esta fórmula, calcular $f''(1{,}3)$ con $f(x) = \ln x$.

(a) Si usamos un polinomio de interpolación de f de segundo grado en x_0, x_1, x_2, siendo $y_j = f(x_j)$, el polinomio de interpolación en la forma de Newton es:

$$p(x) = [y_0] + [y_0, y_1](x - x_0) + [y_0, y_1, y_2](x - x_0)(x - x_1)$$

donde los coeficientes $[y_0]$, $[y_0, y_1]$ y $[y_0, y_1, y_2]$ se obtienen de la tabla de diferencias divididas

$$
\begin{array}{lll}
x & f(x) & \\
x_0 & [y_0] = y_0 & \\
 & & [y_0, y_1] = \dfrac{[y_1] - [y_0]}{x_1 - x_0} \\
x_1 & [y_1] = y_1 & \qquad\qquad [y_0, y_1, y_2] = \dfrac{[y_1, y_2] - [y_0, y_1]}{x_2 - x_0} \\
 & & [y_1, y_2] = \dfrac{[y_2] - [y_1]}{x_2 - x_1} \\
x_2 & [y_2] = y_2 &
\end{array}
$$

Calculemos los elementos de esta tabla. Necesitamos los elementos de arriba de cada columna como coeficientes del polinomio de Newton, pero necesitamos calcular los elementos de la tabla columna a columna.

$$[y_0, y_1] = \frac{y_1 - y_0}{x_1 - x_0} = \frac{y_1 - y_0}{h},$$

$$[y_1, y_2] = \frac{y_2 - y_1}{x_2 - x_1} = \frac{y_2 - y_1}{h},$$

Finalmente

$$[y_0, y_1, y_2] = \frac{[y_1, y_2] - [y_0, y_1]}{x_2 - x_0} = \frac{1}{2h}\Big(\frac{y_2 - y_1}{h} - \frac{y_1 - y_0}{h}\Big) = \frac{1}{2h^2}\big(y_0 - 2y_1 + y_2\big).$$

Y, por lo tanto, los coeficientes del polinomio de Newton son

$$[y_0] = y_0, \quad [y_0, y_1] = \frac{y_1 - y_0}{h}, \quad [y_0, y_1, y_2] = \frac{1}{2h^2}(y_0 - 2y_1 + y_2). \tag{4.1}$$

Y como el polinomio de Newton es

$$p(x) = [y_0] + [y_0, y_1](x - x_0) + [y_0, y_1, y_2](x - x_0)(x - x_1)$$

Si derivamos $p(x)$, teniendo en cuenta que los coeficientes $[y_0]$, $[y_0, y_1]$ y $[y_0, y_1, y_2]$ son valores constantes

$$f'(x) \approx p'(x) = [y_0, y_1] + [y_0, y_1, y_2]((x - x_0) + (x - x_1)), \tag{4.2}$$

y para el caso particular del punto x_2, que es el que corresponde a la fórmula regresiva

$$f'(x_2) \approx p'(x_2) = [y_0, y_1] + [y_0, y_1, y_2]((x_2 - x_0) + (x_2 - x_1)),$$

Si sustituimos ahora los coeficientes por los valores calculados,

$$f'(x_2) \approx \frac{y_1 - y_0}{h} + \frac{1}{2h^2}\big(y_0 - 2y_1 + y_2\big)(2h + h),$$

$$f'(x_2) \approx \frac{2y_1 - 2y_0}{2h} + \frac{3}{2h}\big(y_0 - 2y_1 + y_2\big),$$

$$f'(x_2) \approx \frac{y_0 - 4y_1 + 3y_2}{2h}.$$

(b) Comprobemos que la fórmula

$$f'(x_2) \approx \frac{y_0 - 4y_1 + 3y_2}{2h} \tag{4.3}$$

es de orden 2. Es decir, que el error es proporcional a h^2. La fórmula de Taylor se puede escribir

$$f(x) = f(x_2) + f'(x_2)\frac{x - x_2}{1!} + f''(x_2)\frac{(x - x_2)^2}{2!} + f'''(c)\frac{(x - x_2)^3}{3!} \quad c \in (x, x_2) \text{ o } c \in (x_2, x)$$

Si $x_2 = a$ y $x = a + h$ entonces $h = x - x_2$ y podemos reescribir la fórmula como

$$f(a + h) = f(a) + f'(a)\frac{h}{1!} + f''(a)\frac{h^2}{2!} + f'''(c)\frac{h^3}{3!} \quad c \in (a + h, a) \text{ o } c \in (a, a + h)$$

Como $a = x_2$ se tiene que $y_0 = f(a - 2h)$, $y_1 = f(a - h)$ y $y_2 = f(a)$.

$$\begin{aligned} f(a - 2h) &= f(a) + f'(a)\frac{(-2h)}{1!} + f''(a)\frac{(-2h)^2}{2!} + f'''(c_\text{»})\frac{(-2h)^3}{3!} \\ f(a - h) &= f(a) + f'(a)\frac{(-h)}{1!} + f''(a)\frac{(-h)^2}{2!} + f'''(c_1)\frac{(-h)^3}{3!} \\ f(a) &= f(a) \end{aligned}$$

Si tenemos en cuenta la fórmula (4.3)

$$\begin{aligned} f(a - 2h) &= f(a) + f'(a)\frac{(-2h)}{1!} + f''(a)\frac{(-2h)^2}{2!} + f'''(c_\text{»})\frac{(-2h)^3}{3!} \\ -4f(a - h) &= -4f(a) - 4f'(a)\frac{(-h)}{1!} - 4f''(a)\frac{(-h)^2}{2!} - 4f'''(c_1)\frac{(-h)^3}{3!} \\ 3f(a) &= 3f(a) \end{aligned}$$

Y sumando miembro a miembro tenemos

$$y_0 - 4y_1 + 3y_2 = 2h\, f'(a) + k_1\, f'''(c_3)\, h^3$$

Y dividiendo ambos términos por $2h$ y reordenando los términos

$$f'(a) = \frac{y_0 - 4y_1 + 3y_2}{2h} - k_1\, f'''(c_3)\, h^2$$

Es decir, el error de la fórmula es de la forma

$$E_h = -k_1\, f'''(c_3)\, h^2 = K\, h^2$$

y como el exponente de la h nos da el orden de la fórmula, esta es una fórmula de orden 2.

(c) Para calcular $f'(1{,}3)$ con la fórmula regresiva, usaremos los valores

k	0	1	2
x_k	1,1	1,2	1,3
$y_k = \ln x_k$	0,09531	0,1823	0,2624

Si $x_2 = 1{,}3$

$$f'(x_2) \approx \frac{y_0 - 4y_1 + 3y_2}{2h} = \frac{(0{,}09531) - 4(0{,}1823) + 3(0{,}2624)}{2(0{,}1)} = 0{,}7666$$

(Valor exacto 0,7692)

(d) Usaremos los valores

k	0	1	2
x_k	2,3	2,4	2,5
$y_k = \sqrt{x_k}$	1,52	1,55	1,58

Si $x_2 = 2{,}5$

$$f'(x_2) \approx \frac{y_0 - 4y_1 + 3y_2}{2h} = \frac{(1{,}52) - 4(1{,}55) + 3(1{,}58)}{2(0{,}1)} = 0{,}300$$

(Valor exacto 0.316)

(e) Para el caso particular del punto x_0, que es el que corresponde a la fórmula progresiva, teniendo en cuenta la fórmula de la derivada (4.2) y los valores de las diferencias divididas (4.1)

$$f'(x_0) \approx p'(x_0) = [y_0, y_1] + [y_0, y_1, y_2]((x_0 - x_0) + (x_0 - x_1)),$$

$$f'(x_0) \approx \frac{y_1 - y_0}{h} + \frac{1}{2h^2}\big(y_0 - 2y_1 + y_2\big)(-h),$$

$$f'(x_0) \approx \frac{-3y_0 + 4y_1 - y_2}{2h}.$$

(f) Comprobemos que la fórmula

$$f'(x_0) \approx \frac{-3y_0 + 4y_1 - y_2}{2h} \tag{4.4}$$

Es de orden 2. Es decir, que el error es proporcional a h^2. La fórmula de Taylor es

$$f(x) = f(x_0) + f'(x_0)\frac{x - x_0}{1!} + f''(x_0)\frac{(x - x_0)^2}{2!} + f'''(c)\frac{(x - x_0)^3}{3!} \quad c \in (x, x_0) \text{ or } c \in (x_0, x)$$

Si $x_0 = a$ y $x = a + h$ entonces $h = x - x_0$ y podemos reescribir la fórmula como

$$f(a + h) = f(a) + f'(a)\frac{h}{1!} + f''(a)\frac{h^2}{2!} + f'''(c)\frac{h^3}{3!} \quad c \in (a + h, a) \text{ o } c \in (a, a + h)$$

Como $a = x_0$ se tiene que $y_0 = f(a)$, $y_1 = f(a + h)$ y $y_2 = f(a + 2h)$.

$$\begin{aligned} f(a) &= f(a) \\ f(a + h) &= f(a) + f'(a)\frac{h}{1!} + f''(a)\frac{h^2}{2!} + f'''(c_1)\frac{h^3}{3!} \\ f(a + 2h) &= f(a) + f'(a)\frac{2h}{1!} + f''(a)\frac{(2h)^2}{2!} + f'''(c_2)\frac{(2h)^3}{3!} \end{aligned}$$

Si tenemos en cuenta la fórmula (4.4)

$$\begin{aligned} -3f(a) &= -3f(a) \\ 4f(a + h) &= 4f(a) + 4f'(a)\frac{h}{1!} + 4f''(a)\frac{h^2}{2!} + 4f'''(c_1)\frac{h^3}{3!} \\ -f(a + 2h) &= -f(a) - f'(a)\frac{2h}{1!} - f''(a)\frac{(2h)^2}{2!} - f'''(c_2)\frac{(2h)^3}{3!} \end{aligned}$$

Y sumando miembro a miembro tenemos

$$-3y_0 + 4y_1 - y_2 = 2h\, f'(a) + k_1\, f'''(c_3)\, h^3$$

Y dividiendo ambos términos por $2h$ y reordenando los términos

$$f'(a) = \frac{-3y_0 + 4y_1 - y_2}{2h} - k_1\, f'''(c_3)\, h^2$$

Es decir, el error de la fórmula es de la forma

$$E_h = -k_1\, f'''(c_3)\, h^2 = K\, h^2$$

y como el exponente de la h nos da el orden de la fórmula, esta es una fórmula de orden 2.

(g) Usaremos los valores

k	0	1	2
x	1,3	1,4	1,5
$\ln x$	0,2624	0,3365	0,4055

Si $x_0 = 1{,}3$

$$f'(x_0) \approx \frac{-3y_0 + 4y_1 - y_2}{2h} = \frac{-3(0{,}2624) + 4(0{,}3365) - (0{,}4055)}{2(0{,}1)} = 0{,}7665$$

(Valor exacto 0.7692)

(h) Para aproximar la derivada segunda, utilizaremos la derivada segunda del polinomio interpolante. Teniendo en cuenta que

$$f'(x) \approx p'(x) = [y_0, y_1] + [y_0, y_1, y_2]((x - x_0) + (x - x_1))$$

derivamos otra vez respecto a x

$$f''(x) \approx p''(x) = [y_0, y_1, y_2](1 + 1)$$

y sustituyendo los valores de las diferencias divididas (4.1)

$$f''(x) \approx \frac{1}{2h^2}\big(y_0 - 2y_1 + y_2\big)(2),$$

$$f''(x) \approx \frac{y_0 - 2y_1 + y_2}{h^2},$$

Que es un valor constante. Lo utilizaremos para aproximar la derivada segunda en el punto medio

$$f''(x_1) \approx \frac{y_0 - 2y_1 + y_2}{h^2},$$

Y si llamado $x_1 = a$ entonces $x_0 = a - h$ y $x_1 = a + h$ y una forma alternativa de la fórmula sería

$$f''(a) \approx \frac{f(a+h) - 2f(a) + f(a-h)}{h^2},$$

Para calcular $f''(1{,}3)$ usaremos los valores

x	1,2	1,3	1,4
$\ln x$	0,1823	0,2624	0,3365

y

$$f''(1{,}3) \approx \frac{f(1{,}3+0{,}1) - 2f(1{,}3) + f(1{,}3-0{,}1)}{0{,}1^2} = \frac{f(1{,}4) - 2f(1{,}3) + f(1{,}2)}{0{,}01}$$

Sustituyendo valores

$$f''(1{,}3) \approx \frac{0{,}3365 - 2(0{,}2624) + 0{,}1823}{0{,}01} = -0{,}6000$$

(El valor exacto con cuatro cifras es $-0{,}5917$)

4.2. Derivación numérica: funciones de dos variables

Problema 4.3:

Calcular una aproximación de la divergencia $\text{div}(\mathbf{f})$

(a) Dada la función $\mathbf{f}(x,y) = (f_1, f_2) = (x\,y, x^3 - y^2)$ para $(x,y) = (3,2)$ con $h_x = h_y = 0{,}1$.

(b) Dada la función $\mathbf{f}(x,y) = (f_1, f_2) = (x^2 + y, x^2 - y)$ para $(x,y) = (2,1)$ con $h_x = h_y = 0{,}1$.

(c) Dada la función $\mathbf{f}(x,y) = (xy, -xy)$ para para $(x_m, y_n) = (1,-1)$ con $h_x = h_y = 0{,}2$.

Si

$$\mathbf{f}(x,y) = (f_1(x,y), f_2(x,y))$$

se tiene que

$$\text{div}(\mathbf{f}) = \frac{\partial f_1}{\partial x} + \frac{\partial f_2}{\partial y}$$

(a) Podemos aproximar el primer término con la fórmula centrada de la derivada primera:

$$\begin{aligned}\partial_x f_1(x_m, y_n) \approx & \frac{f_1(x_m + h_x, y_n) - f_1(x_m - h_x, y_n)}{2h_x} = \frac{f_1(3 + 0{,}1, 2) - f_1(3 - 0{,}1, 2)}{2\,(0{,}1)} = \\ = & \frac{((3 + 0{,}1) \times 2) - ((3 - 0{,}1) \times 2)}{2\,(0{,}1)} = \frac{(3{,}1 \times 2) - (2{,}9 \times 2)}{2\,(0{,}1)} = \\ = & \frac{6{,}2 - 5{,}8}{0{,}2} = \frac{0{,}4}{0{,}2} = 2\end{aligned}$$

Y análogamente el segundo término

$$\begin{aligned}\partial_y f_2(x_m, y_n) \approx & \frac{f_2(x_m, y_n + h_y) - f_2(x_m, y_n - h_y)}{2h_y} = \frac{f_2(3, 2 + 0{,}1) - f_2(3, 2 - 0{,}1)}{2\,(0{,}1)} = \\ = & \frac{\left(3^3 - (2 + 0{,}1)^2\right) - \left(3^3 - (2 - 0{,}1)^2\right)}{2\,(0{,}1)} = \frac{(3^3 - 2{,}1^2) - (3^3 - 1{,}9^2)}{2\,(0{,}1)} = \\ = & \frac{22{,}59 - 23{,}39}{0{,}2} = \frac{-0{,}8}{0{,}2} = -4\end{aligned}$$

Por lo tanto

$$\text{div}(\mathbf{f}(3,2)) = \partial_x f_1(3,2) + \partial_y f_2(3,2) \approx 2 - 4 = -2$$

(b) El primer término será:

$$\begin{aligned}\partial_x f_1(x_m, y_n) \approx & \frac{f_1(x_m + h_x, y_n) - f_1(x_m - h_x, y_n)}{2h_x} = \frac{f_1(2 + 0{,}1, 1) - f_1(2 - 0{,}1, 1)}{2(0{,}1)} = \\ = & \frac{\left((2 + 0{,}1)^2 + 1\right) - \left((2 - 0{,}1)^2 + 1\right)}{2(0{,}1)} = \\ = & \frac{5{,}41 - 4{,}61}{0{,}2} = \frac{0{,}8}{0{,}2} = 4\end{aligned}$$

Y el segundo término

$$\begin{aligned}\partial_y f_2(x_m, y_n) \approx & \frac{f_2(x_m, y_n + h_y) - f_2(x_m, y_n - h_y)}{2h_y} = \frac{f_2(2, 1 + 0{,}1) - f_2(2, 1 - 0{,}1)}{2(0{,}1)} = \\ = & \frac{(2^2 - (1 + 0{,}1)) - (2^2 - (1 - 0{,}1))}{2(0{,}1)} = \\ = & \frac{2{,}9 - 3{,}1}{0{,}2} = \frac{-0{,}2}{0{,}2} = -1\end{aligned}$$

Por lo tanto

$$\operatorname{div}(\mathbf{f}(2, 1)) = \partial_x f_1(2, 1) + \partial_y f_2(2, 1) \approx 4 - 1 = 3$$

(c) El primer término será:

$$\begin{aligned}\partial_x f_1(x_m, y_n) \approx & \frac{f_1(x_m + h_x, y_n) - f_1(x_m - h_x, y_n)}{2h_x} = \frac{f_1(1 + 0{,}2, -1) - f_1(1 - 0{,}2, -1)}{2(0{,}2)} = \\ = & \frac{(1 + 0{,}2)(-1) - (1 - 0{,}2)(-1)}{2(0{,}2)} = \\ = & \frac{-1{,}2 - (-0{,}8)}{0{,}4} = \frac{-0{,}4}{0{,}4} = -1.\end{aligned}$$

Y el segundo término

$$\begin{aligned}\partial_y f_2(x_m, y_n) \approx & \frac{f_2(x_m, y_n + h_y) - f_2(x_m, y_n - h_y)}{2h_y} = \frac{f_2(1, -1 + 0{,}2) - f_2(1, -1 - 0{,}2)}{2(0{,}2)} = \\ = & \frac{-(1)(-1 + 0{,}2) - (-(1)(-1 - 0{,}2))}{2(0{,}2)} = \\ = & \frac{0{,}8 - 1{,}2}{0{,}4} = \frac{-0{,}4}{0{,}4} = -1\end{aligned}$$

Por lo tanto

$$\operatorname{div}(\mathbf{f}(1, -1)) = \partial_x f_1(1, -1) + \partial_y f_2(1, -1) \approx -1 - 1 = -2$$

Problema 4.4:

Calcular la aproximación de $\Delta f(x,y)$ para

(a) La función $f(x,y) = 2x^2 + 3y^2$, en $(x,y) = (1,0)$ con $h_x = h_y = 0{,}2$.

(b) La función $f(x,y) = x^2 + y^2$, en $(x,y) = (1,-1)$ con $h_x = h_y = 0{,}2$.

El Laplaciano viene dado por

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

(a) Si aproximamos la derivada segunda con la fórmula centrada, la primera componente del vector será:

$$\begin{aligned}\frac{\partial^2 f}{\partial x^2} \approx & \frac{f(x_m + h_x, y_n) - 2f(x_m, y_n) + f(x_m - h_x, y_n)}{h_x^2} = \\ = & \frac{f(1 + 0{,}2, 0) - 2f(1,0) + f(1 - 0{,}2, 0)}{(0{,}2)^2} = \\ = & \frac{\left(2(1 + 0{,}2)^2 + 3(0)^2\right) - 2\left(2(1)^2 + 3(0)^2\right) + \left(2(1 - 0{,}2)^2 + 3(0)^2\right)}{(0{,}2)^2} = \\ = & \frac{2{,}88 - 4 + 1{,}28}{0{,}04} = \frac{0{,}16}{0{,}04} = 4.\end{aligned}$$

Y la segunda componente

$$\begin{aligned}\frac{\partial^2 f}{\partial y^2} \approx & \frac{f(x_m, y_n + h_y) - 2f(x_m, y_n) + f(x_m, y_n - h_y)}{h_y^2} = \\ = & \frac{f(1, 0 + 0{,}2) - 2f(1,0) + f(1, 0 - 0{,}2)}{(0{,}2)^2} = \\ = & \frac{\left(2(1)^2 + 3(0{,}2)^2\right) - 2\left(2(1)^2 + (0)^2\right) + \left(2(1)^2 + 3(0 - 0{,}2)^2\right)}{(0{,}2)^2} = \\ = & \frac{2{,}12 - 4 + 2{,}12}{0{,}04} = \frac{0{,}24}{0{,}04} = 6\end{aligned}$$

Por lo tanto

$$\Delta f(1,0) \approx 4 + 6 = 10$$

(b) La primera componente del vector será:

$$\begin{aligned}\frac{\partial^2 f}{\partial x^2} \approx & \frac{f\left(x_m + h_x, y_n\right) - 2f\left(x_m, y_n\right) + f\left(x_m - h_x, y_n\right)}{h_x^2} = \\ = & \frac{f\left(1 + 0{,}2, -1\right) - 2f\left(1, -1\right) + f\left(1 - 0{,}2, -1\right)}{\left(0{,}2\right)^2} = \\ = & \frac{\left(\left(1 + 0{,}2\right)^2 + \left(-1\right)^2\right) - 2\left(\left(1\right)^2 + \left(-1\right)^2\right) + \left(\left(1 - 0{,}2\right)^2 + \left(-1\right)^2\right)}{\left(0{,}2\right)^2} = \\ = & \frac{2{,}44 - 4 + 1{,}64}{0{,}04} = \frac{0{,}08}{0{,}04} = 2.\end{aligned}$$

Y la segunda componente

$$\begin{aligned}\frac{\partial^2 f}{\partial y^2} \approx & \frac{f\left(x_m, y_n + h_y\right) - 2f\left(x_m, y_n\right) + f\left(x_m, y_n - h_y\right)}{h_y^2} = \\ = & \frac{f\left(1, -1 + 0{,}2\right) - 2f\left(1, -1\right) + f\left(1, -1 - 0{,}2\right)}{\left(0{,}2\right)^2} = \\ = & \frac{\left(\left(1\right)^2 + \left(-1 + 0{,}2\right)^2\right) - 2\left(\left(1\right)^2 + \left(-1\right)^2\right) + \left(\left(1\right)^2 + \left(-1 - 0{,}2\right)^2\right)}{\left(0{,}2\right)^2} = \\ = & \frac{1{,}64 - 4 + 2{,}44}{0{,}4} = \frac{0{,}08}{0{,}04} = 2\end{aligned}$$

Por lo tanto

$$\Delta f(1, -1) \approx 2 + 2 = 4$$

Problema 4.5:

Calcular la aproximación del gradiente de f $\nabla f(x_0, y_0)$ utilizando fórmulas centradas para

(a) La función $f(x, y) = x^2 + 3y$ en $(x_0, y_0) = (1, 0)$ con $h_x = h_y = 0{,}2$.

(b) La función $f(x, y) = x^2 + y^2$ en $(x_m, y_n) = (1, -1)$ con $h_x = h_y = 0{,}2$

El gradiente viene dado por

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$$

(a) Podemos aproximar el primer término con la fórmula centrada de la derivada primera:

$$\begin{aligned}\frac{\partial f}{\partial x}(x_m, y_n) \approx & \frac{f(x_m + h_x, y_n) - f(x_m - h_x, y_n)}{2h_x} = \frac{f(1+0{,}2, 0) - f(1-0{,}2, 0)}{2(0{,}2)} = \\ = & \frac{\left((1{,}2)^2 + 3(0)\right) - \left((0{,}8)^2 + 3(0)\right)}{2(0{,}2)} = \\ = & \frac{1{,}44 - 0{,}64}{0{,}4} = \frac{0{,}8}{0{,}4} = 2\end{aligned}$$

Y el segundo término

$$\begin{aligned}\frac{\partial f}{\partial y}(x_m, y_n) \approx & \frac{f(x_m, y_n + h_y) - f(x_m, y_n - h_y)}{2h_y} = \frac{f(1, 0+0{,}2) - f(1, 0-0{,}2)}{2(0{,}2)} = \\ = & \frac{(1^2 + 3(0{,}2)) - (1^2 - 3(0{,}2))}{2(0{,}2)} = \\ = & \frac{1{,}2}{0{,}4} = 3\end{aligned}$$

Por lo tanto

$$\nabla f(1,0) = (\partial_x f_1(1,0), \partial_y f_2(1,0)) \approx (2,3)$$

(b) La primera componente del vector será:

$$\begin{aligned}\partial_x f(x_m, y_n) \approx & \frac{f(x_m + h_x, y_n) - f(x_m - h_x, y_n)}{2h_x} = \frac{f(1+0{,}2, -1) - f(1-0{,}2, -1)}{2(0{,}2)} = \\ = & \frac{\left((1+0{,}2)^2 + (-1)^2\right) - \left((1-0{,}2)^2 + (-1)^2\right)}{2(0{,}2)} = \\ = & \frac{2{,}44 - 1{,}64}{0{,}4} = \frac{0{,}8}{0{,}4} = 2.\end{aligned}$$

Y la segunda componente

$$\begin{aligned}\partial_x f(x_m, y_n) \approx & \frac{f(x_m, y_n + h_y) - f(x_m, y_n - h_y)}{2h_y} = \frac{f(1, -1+0{,}2) - f(1, -1-0{,}2)}{2(0{,}2)} = \\ = & \frac{\left((1)^2 + (-1+0{,}2)^2\right) - \left((1)^2 + (-1-0{,}2)^2\right)}{2(0{,}2)} = \\ = & \frac{1{,}64 - 2{,}44}{0{,}4} = \frac{-0{,}8}{0{,}4} = -2\end{aligned}$$

Por lo tanto

$$\nabla f(1,-1) \approx (2,-2)$$

4.3. Integración numérica: fórmulas de Newton-Cotes

Problema 4.6:

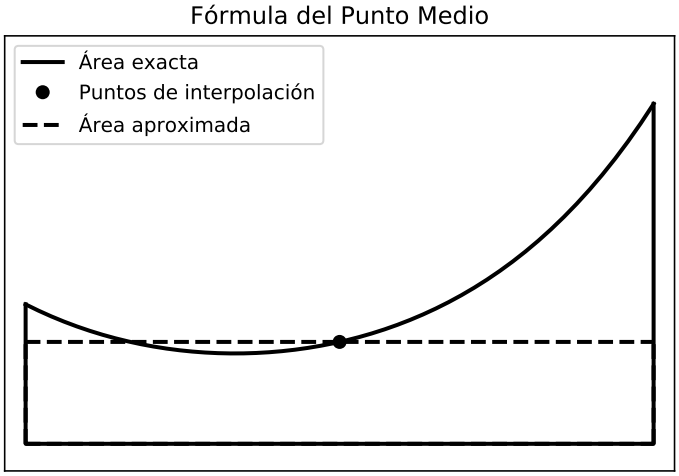
(a) Deducir la fórmula del Punto Medio Simple para aproximar la integral

$$\int_a^b f(x)\,dx$$

(b) Demostrar que su orden de precisión es 1.

(c) Aplicar la fórmula compuesta del Punto Medio con tres subintervalos para calcular las integrales

$$I_1 = \int_{-1,5}^{1,5} \left(2 + x^2\right) dx \qquad I_2 = \int_{-1}^{2} \left(1 + x^3\right) dx$$



(a) La fórmula del Punto Medio se obtiene integrando el polinomio de grado 0 que pasa por el punto medio del intervalo de integración. Si escribimos este polinomio en la forma de Lagrange

$$P_0(x) = f\left(\frac{a+b}{2}\right)$$

y entonces

$$\int_a^b f(x)\,dx \approx \int_a^b P_0(x)\,dx = \int_a^b f\left(\frac{a+b}{2}\right) dx =$$

$$= f\left(\frac{a+b}{2}\right) \int_a^b 1\,dx = f\left(\frac{a+b}{2}\right) [x]_a^b = f\left(\frac{a+b}{2}\right) (b-a)$$

Por lo tanto, la regla del punto medio simple es

$$\int_a^b f(x)\,dx \approx (b-a)\,f\left(\frac{a+b}{2}\right)$$

Y si llamamos $h = b-a$ a la longitud del intervalo de integración, podemos escribir la fórmula

$$\int_a^b f(x)\,dx \approx hf\left(\frac{a+b}{2}\right)$$

(b) Para que una fórmula de cuadratura sea exacta para un polinomio P_n de grado n, o lo que es lo mismo, tenga precisión n, dicha fórmula ha de ser exacta para las funciones 1, x, x^2, ..., x^n y no serlo para x^{n+1}.

Veamos la fórmula del Punto Medio:

¿Es exacta para $f(x)$ =1? Sí, porque

$$\int_a^b 1\,dx = b-a$$

y para este intervalo y esta función la fórmula del punto medio es

$$(b-a)\,f\left(\frac{a+b}{2}\right) = (b-a)\,1 = b-a$$

¿Es exacta para $f(x) = x$? Sí, porque

$$\int_a^b x\,dx = \frac{b^2-a^2}{2}$$

y para este intervalo y esta función la fórmula de los trapecios es

$$(b-a)\,f\left(\frac{a+b}{2}\right) = (b-a)\,\frac{a+b}{2} = \frac{(b+a)\,(b-a)}{2} = \frac{b^2-a^2}{2}$$

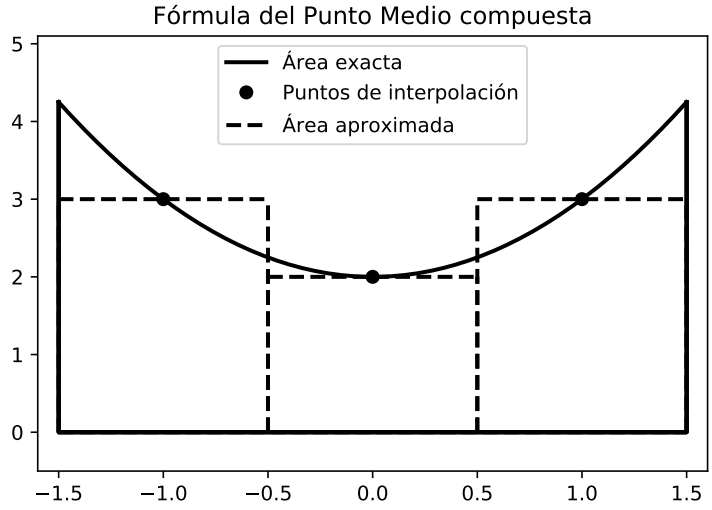
¿Es exacta para $f(x) = x^2$? No, porque

$$\int_a^b x^2\,dx = \frac{b^3-a^3}{3}$$

y para este intervalo y esta función la fórmula de los trapecios es

$$(b-a)\,f\left(\frac{a+b}{2}\right) = (b-a)\left(\frac{a+b}{2}\right)^2 = \frac{b^3+ab^2-a^2b-a^3}{4}$$

Como es exacta para 1 y x pero no para x^2, es exacta para polinomios de hasta grado 1 pero no para grado 2 y la precisión de la fórmula es 1.



(c) Aplicar la regla del Punto Medio Compuesta con tres subintervalos equivale a decir que dividamos el intervalo de intergración, en este caso $[-1{,}5, 1{,}5]$, en tres subintervalos iguales y apliquemos la fórmula del Punto Medio simple en cada uno de ellos. Para $n = 3$ intervalos, la longitud de cada subintervalo será

$$h = \frac{b-a}{3} = \frac{1{,}5-(-1{,}5)}{3} = \frac{3}{3} = 1$$

Los nodos se calculan

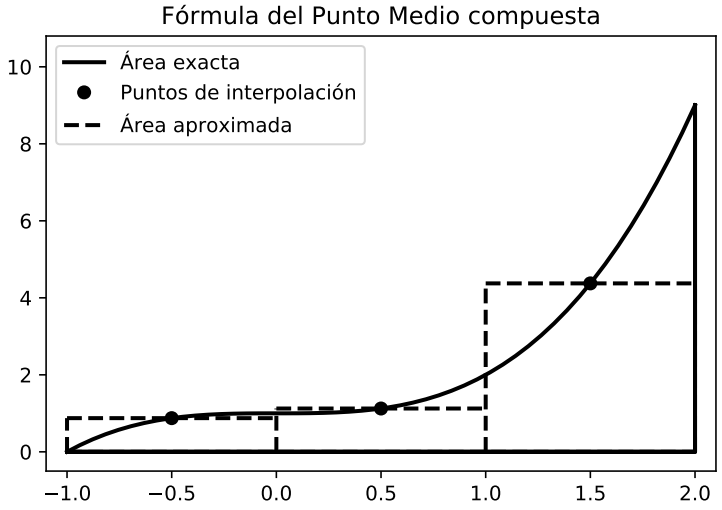
$$\begin{aligned} x_0 &= a + h/2 = -1{,}5 + 1/2 = -1 \\ x_1 &= x_0 + h = -1 + 1 = 0 \\ x_2 &= x_1 + h = 0 + 1 = 1 \end{aligned}$$

Aplicando la regla del Punto Medio a cada uno de los tres subintervalos tenemos

$$I_1 = \int_{-1{,}5}^{1{,}5} f(x)\,dx = \int_{-1{,}5}^{-0{,}5} f(x)\,dx + \int_{-0{,}5}^{0{,}5} f(x)\,dx + \int_{0{,}5}^{1{,}5} f(x)\,dx \approx$$

$$\approx h\,f(x_0) + h\,f(x_1) + h\,f(x_2) = h\,(f(x_0) + f(x_1) + f(x_2)) =$$

$$= h\left(\left(2 + x_0^2\right) + \left(2 + x_1^2\right) + \left(2 + x_2^2\right)\right) = 1\left(6 + (-1)^2 + 0^2 + 1^2\right) = 8$$



Para calcular I_2, dividimos el intervalo de integración $[-1, 2]$, en tres subintervalos iguales y aplicamos la fórmula del Punto Medio simple en cada uno de ellos. Para $n = 3$ intervalos, la longitud de cada subintervalo será

$$h = \frac{b-a}{3} = \frac{2-(-1)}{3} = 1$$

Los nodos se calculan

$$\begin{aligned} x_0 &= a + h/2 = -1 + 1/2 = -0{,}5 \\ x_1 &= x_0 + h = -0{,}5 + 1 = 0{,}5 \\ x_2 &= x_1 + h = 0{,}5 + 1 = 1{,}5 \end{aligned}$$

Aplicando la regla del Punto Medio a cada uno de los tres subintervalos tenemos

$$I_2 = \int_{-1}^{2} f(x)\,dx = \int_{-1}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{2} f(x)\,dx \approx$$

$$\approx h\,f(x_0) + h\,f(x_1) + h\,f(x_2) = h\,(f(x_0) + f(x_1) + f(x_2)) =$$

$$= h\left(\left(1 + x_0^3\right) + \left(1 + x_1^3\right) + \left(1 + x_2^3\right)\right) = 1\left(3 + (-0{,}5)^3 + 0{,}5^3 + 1{,}5^3\right) = 3{+}1{,}5^3 = 6{,}375$$

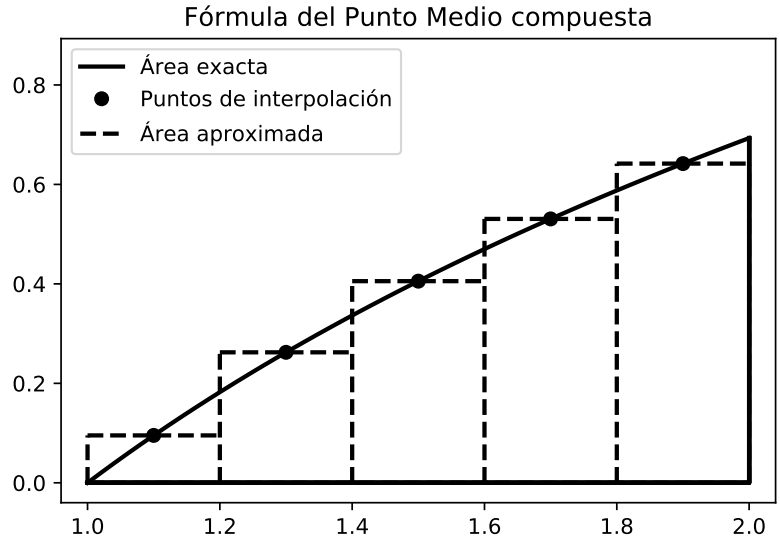
Problema 4.7:

Calcular, utilizando la regla del punto medio compuesta con 5 subintervalos, la integral.

$$I = \int_{1}^{2} \ln x\,dx.$$

La fórmula simple del punto medio, si $h = x_j - x_i$ y se tiene que $\bar{x}_j = \dfrac{x_i + x_j}{2}$

$$\int_{x_i}^{x_j} f(x)dx \approx hf(\bar{x}_j)$$



Dividimos el intervalo $[1, 2]$ en n subintervalos, cada uno de ellos de longitud h. Se verifica

$$h = \frac{b-a}{n} = \frac{2-1}{5} = \frac{1}{5} = 0{,}2.$$

Los extremos de los intervalos serán

$$\begin{aligned}
x_0 =& a = 1\\
x_1 =& x_0 + h = 1 + 0{,}2 = 1{,}2\\
x_2 =& x_1 + h = 1{,}2 + 0{,}2 = 1{,}4\\
x_3 =& x_2 + h = 1{,}4 + 0{,}2 = 1{,}6\\
x_4 =& x_3 + h = 1{,}6 + 0{,}2 = 1{,}8\\
x_5 =& x_4 + h = 1{,}8 + 0{,}2 = 2 = b
\end{aligned}$$

Y los puntos medios de los intervalos seránLos extremos de los intervalos serán

$$\begin{aligned}
\bar{x}_1 =& (x_0 + x_1)/2 = (1 + 1{,}2)/2 = 1{,}1\\
\bar{x}_2 =& (x_1 + x_2)/2 = (1{,}2 + 1{,}4)/2 = 1{,}3\\
\bar{x}_3 =& (x_2 + x_3)/2 = (1{,}4 + 1{,}6)/2 = 1{,}5\\
\bar{x}_4 =& (x_3 + x_4)/2 = (1{,}6 + 1{,}8)/2 = 1{,}7\\
\bar{x}_5 =& (x_1 + x_2)/2 = (1{,}8 + 2)/2 = 1{,}9
\end{aligned}$$

Aplicamos la fórmula simple del punto medio 5 veces:

$$\begin{aligned}
\int_1^2 \ln x dx =& \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \int_{x_2}^{x_3} f(x)dx + \int_{x_3}^{x_4} f(x)dx + \int_{x_4}^{x_5} f(x)dx \approx\\
\approx& h f(\bar{x}_1) + h f(\bar{x}_2) + h f(\bar{x}_3) + h f(\bar{x}_4) + h f(\bar{x}_5) =\\
=& h\left(f(\bar{x}_1) + f(\bar{x}_2) + f(\bar{x}_3) + f(\bar{x}_4) + f(\bar{x}_5)\right) =\\
=& 0{,}2\left(\ln 1{,}1 + \ln 1{,}3 + \ln 1{,}5 + \ln 1{,}7 + \ln 1{,}9\right) = 0{,}3871
\end{aligned}$$

(El valor exacto es 0,3863)

Problema 4.8:

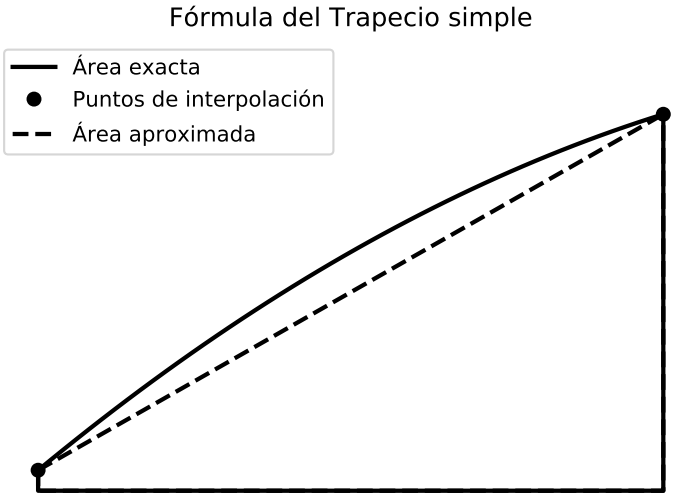
(a) Deducir la fórmula de los Trapecios Simple para aproximar la integral

$$\int_a^b f(x)\,dx$$

(b) Demostrar que su orden de precisión es 1.

(c) Aplicar la fórmula compuesta de los Trapecios con tres subintervalos al cálculo de la integral

$$\int_{-1}^{2} x^3 dx$$



(a) La fórmula de los Trapecios se obtiene integrando el polinomio de grado 1 que pasa por los extremos del intervalo de integración. Si escribimos este polinomio en la forma de Lagrange

$$P_1(x) = f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a}$$

y entonces

$$\int_a^b f(x)\,dx \approx \int_a^b P_1(x)\,dx = \frac{f(a)}{a-b}\int_a^b (x-b)\,dx + \frac{f(b)}{b-a}\int_a^b (x-a)\,dx =$$

$$= \frac{f(a)}{a-b}\left[\frac{(x-b)^2}{2}\right]_a^b + \frac{f(b)}{b-a}\left[\frac{(x-a)^2}{2}\right]_a^b = \frac{f(a)}{a-b}\,\frac{-(a-b)^2}{2} + \frac{f(b)}{b-a}\,\frac{(b-a)^2}{2} =$$

$$= f(a)\,\frac{-(a-b)}{2} + f(b)\,\frac{(b-a)}{2} = \frac{b-a}{2}\,(f(a)+f(b))$$

Por lo tanto, la regla del trapecio simple es

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}\left(f(a)+f(b)\right)$$

Y si llamamos $h = b-a$ a la longitud del intervalo de integración, podemos escribir la fórmula

$$\int_a^b f(x)\,dx \approx \frac{h}{2}\left(f(a)+f(b)\right)$$

(b) Para que una fórmula de cuadratura sea exacta para un polinomio P_n de grado n, o lo que es lo mismo, tenga precisión n, dicha fórmula ha de ser exacta para las funciones 1, x, x^2, ..., x^n y no serlo para x^{n+1}.

Veamos la fórmula de los Trapecios:

¿Es exacta para $f(x)$ =1? Sí, porque

$$\int_a^b 1dx = b-a$$

y para este intervalo y esta función la fórmula de los trapecios es

$$\frac{b-a}{2}\left(f(a)+f(b)\right) = \frac{b-a}{2}\left(1+1\right) = b-a$$

¿Es exacta para $f(x) = x$? Sí, porque

$$\int_a^b x\,dx = \frac{b^2-a^2}{2}$$

y para este intervalo y esta función la fórmula de los trapecios es

$$\frac{b-a}{2}\left(f(a)+f(b)\right) = \frac{b-a}{2}\left(a+b\right) = \frac{\left(b+a\right)\left(b-a\right)}{2} = \frac{b^2-a^2}{2}$$

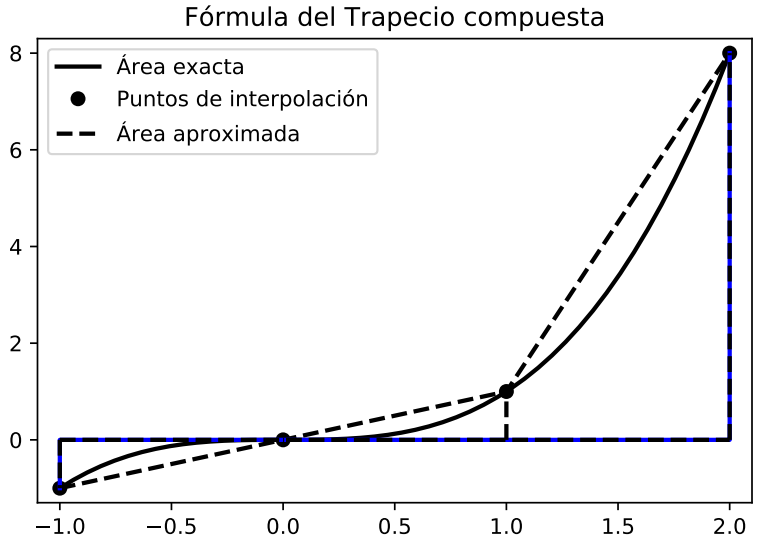
¿Es exacta para $f(x) = x^2$? No, porque

$$\int_a^b x^2\,dx = \frac{b^3-a^3}{3}$$

y para este intervalo y esta función la fórmula de los trapecios es

$$\frac{b-a}{2}\left(f(a)+f(b)\right) = \frac{b-a}{2}\left(a^2+b^2\right) = \frac{b^3-ab^2+a^2b-a^3}{2}$$

Como es exacta para 1 y x pero no para x^2, es exacta para polinomios de hasta grado 1 pero no para grado 2 y la precisión de la fórmula es 1.



(c) Aplicar la regla del Trapecio Compuesta con tres subintervalos equivale a decir que dividimos el intervalo de intergración, en este caso $[-1, 2]$, en tres subintervalos iguales y aplicamos la fórmula de los Trapecios simple en cada uno de ellos. Para $n = 3$ intervalos, la longitud de cada subintervalo será

$$h = \frac{b-a}{3} = \frac{2-(-1)}{3} = 1$$

Los nodos se calculan

$$\begin{aligned} x_0 &= a = -1 \\ x_1 &= x_0 + h = -1 + 1 = 0 \\ x_2 &= x_1 + h = 0 + 1 = 1 \\ x_3 &= x_2 + h = 1 + 1 = 2 \end{aligned}$$

Aplicando la regla del Trapecio a cada uno de los tres subintervalos tenemos

$$\int_{-1}^{2} f(x)\,dx = \int_{x_0}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \int_{x_2}^{x_3} f(x)\,dx \approx$$

$$\approx \frac{h}{2}\left(f(x_0) + f(x_1)\right) + \frac{h}{2}\left(f(x_1) + f(x_2)\right) + \frac{h}{2}\left(f(x_2) + f(x_3)\right) =$$

$$= \frac{h}{2}\left(f(x_0) + 2f(x_1) + 2f(x_2) + f(x_3)\right) = \frac{h}{2}\left(f(x_0) + 2\left(f(x_1) + f(x_2)\right) + f(x_3)\right) =$$

$$= \frac{1}{2}\left(x_0^3 + 2\left(x_1^3 + x_2^3\right) + x_3^3\right) = \frac{1}{2}\left((-1)^3 + 2\left(0^3 + 1^3\right) + 2^3\right) = \frac{1}{2}\left(-1 + 2 + 8\right) = 4{,}5$$

Problema 4.9:

Deducir la regla del trapecio compuesta a partir de la regla del trapecio simple.

La fórmula simple del trapecio, si $h = x_j - x_i$, es

$$\int_{x_i}^{x_j} f(x)dx \approx \frac{h}{2}\left(f(x_i) + f(x_j)\right)$$

Si dividimos el intervalo $[a, b]$ en n subintervalos, cada uno de ellos de longitud h. Se verifica

$$h = \frac{b-a}{n}$$

y entonces los nodos serían

$$x_i = a + i\,h \quad \text{con} \quad i = 0, 1, 2, \ldots, n$$

Si aplicamos la fórmula simple del trapecio n veces:

$$\begin{aligned}
\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \int_{x_2}^{x_3} f(x)dx + \cdots + \int_{x_{n-1}}^{x_n} f(x)dx + \int_{x_{n-1}}^{x_n} f(x)dx \approx \\
\approx \frac{h}{2}\left(f(x_0) + f(x_1)\right) + \frac{h}{2}\left(f(x_1) + f(x_2)\right) + \frac{h}{2}\left(f(x_2) + f(x_3)\right) + \\
+ \cdots + \frac{h}{2}\left(f(x_{n-2}) + f(x_{n-1})\right) + \frac{h}{2}\left(f(x_{n-1}) + f(x_n)\right) = \\
= \frac{h}{2}\left(f(x_0) + 2f(x_1) + 2f(x_2) + 2f(x_3) + \cdots + 2f(x_{n-1}) + f(x_n)\right) = \\
= \frac{h}{2}\left(f(x_0) + 2\left(f(x_1) + f(x_2) + f(x_3) + \cdots + f(x_{n-1})\right) + f(x_n)\right) = \\
= \frac{h}{2}\left(f(a) + 2\sum_{i=1}^{n-1} f(x_i) + f(b)\right)
\end{aligned}$$

y la fórmula del trapecio compuesta es

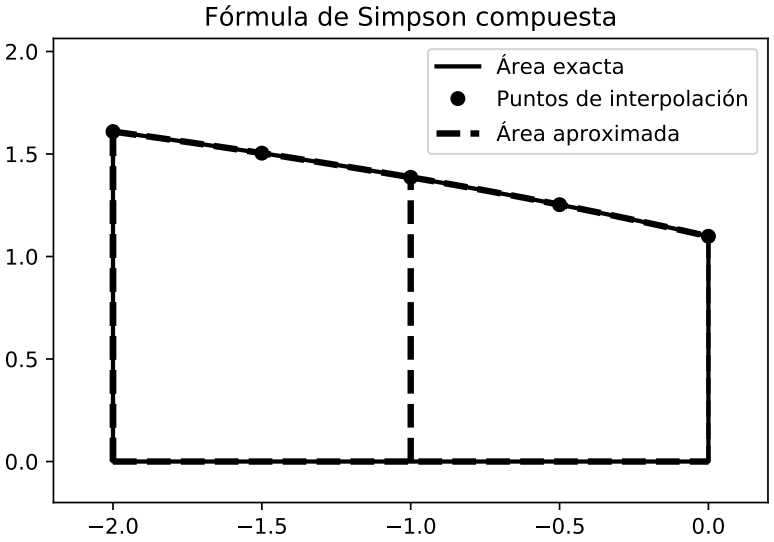
$$\int_a^b f(x)dx \approx \frac{h}{2}\left(f(a) + 2\sum_{i=1}^{n-1} f(x_i) + f(b)\right)$$

con $h = \dfrac{b-a}{n}$ y $x_i = a + i\,h$ para $i = 0, 1, 2, \ldots, n$.

Problema 4.10:

Aproximar, con la regla de Simpson compuesta, usando cinco nodos, la integral

$$\int_{-2}^{0} \ln(3-x)dx,$$



Aplicar la fórmula compuesta de Simpson equivale a:

1. Dividir el intervalo de integración en subintervalos iguales.
2. Aplicar la fórmula simple de Simpson a cada uno de ellos.
3. Sumar los resultados anteriores.

La fórmula simple del Simpson necesita tres nodos equiespaciados en el intervalo de integración y es

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left(f(a)+f\left(\frac{a+b}{2}\right)+f(b)\right)$$

Para tener 5 nodos igualmente separados hemos de dividir el intervalo $[-2,0]$ en 4 trozos de longitud

$$h=\frac{b-a}{4}=\frac{0-(-2)}{4}=0{,}5.$$

Los nodos serán entonces

$$\begin{aligned}
x_0 &= a = -2\\
x_1 &= x_0 + h = -2 + 0{,}5 = -1{,}5\\
x_2 &= x_1 + h = -1{,}5 + 0{,}5 = -1\\
x_3 &= x_3 + h = -1 + 0{,}5 = -0{,}5\\
x_4 &= x_3 + h = -0{,}5 + 0{,}5 = 0 = b
\end{aligned}$$

Cada vez que aplicamos la fórmula simple de Simpson utilizamos 3 nodos. Por lo tanto:

$$\int_{-2}^{0} \ln(3-x)dx \approx \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx =$$

$$\begin{aligned}
=&\frac{x_2-x_0}{6}\left(f(x_0)+4f(x_1)+f(x_2)\right)+\frac{x_4-x_2}{6}\left(f(x_2)+4f(x_3)+f(x_4)\right)=\\
=&\frac{2h}{6}\left(f(x_0)+4f(x_1)+f(x_2)\right)+\frac{2h}{6}\left(f(x_2)+4f(x_3)+f(x_4)\right)=\\
=&\frac{2h}{6}\left(f(x_0)+4\left(f(x_1)+f(x_3)\right)+2f(x_2)+f(x_4)\right)=\\
=&\frac{2(0{,}5)}{6}\left(\ln 5+4\left(\ln 4{,}5+\ln 3{,}5\right)+2\ln 4+\ln 3\right)=\\
=&\frac{1}{6}\left(1{,}60944+4\left(1{,}50408+1{,}25276\right)+2(1{,}38629)+1{,}09861\right)=2{,}75133
\end{aligned}$$

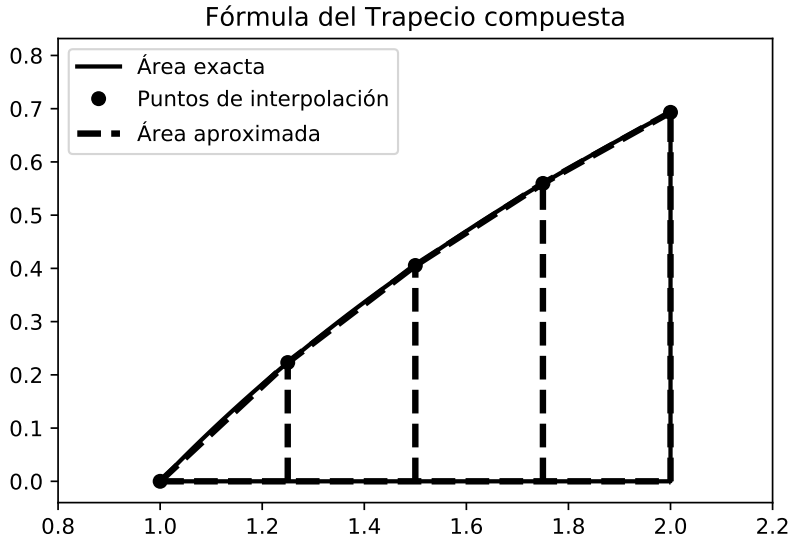
Problema 4.11:

Dada la integral

$$I=\int_1^2 \ln x\, dx.$$

(a) Calcularla utilizando 4 subintervalos y la regla del Trapecio compuesta.

(b) Determinar el número de subintervalos suficientes para que la fórmula del Trapecio compuesta proporcione un valor aproximado de I con un error menor que 10^{-2}.



(a) La regla del trapecio simple es

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}\left(f(a)+f(b)\right)$$

Y aplicar la fórmula compuesta de los trapecios equivale a:

1. Dividir el intervalo de integración en subintervalos iguales.
2. Aplicar la fórmula simple del trapecio a cada uno de ellos.
3. Sumar los resultados anteriores.

Por lo tanto, dividimos el intervalo $[1,2]$ en n subintervalos, cada uno de ellos de longitud h. Se verifica

$$h = \frac{b-a}{n} = \frac{2-1}{4} = \frac{1}{4} = 0{,}25.$$

Los nodos serán

$$\begin{aligned} x_0 =& a = 1\\ x_1 =& x_0 + h = 1 + 0{,}25 = 1{,}25\\ x_2 =& x_1 + h = 1{,}25 + 0{,}25 = 1{,}5\\ x_3 =& x_2 + h = 1{,}5 + 0{,}25 = 1{,}75\\ x_4 =& x_3 + h = 1{,}75 + 0{,}25 = 2 = b \end{aligned}$$

Aplicamos la fórmula simple del Trapecio 4 veces:

$$\begin{aligned} \int_1^2 \ln x dx \approx& \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \int_{x_2}^{x_3} f(x)dx + \int_{x_3}^{x_4} f(x)dx =\\ =& \frac{x_1-x_0}{2}\left(f(x_0)+f(x_2)\right) + \frac{x_2-x_1}{2}\left(f(x_1)+f(x_2)\right) +\\ &+ \frac{x_3-x_2}{2}\left(f(x_2)+f(x_3)\right) + \frac{x_4-x_3}{2}\left(f(x_3)+f(x_4)\right)\\ =& \frac{h}{2}\left(f(x_0) + 2\left(f(x_1)+f(x_2)+f(x_3)\right) + f(x_4)\right) =\\ =& \frac{0{,}25}{2}\left(\ln 1 + 2\left(\ln 1{,}25 + \ln 1{,}5 + \ln 1{,}75\right) + \ln 2\right) = 0{,}3837 \end{aligned}$$

(b) La fórmula del error de la regla de los Trapecios compuesta es:

$$E_h^T = -(b-a)\frac{h^2}{12} f''(c), \quad c \in (a,b)$$

En este caso, tenemos

$$a = 1,\ b = 2,\ h = \frac{b-a}{n} = \frac{1}{n}$$

Y además

$$f(c) = \ln c,\ f'(c) = \frac{1}{c},\ f''(c) = -\frac{1}{c^2}.$$

Y como $\left|f''(c)\right|$ es una función estrictamente decreciente en $[1,2]$ el valor en 1 es una cota superior del valor de la función en el intervalo

$$|f''(c)| = \frac{1}{c^2} < \frac{1}{1^2} = 1.$$

Utilizando la fórmula del error

$$\left|E_h^T\right| = (b-a)\frac{h^2}{12}\left|f''(c)\right| \leq \frac{h^2}{12} \times 1 = \frac{1}{12}\frac{1}{(n)^2} < 0{,}01,$$

$$\frac{1}{12 \times n^2} < 0{,}01$$

$$\frac{100}{12} < n^2$$

$$\sqrt{\frac{100}{12}} < n \Longrightarrow 2{,}88 < n$$

Y si tomamos $n = 3$ (aplicamos tres veces la regla de los trapecios simple) el número de nodos será $n + 1 = 4$ y podemos garantizar que el error al aproximar la integral con la regla de los Trapecios compuesta va a ser menor que 10^{-2}.

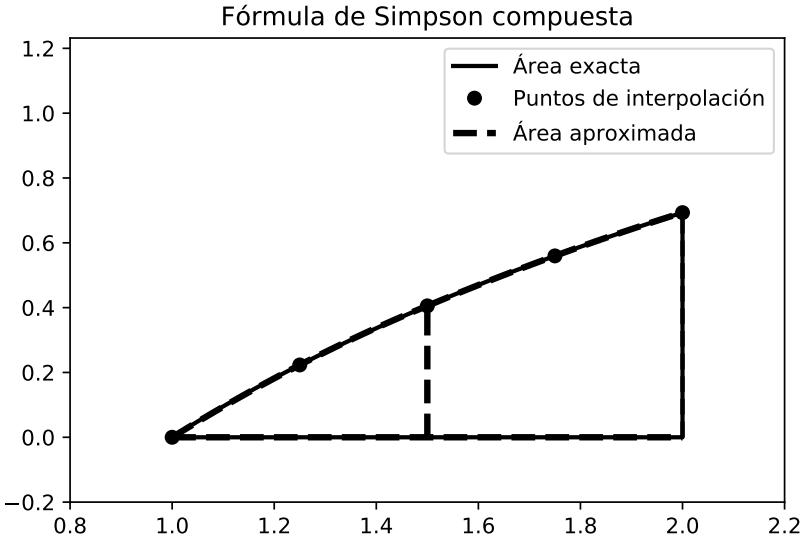
Problema 4.12:

Dada la integral

$$I = \int_1^2 \ln x \, dx.$$

(a) Calcularla utilizando 5 nodos y la regla de Simpson compuesta.

(b) Determinar el número de subintervalos, n, suficientes para que la fórmula de Simpson compuesta proporcione un valor aproximado de I con un error menor que 10^{-4}.



(a) Aplicar la fórmula compuesta de Simpson equivale a:

1. Dividir el intervalo de integración en subintervalos iguales.
2. Aplicar la fórmula simple de Simpson a cada uno de ellos.
3. Sumar los resultados anteriores.

La fórmula simple del Simpson necesita tres nodos equiespaciados en el intervalo de integración y es

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left(f(a)+f\left(\frac{a+b}{2}\right)+f(b)\right)$$

Por lo tanto, dividimos el intervalo $[1,2]$ en n subintervalos, cada uno de ellos de longitud $2h$. Se verifica

$$h=\frac{b-a}{2n}=\frac{2-1}{2(2)}=\frac{1}{4}=0{,}25.$$

Los nodos serán

$$\begin{aligned}
x_0 =& a=1\\
x_1 =& x_0+h=1+0{,}25=1{,}25\\
x_2 =& x_1+h=1{,}25+0{,}25=1{,}5\\
x_3 =& x_2+h=1{,}5+0{,}25=1{,}75\\
x_4 =& x_3+h=1{,}75+0{,}25=2=b
\end{aligned}$$

Aplicamos la fórmula simple de Simpson 2 veces:

$$\begin{aligned}
\int_1^2 \ln x dx \approx& \int_{x_0}^{x_2} f(x)dx+\int_{x_2}^{4_3} f(x)dx=\\
=&\frac{x_2-x_0}{6}\left(f(x_0)+4f(x_1)+f(x_2)\right)+\frac{x_4-x_2}{6}\left(f(x_2)+4f(x_3)+f(x_4)\right)=\\
=&\frac{2h}{6}\left(f(x_0)+4\left(f(x_1)+f(x_3)\right)+2f(x_2)+f(x_4)\right)=\\
=&\frac{2(0{,}25)}{6}\left(\ln 1+4\left(\ln 1{,}25+\ln 1{,}75\right)+2\ln 1{,}5+\ln 2\right)=0{,}3863
\end{aligned}$$

(b) La fórmula del error de la regla de Simpson es:

$$E_h^S=-(b-a)\frac{h^4}{180}f^{(4)}(c),\quad c\in(a,b)$$

En este caso, tenemos

$$a=1,\ b=2,\ h=\frac{b-a}{2n}=\frac{1}{2n}$$

Y además

$$f(c)=\ln c,\ f'(c)=\frac{1}{c},\ f''(c)=-\frac{1}{c^2},\ f'''(c)=\frac{2}{c^3},\ f^{(4)}(c)=-\frac{6}{c^4}.$$

Y como $\left|f^{(4)}(c)\right|$ es una función estrictamente decreciente en $[1,2]$ el valor en 1 es una cota superior del valor de la función en el intervalo

$$\left|f^{(4)}(c)\right| = \frac{6}{c^4} < \frac{6}{1^4} = 6.$$

Utilizando la fórmula del error

$$\left|E_h^S\right| = (b-a)\frac{h^4}{180}\left|f^{(4)}(c)\right| \leq \frac{h^4}{180} \times 6 = \frac{1}{15}\frac{1}{(2n)^4} < 0{,}0001,$$

$$\frac{1}{15 \times 16 \times n^4} < 0{,}0001$$

$$\frac{10000}{15 \times 16} < n^4$$

$$\sqrt[4]{\frac{10000}{15 \times 16}} < n \Longrightarrow 2{,}54 < n$$

Y si tomamos $n = 3$ (aplicamos tres veces la regla de Simpson simple) el número de nodos será $2n+1 = 7$ y podemos garantizar que el error al aproximar la integral con la regla del Simpson compuesta va a ser menor que 10^{-4}.

4.4. Integración numérica: fórmulas gaussianas

Problema 4.13:

Sea f una función continua:

(a) Obtener ω_0 y x_0 para que la fórmula de cuadratura

$$\int_{-1}^{1} f(x)\,dx \simeq \omega_0\, f(x_0)$$

tenga grado de precisión al menos uno.

(b) ¿Cuál es su grado de precisión?

(c) Usándola, calcular

$$\int_0^{\frac{\pi}{2}} \sin(x)\,dx$$

(a) Para que una fórmula de cuadratura tenga precisión n, dicha fórmula ha de ser exacta para las funciones 1, x, x^2, ..., x^n y no serlo para x^{n+1}. Si buscamos una fórmula de precisión al menos 1, habrá de ser exacta para las funciones 1 y x. Si imponemos estas dos condiciones a la fórmula

$$\begin{array}{lll} \textit{Exacta para } f(x)=1 & \int_{-1}^{1} 1dx = 2 = \omega_0\, f(x_0) = \omega_0 & \omega_0 = 2 \\ \textit{Exacta para } f(x)=x & \int_{-1}^{1} x dx = 0 = \omega_0\, f(x_0) = \omega_0\, x_0 & \omega_0\, x_0 = 0 \end{array}$$

Por lo tanto

$$\omega_0 = 2 \quad x_0 = 0$$

y la fórmula es

$$\int_{-1}^{1} f(x)\,dx \approx 2\, f\,(0)$$

que es la *fórmula de cuadratura gaussiana con un nodo*. (Vemos que coincide con la fórmula del punto medio).

(b) Sabemos que tiene precisión al menos 1. Estudiemos qué precisión tiene:

$$\textit{¿Exacta para } f(x)=x^2\textit{?} \qquad \int_{-1}^{1} x^2 dx = \frac{2}{3} \stackrel{?}{=} 2\, f(0) = 2 \times 0^2 = 0$$

Por lo tanto la fórmula no es exacta para polinomios de grado 2. Como es exacta para 1 y x pero no para x^2, es exacta para polinomios de hasta grado 1 pero no para grado 2 y la precisión de la fórmula es 1.

(c) Queremos hacer un cambio de variable que nos lleve del intervalo $[a, b]$ al intervalo donde está definida la fórmula de cuadratura gaussiana $[-1, 1]$. Si hacemos un cambio de variable lineal de la forma

$$x = m\,t + n,$$

si $x = a$, entonces $t = -1$ y

$$a = -m + n.$$

Y si $x = b$, entonces $t = 1$ y

$$b = m + n.$$

La solución a este sistema es

$$m = \frac{b-a}{2} \quad n = \frac{a+b}{2}.$$

Y por lo tanto, el cambio de variable es

$$x = \frac{b-a}{2}t + \frac{a+b}{2} \qquad dx = \frac{b-a}{2}dt.$$

Y

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{a+b}{2}\right)dt$$

Y utilizando la fórmula de cuadratura gaussiana de un nodo, la aproximación será

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}\,\omega_0\, f\left(\frac{b-a}{2}t_0 + \frac{a+b}{2}\right)$$

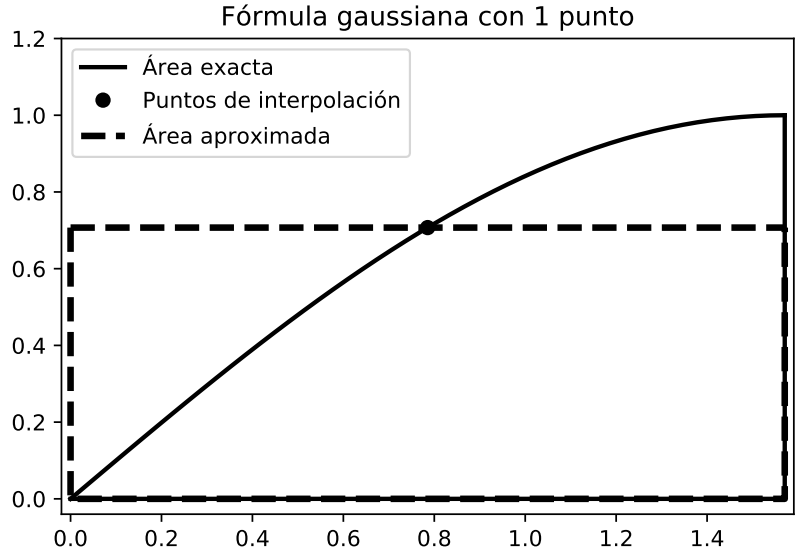
En este caso, como $[a, b] = [0, \frac{\pi}{2}]$

$$\frac{b-a}{2} = \frac{\pi}{4} \qquad \frac{a+b}{2} = \frac{\pi}{4}$$

y tenemos que

$$\int_0^{\frac{\pi}{2}} f(x)\,dx = \left(\frac{\pi}{4}\right)(2)\int_{-1}^{1} f\left(\frac{\pi}{4}\,t + \frac{\pi}{4}\right)dt \approx$$

$$\approx \left(\frac{\pi}{4}\right)(2)\, f\left(\frac{\pi}{4}(0) + \frac{\pi}{4}\right) = \left(\frac{\pi}{4}\right)(2)\, f\left(\frac{\pi}{4}\right) = \left(\frac{\pi}{4}\right)(2)\sin\left(\frac{\pi}{4}\right) = \left(\frac{\pi}{4}\right)(2)\,\frac{\sqrt{2}}{2} = 1{,}11$$



Problema 4.14:

La fórmula de cuadratura de Gauss-Legendre con dos nodos es

$$\int_{-1}^{1} f(x)\,dx \simeq f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) \tag{4.5}$$

(a) Estudiar cuál es su grado de precisión.

(b) Usar la fórmula obtenida para calcular un valor aproximado de:

$$\int_{-1}^{2} (x^3 - x^2 + 2)\,dx$$

realizando previamente un cambio de variable adecuado

(c) ¿Qué error se comete al aplicar dicha fórmula en este ejercicio? Justificar la respuesta sin hacer la integral exacta.

(a) Para que una fórmula de cuadratura tenga precisión n, dicha fórmula ha de ser exacta para las funciones 1, x, x^2, ..., x^n y no serlo para x^{n+1}. Estudiemos la precisión de esta fórmula

¿Exacta para $f(x) = 1$?

$$\int_{-1}^{1} 1\,dx = 2 \stackrel{?}{=} f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) = 1 + 1 = 2 \qquad Si$$

¿Exacta para $f(x) = x$?

$$\int_{-1}^{1} x\,dx = 0 \stackrel{?}{=} f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) = -\frac{1}{\sqrt{3}} + \frac{1}{\sqrt{3}} = 0 \qquad Si$$

¿Exacta para $f(x) = x^2$?

$$\int_{-1}^{1} x^2 dx = \frac{2}{3} \stackrel{?}{=} f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) = \left(-\frac{1}{\sqrt{3}}\right)^2 + \left(\frac{1}{\sqrt{3}}\right)^2 = \frac{2}{3} \qquad Si$$

¿Exacta para $f(x) = x^3$?

$$\int_{-1}^{1} x^3 dx = 0 \stackrel{?}{=} f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) = \left(-\frac{1}{\sqrt{3}}\right)^3 + \left(\frac{1}{\sqrt{3}}\right)^3 = 0 \qquad Si$$

¿Exacta para $f(x) = x^4$?

$$\int_{-1}^{1} x^4 dx = \frac{2}{5} \stackrel{?}{=} f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) = \left(-\frac{1}{\sqrt{3}}\right)^4 + \left(\frac{1}{\sqrt{3}}\right)^4 = \frac{2}{9} \qquad No$$

Como es exacta para 1, x, x^2 y x^3 pero no para x^4, es exacta para polinomios de hasta grado 3 pero no para grado 4 y la precisión de la fórmula es 3.

(b) Queremos hacer un cambio de variable que nos lleve del intervalo $[a, b]$ al intervalo donde está definida la fórmula de cuadratura gaussiana $[-1, 1]$. Si hacemos un cambio de variable lineal de la forma

$$x = m\,t + n,$$

si $x = a$, entonces $t = -1$ y

$$a = -m + n.$$

Y si $x = b$, entonces $t = 1$ y

$$b = m + n.$$

La solución a este sistema es

$$m = \frac{b-a}{2} \quad n = \frac{a+b}{2}.$$

Y por lo tanto, el cambio de variable es

$$x = \frac{b-a}{2}t + \frac{a+b}{2} \qquad dx = \frac{b-a}{2}dt.$$

Y

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{a+b}{2}\right)dt$$

Y utilizando la fórmula de cuadratura gaussiana de dos nodos, la aproximación será

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}\sum_{i=0}^{1}\omega_i\, f\left(\frac{b-a}{2}t_i + \frac{a+b}{2}\right)$$

siendo ω_i los pesos y t_i los nodos de la fórmula. En este caso, como $[a, b] = [-1, 2]$

$$\frac{b-a}{2} = 1{,}5 \qquad \frac{a+b}{2} = 0{,}5$$

y tenemos que

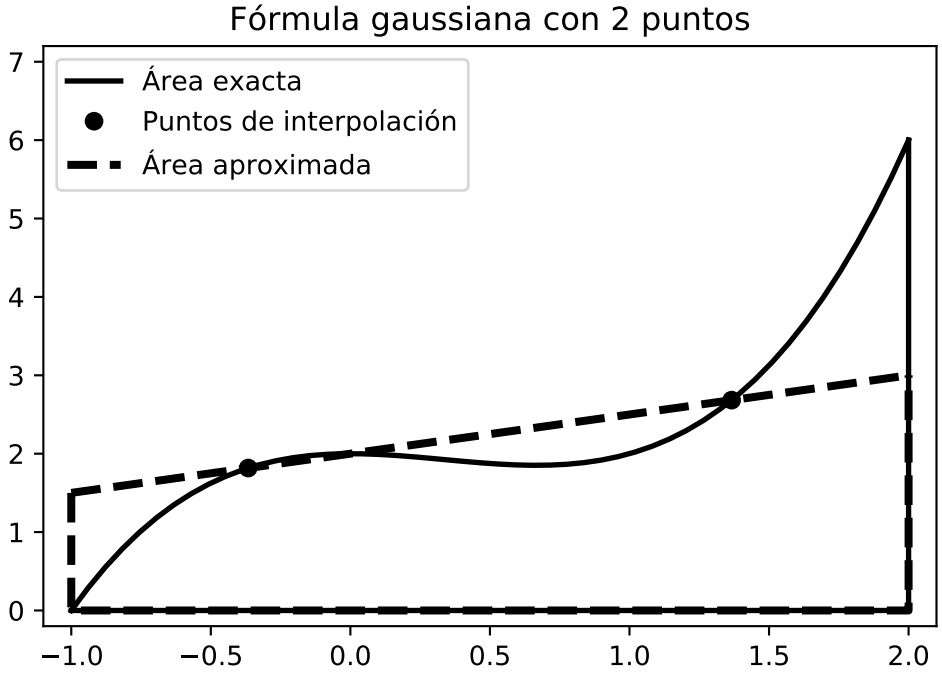
$$\int_{-1}^{2} f(x)\,dx = 1{,}5\int_{-1}^{1} f\left(1{,}5\,t + 0{,}5\right)dt \approx$$

$$\approx 1{,}5\left(f\left(-1{,}5\frac{1}{\sqrt{3}} + 0{,}5\right) + f\left(1{,}5\frac{1}{\sqrt{3}} + 0{,}5\right)\right) =$$

$$= 1{,}5\left(f(-0{,}366) + f(1{,}366)\right) =$$

$$= 1{,}5\left(1{,}817 + 2{,}683\right) = 6{,}75$$

(c) Como la fórmula tiene precisión 3 y la función integrando es un polinomio de grado 3, la integral es exacta y el error es cero.



Problema 4.15:

Calcular

$$\int_0^1 \cos(x^2)dx$$

usando la fórmula de cuadratura gaussiana con tres nodos

$$\int_{-1}^{1} f(x)\,dx \approx \frac{5}{9}\,f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}\,f\left(0\right) + \frac{5}{9}\,f\left(\sqrt{\frac{3}{5}}\right)$$

Queremos hacer un cambio de variable que nos lleve del intervalo $[a, b]$ al intervalo donde está definida la fórmula de cuadratura gaussiana $[-1, 1]$. Si hacemos un cambio de variable lineal de la forma

$$x = m\,t + n,$$

si $x = a$, entonces $t = -1$ y

$$a = -m + n.$$

Y si $x = b$, entonces $t = 1$ y

$$b = m + n.$$

La solución a este sistema es

$$m = \frac{b-a}{2} \quad n = \frac{a+b}{2}.$$

Y por lo tanto, el cambio de variable es

$$x = \frac{b-a}{2}t + \frac{a+b}{2} \qquad dx = \frac{b-a}{2}dt.$$

Y

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}t + \frac{a+b}{2}\right)dt$$

Y utilizando la fórmula de cuadratura gaussiana de tres nodos, la aproximación será

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}\sum_{i=0}^{2}\omega_i\, f\left(\frac{b-a}{2}t_i + \frac{a+b}{2}\right)$$

siendo ω_i los pesos y t_i los nodos de la fórmula. En este caso, como $[a,b] = [0,1]$

$$\frac{b-a}{2} = 0{,}5 \qquad \frac{a+b}{2} = 0{,}5$$

y tenemos que

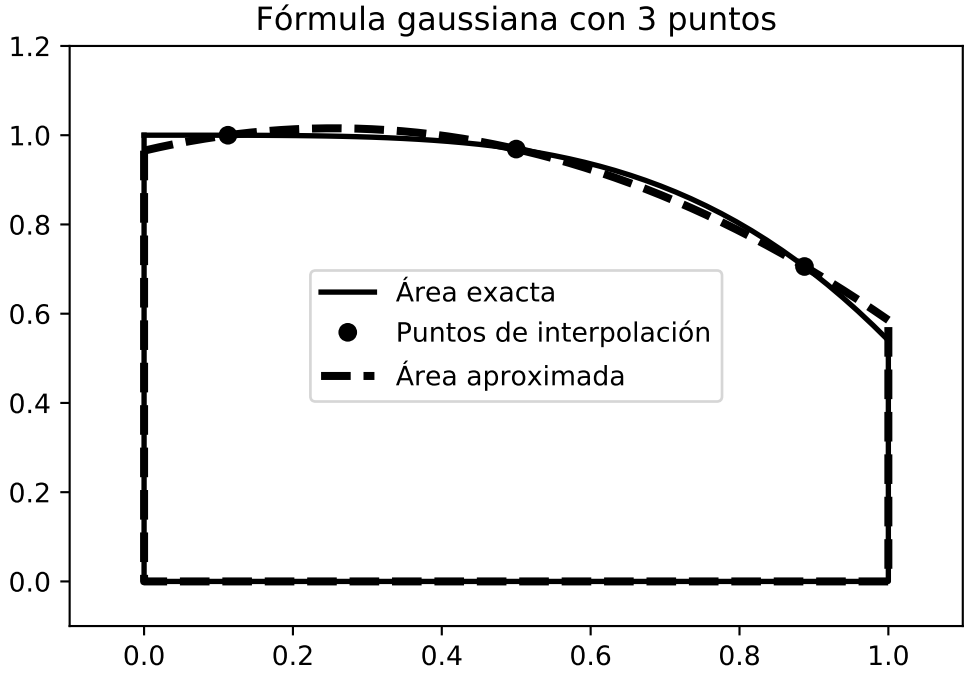
$$\int_0^1 f(x)\,dx = 0{,}5\int_{-1}^{1} f\left(0{,}5\,t + 0{,}5\right)dt \approx$$

$$\approx 0{,}5\left(\frac{5}{9}\,f\left(-0{,}5\sqrt{\frac{3}{5}} + 0{,}5\right) + \frac{8}{9}\,f\left(0{,}5(0) + 0{,}5\right) + \frac{5}{9}\,f\left(0{,}5\sqrt{\frac{3}{5}} + 0{,}5\right)\right) =$$

$$= 0{,}5\left(\frac{5}{9}\,f(0{,}1127) + \frac{8}{9}\,f(0{,}5) + \frac{5}{9}\,f(0{,}8873)\right) =$$

$$0{,}5\left(\frac{5}{9}\,\cos(0{,}1127^2) + \frac{8}{9}\,\cos(0{,}5^2) + \frac{5}{9}\,\cos(0{,}8873^2)\right) = 0{,}9044$$

(Solución exacta $I_e = 0{,}9045$)



Tema 5

Sistemas de ecuaciones lineales

5.1. Métodos directos

Problema 5.1:

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 1 \\ 1 & 1 & 1 & 2 \\ 2 & 1 & 1 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 4 \\ 1 \\ 4 \end{pmatrix}$$

(a) Calcular $\mathbf{x}$ utilizando el método de Gauss.

(b) Calcular el determinante de A utilizando el método de Gauss.

(a) El método para resolución de sistemas de Gauss consta de dos pasos:

1. Triangularización de la matriz A.
2. Resolución del sistema triangular equivalente por sustitución regresiva.

Resolvamos el sistema.

1. Construimos la matriz aumentada y hacemos ceros por debajo del pivote, a_{11}, sumando la primera fila multiplicada por un factor. Este factor se construye dividiendo el elemento de la fila debajo del pivote dividido por el pivote.

$$\begin{matrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{matrix} \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 2 & 1 & 0 & 1 & 4 \\ 1 & 1 & 1 & 2 & 1 \\ 2 & 1 & 1 & 1 & 4 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \leftarrow & f_1 \\ f_2 & \leftarrow & f_2 - \boxed{2/1}\ f_1 \\ f_3 & \leftarrow & f_3 - \boxed{1/1}\ f_1 \\ f_4 & \leftarrow & f_4 - \boxed{2/1}\ f_1 \end{array}$$

La matriz transformada es la siguiente.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \\ f_4 \end{array} \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & -2 & -1 & 2 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & -1 & -1 & 2 \end{pmatrix}$$

El pivote es ahora a_{22}. Hacemos ceros por debajo del pivote

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & -2 & -1 & 2 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & -1 & -1 & 2 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \leftarrow & f_1 \\ f_2 & \leftarrow & f_2 \\ f_3 & \leftarrow & f_3 - \boxed{1/1}\ f_2 \\ f_4 & \leftarrow & f_4 - \boxed{1/1}\ f_2 \end{array}$$

La matriz transformada es la siguiente.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \\ f_4 \end{array} \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & -2 & -1 & 2 \\ 0 & 0 & 2 & 2 & -2 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Ahora el pivote es a_{33}. Hacemos ceros por debajo del pivote.

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & -2 & -1 & 2 \\ 0 & 0 & 2 & 2 & -2 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \leftarrow & f_1 \\ f_2 & \leftarrow & f_2 \\ f_3 & \leftarrow & f_3 \\ f_4 & \leftarrow & f_4 - \boxed{1/2}\ f_3 \end{array}$$

La matriz transformada es la siguiente.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \\ f_4 \end{array} \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & -2 & -1 & 2 \\ 0 & 0 & 2 & 2 & -2 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

Y las matrices transformadas A' y b' son:

$$A' = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & -2 & -1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & -1 \end{pmatrix} \qquad b' = \begin{pmatrix} 1 \\ 2 \\ -2 \\ 1 \end{pmatrix}$$

2. Resolvemos el sistema triangular superior, equivalente al sistema inicial, $A'x = b'$ y obtenemos x, que era lo que buscábamos.

$$\begin{array}{rrrcrcl} x_1 & & +x_3 & +x_4 & = & 1 & x_4 = -1 \\ & x_2 & -2x_3 & -x_4 & = & 2 & x_3 = (-2-2x_4)/2 = (-2+2)/2 = 0 \\ & & 2x_3 & +2x_4 & = & -2 & x_2 = 2+2x_3+x_4 = 2+2(0)-1 = 1 \\ & & & -x_4 & = & 1 & x_1 = 1-x_3-x_4 = 1-0+1 = 2 \end{array}$$

Y la solución del sistema $Ax = b$ es $x^T = (2, 1, 0, -1)$

(b) Cálculo del determinante de A. El método de Gauss para el cálculo de determinantes consta de dos pasos:

1. Triangularización de la matriz.
2. Cálculo del determinante multiplicando los elementos de la diagonal de la matriz triangular.

Como solo se han realizado operaciones por filas, $f_i \leftarrow f_i + \lambda f_j$, el determinante de A y el de la matriz transformada A' es el mismo.
El primer paso ya lo hicimos en el apartado anterior por lo que nos queda únicamente multiplicar los elementos de la diagonal de A'

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

y el determinante es

$$|A| = (1) \times (1) \times (2) \times (-1) = -2$$

Problema 5.2:

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 1 \\ 1 & 1 & 1 & 2 \\ 3 & 3 & 3 & 3 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 2 \\ 4 \\ 4 \\ 6 \end{pmatrix}$$

(a) Calcular $\mathbf{x}$ utilizando Gauss con pivote parcial.

(b) Calcular el determinante de A utilizando el método de Gauss con pivote parcial.

(a) El método de Gauss con pivote parcial para la resolución de sistemas lineales consta de dos pasos:

1. Triangularización de la matriz A usando la estrategia del pivote parcial.
2. Resolución del sistema triangular equivalente por sustitución regresiva.

Resolvamos el sistema.

1. Construimos la matriz aumentada. Escogemos el pivote entre el elemento de la diagonal principal (pivote por defecto) y los elementos que están debajo, es decir, a_{11}. Seleccionamos como pivote, el mayor en valor absoluto. En este caso, en la columna 1 este elemento es a_{41}. Intercambiamos las filas

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 2 \\ 2 & 1 & 0 & 1 & 4 \\ 1 & 1 & 1 & 2 & 4 \\ 3 & 3 & 3 & 3 & 6 \end{pmatrix} \quad \begin{matrix} f_1 \leftarrow f_4 \\ \\ \\ f_4 \leftarrow f_1 \end{matrix}$$

Hacemos ceros por debajo del pivote sumando la primera fila multiplicada por un número.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \\ f_4 \end{array} \begin{pmatrix} 3 & 3 & 3 & 3 & 6 \\ 2 & 1 & 0 & 1 & 4 \\ 1 & 1 & 1 & 2 & 4 \\ 1 & 0 & 1 & 1 & 2 \end{pmatrix} \quad \begin{array}{lllll} f_1 \leftarrow f_1 & & & & \\ f_2 \leftarrow f_2 & - & \boxed{2/1} & f_1 \\ f_3 \leftarrow f_3 & - & \boxed{1/1} & f_1 \\ f_4 \leftarrow f_4 & - & \boxed{1/1} & f_1 \end{array}$$

La matriz transformada es la siguiente.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \\ f_4 \end{array} \begin{pmatrix} 3 & 3 & 3 & 3 & 6 \\ 0 & -1 & -2 & -1 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & -1 & 0 & 0 & 0 \end{pmatrix}$$

Escogemos el nuevo pivote entre los elementos que están debajo del pivote, es decir, el elemento a_{22}. Seleccionamos como pivote, el mayor en valor absoluto, en este caso como como no hay un elemento mayor que a_{22} en valor absoluto no intercambiamos filas. Hacemos ceros por debajo del pivote

$$\begin{pmatrix} 3 & 3 & 3 & 3 & 6 \\ 0 & -1 & -2 & -1 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & -1 & 0 & 0 & 0 \end{pmatrix} \quad \begin{array}{lllll} f_1 \leftarrow f_1 & & & & \\ f_2 \leftarrow f_2 & & & & \\ f_3 \leftarrow f_3 & - & \boxed{(0)/(-1)} & f_2 \\ f_4 \leftarrow f_4 & - & \boxed{(-1)/(-1)} & f_2 \end{array}$$

La matriz transformada es la siguiente. Ahora el pivote es a_{33} que es cero. Por lo tanto hay que intercambiarlo con a_{43}.

$$\begin{pmatrix} 3 & 3 & 3 & 3 & 6 \\ 0 & -1 & -2 & -1 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 2 & 1 & 0 \end{pmatrix} \quad \begin{array}{lll} f_3 \leftarrow f_4 \\ f_4 \leftarrow f_3 \end{array}$$

Ya tenemos la matriz triangularizada. No obstante, un ordenador realizaría la operación por filas

$$\begin{pmatrix} 3 & 3 & 3 & 3 & 6 \\ 0 & -1 & -2 & -1 & 0 \\ 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \quad \begin{array}{lllll} f_1 \leftarrow f_1 & & & & \\ f_2 \leftarrow f_2 & & & & \\ f_3 \leftarrow f_3 & & & & \\ f_4 \leftarrow f_4 & - & \boxed{0/2} & f_2 \end{array}$$

Y las matrices transformadas A' y b' son:

$$A' = \begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & -1 & -2 & -1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad b' = \begin{pmatrix} 6 \\ 0 \\ 0 \\ 2 \end{pmatrix}$$

2. Resolvemos el sistema triangular superior, equivalente al sistema inicial, $A'x = b'$ y obtenemos x, que era lo que buscábamos.

$$\begin{array}{rrrrcl} 3x_1 & +3x_2 & +3x_3 & +3x_4 & = & 6 \\ & -x_2 & -2x_3 & -x_4 & = & 0 \\ & & 2x_3 & +x_4 & = & 0 \\ & & & x_4 & = & 2 \end{array} \qquad \begin{array}{rcl} x_4 & = & 2 \\ x_3 & = & (0 - x_4)/(2) = -2/2 = -1 \\ x_2 & = & (2x_3 + x_4)/(-1) = (2(-1) + 2)/(-1) = 0 \\ x_1 & = & (6 - 3x_2 - 3x_3 - 3x_4)/3 = \\ & = & (6 + 3 - 6)/3 = 1 \end{array}$$

Y la solución del sistema $Ax = b$ es $x^T = (1, 0, -1, 2)$.

(b) Cálculo de determinante de A. El método de Gauss con pivote parcial para el cálculo de determinantes consta de dos pasos:

1. Triangularización de la matriz usando pivote parcial.
2. Cálculo del determinante multiplicando los elementos de la diagonal de la matriz triangular y multiplicar el resultado por $(-1)^n$, siendo n el número de intercambios de fila.

Como solo se han realizado operaciones por filas, $f_i \leftarrow f_i + \lambda f_j$, e intercambios de las mismas, el determinante de A y el de la matriz transformada A' es el mismo multiplicado por $(-1)^n$.
El primer paso ya lo hicimos en el apartado anterior por lo que nos queda únicamente multiplicar los elementos de la diagonal de A'

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

y como $n = 2$ el determinante es

$$|A| = (-1)^2 \times (3) \times (-1) \times (2) \times (1) = -6$$

Problema 5.3:

Dada la matriz

$$M = \begin{pmatrix} 2 & 1 & 1 \\ 4 & 4 & 4 \\ 1 & 2 & 4 \end{pmatrix}$$

Calcular su determinante usando el método de Gauss con pivote parcial. Enumerar las tres propiedades de los determinantes en las que se basa este método.

Aplicaremos el método de Gauss para triangularizar la matriz. Ello implica hacer ceros por debajo de la diagonal principal usando:

1. Operaciones por filas ($f_i \leftarrow f_i + \lambda f_j$).
2. Intercambiar filas de forma que el pivote (elemento de la diagonal) sea lo mayor posible. Se usan como posibles pivotes los elementos de la diagonal y los elementos por debajo de la diagonal.

Cuando empezamos, la fila del pivote es la primera y, de momento, el pivote es $a_{11} = 2$. Lo comparamos con los elementos que están debajo de él ($a_{12} = 4$ y $a_{13} = 1$) y nos quedamos con el elemento mayor en valor absoluto ($a_{12} = 4$). Intercambiamos la fila que lo contiene con la fila del pivote

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \begin{pmatrix} \boxed{2} & 1 & 1 \\ 4 & 4 & 4 \\ 1 & 2 & 4 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \leftarrow & f_2 \\ f_2 & \leftarrow & f_1 \end{array}$$

Ahora hacemos ceros por debajo del elemento a_{11} sumando la fila del pivote multiplicada por un factor real. Este factor se construye usando como numerador el elemento de la fila que está debajo del pivote y como denominador el pivote con signo negativo.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \begin{pmatrix} \boxed{4} & 4 & 4 \\ 2 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix} \quad \begin{array}{lcllll} f_1 & \rightarrow & f_1 & & & \\ f_2 & \rightarrow & f_2 & - & (2/4) & f_1 \\ f_3 & \rightarrow & f_3 & - & (1/4) & f_1 \end{array}$$

La matriz transformada es

$$\begin{pmatrix} 4 & 4 & 4 \\ 0 & \boxed{-1} & -1 \\ 0 & 1 & 3 \end{pmatrix}$$

Ahora la fila del pivote es la segunda y, de momento, el pivote es $a_{22} = -1$. Lo comparamos con los elementos que están debajo de él ($a_{23} = 1$) y nos quedamos con el elemento mayor en valor absoluto. En este caso, no hay ningún elemento mayor en valor absoluto y no intercambiamos filas. Hacemos ceros por debajo del pivote sumando, a las filas por debajo de la fila del pivote, la fila de pivote multiplicada por un real. No tocamos la fila del pivote y las que están por encima de ella

$$\begin{pmatrix} 4 & 4 & 4 \\ 0 & \boxed{-1} & -1 \\ 0 & 1 & 3 \end{pmatrix} \quad \begin{array}{lcllll} f_1 & \rightarrow & f_1 & & & \\ f_2 & \rightarrow & f_2 & & & \\ f_3 & \rightarrow & f_3 & - & (1/(-1)) & f_2 \end{array}$$

Llegamos a

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

Teniendo en cuenta que:

1. Al intercambiar una columna o una fila, el determinante queda multiplicado por el signo negativo.
2. Las operaciones entre filas ($f_i \leftarrow f_i + \lambda f_j$) no alteran el valor del determinante.
3. El valor del determinante de una matriz triangular o diagonal es el producto de los elementos de la diagonal.

Cómo hemos realizado un intercambio de filas multiplicamos por (-1) el producto de los elementos de la diagonal y

$$|M| = (-1) \times (4) \times (-1) \times (2) = 8$$

Problema 5.4:

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 3 & 3 & 3 \\ 3 & 2 & 2 \\ 3 & 2 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}$$

Calcular $\mathbf{x}$ utilizando la factorización LU. Indicar, en cada paso, las operaciones por fila realizadas.

El método consta de dos pasos

(a) Factorización de la matriz A.

(b) Resolución de dos sistemas: el primero por sustitución progresiva y el segundo por sustitución regresiva.

Resolvamos el sistema.

(a) En el primer paso hacemos ceros por debajo del elemento a_{11} sumando la primera fila multiplicada por un real.

$$\begin{matrix} f_1 \\ f_2 \\ f_3 \end{matrix} \begin{pmatrix} 3 & 3 & 3 \\ 3 & 2 & 2 \\ 3 & 2 & 1 \end{pmatrix} \quad \begin{matrix} f_1 \to f_1 \\ f_2 \to f_2 - \boxed{3/3}\ f_1 \\ f_3 \to f_3 - \boxed{3/3}\ f_1 \end{matrix}$$

Los multiplicadores, que aparecen dentro de los cuadros, son los elementos con los que construímos la matriz L. Los insertamos en la matriz, en lugar de los ceros creados. La matriz transformada es

$$\begin{pmatrix} 3 & 3 & 3 \\ \boxed{1} & -1 & -1 \\ \boxed{1} & -1 & -2 \end{pmatrix} \quad \begin{matrix} f_1 \to f_1 \\ f_2 \to f_2 \\ f_3 \to f_3 - \boxed{\text{(-1)/(-1)}}\ f_2 \end{matrix}$$

Repetimos el proceso creando ceros por debajo de a'_{22} y llegamos a la matriz que almacena simultáneamente L y U.

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

Y las matrices L y U son:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \boxed{1} & 1 & 0 \\ \boxed{1} & \boxed{1} & 1 \end{pmatrix} \quad U = \begin{pmatrix} 3 & 3 & 3 \\ 0 & -1 & -1 \\ 0 & 0 & -1 \end{pmatrix}$$

(b) Ahora, teniendo en cuenta que queremos resolver $Ax = b$ y que $A = LU$ podemos escribir $LUx = b$ y si llamamos $Ux = y$, entonces $Ly = b$:

1. Resolvemos el sistema triangular inferior $Ly = b$ por sustitución progresiva y obtenemos y.

$$\begin{array}{rrrcl} y_1 & & & = & 0 \\ y_1 & +y_2 & & = & 1 \\ y_1 & +y_2 & +y_3 & = & 2 \end{array} \qquad \begin{array}{rcl} y_1 & = & 0 \\ y_2 & = & 1 - y_1 = 1 - 0 = 1 \\ y_3 & = & 2 - y_1 - y_2 = 2 - 0 - 1 = 1 \end{array}$$

2. Resolvemos el sistema triangular superior $Ux = y$ por sustitución regresiva y obtenemos x, que era lo que buscábamos.

$$\begin{array}{rrrcl} 3x_1 & +3x_2 & +3x_3 & = & 0 \\ & -x_2 & -x_3 & = & 1 \\ & & -x_3 & = & 1 \end{array} \qquad \begin{array}{rcll} x_3 & = & 1/(-1) & = & -1 \\ x_2 & = & (1 + x_3)/(-1) & = & 0 \\ x_1 & = & (-3x_2 - 3x_3)/3 & = & 1 \end{array}$$

Y la solución del sistema $Ax = b$ es $x^T = (1, 0, -1)$

Problema 5.5:

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

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

Calcular $\mathbf{x}$ utilizando la factorización LU. Indicar, en cada paso, las operaciones por fila realizadas.

El método consta de dos pasos

(a) Factorización de la matriz A.

(b) Resolución de dos sistemas: el primero por sustitución progresiva y el segundo por sustitución regresiva.

Resolvamos el sistema.

(a) En el primer paso hacemos ceros por debajo del elemento a_{11} sumando la primera fila multiplicada por un real.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \begin{pmatrix} 1 & 2 & 0 \\ 2 & 5 & -1 \\ 0 & -1 & 3 \end{pmatrix} \qquad \begin{array}{rcl} f_1 & \to & f_1 \\ f_2 & \to & f_2 - \boxed{2/1}\ f_1 \\ f_3 & \to & f_3 - \boxed{0/1}\ f_1 \end{array}$$

Los multiplicadores, que aparecen dentro de los cuadros, son los elementos con los que construímos la matriz L. Los insertamos en la matriz, en lugar de los ceros creados. La matriz transformada es

$$\begin{pmatrix} 1 & 2 & 0 \\ \boxed{2} & -3 & -1 \\ \boxed{0} & -1 & 0 \end{pmatrix} \qquad \begin{array}{rcl} f_1 & \to & f_1 \\ f_2 & \to & f_2 \\ f_3 & \to & f_3 - \boxed{\text{-1/-3}}\ f_2 \end{array}$$

Repetimos el proceso creando ceros por debajo de a'_{22} y llegamos a la matriz que almacena simultáneamente L y U.

$$\begin{pmatrix} 1 & 2 & 0 \\ \boxed{2} & -3 & -1 \\ \boxed{0} & \boxed{1/3} & 1/3 \end{pmatrix}$$

Y las matrices L y U son:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \boxed{2} & 1 & 0 \\ \boxed{0} & \boxed{1/3} & 1 \end{pmatrix} \quad U = \begin{pmatrix} 1 & 2 & 0 \\ 0 & -3 & -1 \\ 0 & 0 & 1/3 \end{pmatrix}$$

(b) Ahora, teniendo en cuenta que queremos resolver $Ax = b$ y que $A = LU$ podemos escribir $LUx = b$ y si llamamos $Ux = y$:

1. Resolvemos el sistema triangular inferior $Ly = b$ y obtenemos y.

$$\begin{array}{rrrcrcl} y_1 & & & = & 1 & y_1 & = & 1 \\ 2y_1 & +y_2 & & = & -1 & y_2 & = & -1 - 2y_1 = -3 \\ & +(1/3)y_2 & +y_3 & = & -1 & y_3 & = & -1 - (1/3)y_2 = 0 \end{array}$$

2. Resolvemos el sistema triangular superior $Ux = y$ y obtenemos x, que era lo que buscábamos.

$$\begin{array}{rrrcrcllcl} x_1 & +2x_2 & & = & 1 & x_3 & = & 0/3 & = & 0 \\ & -3x_2 & -x_3 & = & -3 & x_2 & = & (-3 + x_3) \quad /(-3) & = & 1 \\ & & (1/3)x_3 & = & 0 & x_1 & = & 1 - 2x_2 & = & -1 \end{array}$$

Y la solución del sistema $Ax = b$ es $x^T = (-1, 1, 0)$

Problema 5.6:

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 0 \\ 3 \\ 1 \end{pmatrix}$$

Calcular $\mathbf{x}$ utilizando la factorización LU. Indicar, en cada paso, las operaciones por fila realizadas.

El método consta de dos pasos

(a) Factorización de la matriz A.

(b) Resolución de dos sistemas: el primero por sustitución progresiva y el segundo por sustitución regresiva.

Resolvamos el sistema.

(a) En el primer paso hacemos ceros por debajo del elemento a_{11} sumando la primera fila multiplicada por un real.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad \begin{array}{lcllll} f_1 & \to & f_1 & & & \\ f_2 & \to & f_2 & - & \boxed{1/1} & f_1 \\ f_3 & \to & f_3 & - & \boxed{0/1} & f_1 \end{array}$$

Los multiplicadores, que aparecen en los recuadros, son los elementos con los que construímos la matriz L. Los insertamos en la matriz, en lugar de los ceros creados. La matriz transformada es

$$\begin{pmatrix} 1 & 1 & 0 \\ \boxed{1} & 3 & 1 \\ \boxed{0} & 1 & 1 \end{pmatrix} \quad \begin{array}{lcllll} f_1 & \to & f_1 & & & \\ f_2 & \to & f_2 & & & \\ f_3 & \to & f_3 & - & \boxed{1/3} & f_2 \end{array}$$

Repetimos el proceso creando ceros por debajo de a'_{22} y llegamos a la matriz que almacena simultáneamente L y U.

$$\begin{pmatrix} 1 & 1 & 0 \\ \boxed{1} & 3 & 1 \\ \boxed{0} & \boxed{1/3} & 2/3 \end{pmatrix}$$

Y las matrices L y U son:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \boxed{1} & 1 & 0 \\ \boxed{0} & \boxed{1/3} & 1 \end{pmatrix} \quad U = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 2/3 \end{pmatrix}$$

(b) Ahora, teniendo en cuenta que queremos resolver $Ax = b$ y que $A = LU$ podemos escribir $LUx = b$ y si llamamos $Ux = y$:

1. Resolvemos el sistema triangular inferior $Ly = b$ y obtenemos y.

$$\begin{array}{rrrcl} y_1 & & & = & 0 \\ y_1 & +y_2 & & = & 3 \\ & 1/3y_2 & +y_3 & = & 1 \end{array} \qquad \begin{array}{rcl} y_1 & = & 0 \\ y_2 & = & 3 - y_1 = 3 \\ y_3 & = & 1 - (1/3)\, y_2 = 0 \end{array}$$

2. Resolvemos el sistema triangular superior $Ux = y$ y obtenemos x, que era lo que buscábamos.

$$\begin{array}{rrrcl} x_1 & +x_2 & & = & 0 \\ & 3x_2 & +x_3 & = & 3 \\ & & 2/3x_3 & = & 0 \end{array} \qquad \begin{array}{rcl} x_3 & = & 0 \\ x_2 & = & (3 - x_3) \;/3 = \quad 1 \\ x_1 & = & -x_2 \qquad\quad = \; -1 \end{array}$$

Y la solución del sistema $Ax = b$ es $x^T = (-1, 1, 0)$

Problema 5.7:

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 1 & 2 & 1 & -1 \\ 1 & 3 & 3 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & -1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} -3 \\ -2 \\ 1 \\ -2 \end{pmatrix}$$

Calcular $\mathbf{x}$ utilizando la factorización LU con pivote. Indicar, en cada paso, las operaciones por fila realizadas.

El método consta de dos pasos:

(a) Factorización de la matriz A.

(b) Resolución de dos sistemas: el primero por sustitución progresiva y el segundo por sustitución regresiva.

Resolvamos el sistema.

(a) Factorizamos la matriz usando pivote parcial. Para ello escogemos el pivote entre los elementos que están debajo del pivote, es decir, el elemento a_{11}. Seleccionamos como pivote, el mayor en valor absoluto. En este caso, como todos los elementos son iguales, no intercambiamos filas. Hacemos ceros por debajo del pivote sumando la primera fila multiplicada por un real.

$$\begin{matrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{matrix} \begin{pmatrix} 1 & 2 & 1 & -1 \\ 1 & 3 & 3 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & -1 \end{pmatrix} \qquad \begin{array}{lcl} f_1 & \leftarrow & f_1 \\ f_2 & \leftarrow & f_2 - \boxed{1/1}\, f_1 \\ f_3 & \leftarrow & f_3 - \boxed{1/1}\, f_1 \\ f_4 & \leftarrow & f_4 - \boxed{1/1}\, f_1 \end{array}$$

Los multiplicadores, que aparecen dentro de los cuadros, son los elementos con los que construímos la matriz L. Los insertamos en la matriz, en lugar de los ceros creados. La matriz transformada es la siguiente.

$$\begin{matrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{matrix} \begin{pmatrix} 1 & 2 & 1 & -1 \\ \boxed{1} & 1 & 2 & 1 \\ \boxed{1} & -2 & -2 & 1 \\ \boxed{1} & -1 & -1 & 0 \end{pmatrix}$$

Escogemos el pivote entre los elementos que están debajo del pivote, es decir, el elemento a_{22}. Seleccionamos como pivote, el mayor en valor absoluto, el elemento a_{32}. Lo llevamos a la posición del pivote intercambiando filas. Intercambiamos las mismas filas en la matriz de permutaciones P.

$$\begin{matrix} f_1 \\ f_2 \leftarrow f_3 \\ f_3 \leftarrow f_2 \\ f_4 \end{matrix} \begin{pmatrix} 1 & 1 & 1 & -1 \\ \boxed{1} & -2 & -2 & 1 \\ \boxed{1} & 1 & 2 & 1 \\ \boxed{1} & -1 & -1 & 0 \end{pmatrix}$$

$$P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Hacemos ceros por debajo del pivote

$$\begin{pmatrix} 1 & 1 & 1 & -1 \\ \boxed{1} & -2 & -2 & 1 \\ \boxed{1} & 1 & 2 & 1 \\ \boxed{1} & -1 & -1 & 0 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \leftarrow & f_1 \\ f_2 & \leftarrow & f_2 \\ f_3 & \leftarrow & f_3 - \boxed{(1)/(-2)}\ f_2 \\ f_4 & \leftarrow & f_4 - \boxed{(-1)/(-2)}\ f_2 \end{array}$$

La matriz transformada es la siguiente. Ahora el pivote es a_{33} y comparado con los elementos por debajo del pivote es el mayor en valor absoluto y no intercambiamos fila. Como ya tenemos ceros debajo del pivote ya hemos acabado de triangularizar y el multiplicador correspondiente a este paso es cero.

$$\begin{pmatrix} 1 & 2 & 1 & -1 \\ \boxed{1} & -2 & -2 & 1 \\ \boxed{1} & \boxed{-1/2} & 1 & 3/2 \\ \boxed{1} & \boxed{1/2} & 0 & -1/2 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \leftarrow & f_1 \\ f_2 & \leftarrow & f_2 \\ f_3 & \leftarrow & f_3 \\ f_4 & \leftarrow & f_4 - \boxed{0/1}\ f_2 \end{array}$$

La matriz que contiene tanto L como U queda

$$\begin{pmatrix} 1 & 2 & 1 & -1 \\ \boxed{1} & -2 & -2 & 1 \\ \boxed{1} & \boxed{-1/2} & 1 & 3/2 \\ \boxed{1} & \boxed{1/2} & \boxed{0} & -1/2 \end{pmatrix}$$

Y las matrices L, U y P son:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \boxed{1} & 1 & 0 & 0 \\ \boxed{1} & \boxed{-1/2} & 1 & 0 \\ \boxed{1} & \boxed{1/2} & \boxed{0} & 1 \end{pmatrix} \quad U = \begin{pmatrix} 1 & 2 & 1 & -1 \\ 0 & -2 & -2 & 1 \\ 0 & 0 & 1 & 3/2 \\ 0 & 0 & 0 & -1/2 \end{pmatrix} \quad P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(b) Ahora, teniendo en cuenta que queremos resolver $Ax = b$, $PAx = Pb$ y que $PA = LU$ podemos escribir $LUx = Pb$ y si llamamos $Ux = y$:

1. Resolvemos el sistema triangular inferior $Ly = Pb$ y obtenemos y. $Pb = (-3, 1, -2, -2)^t$ y por lo tanto

$$\begin{array}{llllll} y_1 & & & = & -3 & y_1 = -3 \\ y_1 & +y_2 & & = & 1 & y_2 = 1 - y_1 = 1 + 3 = 4 \\ y_1 & +(-1/2)y_2 & +y_3 & = & -2 & y_3 = -2 - y_1 + (1/2)y_2 = -2 + 3 + 2 = 3 \\ y_1 & +(1/2)y_2 & +y_4 & = & -2 & y_4 = -2 - y_1 - (1/2)y_2 = -2 + 3 - 2 = -1 \end{array}$$

2. Resolvemos el sistema triangular superior $Ux = y$ y obtenemos x, que era lo que buscábamos.

$$\begin{array}{rrrrcr} x_1 & +2x_2 & +x_3 & -x_4 & = & -3 \\ & -2x_2 & -2x_3 & +x_4 & = & 4 \\ & & x_3 & +(3/2)x_4 & = & 3 \\ & & & -(1/2)x_4 & = & -1 \end{array} \qquad \begin{array}{rcl} x_4 &=& (-1)\times(-2) = 2 \\ x_3 &=& 3-(3/2)x_4 = 3-3 = 0 \\ x_2 &=& (4+2x_3-x_4)/(-2) = \\ &=& (4-2)/(-2) = -1 \\ x_1 &=& -3-2x_2-x_3+x_4 = \\ &=& -3+2+2 = 1 \end{array}$$

Y la solución del sistema $Ax = b$ es $x^T = (1, -1, 0, 2)$

Problema 5.8:

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 1 & 1 & 0 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 0 & -2 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} -2 \\ -1 \\ -3 \\ -4 \end{pmatrix}$$

(a) Calcular $\mathbf{x}$ utilizando la factorización LU con pivote. Indicar, en cada paso, las operaciones por fila realizadas.

(b) A partir de los determinantes de las matrices L, U y P calcular el determinante de A. Enumerar las propiedades de los determinantes utilizadas.

(a) El método consta de dos pasos:

(i) Factorización de la matriz A.

(ii) Resolución de dos sistemas: el primero por sustitución progresiva y el segundo por sustitución regresiva.

Resolvamos el sistema. Factorizamos la matriz usando pivote parcial. Para ello, escogemos el pivote entre los elementos que están debajo del pivote, es decir, el elemento a_{11}. Seleccionamos como pivote, el mayor en valor absoluto. Lo llevamos a la posición del pivote intercambiando filas. Intercambiamos las mismas filas en la matriz de permutaciones P.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \\ f_4 \end{array} \begin{pmatrix} \mathbf{1} & 1 & 0 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 2 & 3 & 0 & -2 \end{pmatrix} \qquad \begin{array}{l} f_1 \leftarrow f_4 \\ f_2 \\ f_3 \\ f_4 \leftarrow f_1 \end{array} \begin{pmatrix} 2 & 3 & 0 & -2 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 \end{pmatrix}$$

$$P_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad P_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Hacemos ceros por debajo del pivote sumando la primera fila multiplicada por un real.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \\ f_4 \end{array} \begin{pmatrix} 2 & 3 & 0 & -2 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \leftarrow & f_1 \\ f_2 & \leftarrow & f_2 - \boxed{1/2}\ f_1 \\ f_3 & \leftarrow & f_3 - \boxed{1/2}\ f_1 \\ f_4 & \leftarrow & f_4 - \boxed{1/2}\ f_1 \end{array}$$

Los multiplicadores, que aparecen dentro de los cuadros, son los elementos con los que construímos la matriz L. Los insertamos en la matriz, en lugar de los ceros creados. La matriz transformada es la siguiente. Como ahora el pivote, a_{22} y los elementos por debajo del pivote son iguales, no intercambiamos filas. Hacemos ceros por debajo del pivote

$$\begin{pmatrix} 2 & 3 & 0 & -2 \\ \boxed{1/2} & \mathbf{-1/2} & 0 & 1 \\ \boxed{1/2} & -1/2 & 1 & 0 \\ \boxed{1/2} & -1/2 & 0 & 0 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \leftarrow & f_1 \\ f_2 & \leftarrow & f_2 \\ f_3 & \leftarrow & f_3 - \boxed{\text{(-1/2)/(-1/2)}}\ f_2 \\ f_4 & \leftarrow & f_4 - \boxed{\text{(-1/2)/(-1/2)}}\ f_2 \end{array}$$

La matriz transformada es la siguiente. Ahora el pivote es a_{33} y comparado con los elementos por debajo del pivote es el mayor en valor absoluto y no intercambiamos fila.

$$\begin{pmatrix} 2 & 3 & 0 & -2 \\ \boxed{1/2} & -1/2 & 0 & 1 \\ \boxed{1/2} & \boxed{1} & \mathbf{1} & 0 \\ \boxed{1/2} & \boxed{1} & 0 & -1 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \leftarrow & f_1 \\ f_2 & \leftarrow & f_2 \\ f_3 & \leftarrow & f_3 \\ f_4 & \leftarrow & f_4 - \boxed{0/1}\ f_2 \end{array}$$

La matriz que contiene tanto L como U queda

$$\begin{pmatrix} 2 & 3 & 0 & -2 \\ \boxed{1/2} & -1/2 & 0 & 1 \\ \boxed{1/2} & \boxed{1} & 1 & 0 \\ \boxed{1/2} & \boxed{1} & \boxed{0} & -1 \end{pmatrix}$$

Y las matrices L, U y P son:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \boxed{1/2} & 1 & 0 & 0 \\ \boxed{1/2} & \boxed{1} & 1 & 0 \\ \boxed{1/2} & \boxed{1} & \boxed{0} & 1 \end{pmatrix} \quad U = \begin{pmatrix} 2 & 3 & 0 & -2 \\ 0 & -1/2 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Ahora, teniendo en cuenta que queremos resolver $Ax = b$, $PAx = Pb$ y que $PA = LU$ podemos escribir $LUx = Pb$ y si llamamos $Ux = y$:

1. Resolvemos el sistema triangular inferior $Ly = Pb$ y obtenemos y. $Pb = (-4, -1, -3, -2)^t$ y por lo tanto

$$\begin{array}{rrrrcl|rcl} y_1 & & & & = & -4 & y_1 & = & -4 \\ (1/2)y_1 & +y_2 & & & = & -1 & y_2 & = & -1 - (1/2)y_1 = -1 + 2 = 1 \\ (1/2)y_1 & +y_2 & +y_3 & & = & -3 & y_3 & = & -3 - (1/2)y_1 - y_2 = -3 + 2 - 1 = -2 \\ (1/2)y_1 & +y_2 & & +y_4 & = & -2 & y_4 & = & -2 - (1/2)y_1 - y_2 = -2 + 2 - 1 = -1 \end{array}$$

2. Resolvemos el sistema triangular superior $Ux = y$ y obtenemos x, que era lo que buscábamos.

$$\begin{array}{rrrrrr} 2x_1 & +3x_2 & -2x_4 & = & -4 \\ & (-1/2)x_2 & +x_4 & = & 1 \\ & x_3 & & = & -2 \\ & & -x_4 & = & -1 \end{array} \qquad \begin{array}{rcl} x_4 & = & 1 \\ x_3 & = & -2 \\ x_2 & = & (1-x_4)/(-1/2) = (1-1)/(-1/2) = 0 \\ x_1 & = & (-4-3x_2+2x_4)/2 = \\ & = & (-4+0+2)/2 = -1 \end{array}$$

Y la solución del sistema $Ax = b$ es $x^T = (-1, 0, -2, 1)$

(b) Si queremos calcular el determinante de A tenemos que

$$PA = LU \quad \Longrightarrow |A| = (-1)^n |L| \, |U|$$

siendo n el número de permutaciones de filas. Por lo tanto

$$|A| = (-1)^1(1 \times 1 \times 1 \times 1)(2 \times (-1/2) \times 1 \times (-1)) = -1$$

Las propiedades usadas son:

1. El determinante de un producto de matrices es el producto de los determinantes.
2. Si se intercambian dos filas el determinante cambia de signo.
3. El determinante de una matriz triangular es igual al producto de los elementos de su diagonal.

5.2. Métodos iterativos

Problema 5.9:

Dado el sistema $Ax = b$, donde

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

(a) Calcular la matriz de iteración de Jacobi B_J.

(b) Calcular la norma infinito de B_J ¿Qué podemos concluir acerca de la convergencia del método para resolver el sistema?

(c) Calcular la norma uno de B_J ¿Qué podemos concluir acerca de la convergencia del método para resolver el sistema?

(d) Determinar si el método de Jacobi para resolver este sistema converge o no.

(e) Realizar dos iteraciones tomando como punto inicial $(0, 0, 0)^T$.

(a) Se tiene que $B_J = -D^{-1}(L + U)$. O también:

1. Dividimos cada fila por el correspondiente elemento de la diagonal.

$$\begin{pmatrix} 1 & 1 & 0 \\ \frac{1}{3} & 1 & \frac{1}{3} \\ 0 & 1 & 1 \end{pmatrix}$$

2. Cambiamos todos los elementos de signo.

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

3. Ponemos ceros en la diagonal principal.

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

(b) Calculamos la norma infinito

$$B_J = \begin{pmatrix} 0 & -1 & 0 \\ -\frac{1}{3} & 0 & -\frac{1}{3} \\ 0 & -1 & 0 \end{pmatrix} \begin{matrix} 0+1+0=1 \\ \frac{1}{3}+0+\frac{1}{3}=\frac{2}{3} \\ 0+1+0=1 \end{matrix}$$

$$||B_J||_\infty = \text{Max}\left(1, \frac{2}{3}, 1\right) = 1$$

Como $||B_J||_\infty < 1$ es condición suficiente, no necesaria, de convergencia, no podemos asegurar ni que converge ni que no converge.

(c) Calculamos la norma uno

$$B_J^T = \begin{pmatrix} 0 & -\frac{1}{3} & 0 \\ -1 & 0 & -1 \\ 0 & -\frac{1}{3} & 0 \end{pmatrix} \begin{matrix} 0+\frac{1}{3}+0=\frac{1}{3} \\ 1+0+1=2 \\ 0+\frac{1}{3}+0=\frac{1}{3} \end{matrix}$$

$$||B_J||_1 = \text{Max}\left(\frac{1}{3}, 2, \frac{1}{3}\right) = 2$$

Como $||B_J||_1 < 1$ es condición suficiente, no necesaria, de convergencia, no podemos asegurar ni que converge ni que no converge.

(d) Para determinar si el método converge o no, calculamos los autovalores de B_J haciendo

$$|B_J - \lambda I| = 0$$

Si

$$\begin{vmatrix} 0-\lambda & -1 & 0 \\ -\frac{1}{3} & 0-\lambda & -\frac{1}{3} \\ 0 & -1 & 0-\lambda \end{vmatrix} = 0$$

calculando el determinante obtenemos el polinomio característico, que igualamos a cero

$$-\lambda^3 - \left(-\frac{1}{3}\lambda - \frac{1}{3}\lambda\right) = -\lambda^3 + \frac{2}{3}\lambda =$$

$$= -\lambda\left(\lambda^2 - \frac{2}{3}\right) = -\lambda\left(\lambda + \sqrt{\frac{2}{3}}\right)\left(\lambda - \sqrt{\frac{2}{3}}\right) = 0$$

Y los autovalores son

$$\lambda_1 = 0, \quad \lambda_2 = \sqrt{\frac{2}{3}}, \quad \lambda_3 = -\sqrt{\frac{2}{3}}$$

La condición necesaria y suficiente para que el método de Jacobi converja es que todos los autovalores, en valor absoluto, sean menores que 1. Como esta condición se cumple, el método converge.

(e) El sistema es

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

En la primera ecuación despejamos la primera incógnita, x, en la segunda, la segunda incógnita, y, y finalmente, z.

$$\begin{array}{rcl} x & = & 1-y \\ y & = & (2-x-z)/3 \\ z & = & -1-y \end{array}$$

Realizamos una iteración

$$\begin{array}{rcl} x^{(1)} & = & 1-y^{(0)} = 1-0 = 1 \\ y^{(1)} & = & \left(2-x^{(0)}-z^{(0)}\right)/3 = (2-0-0)/3 \ = 2/3 = 0{,}67 \\ z^{(1)} & = & -1-y^{(0)} = -1-0 = -1 \end{array}$$

Realizamos otra iteración

$$\begin{array}{rcl} x^{(2)} & = & 1 - y^{(1)} = 1 - 2/3 = 1/3 = 0{,}33 \\ y^{(2)} & = & \left(2 - x^{(1)} - z^{(1)}\right)/3 = (2 - 1 - (-1))/3 = 2/3 = 0{,}67 \\ z^{(2)} & = & -1 - y^{(1)} = -1 - 2/3 = -5/3 = -1{,}67 \end{array}$$

Problema 5.10:

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 3 \\ 4 \\ 2 \end{pmatrix}$$

(a) ¿Es A diagonal dominante por filas? Con este resultado ¿qué podemos concluir acerca de la convergencia del método de Jacobi?

(b) Calcular la matriz de iteración de Jacobi B_J.

(c) Calcular la norma infinito de la matriz de iteración de Jacobi B_J. Con este resultado ¿qué podemos concluir acerca de la convergencia del método de Jacobi?

(d) Calcular la norma uno de B_J. Con este resultado ¿qué podemos concluir acerca de la convergencia del método de Jacobi?

(e) Calcular los autovalores de B_J. Con este resultado ¿qué podemos concluir acerca de la convergencia del método de Jacobi?

(f) Realizar 2 iteraciones con Jacobi. Utilizar $\mathbf{x}^{(0)} = \mathbf{0}$

(a) Estudiemos si la suma de los elementos de una fila, en valor absoluto, excluído el elemento de la diagonal principal, es menor que el elemento de la diagonal principal en valor absoluto

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{array}{l} |1| \not> |1| + |0| \\ |4| > |1| + |1| \\ |1| \not> |0| + |1| \end{array}$$

No, A no es diagonal dominante por filas, puesto que para la primera fila no se verifica $|1| > |1| + |0|$ y para la tercera fila $|1| \not> |0| + |1|$ tampoco y como A no es diagonal dominante, no podemos concluir nada respecto a la convergencia del método de Jacobi.

(b) Se tiene que $B_J = -D^{-1}(L + U)$. O también:

1. Dividimos cada fila por el correspondiente elemento de la diagonal.

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

2. Cambiamos todos los elementos de signo.

$$\begin{pmatrix} -1 & -1 & 0 \\ -1/4 & -1 & -1/4 \\ 0 & -1 & -1 \end{pmatrix}$$

3. Ponemos ceros en la diagonal principal.

$$B_J = \begin{pmatrix} 0 & -1 & 0 \\ -1/4 & 0 & -1/4 \\ 0 & -1 & 0 \end{pmatrix}$$

(c) Calculamos la norma infinito

$$B_J = \begin{pmatrix} 0 & -1 & 0 \\ -1/4 & 0 & -1/4 \\ 0 & -1 & 0 \end{pmatrix} \begin{matrix} 0+1+0=1 \\ 1/4+0+1/4=1/2 \\ 0+1+0=1 \end{matrix}$$

$$||B_J||_\infty = \text{Max}\,(1, 1/4, 1) = 1$$

Y como $||B_J||_\infty \not< 1$ no podemos concluir nada respecto a la convergencia del Método de Jacobi.

(d) Calculamos la norma uno

$$B_J^T = \begin{pmatrix} 0 & -1/4 & 0 \\ -1 & 0 & -1 \\ 0 & -1/4 & 0 \end{pmatrix} \begin{matrix} 0+1/4+0=1/4 \\ 1+0+1=2 \\ 0+1/4+0=1/4 \end{matrix}$$

$$||B_J||_1 = \text{Max}\,(1/4, 2, 1/4) = 2$$

Y como $||B_J||_1 \not< 1$ no podemos concluir nada respecto a la convergencia del Método de Jacobi.

(e) Estudiemos los autovalores de la matriz B_J. Haciendo

$$|B_J - \lambda I| = 0$$

tenemos

$$\begin{vmatrix} 0-\lambda & -1/4 & 0 \\ -1 & 0-\lambda & -1 \\ 0 & -1/4 & 0-\lambda \end{vmatrix} = -\lambda^3 - \left(-\frac{1}{4}\lambda - \frac{1}{4}\lambda\right) = \frac{1}{2}\lambda - \lambda^3 = \lambda\left(\frac{1}{2} - \lambda^2\right) = 0$$

Y los autovalores son

$$\lambda_1 = 0, \quad \lambda_{2,3} = \pm\sqrt{\frac{1}{2}} = \pm 0{,}71$$

Como todos los autovalores de B_J son menores que uno en valor absoluto podemos concluir que el Método de Jacobi será convergente para cualquier valor inicial.

(f) El sistema es

$$\begin{array}{rrrcl} x & +y & & = & 3 \\ x & +4y & +z & = & 4 \\ & y & +z & = & 2 \end{array}$$

En la primera ecuación despejamos la primera incógnita, x, en la segunda, la segunda incógnita, y, y finalmente, z.

$$\begin{array}{rcl} x & = & 3-y \\ y & = & (4-x-z)/4 \\ z & = & 2-y \end{array}$$

Realizamos la primera iteración

$$\begin{array}{rcl} x^{(1)} & = & 3-y^{(0)} = 3-0 = 3 \\ y^{(1)} & = & \left(4-x^{(0)}-z^{(0)}\right)/4 = (4-0-0)/4 = 1 \\ z^{(1)} & = & 2-y^{(0)} = 2-0 = 2 \end{array}$$

y la segunda

$$\begin{array}{rcl} x^{(2)} & = & 3-y^{(1)} = 3-1 = 2 \\ y^{(2)} & = & \left(4-x^{(1)}-z^{(1)}\right)/4 = (4-3-2)/4 = -1/4 = -0{,}25 \\ z^{(2)} & = & 2-y^{(1)} = 2-1 = 1 \end{array}$$

Problema 5.11:

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 3 & 1 & 1 \\ 4 & 6 & 2 \\ 1 & 1 & 4 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix}$$

(a) Si resolvemos el sistema por Jacobi ¿converge?¿Por qué?

(b) Realizar 2 iteraciones con Jacobi. Utilizar $\mathbf{x}^{(0)} = (0,0,0)^T$

(a) Estudiemos si la matriz de coeficientes de este sistema es diagonal dominante por filas

$$A = \begin{pmatrix} 3 & 1 & 1 \\ 4 & 6 & 2 \\ 1 & 1 & 4 \end{pmatrix} \begin{array}{l} |3| > |1|+|1| \\ |6| \not> |4|+|2| \\ |4| > |1|+|1| \end{array}$$

No lo es, así que no podemos concluir nada.

(b) Estudiemos alguna de las normas de la matriz de iteración B_J. Se tiene que $B_J = -D^{-1}(L+U)$. O también:

1. Dividimos cada fila por el correspondiente elemento de la diagonal.

$$\begin{pmatrix} 1 & 1/3 & 1/3 \\ 4/6 & 1 & 2/6 \\ 1/4 & 1/4 & 1 \end{pmatrix}$$

2. Cambiamos todos los elementos de signo.

$$\begin{pmatrix} -1 & -1/3 & -1/3 \\ -4/6 & -1 & -2/6 \\ -1/4 & -1/4 & -1 \end{pmatrix}$$

3. Ponemos ceros en la diagonal principal.

$$B_J = \begin{pmatrix} 0 & -1/3 & -1/3 \\ -4/6 & 0 & -2/6 \\ -1/4 & -1/4 & 0 \end{pmatrix}$$

Calculemos la norma infinito

$$\|B_J\|_\infty = \max_{1\le i\le 3} \sum_{j=1}^{3} |a_{ij}|$$

$$B_J = \begin{pmatrix} 0 & -1/3 & -1/3 \\ -4/6 & 0 & -2/6 \\ -1/4 & -1/4 & 0 \end{pmatrix} \begin{matrix} 0+1/3+1/3 = 1/6 \\ 4/6+2/6 = 1 \\ 1/4+1/4+0 = 1/2 \end{matrix}$$

$$\|B_J\|_\infty = \max(1/6, 1, 1/2) = 1$$

Como esta norma no es estrictamente menor que uno, tampoco decide. Calculemos la norma uno

$$\|B_J\|_1 = \max_{1\le j\le 3} \sum_{i=1}^{3} |a_{ij}|$$

$$B_J = \begin{pmatrix} 0 & -2/3 & -1/4 \\ -1/3 & 0 & -1/4 \\ -1/3 & -1/3 & 0 \end{pmatrix} \begin{matrix} 0+2/3+1/4 = 11/12 \\ 1/3+0+1/4 = 7/12 \\ 1/3+1/3+0 = 2/3 \end{matrix}$$

$$\|B_J\|_1 = \max\left(\frac{11}{12}, \frac{7}{12}, \frac{8}{12}\right) = \frac{11}{12} < 1$$

y como una norma de la matriz de iteración es menor que uno, el método de Jacobi converge.

(c) El sistema es

$$\begin{array}{rrrcr} 3x & +y & +z & = & 1 \\ 4x & +6y & +2z & = & -1 \\ x & +y & +4z & = & 0 \end{array}$$

En la primera ecuación despejamos la primera incógnita, x, en la segunda, la segunda incógnita, y, y finalmente, z.

$$\begin{aligned} x &= (1-y-z)/3 \\ y &= (-1-4x-2z)/6 \\ z &= (-x-y)/4 \end{aligned}$$

Realizamos la primera iteración

$$\begin{aligned} x^{(1)} &= (1-y^{(0)}-z^{(0)})/3 = 1/3 = 0{,}333 \\ y^{(1)} &= \left(-1-4x^{(0)}-2z^{(0)}\right)/6 = -1/6 = -0{,}167 \\ z^{(1)} &= (-x^{(0)}-y^{(0)})/4 = 0 \end{aligned}$$

y la segunda

$$\begin{array}{rcl} x^{(2)} & = & (1 - y^{(1)} - z^{(1)})/3 = (1 - (-0{,}167) - 0)/3 = 0{,}389 \\ y^{(2)} & = & \left(-1 - 4x^{(1)} - 2z^{(1)}\right)/6 = (-1 - 4(0{,}333) - 2(0))/6 = -0{,}389 \\ z^{(2)} & = & (-x^{(1)} - y^{(1)})/4 = (-(0{,}333) - (-0{,}167))/4 = -0{,}0417 \end{array}$$

Problema 5.12:

Dado el sistema $Ax = b$, donde

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

(a) Calcular la matriz de iteración de Gauss-Seidel, B_{G-S}. Utilizar Gauss-Jordan para calcular matrices inversas.

(b) Estudiar la convergencia del método de Gauss-Seidel para resolver este sistema.

(c) Realizar dos iteraciones por Gauss-Seidel. Utilizar $\mathbf{x}^{(0)} = (0, 0, 0)^T$

(a) La matriz $B_{G-S} = -(L + D)^{-1}U$ con

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

Por lo que

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

Calculemos por Gauss-Jordan $(L + D)^{-1}$. Construímos la matriz

$$((L + D)|I)$$

y buscamos, mediante operaciones por filas, llegar a

$$(I|(L + D)^{-1})$$

Empezamos haciendo uno el primer pivote, a_{11}. Para ello dividimos la primera fila por este pivote

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 2 & 0 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 & 1 & 0 \\ 1 & 1 & 2 & 0 & 0 & 1 \end{array}\right) \qquad f_1 \rightarrow f_1/2$$

Ahora hacemos ceros en la columna del pivote a_{11} sumando la fila del pivote multiplicada por un factor a las demás filas

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/2 & 0 & 0 \\ 1 & 2 & 0 & 0 & 1 & 0 \\ 1 & 1 & 2 & 0 & 0 & 1 \end{array}\right) \qquad \begin{array}{lcll} f_1 & \rightarrow & f_1 & \\ f_2 & \rightarrow & f_2 & -(1)f_1 \\ f_3 & \rightarrow & f_3 & -(1)f_1 \end{array}$$

Repetimos este proceso para las demás filas. Ahora, hacemos que el segundo pivote, a_{22} sea uno

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 2 & 0 & -1/2 & 1 & 0 \\ 0 & 1 & 2 & -1/2 & 0 & 1 \end{array}\right) \quad f_2 \rightarrow f_2/2$$

Hacemos ceros en la columna del pivote por encima y por debajo del pivote

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & -1/4 & 1/2 & 0 \\ 0 & 1 & 2 & -1/2 & 0 & 1 \end{array}\right) \quad \begin{array}{lll} f_1 \rightarrow f_1 & -(0)f_2 \\ f_2 \rightarrow f_2 & \\ f_3 \rightarrow f_3 & -(1)f_2 \end{array}$$

Hacemos que el tercer pivote, a_{33} sea uno

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & -1/4 & 1/2 & 0 \\ 0 & 0 & 2 & -1/4 & -1/2 & 1 \end{array}\right) \quad f_3 \rightarrow f_3/2$$

Hacemos ceros en la columna del pivote

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & -1/4 & 1/2 & 0 \\ 0 & 0 & 1 & -1/8 & -1/4 & 1/2 \end{array}\right)$$

Como ya hemos llegado a $(I|(L+D)^{-1})$, se tiene

$$(L+D)^{-1} = \begin{pmatrix} 1/2 & 0 & 0 \\ -1/4 & 1/2 & 0 \\ -1/8 & -1/4 & 1/2 \end{pmatrix}$$

y

$$B_{G-S} = -(L+D)^{-1}U = -\begin{pmatrix} 1/2 & 0 & 0 \\ -1/4 & 1/2 & 0 \\ -1/8 & -1/4 & 1/2 \end{pmatrix}\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1/2 & 0 \\ 0 & 1/4 & -1/2 \\ 0 & 1/8 & 1/4 \end{pmatrix}$$

(b) *Si alguna norma matricial de B_{G-S} es estrictamente menor que uno, el método de Gauss-Seidel converge.*

Probemos con la norma infinito

$$\|B_J\|_\infty = \max_{1\leq i\leq 3} \sum_{j=1}^{3} |a_{ij}|$$

Se tiene que

$$B_{G-S} = \begin{pmatrix} 0 & -1/2 & 0 \\ 0 & 1/4 & -1/2 \\ 0 & 1/8 & 1/4 \end{pmatrix} \begin{array}{l} 0+(1/2)+0 = 1/2 \\ 0+(1/4)+(1/2) = 3/4 \\ 0+(1/8)+(1/4) = 3/8 \end{array}$$

$$||B_{G-S}||_\infty = \text{Max}\left(\frac{1}{2}, \frac{3}{4}, \frac{3}{8}\right) = \text{Max}\left(\frac{4}{8}, \frac{6}{8}, \frac{3}{8}\right) = \frac{6}{8} = 0{,}75$$

Como $||B_{G-S}||_\infty < 1$ es condición suficiente de convergencia, podemos asegurar que converge.

(c) El sistema es

$$\begin{array}{rrrcr} 2x & +y & & = & -2 \\ x & +2y & +z & = & 0 \\ x & +y & +2z & = & 1 \end{array}$$

En la primera ecuación despejamos la primera incógnita, x, en la segunda, la segunda incógnita, y, y finalmente, z.

$$\begin{array}{rcl} x & = & (-2-y)/2 \\ y & = & (-x-z)/2 \\ z & = & (1-x-y)/2 \end{array}$$

Realizamos una iteración

$$\begin{array}{rcl} x^{(1)} & = & \left(-2-y^{(0)}\right)/2 = (-2)/2 = -1 \\ y^{(1)} & = & \left(-x^{(1)}-z^{(0)}\right)/2 = (-(-1))/2 = 0{,}5 \\ z^{(1)} & = & \left(1-x^{(1)}-y^{(1)}\right)/2 = (1-(-1)-(0{,}5))/2 = 0{,}75 \end{array}$$

Realizamos otra iteración

$$\begin{array}{rcl} x^{(2)} & = & \left(-2-y^{(1)}\right)/2 = (-2-0{,}5)/2 = -1{,}25 \\ y^{(2)} & = & \left(-x^{(2)}-z^{(1)}\right)/2 = (-(-1{,}25)-0{,}75)/2 = 0{,}25 \\ z^{(2)} & = & \left(1-x^{(2)}-y^{(2)}\right)/2 = (1-(-1{,}25)-0{,}25)/2 = 1 \end{array}$$

Problema 5.13:

Dado el sistema $Ax = b$, donde

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

(a) Calcular la matriz de iteración de Gauss-Seidel, B_{G-S}. Utilizar Gauss-Jordan para calcular matrices inversas.

(b) Calcular la norma uno e infinito de B_{G-S} ¿Qué podemos concluir acerca de la convergencia del método para resolver el sistema?

(c) Determinar si el método de Gauss-Seidel para resolver este sistema converge o no.

(d) Realizar dos iteraciones.

(a) La matriz $B_{G-S} = -(L+D)^{-1}U$ con

$$A = L + D + U = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Por lo que

$$L + D = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 1 & 1 \end{pmatrix} \quad y \quad U = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Calculemos por Gauss-Jordan $(L+D)^{-1}$

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \quad \begin{array}{lcll} f_1 & \to & f_1 & \\ f_2 & \to & f_2 & -(1/1)f_1 \\ f_3 & \to & f_3 & -(0/1)f_1 \end{array}$$

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 3 & 0 & -1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \quad f_2 \to f_2/3$$

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -\frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right) \quad \begin{array}{lcll} f_1 & \to & f_1 & -(0/1)f_2 \\ f_2 & \to & f_2 & \\ f_3 & \to & f_3 & -(1/1)f_2 \end{array}$$

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -\frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{1}{3} & 1 \end{array}\right)$$

Por lo tanto

$$(L+D)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & -\frac{1}{3} & 1 \end{pmatrix}$$

y

$$B_{G-S} = -(L+D)^{-1}U = -\begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & -\frac{1}{3} & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & \frac{1}{3} & -\frac{1}{3} \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

(b) Calculamos la norma infinito

$$B_{G-S} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & \frac{1}{3} & -\frac{1}{3} \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{array}{l} 0+1+0=1 \\ 0+\frac{1}{3}+\frac{1}{3}=\frac{2}{3} \\ 0+\frac{1}{3}+\frac{1}{3}=\frac{2}{3} \end{array}$$

$$||B_{G-S}||_\infty = \text{Max}\left(1, \frac{2}{3}, \frac{2}{3}\right) = 1$$

Como $||B_{G-S}||_\infty < 1$ es condición suficiente, no necesaria, de convergencia, no podemos asegurar ni que converge ni que no converge.

Calculamos la norma uno

$$B_{G-S}^T = \begin{pmatrix} 0 & 0 & 0 \\ -1 & \frac{1}{3} & -\frac{1}{3} \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{array}{l} 0+0+0=0 \\ 1+\frac{1}{3}+\frac{1}{3}=\frac{5}{3} \\ 0+\frac{1}{3}+\frac{1}{3}=\frac{2}{3} \end{array}$$

$$||B_{G-S}||_1 = \text{Max}\left(0, \frac{5}{3}, \frac{2}{3}\right) = \frac{5}{3}$$

Como $||B_{G-S}||_1 < 1$ es condición suficiente, no necesaria, de convergencia, no podemos asegurar ni que converge ni que no converge.

(c) Para determinar si el método converge o no, calculamos los autovalores de B_{G-S}

$$\begin{vmatrix} 0-\lambda & -1 & 0 \\ 0 & \frac{1}{3}-\lambda & -\frac{1}{3} \\ 0 & -\frac{1}{3} & \frac{1}{3}-\lambda \end{vmatrix} = 0$$

Calculando el determinante obtenemos el polinomio característico, que igualamos a cero

$$-\lambda\left(\frac{1}{3}-\lambda\right)\left(\frac{1}{3}-\lambda\right)-\left(-\frac{1}{9}\lambda\right) = -\lambda\left(\frac{1}{9}-\frac{2}{3}\lambda+\lambda^2-\frac{1}{9}\right) =$$

$$= -\lambda\left(-\frac{2}{9}\lambda+\lambda^2\right) = -\lambda^2\left(-\frac{2}{3}+\lambda\right) = 0$$

Y los autovalores son

$$\lambda_{1,2} = 0, \quad \lambda_3 = \frac{2}{3}$$

La condición necesaria y suficiente para que el método de Gauss-Seidel converja es que todos los autovalores, en valor absoluto, sean menores que 1. Como esta condición se cumple, el método converge.

(d) El sistema es

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

En la primera ecuación despejamos la primera incógnita, x, en la segunda, la segunda incógnita, y, y finalmente, z.

$$\begin{array}{rcl} x & = & 1-y \\ y & = & (2-x-z)/3 \\ z & = & -1-y \end{array}$$

Realizamos una iteración

$$\begin{array}{rcl} x^{(1)} & = & 1-y^{(0)} = 1-0 = 1 \\ y^{(1)} & = & \left(2-x^{(1)}-z^{(0)}\right)/3 = (2-1-0)/3 = 1/3 = 0{,}33 \\ z^{(1)} & = & -1-y^{(1)} = -1-1/3 = -4/3 = -1{,}33 \end{array}$$

Realizamos otra iteración

$$\begin{array}{rcl} x^{(2)} & = & 1-y^{(1)} = 1-1/3 = 2/3 = 0{,}67 \\ y^{(2)} & = & \left(2-x^{(2)}-z^{(1)}\right)/3 = (2-2/3-(-4/3))/3 \ = 8/9 = 0{,}89 \\ z^{(2)} & = & -1-y^{(2)} = -1-8/9 = -17/9 = -1{,}89 \end{array}$$

Problema 5.14:

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 1 & 2 & 5 \\ 2 & 1 & 0 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 4 \\ 21 \\ 2 \end{pmatrix}$$

(a) Resolver el sistema por factorización LU con pivote parcial. Indicar, en cada paso, las operaciones por fila realizadas.

(b) ¿Se puede resolver este sistema por Jacobi?¿Por qué?¿Y por Gauss-Seidel?¿Por qué?

(c) Reordenar las ecuaciones (filas) del sistema para que la matriz A' del sistema equivalente $A'x = b'$ sea diagonal dominante por filas. (Una vez construída A' justificar que es diagonal dominante)

(d) Si resolvemos el sistema $A'x = b'$ ¿Converge por Jacobi?¿Por qué? ¿Y por Gauss-Seidel? ¿Por qué?

(a) El método LU consta de dos pasos

1. Factorización de la matriz A usando pivote parcial.
2. Resolución de dos sistemas: el primero por sustitución progresiva y el segundo por sustitución regresiva.

Por lo tanto, para resolver el sistema:

1. La matriz de permutaciones P será inicialmente

$$P_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Aplicamos la estrategia del pivote: en el primer paso buscamos el elemento de mayor valor absoluto por debajo del pivote a_{11}. Este resulta ser a_{31}, por lo que intercambiamos las filas 3 y 1 tanto en P como en A.

$$A_1 = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 5 \\ 0 & 2 & 1 \end{pmatrix} \quad P_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Ahora, hacemos ceros por debajo del elemento a_{11} restando la primera fila multiplicada por el real construído con el pivote (a_{11}) en el denominador y el elemento de esa fila por debajo del pivote (a_{21} y a_{23} respectivamente) en el numerador.

$$\begin{matrix} f_1 \\ f_2 \\ f_3 \end{matrix} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 5 \\ 0 & 2 & 1 \end{pmatrix} \quad \begin{matrix} f_1 \to f_1 \\ f_2 \to f_2 - \boxed{1/2}\ f_1 \\ f_3 \to f_3 - \boxed{0/2}\ f_1 \end{matrix}$$

Los multiplicadores, que aparecen dentro de los cuadros, son los elementos con los que construímos la matriz L. Los insertamos en la matriz, en lugar de los ceros creados. La matriz transformada es

$$\begin{pmatrix} 2 & 1 & 0 \\ \boxed{1/2} & 3/2 & 5 \\ \boxed{0} & 2 & 1 \end{pmatrix}$$

Volvemos a aplicar la estrategia del pivote: buscamos el elemento de mayor valor absoluto por debajo del pivote a_{22}. Este resulta ser a_{32} por lo que intercambiamos las filas 2 y 3 en P, A y L.

$$A_2|L_2 = \begin{pmatrix} 2 & 1 & 0 \\ \boxed{0} & 2 & 1 \\ \boxed{1/2} & 3/2 & 5 \end{pmatrix} \quad P_2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Ahora, hacemos ceros por debajo del elemento a_{22} restando la segunda fila multiplicada por el real construído con el pivote (a_{22}) en el denominador y el elemento de esa fila por debajo del pivote (a_{32}) en el numerador.

$$\begin{pmatrix} 2 & 1 & 0 \\ \boxed{0} & 2 & 1 \\ \boxed{1/2} & 3/2 & 5 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \rightarrow & f_1 \\ f_2 & \rightarrow & f_2 \\ f_3 & \rightarrow & f_3 - \boxed{(3/2)/2}\ f_2 \end{array}$$

Y llegamos a la matriz que almacena simultáneamente L y U.

$$\begin{pmatrix} 2 & 1 & 0 \\ \boxed{0} & 2 & 1 \\ \boxed{1/2} & \boxed{3/4} & 17/4 \end{pmatrix}$$

Y las matrices L, U y P son:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \boxed{0} & 1 & 0 \\ \boxed{1/2} & \boxed{3/4} & 1 \end{pmatrix} \quad U = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 17/4 \end{pmatrix} \quad P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

2. Ahora, teniendo en cuenta que queremos resolver $Ax = b$, $PAx = Pb$ y que $PA = LU$ podemos escribir $LUx = Pb$ y si llamamos $Ux = y$:
Resolvemos el sistema triangular inferior $Ly = Pb$ $\left(Pb = (2, 4, 21)^T\right)$ y obtenemos y.

$$\begin{array}{rrrcll} y_1 & & & = & 2 & y_1 = 2 \\ & y_2 & & = & 4 & y_2 = 4 \\ (1/2)y_1 & +(3/4)y_2 & +y_3 & = & 21 & y_3 = 21 - (1/2)y_1 - (3/4)y_2 = 17 \end{array}$$

Resolvemos el sistema triangular superior $Ux = y$ y obtenemos x, que era lo que buscábamos.

$$\begin{array}{rrrcll} 2x_1 & +x_2 & & = & 2 & x_3 = 4 \\ & 2x_2 & +x_3 & = & 4 & x_2 = (4 - x_3)/2 = 0 \\ & & (17/4)x_3 & = & 17 & x_1 = (2 - x_2)/2 = 1 \end{array}$$

Y la solución es

$$\mathbf{x} = (1, 0, 4)^T$$

(b) El sistema no se puede resolver por Jacobi ni por Gauss-Seidel, porque la matriz de coeficientes tiene ceros en la diagonal principal y al construir el algoritmo, tanto de Jacobi como de Gauss-Seidel, utilizamos los elementos de la diagonal $a_{ii} \quad i = 1, \ldots, n$ como divisores y por lo tanto no pueden ser cero.

(c) Reordenando las filas tanto de A como de b tenemos:

$$A' = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 2 & 5 \end{pmatrix} \quad \mathbf{b}' = \begin{pmatrix} 2 \\ 4 \\ 21 \end{pmatrix}$$

y la matriz de coeficientes de este sistema es diagonal dominante por filas porque

$$A' = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 2 & 5 \end{pmatrix} \begin{matrix} |2| > |1| + |0| \\ |2| > |0| + |1| \\ |5| > |1| + |2| \end{matrix}$$

Ahora el sistema convergería tanto por Jacobi como por Gauss-Seidel porque una condición suficiente de convergencia del método en ambos casos, es que la matriz de coeficientes sea diagonal dominante.

Tema 6

Optimización

6.1. Condiciones necesarias y suficientes de óptimo

Problema 6.1:

Dados los puntos

x	-1	1	4	5
y	3	-1	-2	3

encontrar el punto central (m_x, m_y) de forma que la suma de las distancias cuadráticas

$$D = \sum_{i=1}^{n} \left[(m_x - x_i)^2 + (m_y - y_i)^2\right]$$

a dicho punto sea mínima.

La función a optimizar es

$$\begin{aligned} D(m_x, m_y) &= d_1^2 + d_2^2 + d_3^2 + d_4^2 \\ D(m_x, m_y) &= [(m_x - (-1))^2 + (m_y - 3)^2] + [(m_x - 1)^2 + (m_y - (-1))^2] + \\ &\quad [(m_x - 4)^2 + (m_y - (-2))^2] + [(m_x - 5)^2 + (m_y - 3)^2] \end{aligned}$$

(a) La condición necesaria de óptimo es que las derivadas parciales sean cero

$$\frac{\partial D}{\partial m_x} = 0$$
$$\frac{\partial D}{\partial m_y} = 0$$

Derivando

$$\begin{aligned} \frac{\partial D}{\partial m_x} = 0 \qquad & 2(m_x - (-1)) + 2(m_x - 1) + 2(m_x - 4) + 2(m_x - 5) = 0 \\ \frac{\partial D}{\partial m_y} = 0 \qquad & 2(m_y - 3) + 2(m_y - (-1)) + 2(m_y - (-2)) + 2(m_y - 3) = 0 \end{aligned}$$

Reorganizamos las ecuaciones

$$\begin{aligned}(-1+1+4+5) &= 4m_x\\ (3-1-2+3) &= 4m_y\end{aligned}$$

La solución es

$$m_x = \frac{-1+1+4+5}{4} = \frac{9}{4} = 2{,}25$$
$$m_y = \frac{3-1-2+3}{4} = \frac{3}{4} = 0{,}75$$

(b) La condición suficiente es que la matriz hessiana sea definida positiva en el punto anterior. La matriz Hessiana para la función f en cualquier punto (x, y) es

$$H_D = \begin{pmatrix} f''_{m_x m_x} & f''_{m_x m_y} \\ f''_{m_y m_x} & f''_{m_y m_y} \end{pmatrix} = \begin{pmatrix} 2+2+2+2 & 0 \\ 0 & 2+2+2+2 \end{pmatrix}.$$

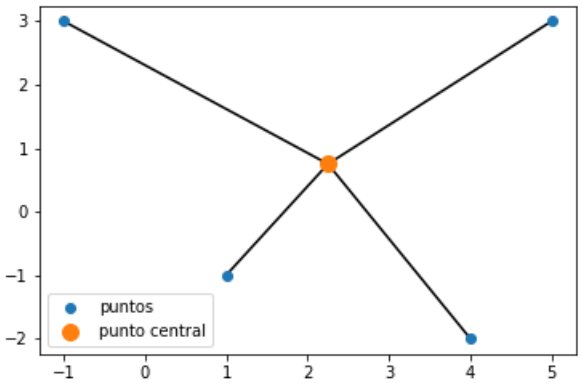
Y en particular, en el punto $(x_0, y_0) = (2{,}25, 0{,}75)$, como en este caso particular, la matriz Hessiana es constante

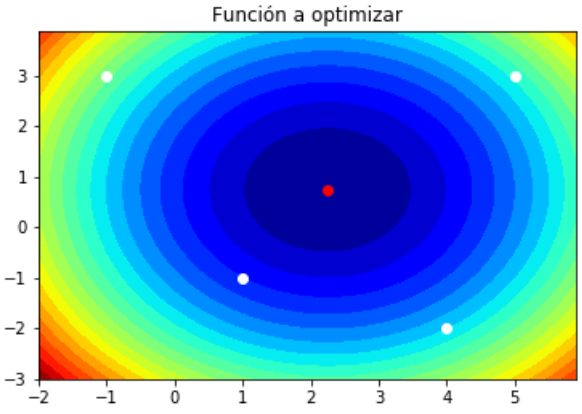
$$H_D(2{,}25, 0{,}75) = \begin{pmatrix} 8 & 0 \\ 0 & 8 \end{pmatrix}.$$

Veamos si es definida positiva. Para ello los determinantes de los menores principales han de ser estrictamente positivos. Es decir

$$|H_D(1,-1)| = \begin{vmatrix} 8 & 0 \\ 0 & 8 \end{vmatrix} = 64 > 0 \quad \text{y} \quad |8| = 8 > 0.$$

Por lo tanto la matriz es definida positiva y se cumple la condición suficiente de mínimo.





6.2. El método de Newton

Problema 6.2:

Sea $f(x,y) = x^2 + xy + y^2 - x + y$

(a) Hallar un mínimo local de f.

(b) Probar que dicho mínimo es, de hecho, global.

(c) Aproximarlo con una iteración por el método de Newton, tomando como punto inicial $(0,0)$.

(a) La condición necesaria de mínimo para un punto (x_m, y_m) es que $\nabla f\,(x_m, y_m) = \left(f'_x, f'_y\right) = (0,0)$. Es decir

$$\begin{cases} f'_x = 0 \\ f'_y = 0 \end{cases} \qquad \begin{cases} 2x + y - 1 = 0 \\ x + 2y + 1 = 0 \end{cases} \qquad \begin{cases} \qquad\qquad 2(-1-2y) + y - 1 = 0 \to \\ x = -1 - 2y \quad \nearrow \end{cases}$$

$$\begin{cases} \to y = -1 \quad \searrow \\ \qquad\qquad x = 1 \end{cases}$$

Y el punto $(x_m, y_m) = (1, -1)$ cumple las condiciones necesarias de mínimo. Veamos si también cumple la condición suficiente, que es que la matriz Hessiana en el punto (x_m, y_m) sea definida positiva. La matriz Hessiana para la función f en cualquier punto (x, y) es

$$H_f = \begin{pmatrix} f''_{xx} & f''_{xy} \\ f''_{yx} & f''_{yy} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Y en particular, en el punto $(x_m, y_m) = (1, -1)$

$$H_f(1,-1) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Veamos si es definida positiva. Para ello, los determinantes de los menores principales han de ser estrictamente positivos. Es decir

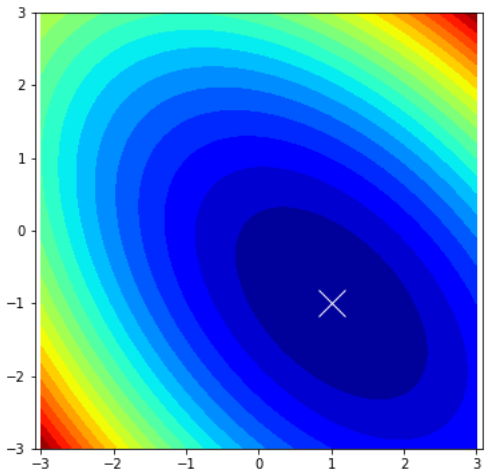
$$|H_f(1,-1)| = \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} = 3 > 0 \quad \text{y} \quad |2| = 2 > 0.$$

Por lo tanto, la matriz es definida positiva y se cumple la condición suficiente de mínimo.

(b) Se tiene que:

- Cualquier mínimo local de una función convexa es también un mínimo absoluto.
- Si para todo elemento del dominio de f, la matriz hessiana $H_f(a)$ es definida positiva, entonces, la función f es estrictamente convexa.

En este caso, la matriz Hessiana f en cualquier punto coincide con la matriz Hessiana en $(1,-1)$ y ya hemos demostrado que es definida positiva, por lo que la función f es convexa y el mínimo local $(1,-1)$ es también mínimo global. La figura siguiente muestra la representación de la función f con curvas de nivel. El mínimo aparece representado con una cruz.



(c) Realizaremos una iteración por el método de Newton usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

Si consideramos que

$$H_{(x_0,y_0)} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \nabla f_{(x_0,y_0)} \tag{6.1}$$

entonces

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

y escribiremos

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{6.2}$$

donde $(c_1, c_2)^T$ es la solución del sistema (6.1).

Tenemos que

$$\nabla f = \left(\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y} \right) = (2x + y - 1, x + 2y + 1)$$

y

$$\nabla f_{(x_0,y_0)} = \nabla f_{(0,0)} = (-1, 1)\,.$$

Además

$$H = \begin{pmatrix} \frac{\partial^2 f(x,y)}{\partial x\, \partial x} & \frac{\partial^2 f(x,y)}{\partial x\, \partial y} \\ \frac{\partial^2 f(x,y)}{\partial y\, \partial x} & \frac{\partial^2 f(x,y)}{\partial y\, \partial y} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

y

$$H_{(x_0,y_0)} = H_{(0,0)} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

El sistema (6.1) es

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

y resolviéndolo

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Por lo tanto (6.2) es

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Y $(-1, 1)$ es la aproximación al mínimo con una iteración de Newton. De hecho, hemos alcanzado el mínimo en una sola iteración.

Problema 6.3:

Sea $f(x, y) = x^2 + xy + y^2 - 2y + 2x$

(a) Hallar un mínimo local de f.

(b) Probar que dicho mínimo es, de hecho, global.

(c) Aproximarlo con una iteración por el método de Newton tomando como punto inicial $(1, 1)$.

(a) La condición necesaria de mínimo para un punto (x_m, y_m) es que $\nabla f(x_m, y_m) = \left(f'_x, f'_y\right) = (0, 0)$. Es decir

$$\begin{cases} f'_x = 0 \\ f'_y = 0 \end{cases} \quad \begin{cases} 2x + y + 2 = 0 \\ x + 2y - 2 = 0 \end{cases} \rightarrow \quad \begin{cases} & 2(2 - 2y) + y = 0 \rightarrow \\ x = 2 - 2y & \nearrow \end{cases}$$

$$\begin{cases} \rightarrow y = 2 & \searrow \\ & x = -2 \end{cases}$$

Y el punto $(x_m, y_m) = (-2, 2)$ cumple las condiciones necesarias de mínimo. Veamos si también cumple la condición suficiente que es que la matriz Hessiana en el punto (x_m, y_m) sea definida positiva. La matriz Hessiana para la función f en cualquier punto (x, y) es

$$H_f = \begin{pmatrix} f''_{xx} & f''_{xy} \\ f''_{yx} & f''_{yy} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Y en particular, en el punto $(x_m, y_m) = (-2, 2)$

$$H_f(-2, 2) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Veamos si es definida positiva. Para ello los determinantes de los menores principales han de ser estrictamente positivos. Es decir

$$|H_f(-1,2)| = \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} = 3 > 0 \quad \text{y} \quad |2| = 2 > 0.$$

Por lo tanto la matriz es definida positiva y se cumple la condición suficiente de mínimo.

(b) Se tiene que:

- Cualquier mínimo local de una función convexa es también un mínimo absoluto.
- Si para todo elemento del dominio de f, la matriz hessiana $H_f(a)$ es definida positiva, entonces, la función f es estrictamente convexa.

En este caso particular, la matriz Hessiana f en cualquier punto coincide con la matriz Hessiana en $(-1,2)$ y ya hemos demostrado que es definida positiva, por lo que la función f es convexa y el mínimo local $(-1,2)$ es también mínimo global.

(c) Realizaremos una iteración por el método de Newton usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

Si consideramos que

$$H_{(x_0,y_0)} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \nabla f_{(x_0,y_0)} \tag{6.3}$$

entonces

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

y escribiremos

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{6.4}$$

donde $(c_1, c_2)^T$ es la solución del sistema (6.3).

Del apartado anterior tenemos que

$$\nabla f = (2x + y + 2, x + 2y - 2)$$

y

$$\nabla f_{(x_0,y_0)} = \nabla f_{(1,1)} = (5,1)\,.$$

Además

$$H_f = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

y

$$H_{(x_0,y_0)} = H_{(1,1)} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

El sistema (6.3) es

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$$

y resolviéndolo

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \end{pmatrix}.$$

Por lo tanto (6.4) es

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 \\ -1 \end{pmatrix} = \begin{pmatrix} -2 \\ 2 \end{pmatrix}.$$

Y $(-2, 2)$ es la aproximación al mínimo con una iteración de Newton. De hecho, hemos alcanzado el mínimo en una sola iteración.

Problema 6.4:

Sea $f(x, y, z) = (x-1)^4 + (y-2)^2 + (z-3)^2$. Aproximar el mínimo utilizando el método de Newton y tomando como punto inicial $\left(x^{(0)}, y^{(0)}, z^{(0)}\right) = (0, 0, 0)$. Realizar una iteración.

Realizaremos una iteración por el método de Newton usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} - H^{-1}_{(x_0,y_0,z_0)} \cdot \nabla f_{(x_0,y_0,z_0)}$$

Si consideramos que

$$H_{(x_0,y_0,z_0)} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \nabla f_{(x_0,y_0,z_0)} \tag{6.5}$$

entonces

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = H^{-1}_{(x_0,y_0,z_0)} \cdot \nabla f_{(x_0,y_0,z_0)}$$

y escribiremos

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \tag{6.6}$$

donde $(c_1, c_2, c_3)^T$ es la solución del sistema (6.5).
Tenemos que

$$\nabla f = \left(\frac{\partial f(x,y,z)}{\partial x}, \frac{\partial f(x,y,z)}{\partial y}, \frac{\partial f(x,y,z)}{\partial z} \right) = \left(4(x-1)^3, 2(y-2), 2(z-3) \right)$$

y

$$\nabla f_{(x_0,y_0,z_0)} = \nabla f_{(0,0,0)} = (-4, -4, -6).$$

Además

$$H = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{pmatrix} = \begin{pmatrix} 12(x-1)^2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

y

$$H_{(x_0,y_0,z_0)} = H_{(0,0,0)} = \begin{pmatrix} 12 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

El sistema (6.5) es

$$\begin{pmatrix} 12 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} -4 \\ -4 \\ -6 \end{pmatrix}$$

y resolviéndolo

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} -1/3 \\ -2 \\ -3 \end{pmatrix}.$$

Por lo tanto (6.6) es

$$\begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} -1/3 \\ -2 \\ -3 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 2 \\ 3 \end{pmatrix}.$$

Y $(1/3, 2, 3)$ es la aproximación al mínimo con una iteración de Newton. Hemos mejorado respecto al punto inicial porque $f(x_0, y_0, z_0) = f(0,0,0) = 1+4+9 = 14$ y $f(x_1, y_1, z_1) = f(1/3, 2, 3) \approx 0{,}2$ que es menor.

Problema 6.5:

Dada la función

$$f(x, y) = x^2y^2 + x^2 + y^2 + x + y$$

Aproximar el mínimo de la función comenzando por el punto inicial $(0, 0)$ y utilizando el método de Newton (una iteración).

Realizaremos una iteración por el método de Newton usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

Si consideramos que

$$H_{(x_0,y_0)} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \nabla f_{(x_0,y_0)} \tag{6.7}$$

entonces

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

y escribiremos

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{6.8}$$

donde $(c_1, c_2)^T$ es la solución del sistema (6.7).

Tenemos que

$$\nabla f = \left(\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y} \right) = \left(1 + 2x + 2xy^2, 1 + 2y + 2x^2 y\right)$$

y

$$\nabla f_{(x_0,y_0)} = \nabla f_{(0,0)} = (1,1)\,.$$

Además

$$H = \begin{pmatrix} \frac{\partial^2 f(x,y)}{\partial x\,\partial x} & \frac{\partial^2 f(x,y)}{\partial x\,\partial y} \\ \frac{\partial^2 f(x,y)}{\partial y\,\partial x} & \frac{\partial^2 f(x,y)}{\partial y\,\partial y} \end{pmatrix} = \begin{pmatrix} 2 + 2y^2 & 4xy \\ 4xy & 2 + 2x^2 \end{pmatrix}$$

y

$$H_{(x_0,y_0)} = H_{(0,0)} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

El sistema (6.7) es

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

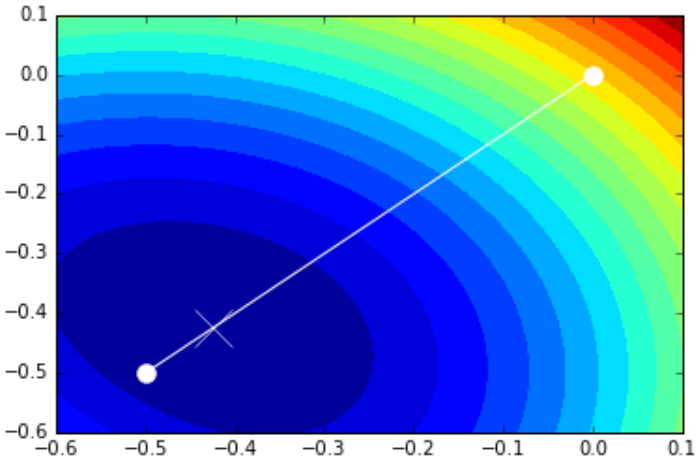
y resolviéndolo

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0{,}5 \\ 0{,}5 \end{pmatrix}.$$

Por lo tanto (6.8) es

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 0{,}5 \\ 0{,}5 \end{pmatrix} = \begin{pmatrix} -0{,}5 \\ -0{,}5 \end{pmatrix}.$$

Y la aproximación del mínimo con una iteración de Newton es $(-0{,}5, -0{,}5)$. Hemos mejorado respecto al punto inicial y ahora estamo más cerca del mínimo porque $f\,(x_0, y_0) = f(0,0) = 0$ y $f\,(x_1, y_1) = f(-0{,}5, -0{,}5) = -0{,}4375$ que es menor. (El mínimo está en $(-0{,}424, -0{,}424)$, señalado con una cruz en el gráfico).



Problema 6.6:

Dada la función

$$f(x,y) = (x-1)^2 + y^2 - 2y$$

Aproximar el mínimo de la función comenzando por el punto inicial $(0,0)$ y utilizando el método de Newton (una iteración).

Realizaremos una iteración por el método de Newton usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

Si consideramos que

$$H_{(x_0,y_0)} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \nabla f_{(x_0,y_0)} \tag{6.9}$$

entonces

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

y escribiremos

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{6.10}$$

donde $(c_1, c_2)^T$ es la solución del sistema (6.5).

Tenemos que

$$\nabla f = \left(\frac{\partial f(x,y)}{\partial x}, \frac{\partial f(x,y)}{\partial y} \right) = (2(x-1), 2y-2)$$

y

$$\nabla f_{(x_0,y_0)} = \nabla f_{(0,0)} = (-2,-2)\,.$$

Además

$$H = \begin{pmatrix} \frac{\partial^2 f(x,y)}{\partial x\,\partial x} & \frac{\partial^2 f(x,y)}{\partial x\,\partial y} \\ \frac{\partial^2 f(x,y)}{\partial y\,\partial x} & \frac{\partial^2 f(x,y)}{\partial y\,\partial y} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

y

$$H_{(x_0,y_0)} = H_{(0,0)} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

El sistema (6.9) es

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} -2 \\ -2 \end{pmatrix}$$

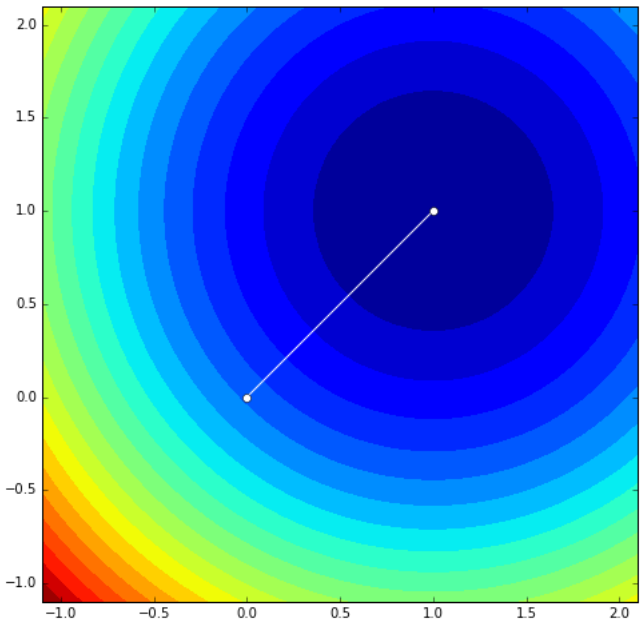
y resolviéndolo

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$$

Por lo tanto (6.10) es

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Y la aproximación del mínimo es $(1, 1)$. Hemos mejorado porque $f(x_0, y_0) = f(0, 0) = 1$ y $f(x_1, y_1) = f(1, 1) = -1$ que es menor. Además, en este caso, coincide con el mínimo global.



6.3. El método del gradiente

Problema 6.7:

Utilizando el método del gradiente, aproximar el mínimo de la función

$$f(x,y) = x^4 + y^4$$

Utilizar el punto $(1,1)$ como punto inicial y realizar una iteración.

Realizaremos una iteración por el método del gradiente usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - h\nabla f_{(x_0,y_0)} \tag{6.11}$$

y buscaremos h para que

$$g(h) = f\left(x_0 - hf_x(x_0,y_0), y_0 - hf_y(x_0,y_0)\right)$$

tenga un valor mínimo.
Tenemos que

$$\nabla f = (f_x, f_y) = \left(4x^3, 4y^3\right)$$

y

$$\nabla f_{(x_0,y_0)} = \nabla f_{(1,1)} = (4,4)\,.$$

$g(h) = f\left(1 - hf_x(1,1), 1 - hf_y(1,1)\right) = f\left(1-4h, 1-4h\right) = (1-4h)^4 + (1-4h)^4 = 2\left(1-4h\right)^4$
Optimicemos esta función:

$$g'(h) = 8(1-4h)^3(-4) = 0 \Longrightarrow 1 - 4h = 0 \Longrightarrow h = \frac{1}{4}$$

y usando este valor en la ecuación (6.11)

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{4}\begin{pmatrix} 4 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Hemos llegado al mínimo porque $f(x_0,y_0) = f(1,1) = 2$ y $f(x_1,y_1) = f(0,0) = 0$ que es el menor valor posible puesto que f es siempre positivo.

Problema 6.8:

Dada la función

$$f(x,y) = (x-1)^2 + y^2$$

(a) Hallar un mínimo local de f.

(b) Probar que dicho mínimo es, de hecho, global.

(c) Aproximar el mínimo de la función comenzando por el punto inicial $(0,1)$ y utilizando el método del gradiente (una iteración).

(a) La condición necesaria de mínimo para un punto (x_m, y_m) es que $\nabla f(x_m, y_m) = \left(f'_x, f'_y\right) = (0,0)$. Es decir

$$\begin{cases} f'_x = 0 \\ f'_y = 0 \end{cases} \quad \begin{cases} 2(x-1) = 0 & \rightarrow \\ 2y = 0 & \rightarrow \end{cases} \quad \begin{cases} x = 1 \\ y = 0 \end{cases}$$

Y el punto $(x_m, y_m) = (1,0)$ cumple las condiciones necesarias de mínimo. Veamos si también cumple la condición suficiente que es que la matriz Hessiana en el punto (x_m, y_m) sea definida positiva. La matriz Hessiana para la función f en cualquier punto (x, y) es

$$H_f = \begin{pmatrix} f''_{xx} & f''_{xy} \\ f''_{yx} & f''_{yy} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Y en particular, en el punto $(x_m, y_m) = (1,0)$

$$H_f(1,0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Veamos si es definida positiva. Para ello los determinantes de los menores principales han de ser estrictamente positivos. Es decir

$$|H_f(1,0)| = \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} = 4 > 0 \quad \text{y} \quad |2| = 2 > 0.$$

Por lo tanto la matriz es definida positiva y se cumple la condición suficiente de mínimo.

(b) Se tiene que:

- Cualquier mínimo local de una función convexa es también un mínimo absoluto.
- Si para todo elemento del dominio de f, la matriz hessiana $H_f(a)$ es definida positiva, entonces, la función f es estrictamente convexa.

En este caso particular, la matriz Hessiana f en cualquier punto coincide con la matriz Hessiana en $(1,0)$ y ya hemos demostrado que es definida positiva, por lo que la función f es convexa y el mínimo local $(1,0)$ es también mínimo global.

(c) Realizaremos una iteración por el método del gradiente usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - h\nabla f_{(x_0,y_0)} \tag{6.12}$$

y buscaremos h para que

$$g(h) = f\left(x_0 - hf_x(x_0,y_0), y_0 - hf_y(x_0,y_0)\right)$$

tenga un valor mínimo.
Tenemos que

$$\nabla f = (f_x, f_y) = (2(x-1), 2y)$$

y

$$\nabla f_{(x_0,y_0)} = \nabla f_{(1,1)} = (-2, 2).$$

Sustituyendo $x_0 = 0$ e $y_0 = 1$ en $g(h)$ tenemos

$$g(h) = f\left(0 - hf_x(0,1), 1 - hf_y(0,1)\right) = f\left(0 + 2h, 1 - 2h\right) =$$

$$= (2h-1)^2 + (1-2h)^2 = 2\left(1-2h\right)^2$$

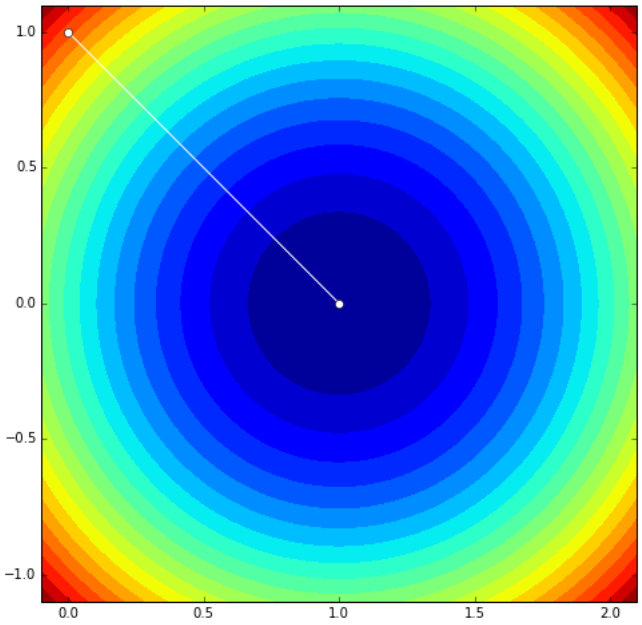
Optimicemos esta función:

$$g'(h) = 4\left(1-2h\right)(-2) = 0 \Longrightarrow 1 - 2h = 0 \Longrightarrow h = \frac{1}{2}$$

y usando este valor en la ecuación (6.12)

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} -2 \\ 2 \end{pmatrix} = \begin{pmatrix} 0+1 \\ 1-1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Hemos llegado al mínimo porque $f\left(x_0, y_0\right) = f(0,1) = 2$ y $f\left(x_1, y_1\right) = f(1,0) = 0$ que es el menor valor posible puesto que f es siempre positivo.



6.4. Optimización con restriciones

Problema 6.9:

Hallar el mínimo sujeto a las restricciones $x \geq 0$ e $y \geq 0$ para la función

$$f(x,y) = x^2 + xy + y^2 - x + y$$

Necesitamos calcular los extremos en:

(a) Todo el dominio de la función.

(b) Las fronteras de la zona donde buscamos el mínimo.

(c) La intersección de las fronteras.

Estudiemos entonces el mínimo por partes:

(a) *Todo el dominio de la función.* La condición necesaria de mínimo para un punto (x_m, y_m) es que $\nabla f\,(x_m, y_m) = \left(f'_x, f'_y\right) = (0,0)$. Es decir

$$\begin{cases} f'_x = 0 \\ f'_y = 0 \end{cases} \quad \begin{cases} 2x + y - 1 = 0 \\ x + 2y + 1 = 0 \end{cases} \quad \begin{cases} \qquad\qquad 2(-1-2y) + y - 1 = 0 \rightarrow \\ x = -1 - 2y \quad \nearrow \end{cases}$$

$$\begin{cases} \rightarrow y = -1 \quad \searrow \\ \qquad\qquad x = 1 \end{cases}$$

Y el punto $(x_m, y_m) = (1, -1)$ cumple las condiciones necesarias de mínimo.

(b) *Las fronteras de la zona donde buscamos el mínimo.* Las fronteras son $y = 0$ y $x = 0$.

$$\begin{cases} y = 0 \rightarrow \quad g_1(x) = f(x,0) = x^2 - x \\ x = 0 \rightarrow \quad g_2(y) = f(0,y) = y^2 + y \end{cases}$$

Si evaluamos las condiciones necesarias para cada función

$$\begin{cases} y = 0 \rightarrow \quad g'_1(x) = 2x - 1 = 0 \rightarrow x = 1/2 \\ x = 0 \rightarrow \quad g'_2(y) = 2y + 1 = 0 \rightarrow y = -1/2 \end{cases}$$

Y las condiciones suficientes

$$\begin{cases} y = 0 \rightarrow \quad g''_1(x) = 2 \rightarrow g''_1(1/2) = 2 > 0 \rightarrow min \\ x = 0 \rightarrow \quad g''_2(y) = 2 \rightarrow g''_2(-1/2) = 2 > 0 \rightarrow min \end{cases}$$

Por lo tanto

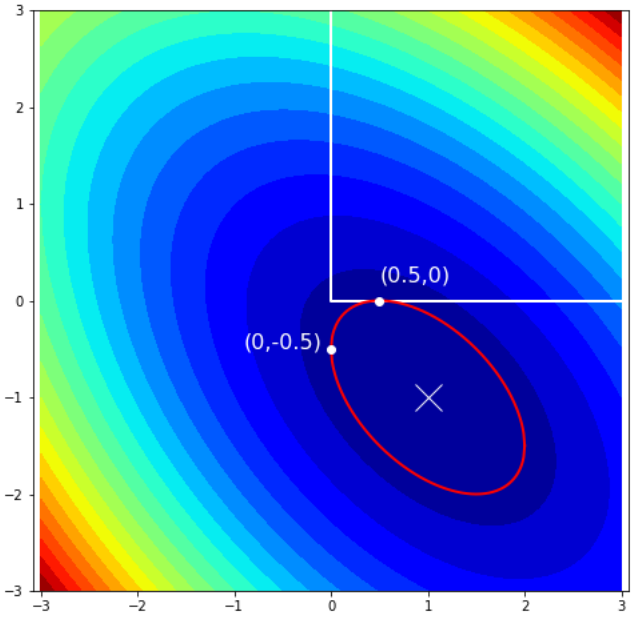
$$\begin{cases} y = 0 \rightarrow \quad (1/2, 0) \rightarrow min \\ x = 0 \rightarrow \quad (0, -1/2) \rightarrow min \end{cases}$$

(c) *La intersección de las fronteras.* Es el punto $(0,0)$ que también es un mínimo para la frontera $y = 0$.

Resumiendo, hemos obtenido los puntos

(x,y)	¿Pertenece a la región?	$f(x,y)$
$(1,-1)$	No	$-$
$\mathbf{(0{,}5,0)}$	**Si**	$\mathbf{-0{,}25}$
$(0,-0{,}5)$	No	$-$
$(0,0)$	Si	0

Como de los puntos que pertenecen a la región el mínimo valor le corresponde a $(0{,}5,0)$ este es el mínimo global para la región que cumple las restricciones $x \geq 0,\ y \geq 0$.
La representación gráfica de la función f, la frontera, los puntos calculados y la isolínea de la función que pasa por el mínimo es



Problema 6.10:

Proponer una función objetivo para minimizar la función

$$f(x,y) = 22 - 4x + x^2 - 12y + 2y^2$$

con las restricciones $x \geq 0$ y $y \geq 0$ utilizando el método de penalización.

La idea del método de penalización es reemplazar la función objetivo f por otra función

$$F(x,y) = f(x,y) + cP(x,y)$$

y resolver el problema sin restricciones. Para ello tomamos c como una constante positiva y P que satisface:

- P es continua en el dominio de f.
- $P(x, y) \geq 0$ para todo punto del domino de f, y
- $P(x, y) = 0$ si y solo si el punto (x, y) satisface las restricciones.

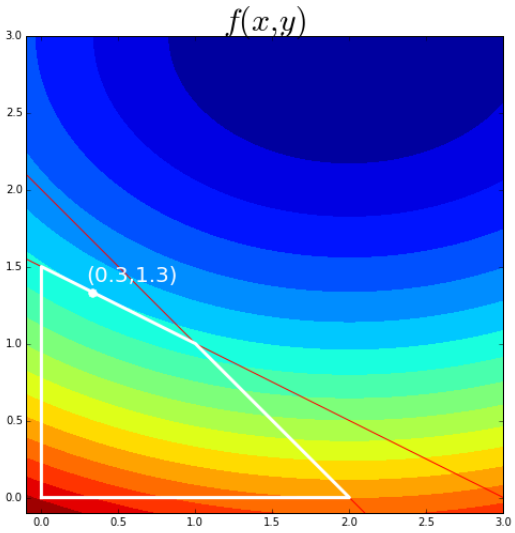
Una posible función para aproximar el mínimo con restricciones sería

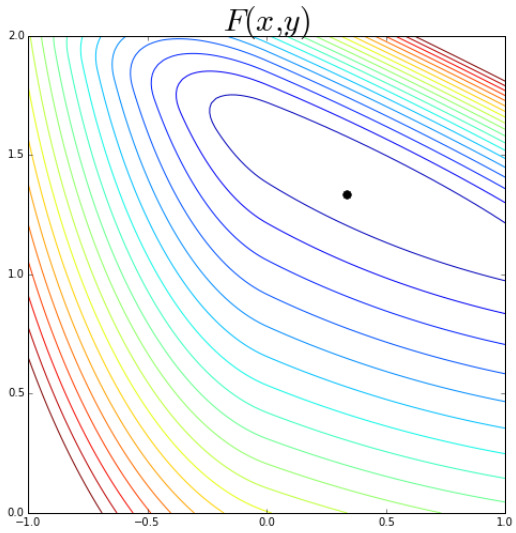
$$F(x, y) = f(x, y) + 10g_1(x, y) + 10g_2(x, y) + 10g_3(x, y) + 10g_4(x, y)$$

$$g_1(x, y) = \begin{cases} 0 & \text{si } x \geq 0 \\ x^2 & \text{si } x < 0 \end{cases} \qquad g_2(x, y) = \begin{cases} 0 & \text{si } y \geq 0 \\ y^2 & \text{si } y < 0 \end{cases}$$

$$g_3(x, y) = \begin{cases} 0 & \text{si } x + y \leq 2 \\ (x + y - 2)^2 & \text{si } x + y > 2 \end{cases} \qquad g_4(x, y) = \begin{cases} 0 & \text{si } x + 2y \leq 3 \\ (x + 2y - 3)^2 & \text{si } x + 2y > 3 \end{cases}$$

Si representamos la función $f(x, y)$ el mínimo está donde el contorno de la zona donde buscamos el mínimo con restricciones es tangente a la curva de nivel, en el punto (0,3, 1,3). Si representamos la función $F(x, y)$ vemos que tiene un mínimo cerca de (0,3, 1,3).





Problema 6.11:

Sea $f(x,y) = x^2 + xy + y^2 - 3y$.

(a) Hallar un mínimo local de f.

(b) Probar que dicho mínimo es, de hecho, global.

(c) Aproximarlo con una iteración por el método de Newton tomando como punto inicial $(0,0)$.

(d) Hallar el mínimo sujeto a las restricciones $x \geq 0$ e $y \geq 0$.

(e) Proponer una función objetivo utilizando el método de penalización para aproximar el mínimo del apartado anterior.

(a) La condición necesaria de mínimo para un punto (x_m, y_m) es que $\nabla f\,(x_m, y_m) = \left(f'_x, f'_y\right) = (0,0)$. Es decir

$$\begin{cases} f'_x = 0 \\ f'_y = 0 \end{cases} \quad \begin{cases} 2x + y = 0 \\ x + 2y - 3 = 0 \end{cases} \rightarrow \quad \begin{cases} & 2(3-2y) + y = 0 \rightarrow \\ x = 3 - 2y & \nearrow \end{cases}$$

$$\begin{cases} y = 2 & \searrow \\ & x = -1 \end{cases}$$

Y el punto $(x_m, y_m) = (-1, 2)$ cumple las condiciones necesarias de mínimo. Veamos si también cumple la condición suficiente que es que la matriz Hessiana en el punto (x_m, y_m) sea definida positiva. La matriz Hessiana para la función f en cualquier punto (x, y) es

$$H_f = \begin{pmatrix} f''_{xx} & f''_{xy} \\ f''_{yx} & f''_{yy} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Y en particular, en el punto $(x_m, y_m) = (-1, 2)$

$$H_f(-1,2) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Veamos si es definida positiva. Para ello los determinantes de los menores principales han de ser estrictamente positivos. Es decir

$$|H_f(-1,2)| = \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} = 3 > 0 \quad \text{y} \quad |2| = 2 > 0.$$

Por lo tanto la matriz es definida positiva y se cumple la condición suficiente de mínimo.

(b) Se tiene que:

- Cualquier mínimo local de una función convexa es también un mínimo absoluto.

- Si para todo elemento del dominio de f, la matriz hessiana $H_f(a)$ es definida positiva, entonces, la función f es estrictamente convexa.

En este caso particular, la matriz Hessiana f en cualquier punto coincide con la matriz Hessiana en $(-1,2)$ y ya hemos demostrado que es definida positiva, por lo que la función f es convexa y el mínimo local $(-1,2)$ es también mínimo global.

(c) Realizaremos una iteración por el método de Newton usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

Si consideramos que

$$H_{(x_0,y_0)} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \nabla f_{(x_0,y_0)} \tag{6.13}$$

entonces

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

y escribiremos

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{6.14}$$

donde $(c_1, c_2)^T$ es la solución del sistema (6.13).
Del apartado anterior tenemos que

$$\nabla f = (2x + y, x + 2y - 3)$$

y

$$\nabla f_{(x_0,y_0)} = \nabla f_{(0,0)} = (0, -3)\,.$$

Además

$$H_f = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

y

$$H_{(x_0,y_0)} = H_{(0,0)} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

El sistema (6.13) es

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -3 \end{pmatrix}$$

y resolviéndolo

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

Por lo tanto (6.14) es

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

Hemos mejorado porque $f(x_0, y_0) = f(0,0) = 0$ y $f(x_1, y_1) = f(-1,2) = -3$. Además $(-1,2)$ es el mínimo global.

(d) Necesitamos calcular los extremos en

- Todo el dominio de la función.
- Las fronteras de la zona donde buscamos el mínimo.
- La intersección de las fronteras.

Todo el dominio de la función. Ya está calculado en el apartado a) y es el punto $(-1, 2)$.
Las fronteras de la zona donde buscamos el mínimo. Las fronteras son $y = 0$ y $x = 0$.

$$\begin{cases} y = 0 \rightarrow & g_1(x) = f(x, 0) = x^2 \\ x = 0 \rightarrow & g_2(y) = f(0, y) = y^2 - 3y \end{cases}$$

Si evaluamos las condiciones necesarias para cada función

$$\begin{cases} y = 0 \rightarrow & g_1'(x) = 2x = 0 \rightarrow x = 0 \\ x = 0 \rightarrow & g_2'(y) = 2y - 3 = 0 \rightarrow y = 3/2 \end{cases}$$

Y las condiciones suficientes

$$\begin{cases} y = 0 \rightarrow & g_1''(x) = 2 \rightarrow g_1''(0) = 2 > 0 \rightarrow min \\ x = 0 \rightarrow & g_2''(y) = 2 \rightarrow g_2''(3/2) = 2 > 0 \rightarrow min \end{cases}$$

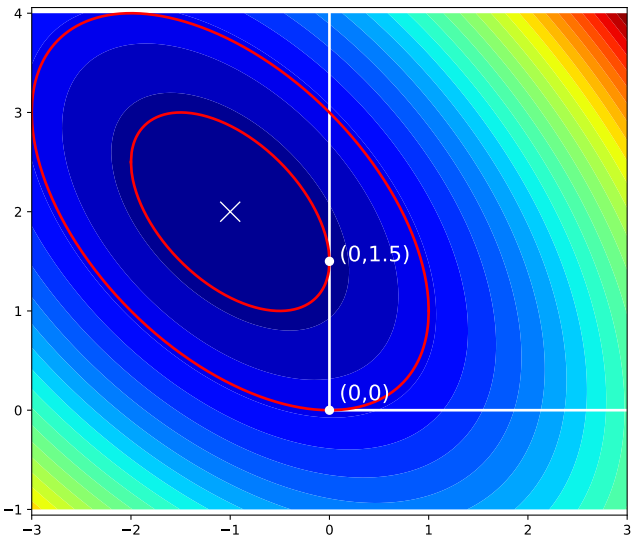
Por lo tanto

$$\begin{cases} y = 0 \rightarrow & (0, 0) \rightarrow min \\ x = 0 \rightarrow & (0, 3/2) \rightarrow min \end{cases}$$

La intersección de las fronteras. Es el punto $(0, 0)$ que también es un mínimo para la frontera $y = 0$.
Resumiendo, hemos obtenido los puntos

(x, y)	¿Pertenece a la región?	$f(x, y)$
$(-1, 2)$	No	$-$
$\mathbf{(0, 3/2)}$	**Si**	$\mathbf{-9/4}$
$(0, 0)$	Si	0

Como de los puntos que pertenecen a la región el mínimo valor le corresponde a $(0, 3/2)$ este es el mínimo global para la región que cumple las restricciones $x \geq 0$, $y \geq 0$.
La representación gráfica de la función f, la frontera, los puntos calculados y las isolíneas tangentes a las fronteras de la zona donde calculamos el mínimo se representan en la figura siguiente



(e) La idea del método de penalización es reemplazar la función objetivo f por otra función

$$F(x,y) = f(x,y) + cP(x,y)$$

y resolver el problema sin restricciones. Para ello tomamos c como una constante positiva y P satisfaciendo:

- P es continua en el dominio de f.
- $P(x,y) \geq 0$ para todo punto del domino de f, y
- $P(x,y) = 0$ si y solo si el punto (x,y) satisface las restricciones.

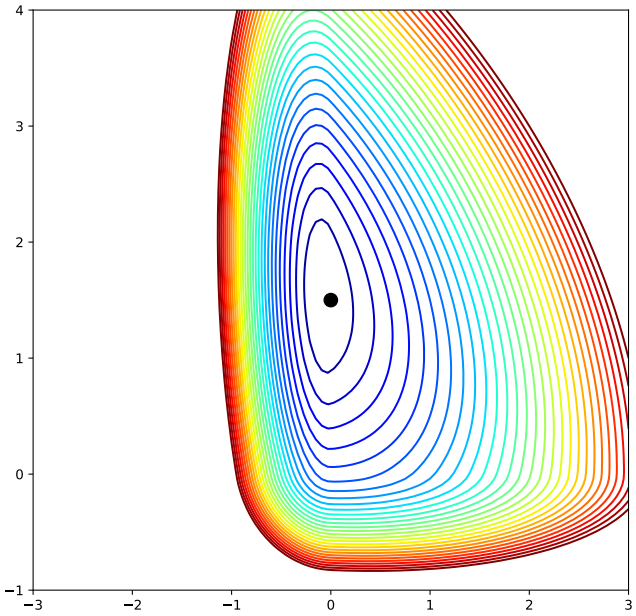
De acuerdo con ello, una función F podria ser

$$F(x,y) = f(x,y) + 10\phi_1(x,y) + 10\phi_2(x,y)$$

con

$$\phi_1(x,y) = \begin{cases} x^2 & \text{si} \quad x < 0 \\ 0 & \text{si} \quad x \geq 0 \end{cases} \qquad \phi_2(x,y) = \begin{cases} y^2 & \text{si} \quad y < 0 \\ 0 & \text{si} \quad y \geq 0 \end{cases}$$

Su representación gráfica es



Problema 6.12:

Sea $f(x,y) = x^2 + xy + y^2 - x + y$.

(a) Hallar un mínimo local de f.

(b) Probar que dicho mínimo es, de hecho, global.

(c) Aproximarlo con una iteración por el método de gradiente con el descenso más pronunciado, tomando como punto inicial $(0,0)$.

(d) Aproximarlo con dos iteraciones por el método de gradiente con tasa de aprendizaje $\eta = 0{,}5$, tomando como punto inicial $(0,0)$.

(e) Hallar el mínimo sujeto a las restricciones $x \geq 0$ e $y \geq 0$.

(f) Proponer una función objetivo utilizando el método de penalización para aproximar el mínimo del apartado anterior.

(a) La condición necesaria de mínimo para un punto (x_m, y_m) es que $\nabla f\,(x_m, y_m) = \left(f'_x, f'_y\right) = (0,0)$. Es decir

$$\begin{cases} f'_x = 0 \\ f'_y = 0 \end{cases} \qquad \begin{cases} 2x + y - 1 = 0 \\ x + 2y + 1 = 0 \end{cases} \qquad \begin{cases} \\ x = -1 - 2y \end{cases} \quad 2(-1-2y) + y - 1 = 0 \rightarrow \nearrow$$

$$\begin{cases} y = -1 & \searrow \\ & x = 1 \end{cases}$$

Y el punto $(x_m, y_m) = (1, -1)$ cumple las condiciones necesarias de mínimo. Veamos si también cumple la condición suficiente que es que la matriz Hessiana en el punto (x_m, y_m) sea definida positiva. La matriz Hessiana para la función f en cualquier punto (x, y) es

$$H_f = \begin{pmatrix} f''_{xx} & f''_{xy} \\ f''_{yx} & f''_{yy} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Y en particular, en el punto $(x_m, y_m) = (1, -1)$

$$H_f(1, -1) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Veamos si es definida positiva. Para ello los determinantes de los menores principales han de ser estrictamente positivos. Es decir

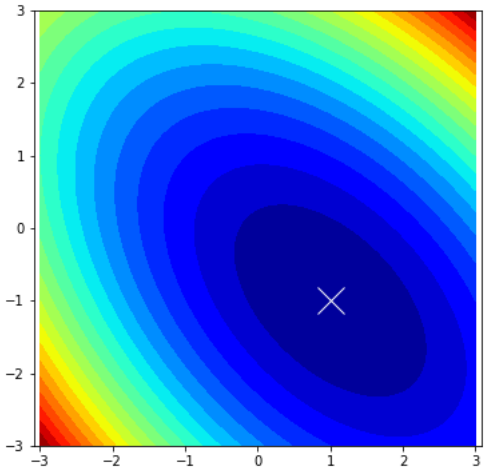
$$|H_f(1, -1)| = \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} = 3 > 0 \quad \text{y} \quad |2| = 2 > 0.$$

Por lo tanto la matriz es definida positiva y se cumple la condición suficiente de mínimo en $(1, -1)$.

(b) Se tiene que:

- Cualquier mínimo local de una función convexa es también un mínimo absoluto.
- Si para todo elemento del dominio de f, la matriz hessiana $H_f(a)$ es definida positiva, entonces, la función f es estrictamente convexa.

En este caso, la matriz Hessiana f en cualquier punto coincide con la matriz Hessiana en $(1, -1)$ y ya hemos demostrado que es definida positiva, por lo que la función f es convexa y el mínimo local $(1, -1)$ es también mínimo global.



(c) Realizaremos una iteración por el método del gradiente con el descenso más pronunciado usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - h\nabla f_{(x_0, y_0)} \tag{6.15}$$

y buscaremos h para que

$$g(h) = f\left(x_0 - hf_x(x_0, y_0), y_0 - hf_y(x_0, y_0)\right)$$

tenga un valor mínimo.
Tenemos que

$$\nabla f = (f_x, f_y) = (2x + y - 1, x + 2y + 1)$$

y

$$\nabla f_{(x_0,y_0)} = \nabla f_{(0,0)} = (-1, 1)\,.$$

$g(h) = f\left(x_0 - hf_x(0,0), y_0 - hf_y(0,0)\right) = f\left(0 + h, 0 - h\right) = h^2 - h^2 + h^2 - h - h = h^2 - 2h = 0$

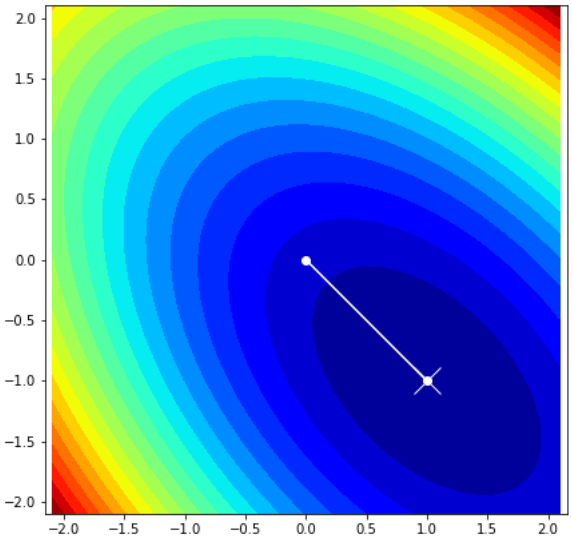
Optimicemos esta función:

$$g'(h) = 2h - 2 = 0 \Longrightarrow h = 1$$

y usando este valor en la ecuación (6.15)

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - (1) \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

y la aproximación del mínimo es $(1, -1)$ que, en este caso particular, es el mínimo global.



(d) Ahora realizaremos dos iteraciones con el método del gradiente con tasa de aprendizaje $\eta = 0{,}5$. La primera iteración

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \eta \nabla f_{(x_0,y_0)}$$

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \eta \begin{pmatrix} 2x_0 + y_0 - 1 \\ x_0 + 2y_0 + 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 0{,}5 \begin{pmatrix} 2(0) + (0) - 1 \\ (0) + 2(0) + 1 \end{pmatrix} = \begin{pmatrix} 0{,}5 \\ -0{,}5 \end{pmatrix}$$

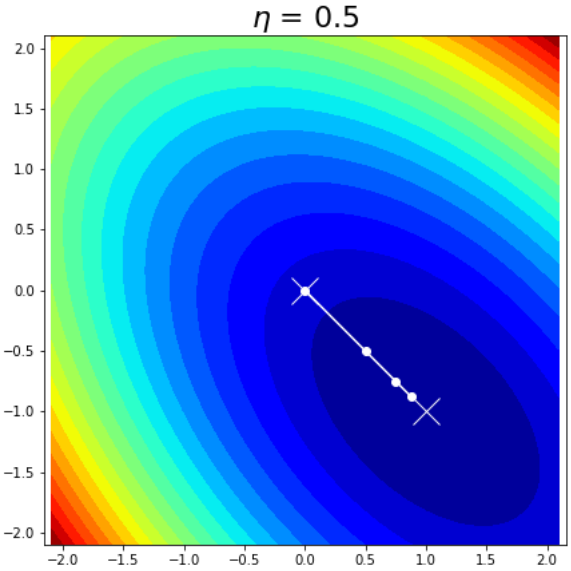
Y la segunda iteración

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \eta \nabla f_{(x_1, y_1)}$$

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \eta \begin{pmatrix} 2x_1 + y_1 - 1 \\ x_1 + 2y_1 + 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}5 \\ -0{,}5 \end{pmatrix} - 0{,}5 \begin{pmatrix} 2(0{,}5) + (-0{,}5) - 1 \\ (0{,}5) + 2(-0{,}5) + 1 \end{pmatrix} = \begin{pmatrix} 0{,}75 \\ -0{,}75 \end{pmatrix}$$

Y la aproximación al mínimo después de dos iteraciones es $(0{,}75, -0{,}75)$. En la siguiente figura se representan las tres primeras iteraciones, el punto inicial y el mínimo absoluto



(e) Necesitamos calcular los extremos en

- Todo el dominio de la función.
- Las fronteras de la zona donde buscamos el mínimo.
- La intersección de las fronteras.

Todo el dominio de la función. Ya está calculado en el apartado a) y es el punto $(1, -1)$.
Las fronteras de la zona donde buscamos el mínimo. Las fronteras son $y = 0$ y $x = 0$.

$$\begin{cases} y = 0 \rightarrow & g_1(x) = f(x, 0) = x^2 - x \\ x = 0 \rightarrow & g_2(y) = f(0, y) = y^2 + y \end{cases}$$

Si evaluamos las condiciones necesarias para cada función

$$\begin{cases} y = 0 \rightarrow & g_1'(x) = 2x - 1 = 0 \rightarrow x = 1/2 \\ x = 0 \rightarrow & g_2'(y) = 2y + 1 = 0 \rightarrow y = -1/2 \end{cases}$$

Y las condiciones suficientes

$$\begin{cases} y = 0 \rightarrow & g_1''(x) = 2 \rightarrow g_1''(1/2) = 2 > 0 \rightarrow min \\ x = 0 \rightarrow & g_2''(y) = 2 \rightarrow g_2''(-1/2) = 2 > 0 \rightarrow min \end{cases}$$

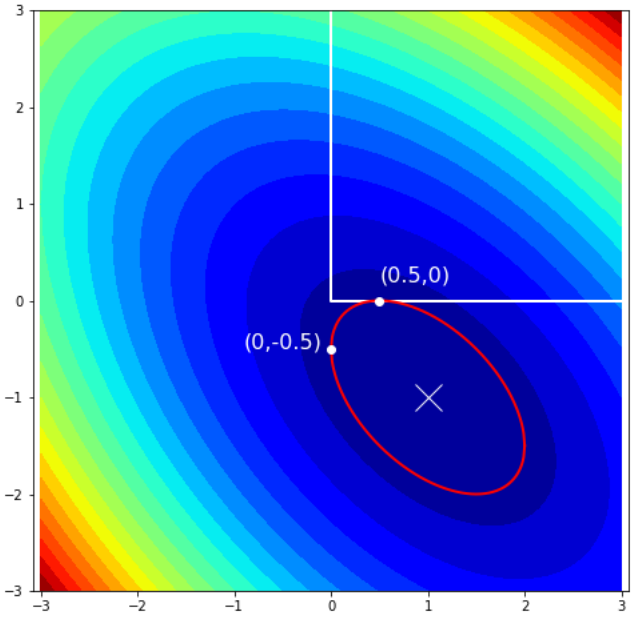
Por lo tanto

$$\begin{cases} y = 0 \rightarrow & (1/2, 0) \rightarrow min \\ x = 0 \rightarrow & (0, -1/2) \rightarrow min \end{cases}$$

La intersección de las fronteras. Es el punto $(0, 0)$ que también es un mínimo para la frontera $y = 0$.
Resumiendo, hemos obtenido los puntos

(x, y)	¿Pertenece a la región?	$f(x, y)$
$(1, -1)$	No	–
$\mathbf{(1/2, 0)}$	**Si**	$\mathbf{-1/4}$
$(0, -1/2)$	No	–
$(0, 0)$	Si	0

Como de los puntos que pertenecen a la región el mínimo valor le corresponde a $(1/2, 0)$, éste es el mínimo global para la región que cumple las restricciones $x \geq 0$, $y \geq 0$.
La función f, la frontera, los puntos calculados y la isolínea de la función que pasa por el mínimo se representan en la figura siguiente



(f) La idea del método de penalización es reemplazar la función objetivo f por otra función

$$F(x,y) = f(x,y) + cP(x,y)$$

y resolver el problema sin restricciones. Para ello tomamos como c una constante positiva y P es una función que satisface las condiciones:

- P es continua en el dominio de f.
- $P(x,y) \geq 0$ para todo punto del domino de f, y
- $P(x,y) = 0$ si y solo si el punto (x,y) satisface las restricciones.

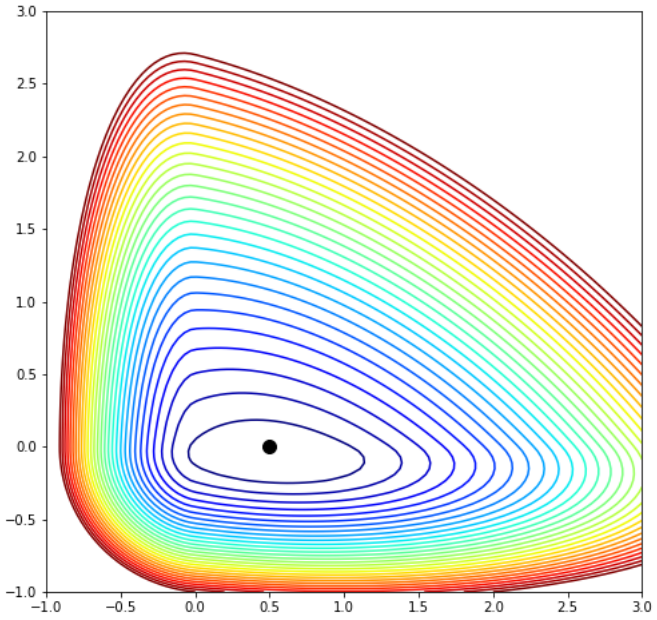
De acuerdo con ello, una función F podria ser

$$F(x,y) = f(x,y) + 10\phi_1(x,y) + 10\phi_2(x,y)$$

con

$$\phi_1(x,y) = \begin{cases} x^2 & \text{si} \quad x < 0 \\ 0 & \text{si} \quad x \geq 0 \end{cases} \qquad \phi_2(x,y) = \begin{cases} y^2 & \text{si} \quad y < 0 \\ 0 & \text{si} \quad y \geq 0 \end{cases}$$

Si representamos esta función F, vemos que es punto $(0{,}5, 0)$ está próximo al mínimo local de esta función.



6.5. Problemas lineales con restricciones lineales

Problema 6.13:

Minimizar la función $f(x,y) = -4x + 3y + 1$ sujeta a las restricciones.

$$\begin{cases} x + y & \leq 4, \\ 2x + y & \leq 5, \\ x, y & \geq 0. \end{cases}$$

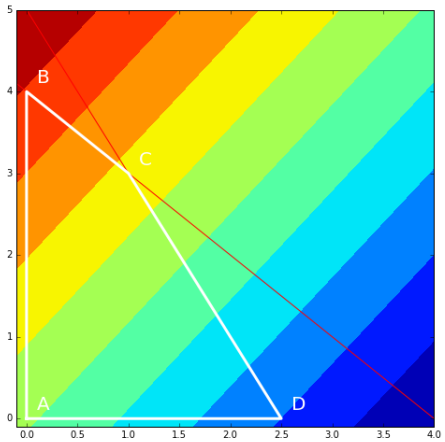
Las condiciones $x \geq 0$ e $y \geq 0$ significan que la región del plano que describen las condiciones está en el primer cuadrante. Dibujamos las demás fronteras de la región donde ha de encontrarse la solución. Tenemos dos rectas, r_1 y r_2

$$r_1 \qquad x + y = 4 \qquad \rightarrow \qquad \frac{x}{4} + \frac{y}{4} = 1$$

$$r_2 \qquad 2x + y = 5 \qquad \rightarrow \qquad \frac{x}{2{,}5} + \frac{y}{5} = 1$$

y entonces r_1 corta al eje OX en $x = 4$ y al eje OY en $y = 4$. Análogamente r_2 corta al eje OX en $x = 2{,}5$ y al eje OY en $y = 5$.

Cada una de estas rectas divide el plano en dos regiones, una que verifica la condición y otra que no. Basta por lo tanto con probar con un punto y si este punto cumple la condición, la región del plano donde se encuentra este punto es pertenece a la región. Por simplicidad, probamos con el origen $(0,0)$. El origen cumple la condición $x+y \leq 4$ porque $0+0 \leq 4$. Y también la condición $2x + y \leq 5$ porque $2(0) + 0 \leq 5$. Así que la región será la representada en la siguiente figura.



El mínimo estará en uno de los vértices del polígono blanco que son

Vértice	x	y	$f(x,y)$
A	0	0	1
B	0	4	13
C	1	3	6
D	**2.5**	**0**	**-9**

El mínimo está en D puesto que es el punto de la región donde la función toma el valor más pequeño.

Problema 6.14:

Minimizar la función $f(x,y) = 4x - 3y + 1$ sujeta a las restricciones.

$$\begin{cases} x + y & \leq 2, \\ x + 2y & \leq 3, \\ x, y & \geq 0. \end{cases}$$

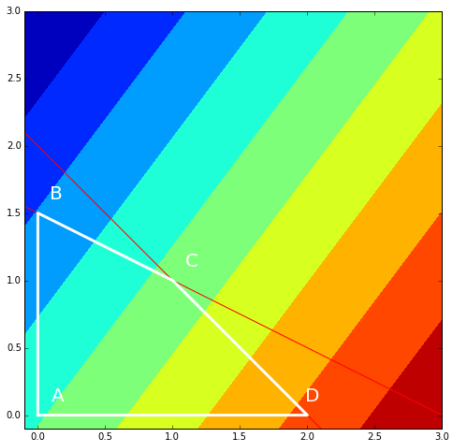
Las condiciones $x \geq 0$ e $y \geq 0$ significan que la región del plano que describen las condiciones está en el primer cuadrante. Dibujamos las demás fronteras de la región donde ha de encontrarse la solución. Tenemos dos rectas, r_1 y r_2

$$r_1 \qquad x + y = 2 \qquad \rightarrow \qquad \frac{x}{2} + \frac{y}{2} = 1$$

$$r_2 \qquad x + 2y = 3 \qquad \rightarrow \qquad \frac{x}{3} + \frac{y}{1{,}5} = 1$$

y entonces r_1 corta al eje OX en $x = 2$ y al eje OY en $y = 2$. Análogamente r_2 corta al eje OX en $x = 3$ y al eje OY en $y = 1{,}5$.

Cada una de estas rectas divide el plano en dos regiones, una que verifica la condición y otra que no. Basta por lo tanto con probar con un punto y si este punto cumple la condición, la región del plano donde se encuentra este punto es pertenece a la región. Por simplicidad, probamos con el origen $(0,0)$. El origen cumple la condición $x+y \leq 2$ porque $0+0 \leq 2$. Y también la condición $x + 2y \leq 3$ porque $0 + 2(0) \leq 3$. Así que la región será la representada en la siguiente figura.



El mínimo estará en uno de los vértices del polígono blanco que son

Vértice	x	y	$f(x,y)$
A	0	0	1
B	**0**	**1.5**	**-3.5**
C	1	1	2
D	2	0	9

El mínimo está en B puesto que es el punto de la región donde la función toma el valor más pequeño.

Problema 6.15:

Dos almacenes, A_1 y A_2, tienen en stock, respectivamente, 30 y 50 unidades de un cierto producto. Este se distribuye a tres ciudades C_1, C_2 y C_3 en cantidades de 25, 25 y 30 unidades, respectivamente. Las ganancias vienen dadas, en euros por unidad, en la tabla siguiente.

	C_1	C_2	C_3
A_1	30	35	20
A_2	20	40	30

Determinar cómo hay que distribuir las unidades para que las ganancias sean máximas.

Las unidades se distribuyen

Unidades	Hasta C_1	Hasta C_2	Hasta C_3	
Desde A_1	x	y	$30-x-y$	30
Desde A_2	$25-x$	$25-y$	$x+y$	50
	25	25	30	

Y teniendo en cuenta el cuadro del enunciado, las ganancias serán

$$G = 30x + 35y + 20(30 - x - y) + 20(25 - x) + 40(25 - y) + 30(x + y)$$

que simplificando queda

$$G = 2100 + 20x + 5y.$$

Por otro lado, si hacemos que todos los elementos del cuadro anterior sean mayores o iguales que cero

$x \geq 0$	$y \geq 0$	$30 - x - y \geq 0$
$25 - x \geq 0$	$25 - y \geq 0$	$x + y \geq 0$

llegaremos a que el problema a resolver es:

Maximizar la función $f(x, y) = 2100 + 20x + 5y$ con las restricciones

$$\begin{cases} x + y & \leq 30, \\ x + y & \geq 0, \\ x & \leq 25, \\ y & \leq 25, \\ x, y & \geq 0. \end{cases}$$

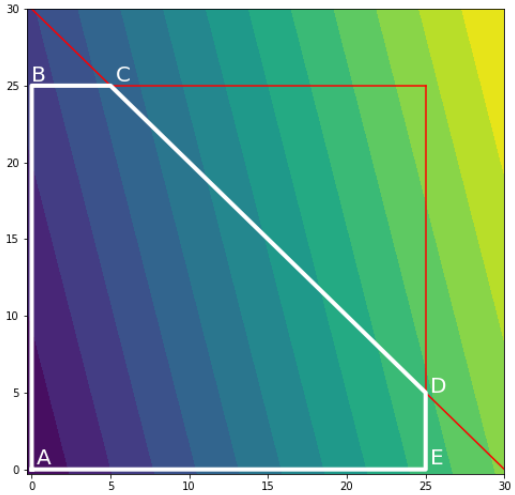
Las condiciones $x \geq 0$ e $y \geq 0$ significan que la región del plano que describen las condiciones está en el primer cuadrante. Dibujamos las demás fronteras de la región donde ha de encontrarse la solución. Tenemos cuatro rectas, r_1, r_2, r_3 y r_4.

$$r_1 \qquad x + y = 30 \qquad \rightarrow \qquad \frac{x}{30} + \frac{y}{30} = 1$$

$$r_2 \qquad x + y = 0 \qquad \rightarrow \qquad y = -x$$

y entonces r_1 corta al eje OX en $x = 30$ y al eje OY en $y = 30$. La recta r_2 pasa por $(0, 0)$ y tiene pendiente -1. La recta $r_3 : x = 25$ es la recta vertical que pasa por el punto $(25, 0)$ y la recta $r_4 : y = 25$ es la recta horizontal que pasa por el punto $(0, 25)$.

Cada una de estas rectas divide el plano en dos regiones, una que verifica la condición y otra que no. Basta por lo tanto con probar con un punto y si este punto cumple la condición, la región del plano donde se encuentra este punto es pertenece a la región. Por simplicidad, probamos con el origen $(0, 0)$. El origen cumple la condición $x + y \leq 30$ porque $0 + 0 \leq 30$. El origen también cumple las codiciones $x \leq 25$ y $y \leq 25$ Así que la región será la representada en la siguiente figura. La recta r_2 no aporta condiciones nuevas, porque si se cumple $x \geq 0$ e $y \geq 0$, se cumple $x + y \geq 0$. Es decir, todos los puntos del primer cuadrante, cumplen esta condición. La región por lo tanto, es la comprendida dentro del polígono blanco de la siguiente figura



El máximo estará en uno de los vértices del polígono blanco que son

Vértice	x	y	$f(x,y)$
A	0	0	2100
B	0	25	2225
C	5	25	2325
D	**25**	**5**	**2625**
E	25	0	2600

El máximo está en D, $x = 25$ $y = 5$ puesto que es el punto de la región donde la función toma el valor mayor. Y finalmente, las unidades se distribuyen

Unidades	Hasta C_1	Hasta C_2	Hasta C_3	
Desde A_1	25	5	0	30
Desde A_2	0	20	30	50
	25	25	30	

Problema 6.16:

Dos almacenes, A_1 y A_2, tienen en stock, respectivamente, 35 y 50 unidades de un cierto producto. Este se distribuye a tres ciudades C_1, C_2 y C_3 en cantidades de 30, 30 y 25 unidades, respectivamente. Las ganancias unitarias vienen dadas, en euros por unidad, en la tabla siguiente.

	C_1	C_2	C_3
A_1	35	45	70
A_2	20	25	35

Determinar cómo hay que distribuir las unidades para que las ganancias sean máximas.

Las unidades se distribuyen

Unidades	Hasta C_1	Hasta C_2	Hasta C_3	
Desde A_1	x	y	$35-x-y$	35
Desde A_2	$30-x$	$30-y$	$x+y-10$	50
	30	30	25	

Y teniendo en cuenta el cuadro del enunciado, las ganancias serán

$$G = 35x + 45y + 70(35 - x - y) + 20(30 - x) + 25(30 - y) + 35(x + y - 10)$$

que simplificando queda

$$G = 3450 - 20x - 15y.$$

Por otro lado, si hacemos que todos los elementos del cuadro anterior sean mayores o iguales que cero

$x \geq 0$	$y \geq 0$	$35 - x - y \geq 0$
$30 - x \geq 0$	$30 - y \geq 0$	$x + y - 10 \geq 0$

llegaremos a que el problema a resolver es:

Maximizar la función $f(x, y) = 3450 - 20x - 15y$ con las restricciones

$$\begin{cases} x + y & \leq 35, \\ x + y & \geq 10, \\ x & \leq 30, \\ y & \leq 30, \\ x, y & \geq 0. \end{cases}$$

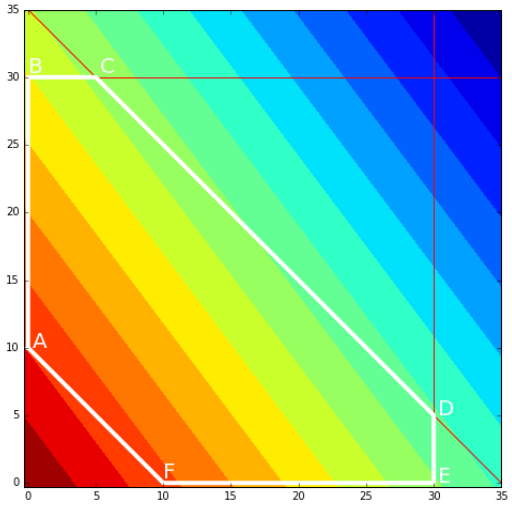
Las condiciones $x \geq 0$ e $y \geq 0$ significan que la región del plano que describen las condiciones está en el primer cuadrante. Dibujamos las demás fronteras de la región donde ha de encontrarse la solución. Tenemos cuatro rectas, r_1, r_2, r_3 y r_4.

$$r_1 \qquad x + y = 35 \qquad \rightarrow \qquad \frac{x}{35} + \frac{y}{35} = 1$$

$$r_2 \qquad x + y = 10 \qquad \rightarrow \qquad \frac{x}{10} + \frac{y}{10} = 1$$

y entonces r_1 corta al eje OX en $x = 35$ y al eje OY en $y = 35$. Análogamente r_2 corta al eje OX en $x = 10$ y al eje OY en $y = 10$. La recta $r_3 : x = 30$ es la recta vertical que pasa por el punto $(30, 0)$ y la recta $r_4 : y = 30$ es la recta horizontal que pasa por el punto $(0, 30)$.

Cada una de estas rectas divide el plano en dos regiones, una que verifica la condición y otra que no. Basta por lo tanto con probar con un punto y si este punto cumple la condición, la región del plano donde se encuentra este punto es pertenece a la región. Por simplicidad, probamos con el origen $(0, 0)$. El origen cumple la condición $x + y \leq 35$ porque $0 + 0 \leq 35$. Y pero no cumple la condición $x + y \geq 10$ porque $0 + 0 \geq 10$. El origen también cumple las codiciones $x \leq 30$ y $y \leq 30$ Así que la región será la representada en la siguiente figura.



El máximo estará en uno de los vértices del polígono blanco que son

Vértice	x	y	$f(x,y)$
A	**0**	**10**	**3450**
B	0	30	3000
C	5	30	2900
D	30	5	2775
E	30	0	2850
F	10	0	3250

El máximo está en A, $x = 0$ $y = 10$ puesto que es el punto de la región donde la función toma el valor mayor.

Y finalmente, las unidades se distribuyen

Unidades	Hasta C_1	Hasta C_2	Hasta C_3	
Desde A_1	0	10	25	35
Desde A_2	30	20	0	50
	30	30	25	