

Curso de Computación Numérica



**Fernando Sánchez Lasheras
Esperanza García Gonzalo**

Curso de Computación Numérica

Fernando Sánchez Lasheras

Departamento de Matemáticas. Universidad de Oviedo

Esperanza García Gonzalo

Departamento de Matemáticas. Universidad de Oviedo

CURSO DE COMPUTACIÓN NUMÉRICA.

© Fernando Sánchez Lasheras

© Esperanza García Gonzalo

Diseño de la cubierta: Javier Arduengo García

Imprime: HiFer A.G., Oviedo. www.hifer.com

I.S.B.N.:978-84-18289-57-6

Dep. Legal: AS-00024-2021



www.elsastredeloslibros.es

No está permitida la reproducción total o parcial de este libro, ni su tratamiento informático, ni la transmisión de ninguna forma o por cualquier medio, ya sea electrónico, mecánico, por fotocopia, por registro u otros métodos, ni su préstamo o alquiler o cualquiera otra forma de cesión de uso del ejemplar, sin permiso previo y por escrito del titular del Copyright.

© El Copyright y todos los demás derechos son propiedad del autor y está debidamente registrado en el Registro General de la Propiedad Intelectual de Asturias.

A mis padres, con todo mi amor.
Fernando

Prólogo

El presente libro expone problemas relativos a los contenidos que se imparten en la asignatura de Computación Numérica perteneciente al grado en Ingeniería Informática del Software de la Universidad de Oviedo. Este trabajo surge de la experiencia adquirida en la docencia de esta asignatura y tiene como objetivo ofrecer a los alumnos de este grado materiales para su estudio de la asignatura. La obra se estructura en seis capítulos: aritmética finita y teoría del error, raíces de ecuaciones lineales, aproximación de funciones, derivación e integración numérica, sistemas de ecuaciones lineales y optimización.

A la hora de explicar los problemas que se proponen, estos se relacionan con los contenidos teóricos que permiten su resolución. Por tanto, a través de este libro el estudiante no solo aprende a resolver problemas, sino que también afianza los conocimientos teóricos estudiados.

Finalmente, cabe mencionar que el Análisis Numérico es una de las ramas de las Matemáticas que más se han beneficiado de los progresos conseguidos en las últimas décadas en la Informática. La resolución de problemas matemáticos con la ayuda de un ordenador es una realidad a la que los graduados en Ingeniería Informática del Software se tendrán que enfrentar a lo largo de su carrera profesional. En opinión de los autores, para estos profesionales resulta imprescindible una buena base matemática.

Oviedo, enero de 2021

Los autores

Índice

| | |
|--|----------|
| 1 ARITMÉTICA FINITA Y ANÁLISIS DE ERROR | 1 |
| 1.1 Paso del sistema binario a decimal | |
| Paso del sistema decimal a binario | 1 |
| 1.1.1 Ejercicio | 1 |
| 1.2 La norma IEEE 754 | |
| Formato normalizado | |
| Representación del exponente | |
| Binary32 (precisión simple) | |
| Binary64 (precisión doble) | 5 |
| 1.2.1 Ejercicio | 5 |
| 1.2.2 Ejercicio | 11 |
| 1.3 Números normalizados | |
| Números desnormalizados | |
| El epsilon de máquina | |
| Errores de redondeo | |
| Máximo entero | |
| Valores especiales | 14 |
| 1.3.1 Ejercicio | 14 |
| 1.3.2 Ejercicio | 26 |
| 1.4 Error absoluto y relativo | 35 |
| 1.4.1 Ejercicio | 35 |
| 1.5 Redondeo con la norma IEEE 754 | 37 |
| 1.5.1 Ejercicio | 37 |
| 1.5.2 Ejercicio | 44 |

| | |
|--|-----------|
| 2 RAÍCES DE ECUACIONES NO LINEALES | 49 |
| 2.1 Raíz de una función | |
| Separación de raíces | |
| Métodos numéricos de cálculo de raíces | |
| Método de bisección | 49 |
| 2.1.1 Ejercicio: cálculo de raíces | 49 |
| 2.1.2 Ejercicio: cálculo de extremos | 59 |
| 2.2 Método de Newton-Raphson | 64 |
| 2.2.1 Ejercicio: cálculo de raíces | 64 |
| 2.2.2 Ejercicio: cálculo de extremos | 67 |
| 2.3 Método de la secante | 71 |
| 2.3.1 Ejercicio | 71 |
| 2.4 Método de Regula-Falsi | 75 |
| 2.4.1 Ejercicio | 75 |
| 2.5 Método de punto fijo | 79 |
| 2.5.1 Ejercicio | 79 |

| | |
|---|-----------|
| 3 APROXIMACIÓN DE FUNCIONES | 89 |
| 3.1 Polinomios de interpolación de Lagrange | 89 |
| 3.1.1 Ejercicio: polinomios fundamentales de Lagrange | 89 |
| 3.1.2 Ejercicio: diferencias divididas. Forma de Newton | 101 |
| 3.2 Interpolación polinómica a trozos | 105 |
| 3.2.1 Ejercicio: interpolación a trozos lineal | 105 |
| 3.2.2 Ejercicio: interpolación con splines cúbicos | 110 |
| 3.3 Recta de regresión por mínimos cuadrados | 115 |
| 3.3.1 Ejercicio | 115 |
| 3.4 Ajuste de funciones previa linearización | 123 |
| 3.4.1 Ejercicio: función exponencial | 123 |
| 3.4.2 Ejercicio: función potencial | 127 |
| 3.4.3 Ejercicio: hipérbola | 128 |
| 3.4.4 Ejercicio: función logística | 130 |
| 3.5 Ajuste de funciones con polinomios | 133 |
| 3.5.1 Ejercicio: caso continuo, polinomios ortogonales | 133 |
| 3.5.2 Ejercicio: caso discreto | 138 |

| | |
|---|------------|
| 4 DERIVACIÓN E INTEGRACIÓN NUMÉRICA | 143 |
| 4.1 Fórmulas en diferencias finitas | 143 |
| 4.1.1 Ejercicio: fórmulas progresiva, regresiva y centrada . | 143 |
| 4.2 Fórmulas interpolatorias de derivación numérica | 151 |
| 4.2.1 Ejercicio: Fórmulas de orden dos | 151 |
| 4.3 Derivación numérica de funciones de dos variables | 158 |
| 4.3.1 Ejercicio: gradiente y laplaciano | 158 |
| 4.4 Fórmulas de cuadratura | |
| Fórmulas de Newton-Cotes | 160 |
| 4.4.1 Ejercicio: fórmulas del trapecio y el punto medio . | 160 |
| 4.4.2 Ejercicio: aplicación fórmulas simples | 166 |
| 4.4.3 Ejercicio: aplicación fórmulas compuestas | 169 |
| 4.4.4 Ejercicio: error regla trapecio | 173 |
| 4.4.5 Ejercicio: error regla trapecio | 177 |
| 4.5 Fórmulas de cuadratura gaussianas | 180 |
| 4.5.1 Ejercicio | 180 |

| | |
|---|------------|
| 5 SISTEMAS DE ECUACIONES LINEALES | 185 |
| 5.1 Método de Gauss | 185 |
| 5.1.1 Ejercicio: sin pivote | 185 |
| 5.1.2 Ejercicio: con pivote parcial | 189 |
| 5.2 Método de Gauss-Jordan | 191 |
| 5.2.1 Ejercicio | 191 |
| 5.3 Factorización LU | 194 |
| 5.3.1 Ejercicio: sin pivote | 194 |
| 5.3.2 Ejercicio: con pivote parcial | 200 |
| 5.4 Determinantes | 203 |
| 5.4.1 Ejercicio | 203 |
| 5.5 Método de Jacobi | 205 |
| 5.5.1 Ejercicio: convergencia | 205 |
| 5.5.2 Ejercicio: algoritmo | 211 |
| 5.6 Método de Gauss-Seidel | 215 |
| 5.6.1 Ejercicio: convergencia | 215 |
| 5.6.2 Ejercicio: algoritmo | 222 |
| 5.7 Condicionamiento de matrices | 226 |
| 5.7.1 Ejercicio | 226 |

| | | |
|----------|--|------------|
| 6 | OPTIMIZACIÓN | 229 |
| 6.1 | Mínimo local y global. Cálculo simbólico | 229 |
| 6.1.1 | Ejercicio: mínimo de una función de tres variables . . | 229 |
| 6.1.2 | Ejercicio: mínima distancia a 4 puntos | 239 |
| 6.2 | Mínimo local. Cálculo numérico. Métodos del gradiente, de Newton y de la sección áurea | 241 |
| 6.2.1 | Ejercicio | 241 |
| 6.3 | Mínimo con restricciones de igualdad. Cálculo simbólico: multiplicadores de Lagrange. Cálculo numérico: método de penalización | 251 |
| 6.3.1 | Ejercicio | 251 |
| 6.3.2 | Ejercicio | 254 |
| 6.4 | Mínimo local con restricciones de desigualdad. Cálculo simbólico y numérico | 256 |
| 6.4.1 | Ejercicio | 256 |
| 6.4.2 | Ejercicio | 260 |
| 6.5 | Extremo de función lineal con restricciones lineales | 265 |
| 6.5.1 | Ejercicio | 265 |
| 6.5.2 | Ejercicio: el problema del transporte | 267 |

TEMA 1

ARITMÉTICA FINITA Y ANÁLISIS DE ERROR

1.1 Paso del sistema binario a decimal

Paso del sistema decimal a binario

Ejercicio 1.1.1

Cambio de base:

1. Dado el número 101100,001 en base 2 calcular su representación decimal.
2. Dado el número 51 en base 10 calcular su representación binaria.
3. Dado el número 51,65625 en base 10 calcular su representación binaria.

INTRODUCCIÓN

En este curso presentaremos los métodos numéricos básicos que resuelven un conjunto de problemas matemáticos clásicos. Los ordenadores son una herramienta necesaria en el uso eficiente de los métodos numéricos. Por lo tanto, veremos cómo los números, que pueden tener infinitos dígitos, se almacenan en el ordenador, que es un dispositivo finito.

Los números se almacenan en los ordenadores con los formatos:

- *Números enteros*
- *Números en coma flotante*

Los números enteros se almacenan de forma exacta como números enteros y por lo tanto no hay error (salvo error de desbordamiento) y no los vamos a estudiar. Únicamente veremos el almacenamiento de enteros sin signo porque es la forma más sencilla y el almacenamiento sesgado de los

enteros con signo porque esta última es la forma en que se almacenan los exponentes de los números en coma flotante.

Estudiaremos el almacenamiento de números en binario. Por lo tanto, el primer paso que realizaremos será repasar cómo convertir un número de decimal a binario y de binario a decimal.

Paso del sistema binario a decimal

En el sistema decimal el número 107,625 significa:

$$107,625 = 1 \cdot 10^2 + 7 \cdot 10^0 + 6 \cdot 10^{-1} + 2 \cdot 10^{-2} + 5 \cdot 10^{-3}.$$

En general, los ordenadores usan el sistema binario: solo se almacenan 0 y 1. En el sistema binario, los números representan potencias de 2. El paso de binario a decimal es directo.

$$(1101011,101)_2$$

Necesitamos conocer la posición de cada dígito respecto de la coma

$$\begin{array}{ccccccccccccc} (6) & (5) & (4) & (3) & (2) & (1) & (0) & & (-1) & (-2) & (-3) \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 & , & 1 & 0 & 1 \end{array}$$

que expresa el número

$$1 \cdot 2^6 + 1 \cdot 2^5 + 0 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3}$$

que es

$$(107,625)_{10}$$

Paso del sistema decimal a binario

Parte entera

Dividimos sucesivamente por 2 hasta que el cociente es cero y los restos son los dígitos en base 2.

| Dividendo | Divisor | Cociente | Resto | |
|-----------|---------|----------|-------|---|
| 107 | 2 | 53 | 1 | ↑ |
| 53 | 2 | 26 | 1 | ↑ |
| 26 | 2 | 13 | 0 | ↑ |
| 13 | 2 | 6 | 1 | ↑ |
| 6 | 2 | 3 | 0 | ↑ |
| 3 | 2 | 1 | 1 | ↑ |
| 1 | 2 | 0 | 1 | ↑ |

El número empieza por el último resto y el número en binario es 1101011.

Parte fraccionaria

Multiplicamos por 2, restamos la parte entera y repetimos hasta que la parte fraccionaria sea cero.

$$\begin{array}{rcl} 0,625 & \times & 2 = 1,25 \rightarrow 1 \downarrow \\ 0,25 & \times & 2 = 0,5 \rightarrow 0 \downarrow \\ 0,5 & \times & 2 = 1,0 \rightarrow 1 \downarrow \end{array}$$

Empezamos por el la primera parte entera y el número en binario es 0.101

EJERCICIO**1. Dado el número 101100,001 en base 2 calcular su representación decimal**

Necesitamos tener en cuenta la posición de cada dígito respecto a la coma

$$\begin{array}{cccccccccc} (5) & (4) & (3) & (2) & (1) & (0) & & (-1) & (-2) & (-3) \\ 1 & 0 & 1 & 1 & 0 & 0 & , & 0 & 0 & 1 \end{array}$$

Y entonces, el valor de este número en base 10 es

$$\begin{aligned} 1 \cdot 2^5 + 0 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 0 \cdot 2^0 + 0 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} = \\ = 2^5 + 2^3 + 2^2 + 2^{-3} = \boxed{44,125} \end{aligned}$$

$$(101100,001)_2 = (44,125)_{10}$$

2. Dado el número 51 en base 10 calcular su representación binaria

Dividimos sucesivamente por 2 hasta que el cociente es cero y los restos son los dígitos en base 2.

| Dividendo | Divisor | Cociente | Resto | |
|-----------|---------|----------|-------|---|
| 51 | 2 | 25 | 1 | ↑ |
| 25 | 2 | 12 | 1 | ↑ |
| 12 | 2 | 6 | 0 | ↑ |
| 6 | 2 | 3 | 0 | ↑ |
| 3 | 2 | 1 | 1 | ↑ |
| 1 | 2 | 0 | 1 | ↑ |

El número empieza por el último resto y el número en binario es 110011.

$$(51)_{10} = (110011)_2$$

3. Dado el número 51,65625 en base 10 calcular su representación binaria.

La parte entera ya está convertida en el apartado anterior. Convirtamos ahora su parte fraccionaria.

Multiplicamos por 2, restamos la parte entera y repetimos hasta que la parte fraccionaria sea cero.

$$\begin{array}{rcl} 0,65625 & \times & 2 = 1,3125 \rightarrow 1 \downarrow \\ 0,3125 & \times & 2 = 0,625 \rightarrow 0 \downarrow \\ 0,625 & \times & 2 = 1,25 \rightarrow 1 \downarrow \\ 0,25 & \times & 2 = 0,5 \rightarrow 0 \downarrow \\ 0,5 & \times & 2 = 1,0 \rightarrow 1 \downarrow \end{array}$$

Empezamos por el la primera parte entera y el número en binario es 0,10101

Y teniendo en cuenta la parte entera

$$(51,65625)_{10} = (110011,10101)_2$$

1.2 La norma IEEE 754

Formato normalizado

Representación del exponente

Binary32 (precisión simple)

Binary64 (precisión doble)

Ejercicio 1.2.1

Si el número

| Signo 1 bit | Exponente 8 bits | Mantisa 23 bits |
|----------------|---------------------|------------------------------|
| 1 | 1000 1101 | 0110 1000 0000 0000 0000 000 |

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

INTRODUCCIÓN

La norma IEEE 754

El formato en coma flotante usado habitualmente es el formato IEEE 754. IEEE significa Institute of Electrical and Electronic Engineers. Este estándar es el usado actualmente por casi todos los procesadores.

La primera norma IEEE 754 se publicó en 1985 e incluía únicamente la representación de números en binario. Sus formatos básicos eran simple y doble precisión. En el 2008 se publicó una segunda versión donde se incluía también la representación de números en decimal con dos formatos básicos y se añadía un formato básico binario con precisión cuádruple. En el 2019 se publicó una tercera versión con modificaciones menores.

Los cinco formatos básicos y sus parámetros más importantes son

| parámetro | Formatos binarios ($b = 2$) | Formatos decimales ($b = 10$) | | | |
|------------------|-------------------------------|---------------------------------|---------------|------------|-------------|
| precisión(p) | binary32 | binary64 | binary128 | decimal64 | decimal128 |
| e_{max} | 24 +127 | 53 +1023 | 113 +16383 | 16 +384 | 34 +6144 |

La *precisión* es el número de dígitos de la mantisa y e_{max} es el exponente máximo del formato. El formato por defecto en computación numérica es actualmente el binario en doble precisión. Vamos a estudiar en más detalle los formatos binarios en simple y doble precisión.

Formato normalizado

Número decimal

Sea el número decimal 314,15. Para normalizarlo

1. Movemos la coma de forma que aparezca un único dígito distinto de cero a su izquierda.
2. Al mover la coma tenemos que multiplicar por 10^n siendo n el número de posiciones que hemos movido la coma a la izquierda o 10^{-n} siendo n el número de posiciones que hemos movido la coma a la derecha.
3. Le añadimos el signo.

El número anterior, normalizado sería

$$+3,1415 \times 10^2$$

cuyos elementos son

- Signo: +
- Mantisa: 3,1415
- Exponente: 2

Además tenemos que la base es 10.

Número binario

Sea el número binario 10101,11001. Para normalizarlo aplicaríamos los mismos pasos que en el caso anterior pero teniendo en cuenta que la base ahora es 2 y por lo tanto, este número normalizado sería

$$+1,010111001 \times 2^4$$

con

- Signo: +
- Mantisa: 1,010111001
- Exponente: 4 (tendríamos que expresarlo en binario)

Además tenemos que la base es 2.

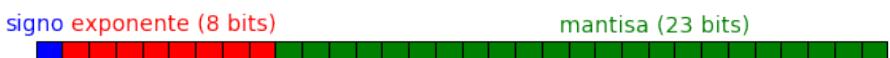
Representación del exponente

El exponente en esta norma será siempre un número entero con representación *sesgada*. Veamos cómo funciona con un ejemplo. Supongamos que tenemos $m = 4$ bits donde podemos almacenar 4 dígitos binarios. Podríamos almacenar $2^m = 2^4 = 16$ combinaciones diferentes con las que podríamos representar los siguientes números

| Representación binaria ($m = 4$ bits) | Valor nominal | Enteros con signo (Expo.) $sesgo = 2^{m-1} - 1$ |
|---|---------------|--|
| 0000 | 0 | Reservado |
| 0001 | 1 | -6 |
| 0010 | 2 | -5 |
| 0011 | 3 | -4 |
| 0100 | 4 | -3 |
| 0101 | 5 | -2 |
| 0110 | 6 | sesgo |
| 0111 | 7 | → 0 |
| 1000 | 8 | -7 |
| 1001 | 9 | 1 |
| 1010 | 10 | 2 |
| 1011 | 11 | 3 |
| 1100 | 12 | 4 |
| 1101 | 13 | 5 |
| 1110 | 14 | 6 |
| 1111 | 15 | 7 |
| | | Reservado |

Binary32 (precisión simple)

En el caso de precisión simple, se utilizan 32 bits (4 bytes) distribuidos de la forma siguiente:



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

Signo

Se utiliza un 0 si el signo es positivo y un 1 si el signo es negativo.

Exponente

Tenemos $m = 8$ bits para el exponente. Por lo tanto hay $2^m = 2^8 = 256$ combinaciones distintas y, en principio, podemos representar 256 números. Como empezamos el valor nominal en 0 acabará en 255. El primer número, 0000 0000, y el último, 1111 1111 se reservan (ya veremos luego para qué). Y como la representación es sesgada, tenemos que restar el

$$\text{sesgo} = 2^{m-1} - 1 = 2^{8-1} - 1 = 2^7 - 1 = 128 - 1 = 127$$

para obtener el valor representado.

| Número binario | Valor nominal | Valor representado |
|----------------|---------------|--------------------|
| 0000 0000 | 0 | R |
| 0000 0001 | 1 | -126 |
| 0000 0010 | 2 | -125 |
| 0000 0011 | 3 | -124 |
| ... | ... | -127 |
| ... | ... | → |
| 1111 1100 | 252 | 125 |
| 1111 1101 | 253 | 126 |
| 1111 1110 | 254 | 127 |
| 1111 1111 | 255 | R |

Por lo tanto el exponente mínimo representado es $e_{min} = -126$ y el máximo $e_{max} = 127$.

Mantisa

En binario, en la representación normalizada, el dígito a la izquierda de la coma es siempre 1, no se representa y se llama *el bit escondido*. Por lo tanto, aunque tenemos 23 dígitos binarios para almacenar estamos usando uno más y la precisión es 24.

Binary64 (precisión doble)

En el caso de doble precisión, se utilizan 64 bits (8 bytes) distribuidos de la forma siguiente:



- 1 bit para el signo.
- 11 bits para el exponente.
- 52 bits para la mantisa.

Signo

Se utiliza un 0 si el signo es positivo y un 1 si el signo es negativo.

Exponente

Tenemos $m = 11$ bits para el exponente. Por lo tanto hay $2^m = 2^{11} = 2048$ combinaciones distintas, y, en principio, podemos representar 2048 números. Como empezamos el valor nominal empieza en 0 acabará en 2047. El primer número, 0000 0000 000, y el último, 1111 1111 111 se reservan. Y como la representación es sesgada, tenemos que restar el

$$\text{sesgo} = 2^{m-1} - 1 = 2^{11-1} - 1 = 2^{10} - 1 = 1024 - 1 = 1023$$

para obtener el valor representado.

| Número binario | Valor nominal | Valor representado |
|----------------|---------------|--------------------|
| 0000 0000 000 | 0 | R |
| 0000 0000 001 | 1 | -1022 |
| 0000 0000 010 | 2 | -1021 |
| 0000 0000 011 | 3 | -1020 |
| ... | ... | -1023 |
| ... | ... | → ... |
| 1111 1111 100 | 2044 | 1021 |
| 1111 1111 101 | 2045 | 1022 |
| 1111 1111 110 | 2046 | 1023 |
| 1111 1111 111 | 2047 | R |

Por lo tanto el exponente mínimo representado es $e_{min} = -1022$ y el máximo $e_{max} = 1023$.

Mantisa

En binario, en la representación normalizada, el dígito a la izquierda de la coma es siempre 1, no se representa y se llama *el bit escondido*. Por lo tanto, aunque tenemos 52 dígitos binarios para almacenar estamos usando uno más y la precisión es 53.

EJERCICIO

| Signo 1 bit | Exponente 8 bits | Mantisa 23 bits |
|----------------|---------------------|------------------------------|
| 1 | 1000 1101 | 0110 1000 0000 0000 0000 000 |

Signo

Como es 1 → signo negativo

Exponente

El valor nominal del exponente 1000 1101 teniendo en cuenta la posición de cada dígito

$$\begin{array}{cccccccc} (7) & (6) & (5) & (4) & (3) & (2) & (1) & (0) \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{array}$$

es

$$2^7 + 2^3 + 2^2 + 2^0 = 128 + 8 + 4 + 1 = 141$$

Y si tenemos en cuenta que para $m = 8$ bits,

$$\text{sesgo} = 2^{m-1} - 1 = 2^{8-1} - 1 = 2^7 - 1 = 128 - 1 = 127$$

El valor del exponente es $141 - 127 = 14$.

| Número binario | Valor nominal | Valor representado |
|----------------|---------------|--------------------|
| 0000 0000 | 0 | R |
| 0000 0001 | 1 | -126 |
| 0000 0010 | 2 | -125 |
| 0000 0011 | 3 | -124 |
| ... | ... | -127 |
| ... | ... | → |
| 1000 1101 | 141 | 14 |
| ... | ... | ... |
| ... | ... | ... |
| 1111 1100 | 252 | 125 |
| 1111 1101 | 253 | 126 |
| 1111 1110 | 254 | 127 |
| 1111 1111 | 255 | R |

Mantisa

Los dígitos almacenados de la mantisa son $0110\ 1000\dots 0000$ y teniendo en cuenta el bit escondido, que es uno, la mantisa es

$$1,0110\ 1000\dots 0000$$

o también

$$1,0110\ 1 \quad \begin{matrix} (0) & (-1) & (-2) & (-3) & (-4) & (-5) \\ 1 & , & 0 & 1 & 1 & 0 & 1 \end{matrix}$$

Número

Por lo tanto, si escribimos el número este es

$$-1,0110\ 1 \times 2^{14} \longrightarrow -(1 + 2^{-2} + 2^{-3} + 2^{-5}) \times 2^{14} = \boxed{-23040}$$

Ejercicio 1.2.2

¿Cómo se almacenaría en precisión simple según la norma IEEE 754 el número 120,875?

Paso del sistema decimal a binario*Parte entera*

Dividimos sucesivamente por 2 hasta que el cociente es cero y los restos son los dígitos en base 2.

| Dividendo | Divisor | Cociente | Resto | |
|-----------|---------|----------|-------|---|
| 120 | 2 | 60 | 0 | ↑ |
| 60 | 2 | 30 | 0 | ↑ |
| 30 | 2 | 15 | 0 | ↑ |
| 15 | 2 | 7 | 1 | ↑ |
| 7 | 2 | 3 | 1 | ↑ |
| 3 | 2 | 1 | 1 | ↑ |
| 1 | 2 | 0 | 1 | ↑ |

El número empieza por el último resto y el número en binario es 1111000.

Parte fraccionaria

Multiplicamos por 2, restamos la parte entera y repetimos hasta que la parte fraccionaria sea cero.

$$\begin{array}{r}
 0,875 \times 2 = 1,750 \rightarrow 1 \downarrow \\
 0,75 \times 2 = 1,5 \rightarrow 1 \downarrow \\
 0,5 \times 2 = 1,0 \rightarrow 1 \downarrow
 \end{array}$$

Empezamos por el la primera parte entera y el número en binario es 0,111

Y el número completo es

$$(120,875)_{10} = (1111000,111)_2$$

Normalización

1. Movemos la coma de forma que aparezca un único dígito distinto de cero a su izquierda.
2. Al mover la coma tenemos que multiplicar por 2^n siendo n el número de posiciones que hemos movido la coma a la izquierda o 2^{-n} siendo n el número de posiciones que hemos movido la coma a la derecha.
3. Le añadimos el signo.

Este número normalizado sería

$$+1,1110\ 0011\ 1 \times 2^6$$

con

- Signo: +
- Mantisa: 1,1110 0011 1
- Exponente: 6

Signo

Como el signo es positivo \rightarrow signo 0

Exponente

Tenemos $m = 8$ bits para el exponente. Por lo tanto hay $2^m = 2^8 = 256$ combinaciones distintas y, en principio, podemos representar 512 números. Como empezamos el valor nominal empieza en 0 acabará en 255. El primer número, 0000 0000, y el último, 1111 1111 se reservan (ya veremos luego para qué). Y como la representación es sesgada, tenemos que restar el

$$\text{sesgo} = 2^{m-1} - 1 = 128 - 1 = 127$$

para obtener el valor representado.

El valor del exponente es 6. Para obtener el valor nominal tenemos que sumarle el sesgo y $6 + 127 = 133$ que en binario sería

| Dividendo | Divisor | Cociente | Resto | |
|-----------|---------|----------|-------|---|
| 133 | 2 | 66 | 1 | ↑ |
| 66 | 2 | 33 | 0 | ↑ |
| 33 | 2 | 16 | 1 | ↑ |
| 16 | 2 | 8 | 0 | ↑ |
| 8 | 2 | 4 | 0 | ↑ |
| 4 | 2 | 2 | 0 | ↑ |
| 2 | 2 | 1 | 0 | ↑ |
| 1 | 2 | 0 | 1 | ↑ |

es decir

$$(133)_{10} = (1000\ 0101)_2$$

| Número binario | Valor nominal | Valor representado |
|------------------|---------------|--------------------|
| 0000 0000 | 0 | R |
| 0000 0001 | 1 | -126 |
| 0000 0010 | 2 | -125 |
| 0000 0011 | 3 | -124 |
| ... | ... | -127 |
| ... | ... | → |
| 1000 0101 | 133 | 6 |
| ... | ... | +127 |
| ... | ... | ← |
| 1111 1100 | 252 | 126 |
| 1111 1101 | 253 | 126 |
| 1111 1110 | 254 | 127 |
| 1111 1111 | 255 | R |

Mantisa

$$(120,875)_{10}$$

Vimos que la mantisa era

$$1, \textcolor{teal}{1110\ 0011\ 1}$$

Hemos de tener en cuenta el bit escondido, que no almacenamos, y que rellenamos con ceros por la derecha hasta tener 23 bits.

Número

El número 120,875 en precisión sencilla se almacena

| signo | exponente | mantisa |
|-------|------------------|-------------------------------------|
| 0 | 1000 0101 | 1110 0011 1000 0000 0000 000 |

1.3 Números normalizados

Números desnormalizados

El epsilon de máquina

Errores de redondeo

Máximo entero

Valores especiales

Ejercicio 1.3.1

Sea el formato de 6 bits que funciona de forma similar al descrito por la norma IEEE 754. El primer bit corresponde al signo, los tres siguientes al exponente y los dos últimos a la mantisa.

| signo | exponente | mantisa |
|-------|--|-------------------------------|
| s | e ₁ e ₂ e ₃ | m ₁ m ₂ |

- Calcular los exponentes máximo y mínimo.
- Determinar los números positivos normalizados representables y dibujarlos sobre la recta real. ¿Cuántos números positivos normalizados se pueden representar con este formato?
- Escribir todos los números desnormalizados positivos en este sistema. ¿Cuántos números positivos desnormalizados se pueden representar con este formato?
- Calcular el ϵ de máquina.
- Dibujar los errores absolutos y relativos cometidos al almacenar los números utilizando el redondeo *al par más cercano* y el redondeo *hacia el cero*.
- Calcular el máximo entero representable de forma exacta de forma que el siguiente entero no es representable de forma exacta.
- ¿Cómo se representaría el cero? ¿E infinito? ¿Y NaN?

1. Exponente máximo y mínimo

Recordemos que el formato es

| signo | exponente | mantisa |
|-------|--|---------------------------------|
| s | e ₁ e ₂ e ₃ | 1,m ₁ m ₂ |

Tenemos $m = 3$ bits para el exponente. Por lo tanto hay $2^m = 2^3 = 8$ combinaciones distintas y, en principio, podemos representar 8 números. Como empezamos el valor nominal empieza en 0 acabará en 7. El primer

número, 000, y el último, 111 se reservan. Y como la representación es sesgada, tenemos que restar el

$$\text{sesgo} = 2^{m-1} - 1 = 2^{3-1} - 1 = 2^2 - 1 = 4 - 1 = 3$$

para obtener el valor representado

| Número binario | Valor nominal | Valor representado |
|----------------|---------------|--------------------|
| 000 | 0 | R |
| 001 | 1 | -2 |
| 010 | 2 | -1 |
| 011 | 3 | sesgo |
| 100 | 4 | 0 |
| 101 | 5 | -3 |
| 110 | 6 | 1 |
| 111 | 7 | 2 |
| | | 3 |
| | | R |

Por lo tanto, en este estándar, el exponente mínimo es $e_{\min} = -2$ y el máximo es $e_{\max} = 3$.

2. Números positivos normalizados

Recordemos que el formato es

| signo | exponente | mantisa |
|-------|---------------|--------------|
| s | $e_1 e_2 e_3$ | $1, m_1 m_2$ |

- Como los números son positivos, el primer bit, el del signo, es siempre cero.
- Los tres bits siguientes son los del exponente.
- Si el número de bits de la mantisa es $n = 2$, escribimos cada exponente (del apartado anterior) con las $2^n = 2^2 = 4$ mantisas posibles. Recordemos que el bit escondido no se almacena y es siempre 1 para los números normalizados.

| Num(bin) | Num(dec) | distancia |
|----------|----------------------|-------------------|
| 0 001 00 | $1,00 \times 2^{-2}$ | 0,25 |
| 0 001 01 | $1,01 \times 2^{-2}$ | 0,3125 |
| 0 001 10 | $1,10 \times 2^{-2}$ | 0,375 |
| 0 001 11 | $1,11 \times 2^{-2}$ | 0,4375 |
| | | $2^{-4} = 0,0625$ |
| 0 010 00 | $1,00 \times 2^{-1}$ | 0,5 |
| 0 010 01 | $1,01 \times 2^{-1}$ | 0,625 |
| 0 010 10 | $1,10 \times 2^{-1}$ | 0,75 |
| 0 010 11 | $1,11 \times 2^{-1}$ | 0,875 |
| | | $2^{-3} = 0,125$ |
| 0 011 00 | $1,00 \times 2^0$ | 1 |
| 0 011 01 | $1,01 \times 2^0$ | 1,25 |
| 0 011 10 | $1,10 \times 2^0$ | 1,5 |
| 0 011 11 | $1,11 \times 2^0$ | 1,75 |
| | | $2^{-2} = 0,25$ |
| 0 100 00 | $1,00 \times 2^1$ | 2 |
| 0 100 01 | $1,01 \times 2^1$ | 2,5 |
| 0 100 10 | $1,10 \times 2^1$ | 3 |
| 0 100 11 | $1,11 \times 2^1$ | 3,5 |
| | | $2^{-1} = 0,5$ |
| 0 101 00 | $1,00 \times 2^2$ | 4 |
| 0 101 01 | $1,01 \times 2^2$ | 5 |
| 0 101 10 | $1,10 \times 2^2$ | 6 |
| 0 101 11 | $1,11 \times 2^2$ | 7 |
| | | $2^0 = 1$ |
| 0 110 00 | $1,00 \times 2^3$ | 8 |
| 0 110 01 | $1,01 \times 2^3$ | 10 |
| 0 110 10 | $1,10 \times 2^3$ | 12 |
| 0 110 11 | $1,11 \times 2^3$ | 14 |
| | | $2^1 = 2$ |

Para $m = 3$ bits de exponente tenemos 2^m números distintos pero como el primero 000 y el último 111 están reservados tenemos

$$2^m - 2 = 2^3 - 2 = 8 - 2 = 6 \text{ exponentes}$$

Para $n = 2$ bits de mantisa tenemos 2^n números distintos. El bit escondido es siempre 1, así que no aporta nuevos valores y tenemos

$$2^n = 2^2 = 4 \text{ mantisas}$$

y en total

$$6 \text{ exponentes} \times 4 \text{ mantisas} = \boxed{24 \text{ números normalizados positivos}}$$

Si representamos estos números sobre la recta real



Vemos que:

- El espacio entre números aumenta cuando nos desplazamos a la derecha. De hecho, cada vez que cambiamos de exponente, el espacio entre números se duplica.
- Hay un espacio llamativo entre el número normalizado más pequeño y el cero.
- Hemos señalado en verde el número 1 y el siguiente número representable en este estándar.
- El mínimo número normalizado a representar de forma exacta es 0,25 y el máximo es 14.

3. Números desnormalizados

Recordemos que el formato normalizado es

| signo | exponente | mantisa |
|-------|---------------|--------------|
| s | $e_1 e_2 e_3$ | $1, m_1 m_2$ |

y los números desnormalizados son

| signo | exponente | mantisa |
|-------|-----------|--------------|
| s | 000 | $0, m_1 m_2$ |

Números desnormalizados en precisión simple

Si consideramos que el bit escondido es cero, podemos representar números menores que el mínimo número normalizado.

Estos números, en la norma IEEE 754, se representan con exponente 0000 0000 en precisión simple y con exponente 0000 0000 000 en precisión doble. Pero se interpreta que el valor de su exponente es el exponente mínimo, es decir -126 en precisión sencilla y -1022 en doble precisión.

- El inconveniente de estos números es que su precisión es menor que 24 en precisión simple y menor que 53 en precisión doble.
- Y su ventaja es que aumentan el rango de números a representar llenando el espacio entre el menor número normalizado y el cero.

Veamos un ejemplo

El número siguiente está representado en precisión simple según la norma IEEE 754.

| signo | exponente | mantisa |
|-------|-----------|-------------------------------|
| 0 | 0000 0000 | 0001 0110 0000 0000 0000 0000 |

¿Cuál es su valor en base 10? ¿Cuál es la precisión del número representado?

Como su exponente es 0000 0000 y su mantisa no es cero, es un número desnormalizado:

- Su exponente es el exponente mínimo del estándar –126.
- Su bit escondido es 0.

Por lo tanto representa el número

$$0,0001011 \cdot 2^{-126}$$

que se corresponde con el número en base 10

$$(2^{-4} + 2^{-6} + 2^{-7}) \cdot 2^{-126} \approx 1,0102 \cdot 10^{-39}.$$

Este número tiene una precisión de solo 20 (no 24 como corresponde a los números normalizados en precisión simple), puesto que a efectos de precisión no cuentan los tres ceros a la izquierda del primer uno de la mantisa (los tres ceros en rojo 000).

Números desnormalizados en este estándar

Volvemos a nuestro estándar. Recordemos que el formato es

| signo | exponente | mantisa |
|-------|--|---------------------------------|
| s | e ₁ e ₂ e ₃ | 1,m ₁ m ₂ |

y para números desnormalizados

| signo | exponente | mantisa |
|-------|-----------|---------------------------------|
| s | 000 | 0,m ₁ m ₂ |

Como los números que vamos a representar son desnormalizados, su exponente es 000 y su mantisa no puede ser cero:

- El valor del exponente es el exponente mínimo del estándar –2.
- Su bit escondido es 0.

En este estandar los números desnormalizados son

| Num(bin) | Num(dec) | espacio |
|----------|----------------------|--------------------------|
| 0 000 01 | $0,01 \times 2^{-2}$ | 0,0625 |
| 0 000 10 | $0,10 \times 2^{-2}$ | 2 ⁻⁴ = 0,0625 |
| 0 000 11 | $0,11 \times 2^{-2}$ | 0,1875 |

Como el exponente, el signo (estamos contando solo los positivos) y el bit escondido no varían sólo hemos de tener en cuenta el número de bits de la mantisa $n = 2$ y le hemos de descontar el caso donde la mantisa es todo ceros. Por lo tanto

$$2^n - 1 = 2^2 - 1 = \boxed{3 \text{ números desnormalizados}}$$

Vemos que:

- El espacio entre números es constante y coincide con el espacio entre números normalizados más pequeños.
- El mínimo número desnormalizado a representar de forma exacta es 0,0625 y el máximo es 0,1875 que están entre el cero y el menor número desnormalizado.
- La precisión es 1 para el primer número y dos para los dos siguientes (la precisión de este estándar para los números normalizados es tres, los dos bits más el uno del bit escondido).

4. El epsilon de máquina

Recordemos que el formato es

| signo | exponente | mantisa |
|-------|--|---------------------------------|
| s | e ₁ e ₂ e ₃ | 1,m ₁ m ₂ |

Se define el ϵ de máquina como la distancia entre el número 1 y el siguiente número representable de forma exacta en este estándar. Recordemos que en este estándar el 1 y el siguiente número son

$$\begin{array}{ll} 1,00 \times 2^0 & 1 \\ 1,01 \times 2^0 & 1,25 \end{array}$$

Calculamos la distancia restando estos dos valores

$$\begin{array}{r} 1,01 \times 2^0 & 1,25 \\ - 1,00 \times 2^0 & 1 \\ \hline \epsilon \rightarrow 0,01 \times 2^0 \rightarrow 0,25 \end{array}$$

Es decir, para este estándar

$$\boxed{\epsilon = 0,25}$$

El ϵ de máquina es una cota superior del error relativo de redondeo que se comete al almacenar cualquier número en este estándar. Comprobaremos esto en el apartado siguiente.

El epsilon de máquina en precisión simple

Si aplicamos la definición para precisión sencilla, la representación de uno y el siguiente número representable de forma exacta, teniendo en cuenta que tenemos 23 bits de mantisa más el bit escondido, es

$$\begin{array}{r} 1.0000\ 0000\ 0000\ 0000\ 000 \times 2^0 \\ 1.0000\ 0000\ 0000\ 0000\ 001 \times 2^0 \end{array}$$

Si restamos estos dos números

$$\begin{array}{r} 1.0000\ 0000\ 0000\ 0000\ 001 \times 2^0 \\ - 1.0000\ 0000\ 0000\ 0000\ 000 \times 2^0 \\ \hline \epsilon \rightarrow 0.0000\ 0000\ 0000\ 0000\ 001 \times 2^0 \rightarrow [2^{-23} \approx 1.19 \times 10^{-7}] \end{array}$$

El epsilon de máquina en precisión doble

Si aplicamos la definición para precisión doble, la representación de uno y el siguiente número representable de forma exacta, teniendo en cuenta que tenemos 52 bits de mantisa más el bit escondido, es

$$\begin{array}{r} \overbrace{1.0000\ 0000\ \dots\ 00}^{52\text{ bits}} \times 2^0 \\ - 1.0000\ 0000\ \dots\ 01 \times 2^0 \end{array}$$

Y la diferencia es

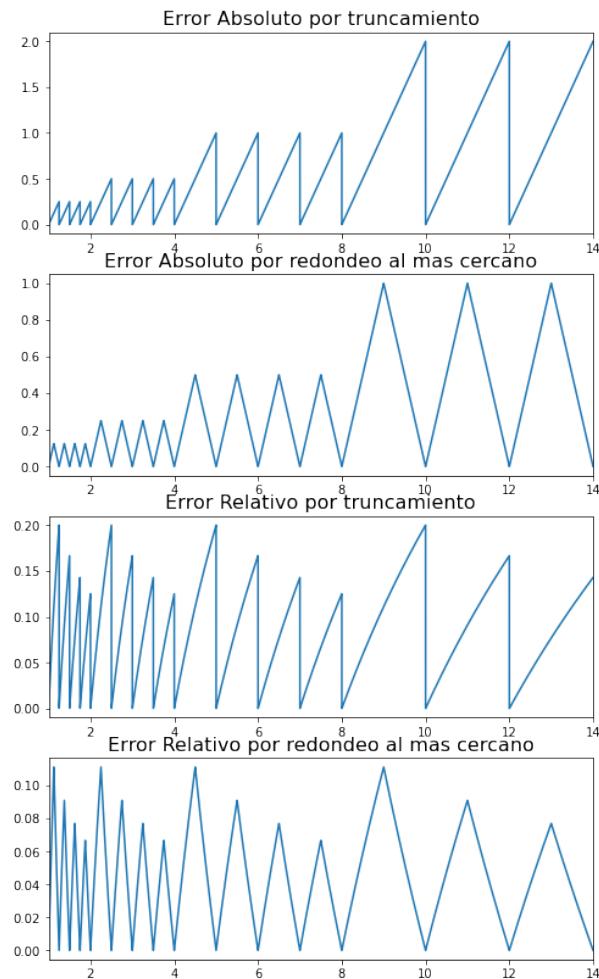
$$\begin{array}{r} \overbrace{1.0000\ 0000\ \dots\ 01}^{52\text{ bits}} \times 2^0 \\ - 1.0000\ 0000\ \dots\ 00 \times 2^0 \\ \hline \epsilon \rightarrow 0.0000\ 0000\ \dots\ 01 \times 2^0 \rightarrow [2^{-52} \approx 2.22 \times 10^{-16}] \end{array}$$

5. Errores de redondeo

De momento, simplemente dibujaremos los errores absolutos y relativos que se comenten al almacenar números con este estándar. Hemos visto que podemos representar números entre 0 y 14. Cuando queremos almacenar un número en este rango, si no coincide con alguno de los números de la tabla, lo almacenamos como un número de la tabla.

- Si lo almacenamos como el número más próximo por debajo, cometeremos un *error de truncamiento o hacia el cero* (matizaremos más adelante estas denominaciones) que será la distancia del número que queremos representar al número que está por debajo más próximo.
- El máximo error se dará justo por debajo de los números representables de forma exacta.
- Si lo almacenamos al número más próximo, cometeremos un *error de redondeo o al par más cercano* que será la distancia al número más cercano.

- El máximo error se en el punto medio entre dos números representables de forma exacta.



Podemos ver que:

- El error absoluto crece con el número, pero el error relativo, que es el importante, no.
- El error máximo por truncamiento es el doble del error máximo por redondeo.
- **Todos** los errores relativos son menores que $\epsilon = 0,25$ como habíamos dicho.

6. Máximo entero

Queremos calcular el máximo entero representable de forma exacta de forma que el siguiente entero no es representable de forma exacta.

Recordemos que el formato es

| signo | exponente | mantisa |
|-------|--|----------------------------------|
| s | e ₁ e ₂ e ₃ | 1, m ₁ m ₂ |



Si nos fijamos en los enteros representados vemos que podemos representar del 1 al 8 todos los enteros pero el 9 ya no lo podemos representar de forma exacta. Por lo tanto, la respuesta a la pregunta sería "el 8".

¿Cómo podemos razonar esto de forma general? La idea es que si en un formato podemos almacenar todos los dígitos, no cometemos error. Si tenemos que prescindir de algún dígito, es posible que haya error. Para el actual formato tenemos dos dígitos de mantisa más el bit escondido.

| entero | binario | |
|--------|---------|-------------------|
| 1 | 1 | $1,00 \times 2^0$ |
| 2 | 10 | $1,00 \times 2^1$ |
| 3 | 11 | $1,10 \times 2^1$ |
| 4 | 100 | $1,00 \times 2^2$ |
| 5 | 101 | $1,01 \times 2^2$ |
| 6 | 110 | $1,10 \times 2^2$ |
| 7 | 111 | $1,11 \times 2^2$ |
| 8 | 1000 | $1,00 \times 2^3$ |
| 9 | 1001 | $1,00 \times 2^3$ |
| 10 | 1010 | 1.01×2^3 |
| 11 | 1011 | $1,10 \times 2^3$ |
| 12 | 1100 | $1,10 \times 2^3$ |
| 13 | 1101 | $1,10 \times 2^3$ |
| 14 | 1110 | $1,11 \times 2^3$ |

Podemos ver que:

- Hasta el número 7 no hay problema porque tenemos espacio para almacenar todos los dígitos significativos.
- Para el número 8 nos falta espacio para almacenar el último dígito, pero como es un 0, no se comete error.
- El 9 ya no se puede almacenar de forma exacta porque el dígito que no podemos almacenar es un 1. Tendremos que *redondearlo* al 8 o al 10. Como están a la misma distancia de los dos se redondea al que en representación binaria la mantisa acabe en 0.

- Para el 10 el caso es parecido al 8. No tenemos espacio para todos los dígitos, pero como el que no nos cabe es un **0** no se comete error.
- El 11 no se puede almacenar de forma exacta porque el dígito que no podemos almacenar es un **1**. Tendremos que *redondearlo* al 10 o al 12. Como están a la misma distancia de los dos se redondea al que en representación binaria la mantisa acabe en 0. En este caso hacia 1.10×2^3 .
- Y así sucesivamente.

Este valor, 8 para este formato, nos da una idea de la capacidad de este formato para almacenar enteros de forma exacta.

Máximo entero en precisión simple

Busquemos el máximo entero almacenado de forma exacta de forma que todos los enteros por debajo de él se almacenan de forma exacta.

Si hacemos el mismo razonamiento que en el caso anterior, teniendo en cuenta de que en este formato disponemos de 23 bits más el bit escondido, es decir, que podemos almacenar 24 dígitos significativos

| int | bin |
|----------|---|
| 1 | 1 1.0000 0000 0000 0000 0000 000 $\times 2^0$ |
| 2 | 10 1.0000 0000 0000 0000 0000 000 $\times 2^1$ |
| 3 | 11 1.1000 0000 0000 0000 0000 000 $\times 2^1$ |
| 4 | 100 1.0000 0000 0000 0000 0000 000 $\times 2^2$ |
| 5 | 101 1.0100 0000 0000 0000 0000 000 $\times 2^2$ |
| 6 | 110 1.1000 0000 0000 0000 0000 000 $\times 2^2$ |
| 7 | 111 1.1100 0000 0000 0000 0000 000 $\times 2^2$ |
| : | : |
| 16777215 | 1 1111 1111 1111 1111 1111 111 1.1111 1111 1111 1111 1111 111 $\times 2^{23}$ |
| 16777216 | 10 0000 0000 0000 0000 0000 0000 0000 0 $\times 2^{24}$ |
| 16777217 | 10 0000 0000 0000 0000 0000 0000 0001 1 1.0000 0000 0000 0000 000 $\times 2^{24}$ |
| 16777218 | 10 0000 0000 0000 0000 0000 0000 0001 0 1.0000 0000 0000 0000 0000 0001 $\times 2^{24}$ |
| 16777219 | 10 0000 0000 0000 0000 0000 0000 0011 1 1.0000 0000 0000 0000 0000 0010 $\times 2^{24}$ |
| 16777220 | 10 0000 0000 0000 0000 0000 0000 0100 0 1.0000 0000 0000 0000 0000 0000 010 $\times 2^{24}$ |
| : | : |

Es decir, disponemos de 23 bits más el bit escondido para almacenar dígitos. Por lo tanto, el mayor entero del que podemos almacenar todos los dígitos es

$$1\ 1111\ 1111\ 1111\ 1111\ 1111\ 1111$$

es decir

$$1 \underbrace{1111 \cdots 111}_{23 \text{ bits}} \rightarrow 1, \underbrace{1111 \cdots 111}_{23 \text{ bits}} \times 2^{23}$$

El número siguiente es

$$10 \overbrace{0000 \cdots 00}^{23 \text{ bits}} = 1 \overbrace{0000 \cdots 00}^{23 \text{ bits}} 0 \rightarrow 1, \overbrace{0000 \cdots 00}^{23 \text{ bits}} \times 2^{24}$$

del cual no podemos almacenar el cero **0**. Y como es un cero no hay error. Este número es

$$2^{24} = 16777216$$

El número entero siguiente es

$$10 \overbrace{0000 \cdots 00}^{23 \text{ bits}} 1 = 1 \overbrace{0000 \cdots 00}^{23 \text{ bits}} 1 \rightarrow 1, \overbrace{0000 \cdots 00}^{23 \text{ bits}} \times 2^{24}$$

y no podemos almacenar el uno **1** y lo tenemos que redondear al número anterior o al siguiente (en este caso al anterior) y no lo podemos representar de forma exacta.

Máximo entero en precisión doble

Busquemos el máximo entero almacenado de forma exacta de forma que todos los enteros por debajo de él se almacenan de forma exacta.

Hagamos el mismo razonamiento que en el caso anterior, teniendo en cuenta de que en este formato disponemos de 52 bits más el bit escondido, es decir, que podemos almacenar 53 dígitos significativos

Por lo tanto, el mayor entero del que podemos almacenar todos los dígitos es

$$1 \overbrace{1111 \cdots 11}^{52 \text{ bits}} \rightarrow 1, \overbrace{1111 \cdots 11}^{52 \text{ bits}} \times 2^{52}$$

El número siguiente es

$$10 \overbrace{0000 \cdots 00}^{52 \text{ bits}} = 1 \overbrace{0000 \cdots 00}^{52 \text{ bits}} 0 \rightarrow 1, \overbrace{0000 \cdots 00}^{52 \text{ bits}} \times 2^{53}$$

del cual no podemos almacenar el cero **0**. Y como es un cero no hay error. Este número es

$$2^{53} = 9007199254740992$$

El número entero siguiente

$$10 \overbrace{0000 \cdots 01}^{52 \text{ bits}} = 1 \overbrace{0000 \cdots 00}^{52 \text{ bits}} 1 \rightarrow 1, \overbrace{0000 \cdots 00}^{52 \text{ bits}} \times 2^{53}$$

y no podemos almacenar el uno **1** y lo tenemos que redondear al número anterior o al siguiente (en este caso al anterior) y no lo podemos representar de forma exacta.

7. Valores especiales

Recordemos que el formato es

| signo | exponente | mantisa |
|-------|---------------|--------------|
| s | $e_1 e_2 e_3$ | $1, m_1 m_2$ |

Por convenio, el cero se representa con todos los bits del exponente y la mantisa cero.

| | signo | exponente | mantisa |
|----|-------|-----------|---------|
| +0 | 0 | 000 | 00 |
| -0 | 1 | 000 | 00 |

Por convenio, el infinito se representa con todos los bits del exponente 1 y todos los bits de la mantisa 0

| | signo | exponente | mantisa |
|-----------|-------|-----------|---------|
| $+\infty$ | 0 | 111 | 00 |
| $-\infty$ | 1 | 111 | 00 |

Por convenio, NaN (Not a Number) se representa con todos los bits del exponente 1 y los bits de la mantisa con cualquier combinación que no sean todos ceros, por ejemplo

| | signo | exponente | mantisa |
|-----|-------|-----------|---------|
| NaN | 0 | 111 | 01 |

Ejercicio 1.3.2

Una máquina almacena números en punto flotante en 10 bits. El primer bit se usa para el signo del número, los cuatro siguientes para el exponente sesgado y los últimos cinco bits para la magnitud de la mantisa. Si se sigue un criterio similar al de la norma IEEE 754:

| signo | exponente | mantisa |
|-------|---|--|
| s | e ₁ e ₂ e ₃ e ₄ | m ₁ m ₂ m ₃ m ₄ m ₅ |

1. Calcular el número $(1011011010)_2$ en base 10.

| signo | exponente | mantisa |
|-------|-----------|---------|
| 1 | 0110 | 11010 |

2. ¿Cuál sería el ϵ de máquina expresado en base 10?
3. ¿Cuál es el mayor entero que se puede almacenar de forma exacta de forma que el siguiente entero no se puede almacenar de forma exacta?
4. ¿Cuales son el menor y el mayor real positivo que se almacena en forma normalizada? ¿Cómo se almacenarían en binario? ¿Qué precisión tendrían? ¿Qué separación tienen con el siguiente/anterior número almacenable?
5. ¿Cuántos números normalizados se pueden representar con este sistema?
6. ¿Cuales son el menor y el mayor real positivo que se almacena en forma desnormalizada? ¿Cómo se almacenarían en binario? ¿Qué precisión tendría cada uno? ¿Qué separación tienen con el siguiente/anterior número almacenable?
7. ¿Cuántos números desnormalizados se pueden representar con este sistema?
8. ¿Cuál sería la representación de 0, $+\infty$, $-\infty$?
9. Da un ejemplo de representación de NaN.
10. ¿Cuál sería la representación del número $-1,5625$ en este sistema?

1. Calcular el número 1011011010 en base 10.

| signo | exponente | mantisa |
|-------|-----------|---------|
| 1 | 0110 | 11010 |

Signo

Como es 1 → signo negativo

Exponente

Tenemos $m = 4$ bits para el exponente. Por lo tanto hay $2^m = 2^4 = 16$ combinaciones distintas y, en principio, podemos representar 16 números. Como empezamos el valor nominal empieza en 0 acabará en 15. El primer número, 0000, y el último, 1111 se reservan. Y como la representación es sesgada, tenemos que restar el

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

para obtener el valor representado

| Número binario | Valor nominal | Valor representado |
|----------------|---------------|--------------------|
| 0000 | 0 | R |
| 0001 | 1 | -6 |
| 0010 | 2 | -5 |
| ⋮ | ⋮ | ⋮ |
| 0110 | 6 | -1 |
| ⋮ | ⋮ | ⋮ |
| 1101 | 13 | 6 |
| 1110 | 14 | 7 |
| 1111 | 15 | R |

El valor nominal del exponente 0110 teniendo en cuenta la posición de cada dígito

$$\begin{array}{cccc} (3) & (2) & (1) & (0) \\ 0 & 1 & 1 & 0 \end{array}$$

es

$$2^2 + 2^1 = 4 + 2 = 6$$

Y si tenemos en cuenta que para $m = 4$ bits, sesgo = 7

El valor del exponente es $6 - 7 = -1$.

Mantisa

Los dígitos almacenados de la mantisa son 11010 y teniendo en cuenta el bit escondido, que es uno, la mantisa es

$$1, \textcolor{teal}{11010}$$

que teniendo en cuenta la posición respecto a al coma

$$\begin{array}{ccccccc} (0) & (-1) & (-2) & (-3) & (-4) & (-5) \\ 1 & , & \textcolor{red}{1} & \textcolor{red}{1} & 0 & \textcolor{red}{1} & 0 \end{array}$$

Número

Por lo tanto, si escribimos el número este es

$$-1, \textcolor{red}{11010} \times 2^{-1} \rightarrow -(1 + 2^{-1} + 2^{-2} + 2^{-4}) \times 2^{-1} = \boxed{-0,90625}$$

2. Épsilon de máquina

Recordemos que el formato es

| signo | exponente | mantisa |
|-------|---|---|
| s | e ₁ e ₂ e ₃ e ₄ | 1, m ₁ m ₂ m ₃ m ₄ m ₅ |

Se define el ϵ de máquina como la distancia entre el número 1 y es siguiente número representable de forma exacta en este estándar.

Recordemos que en este estándar el 1 y el siguiente número son

$$\begin{array}{l} 1,00000 \times 2^0 \\ 1,00001 \times 2^0 \end{array}$$

Calculamos la distancia restando estos dos valores

$$\begin{array}{r} 1,00001 \times 2^0 \\ - 1,00000 \times 2^0 \\ \hline \epsilon \rightarrow 0,00001 \times 2^0 \rightarrow \boxed{2^{-5} \times 2^0 = 2^{-5} = 0,03125} \end{array}$$

3. Mayor entero

Mayor entero que se puede almacenar de forma exacta de forma que el siguiente entero no se puede almacenar de forma exacta

| signo | exponente | mantisa |
|-------|---|---|
| s | e ₁ e ₂ e ₃ e ₄ | 1, m ₁ m ₂ m ₃ m ₄ m ₅ |

En este formato disponemos de 5 bits más el bit escondido, es decir, que podemos almacenar 6 dígitos significativos.

Por lo tanto, el mayor entero del que podemos almacenar todos los dígitos es

$$1\ 11111 \rightarrow 1, 11111 \times 2^5$$

El número siguiente es

$$10 \overbrace{00000}^{5 \text{ bits}} = 1 \overbrace{00000}^{5 \text{ bits}} 0 \rightarrow 1, \overbrace{00000}^{5 \text{ bits}} \times 2^6$$

del cual no podemos almacenar el cero **0**. Y como es un cero no hay error. Este número es

$$2^6 = 64$$

El número entero siguiente

$$\overbrace{10\ 0000}^{5 \text{ bits}} \overbrace{1}^{5 \text{ bits}} \rightarrow 1, \overbrace{00000}^{5 \text{ bits}} \times 2^6$$

y no podemos almacenar el uno **1** y lo tenemos que redondear al número anterior o al siguiente (en este caso al anterior) y no lo podemos representar de forma exacta.

4. Números normalizados

| signo | exponente | mantisa |
|-------|---|---|
| s | e ₁ e ₂ e ₃ e ₄ | 1, m ₁ m ₂ m ₃ m ₄ m ₅ |

Si volvemos al primer apartado, podemos ver que el exponente mínimo es $e_{min} = -6$ y el máximo es $e_{max} = 7$.

El menor real positivo normalizado tendrá mantisa mínima y exponente mínimo. En este formato

| signo | exponente | mantisa |
|-------|-----------|---------|
| 0 | 0001 | 00000 |

El exponente no puede ser 0000 porque está reservado. Para este número

$$1,00000 \times 2^{-6} \rightarrow 2^{-6} = 0,015625$$

Su precisión es $p = 6$ porque tiene 6 dígitos, 5 que se almacenan más el bit escondido.

El siguiente número representable de forma exacta sería

$$1,00001 \times 2^{-6}$$

Y el espacio (diferencia) entre los dos sería

$$0,00001 \times 2^{-6} \rightarrow 2^{-5} \times 2^{-6} = 2^{-11} = 0,00048828125$$

El mayor real positivo normalizado tendrá mantisa máxima y exponente máximo. En este formato

| signo | exponente | mantisa |
|-------|-----------|---------|
| 0 | 1110 | 11111 |

El exponente no puede ser 1111 porque está reservado. Para este número

$$1,11111 \times 2^7 \longrightarrow (1 + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} + 2^{-5}) \times 2^7 = \boxed{252}$$

Su precisión es $\boxed{p = 6}$ porque tiene 6 dígitos, 5 que se almacenan más el bit escondido.

El anterior número representable de forma exacta sería

$$1,11110 \times 2^7$$

Y el espacio (diferencia) entre los dos sería

$$0,00001 \times 2^7 \longrightarrow 2^{-5} \times 2^7 = 2^2 = \boxed{4}$$

5. ¿Cuántos números normalizados?

Para $m = 4$ bits de exponente tenemos 2^m números distintos pero como el primero 0000 y el último 1111 están reservados tenemos

$$2^m - 2 = 2^4 - 2 = 16 - 2 = 14 \text{ exponentes}$$

Para $n = 5$ bits de mantisa tenemos 2^n números distintos. El bit escondido es siempre 1, así que no aporta nuevos valores y tenemos

$$2^n = 2^5 = 32 \text{ mantisas}$$

y en total

$$14 \text{ exponentes} \times 32 \text{ mantisas} = \boxed{448 \text{ números normalizados positivos}}$$

6. Números desnormalizados

| signo | exponente | mantisa |
|-------|-----------|---|
| s | 0000 | 0, m ₁ m ₂ m ₃ m ₄ m ₅ |

Los números desnormalizados se caracterizan porque:

- Los dígitos del exponente son todos cero.
- El exponente representado es el mínimo del sistema, en este caso $e_{min} = -6$.
- El bit escondido es cero.

El **menor real positivo desnormalizado** tendrá exponente todo ceros y mantisa mínima (no vale todo ceros que respresentaría el cero)

| signo | exponente | mantisa |
|-------|-----------|---------|
| 0 | 0000 | 00001 |

Que representa

$$0,00001 \times 2^{-6} \longrightarrow 2^{-5} \times 2^{-6} = 2^{-11} = \boxed{0,00048828125}$$

Su precisión es $p = 1$ porque tiene 1 dígito significativo (los ceros a la izquierda no cuentan).

El siguiente número representable de forma exacta sería

$$0,00010 \times 2^{-6}$$

Y el espacio (diferencia) entre los dos sería

$$0,00001 \times 2^{-6} \longrightarrow 2^{-5} \times 2^{-6} = 2^{-11} = \boxed{0,00048828125}$$

El mayor real positivo desnormalizado tendrá exponente todo ceros y mantisa máxima

| signo | exponente | mantisa |
|-------|-----------|---------|
| 0 | 0000 | 11111 |

Que representa

$$0,11111 \times 2^{-6} \longrightarrow (2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} + 2^{-5}) \times 2^{-6} = \boxed{0,01513671875}$$

Su precisión es $p = 5$ porque tiene 5 dígitos significativos.

El anterior número representable de forma exacta sería

$$0,11110 \times 2^{-6}$$

Y el espacio (diferencia) entre los dos sería

$$0,00001 \times 2^{-6} \longrightarrow 2^{-5} \times 2^{-6} = 2^{-11} = \boxed{0,00048828125}$$

7. ¿Cuántos números desnormalizados?

Tenemos un único exponente que es 0000.

Para $n = 5$ bits de mantisa tenemos 2^n números distintos. El bit escondido es siempre 0, así que no aporta nuevos valores y tenemos

Y tenemos

$$2^n - 1 = 2^5 - 1 = 32 - 1 \text{ mantisas}$$

porque hemos de quitarle la que es todo ceros que representa el cero.

Y en total

$$1 \text{ exponente} \times 31 \text{ mantisas} = \boxed{31 \text{ números desnormalizados positivos}}$$

8. Valores especiales

Recordemos que el formato es

| signo | exponente | mantisa |
|-------|---|--|
| s | e ₁ e ₂ e ₃ e ₄ | m ₁ m ₂ m ₃ m ₄ m ₅ |

Por convenio, el cero se representa con todos los bits del exponente y la mantisa cero.

| | signo | exponente | mantisa |
|----|-------|-----------|---------|
| +0 | 0 | 0000 | 00000 |
| -0 | 1 | 0000 | 00000 |

Por convenio, el infinito se representa con todos los bits del exponente 1 y todos los bits de la mantisa 0

| | signo | exponente | mantisa |
|----|-------|-----------|---------|
| +∞ | 0 | 1111 | 00000 |
| -∞ | 1 | 1111 | 00000 |

Por convenio, NaN se representa con todos los bits del exponente 1 y los bits de la mantisa con cualquier combinación que no sean todos ceros, por ejemplo

| | signo | exponente | mantisa |
|-----|-------|-----------|---------|
| NaN | 0 | 1111 | 01101 |

10. Representa -1.5625 en este sistema

Paso del sistema decimal a binario

Parte entera

La parte entera es 1 en base 2 y en base 10

Parte fraccionaria

Multiplicamos por 2, restamos la parte entera y repetimos hasta que la parte fraccionaria sea cero.

$$\begin{array}{r}
 0,5625 \times 2 = 1,125 \rightarrow 1 \downarrow \\
 0,125 \times 2 = 0,25 \rightarrow 0 \downarrow \\
 0,25 \times 2 = 0,5 \rightarrow 0 \downarrow \\
 0,5 \times 2 = 1,0 \rightarrow 1 \downarrow
 \end{array}$$

Empezamos por la primera parte entera y el número en binario es 0,1001

Y el número completo es

$$(1,5625)_{10} = (1,1001)_2$$

Normalización

1. Movemos la coma de forma que aparezca un único dígito distinto de cero a su izquierda.
2. Al mover la coma tenemos que multiplicar por 2^n siendo n el número de posiciones que hemos movido la coma a la izquierda o 2^{-n} siendo n el número de posiciones que hemos movido la coma a la derecha.
3. Le añadimos el signo.

Este número normalizado sería

$$-1,1001 \times 2^0$$

con

- Signo: -
- Mantisa: 1,1001
- Exponente: 0

Signo

Como el signo es negativo \rightarrow signo 1

Exponente

Tenemos $m = 4$ bits para el exponente. Por lo tanto hay $2^m = 2^4 = 16$ combinaciones distintas y, en principio, podemos representar 16 números. Como empezamos el valor nominal empieza en 0 acabará en 15. El primer número, 0000, y el último, 1111 se reservan. Y como la representación es sesgada, tenemos que restar el

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

para obtener el valor representado.

El valor del exponente es 0. Para obtener el valor nominal tenemos que sumarle el sesgo y $0 + 7 = 7$ que en binario sería

| Dividendo | Divisor | Cociente | Resto | |
|-----------|---------|----------|-------|---|
| 7 | 2 | 3 | 1 | ↑ |
| 3 | 2 | 1 | 1 | ↑ |
| 1 | 2 | 0 | 1 | ↑ |

es decir

$$(7)_{10} = (111)_2$$

Y completamos con un cero a la izquierda

| Número binario | Valor nominal | Valor representado | |
|----------------|---------------|--------------------|----|
| 0000 | 0 | | R |
| 0001 | 1 | | -6 |
| 0010 | 2 | | -5 |
| : | : | sesgo | : |
| 0111 | 7 | +7 | 0 |
| : | : | ← | : |
| 1101 | 13 | | 6 |
| 1110 | 14 | | 7 |
| 1111 | 15 | | R |

Mantisa

La mantisa es

1, 1001.

Hemos de tener en cuenta el bit escondido, que no almacenamos, y que rellenamos con ceros por la derecha hasta tener 5 bits.

Número

El número -1,5621 en este formato, se almacena

| signo | exponente | mantisa |
|-------|-----------|---------|
| 1 | 0111 | 10010 |

1.4 Error absoluto y relativo

Ejercicio 1.4.1

La siguientes gasolineras fueron denunciadas por engaño en las cantidades dispensadas al Ministerio de Industria. ¿Cuál crees que engaño más?

| Gasolinera | Entregado | Cobrado |
|------------|-----------|---------|
| Ser | 9,90 | 10,00 |
| Cit | 19,80 | 20,00 |
| Has | 29,10 | 30,00 |
| Shol | 28,90 | 30,00 |

Vamos a utilizar como criterio el error relativo.

Si x es el valor real y x^* el aproximado:

- Error absoluto

$$e_a = |x - x^*|$$

- Error relativo

$$e_r = \frac{e_a}{|x|}$$

El error relativo está dado en tanto por uno. Si lo multiplicamos por cien vendrá dado en porcentaje.

Ser

- Error absoluto

$$e_a = |x - x^*| = |9,90 - 10,00| = 0,10$$

- Error relativo

$$e_r = \frac{e_a}{|x|} = \frac{0,10}{9,90} \approx 0,01 = 1\%$$

Cit

- Error absoluto

$$e_a = |x - x^*| = |19,80 - 20,00| = 0,20$$

- Error relativo

$$e_r = \frac{e_a}{|x|} = \frac{0,20}{19,80} \approx 0,01 = 1\%$$

Has

- Error absoluto

$$e_a = |x - x^*| = |29,10 - 30,00| = 0,90$$

- Error relativo

$$e_r = \frac{e_a}{|x|} = \frac{0,90}{29,10} \approx 0,03 = 3\%$$

Shol

- Error absoluto

$$e_a = |x - x^*| = |28,90 - 30,00| = 1,10$$

- Error relativo

$$e_r = \frac{e_a}{|x|} = \frac{1,10}{28,90} \approx 0,04 = 4\%$$

Resumiendo, tenemos

| Gasolinera | Entregado | Cobrado | e_a | e_r |
|------------|-----------|---------|-------|-------|
| Ser | 9,90 | 10,00 | 0,10 | 1% |
| Cit | 19,80 | 20,00 | 0,20 | 1% |
| Has | 29,10 | 30,00 | 0,90 | 3% |
| Shol | 28,90 | 30,00 | 1,10 | 4% |

Y la mayor diferencia relativa entre lo entregado y lo cobrado se produce para la gasolinera Shol y podemos considerar que es la que más estafó.

1.5 Redondeo con la norma IEEE 754

Ejercicio 1.5.1

Truncar y redondear al par más cercano, si la precisión es 5, los siguientes números:

| | | | | | | |
|---------|----------|----------|----------|----------|----------|----------|
| base 10 | 1,999956 | 1,999943 | 2,462150 | 2,462250 | 2,462151 | 2,462149 |
| base 2 | 1,111111 | 1,111101 | 1,010110 | 1,010010 | 1,010011 | 1,010001 |

INTRODUCCIÓN

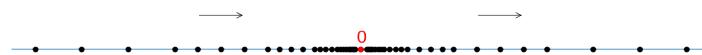
A menudo, para un número real x , no existe una representación exacta en coma flotante y está entre dos números consecutivos que se pueden representar de forma exacta

$$x^- < x < x^+$$



Como representación de x elegiremos uno de los dos dependiendo del método de redondeo usado. La norma IEEE 754 propone 5 sistemas de redondeo.

- *Hacia arriba*: redondea hacia el número mayor más próximo.



- *Hacia abajo*: redondea hacia el número menor más próximo.



- *Hacia el cero* (“truncamiento”).

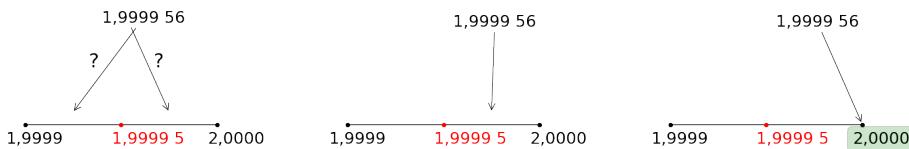


- Hacia el más cercano, y a igualdad de distancia, hacia el mayor en valor absoluto.
- Hacia el más cercano, y a igualdad de distancia, el que tiene el último dígito (el menos significativo) par (“redondeo al par más cercano”). Es decir, puede redondear hacia arriba o hacia abajo.

Este último es el método de redondeo más utilizado y uno de los que vamos a utilizar en este ejercicio, llamándolo “Redondeo”.

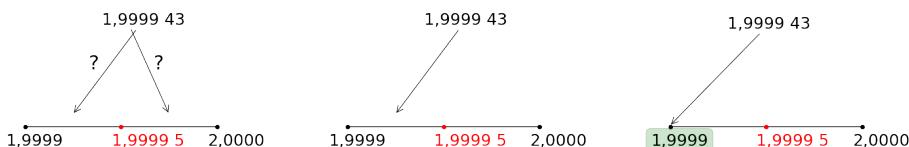
BASE 10

1,999956



- **Truncamiento:** Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 1,9999
- **Redondeo:** Tomamos el siguiente número en esta precisión que es 2,0000. Tomamos el punto medio entre los dos que es 1,99995. Como el número que queremos redondear es mayor que este punto medio está más cerca de 2,0000 y redondea hacia arriba, a 2,0000

1,999943



- **Truncamiento:** Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 1,9999
- **Redondeo:** Tomamos el siguiente número en esta precisión que es 2,0000. Tomamos el punto medio entre los dos que es 1,99995. Como el número que queremos redondear es menor que este punto medio está más cerca de 1,9999 y redondea hacia abajo, a 1,9999

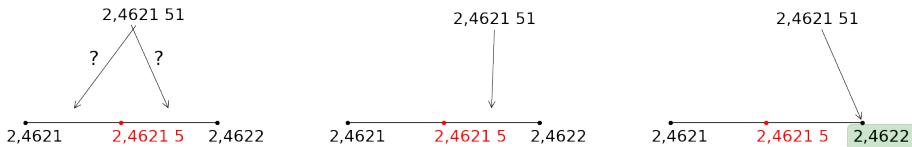
2,462150

- **Truncamiento:** Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 2,4621
- **Redondeo:** Tomamos el siguiente número en esta precisión que es 2,4622. Tomamos el punto medio entre los dos que es 2,46215. Como el número que queremos redondear es exactamente este punto medio, está a igual distancia de los dos y redondea al que acaba en número par 2,4622

2,462250

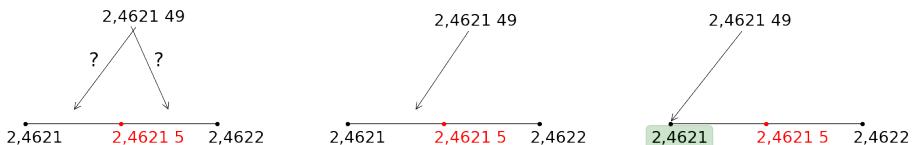
- **Truncamiento:** Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 2,4622
- **Redondeo:** Tomamos el siguiente número en esta precisión que es 2,4623. Tomamos el punto medio entre los dos que es 2,46225. Como el número que queremos redondear es exactamente este punto medio, está a igual distancia de los dos y redondea al que acaba en número par 2,4622

2,462151

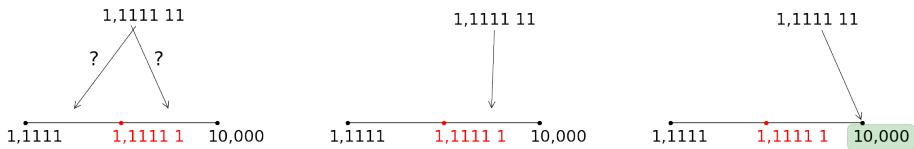


- *Truncamiento:* Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 2,4621
- *Redondeo:* Tomamos el siguiente número en esta precisión que es 2,4622. Tomamos el punto medio entre los dos que es 2,46215. Como el número que queremos redondear es mayor que el punto medio redondea hacia arriba, a 2,4622

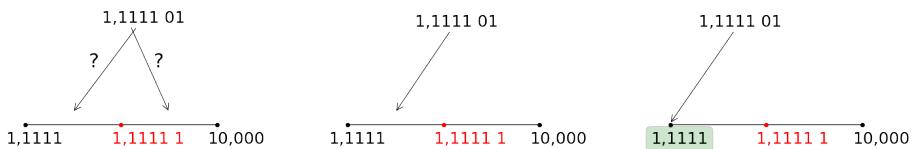
2,462149



- *Truncamiento:* Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 2,4621
- *Redondeo:* Tomamos el siguiente número en esta precisión que es 2,4622. Tomamos el punto medio entre los dos que es 2,46215. Como el número que queremos redondear es menor que el punto medio redondea hacia abajo, a 2,4621

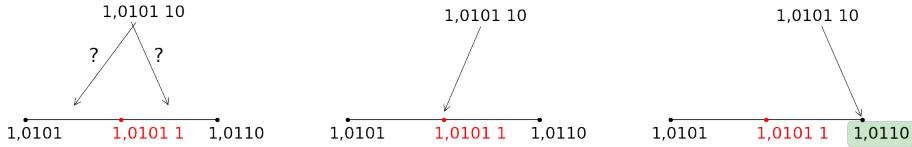
BASE 2**1,111111**

- *Truncamiento:* Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 1,1111
- *Redondeo:* Tomamos el siguiente número en esta precisión que es 10,000. Tomamos el punto medio entre los dos que es 1,11111. Como el número que queremos redondear es mayor que el punto medio redondea hacia arriba, a 10,000

1,111101

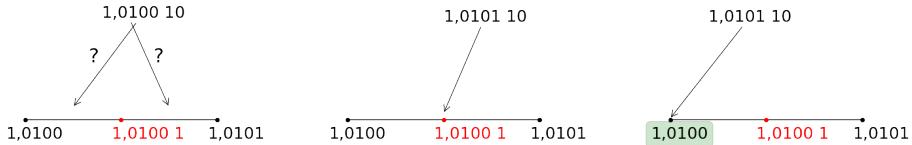
- *Truncamiento:* Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 1,1111
- *Redondeo:* Tomamos el siguiente número en esta precisión que es 10,000. Tomamos el punto medio entre los dos que es 1,11111. Como el número que queremos redondear es menor que el punto medio redondea hacia abajo, a 1,1111

1,010110



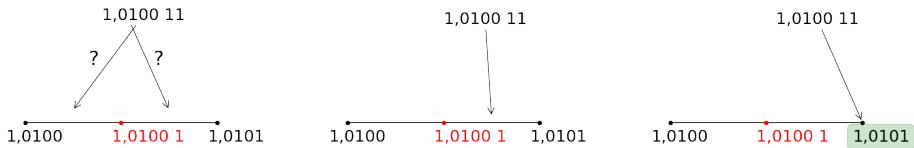
- *Truncamiento:* Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 1,0101
- *Redondeo:* Tomamos el siguiente número en esta precisión que es 1,0110. Tomamos el punto medio entre los dos que es 1,01101. Como el número que queremos redondear es exactamente el punto medio, redondea hacia el que acaba en par, es decir, en cero a 1,0110

1,010010



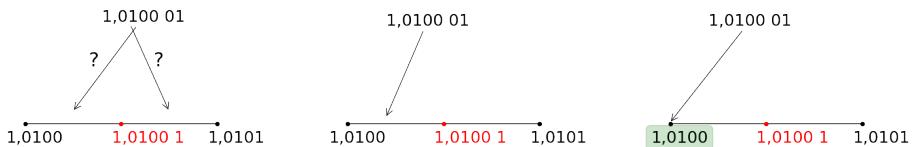
- *Truncamiento:* Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 1,0100
- *Redondeo:* Tomamos el siguiente número en esta precisión que es 1,0101. Tomamos el punto medio entre los dos que es 1,01001. Como el número que queremos redondear es exactamente el punto medio, redondea hacia el que acaba en par, es decir, en cero a 1,0100

1,010011



- **Truncamiento:** Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 1,0100
- **Redondeo:** Tomamos el siguiente número en esta precisión que es 1,0101. Tomamos el punto medio entre los dos que es 1,01001. Como el número que queremos redondear es mayor que punto medio, redondea hacia arriba, a 1,0101

1,010001



- **Truncamiento:** Le quitamos los dos últimos dígitos para dejar el número de dígitos en 5 1,0100
- **Redondeo:** Tomamos el siguiente número en esta precisión que es 1,0101. Tomamos el punto medio entre los dos que es 1,01001. Como el número que queremos redondear es menor que punto medio, redondea hacia abajo, a 1,0100

Ejercicio 1.5.2

1. Representar 0,3 en precisión simple. Redondear al par más cercano.
2. Dar el error absoluto y relativo en base 10. Comprobar que el error relativo es menor que el épsilon de máquina.

1. Representar 0,3 en precisión simple. Redondear al par más cercano**Paso del sistema decimal a binario**

Este número no tiene parte entera, solo parte fraccionaria. Para convertirla a base 2 multiplicamos por 2, restamos la parte entera, que guardamos porque van a ser nuestros dígitos binarios, y repetimos.

$$\begin{array}{r}
 \begin{array}{rcllll}
 0,3 & \times & 2 & = & 0,6 & \rightarrow & 0 \downarrow \\
 0,6 & \times & 2 & = & 1,2 & \rightarrow & 1 \downarrow \\
 0,2 & \times & 2 & = & 0,4 & \rightarrow & 0 \downarrow \\
 0,4 & \times & 2 & = & 0,8 & \rightarrow & 0 \downarrow \\
 0,8 & \times & 2 & = & 1,6 & \rightarrow & 1 \downarrow \\
 \hline
 0,6 & \times & 2 & = & 1,2 & \rightarrow & 1 \downarrow \\
 0,2 & \times & 2 & = & 0,4 & \rightarrow & 0 \downarrow \\
 0,4 & \times & 2 & = & 0,8 & \rightarrow & 0 \downarrow \\
 0,8 & \times & 2 & = & 1,6 & \rightarrow & 1 \downarrow \\
 \end{array} \\
 \vdots \qquad \vdots \qquad \vdots
 \end{array}$$

En este caso, podemos seguir obteniendo dígitos hasta el infinito. Es un número binario con parte fraccionaria periódica. Guardamos los dígitos empezando por el de arriba

$$(0,3)_{10} = (0,0\overline{1001100110011001100110011001\dots})_2$$

Normalización

1. Movemos la coma de forma que aparezca un único dígito distinto de cero a su izquierda.
2. Al mover la coma tenemos que multiplicar por 2^n siendo n el número de posiciones que hemos movido la coma a la izquierda o 2^{-n} siendo n el número de posiciones que hemos movido la coma a la derecha.
3. Le añadimos el signo.

Este número normalizado sería

$$+1,001100110011001100110011001\dots \times 2^{-2}$$

Signo

Como el signo es positivo \rightarrow signo 0

Exponente

Tenemos $m = 8$ bits para el exponente. Por lo tanto hay $2^m = 2^8 = 256$ combinaciones distintas y, en principio, podemos representar 256 números. Como empezamos el valor nominal empieza en 0 acabará en 255. El primer número, 0000 0000, y el último, 1111 1111 se reservan (ya veremos luego para qué). Y como la representación es sesgada, tenemos que restar el

$$\text{sesgo} = 2^{m-1} - 1 = 128 - 1 = 127$$

para obtener el valor representado.

El valor del exponente es -2 . Para obtener el valor nominal tenemos que sumarle el sesgo y $-2 + 127 = 125$ que en binario sería

| Dividendo | Divisor | Cociente | Resto | |
|-----------|---------|----------|-------|---|
| 125 | 2 | 62 | 1 | ↑ |
| 62 | 2 | 31 | 0 | ↑ |
| 31 | 2 | 15 | 1 | ↑ |
| 15 | 2 | 7 | 1 | ↑ |
| 7 | 2 | 3 | 1 | ↑ |
| 3 | 2 | 1 | 1 | ↑ |
| 1 | 2 | 0 | 1 | ↑ |

es decir

$$(125)_{10} = (111\ 1101)_2$$

| Número binario | Valor nominal | Valor representado |
|----------------|---------------|--------------------|
| 0000 0000 | 0 | R |
| 0000 0001 | 1 | -126 |
| 0000 0010 | 2 | -125 |
| 0000 0011 | 3 | -124 |
| ... | ... | -127 |
| ... | ... | → |
| 0111 1101 | 125 | -2 |
| ... | ... | +127 |
| ... | ... | ← |
| 1111 1100 | 252 | 126 |
| 1111 1101 | 253 | 126 |
| 1111 1110 | 254 | 127 |
| 1111 1111 | 255 | R |

Mantisa

$$(0,3)_{10}$$

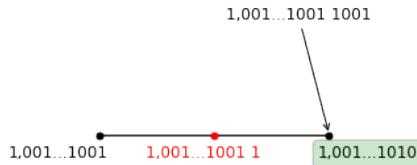
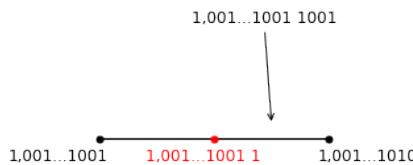
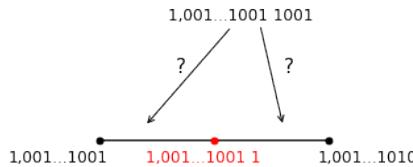
La mantisa es

$$1,001\text{ }1001\text{ }1001\text{ }1001\text{ }1001\text{ }1001\text{ }1001\text{ }1001\ldots$$

Hemos de tener en cuenta el bit escondido, que no almacenamos.

Los dígitos $1001\ldots$ no los vamos a almacenar, pero tenemos que tenerlos en cuenta para redondear el número.

El método de redondeo será *al número par más cercano*.



- *Truncamiento:* Le quitamos los últimos dígitos para dejar el número de dígitos en 24

$$1,001\text{ }1001\text{ }1001\text{ }1001\text{ }1001\text{ }1001$$

- *Redondeo:* Tomamos el siguiente número en esta precisión que es

$$\begin{array}{r}
 1,001\text{ }1001\text{ }1001\text{ }1001\text{ }1001\text{ }1001 \\
 + \\
 0,000\text{ }0000\text{ }0000\text{ }0000\text{ }0000\text{ }0001 \\
 \hline
 1,001\text{ }1001\text{ }1001\text{ }1001\text{ }1010
 \end{array}$$

Tomamos el punto medio entre los dos que es

$$1,001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1$$

Como el número que queremos redondear es mayor que este punto medio, redondea hacia arriba, a

$$1,001\ 1001\ 1001\ 1001\ 1001\ 1010$$

Almacenamos 23 dígitos. El uno a la izquierda de la coma no se almacena. Es el bit escondido.

Número

El número 0,3 en precisión sencilla se almacena

| signo | exponente | mantisa |
|-------|-----------|------------------------------|
| 0 | 0111 1101 | 001 1001 1001 1001 1001 1010 |

2. Dar el error absoluto y relativo en base 10

El número almacenado es

$$1,001\ 1001\ 1001\ 1001\ 1001\ 1010 \times 2^{-2}$$

Que en base 10 es

$$(1 + 2^{-3} + 2^{-4} + 2^{-7} + 2^{-8} + 2^{-11} + 2^{-12} + 2^{-15} + 2^{-16} + 2^{-19} + 2^{-20} + 2^{-22}) \times 2^{-2}$$

Operando, este número es

$$x^* = 0,30000001192092896$$

Y como $x = 0,3$

- Error absoluto

$$e_a = |x - x^*| = 0,00000001192092896 \approx 1,2 \times 10^{-8}$$

- Error relativo

$$e_r = \frac{e_a}{|x|} \approx 4 \times 10^{-8}$$

Y habíamos dicho que este número tenía que ser menor que el épsilon de máquina. Como para precisión simple el épsilon de máquina es

$$\epsilon = 2^{-23} \approx 1,2 \times 10^{-7}$$

tenemos que

$$e_r = 4 \times 10^{-8} < 12 \times 10^{-8} = 1,2 \times 10^{-7} = \epsilon$$

TEMA 2

RAÍCES DE ECUACIONES NO LINEALES

2.1 Raíz de una función

Separación de raíces

Métodos numéricos de cálculo de raíces

Método de bisección

Ejercicio 2.1.1

Sea la ecuación

$$\frac{x}{2}e^{-\frac{x}{2}} + \frac{1}{2} = 0$$

1. Demostrar que en $[-1, 1]$ existe una única raíz.
2. ¿Se puede calcular por el método de bisección partiendo de dicho intervalo?
3. Aproximar la raíz haciendo cuatro iteraciones.
4. Dar una cota del error cometido al calcular esta raíz.
5. ¿Cuántas iteraciones n tendríamos que hacer para garantizar que el error es menor que 10^{-6} ?

INTRODUCCIÓN

Raíz de una función

Una ecuación escalar es una expresión de la forma

$$f(x) = 0$$

donde $f : \Omega \subset \mathbb{R} \rightarrow \mathbb{R}$ es una función.

Cualquier número α tal que $f(\alpha) = 0$ se dice que es una *solución* de f o *una raíz* de f .

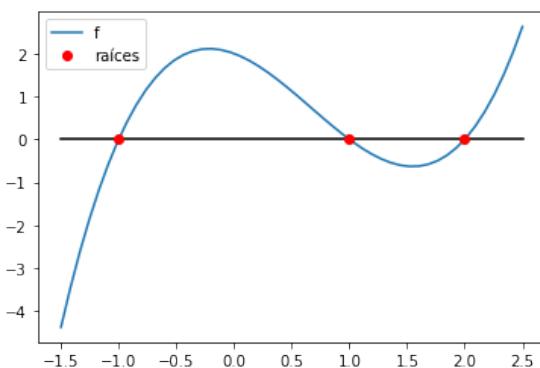
Por ejemplo, la ecuación

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

tiene como raíces $x = -1$, $x = 1$ y $x = 2$ porque

- $f(-1) = (-1)^3 - 2(-1)^2 - (-1) + 2 = -1 - 2 + 1 + 2 = 0$
- $f(1) = (1)^3 - 2(1)^2 - (1) + 2 = 1 - 2 - 1 + 2 = 0$
- $f(2) = (2)^3 - 2(2)^2 - (2) + 2 = 8 - 8 - 2 + 2 = 0$

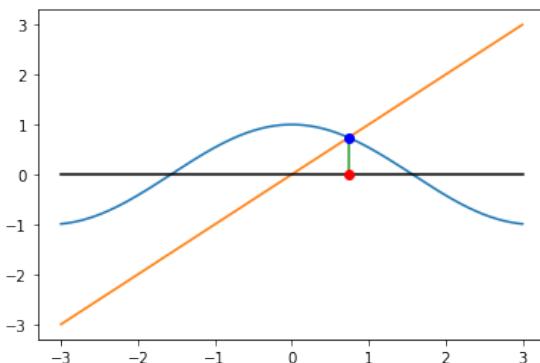
Gráficamente, las raíces son los valores de x para los cuales la curva f corta al eje OX .



Los métodos numéricos no solo son métodos alternativos a los analíticos, sino que en algunos casos son los únicos posibles. Por ejemplo la ecuación

$$\cos x = x$$

tiene una única solución, pero no hay forma de despejar la x y resolver analíticamente el sistema.



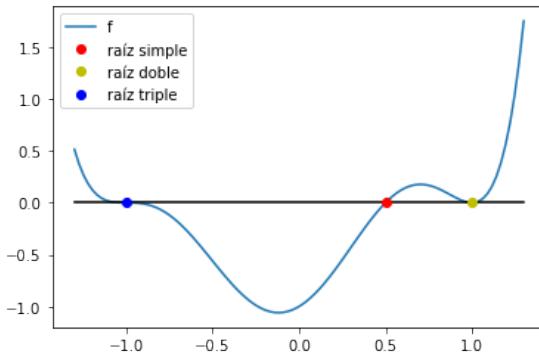
Decimos que una raíz de f , α , es de multiplicidad m si

$$f(\alpha) = f'(\alpha) = \dots = f^{(m-1)}(\alpha) = 0, \quad \text{y} \quad f^{(m)}(\alpha) \neq 0,$$

En el caso de que $m = 1$ decimos que la raíz es *simple*.

Por ejemplo:

$$f(x) = 2(x+1)^3(x-1)^2(x-0.5) = -1 + x + 4x^2 - 2x^3 - 5x^4 + x^5 + 2x^6$$



Separación de raíces

Separar una raíz de f consiste en encontrar un intervalo $[a, b]$ que contiene una y solo una raíz de f .

Para separar raíces tendremos en cuenta el **Teorema de Bolzano**:

Sea $f : [a, b] \rightarrow \mathbb{R}$ una función continua con $f(a)f(b) < 0$, es decir, con signo distinto en los extremos del intervalo. Entonces existe al menos una raíz de f en $[a, b]$.

Si se cumple el teorema de Bolzano y además la función es:

- Estrictamente creciente en $[a, b]$, es decir, $f'(x) > 0$ para todo $x \in (a, b)$ o
- Estrictamente decreciente en $[a, b]$, es decir, $f'(x) < 0$ para todo $x \in (a, b)$.

La función tiene una única raíz en $[a, b]$

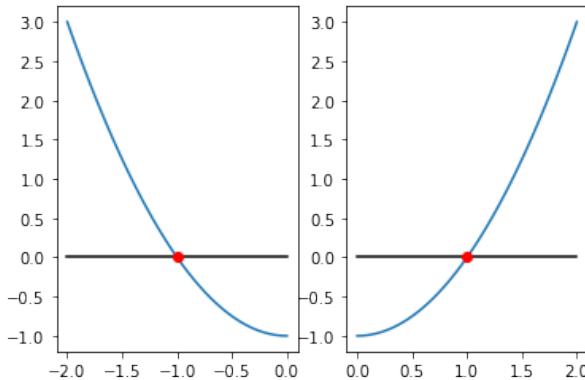
Por ejemplo, si $f(x) = x^2 - 1$ podemos separar sus raíces en los intervalos $[-2, 0]$ y $[0, 2]$.

En el caso de $[-2, 0]$:

- $f(-2) = 3 > 0$
- $f(0) = -1 < 0$
- f es decreciente porque $f'(x) = 2x < 0$ porque x es negativo si $x \in (-2, 0)$.

Para $[0, 2]$:

- $f(0) = -1 < 0$
- $f(2) = 3 > 0$
- f es creciente porque $f'(x) = 2x > 0$ porque x es positivo si $x \in (0, 2)$.



Métodos numéricos de cálculo de raíces y orden de convergencia

Los métodos numéricos de cálculo de raíces de una función son métodos iterativos, es decir, construimos una sucesión

$$x_0, x_1, \dots, x_k, \dots \rightarrow \alpha$$

tal que

$$\lim_{k \rightarrow \infty} f(x_k) = 0.$$

El **orden de convergencia** de un métodos está relacionado con la velocidad de convergencia de la sucesión con respecto a k .

Supongamos que la sucesión x_k converge a $\alpha \in \mathbb{R}$. Decimos que x_k converge a α con orden de convergencia p si

$$\lim_{k \rightarrow \infty} \frac{e_k}{e_{k-1}^p} = \lambda \neq 0$$

siendo $e_k = |x_k - \alpha|$ el error absoluto en el paso k . Es decir

$$e_k \approx \lambda e_{k-1}^p$$

En los casos particulares

- $p = 1$, decimos que la convergencia es lineal.
- $p = 2$, decimos que la convergencia es cuadrática.

Un método numérico se dice de orden p si genera una sucesión que converge a la solución con un orden de convergencia p .

Así, por ejemplo, supongamos que nuestro método tiene convergencia **lineal** con $\lambda = \frac{1}{2}$ y que $e_k = 0.1$. Tenemos que

$$e_{k+1} \approx \frac{e_k}{2} = 0.05 \quad e_{k+2} \approx \frac{e_{k+1}}{2} = 0.025 \quad e_{k+3} \approx \frac{e_{k+2}}{2} = 0.0125 \quad \dots$$

Es decir, a cada paso, el error disminuye a la mitad.

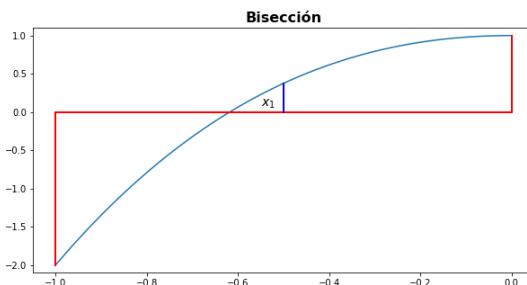
Supongamos ahora que nuestro método tiene convergencia **cuadrática** con $\lambda = 1$ y que $e_k = 0.1$. Tenemos que

$$e_{k+1} \approx e_k^2 = 0.01 \quad e_{k+2} \approx e_{k+1}^2 = 0.0001 \quad e_{k+3} \approx e_{k+2}^2 = 0.00000001 \quad \dots$$

Es decir, a cada paso, el error disminuye de orden de magnitud, primero a las centésimas, luego a las diezmilésimas, etc. Como se puede apreciar el error disminuye mucho más rápidamente que en el caso lineal.

Método de bisección

Sea la función continua $f : [a, b] \rightarrow \mathbb{R}$ tal que $f(a)f(b) < 0$. El teorema de Bolzano garantiza la existencia de una raíz de f en (a, b) .



Algoritmo de bisección

- Sea $a_1 = a, b_1 = b$.
- Para $k = 1, 2, \dots, \text{MaxNumIter}$
 - Calcular el punto medio $x_k = \frac{a_k + b_k}{2}$.
 - Si x_k satisface el criterio de parada, parar.
 - En el caso contrario,
 - * si $f(a_k)f(x_k) < 0$ entonces: $a_{k+1} = a_k$ y $b_{k+1} = x_k$,
 - * si $f(x_k)f(b_k) < 0$ entonces: $a_{k+1} = x_k$ y $b_{k+1} = b_k$,
 - * en otro caso, acabar.

Cota de error del método de Bisección

Supongamos que el intervalo inicial $[a_0, b_0]$ cumple las condiciones del teorema de Bolzano y, sin pérdida de generalidad, que el intervalo contiene una sola raíz. Entonces todos los demás intervalos obtenidos mediante el algoritmo de bisección contienen una única raíz.

Tomaremos como solución aproximada el último x_k obtenido, que será extremo del siguiente intervalo. **Una cota del error es la longitud del intervalo:** el peor caso es que la raíz esté muy cerca del otro extremo del intervalo y entonces el error sería un poco menor que la longitud del intervalo.

Así, tras hacer una iteración, el intervalo es $[a_1, b_1]$ y su longitud es la mitad de la de $[a_0, b_0]$. Es decir, la cota de error cuando hemos hecho una iteración es

$$c_1 = \frac{b_0 - a_0}{2}$$

Si hacemos otra iteración, la longitud del intervalo se reduce a la mitad y la cota de error es

$$c_2 = \frac{1}{2} \times \frac{b_0 - a_0}{2} = \frac{b_0 - a_0}{2^2}$$

Para la iteración 3

$$c_3 = \frac{1}{2} \times \frac{b_0 - a_0}{2^2} = \frac{b_0 - a_0}{2^3}$$

y para la iteración k

$$c_k = \frac{b_0 - a_0}{2^k}$$

Así que dado el número de iteraciones, podemos dar una cota del error. Y viceversa, dada una tolerancia, podemos dar el número de iteraciones que garantizan un error menor que la tolerancia.

Ventajas del método de bisección

- Fácil de programar.
- Si el intervalo inicial cumple el teorema de Bolzano:
 - El método es convergente.
 - Es fácil estimar una cota del error absoluto.
 - Podemos saber a priori cuantas iteraciones tenemos que realizar.

Desventajas del método de bisección

- La velocidad de convergencia es lenta.
- Podemos estar cerca de la raíz y en la siguiente iteración alejarnos.

En la práctica, el método de bisección se utiliza para inicializar métodos más rápidos.

EJERCICIO

Sea la ecuación

$$\frac{x}{2}e^{-\frac{x}{2}} + \frac{1}{2} = 0$$

1. Demostrar que en $[-1, 1]$ existe una única raíz.

Podemos definir la función $f(x) = \frac{x}{2}e^{-\frac{x}{2}} + \frac{1}{2}$.

Vemos que se cumplen las tres condiciones suficientes para que exista una única raíz en el intervalo:

1. f es continua.
2. f tiene distinto signo en los extremos.
3. f es estrictamente creciente (o decreciente) en este intervalo.

Demostrémoslo también analíticamente:

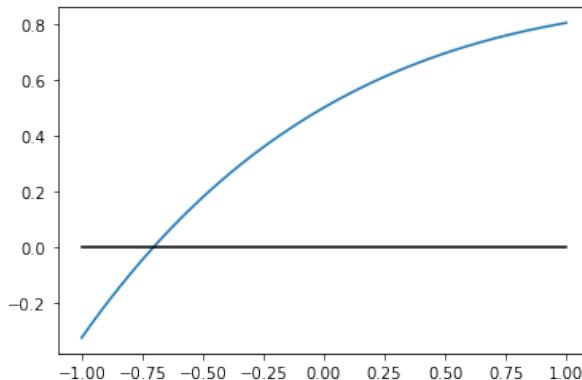
1. f es la suma y producto de funciones continuas, por lo tanto es continua.
2. $f(-1) = -0.3$ y $f(1) = 0.8$
3. Si derivamos la función f

$$f'(x) = \frac{1}{2}e^{-\frac{x}{2}} + \frac{x}{2}e^{-\frac{x}{2}} \frac{(-1)}{2} = \frac{2}{4}e^{-\frac{x}{2}} - \frac{x}{4}e^{-\frac{x}{2}} = \frac{1}{4}e^{-\frac{x}{2}}(2-x)$$

y entonces

$$f'(x) = \frac{1}{4}e^{-\frac{x}{2}}(2-x) = (+)(+) > 0$$

porque $(2-x)$ es positiva en $(-1, 1)$ y una función exponencial es siempre positiva. Por ello $f' > 0$ para todo valor del intervalo y f es estrictamente creciente en el intervalo.



2. ¿Se puede calcular por el método de bisección partiendo de dicho intervalo?

Las condiciones para aplicar el método de Bisección son las condiciones de Bolzano, es decir, las condiciones 1 y 2 del apartado anterior.

Como demostramos, se cumplen estas dos condiciones y por lo tanto, podemos aplicar el método de Bisección.

3. Aproximar la raíz haciendo cuatro iteraciones.

Hacemos las iteraciones teniendo en cuenta que dados a y b

$$m = \frac{a + b}{2}$$

y que la cota de error viene dada por

$$c = b - a$$

pero de la iteración siguiente.

Iteración 1

El intervalo es $[-1, 1]$ y su punto medio es

$$m = \frac{a + b}{2} = \frac{(-1) + 1}{2} = 0$$

Como $f(-1) = -0.32$, $f(0) = 0.50$ tienen signo distinto el siguiente intervalo es $[-1, 0]$ y la cota de error es la longitud de este nuevo intervalo, 1.

Iteración 2

El intervalo es $[-1, 0]$ y su punto medio es

$$m = \frac{a + b}{2} = \frac{(-1) + 0}{2} = -0.5$$

Como $f(-1) = -0.32$, $f(-0.5) = 0.18$ tienen signo distinto el siguiente intervalo es $[-1, -0.5]$ y la cota de error es la longitud de este nuevo intervalo, 0.5.

Iteración 3

El intervalo es $[-1, -0.5]$ y su punto medio es

$$m = \frac{a + b}{2} = \frac{(-1) + (-0.5)}{2} = -0.75$$

Como $f(-0.75) = -0.05$, $f(-0.5) = 0.18$ tienen signo distinto el siguiente intervalo es $[-0.75, -0.5]$ y la cota de error es la longitud de este nuevo intervalo, 0.25.

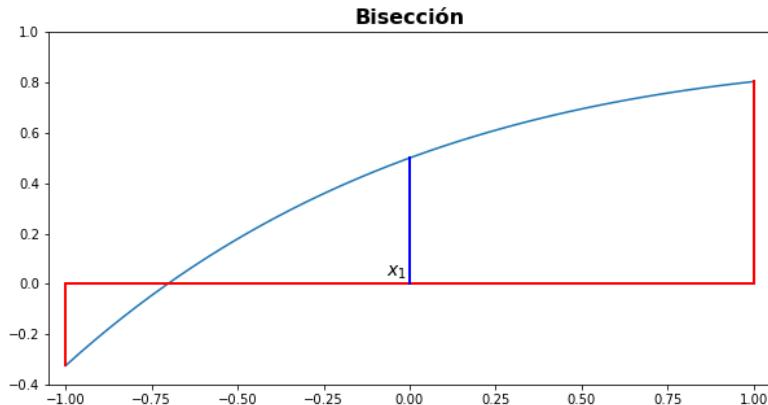
Iteración 4

El intervalo es $[-0.75, -0.5]$ y su punto medio es

$$m = \frac{a + b}{2} = \frac{(-1) + (-0.5)}{2} = -0.625$$

y la cota de error es la mitad de la longitud del intervalo anterior, es decir, 0.125.

| iteración | a | m | b | $f(a)$ | $f(m)$ | $f(b)$ | cota error |
|-----------|--------|--------|--------|--------|--------|--------|------------|
| 1 | -1.000 | 0.000 | 1.000 | -0.32 | 0.50 | 0.80 | 1.000 |
| 2 | -1.000 | -0.500 | 0.000 | -0.32 | 0.18 | 0.50 | 0.500 |
| 3 | -1.000 | -0.750 | -0.500 | -0.32 | -0.05 | 0.18 | 0.250 |
| 4 | -0.750 | -0.625 | -0.500 | -0.05 | | 0.18 | 0.125 |

**4. Dar una cota del error cometido al calcular esta raíz.**

Ya calculamos la cota de error conforme calculábamos las iteraciones, pero, teniendo en cuenta que el número de iteraciones es $k = 4$, también podíamos haber usado la fórmula

$$c_k = \frac{b_0 - a_0}{2^k} = \frac{1 - (-1)}{2^4} = \frac{2}{2^4} = \frac{1}{2^3} = \frac{1}{8} = 0.125$$

Comprobemos que el error es menor que la cota de error. Como la solución exacta es $\alpha = -0.703$, el error absoluto es

$$e_4 = |x_4 - \alpha| = |-0.625 - (-0.703)| = 0.078$$

que es menor que la cota de error, como era de esperar.

5. ¿Cuántas iteraciones n tendríamos que hacer para garantizar que el error es menor que 10^{-6} ?

Buscamos que

$$e_a = |\alpha - x_n| < 10^{-6}.$$

Como se verifica que el error (desconocido) es menor que la cota de error (conocida)

$$e_a = |\alpha - x_n| < \frac{b_0 - a_0}{2^n}.$$

una condición suficiente para que el error sea menor que 10^{-6} es que la cota de error sea menor que 10^{-6}

$$\frac{b_0 - a_0}{2^n} < 10^{-6}.$$

Trabajaremos con esta desigualdad y aplicaremos las siguientes propiedades

1. Si $a < b$ y $c > 0 \implies ac < bc$
2. Si f es una función estrictamente creciente se tiene que

$$x < y \implies f(x) < f(y)$$

$$3. \log A^B = B \log A$$

Teniendo en cuenta la propiedad 1 y multiplicando ambos miembros de la desigualdad primero por 2^n y luego 10^6 tenemos que

$$\frac{1 - (-1)}{2^n} < 10^{-6} \iff 2 < 2^n 10^{-6} \iff 2 \times 10^6 < 2^n$$

Como $f(x) = \log(x)$ es una función estrictamente creciente, aplicando la propiedad 2 se tiene que

$$\log(2 \times 10^6) < \log(2^n)$$

y teniendo en cuenta la propiedad 3

$$\log(2 \times 10^6) < n \log 2.$$

Como $\log 2 > 0$, aplicando la propiedad 1 con $c = 1 / \log 2$

$$\frac{\log(2 \times 10^6)}{\log 2} < n$$

log representa aquí un logaritmo en cualquier base. Usaremos, por ejemplo, logaritmos neperianos

$$20.9 < n$$

Y si hacemos $n = 21$ iteraciones, podemos garantizar que el error es menor que 10^{-6} .

Ejercicio 2.1.2

Sea la función

$$h(x) = (x^3 - x) e^{-x}$$

1. Demostrar que esta función tiene un único extremo en [3,4]
2. ¿Se puede calcular por el método de bisección partiendo de dicho intervalo?
3. Aproximar el extremo haciendo cuatro iteraciones.
4. Dar una cota del error cometido al calcular esta raíz.
5. ¿Cuántas iteraciones n tendríamos que hacer para garantizar que el error es menor que 10^{-8} ?

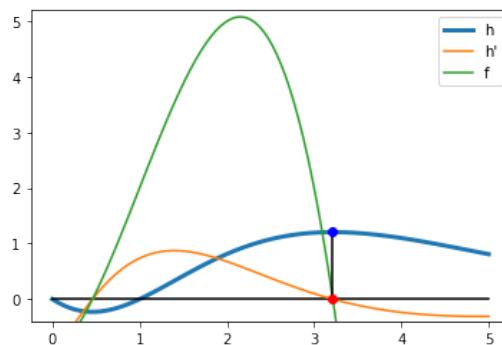
1. Demostrar que esta función tiene un único extremo en [3,4]

Si buscamos los extremos de h , como la condición necesaria de extremos para una función derivable es que $h'(x) = 0$, estamos buscando las raíces de h' .

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

Como el exponencial es siempre positivo y distinto de cero, si queremos las raíces de h' tenemos que buscar las raíces de

$$f(x) = -x^3 + 3x^2 + x - 1$$



Vemos que se cumplen las tres condiciones suficientes para que exista una única raíz en el intervalo:

1. f es continua.
2. f tiene distinto signo en los extremos.
3. f es estrictamente creciente (o decreciente) en este intervalo.

Demostrémoslo también analíticamente:

1. f es un polinomio y, por lo tanto, es continua.
2. $f(3) = 2$ y $f(4) = -13$
3. f es estrictamente decreciente en $[3, 4]$ porque $f' < 0$ en $(3, 4)$, ya que si factorizamos

$$f'(x) = -3x^2 + 6x + 1$$

Para ello calculamos las raíces de $ax^2 + bx + c = 0$ que son

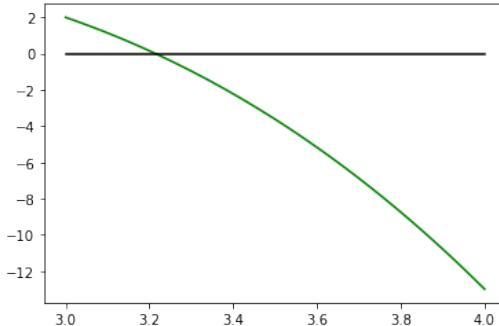
$$x_{1,2} = \frac{b \pm \sqrt{b^2 - 4ac}}{2a}$$

Como $a = -3$, $b = 6$ y $c = 1$

$$x_{1,2} = \frac{6 \pm \sqrt{6^2 - 4(-3)}}{2(-3)} \quad x_1 = -0.15 \quad x_2 = 2.15$$

Por lo tanto $f'(x) = a(x - x_1)(x - x_2) = -3(x + 0.15)(x - 2.2)$ y como el primer factor, -3 , es negativo, y el segundo, $(x + 0.15)$ y el tercero, $(x - 2.2)$ positivos en $(3, 4)$

$$f'(x) = -3(x + 0.15)(x - 2.2) = (-)(+)(+) < 0 \quad \text{en } (3, 4)$$



2. ¿Se puede calcular por el método de bisección partiendo de dicho intervalo?

Las condiciones para aplicar el método de Bisección son las condiciones de Bolzano, es decir, las condiciones 1 y 2 del apartado anterior.

Como demostramos, se cumplen estas dos condiciones y por lo tanto, podemos aplicar el método de Bisección.

3. Aproximar el extremo haciendo cuatro iteraciones.

Hacemos las iteraciones teniendo en cuenta que dados a y b

$$m = \frac{a + b}{2}$$

y que la cota de error viene dada por

$$c = b - a$$

pero de la iteración siguiente.

Iteración 1

El intervalo es $[3, 4]$ y su punto medio es

$$m = \frac{a + b}{2} = \frac{3 + 4}{2} = 3.5$$

Como $f(3) = 2, f(3.5) = -3.63$ tienen signo distinto el siguiente intervalo es $[3, 3.5]$ y la cota de error es la longitud de este nuevo intervalo, 0.5.

Iteración 2

El intervalo es $[3, 3.5]$ y su punto medio es

$$m = \frac{a + b}{2} = \frac{3 + 3.5}{2} = 3.25$$

Como $f(3) = 2, f(3.25) = -0.39$ tienen signo distinto el siguiente intervalo es $[3, 3.25]$ y la cota de error es la longitud de este nuevo intervalo, 0.25.

Iteración 3

El intervalo es $[3, 3.25]$ y su punto medio es

$$m = \frac{a + b}{2} = \frac{3 + 3.25}{2} = 3.125$$

Como $f(3.125) = 0.9, f(3.25) = -0.39$ tienen signo distinto el siguiente intervalo es $[3.125, 3.25]$ y la cota de error es la longitud de este nuevo intervalo, 0.125.

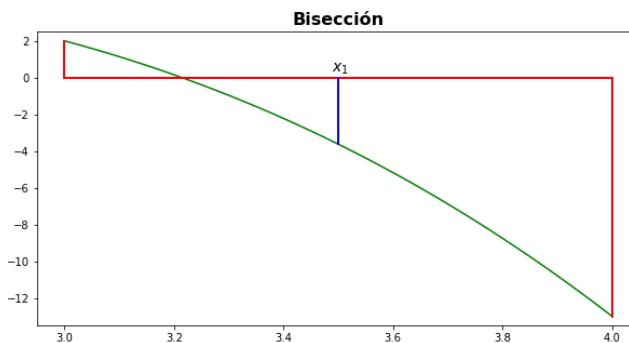
Iteración 4

El intervalo es $[3.125, 3.25]$ y su punto medio es

$$m = \frac{a + b}{2} = \frac{3.125 + 3.25}{2} = 3.1875$$

y la cota de error es la mitad de la longitud del intervalo anterior, es decir, 0.0625.

| iteración | a | m | b | $f(a)$ | $f(m)$ | $f(b)$ | cota error |
|-----------|-------|--------|-------|--------|--------|--------|------------|
| 1 | 3.000 | 3.5000 | 4.000 | 2.00 | -3.63 | -13.00 | 0.500 |
| 2 | 3.000 | 3.2500 | 3.500 | 2.00 | -0.39 | -3.62 | 0.250 |
| 3 | 3.000 | 3.1250 | 3.250 | 2.00 | 0.90 | -0.39 | 0.125 |
| 4 | 3.125 | 3.1875 | 3.250 | 0.90 | | -0.39 | 0.0625 |



4. Dar una cota del error cometido al calcular esta raíz.

Ya calculamos la cota de error conforme calculábamos las iteraciones, pero, teniendo en cuenta que el número de iteraciones es $k = 4$, también podíamos haber usado la fórmula

$$c_k = \frac{b_0 - a_0}{2^k} = \frac{4 - 3}{2^4} = \frac{1}{2^4} = \frac{1}{2^4} = \frac{1}{16} = 0.0625$$

Comprobemos que el error es menor que la cota de error. Como la solución exacta es $\alpha = 3.2143$, el error absoluto es

$$e_4 = |x_4 - \alpha| = |3.1875 - 3.2143| = 0.0268$$

que es menor que la cota de error, como era de esperar.

5. ¿Cuántas iteraciones n tendríamos que hacer para garantizar que el error es menor que 10^{-8} ?

Buscamos que

$$e_n = |\alpha - x_n| < 10^{-8}.$$

Como se verifica que el error (desconocido) es menor que la cota de error (conocida)

$$e_n = |\alpha - x_n| < \frac{b_0 - a_0}{2^n}.$$

una condición suficiente para que el error sea menor que 10^{-8} es que la cota de error sea menor que 10^{-8}

$$\frac{b_0 - a_0}{2^n} < 10^{-8}.$$

Trabajaremos con esta desigualdad y aplicaremos las siguientes propiedades

1. Si $a < b$ y $c > 0 \implies ac < bc$

2. Si f es una función estrictamente creciente se tiene que

$$x < y \implies f(x) < f(y)$$

3. $\log A^B = B \log A$

Teniendo en cuenta la propiedad 1 y multiplicando ambos miembros de la desigualdad primero por 2^n y luego 10^8 tenemos que

$$\frac{4 - 3}{2^n} < 10^{-8} \iff 1 < 2^n 10^{-8} \iff 10^8 < 2^n$$

Como $f(x) = \log(x)$ es una función estrictamente creciente, aplicando la propiedad 2 se tiene que

$$\log(10^8) < \log(2^n)$$

y teniendo en cuenta la propiedad 3

$$\log(10^8) < n \log 2.$$

Como $\log 2 > 0$, aplicando la propiedad 1 con $c = 1 / \log 2$

$$\frac{\log(10^8)}{\log 2} < n$$

\log representa aquí un logaritmo en cualquier base. Usaremos, por ejemplo, logaritmos neperianos

$$26.6 < n$$

Y si hacemos $n = 27$ iteraciones, podemos garantizar que el error es menor que 10^{-8} .

2.2 Método de Newton-Raphson

Ejercicio 2.2.1

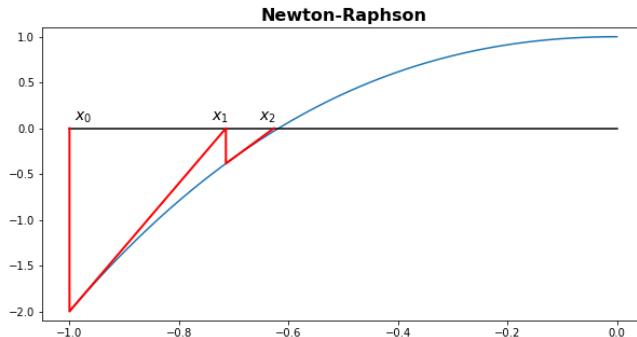
Aproximar utilizando el método de Newton $r = \sqrt{3}$. Utilizar como punto inicial $x_0 = 1$, realizar tres iteraciones y calcular el residuo. Calcular el error absoluto de la aproximación.

INTRODUCCIÓN

Método de Newton-Raphson

El método de Newton se puede explicar como un método iterativo donde

- La curva se sustituye por su recta tangente en un punto.
- Se approxima la raíz de la curva con la raíz de la recta.
- Tomamos la raíz de la recta tangente en este punto como aproximación de la raíz.
- Repetimos hasta que se cumpla una condición de parada.



La ecuación de la recta tangente a la curva f en el punto x_0 viene dada por

$$y = f'(x_0)(x - x_0) + f(x_0)$$

La raíz de esta recta es x_1 y la obtenemos haciendo $y = 0$ en esta ecuación, que nos da el punto donde la recta corta al eje OX .

$$0 = f'(x_0)(x_1 - x_0) + f(x_0)$$

Si despejamos x_1

$$-f(x_0) = f'(x_0)(x_1 - x_0) \Rightarrow -\frac{f(x_0)}{f'(x_0)} = x_1 - x_0$$

y

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Análogamente, para obtener x_2

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

y así sucesivamente.

Algoritmo

- Sea x_0 un punto inicial.
- Para $k = 1, 2, \dots, \text{MaxNumIter}$:
 - Calcular $x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})}$
 - Si x_k satisface el criterio de parada, parar.
 - En el caso contrario, hacer otra iteración.

Convergencia

- En general, el método sólo converge si el punto inicial, x_0 está **suficientemente próximo** a la raíz, α . En la práctica, se usa el método de bisección unas cuantas veces, o se selecciona sobre una gráfica de f para inicializar el algoritmo.
- En general, aunque no siempre, el método de Newton tiene orden de convergencia 2.

Ventajas

- Es un algoritmo que converge rápidamente.
- Es sencillo de programar.

Inconvenientes

- Dependiendo del punto inicial, a veces converge, a veces no.
- Necesitamos la derivada exacta de la función, que no siempre es posible suministrar.

EJERCICIO

Nuestra ecuación es

$$x = \sqrt{3} \Rightarrow x^2 = 3 \Rightarrow x^2 - 3 = 0$$

Por lo tanto $f(x) = x^2 - 3$ y $f'(x) = 2x$

La función de iteración viene dada por

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

En este caso

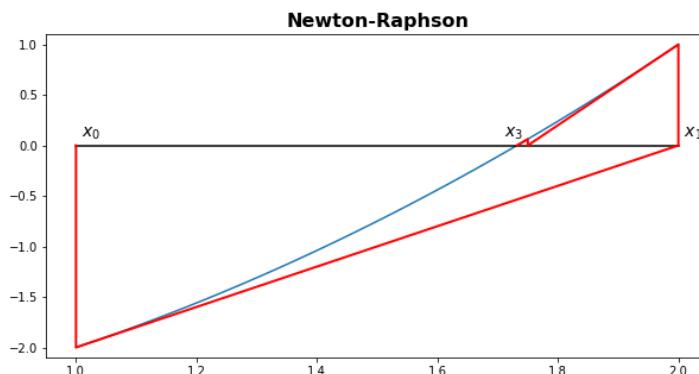
$$x_{k+1} = x_k - \frac{x_k^2 - 3}{2x_k} = \frac{2x_k^2 - x_k^2 + 3}{2x_k} = \frac{x_k^2 + 3}{2x_k}$$

Es decir

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

Iteraciones

- $x_0 = 1, x_1 = \frac{x_0^2 + 3}{2x_0} = \frac{1^2 + 3}{2} = 2$
- $x_1 = 2, x_2 = \frac{x_1^2 + 3}{2x_1} = \frac{2^2 + 3}{2(2)} = 1.75$
- $x_2 = 1.75, x_3 = \frac{x_2^2 + 3}{2x_2} = \frac{1.75^2 + 3}{2(1.75)} = 1.732143$



El valor de la función en la raíz es cero. El valor de la función en la aproximación se llama **residuo** y, si es una buena aproximación, es un número pequeño

$$r = |f(x_3)| = |1.732143^2 - 3| = 0.0003$$

Además, el residuo en el paso k , es conocido, mientras que el error es desconocido (necesitamos la raíz exacta, que es lo que estamos buscando).

Como $\alpha = 1.732051$ el error absoluto es

$$e_a = |x_3 - \alpha| = |1.732143 - 1.732051| = 0.000092 \approx 0.0001$$

Ejercicio 2.2.2

Sea la función

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

1. Demostrar que esta función tiene un único extremo en $[1, 2]$.
2. Aproximar el extremo utilizando el método de Newton.
Utilizar como punto inicial $x_0 = 1$ y realizar 4 iteraciones.
3. Calcular el residuo.

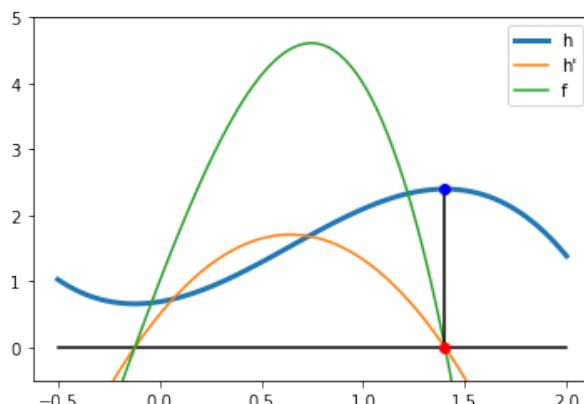
1. Demostrar que esta función tiene un único extremo en $[1, 2]$

Si buscamos los extremos de h , como la condición necesaria de extremos para una función derivable es que $h'(x) = 0$, estamos buscando las raíces de h' .

$$h'(x) = 4x - 3x^2 + \frac{1}{2+x} = \frac{(4x - 3x^2)(2+x) + 1}{2+x} = \frac{-3x^3 - 2x^2 + 8x + 1}{2+x}$$

Las raíces de h' son las raíces de f , donde

$$f(x) = -3x^3 - 2x^2 + 8x + 1$$



$$f(x) = -3x^3 - 2x^2 + 8x + 1$$

Vemos que se cumplen las tres condiciones suficientes para que exista una única raíz en el intervalo:

1. f es continua.
2. f tiene distinto signo en los extremos.
3. f es estrictamente creciente (o decreciente) en este intervalo.

Demostrémos que se cumplen:

1. f es un polinomio y, por lo tanto, es continua.
2. $f(1) = 4$ y $f(2) = -15$
3. f es estrictamente decreciente en $[1,2]$ porque $f' < 0$ en $(1,2)$, ya que si factorizamos

$$f'(x) = -9x^2 - 4x + 8$$

Para ello calculamos las raíces del polinomio $ax^2 + bx + c = 0$

$$x_{1,2} = \frac{b \pm \sqrt{b^2 - 4ac}}{2a}$$

En este caso, $a = -9$, $b = -4$ y $c = 8$

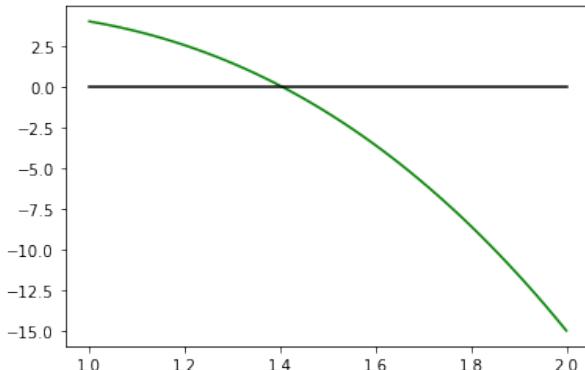
$$x_{1,2} = \frac{-4 \pm \sqrt{(-4)^2 - 4(-9)(8)}}{2(-9)} \quad x_1 = -1.19 \quad x_2 = 0.75$$

Por lo tanto

$$f'(x) = a(x - x_1)(x - x_2) = -9(x + 1.19)(x - 0.75)$$

y como el primer factor, -9 es negativo, y el segundo, $(x + 1.2)$, y el tercero, $(x - 0.75)$, positivos en $(1,2)$

$$f'(x) = -9(x + 1.2)(x - 0.74) = (-)(+)(+) < 0 \quad \text{en } (1,2)$$



2. Aproximar el extremo utilizando el método de Newton. Utilizar como punto inicial $x_0 = 1$ y realizar 4 iteraciones.

La función de iteración viene dada por

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

En este caso

$$f(x) = -3x^3 - 2x^2 + 8x + 1 \quad f'(x) = -9x^2 - 4x + 8$$

Es decir

$$x_{k+1} = x_k - \frac{-3x_k^3 - 2x_k^2 + 8x_k + 1}{-9x_k^2 - 4x_k + 8}$$

Iteraciones

- $x_0 = 1$,

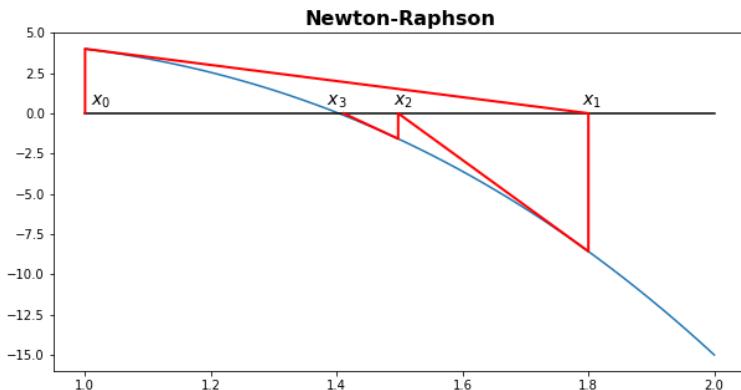
$$x_1 = 1 - \frac{-3(1)^3 - 2(1)^2 + 8(1) + 1}{-9(1)^2 - 4(1) + 8} = 1 - \frac{-3 - 2 + 8 + 1}{-9 - 4 + 8} = 1 + \frac{4}{5} = 1.8$$

- $x_1 = 1.8$,

$$x_2 = 1.8 - \frac{-3(1.8)^3 - 2(1.8)^2 + 8(1.8) + 1}{-9(1.8)^2 - 4(1.8) + 8} = 1.497602$$

Y así sucesivamente

| k | x_k |
|-----|----------|
| 0 | 1.000000 |
| 1 | 1.800000 |
| 2 | 1.497602 |
| 3 | 1.410600 |
| 4 | 1.403193 |



4. Calcular el residuo

El valor de la función en la raíz es cero. El valor de la función en la aproximación se llama **residuo** y, si es una buena aproximación, es un número pequeño

$$r = |f(x_4)| = 0.0008$$

Además, el residuo en el paso k , es conocido, mientras que el error es desconocido (necesitamos la raíz exacta, que es lo que estamos buscando).

Como $\alpha = 1.403140$ el error absoluto es

$$e_a = |x_4 - \alpha| = |1.403193 - 1.403140| = 0.00005$$

2.3 Método de la secante

Ejercicio 2.3.1

1. Aproximar utilizando el método de la secante. Utilizar como puntos iniciales $x_0 = 1$ y $x_1 = 2$. Realizar 3 iteraciones.
2. Calcular el residuo.
3. Usar la calculadora para estimar el error absoluto de la aproximación.

INTRODUCCIÓN

Método de la secante

El método de la secante se puede considerar una variante del método de Newton en el que se sustituye la derivada por una aproximación.

Podemos definir la derivada de una función f en un punto a

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}$$

Y si tomamos $x = a + h$ entonces $h = x - a$

$$f'(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}$$

o también

$$f'(a) = \lim_{x \rightarrow a} \frac{f(a) - f(x)}{a - x}$$

si $a = x_{k-1}$ y $x = x_{k-2}$

$$f'(x_{k-1}) = \lim_{x_{k-2} \rightarrow x_{k-1}} \frac{f(x_{k-1}) - f(x_{k-2})}{x_{k-1} - x_{k-2}}$$

Y quitando el límite obtenemos la fórmula de diferencias divididas

$$f'(x_{k-1}) \approx \frac{f(x_{k-1}) - f(x_{k-2})}{x_{k-1} - x_{k-2}}$$

Y si en la fórmula del método de Newton

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

sustituimos la derivada por su aproximación

$$x_k = x_{k-1} - f(x_{k-1}) \frac{x_{k-1} - x_{k-2}}{f(x_{k-1}) - f(x_{k-2})}$$

Gráficamente, el método de la secante se puede explicar como un método iterativo donde

- La curva se sustituye por una recta que pasa por dos de sus puntos (recta secante)
- Se aproxima la raíz de la curva con la raíz de la recta.
- Se descarta el punto más antiguo (¡importante!) y se traza una nueva secante con el punto que nos queda más el último punto calculado.
- Repetimos hasta que se cumpla una condición de parada.

Es decir, empezamos con

- x_0 y x_1 y obtenemos x_2 .

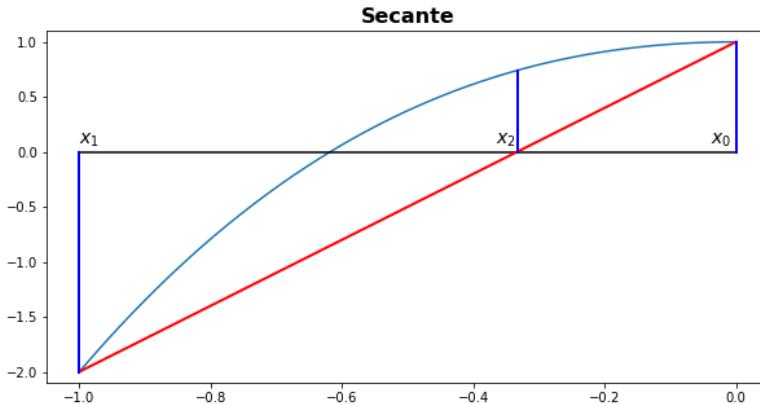
Seguimos con

- x_1 y x_2 y obtenemos x_3 .

y así sucesivamente.

Algoritmo

- Sean x_0 y x_1 los puntos iniciales.
- Para $k = 2, 3, \dots, \text{MaxNumIter}$:
 - Calcular $x_k = x_{k-1} - f(x_{k-1}) \frac{x_{k-1} - x_{k-2}}{f(x_{k-1}) - f(x_{k-2})}$
 - Si x_k satisface el criterio de parada, parar.
 - En el caso contrario, hacer otra iteración.



Inconvenientes

- Necesitamos dos puntos iniciales.
- En general, converge más lentamente que Newton (pero más rápidamente que bisección).
- Si no seleccionamos bien los puntos iniciales, puede no converger.

Ventajas

- No necesitamos conocer la derivada exacta, solo el valor de la función.
- La velocidad de convergencia es relativamente rápida. El orden de convergencia es $p \approx 1.6$

EJERCICIO

1. Aproximar utilizando el método de la secante $r = \sqrt{3}$. Utilizar como puntos iniciales $x_0 = 1$ y $x_1 = 2$. Realizar 3 iteraciones.

Nuestra ecuación es

$$x = \sqrt{3} \Rightarrow x^2 = 3 \Rightarrow x^2 - 3 = 0$$

Por lo tanto $f(x) = x^2 - 3$

La función de iteración viene dada por

$$x_k = x_{k-1} - f(x_{k-1}) \frac{x_{k-1} - x_{k-2}}{f(x_{k-1}) - f(x_{k-2})}$$

En este caso

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

Iteraciones

- $x_0 = 1, x_1 = 2$

$$x_2 = x_1 - (x_1^2 - 3) \frac{x_1 - x_0}{(x_1^2 - 3) - (x_0^2 - 3)} = 2 - (2^2 - 3) \frac{2 - 1}{(2^2 - 3) - (1^2 - 3)}$$

$$x_2 = 2 - (1) \frac{1}{(1) - (-2)} = 2 - \frac{1}{3} = 1.666667$$

- $x_1 = 2, x_2 = 1.666667$,

$$x_3 = x_2 - (x_2^2 - 3) \frac{x_2 - x_1}{(x_2^2 - 3) - (x_1^2 - 3)}$$

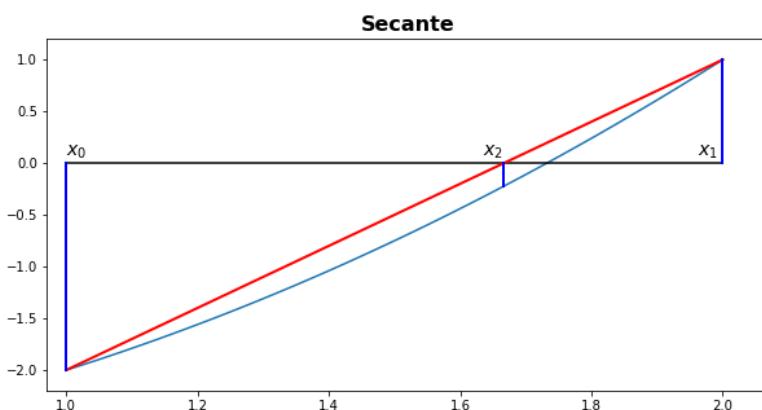
$$x_3 = 1.666667 - (1.666667^2 - 3) \frac{1.666667 - 2}{(1.666667^2 - 3) - (2^2 - 3)} = 1.727273$$

- $x_2 = 1.666667, x_3 = 1.727273$,

$$x_4 = x_3 - (x_3^2 - 3) \frac{x_3 - x_2}{(x_3^2 - 3) - (x_2^2 - 3)}$$

$$x_4 = 1.727273 - (1.727273^2 - 3) \frac{1.727273 - 1.666667}{(1.727273^2 - 3) - (1.666667^2 - 3)} = 1.732143$$

| k | x_k |
|-----|----------|
| 0 | 1.000000 |
| 1 | 2.000000 |
| 2 | 1.666667 |
| 3 | 1.727273 |
| 4 | 1.732143 |



2. Calcular el residuo.

El valor de la función en la raíz es cero. El valor de la función en la aproximación se llama **residuo** y, si es una buena aproximación, es un número pequeño

$$r = |f(x_3)| = |1.732143^2 - 3| = 0.0003$$

El residuo en el paso k , es conocido, mientras que el error es desconocido (necesitamos la raíz exacta, que es lo que estamos buscando).

3. Usar la calculadora para estimar el error absoluto de la aproximación.

Como $\alpha = 1.732051$ el error absoluto es

$$e_a = |x_3 - \alpha| = |1.732143 - 1.732051| = 0.000092 \approx 0.0001$$

2.4 Método de Regula-Falsi

Ejercicio 2.4.1

1. Aproximar utilizando el método de la Regula-Falsi $r = \sqrt{3}$. Utilizar el intervalo inicial $[1, 2]$. Realizar 3 iteraciones.
2. Calcular el residuo.
3. Dar una cota del error cometido al calcular esta raíz. ¿Es una buena cota? Por qué?

INTRODUCCIÓN

Método de Regula-Falsi

El método de la Regula-Falsi se puede considerar un híbrido del método de Bisección con el método de la Secante.

- Empezamos con una función f , y dos puntos iniciales a y b que cumplan las condiciones del teorema de Bolzano, como en Bisección.
- En cada iteración calculamos un punto nuevo c con la misma fórmula que la Secante, a partir de los dos puntos anteriores.
- Descartamos uno de los dos puntos de la iteración anterior usando los mismos criterios que para Bisección, es decir, nos quedamos con los dos puntos para los cuales la función cumple el teorema de Bolzano.

La fórmula para calcular la siguiente iteración con el método de la secante es

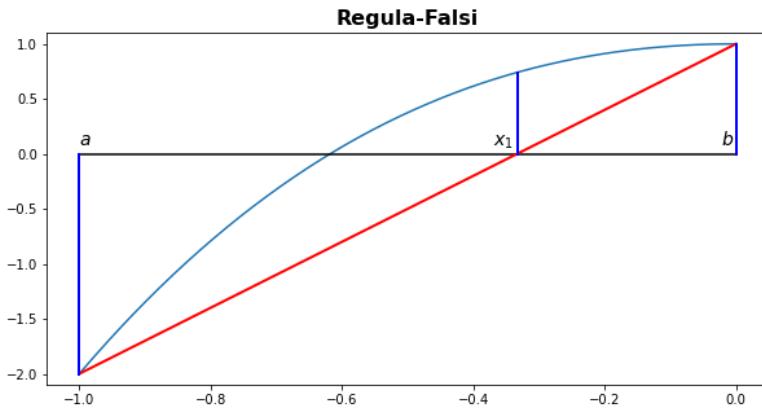
$$x_k = x_{k-1} - f(x_{k-1}) \frac{x_{k-1} - x_{k-2}}{f(x_{k-1}) - f(x_{k-2})}$$

En el método de la Regula-Falsi, el punto siguiente es

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

Algoritmo

- Sea $a_1 = a$, $b_1 = b$.
- Para $k = 1, 2, \dots, \text{MaxNumIter}$
 - Calcular el punto $x_k = \frac{a_k f(b_k) - b_k f(a_k)}{f(b_k) - f(a_k)}$.
 - Si x_k satisface el criterio de parada, parar.
 - En el caso contrario,
 - * si $f(a_k)f(x_k) < 0$ entonces: $a_{k+1} = a_k$, $b_{k+1} = x_k$,
 - * si $f(x_k)f(b_k) < 0$ entonces: $a_{k+1} = x_k$, $b_{k+1} = b_k$,
 - * en otro caso acabar.



Inconvenientes

- En general, converge más lentamente que el método de la Secante.
- El intervalo que contiene la raíz no siempre disminuye de longitud de forma significativa con las iteraciones, como en Bisección. Por lo tanto, su longitud no es una buena cota de error.

Ventajas

- A diferencia del método de la secante, para el método de Regula-Falsi, la convergencia está garantizada.
- En general, converge más rápidamente que el método de Bisección, aunque el orden de convergencias de ambos métodos es lineal.

EJERCICIO

1. Aproximar utilizando el método de la Regula-Falsi $r = \sqrt{3}$. Utilizar el intervalo inicial $[1, 2]$. Realizar 3 iteraciones.

Nuestra ecuación es

$$x = \sqrt{3} \Rightarrow x^2 = 3 \Rightarrow x^2 - 3 = 0$$

Por lo tanto $f(x) = x^2 - 3$

La sucesión que genera regula-falsi utiliza la fórmula

$$x_k = \frac{a_k f(b_k) - b_k f(a_k)}{f(b_k) - f(a_k)}$$

Iteraciones

Iteración 1

El intervalo es $[1, 2]$ y el punto siguiente es

$$c = \frac{af(b) - bf(a)}{f(b) - f(a)} = \frac{1(2^2 - 3) - 2(1^2 - 3)}{(2^2 - 3) - (1^2 - 3)} = \frac{1 + 4}{1 + 2} = \frac{5}{3} = 1.666667$$

Como $f(1.666667) = -0.22$, $f(2) = 1$ tienen signo distinto el siguiente intervalo es $[1.666667, 2]$.

Iteración 2

El intervalo es $[1.666667, 2]$ y su punto medio es

$$c = \frac{1.666667(2^2 - 3) - 2(1.666667^2 - 3)}{(2^2 - 3) - (1.666667^2 - 3)} = 1.727273$$

Como $f(1.727273) = -0.02$, $f(2) = 1$ tienen signo distinto el siguiente intervalo es $[1.727273, 2]$.

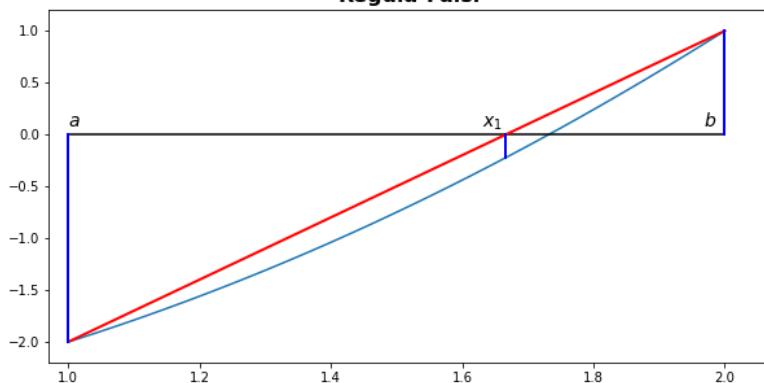
Iteración 3

El intervalo es $[1.727273, 2]$ y el siguiente punto intermedio es

$$c = \frac{1.727273(2^2 - 3) - 2(1.727273^2 - 3)}{(2^2 - 3) - (1.727273^2 - 3)} = 1.731707$$

| iteración | a | c | b | $f(a)$ | $f(c)$ | $f(b)$ | cota error |
|-----------|----------|----------|-----|--------|--------|--------|------------|
| 1 | 1.000000 | 1.666667 | 2.0 | -2.00 | -0.222 | 1.0 | 0.333333 |
| 2 | 1.666667 | 1.727273 | 2.0 | -0.22 | -0.017 | 1.0 | 0.272727 |
| 3 | 1.727273 | 1.731707 | 2.0 | -0.02 | -0.001 | 1.0 | 0.268293 |

Regula-Falsi



2. Calcular el residuo

El valor de la función en la raíz es cero. El valor de la función en la aproximación se llama **residuo** y, si es una buena aproximación, es un número pequeño

$$r = |f(x_3)| = |1.731707^2 - 3| = 0.001$$

El residuo en el paso k , es conocido, mientras que el error es desconocido (necesitamos la raíz exacta, que es lo que estamos buscando).

Como $\alpha = 1.732051$ el error absoluto es

$$e_a = |x_3 - \alpha| = |1.731707 - 1.732051| = 0.0003$$

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

En la tabla hemos calculado una cota del error cometido. Esta, como en el método de Bisección, viene dada por la longitud del último intervalo. Pero en el caso de la regula-falsi, no es muy útil porque, en muchos casos (como este) la longitud del intervalo no disminuye significativamente conforme nos acercamos a la raíz porque el método crea una sucesión que, a partir de cierto momento, se acerca a la raíz siempre desde la izquierda o siempre desde la derecha.

Así, la cota final es aproximadamente 0.28 que es muy distinta del error 0.0003 y, por lo tanto, no aporta información relevante.

Teniendo en cuenta que la raíz que buscamos es $\alpha = 1.732051$ comparemos los tres métodos que hemos usado para resolver este problema

| k | Newton | Secante | Regula – falsi |
|-----|----------|----------|----------------|
| 1 | 2.000000 | 1.666667 | 1.666667 |
| 2 | 1.750000 | 1.727273 | 1.727273 |
| 3 | 1.732143 | 1.732143 | 1.731707 |

En este caso, hemos obtenido los mismos resultados con los métodos de Newton y la Secante. Regula-Falsi obtiene unos resultados un poco peores pero bastante buenos.

2.5 Método de punto fijo

Ejercicio 2.5.1

Para calcular las raíces de $f(x) = x + \ln(x)$ por el método de punto fijo se definen las siguientes funciones de iteración.

$$(i) g_1(x) = -\ln(x), \quad (ii) g_2(x) = e^{-x}, \quad (iii) g_3(x) = \frac{x + e^{-x}}{2}$$

1. Demostrar que la ecuación $f(x) = 0$ tiene la misma raíz que $g_i(x) = x$ con $i = 1, 2, 3$.
2. Estudiar gráficamente si se cumplen las condiciones del teorema de la aplicación contractiva en el intervalo $[0, 1]$
3. Realizar nueve iteraciones con cada uno de las funciones de iteración utilizando como punto inicial $x_0 = 0.5$
4. ¿Qué funciones pueden usarse? ¿Qué función de iteración debería usarse?

INTRODUCCIÓN

El método de punto fijo

Decimos que α es un **punto fijo** de g si

$$g(\alpha) = \alpha$$

Como cualquier ecuación $f(x) = 0$ puede ser reescrita como $g(x) = x$, por ejemplo, $g(x) = f(x) + x$, resolver la primera ecuación es equivalente a encontrar un punto fijo de la segunda función.

Por ejemplo, sea $f(x) = \cos(x) - x$ y buscamos las raíces de f tal que

$$\cos(x) - x = 0.$$

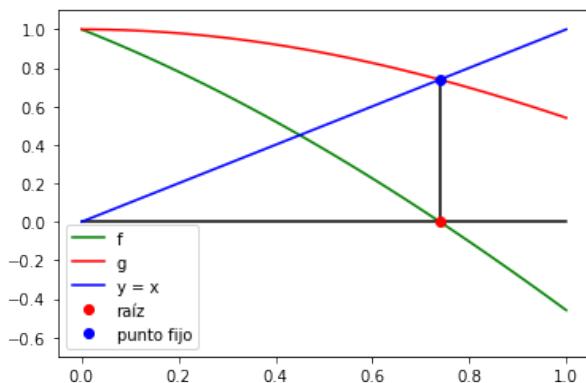
Podemos reorganizar la ecuación como

$$\cos(x) = x.$$

La solución de esta segunda ecuación, también llamada punto fijo de la función

$$g(x) = \cos(x)$$

será también una raíz de f .



Una vez tenemos la función g el procedimiento para calcular el punto fijo es:

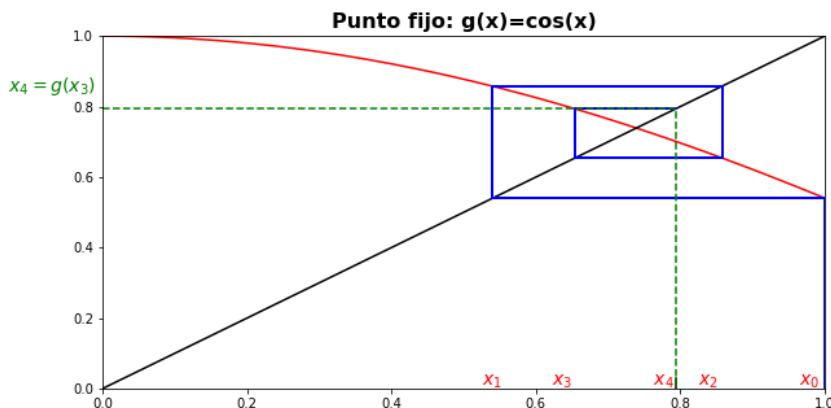
- Tomamos un x_0
- Iteración 1: $x_1 = g(x_0)$
- Iteración 2: $x_2 = g(x_1)$
- Iteración 3: $x_3 = g(x_2)$
- ...

Algoritmo

- Sea x_0 un punto inicial.
- Para $k = 1, 2, \dots, \text{MaxNumIter}$:
 - Calcular $x_k = g(x_{k-1})$
 - Si x_k satisface el criterio de parada, parar.
 - En el caso contrario, hacer otra iteración.

Por ejemplo, si para $g(x) = \cos(x)$ tomamos $x_0 = 1$ (¡Ojo! Las unidades de ángulos en matemáticas y en física son los radianes):

- Iteración 1: $x_1 = \cos(1) = 0.54$
- Iteración 2: $x_2 = \cos(0.54) = 0.86$
- Iteración 3: $x_3 = \cos(0.86) = 0.65$
- ...



Gráficamente la sucesión se construye:

- Dibujamos la curva g (en rojo en la figura).
- Dibujamos la recta $y = x$ (en negro en la figura). Estamos buscando el punto fijo $g(x) = x$, es decir, la intersección de la curva con la recta.
- Obtenemos el punto $(x_0, f(x_0))$ sobre la curva.
- Trazamos la línea horizontal hasta la recta $y = x$. Como está a la misma altura que el punto de la curva, la altura del punto sobre la recta, es decir, la coordenada y es igual a $x_1 = f(x_0)$. Como para los puntos de la recta $y = x$, la coordenada x es la misma, es decir, x_1 .
- Trazamos la línea vertical hasta la curva y obtenemos el punto $(x_1, f(x_1))$ sobre la curva.
- Trazamos la línea horizontal hasta la recta $y = x$. Como está a la misma altura que el punto de la curva, la altura del punto sobre la recta, es decir, la coordenada y es igual a $x_2 = f(x_1)$. Como para los puntos de la recta $y = x$, la coordenada x es la misma, es decir, x_2 .
- Y seguimos así, curva, recta, curva, recta, ...

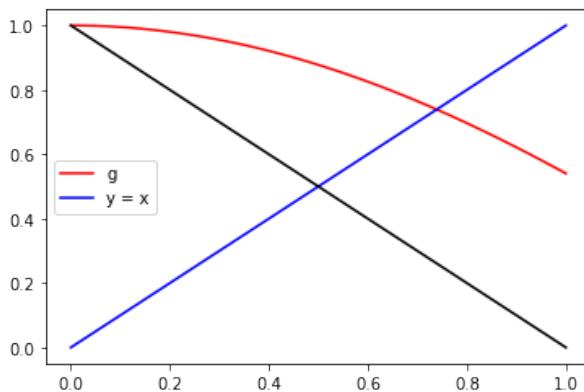
Convergencia

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

- $x \in [a, b] \Rightarrow g(x) \in [a, b]$
- $|g'(x)| \leq k < 1$ para todo $x \in [a, b]$

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

Veamos si se verifican estas condiciones para la función $g(x) = \cos x$ en el intervalo $[0, 1]$



Gráficamente

- Si nos fijamos en el eje OY, vemos que la gráfica de g siempre está contenida entre 0 y 1. Por lo tanto se cumple la primera condición.
- La pendiente de la curva es siempre menor que 1 en valor absoluto en el intervalo $[0, 1]$. La referencia son las diagonales: si la curva es menos pendiente (es más planas o está más tumbada) que la diagonal correspondiente (en este caso la negra, que tiene pendiente -1), cumple la segunda condición.

Analíticamente

- Como $\cos(x)$ es decreciente en $\left[0, \frac{\pi}{2}\right]$ y $[0, 1] \subset \left[0, \frac{\pi}{2}\right] = [0, 1.57]$ entonces $\cos(x)$ es decreciente en $[0, 1]$ y tendrá su máximo en 0 y su mínimo en 1, es decir

$$\cos(1) \leq \cos(x) \leq \cos(0)$$

que es

$$0.54 \leq \cos(x) \leq 1$$

y por lo tanto

$$\text{Si } x \in [0, 1] \implies \cos(x) \in [0.54, 1] \subset [0, 1]$$

- $g'(x) = -\operatorname{sen}(x)$ y $|g'(x)| = \operatorname{sen}(x)$ en $[0, 1]$. Como el $\operatorname{sen}(x)$ es creciente en $[0, 1]$ tendrá su máximo en 1 y su mínimo en 0, es decir

$$\operatorname{sen}(0) \leq \operatorname{sen}(x) \leq \operatorname{sen}(1)$$

que es

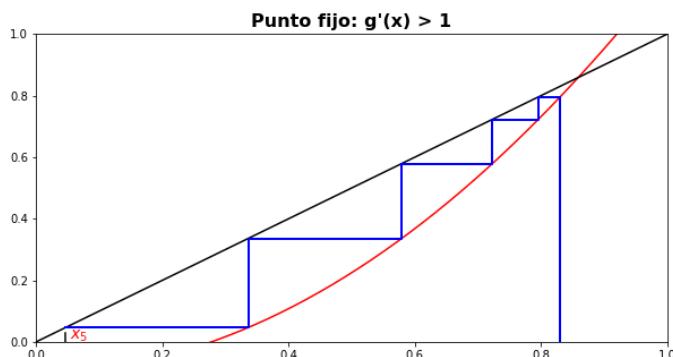
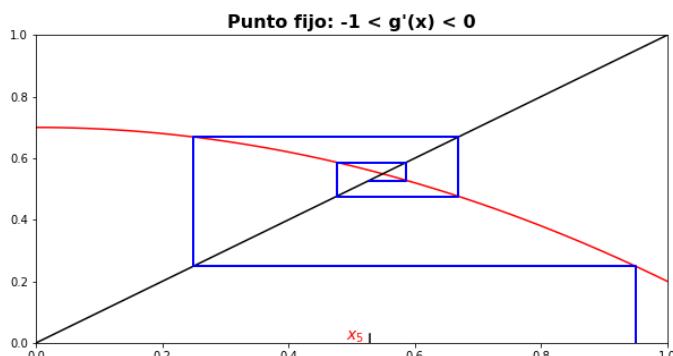
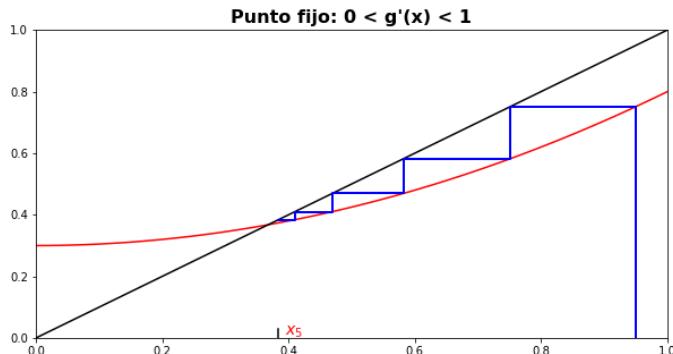
$$0 \leq \operatorname{sen}(x) \leq 0.84$$

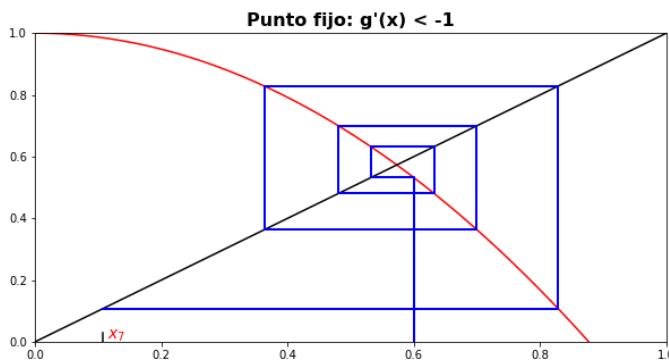
y por lo tanto

$$|g'(x)| \leq 0.84 < 1 \text{ para todo } x \in [0, 1]$$

Como se cumplen las condiciones del teorema de la aplicación contractiva podemos escoger cualquier punto del intervalo $[0,1]$ como valor x_0 y está garantizado que la sucesión generada con la función de iteración g converja.

Existen cuatro posibles casos:





Ventajas

- El método de punto fijo es muy sencillo de implementar.
- Permite un estudio teórico sistematizado de la convergencia.
- Es un conjunto de métodos: por ejemplo, Newton-Raphson es un método de punto fijo donde $g(x) = x - f(x)/f'(x)$.
- Se pueden construir métodos de punto fijo del orden de convergencia deseado.

Inconvenientes

- Métodos con órdenes de convergencia altos requieren la existencia y buen comportamiento de las derivadas de g .
- Si no usamos las derivadas el orden de convergencia es lineal y converge lentamente.

EJERCICIO

Para calcular las raíces de $f(x) = x + \ln(x)$ por el método de punto fijo se definen las siguientes funciones de iteración.

$$(i) g_1(x) = -\ln(x), \quad (ii) g_2(x) = e^{-x}, \quad (iii) g_3(x) = \frac{x + e^{-x}}{2}$$

1. Demostrar que la ecuación $f(x) = 0$ tiene la misma raíz que $g_i(x) = x$ con $i = 1, 2, 3$

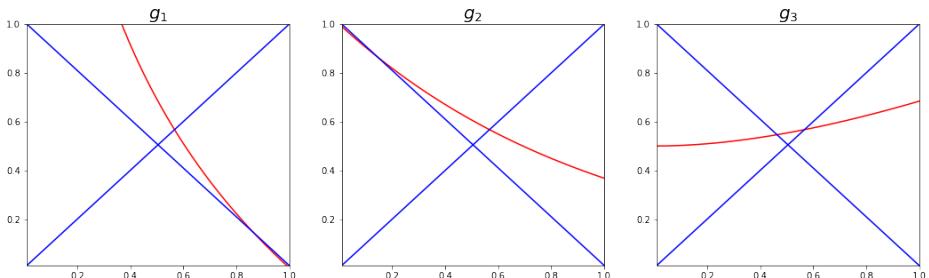
Comprobemos que la raíz de la ecuación es punto fijo de las funciones que se definen a partir de las ecuaciones equivalentes. Es decir

$$f(x) = 0 \iff g(x) = x$$

Empecemos por la función de iteración

- $g_1(x) = -\ln x$
 $x = -\ln x \iff x + \ln x = 0 \iff f(x) = x + \ln x$
- $g_2(x) = e^{-x}$
 $x = e^{-x} \iff \ln x = \ln e^{-x} \iff \ln x = -x \iff \ln x + x = 0 \iff f(x) = x + \ln x$
- $g_3(x) = \frac{x + e^{-x}}{2}$
 $x = \frac{x + e^{-x}}{2} \iff 2x = x + e^{-x} \iff x = e^{-x} \iff f(x) = x + \ln x$

2. Estudiar gráficamente si se cumplen las condiciones del la teorema de la aplicación contractiva en el intervalo $[0, 1]$



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

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

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

Viendo las gráficas:

- g_1 no cumple la condición 2 (del teorema de la aplicación contractiva) en ningún punto.
- g_2 cumple la condición 2 excepto quizás en 0 por lo que habría que tomar otro intervalo más reducido.
- g_3 cumple las dos condiciones para el intervalo $[0, 1]$ y además vemos que es la función con la menor derivada (la más horizontal).

3. Realizar nueve iteraciones con cada uno de las funciones de iteración utilizando como punto inicial $x_0 = 0.5$

Por ejemplo, si para $g_1(x) = -\ln(x)$ tomamos $x_0 = 0.5$:

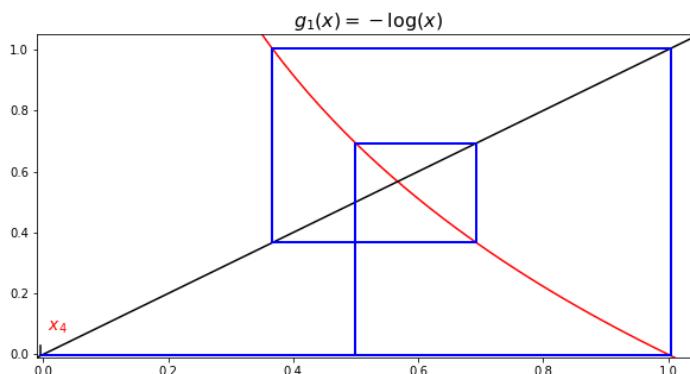
- Iteración 1: $x_1 = -\ln(0.5) = 0.693147$
- Iteración 2: $x_2 = -\ln(0.693147) = 0.366513$
- Iteración 3: $x_3 = -\ln(0.366513) = 1.003722$
- ...

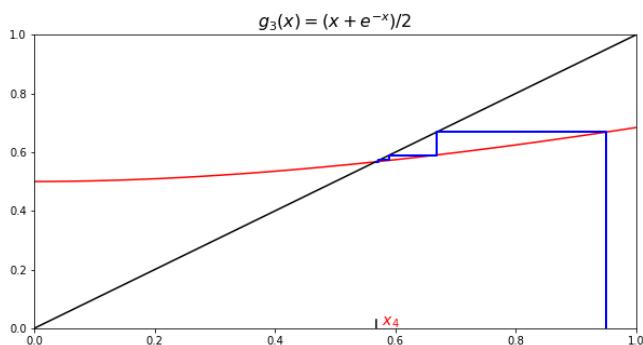
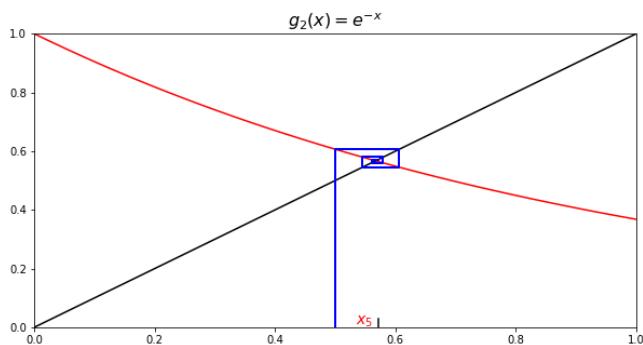
Para las demás funciones

| k | g_1 | g_2 | g_3 |
|-----|-----------|----------|----------|
| 0 | 0.500000 | 0.500000 | 0.500000 |
| 1 | 0.693147 | 0.606531 | 0.553265 |
| 2 | 0.366513 | 0.545239 | 0.564167 |
| 3 | 1.003722 | 0.579703 | 0.566500 |
| 4 | -0.003715 | 0.560065 | 0.567004 |
| 5 | NaN | 0.571172 | 0.567113 |
| 6 | NaN | 0.564863 | 0.567137 |
| 7 | NaN | 0.568438 | 0.567142 |
| 8 | NaN | 0.566409 | 0.567143 |
| 9 | NaN | 0.567560 | 0.567143 |

Vemos que la función g_1 , como no está definida para valores negativos, a partir de cierto punto, la sucesión no está definida.

Podemos asumir que si una serie de decimales se repiten en iteraciones sucesivas son los correctos. Así, las sucesiones generadas con g_2 y g_3 convergen, pero converge más rápido la segunda, porque podemos asumir que como para 0.567143 se repiten los 6 decimales en las dos últimas iteraciones, esta es la solución correcta para 6 decimales. Y, si esta es la solución correcta, con 9 iteraciones g_2 ha conseguido dos decimales correctos.





4. ¿Qué funciones pueden usarse? ¿Qué función de iteración debería usarse?

Teniendo en cuenta los resultados anteriores, podemos usar g_2 y g_3 . La función g_1 no cumple las condiciones de convergencia y la sucesión generada por ella no converge.

La mejor función de iteración es la de derivada más baja en el intervalo porque converge más rápidamente, es decir, g_3 .

TEMA 3

APROXIMACIÓN DE FUNCIONES

3.1 Polinomios de interpolación de Lagrange

Ejercicio 3.1.1

Dados los nodos $x_0 = -1$, $x_1 = 1$, $x_2 = 3$ y $x_3 = 5$ y la función

$$f(x) = \sin\left(\frac{\pi}{6}x\right)$$

1. Calcular los polinomios fundamentales de Lagrange y dibujarlos.
2. Calcular el polinomio interpolante por el método de Lagrange.
3. Aproximar el valor en el punto $x = 2$, calcular la cota de error y compararla con el error.

INTRODUCCIÓN

Interpolación

En ocasiones necesitamos calcular el valor de una función en varios puntos pero

- Puede que no tengamos el valor de la función para todos los puntos sino solo en algunos porque, por ejemplo, hemos obtenido los valores que tenemos a partir de algún experimento.
- O puede que la función sea muy costosa de calcular desde el punto de vista computacional.
- O puede que queramos una función, que siendo aproximada a la nuestra, sea más fácil de derivar o integrar o realizar otro tipo de operación.

Es en estos casos cuando la técnica de interpolación se hace útil. En interpolación utilizamos los datos (valor, derivada) de nuestra función en varios puntos de la función a aproximar y construimos una nueva función. Ejemplos de tipos de funciones que se emplean para construir estas nuevas funciones aproximadas son los polinomios y las funciones trigonométricas.

Vamos a ver:

- Interpolación polinómica de Lagrange.
 - Utilizando los polinomios fundamentales de Lagrange.
 - Utilizando la forma de Newton con diferencias divididas.
- Interpolación polinómica a trozos.

Interpolación polinómica de Lagrange

En este tipo de interpolación disponemos como datos de:

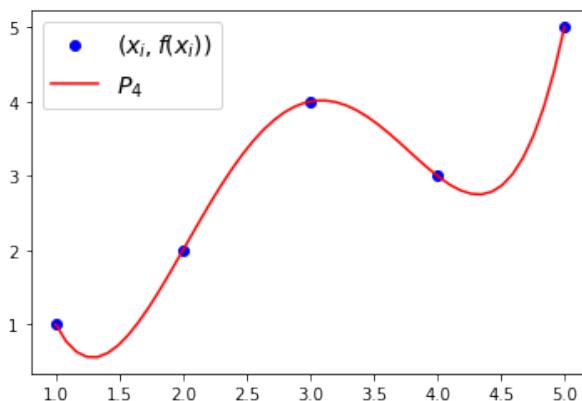
- **Nodos de interpolación:** son $n + 1$ puntos $x_0, x_1, x_2, \dots, x_n$
- Valores de la función f en los nodos: $f(x_0), f(x_1), f(x_2), \dots, f(x_n)$

Y buscamos un polinomio que pase por todos estos puntos. Esta puede ser toda la información que tengamos de f y no necesariamente existe una expresión analítica de f .

Por ejemplo, sean los puntos

| k | 0 | 1 | 2 | 3 | 4 |
|-----|---|---|---|---|---|
| x | 1 | 2 | 3 | 4 | 5 |
| y | 1 | 2 | 4 | 3 | 5 |

Buscamos un polinomio que pase por todos estos puntos



Existencia y unicidad del polinomio de Lagrange

En el ejemplo del gráfico anterior, como tenemos 5 puntos podemos plantear 5 ecuaciones y necesitamos 5 incógnitas que serán los coeficientes del polinomio. Por lo tanto:

$$P_4(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$$

es decir a_0, a_1, a_2, a_3 y a_4 son las 5 incógnitas y el polinomio es, en principio, de grado 4. En general, para los puntos $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$ el polinomio interpolante será de grado $\leq n$.

Las ecuaciones serían:

$$\begin{aligned} P_4(x_0) = y_0 & \quad a_0 + a_1x_0 + a_2x_0^2 + a_3x_0^3 + a_4x_0^4 = y_0 \\ P_4(x_1) = y_1 & \quad a_0 + a_1x_1 + a_2x_1^2 + a_3x_1^3 + a_4x_1^4 = y_1 \\ P_4(x_2) = y_2 & \quad a_0 + a_1x_2 + a_2x_2^2 + a_3x_2^3 + a_4x_2^4 = y_2 \\ P_4(x_3) = y_3 & \quad a_0 + a_1x_3 + a_2x_3^2 + a_3x_3^3 + a_4x_3^4 = y_3 \\ P_4(x_4) = y_4 & \quad a_0 + a_1x_4 + a_2x_4^2 + a_3x_4^3 + a_4x_4^4 = y_4 \end{aligned}$$

Y el sistema a resolver es:

$$\left(\begin{array}{ccccc} 1 & x_0 & x_0^2 & x_0^3 & x_0^4 \\ 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ 1 & x_3 & x_3^2 & x_3^3 & x_3^4 \\ 1 & x_4 & x_4^2 & x_4^3 & x_4^4 \end{array} \right) \left(\begin{array}{c} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array} \right) = \left(\begin{array}{c} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{array} \right) \quad (1)$$

La matriz de coeficientes del sistema se llama *Matriz de Vandermonde* y su determinante es

$$\det(A) = \prod_{0 \leq l \leq k < n} (x_k - x_l) \neq 0 \quad \text{si } x_k \neq x_l$$

es decir

$$\begin{aligned} \det(A) &= (x_1 - x_0)(x_2 - x_0)(x_3 - x_0)(x_4 - x_0)(x_2 - x_1) \\ &\quad (x_3 - x_1)(x_4 - x_1)(x_3 - x_2)(x_4 - x_2)(x_4 - x_3) \end{aligned}$$

que es distinto de cero si no hay dos x_i iguales. Y en este caso, la solución del sistema existe y es única porque es un sistema determinado.

En el ejemplo anterior los puntos (x_k, y_k) , con $y_k = f(x_k)$ son

| k | 0 | 1 | 2 | 3 | 4 |
|-----|---|---|---|---|---|
| x | 1 | 2 | 3 | 4 | 5 |
| y | 1 | 2 | 4 | 3 | 5 |

Sustituyendo el valor de los nodos en el sistema lineal (1)

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 & 16 \\ 1 & 3 & 9 & 27 & 81 \\ 1 & 4 & 16 & 64 & 256 \\ 1 & 5 & 25 & 125 & 625 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \\ 3 \\ 5 \end{pmatrix}$$

Como tenemos 5 nodos distintos el determinante no se anula

$$\det(A) = (2-1)(3-1)(4-1)(5-1)(3-2)(4-2)(5-2)(4-3)(5-3)(5-4) = 288$$

Y, resolviendo el sistema, el polinomio interpolador de Lagrange es

$$P_5(x) = 15 - 28.6666667x + 19.08333333x^2 - 4.83333333x^3 + 0.41666667x^4$$

Una forma de obtener el polinomio interpolador de Lagrange es resolver el sistema anterior. Pero este sistema está mal condicionado (pequeñas errores en los datos pueden producir grandes errores en los resultados) y es comparativamente costoso.

Vamos a ver dos formas alternativas de calcular el polinomio interpolador de Lagrange:

- Mediante los polinomios fundamentales de Lagrange.
- Mediante las diferencias divididas de Newton.

EJERCICIO

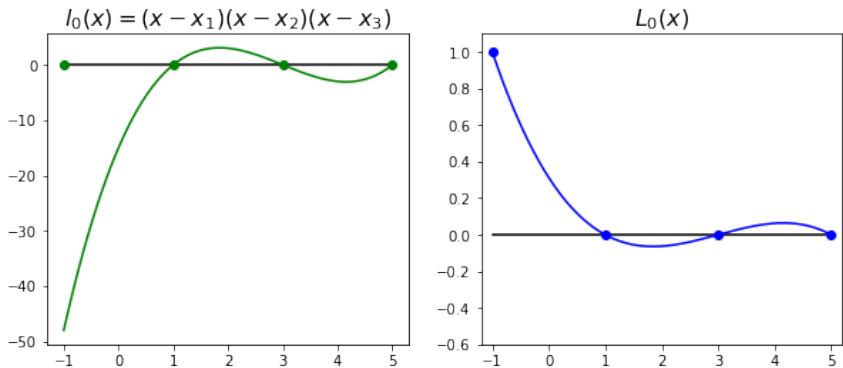
Dados los nodos $x_0 = -1$, $x_1 = 1$, $x_2 = 3$ y $x_3 = 5$ y la función

$$f(x) = \sin\left(\frac{\pi}{6}x\right)$$

1. Calcular los polinomios fundamentales de Lagrange y dibujarlos

Los polinomios fundamentales de lagrange tienen un valor 0 en todos los nodos excepto en uno de ellos, donde vale 1. Hay tantos polinomios fundamentales de Lagrange como nodos y los llamaremos $L_0(x)$, $L_1(x)$, $L_2(x)$ y $L_3(x)$.

Veamos como se construyen. Primero hacemos que los polinomios valgan cero en todos los nodos de interpolación menos en uno y tendremos



Los nodos son

$$x_0 = -1 \quad x_1 = 1 \quad x_2 = 3 \quad x_3 = 5$$

Y un polinomio que es cero en los tres últimos los nodos es

$$l_0(x) = (x - x_1)(x - x_2)(x - x_3) = (x - 1)(x - 3)(x - 5)$$

Si dividimos este polinomio por $l_0(x_0)$

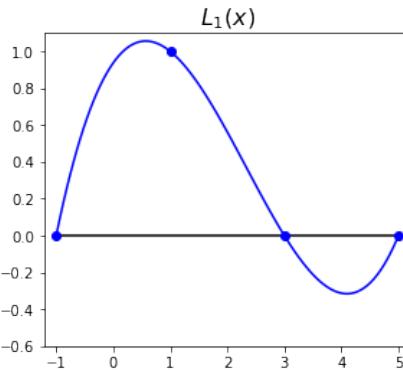
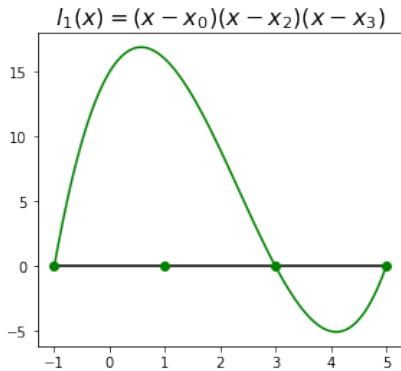
$$L_0(x) = \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} = \frac{(x - 1)(x - 3)(x - 5)}{(-1 - 1)(-1 - 3)(-1 - 5)}$$

Estamos omitiendo en el numerador y el denominador el término en x_0 , que es $(x - x_0)$ y

$$L_0(x_0) = 1 \quad L_0(x_1) = 0 \quad L_0(x_2) = 0 \quad L_0(x_3) = 0$$

es decir

$$L_k(x_i) = \begin{cases} 0 & \text{si } k \neq i \\ 1 & \text{si } k = i \end{cases}$$



Los nodos son

$$x_0 = -1 \quad x_1 = 1 \quad x_2 = 3 \quad x_3 = 5$$

Y un polinomio que es cero en los nodos primero, tercero y cuarto es

$$I_1(x) = (x - x_0)(x - x_2)(x - x_3) = (x + 1)(x - 3)(x - 5)$$

Si dividimos este polinomio por $I_1(x_1)$

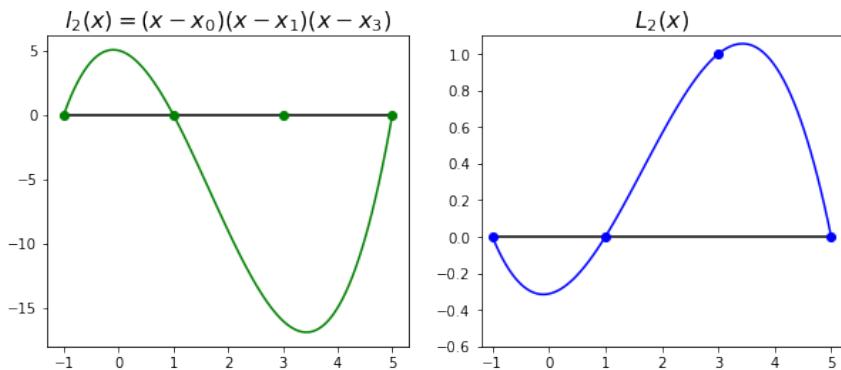
$$L_1(x) = \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} = \frac{(x + 1)(x - 3)(x - 5)}{(1 + 1)(1 - 3)(1 - 5)} \quad k = 0, 1, \dots, n$$

Estamos omitiendo en el numerador y el denominador el término en x_1 , que es $(x - x_1)$ y tenemos que

$$L_1(x_0) = 0 \quad L_1(x_1) = 1 \quad L_1(x_2) = 0 \quad L_1(x_3) = 0$$

es decir

$$L_k(x_i) = \begin{cases} 0 & \text{si } k \neq i \\ 1 & \text{si } k = i \end{cases}$$



Los nodos son

$$x_0 = -1 \quad x_1 = 1 \quad x_2 = 3 \quad x_3 = 5$$

Y un polinomio que es cero en todos los nodos menos el tercero es

$$l_2(x) = (x - x_0)(x - x_1)(x - x_3) = (x + 1)(x - 1)(x - 5)$$

Si dividimos este polinomio por $l_2(x_2)$

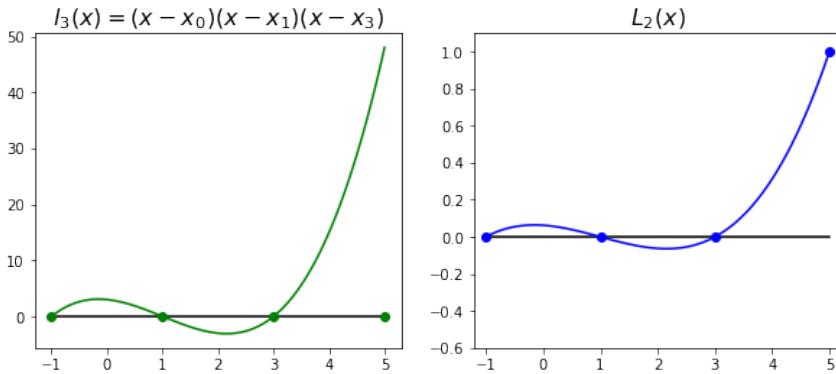
$$L_2(x) = \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} = \frac{(x + 1)(x - 1)(x - 5)}{(3 + 1)(3 - 1)(3 - 5)}$$

Estamos omitiendo en el numerador y el denominador el término en x_2 , que es $(x - x_2)$ y

$$L_2(x_0) = 0 \quad L_2(x_1) = 0 \quad L_2(x_2) = 1 \quad L_2(x_3) = 0$$

es decir

$$L_k(x_i) = \begin{cases} 0 & \text{si } k \neq i \\ 1 & \text{si } k = i \end{cases}$$



Los nodos son

$$x_0 = -1 \quad x_1 = 1 \quad x_2 = 3 \quad x_3 = 5$$

Y un polinomio que es cero en todos los nodos menos el último es

$$l_3(x) = (x - x_0)(x - x_1)(x - x_2) = (x + 1)(x - 1)(x - 3)$$

Si dividimos este polinomio por $l_3(x_3)$

$$L_3(x) = \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} = \frac{(x + 1)(x - 1)(x - 3)}{(5 + 1)(5 - 1)(5 - 3)}$$

Estamos omitiendo en el numerador y el denominador el término en x_3 , que es $(x - x_3)$ y

$$L_3(x_0) = 0 \quad L_3(x_1) = 0 \quad L_3(x_2) = 0 \quad L_3(x_3) = 1$$

es decir

$$L_k(x_i) = \begin{cases} 0 & \text{si } k \neq i \\ 1 & \text{si } k = i \end{cases}$$

Ya hemos visto los polinomios fundamentales de Lagrange para nuestros nodos (4 nodos, 4 polinomios fundamentales). En general, para x_0, x_1, \dots, x_n nodos tenemos $n + 1$ polinomios fundamentales de Lagrange que son

$$L_k(x) = \frac{(x - x_0)}{(x_k - x_0)} \cdots \frac{(x - x_{k-1})}{(x_k - x_{k-1})} \frac{(x - x_{k+1})}{(x_k - x_{k+1})} \cdots \frac{(x - x_n)}{(x_k - x_n)}$$

o expresado con productorios

$$L_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^n \frac{x - x_j}{x_k - x_j} \quad k = 0, \dots, n$$

Estos polinomios verifican

$$L_k(x_i) = \begin{cases} 0 & \text{si } k \neq i \\ 1 & \text{si } k = i \end{cases}$$

2. Calcular el polinomio interpolante por el método de Lagrange

El polinomio interpolante es de la forma

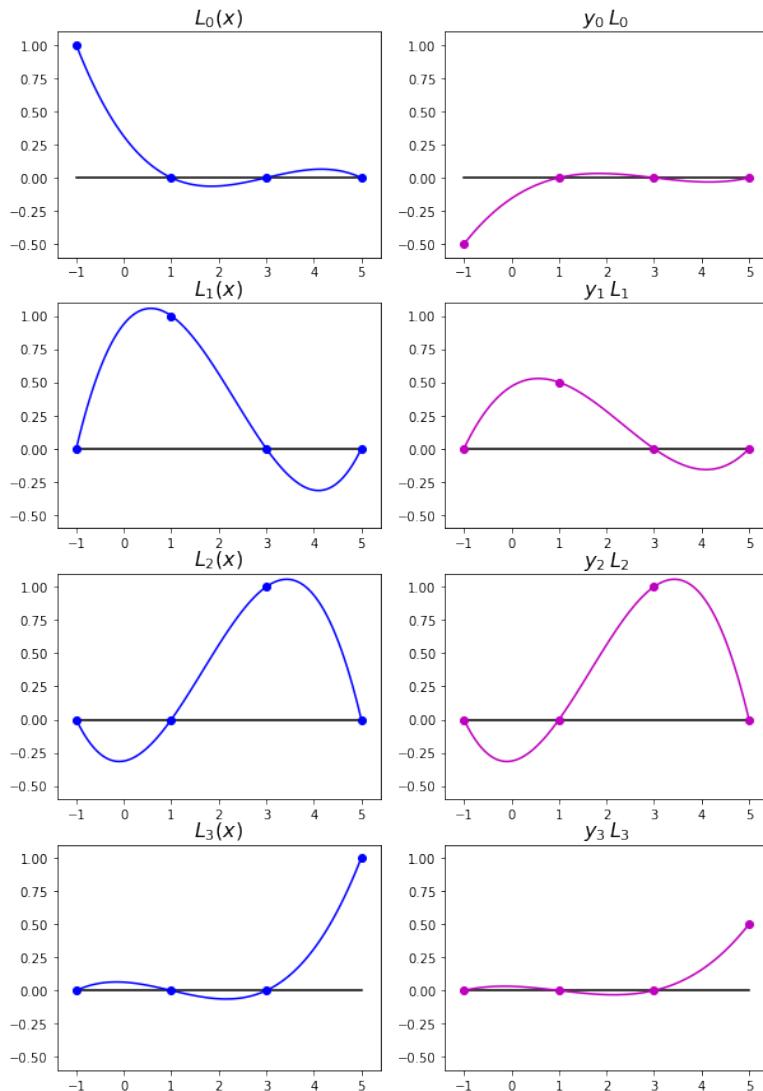
$$P_3(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) + y_3 L_3(x)$$

Los polinomios fundamentales de Lagrange se construyen solo con los nodos. Pero ahora, para calcular el polinomio de interpolación $P_3(x)$ necesitamos los valores de la función en los nodos

$$f(x) = \operatorname{sen}\left(\frac{\pi}{6}x\right)$$

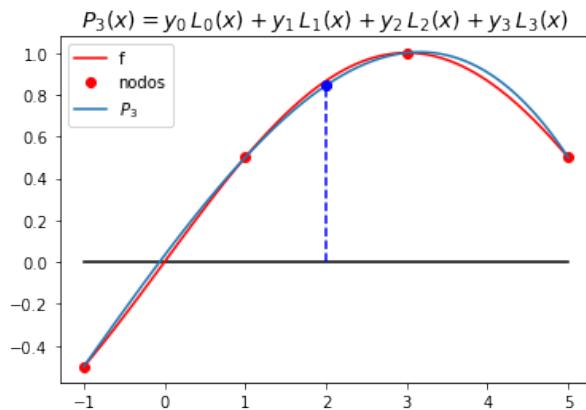
$$\begin{aligned} y_0 &= f(-1) = \operatorname{sen}\left(-\frac{\pi}{6}\right) = -\frac{1}{2} & y_1 &= f(1) = \operatorname{sen}\left(\frac{\pi}{6}\right) = \frac{1}{2} \\ y_2 &= f(3) = \operatorname{sen}\left(3\frac{\pi}{6}\right) = \operatorname{sen}\left(\frac{\pi}{2}\right) = 1 & y_3 &= f(5) = \operatorname{sen}\left(\frac{5\pi}{6}\right) = \frac{1}{2} \end{aligned}$$

| | | | | |
|-----|----------------|---------------|---|---------------|
| k | 0 | 1 | 2 | 3 |
| x | -1 | 1 | 3 | 5 |
| y | $-\frac{1}{2}$ | $\frac{1}{2}$ | 1 | $\frac{1}{2}$ |



- Si multiplicamos cada polinomio fundamental de Lagrange $L_i(x)$ por el valor de la función correspondiente, y_i , el polinomio $y_i L_i(x)$ toma el valor y_i en el nodo i y cero en los otros nodos.
- Así conseguimos 4 polinomios, cada uno de los cuales pasa por uno de los puntos a interpolar.
- Si sumamos estos 4 polinomios, como la suma se hace punto a punto, el valor en los nodos será el valor de la función, y_i , por lo que este polinomio suma, $P_3(x)$ es el polinomio de interpolación que estábamos buscando.
- Todos los polinomios son de grado 3, por lo que al sumarlos $P_3(x)$

será de grado menor o igual que 3. En este caso el polinomio es de grado tres pero si, por ejemplo, los 4 puntos estuvieran alineados, el polinomio sería de grado 1. O si los 4 puntos estuvieran sobre una recta horizontal el polinomio sería de grado cero. O si estuvieran sobre una parábola, sería de grado 2.



| k | 0 | 1 | 2 | 3 |
|-----|----------------|---------------|---|---------------|
| x | -1 | 1 | 3 | 5 |
| y | $-\frac{1}{2}$ | $\frac{1}{2}$ | 1 | $\frac{1}{2}$ |

Y teniendo en cuenta los polinomios fundamentales de Lagrange que calculamos arriba, el polinomio de interpolación es

$$P_3(x) = -\frac{1}{2}L_0(x) + \frac{1}{2}L_1(x) + (1)L_2(x) + \frac{1}{2}L_3(x)$$

El polinomio de interpolación, para un caso general, donde tenemos $n + 1$ nodos x_0, x_1, \dots, x_n será de la forma

$$P_n(x) = y_0 L_0(x) + y_1 L_1(x) + \cdots + y_n L_n(x) = \sum_{k=0}^n y_k L_k(x)$$

donde $P_n(x)$ es un polinomio de grado menor o igual que n .

3. Aproximar el valor en el punto $x = 2$, calcular la cota de error y compararla con el error

$$P_3(x) = -\frac{1}{2}L_0(x) + \frac{1}{2}L_1(x) + (1)L_2(x) + \frac{1}{2}L_3(x)$$

Sustituyendo $x = 2$ en las expresiones que calculamos para los polinomios fundamentales de Lagrange

$$P_3(2) = -\frac{1}{2}L_0(2) + \frac{1}{2}L_1(2) + (1)L_2(2) + \frac{1}{2}L_3(2)$$

$$P_3(2) = -\frac{1}{2}(-0.0625) + \frac{1}{2}(0.5625) + (1)(0.5625) + \frac{1}{2}(-0.0625)$$

$$\operatorname{sen}\left(2 \frac{\pi}{6}\right)=\operatorname{sen}\left(\frac{\pi}{3}\right) \approx P_3(2)=0.84375$$

Error de interpolación

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 y c un punto del intervalo de interpolación. En este caso, como tenemos cuatro nodos de interpolación

$$|E(x)| = |f(x) - P_3(x)| = \left|f^{(4)}(c)\right| \frac{|(x-x_0)(x-x_1)(x-x_2)(x-x_3)|}{4!}.$$

Como c es desconocido, aunque sabemos que está en el intervalo de interpolación, en nuestro caso $c \in (-1, 5)$, tenemos que encontrar una cota para ese valor

$$f(x) = \operatorname{sen}\left(\frac{\pi}{6}x\right) \quad f'(x) = \frac{\pi}{6} \cos\left(\frac{\pi}{6}x\right) \quad f''(x) = -\frac{\pi^2}{6^2} \operatorname{sen}\left(\frac{\pi}{6}x\right)$$

$$f'''(x) = -\frac{\pi^3}{6^3} \cos\left(\frac{\pi}{6}x\right) \quad f^{(4)}(x) = \frac{\pi^4}{6^4} \operatorname{sen}\left(\frac{\pi}{6}x\right)$$

Y como

$$\left|\operatorname{sen}\left(\frac{\pi}{6}x\right)\right| \leq 1$$

Entonces

$$\left|f^{(4)}(c)\right| \leq \frac{\pi^4}{6^4}(1) = \frac{\pi^4}{6^4}$$

Y por lo tanto podemos dar como cota del error

$$|E(2)| \leq \frac{\pi^4}{6^4} \frac{|(2-x_0)(2-x_1)(2-x_2)(2-x_3)|}{4!}$$

$$|E(2)| \leq \frac{\pi^4}{6^4} \frac{|(2+1)(2-1)(2-3)(2-5)|}{4!} = 0.0282$$

$$\text{Como } f(2) = \operatorname{sen}\left(\frac{\pi}{6}(2)\right) = \operatorname{sen}\left(\frac{\pi}{3}\right) = 0.86603$$

$$\text{Error} = |f(2) - P_3(2)| = |0.86603 - 0.84375| = 0.0223 < 0.0282$$

Ejercicio 3.1.2

Dados los nodos $x_0 = -1$, $x_1 = 1$, $x_2 = 3$ y $x_3 = 5$ y la función

$$f(x) = \operatorname{sen}\left(\frac{\pi}{6}x\right)$$

1. Construir la tabla de diferencias divididas.
2. Usando las diferencias divididas, calcular el polinomio interpolante en la forma de Newton.
3. Aproximar el valor de la función en el punto $x = 2$.

1. Tabla de diferencias divididas

Para 4 puntos la tabla de diferencias divididas se construye

$$\begin{array}{cc} x_0 & y_0 \\ \hline & [y_0, y_1] = \frac{y_1 - y_0}{x_1 - x_0} \\ x_1 & y_1 \\ \hline & [y_0, y_1, y_2] = \frac{[y_1, y_2] - [y_0, y_1]}{x_2 - x_0} \\ & [y_1, y_2] = \frac{y_2 - y_1}{x_2 - x_1} \\ x_2 & y_2 \\ \hline & [y_1, y_2, y_3] = \frac{[y_2, y_3] - [y_1, y_2]}{x_3 - x_1} \\ & [y_2, y_3] = \frac{y_3 - y_2}{x_3 - x_2} \\ x_3 & y_3 \end{array}$$

$$\begin{aligned} [y_0, y_1, y_2, y_4] &= \frac{[y_1, y_2, y_3] - [y_0, y_1, y_2]}{x_3 - x_0} \end{aligned}$$

Los nodos x vienen dados y las y correspondientes se calculan a partir de la función f .

$$f(x) = \operatorname{sen}\left(\frac{\pi}{6}x\right)$$

$$\begin{aligned} y_0 &= f(-1) = \operatorname{sen}\left(-\frac{\pi}{6}\right) = -\frac{1}{2} & y_1 &= f(1) = \operatorname{sen}\left(\frac{\pi}{6}\right) = \frac{1}{2} \\ y_2 &= f(3) = \operatorname{sen}\left(3\frac{\pi}{6}\right) = \operatorname{sen}\left(\frac{\pi}{2}\right) = 1 & y_3 &= f(5) = \operatorname{sen}\left(\frac{5\pi}{6}\right) = \frac{1}{2} \end{aligned}$$

| | | | | |
|-----|----------------|---------------|---|---------------|
| k | 0 | 1 | 2 | 3 |
| x | -1 | 1 | 3 | 5 |
| y | $-\frac{1}{2}$ | $\frac{1}{2}$ | 1 | $\frac{1}{2}$ |

Y la tabla de diferencias divididas en este caso es

| x | y | c_0 |
|-----|------------------------|--|
| -1 | $\boxed{-\frac{1}{2}}$ | |
| 1 | $\frac{1}{2}$ | $\frac{\frac{1}{2} - (-\frac{1}{2})}{1 - (-1)} = \boxed{\frac{1}{2}}^{c_1}$ |
| 3 | 1 | $\frac{\frac{1}{4} - \frac{1}{2}}{3 - (-1)} = \boxed{-\frac{1}{16}}^{c_2}$ |
| 5 | $\frac{1}{2}$ | $\frac{\frac{1}{8} - (-\frac{1}{16})}{5 - (-1)} = \boxed{-\frac{1}{96}}^{c_3}$ |

O utilizando notación decimal

| x | y | c_0 |
|-----|----------------|---|
| -1 | $\boxed{-0.5}$ | |
| 1 | 0.5 | $\frac{0.5 - (-0.5)}{1 - (-1)} = \boxed{0.5}^{c_1}$ |
| 3 | 1 | $\frac{0.25 - 0.5}{3 - (-1)} = \boxed{-0.0625}^{c_2}$ |
| 5 | 0.5 | $\frac{-0.125 - (-0.0625)}{5 - (-1)} = \boxed{-0.010417}^{c_3}$ |

2. Usando las diferencias divididas, calcular el polinomio interpolante en la forma de Newton

El polinomio de interpolación de Lagrange en los nodos x_0, x_1, x_2 y x_3 donde la función toma los valores y_0, y_1, y_2 y y_3 es

$$P_3(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + c_3(x - x_0)(x - x_1)(x - x_2)$$

Si sustituimos los valores de los nodos y las diferencias divididas

$$P_3(x) = -\frac{1}{2} + \frac{1}{2}(x + 1) - \frac{1}{16}(x + 1)(x - 1) - \frac{1}{96}(x + 1)(x - 1)(x - 3)$$

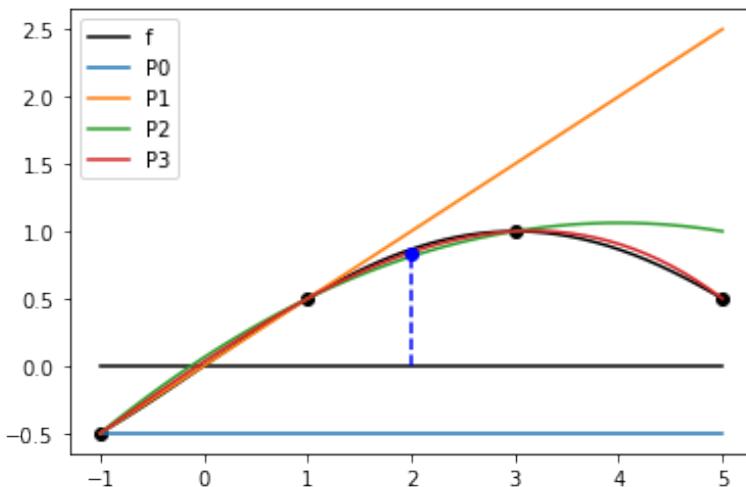
El polinomio interpolante en la forma de Newton

El polinomio interpolante de Lagrange en la forma de Newton se construye añadiendo un nodo con cada término que se suma

- $P_0(x) = -\frac{1}{2}$ es el polinomio de interpolación para x_0 .
- $P_1(x) = P_0(x) + \frac{1}{2}(x + 1)$ es el polinomio de interpolación para x_0 , y x_1 .
- $P_2(x) = P_1(x) - \frac{1}{16}(x + 1)(x - 1)$ es el polinomio de interpolación para x_0 , x_1 y x_2 .
- $P_3(x) = P_2(x) - \frac{1}{96}(x + 1)(x - 1)(x - 3)$ es el polinomio de interpolación para x_0 , x_1 , x_2 y x_3 .

$$P_3(x) = \boxed{\boxed{-\frac{1}{2} + \frac{1}{2}(x + 1)} - \frac{1}{16}(x + 1)(x - 1)} - \frac{1}{96}(x + 1)(x - 1)(x - 3)$$

Una consecuencia de esto es que si la última diferencia dividida (c_3) fuera cero, el polinomio interpolante ya no sería de grado 3 sino de grado 2. Y si las dos últimas fueran cero (c_2 y c_3) el polinomio sería de grado uno. Y así sucesivamente.



Así que una ventaja del polinomio interpolante en la forma de Newton es que se pueden ir añadiendo nodos uno a uno y no hay que reconstruir el polinomio completamente cada vez que se añade un nodo, como en la forma de Lagrange.

3. Aproximar el valor de la función en el punto $x = 2$

El polinomio interpolante en la forma de Newton es

$$P_3(x) = -\frac{1}{2} + \frac{1}{2}(x+1) - \frac{1}{16}(x+1)(x-1) - \frac{1}{96}(x+1)(x-1)(x-3)$$

Y aproximamos el valor de la función con el valor del polinomio

$$\begin{aligned} P_3(2) &= -\frac{1}{2} + \frac{1}{2}(2+1) - \frac{1}{16}(2+1)(2-1) - \frac{1}{96}(2+1)(2-1)(2-3) = \\ &= -\frac{1}{2} + \frac{3}{2} - \frac{3}{16} + \frac{3}{96} = 0.84375 \end{aligned}$$

$$\operatorname{sen}\left(2 \frac{\pi}{6}\right) = \operatorname{sen}\left(\frac{\pi}{3}\right) \approx P_3(2) = 0.84375$$

Que es el mismo resultado que habíamos obtenido para el polinomio interpolante en la forma de Lagrange **porque es el mismo polinomio** escrito de otra forma.

3.2 Interpolación polinómica a trozos

Ejercicio 3.2.1

Dados los nodos $x_0 = -1, x_1 = 1, x_2 = 3$ y $x_3 = 5$ y la función

$$f(x) = \operatorname{sen}\left(\frac{\pi}{6}x\right)$$

1. Aproximar el valor en el punto $x = 2$ utilizando interpolación lineal a trozos.
2. Calcular una cota de error y compararla con el error.

INTRODUCCIÓN

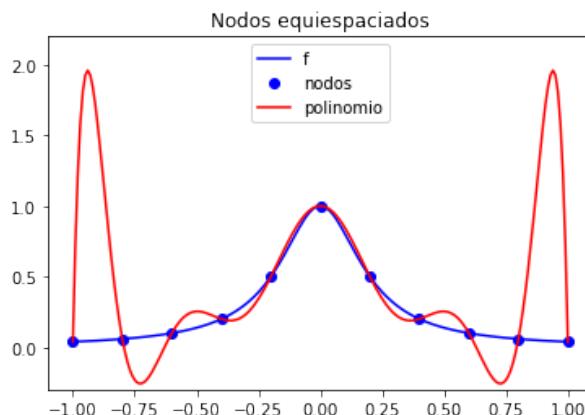
Interpolación lineal a trozos

Ya hemos visto como se construye el polinomio de interpolación de Lagrange. ¿Cómo podría mejorarse el resultado? Es decir, ¿cómo conseguir que el error sea más pequeño? ¿Añadiendo nodos? El problema es que si aumentamos el número de nodos, aumenta el grado del polinomio de interpolación. Y los polinomios de grados altos pueden dar errores muy grandes debido a grandes oscilaciones.

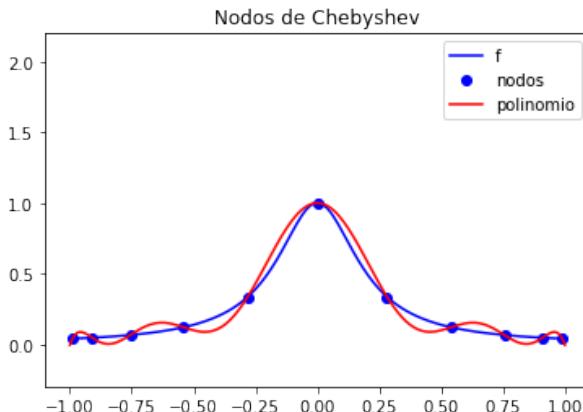
Por ejemplo, en el ejemplo siguiente, interpolamos la función

$$f(x) = \frac{1}{1 + 25x^2}$$

con 11 nodos equiespaciados en el intervalo $[-1, 1]$ y el error relativo en puntos cercanos a los extremos es enorme.



Por esta razón no se suelen emplear polinomios de interpolación de grados altos. Una solución es usar los *nodos de Chebysev*, que son las raíces del *polinomio de Chebysev* (están en el intervalo $[-1, 1]$ pero con una transformación lineal se pueden obtener para cualquier intervalo cerrado).



El problema es que no siempre podemos elegir los nodos.

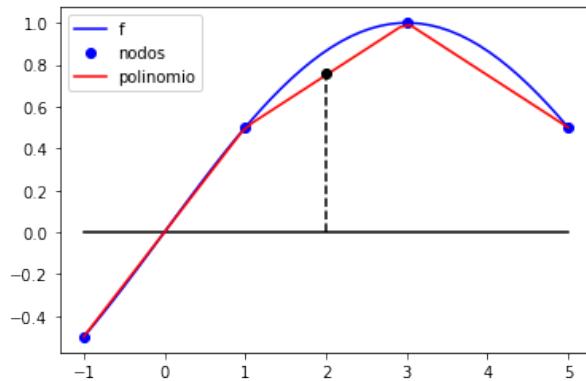
La solución más comúnmente utilizada es la **interpolación a trozos**. Se divide el intervalo en subintervalos y en cada uno de ellos se interpola con un polinomio de grado 1, 2 ó 3. En la práctica lo más común es

- *La interpolación lineal a trozos:* se utilizan polinomios de grado uno en cada subintervalo. Este método tiene la ventaja de que es muy sencillo y requiere muy pocos recursos.
- *Interpolación con splines cúbicos:* se utilizan polinomios de grado tres en cada subintervalo. Este método es más complejo pero suele dar muy buenos resultados.

EJERCICIO

- 1. Aproximar el valor en el punto $x = 2$ utilizando interpolación lineal a trozos.**

La interpolación lineal a trozos consiste en unir los nodos con segmentos de recta y aproximar el valor de la función con su valor sobre estas rectas.



Queremos aproximar la función en $x = 2$. Primero tenemos que escoger los nodos que usaremos para nuestro segmento de recta. Tomaremos los dos nodos más próximos a este valor que son 1 y 3. A partir de ahora nos olvidamos de los otros nodos y renombramos estos nodos como

$$x_0 = 1 \quad x_1 = 3$$

El valor de la función en estos nodos es

$$f(x) = \sin\left(\frac{\pi}{6}x\right)$$

$$y_0 = f(1) = \sin\left(\frac{\pi}{6}\right) = \frac{1}{2} \quad y_1 = f(3) = \sin\left(3\frac{\pi}{6}\right) = \sin\left(\frac{\pi}{2}\right) = 1$$

Forma de Lagrange

Podemos interpolar utilizando la forma de Lagrange y entonces el polinomio de interpolación lineal en el intervalo $[1, 3]$ tiene la forma

$$P_1(x) = y_0 L_0(x) + y_1 L_1(x)$$

$$P_1(x) = y_0 \frac{x - x_1}{x_0 - x_1} + y_1 \frac{x - x_0}{x_1 - x_0}$$

$$P_1(x) = \frac{1}{2} \frac{x - 3}{1 - 3} + (1) \frac{x - 1}{3 - 1}$$

Y el valor en $x = 2$ es

$$P_1(2) = \frac{1}{2} \frac{2 - 3}{1 - 3} + (1) \frac{2 - 1}{3 - 1} = \frac{1}{2} \left(\frac{-1}{-2} \right) + (1) \frac{1}{2} = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} = 0.75$$

Forma de Newton

La tabla de diferencias divididas es

| x | y |
|-----|--|
| 1 | $\begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix}^{c_0}$ |
| 3 | $\frac{1 - \frac{1}{2}}{3 - 1} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}^{c_1}$ |

Y el polinomio de interpolación en la forma de Newton

$$P_1(x) = c_0 + c_1(x - x_0)$$

$$P_1(x) = \frac{1}{2} + \frac{1}{4}(x - 1)$$

Y el valor en $x = 2$ es

$$P_1(2) = \frac{1}{2} + \frac{1}{4}(2 - 1) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4} = 0.75$$

2. Calcular una cota de error y compararla con el error

El error de interpolación viene dado por la fórmula

$$E(x) = f(x) - P_n(x) = f^{(n+1)}(c) \frac{(x - x_0) \dots (x - x_n)}{(n+1)!}$$

que es la misma que en la interpolación Lagrangiana puesto que si solo consideramos el intervalo $[1, 3]$ y que usamos solo dos nodos, es interpolación lagrangiana. Para dos nodos, esta fórmula es

$$E(x) = f(x) - P_1(x) = f''(c) \frac{(x - x_0)(x - x_1)}{2!}$$

El valor c es desconocido, aunque sabemos que está en el intervalo de interpolación, en nuestro caso $c \in (1, 3)$, y tenemos que encontrar una cota para ese valor. Calculamos la derivada segunda de nuestra función

$$f(x) = \sin\left(\frac{\pi}{6}x\right) \quad f'(x) = \frac{\pi}{6} \cos\left(\frac{\pi}{6}x\right) \quad f''(x) = -\frac{\pi^2}{6^2} \sin\left(\frac{\pi}{6}x\right)$$

Cuando $c \in (1, 3)$ el valor de

$$\frac{\pi}{6}c \in \left(\frac{\pi}{6}(1), \frac{\pi}{6}(3)\right) = \left(\frac{\pi}{6}, \frac{\pi}{2}\right) = (30^\circ, 90^\circ)$$

Y como la función seno toma el valor máximo para 90°

$$\left|\sin\left(\frac{\pi}{6}c\right)\right| < 1$$

Entonces

$$|f''(c)| < \frac{\pi^2}{6^2}$$

Y por lo tanto podemos dar como cota del error

$$|E(2)| < \frac{\pi^2}{6^2} \frac{|(2-1)(2-3)|}{2!} = 0.13707$$

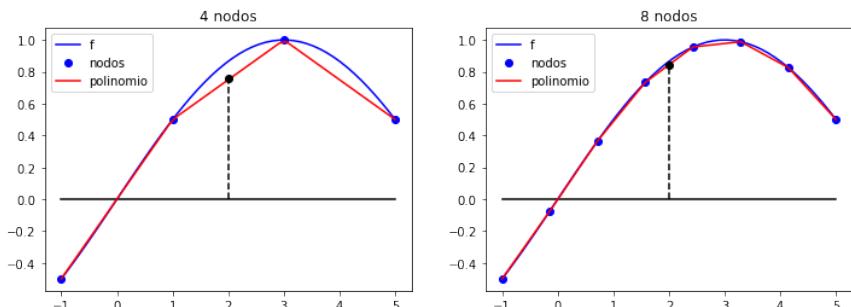
$$\text{Como } f(2) = \sin\left(\frac{\pi}{6}(2)\right) = \sin\left(\frac{\pi}{3}\right) = 0.86603$$

$$\text{Error} = |f(2) - P_1(2)| = |0.86603 - 0.75000| = 0.11603 < 0.13707$$

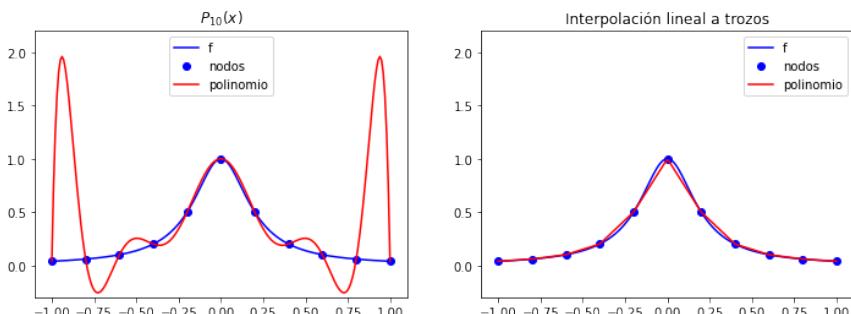
Vemos que

- El error es bastante mayor que el obtenido utilizando los cuatro nodos, pero también el procedimiento es mucho más sencillo.
- Tenemos una buena cota de error, porque es del mismo orden de magnitud que el error.

¿Podemos mejorar este resultado? Sí, aumentando el número de nodos. Por ejemplo, si en lugar de 4 nodos equiespaciados usamos 8 nodos equiespaciados el resultado de la interpolación lineal, gráficamente, será



Comparemos ahora, para el mismo número de nodos, la interpolación con polinomio de grado alto y la interpolación lineal a trozos. Vemos como, en general, es mejor la interpolación lineal a trozos.



Ejercicio 3.2.2

Dados los nodos $x_0 = -1$, $x_1 = 1$, $x_2 = 3$ y $x_3 = 5$, y la función

$$f(x) = \operatorname{sen}\left(\frac{\pi}{6}x\right)$$

Si escribimos el spline cúbico natural que los ajusta como

$$s(x) = \begin{cases} s_1(x) = a(x+1)^3 + b(x+1)^2 + c(x+1) + d & \text{si } x \in [-1, 1] \\ s_2(x) = e(x-1)^3 + f(x-1)^2 + g(x-1) + h & \text{si } x \in [1, 3] \\ s_3(x) = i(x-3)^3 + j(x-3)^2 + k(x-3) + l & \text{si } x \in [3, 5] \end{cases}$$

1. Calcular los coeficientes de los polinomios.
2. Calcular el valor de la spline en el punto $x = 2$.

1. Calcular los coeficientes de los polinomios

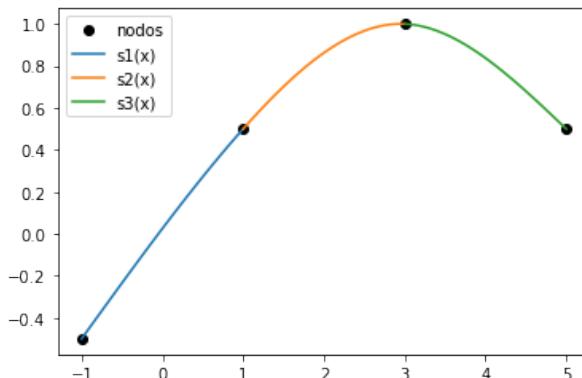
Para calcular el spline necesitamos los valores de la función en los nodos

$$f(x) = \operatorname{sen}\left(\frac{\pi}{6}x\right)$$

$$y_0 = f(-1) = \operatorname{sen}\left(-\frac{\pi}{6}\right) = -\frac{1}{2} \quad y_1 = f(1) = \operatorname{sen}\left(\frac{\pi}{6}\right) = \frac{1}{2}$$

$$y_2 = f(3) = \operatorname{sen}\left(3\frac{\pi}{6}\right) = \operatorname{sen}\left(\frac{\pi}{2}\right) = 1 \quad y_3 = f(5) = \operatorname{sen}\left(\frac{5\pi}{6}\right) = \frac{1}{2}$$

| | | | | |
|-----|----------------|---------------|---|---------------|
| k | 0 | 1 | 2 | 3 |
| x | -1 | 1 | 3 | 5 |
| y | $-\frac{1}{2}$ | $\frac{1}{2}$ | 1 | $\frac{1}{2}$ |



Se tienen que cumplir las siguientes condiciones:

- La curva ha de pasar por los tres puntos. Por lo tanto:

$$s_1(-1) = -\frac{1}{2}, s_1(1) = \frac{1}{2}, s_2(1) = \frac{1}{2}, s_2(3) = 1, s_3(3) = 1, s_3(5) = \frac{1}{2}.$$

- Han de coincidir las derivadas primera y segunda en los puntos intermedios:

$$s'_1(1) = s'_2(1), s'_2(3) = s'_3(3), s''_1(1) = s''_2(1), s''_2(3) = s''_3(3)$$

- 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:

$$s''_1(-1) = 0, s''_3(5) = 0.$$

Tenemos

$$s(x) = \begin{cases} s_1(x) = a(x+1)^3 + b(x+1)^2 + c(x+1) + d & \text{si } x \in [-1, 1] \\ s_2(x) = e(x-1)^3 + f(x-1)^2 + g(x-1) + h & \text{si } x \in [1, 3] \\ s_3(x) = i(x-3)^3 + j(x-3)^2 + k(x-3) + l & \text{si } x \in [3, 5] \end{cases}$$

Calculamos las derivadas primeras

$$s'(x) = \begin{cases} s'_1(x) = 3a(x+1)^2 + 2b(x+1) + c & \text{si } x \in [-1, 1] \\ s'_2(x) = 3e(x-1)^2 + 2f(x-1) + g & \text{si } x \in [1, 3] \\ s'_3(x) = 3i(x-3)^2 + 2j(x-3) + k & \text{si } x \in [3, 5] \end{cases}$$

y segundas

$$s''(x) = \begin{cases} s''_1(x) = 6a(x+1) + 2b & \text{si } x \in [-1, 1] \\ s''_2(x) = 6e(x-1) + 2f & \text{si } x \in [1, 3] \\ s''_3(x) = 6i(x-3) + 2j & \text{si } x \in [3, 5] \end{cases}$$

y las ecuaciones son:

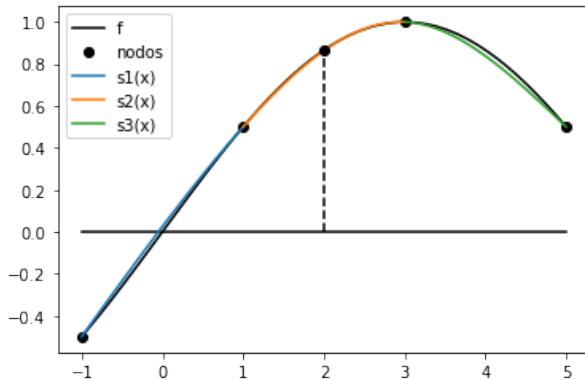
| | | |
|----|--------------------------|----------------------------------|
| 1 | $s_1(-1) = -\frac{1}{2}$ | $d = -\frac{1}{2}$ |
| 2 | $s_1(1) = \frac{1}{2}$ | $8a + 4b + 2c + d = \frac{1}{2}$ |
| 3 | $s_2(1) = \frac{1}{2}$ | $h = \frac{1}{2}$ |
| 4 | $s_2(3) = 1$ | $8e + 4f + 2g + h = 1$ |
| 5 | $s_3(3) = 1$ | $l = 1$ |
| 6 | $s_3(5) = \frac{1}{2}$ | $8i + 4j + 2k + l = \frac{1}{2}$ |
| 7 | $s'_1(1) = s'_2(1)$ | $12a + 4b + c = g$ |
| 8 | $s'_2(3) = s'_3(3)$ | $12e + 4f + g = k$ |
| 9 | $s''_1(1) = s''_2(1)$ | $12a + 2b = 2f$ |
| 10 | $s''_2(3) = s''_3(3)$ | $12e + 2f = 2j$ |
| 11 | $s''_1(-1) = 0$ | $2b = 0$ |
| 12 | $s''_3(5) = 0$ | $12i + 2j = 0$ |

Y tenemos un sistema lineal de 12 ecuaciones con 12 incógnitas, que expresado en forma matricial es

$$\left(\begin{array}{cccccccccccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 8 & 4 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 4 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 4 & 2 & 1 \\ 12 & 4 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 12 & 4 & 1 & 0 & 0 & 0 & -1 & 0 \\ 12 & 2 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 12 & 2 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12 & 2 & 0 & 0 \end{array} \right) \begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ i \\ j \\ k \\ l \end{pmatrix} = \begin{pmatrix} -1/2 \\ 1/2 \\ 1/2 \\ 1 \\ 1 \\ 1/2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

La solución de este sistema es $a = -1/120$, $b = 0$, $c = 8/15$, $d = -1/2$, $e = -1/48$, $f = -1/20$, $g = 13/30$, $h = 1/2$, $i = 7/240$, $j = -7/40$, $k = -1/60$ y $l = 1$. Por lo tanto la spline cúbica es

$$s(x) = \begin{cases} s_1(x) = -\frac{1}{120}(x+1)^3 + \frac{8}{15}(x+1) - \frac{1}{2} & \text{si } x \in [-1, 1] \\ s_2(x) = -\frac{1}{48}(x-1)^3 - \frac{1}{20}(x-1)^2 + \frac{13}{30}(x-1) + \frac{1}{2} & \text{si } x \in [1, 3] \\ s_3(x) = \frac{7}{240}(x-3)^3 - \frac{7}{40}(x-3)^2 - \frac{1}{60}(x-3) + 1 & \text{si } x \in [3, 5] \end{cases}$$



Este es un ejemplo de qué es una spline cúbica, no es un ejemplo de algoritmo eficaz de construcción del spline. En el algoritmo usual, la forma de los polinomios de grado 3 en cada subintervalo es distinta de la que se ha dado y busca que el sistema lineal sea *tridiagonal*, es decir que la matriz de coeficientes solo tiene elementos en tres diagonales: la principal, la que está por encima de la diagonal principal y la que está por debajo. Los sistemas tridiagonales tienen algoritmos específicos de almacenamiento y resolución más sencillos que los de matriz llena.

Independientemente del número de nodos, al construir una spline cúbica siempre faltan dos ecuaciones. Se necesitan dos ecuaciones adicionales que se suelen aplicar a los nodos de los bordes del intervalo y por ello se habla de *condiciones de contorno*. Dependiendo de ellas la spline se llama

- *Natural:* $s_1''(x_0) = 0$ y $s_n''(x_n) = 0$
- *Sujeta:* $s_1'(x_0) = y'_0$ y $s_n'(x_n) = y'_n$
- *Periódica:* Se tiene que verificar que $s_1(x_0) = s_n(x_n)$ y se impone que $s_1'(x_0) = s_n'(x_n)$ y $s_1''(x_0) = s_n''(x_n)$

2. Calcular el valor de la spline en el punto $x = 2$ Como el punto 2 está en el intervalo $[1, 3]$ utilizaremos el polinomio

$$s_2(x) = -\frac{1}{48}(x-1)^3 - \frac{1}{20}(x-1)^2 + \frac{13}{30}(x-1) + \frac{1}{2}$$

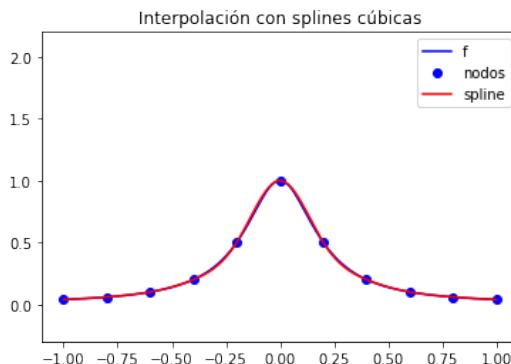
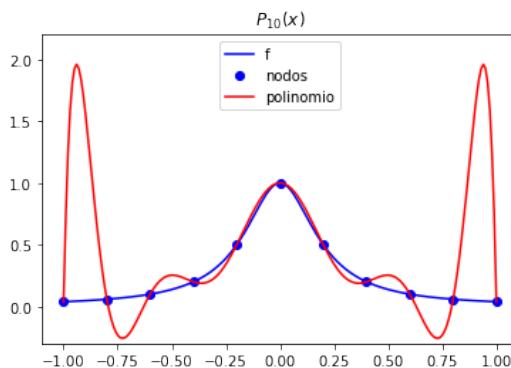
$$\begin{aligned} s(2) &= s_2(2) = -\frac{1}{48}(2-1)^3 - \frac{1}{20}(2-1)^2 + \frac{13}{30}(2-1) + \frac{1}{2} \\ s(2) &= -\frac{1}{48} - \frac{1}{20} + \frac{13}{30} + \frac{1}{2} = 0.8625 \end{aligned}$$

$$\text{Como } f(2) = \operatorname{sen}\left(\frac{\pi}{6}(2)\right) = \operatorname{sen}\left(\frac{\pi}{3}\right) = 0.86603$$

$$\text{Error} = |f(2) - s(2)| = |0.86603 - 0.86250| = 0.00353$$

Y vemos que con el mismo número de nodos, en este caso, el error es menor que para el polinomio de interpolación de Lagrange con 4 nodos.

Comparemos también el resultado para una función donde la interpolación de Lagrange falla porque el grado del polinomio es muy alto y el polinomio interpolatorio oscila.



3.3 Recta de regresión por mínimos cuadrados

Ejercicio 3.3.1

Dada la tabla de valores

| | | | | | |
|-----|---|---|---|----|----|
| x | 0 | 1 | 2 | 3 | 4 |
| y | 2 | 5 | 8 | 13 | 18 |

1. Hallar la recta de ajuste (recta de regresión) por mínimos cuadrados.
2. Calcular los residuos.

INTRODUCCIÓN

Ajuste de datos

Hemos visto la aproximación de funciones mediante interpolación. En interpolación, suponíamos que nuestros datos eran exactos. Por ejemplo, en algunos ejercicios de interpolación, obteníamos los datos a partir de la expresión de la función que queríamos aproximar.

El ajuste de datos es otra técnica de aproximación de funciones donde, en general, asumimos que nuestros datos contienen errores. Es el caso, por ejemplo, de datos obtenidos mediante mediciones en experimentos. A veces conocemos la forma de la función que subyace a los datos porque hay una ley física demostrada y, en ese caso, intentaremos utilizar esa función para aproximar los datos. En otros casos, no conocemos la función, e intentaremos aproximar con funciones genéricas, como polinomios o funciones trigonométricas.

El ejemplo más sencillo de ajuste o regresión es la **recta de regresión por mínimos cuadrados**. La recta o polinomio de grado uno, es la función con la que ajustamos. *Mínimos cuadrados* alude a la función que utilizamos para evaluar el error del ajuste. En este tema será la única función de error que usaremos. La ventaja que tiene usar el *error mínimo cuadrático* con funciones de ajuste lineales, que también serán las únicas que veremos en este tema, es que la solución del problema es un sistema de ecuaciones lineales.

EJERCICIO

1. Hallar la recta de ajuste (recta de regresión) por mínimos cuadrados

El problema de regresión por mínimos cuadrados o ajuste de datos se puede plantear de diferentes maneras:

- Como un problema de optimización.
- Como un problema de estadística.
- Como un problema de álgebra.

En este curso el enfoque que corresponde es el primero.

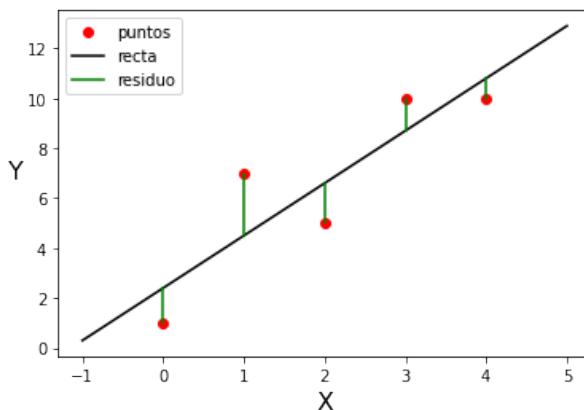
Notación expandida

La recta es un polinomio de grado uno y será de la forma:

$$P_1(x) = a_0 + a_1x$$

Nuestras incógnitas son a_0 y a_1 . Queremos calcular la recta que minimiza la suma de los residuos (errores) cuadráticos.

(Error y residuo son ideas parecidas pero se usan en contextos distintos. Ahora vamos a construir el modelo, la recta de regresión, y a la diferencia entre los valores de nuestros datos y el correspondiente valor del modelo se llama *residuo*. Cuando utilicemos este modelo para predecir un valor de la función hablaremos de *errores*. Pero en ambos casos es $y_i - P_1(x_i)$. Si x_i estaba entre los puntos con los que construimos el modelo esta diferencia será un *residuo*. Si x_i es un punto nuevo será un *error*.)



$$E = r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2$$

con

$$r_1 = y_1 - P_1(x_1) \quad r_2 = y_2 - P_1(x_2) \quad \dots \quad r_5 = y_5 - P_1(x_5)$$

- Parece que tendría más sentido minimizar

$$E = |r_1| + |r_2| + |r_3| + |r_4| + |r_5|,$$

que sería la suma de distancias entre los puntos y la recta. Pero este es un problema de solución más complicada.

- Al elevar al cuadrado estas distancias, garantizamos que son positivas y que no se compensan residuos positivos con negativos.
- Como E es una suma de cuadrados y vamos a minimizar esta función, la recta que obtendremos se llama *recta de regresión de mínimos cuadrados*.
- La notación para los puntos que estamos ajustando ahora empieza en 1. Los puntos son x_1, x_2, x_3, x_4 y x_5 y las y análogamente.
- El error total E depende de a_0 y a_1 que, de momento, son desconocidas.

Usaremos la función de error

$$E = r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2 \quad r_k = y_k - P_1(x_k)$$

Para los puntos

| | | | | | |
|-----|---|---|---|----|----|
| x | 0 | 1 | 2 | 3 | 4 |
| y | 2 | 5 | 8 | 13 | 18 |

Y la función de ajuste

$$P_1(x) = a_0 + a_1 x$$

Tenemos

$$\begin{aligned} E(a_0, a_1) &= (P_1(0) - 2)^2 + (P_1(1) - 5)^2 + (P_1(2) - 8)^2 + \\ &\quad + (P_1(3) - 13)^2 + (P_1(4) - 18)^2 = \\ &= (a_0 + a_1(0) - 2)^2 + (a_0 + a_1(1) - 5)^2 + (a_0 + a_1(2) - 8)^2 + \\ &\quad + (a_0 + a_1(3) - 13)^2 + (a_0 + a_1(4) - 18)^2 \end{aligned}$$

Para hallar el error mínimo calculamos las derivadas parciales respecto a las dos variables y las igualamos a cero:

$$\begin{aligned} \frac{\partial E}{\partial a_0} &= 2(a_0 + a_1(0) - 2) + 2(a_0 + a_1(1) - 5) + 2(a_0 + a_1(2) - 8) + \\ &\quad 2(a_0 + a_1(3) - 13) + 2(a_0 + a_1(4) - 18) = 0 \end{aligned}$$

$$\begin{aligned} \frac{\partial E}{\partial a_1} &= 2(a_0(0) + a_1(0)^2 - (0)(2)) + 2(a_0(1) + a_1(1)^2 - (1)(5)) + \\ &\quad + 2(a_0(2) + a_1(2)^2 - (2)(8)) + 2(a_0(3) + a_1(3)^2 - (3)(13)) + \\ &\quad + 2(a_0(4) + a_1(4)^2 - (4)(18)) = 0 \end{aligned}$$

Tomando a_0 y a_1 factor común

$$a_0(1 + 1 + 1 + 1 + 1) + a_1(0 + 1 + 2 + 3 + 4) = 2 + 5 + 8 + 13 + 18$$

$$\begin{aligned} a_0(0 + 1 + 2 + 3 + 4) + a_1(0^2 + 1^2 + 2^2 + 3^2 + 4^2) &= \\ &= (0)(2) + (1)(5) + (2)(8) + (3)(13) + (4)(18) \end{aligned}$$

Que es

$$\begin{array}{rcl} 5a_0 + 10a_1 & = & 46 \\ 10a_0 + 30a_1 & = & 132 \end{array}$$

Resolvemos el sistema por Gauss: la segunda ecuación $e_2 \rightarrow e_2 - 2e_1$

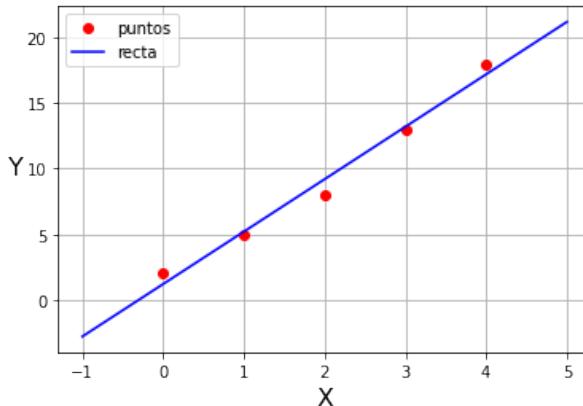
$$\begin{array}{rcl} 5a_0 + 10a_1 & = & 46 \\ 10a_1 & = & 40 \end{array}$$

y por sustitución reversiva

$$\begin{array}{rcl} a_1 & = & 40/10 = 4 \\ a_0 & = & (46 - 10a_1)/5 = 1.2 \end{array}$$

Y la recta de regresión mínimo cuadrática es

$$P_1(x) = 1.2 + 4x$$



Notación con sumatorios

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:

$$E(a_0, a_1) = \sum_{k=1}^5 (P_1(x_k) - y_k)^2 = \sum_{k=1}^5 (a_0 + a_1 x_k - y_k)^2.$$

Para hallar el error mínimo calculamos las derivadas parciales respecto a las dos variables y las igualamos a cero:

$$\begin{aligned}\frac{\partial E}{\partial a_0} &= \sum_{k=1}^5 2(a_0 + a_1 x_k - y_k) = 0 \\ \frac{\partial E}{\partial a_1} &= \sum_{k=1}^5 2(a_0 + a_1 x_k - y_k)x_k = \sum_{k=1}^5 2(a_0 x_k + a_1 x_k^2 - x_k y_k) = 0\end{aligned}$$

Que equivale a

$$\begin{aligned}a_0 \sum_{k=1}^5 1 + a_1 \sum_{k=1}^5 x_k &= \sum_{k=1}^5 y_k \\ a_0 \sum_{k=1}^5 x_k + a_1 \sum_{k=1}^5 x_k^2 &= \sum_{k=1}^5 x_k y_k\end{aligned}$$

Sistema, que expresado matricialmente es:

$$\begin{pmatrix} \sum_{k=1}^5 1 & \sum_{k=1}^5 x_k \\ \sum_{k=1}^5 x_k & \sum_{k=1}^5 x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 y_k \\ \sum_{k=1}^5 x_k y_k \end{pmatrix}$$

Calculamos los datos

| | 1 | x_k | x_k^2 | y_k | $x_k y_k$ |
|----------|---|-------|---------|-------|-----------|
| 1 | 0 | 0 | 2 | 0 | |
| 1 | 1 | 1 | 5 | 5 | |
| 1 | 2 | 4 | 8 | 16 | |
| 1 | 3 | 9 | 13 | 39 | |
| 1 | 4 | 16 | 18 | 72 | |
| Σ | 5 | 10 | 30 | 46 | 132 |

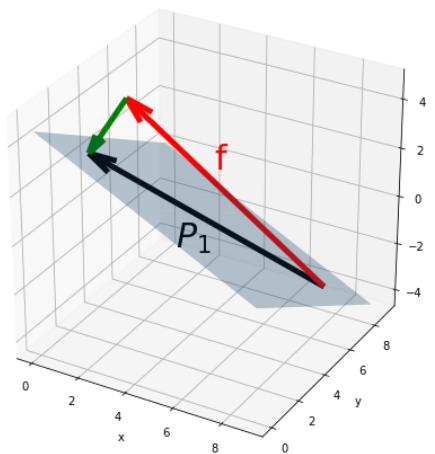
Y los sustituimos en el sistema

$$\begin{aligned}5a_0 + 10a_1 &= 46 \\ 10a_0 + 30a_1 &= 132\end{aligned}$$

Y la recta de regresión mínimo cuadrática es

$P_1(x) = 1.2 + 4x$

Planteamiento del problema como una proyección



El problema de aproximación por mínimos cuadrados también se puede plantear como una proyección ortogonal sobre un subespacio de funciones de base B . Entonces la solución vendría dada resolviendo el sistema lineal anterior.

Queremos aproximar los puntos usando la base de funciones polinómicas

$$B = \{P_0(x), P_1(x)\} = \{1, x\}$$

Es decir, queremos obtener un polinomio

$$P(x) = a_0 \cdot P_0(x) + a_1 \cdot P_1(x) = a_0 \cdot 1 + a_1 \cdot x$$

Obtenemos los coeficientes a_0 y a_1 como solución del sistema lineal

$$\begin{pmatrix} \langle P_0, P_0 \rangle & \langle P_0, P_1 \rangle \\ \langle P_1, P_0 \rangle & \langle P_1, P_1 \rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \langle P_0, f(x) \rangle \\ \langle P_1, f(x) \rangle \end{pmatrix}$$

En el caso discreto, el producto escalar más habitual es

$$\langle f(x), g(x) \rangle = \sum_{k=1}^n f(x_k)g(x_k)$$

Que para el caso anterior nos daría el sistema

$$\begin{pmatrix} \sum_{k=1}^5 1 \cdot 1 & \sum_{k=1}^5 1 \cdot x_k \\ \sum_{k=1}^5 x_k \cdot 1 & \sum_{k=1}^5 x_k \cdot x_k \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 1 \cdot y_k \\ \sum_{k=1}^5 x_k \cdot y_k \end{pmatrix}$$

O lo que es lo mismo

$$\begin{pmatrix} \sum_{k=1}^5 1 & \sum_{k=1}^5 x_k \\ \sum_{k=1}^5 x_k & \sum_{k=1}^5 x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 y_k \\ \sum_{k=1}^5 x_k y_k \end{pmatrix}$$

Sustituyendo los datos y operando

$$\begin{array}{lcl} 5a_0 + 10a_1 & = & 46 \\ 10a_0 + 30a_1 & = & 132 \end{array}$$

Y la recta de regresión mínimo cuadrática es

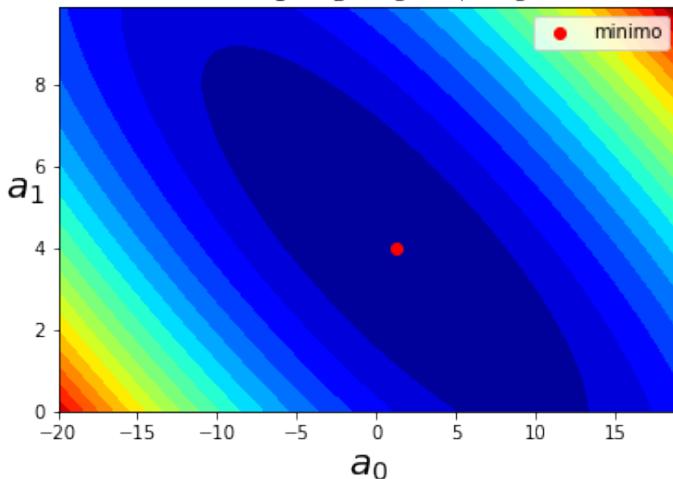
$$P_1(x) = 1.2 + 4x$$

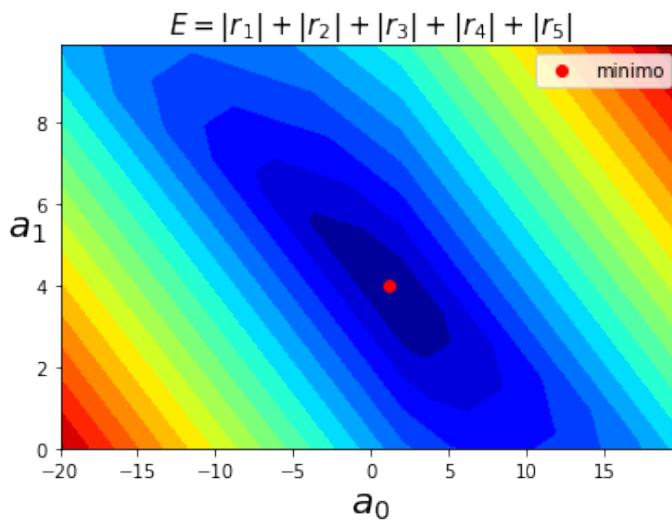
¿Por qué no resolvimos el problema usando los errores en lugar de los errores al cuadrado? Porque, como vimos, la derivada del error cuadrático es una función lineal, y obtenemos varias ecuaciones lineales y la solución es un sistema lineal. Es decir, la solución es sencilla y rápida.

Sin embargo, si utilizamos la suma de los residuales, la función E , contiene valores absolutos, que es una función que no siempre es derivable, es decir, no es suave, y el plantamiento que hicimos de obtener las derivadas parciales, ya no vale y la solución sería más complicada.

Actualmente sí se utiliza el valor absoluto en la función de error (y otras muchas funciones de error) gracias a la existencia de los ordenadores que hace que los cálculos se compliquen no sea un problema.

$$E = r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2$$

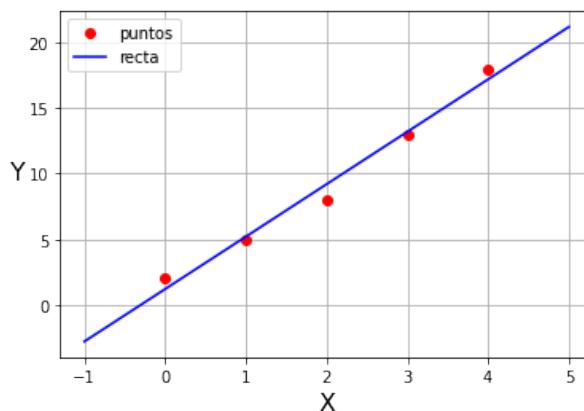




2. Calcular los residuos

Calculamos la suma de residuos, r_k , que es lo que hemos minimizado.

| x_k | $P_1(x_k)$ | y_k | $r_k = y_k - P_1(x_k)$ | r_k^2 |
|----------|------------|-------|------------------------|---------|
| 0 | 1.2 | 2 | 0.8 | 0.64 |
| 1 | 5.2 | 5 | -0.2 | 0.04 |
| 2 | 9.2 | 8 | -1.2 | 1.44 |
| 3 | 13.2 | 13 | -0.2 | 0.04 |
| 4 | 17.2 | 18 | 0.8 | 0.64 |
| Σ | | | | 2.80 |



3.4 Ajuste de funciones previa linearización

Ejercicio 3.4.1

Dada la tabla de valores

| | | | | | |
|-----|----|----|-----|------|------|
| t | 1 | 2 | 3 | 4 | 5 |
| Q | 14 | 58 | 260 | 1140 | 5660 |

ajustar la curva $Q(t) = r e^{st}$, calculando los valores r y s utilizando el criterio de los mínimos cuadrados.

INTRODUCCIÓN

La función exponencial

La función exponencial se utiliza en modelos sencillos de poblaciones y epidemias. La hipótesis de partida es que la variación de una población P con el tiempo, $\frac{dP}{dt}$, es proporcional a la población existente. Y en un modelo de epidemia, la aparición de nuevos casos de infectados es proporcional al número de personas ya contagiadas

$$\frac{dP}{dt} = kP$$

En una epidemia la constante k depende de la probabilidad de que una persona contagie a otras. Este es un modelo sencillo que supone la k constante. Pero variando las condiciones, la probabilidad de contagio puede cambiar.

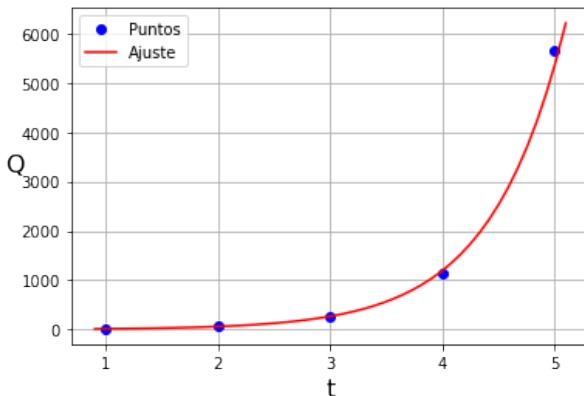
Sin perder generalidad, podemos suponer que $t_0 = 0$ y que $P(t_0) = P_0$ y resolviendo la ecuación diferencial

$$\begin{aligned} \frac{dP}{dt} = kP &\implies \frac{dP}{P} = kdt \implies \int_{P_0}^{P(t)} \frac{dP}{P} = \int_0^t kdt \implies \\ &\implies (\ln P)_{P_0}^{P(t)} t = kt \implies \ln P(t) - \ln P_0 = kt \implies \\ &\implies \ln \frac{P(t)}{P_0} = kt \implies \frac{P(t)}{P_0} = e^{kt} \implies P(t) = P_0 e^{kt} \end{aligned}$$

Y vemos que la función exponencial, escrita en esta forma, depende de dos parámetros. En el caso de que modele una población, P_0 es la población en el momento considerado inicial y k es la *constante de crecimiento relativo* de la población. En el caso de una epidemia, P_0 serían los infectados en el momento considerado inicial y k forma parte del *número reproductivo básico*, que es el número promedio de casos que genera un infectado, y es

$R_0 = e^{k\tau}$ siendo τ el periodo infecioso. Si $R_0 < 1$ la epidemia declina mientras que si $R_0 > 1$ se propaga.

Linealización de la función de ajuste



Si los datos cumplen (aproximadamente) que $Q = r e^{st}$ el ajuste (más adecuado) no va a ser lineal. Tenemos dos opciones:

- Intentar ajustar los datos con esta función de aproximación. Es decir, utilizar una función de error cuadrática

$$E(r, s) = r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2$$

que utilizando los puntos

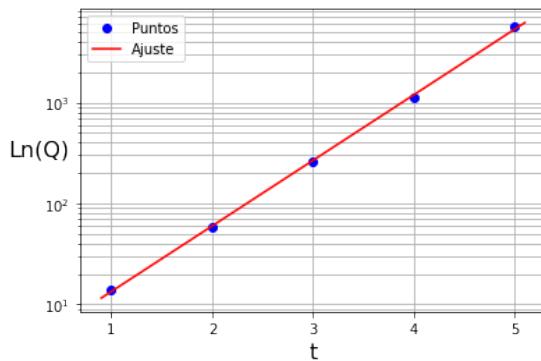
| t | 1 | 2 | 3 | 4 | 5 |
|---|----|----|-----|------|------|
| Q | 14 | 58 | 260 | 1140 | 5660 |

es

$$\begin{aligned} E(r, s) &= (Q(1) - 14)^2 + (Q(2) - 58)^2 + (Q(3) - 260)^2 + \\ &\quad + (Q(4) - 1140)^2 + (Q(5) - 5660)^2 = \\ &= (r e^s - 18)^2 + (r e^{2s} - 90)^2 + (r e^{3s} - 260)^2 + \\ &\quad + (r e^{4s} - 510)^2 + (r e^{5s} - 990)^2 \end{aligned}$$

Esta función es derivable pero el problema es que si calculamos las derivadas parciales respecto de r y s obtenemos un sistema de ecuaciones no lineales que son, en general, complicados de resolver. Como las ecuaciones no lineales, los sistemas de ecuaciones se suelen resolver de forma iterativa y pueden ser sensibles al punto inicial escogido.

Otra forma de resolverlo sería minimizar esta función directamente con un algoritmo de optimización. Estos algoritmos también suelen ser iterativos y los resultados depender del punto inicial.



- Linealizar la función de ajuste y , aplicando las transformaciones correspondientes a los datos, hacer una **regresión lineal por mínimos cuadrados**.

La ventaja de este método es que (si tenemos puntos suficientes y estos cumplen unas determinadas condiciones) la solución está determinada y se calcula resolviendo un sistema de ecuaciones lineales. La solución no coincide totalmente con la solución del primer planteamiento, pero da muy buen resultado de forma mucho más sencilla.

EJERCICIO

Así que vamos a linealizar la función a ajustar. Si tenemos en cuenta que

$$\ln(AB) = \ln A + \ln B \quad \ln A^B = B \ln A \quad \ln e = 1$$

entonces

$$\begin{aligned} Q = r e^{st} &\implies \ln Q = \ln(r e^{st}) \implies \ln Q = \ln r + \ln(e^{st}) \implies \\ &\implies \ln Q = \ln r + st \ln e \implies \ln Q = \ln r + s t \end{aligned}$$

Y si llamamos

$$y_k = \ln Q_k, \quad x_k = t_k, \quad a_0 = \ln r, \quad a_1 = s$$

tenemos

$$\ln Q_k \approx \ln r + s t_k \implies y_k \approx a_0 + a_1 x_k$$

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) , $k = 1, \dots, 5$ con el sistema

$$\begin{pmatrix} \sum_{k=1}^5 1 & \sum_{k=1}^5 x_k \\ \sum_{k=1}^5 x_k & \sum_{k=1}^5 x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 y_k \\ \sum_{k=1}^5 x_k y_k \end{pmatrix}$$

Calculamos los elementos del sistema

| | $x_k = t_k$ | Q_k | $y_k = \ln Q_k$ | x_k^2 | $x_k y_k$ |
|----------|-------------|-------|-----------------|---------|-----------|
| | 1 | 14 | 2.639 | 1 | 2.639 |
| | 2 | 58 | 4.060 | 4 | 8.121 |
| | 3 | 260 | 5.561 | 9 | 16.68 |
| | 4 | 1140 | 7.039 | 16 | 28.16 |
| | 5 | 5660 | 8.641 | 25 | 43.21 |
| Σ | 15 | | 27.94 | 55 | 98.80 |

Sustituyendo los datos y operando

$$\begin{aligned} 5a_0 + 15a_1 &= 27.94 \\ 15a_0 + 55a_1 &= 98.80 \end{aligned}$$

Y la solución de este sistema es $a_0 = 1.093$ $a_1 = 1.498$. Como

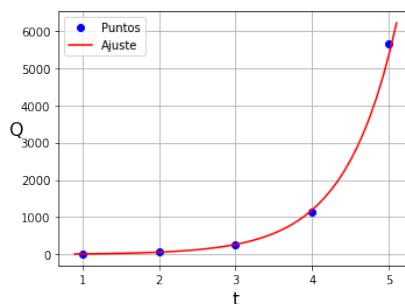
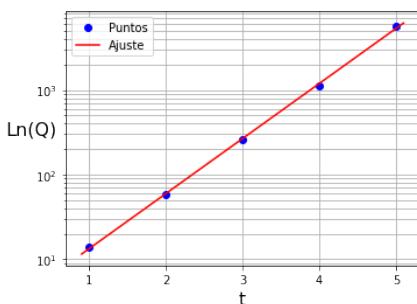
$$a_0 = \ln r, \quad a_1 = s \implies r = e^{a_0} \approx 3 \quad s = a_1 \approx 1.5$$

La curva ajustada es

$$Q(t) = 3e^{1.5t}$$

Resumiendo:

- Aplicamos una transformación que linealice la función que siguen los puntos.
- Aplicamos la misma transformación a los puntos.
- Ajustamos los puntos transformados a una recta por el método de los mínimos cuadrados.
- Utilizando los parámetros de la recta, a_0 y a_1 , deducimos los parámetros de nuestra curva de ajuste r y s y ya tenemos la curva de ajuste.



Ejercicio 3.4.2

Dada la tabla de valores

| | | | | | |
|-----|----|----|-----|-----|-----|
| t | 2 | 4 | 6 | 8 | 10 |
| Q | 18 | 90 | 260 | 510 | 990 |

ajustar la curva $Q(t) = r t^s$, calculando los valores r y s utilizando el criterio de los mínimos cuadrados.

Vamos a linealizar la función a ajustar. Si tenemos en cuenta que

$$\ln(AB) = \ln(A) + \ln(B) \quad \ln(A^B) = B \ln(A)$$

entonces

$$Q = r t^s \implies \ln Q = \ln(r t^s) \implies$$

$$\implies \ln Q = \ln r + \ln(t^s) \implies \ln Q = \ln r + s \ln t$$

Y si llamamos

$$y_k = \ln Q_k, \quad x_k = \ln t_k, \quad a_0 = \ln r, \quad a_1 = s$$

tenemos

$$\ln Q_k \approx \ln r + s \ln t_k \implies y_k \approx a_0 + a_1 x_k$$

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) , $k = 1, \dots, 5$ con el sistema

$$\begin{pmatrix} \sum_{k=1}^5 1 & \sum_{k=1}^5 x_k \\ \sum_{k=1}^5 x_k & \sum_{k=1}^5 x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 y_k \\ \sum_{k=1}^5 x_k y_k \end{pmatrix}$$

| | t_k | Q_k | $y_k = \ln Q_k$ | $x_k = \ln t_k$ | x_k^2 | $x_k y_k$ |
|----------|-------|-------|-----------------|-----------------|---------|-----------|
| | 2 | 18 | 2.890 | 0.6931 | 0.4805 | 2.003 |
| | 4 | 90 | 4.500 | 1.386 | 1.922 | 6.238 |
| | 6 | 260 | 5.561 | 1.792 | 3.210 | 9.963 |
| | 8 | 510 | 6.234 | 2.079 | 4.324 | 12.96 |
| | 10 | 990 | 6.898 | 2.303 | 5.302 | 15.88 |
| Σ | | | 26.08 | 8.253 | 15.24 | 47.05 |

Sustituyendo los datos y operando

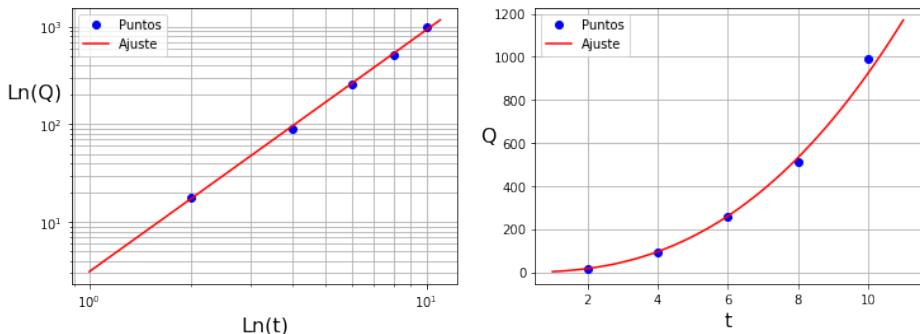
$$\begin{aligned} 5a_0 + 8.253a_1 &= 26.08 \\ 8.253a_0 + 15.24a_1 &= 47.05 \end{aligned}$$

Y la solución de este sistema es $a_0 = 1.312$ $a_1 = 2.475$. Como

$$a_0 = \ln r, \quad a_1 = s \implies r = e^{a_0} \approx 3.1, \quad s = a_1 \approx 2.5$$

La curva ajustada es

$$Q(t) = 3.1 t^{2.5}$$



Ejercicio 3.4.3

Dada la tabla de valores

| t | 1 | 2 | 3 | 4 | 5 |
|-----|------|------|------|------|------|
| Q | 0.47 | 0.27 | 0.20 | 0.15 | 0.12 |

ajustar, utilizando el criterio de los mínimos cuadrados, la curva

$$Q(t) = \frac{Q_0}{1 + Q_0 K t}$$

Vamos a linealizar la función a ajustar.

$$Q = \frac{Q_0}{1 + Q_0 K t} \implies \frac{1}{Q} = \frac{1 + Q_0 K t}{Q_0} \implies \frac{1}{Q} = \frac{1}{Q_0} + K t$$

Y si llamamos

$$y_k = \frac{1}{Q_k}, \quad x_k = t_k, \quad a_0 = \frac{1}{Q_0}, \quad a_1 = K$$

tenemos

$$\frac{1}{Q_k} = \frac{1}{Q_0} + K t_k \implies y_k = a_0 + a_1 x_k$$

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) , $k = 1, \dots, 5$ con el sistema

$$\begin{pmatrix} \sum_{k=1}^5 1 & \sum_{k=1}^5 x_k \\ \sum_{k=1}^5 x_k & \sum_{k=1}^5 x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 y_k \\ \sum_{k=1}^5 x_k y_k \end{pmatrix}$$

Calculamos los elementos de estas matrices

| | $x_k = t_k$ | Q_k | $y_k = 1/Q_k$ | x_k^2 | $x_k y_k$ |
|----------|-------------|-------|---------------|---------|-----------|
| | 1 | 0.47 | 2.128 | 1 | 2.127 |
| | 2 | 0.27 | 3.704 | 4 | 7.407 |
| | 3 | 0.20 | 5.000 | 9 | 15.00 |
| | 4 | 0.15 | 6.667 | 16 | 26.67 |
| | 5 | 0.12 | 8.333 | 25 | 41.67 |
| Σ | 15 | | 25.831 | 55 | 92.87 |

Sustituyendo los datos y operando

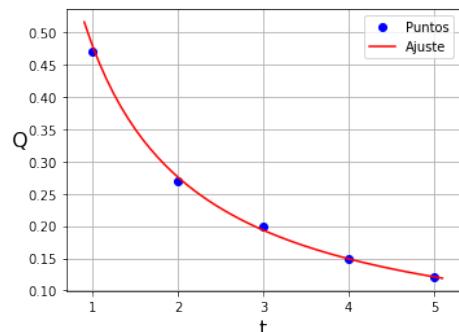
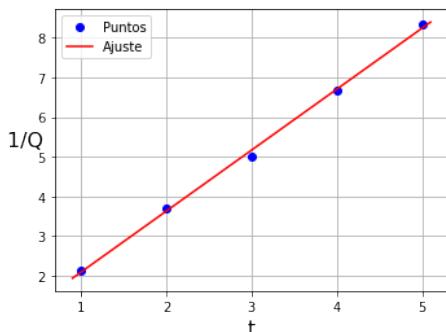
$$\begin{aligned} 5a_0 + 15a_1 &= 25.83 \\ 15a_0 + 55a_1 &= 92.87 \end{aligned}$$

Y la solución de este sistema es $a_0 = 0.5540$ $a_1 = 1.5474$. Como

$$a_0 = \frac{1}{Q_0}, \quad a_1 = K \implies Q_0 = \frac{1}{a_0} \approx 1.8 \quad K = a_1 \approx 1.5$$

La curva ajustada es

$$Q(t) = \frac{1.8}{1 + 1.8 \times 1.5 t}$$



Ejercicio 3.4.4

Para una población que evoluciona con el tiempo según la función $P(t)$, limitada por un valor L , de tipo logístico,

$$P(t) = \frac{L}{1 + c e^{Kt}}$$

se han reunido los siguientes datos

| t | 0 | 1 | 2 | 3 | 4 |
|-----|-----|-----|-----|-----|-----|
| P | 200 | 400 | 650 | 850 | 950 |

Y tomando $L = 1000$, calcular los valores c y K utilizando el criterio de los mínimos cuadrados.

INTRODUCCIÓN

La función logística

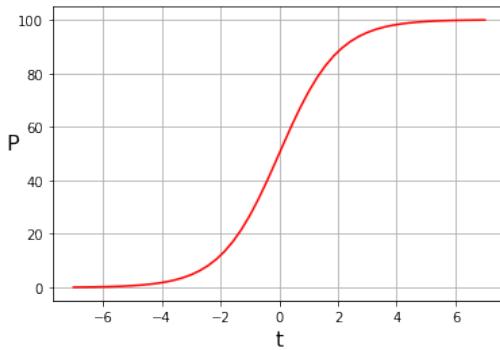
Una función un poco más compleja que la exponencial para modelizar poblaciones y epidemias es la función logística. En el caso de una población, describe la dinámica de una población en un entorno de recursos limitados. En el caso de un modelo epidemiológico, asume que hay un número máximo de individuos L que se pueden contagiar. La ecuación diferencial que la describe es parecida a la exponencial pero aparece un factor que pone de relieve esta limitación

$$\frac{dP}{dt} = KP \left(1 - \frac{P}{L}\right)$$

En los primeros momentos de la epidemia, cuando hay pocos individuos contagiados

$$\frac{P}{L} \approx 0 \quad \frac{dP}{dt} \approx KP$$

la curva se comporta como una exponencial. Cuando la proporción de contagiados crece lo suficiente el proceso se ralentiza porque hay muchos individuos que ya no son susceptibles de contagio porque ya se han infectado y la curva se comporta como una recta. En las últimas fases el número de infectados ya no crece.



EJERCICIO

Así que vamos a linealizar la función a ajustar.

$$\begin{aligned}
 P = \frac{L}{1 + c e^{Kt}} &\implies \frac{1}{P} = \frac{1 + c e^{Kt}}{L} \implies \frac{L}{P} = 1 + c e^{Kt} \implies \\
 &\implies \frac{L}{P} - 1 = c e^{Kt} \implies \ln\left(\frac{L}{P} - 1\right) = \ln(c e^{Kt}) \implies \\
 &\implies \ln \frac{L - P}{P} = \ln c + \ln e^{Kt} \implies \ln \frac{L - P}{P} = \ln c + K t \ln e \implies \\
 &\qquad \ln \frac{L - P}{P} = \ln c + K t
 \end{aligned}$$

Y si llamamos

$$y_k = \ln \frac{L - P_k}{P_k}, \quad x_k = t_k, \quad a_0 = \ln c, \quad a_1 = k$$

tenemos

$$\ln \frac{L - P_k}{P_k} \approx \ln c + K t_k \implies y_k \approx a_0 + a_1 x_k$$

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, \dots, 5\$ con el sistema

$$\begin{pmatrix} \sum_{k=1}^5 1 & \sum_{k=1}^5 x_k \\ \sum_{k=1}^5 x_k & \sum_{k=1}^5 x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 y_k \\ \sum_{k=1}^5 x_k y_k \end{pmatrix}$$

Si calculamos los elementos de estas matrices

| | $x_k = t_k$ | P_k | $y_k = \ln((L - P_k)/P_k)$ | x_k^2 | $x_k y_k$ |
|----------|-------------|-------|----------------------------|---------|-----------|
| | 0 | 200 | 1.386 | 0 | 0. |
| | 1 | 400 | 0.405 | 1 | 0.4055 |
| | 2 | 650 | -0.619 | 4 | -1.238 |
| | 3 | 850 | -1.735 | 9 | -5.204 |
| | 4 | 950 | -2.944 | 16 | -11.78 |
| Σ | 10 | | -3.506 | 30 | -17.81 |

Sustituyendo los datos y operando

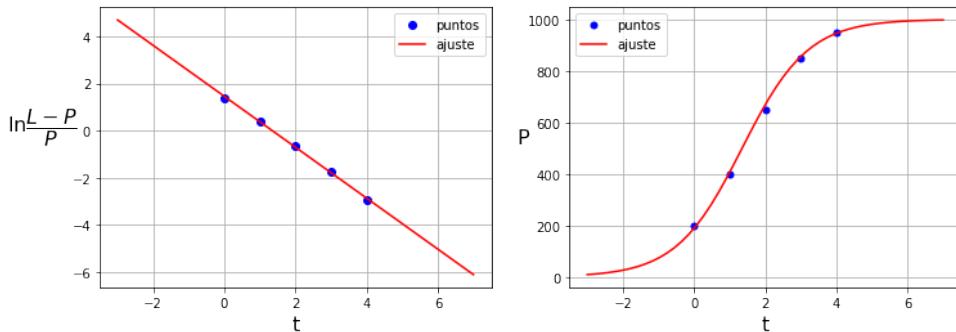
$$\begin{aligned} 5a_0 + 10a_1 &= -3.506 \\ 10a_0 + 30a_1 &= -17.81 \end{aligned}$$

Y la solución de este sistema es $a_0 = 1.459$ $a_1 = -1.080$. Como

$$a_0 = \ln c \quad a_1 = K \implies c = e^{a_0} \approx 4.3 \quad K = a_1 \approx -1.08$$

La curva ajustada es

$$P(t) = \frac{1000}{1 + 4.3 e^{-1.08t}}$$



3.5 Ajuste de funciones con polinomios

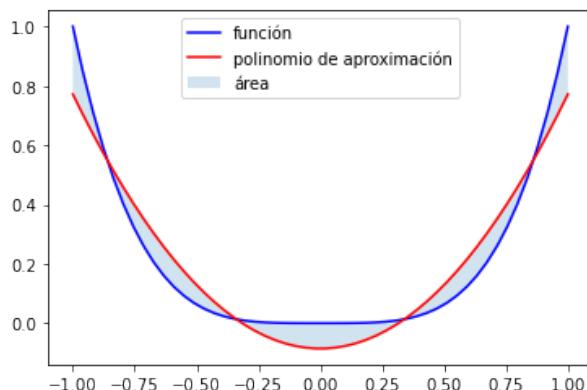
Ejercicio 3.5.1

Dada la función $f(x) = x^4$ en $[-1, 1]$ calcular la parábola que aproxima de forma continua a la función:

1. Utilizando la base de polinomios $\{1, x, x^2\}$.
2. Utilizando la base de polinomios ortogonales $\{1, x, 1 - 3x^2\}$.
3. Aproximar el valor en el punto $x = 0.8$ y calcular el error relativo y absoluto al usar el polinomio de aproximación en lugar de la función.

1. Base de polinomios $\{1, x, x^2\}$

Planteamiento como un problema de optimización



El problema se puede plantear como un problema de minimización.

Una primera propuesta sería minimizar el área entre las dos funciones, que es

$$E(a_0, a_1, a_2) = \int_{-1}^1 |P(x) - f(x)| dx$$

obteniendo las derivadas parciales de E respecto a a_0 , a_1 y a_2 . Pero el problema es que la función valor absoluto no es derivable en todos los puntos. Como en el caso discreto, una solución más sencilla es minimizar el error cuadrático, que es

$$E(a_0, a_1, a_2) = \int_{-1}^1 (P(x) - f(x))^2 dx = \int_{-1}^1 (a_0 + a_1 x + a_2 x^2 - f(x))^2 dx$$

Y obteniendo las derivadas parciales de E respecto a a_0 , a_1 y a_2 e igualándolas a cero obtendríamos un sistema lineal de tres ecuaciones con tres incógnitas que nos daría la solución del problema.

Planteamiento como un problema de proyección

Queremos aproximar la función usando la base de funciones polinómicas

$$B = \{P_0(x), P_1(x), P_2(x)\} = \{1, x, x^2\}$$

Es decir, queremos obtener un polinomio

$$P(x) = a_0 \cdot P_0(x) + a_1 \cdot P_1(x) + a_2 \cdot P_2(x) = a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot x^2$$

Como la base genera el espacio de los polinomios de grado 2 estamos buscando el polinomio de grado dos que mejor ajusta la función.

En el caso continuo, el producto escalar más habitual es

$$\langle g(x), h(x) \rangle = \int_{-1}^1 g(x)h(x)dx$$

Obtenemos los coeficientes a_0 , a_1 y a_2 como solución del sistema lineal

$$\begin{pmatrix} \langle P_0, P_0 \rangle & \langle P_0, P_1 \rangle & \langle P_0, P_2 \rangle \\ \langle P_1, P_0 \rangle & \langle P_1, P_1 \rangle & \langle P_1, P_2 \rangle \\ \langle P_2, P_0 \rangle & \langle P_2, P_1 \rangle & \langle P_2, P_2 \rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f(x) \rangle \\ \langle P_1, f(x) \rangle \\ \langle P_2, f(x) \rangle \end{pmatrix}$$

Sustituyendo el producto escalar con las funciones adecuadas

$$\begin{pmatrix} \int_{-1}^1 1 \cdot 1 dx & \int_{-1}^1 1 \cdot x dx & \int_{-1}^1 1 \cdot x^2 dx \\ \int_{-1}^1 x \cdot 1 dx & \int_{-1}^1 x \cdot x dx & \int_{-1}^1 x \cdot x^2 dx \\ \int_{-1}^1 x^2 \cdot 1 dx & \int_{-1}^1 x^2 \cdot x dx & \int_{-1}^1 x^2 \cdot x^2 dx \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \int_{-1}^1 1 \cdot f(x) dx \\ \int_{-1}^1 x \cdot f(x) dx \\ \int_{-1}^1 x^2 \cdot f(x) dx \end{pmatrix}$$

O lo que es lo mismo

$$\begin{pmatrix} \int_{-1}^1 1 dx & \int_{-1}^1 x dx & \int_{-1}^1 x^2 dx \\ \int_{-1}^1 x dx & \int_{-1}^1 x^2 dx & \int_{-1}^1 x^3 dx \\ \int_{-1}^1 x^2 dx & \int_{-1}^1 x^3 dx & \int_{-1}^1 x^4 dx \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \int_{-1}^1 x^4 dx \\ \int_{-1}^1 x x^4 dx \\ \int_{-1}^1 x^2 x^4 dx \end{pmatrix}$$

Calculando las integrales y sustituyendo los resultados

$$\begin{pmatrix} 2 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 2/5 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2/5 \\ 0 \\ 2/7 \end{pmatrix}$$

Si multiplicamos la primera fila por $-\frac{1}{3}$ y la sumamos a la tercera, es decir

$$e_3 \leftarrow e_3 - \frac{e_1}{3}$$

$$\begin{pmatrix} 2 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 0 & 0 & 8/45 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2/5 \\ 0 \\ 16/105 \end{pmatrix}$$

Que es el sistema triangular

$$\begin{array}{rcl} 2a_0 & +2/3a_2 & = 2/5 \\ 2/3a_1 & & = 0 \\ & +(8/45)a_2 & = 16/105 \end{array}$$

Y si resolvemos empezando por la última ecuación

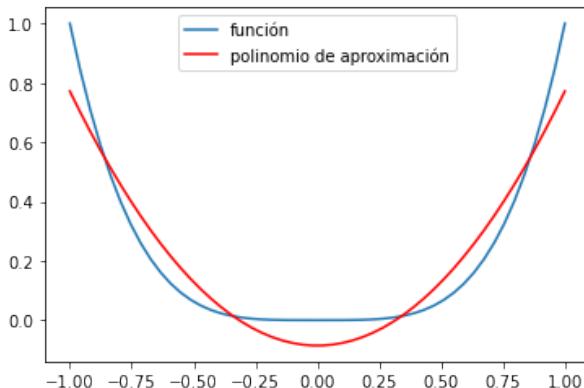
$$\begin{aligned} a_2 &= \frac{(16)(45)}{(8)(105)} = \frac{6}{7} \approx 0.8571 \\ a_1 &= 0 \\ a_0 &= \frac{1}{2} \left(\frac{2}{5} - \frac{2}{3} \frac{6}{7} \right) = -\frac{3}{35} \approx -0.08571 \end{aligned}$$

Y como el polinomio de grado dos que aproxima a esta función en el intervalo $[-1, 1]$ era

$$P(x) = a_0 \cdot P_0(x) + a_1 \cdot P_1(x) + a_2 \cdot P_2(x) = a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot x^2$$

sustituyendo los valores de a_0 , a_1 y a_2 .

$$P(x) = -\frac{3}{35} + \frac{6}{7}x^2$$



2. Base de polinomios ortogonales $\{1, x, 1 - 3x^2\}$

El sistema planteado en el apartado anterior, en general, está mal condicionado (pequeños errores en los datos pueden producir grandes errores en los resultados) por lo que conviene buscar una solución alternativa. Esta puede ser usar una base de polinomios ortogonales, es decir, polinomios, que para el producto escalar dado son ortogonales dos a dos, es decir

$$\langle P_i, P_j \rangle = 0 \quad \text{si} \quad i \neq j$$

Vamos a aproximar la función usando la base de funciones polinómicas

$$B = \{P_0(x), P_1(x), P_2(x)\} = \{1, x, 1 - 3x^2\}$$

Es decir, queremos obtener un polinomio

$$P(x) = a_0 \cdot P_0(x) + a_1 \cdot P_1(x) + a_2 \cdot P_2(x) = a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot (1 - 3x^2)$$

Obtenemos los coeficientes a_0 , a_1 y a_2 como solución del sistema lineal

$$\begin{pmatrix} \langle P_0, P_0 \rangle & \langle P_0, P_1 \rangle & \langle P_0, P_2 \rangle \\ \langle P_1, P_0 \rangle & \langle P_1, P_1 \rangle & \langle P_1, P_2 \rangle \\ \langle P_2, P_0 \rangle & \langle P_2, P_1 \rangle & \langle P_2, P_2 \rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f(x) \rangle \\ \langle P_1, f(x) \rangle \\ \langle P_2, f(x) \rangle \end{pmatrix}$$

Como en este caso la base de polinomios es ortogonal, el producto escalar entre dos polinomios de la base es cero y el sistema queda

$$\begin{pmatrix} \langle P_0, P_0 \rangle & 0 & 0 \\ 0 & \langle P_1, P_1 \rangle & 0 \\ 0 & 0 & \langle P_2, P_2 \rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f(x) \rangle \\ \langle P_1, f(x) \rangle \\ \langle P_2, f(x) \rangle \end{pmatrix}$$

Y la solución de este sistema diagonal es

$$a_0 = \frac{\langle P_0, f(x) \rangle}{\langle P_0, P_0 \rangle} \quad a_1 = \frac{\langle P_1, f(x) \rangle}{\langle P_1, P_1 \rangle} \quad a_2 = \frac{\langle P_2, f(x) \rangle}{\langle P_2, P_2 \rangle}$$

Usando el producto escalar

$$\langle g(x), h(x) \rangle = \int_{-1}^1 g(x)h(x)dx$$

Tenemos

$$a_0 = \frac{\int_{-1}^1 P_0 f(x) dx}{\int_{-1}^1 P_0 P_0 dx} \quad a_1 = \frac{\int_{-1}^1 P_1 f(x) dx}{\int_{-1}^1 P_1 P_1 dx} \quad a_2 = \frac{\int_{-1}^1 P_2 f(x) dx}{\int_{-1}^1 P_2 P_2 dx}$$

Es decir

$$a_0 = \frac{\int_{-1}^1 x^4 dx}{\int_{-1}^1 1 dx} \quad a_1 = \frac{\int_{-1}^1 x^5 dx}{\int_{-1}^1 x^2 dx} \quad a_2 = \frac{\int_{-1}^1 (1 - 3x^2) x^4 dx}{\int_{-1}^1 (1 - 3x^2)^2 dx}$$

Calculando estas integrales y sustituyendo su valor, tenemos

$$a_0 = \frac{2/5}{2} = \frac{1}{5} \quad a_1 = \frac{0}{2/3} = 0 \quad a_2 = \frac{-16/35}{8/5} = -\frac{2}{7}$$

Y como el polinomio buscado era

$$P(x) = a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot (1 - 3x^2)$$

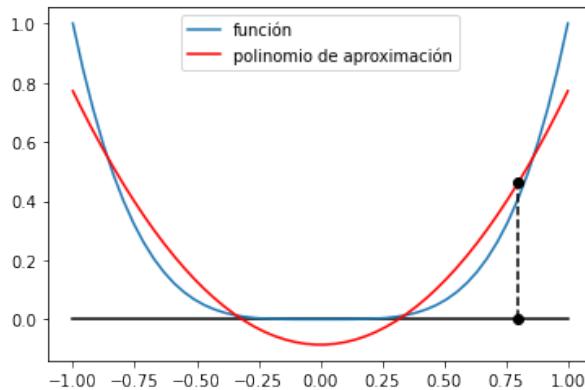
tenemos que la solución es

$$P(x) = \frac{1}{5} - \frac{2}{7}(1 - 3x^2)$$

Vamos a ponerlo en la forma del polinomio del primer apartado para compararlos

$$P(x) = \frac{1}{5} - \frac{2}{7} + \frac{2}{7}3x^2 = \frac{7}{35} - \frac{10}{35} + \frac{6}{7}x^2 = -\frac{3}{35} + \frac{6}{7}x^2$$

Y, efectivamente, es el mismo polinomio que en el caso anterior porque hemos el mismo problema de otra manera.



3. Aproximar el valor en el punto $x = 0.8$ y calcular el error relativo y absoluto al usar el polinomio de aproximación en lugar de la función.

Si x es el valor real y x^* el aproximado:

$$x = f(0.8) = 0.8^4 = 0.4096 \quad x^* = P(0.8) = -\frac{3}{35} + \frac{6}{7}0.8^2 = 0.4629$$

Por lo tanto

- Error absoluto

$$e_a = |x - x^*| = |0.4096 - 0.4629| = 0.0533$$

- Error relativo

$$e_r = \frac{e_a}{|x|} = \frac{e_a}{|x^*|} = \frac{0.053}{0.4096} = 0.13 = 13\%$$

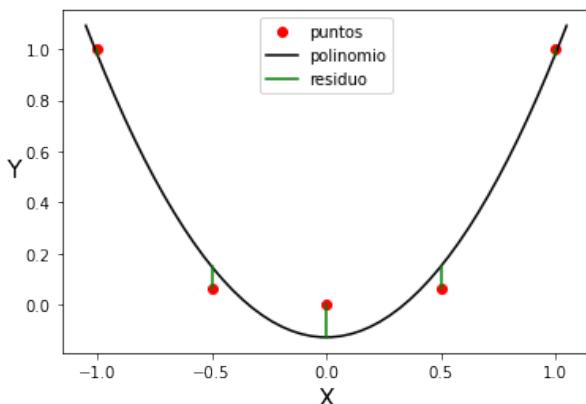
Ejercicio 3.5.2

Dados los puntos $x_1 = -1, x_2 = -0.5, x_3 = 0, x_4 = 0.5$ y $x_5 = 1$ de la función $f(x) = x^4$

1. Calcular la parábola que aproxima los puntos de la función.
2. Aproximar el valor en el punto $x = 0.8$ y calcular el error relativo y absoluto al usar el polinomio de aproximación en lugar de la función.

1. Calcular la parábola que aproxima los puntos de la función

Planteamiento como un problema de optimización



El problema se puede plantear como un problema de minimización. Y entonces minimizaremos

$$E(a_0, a_1, a_2) = r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2$$

donde

$$r_k = y_k - P(x_k) \quad \text{con } P(x) = a_0 + a_1x + a_2x^2 \quad k = 1, \dots, 5$$

Y obteniendo las derivadas parciales de E respecto a a_0 , a_1 y a_2 e igualándolas a cero obtendríamos un sistema lineal de tres ecuaciones con tres incógnitas que nos daría la solución del problema.

Planteamiento como un problema de proyección

Queremos aproximar los puntos usando la base de funciones polinómicas

$$B = \{P_0(x), P_1(x), P_2(x)\} = \{1, x, x^2\}$$

Es decir, queremos obtener un polinomio

$$P(x) = a_0 \cdot P_0(x) + a_1 \cdot P_1(x) + a_2 \cdot P_2(x) = a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot x^2$$

Como la base genera el espacio de los polinomios de grado 2 estamos buscando el polinomio de grado dos que mejor ajusta la función.

En el caso discreto, el producto escalar más habitual es

$$\langle g(x), h(x) \rangle = \sum_{k=1}^n g(x_k)h(x_k)$$

Obtenemos los coeficientes a_0 , a_1 y a_2 como solución del sistema lineal

$$\begin{pmatrix} \langle P_0, P_0 \rangle & \langle P_0, P_1 \rangle & \langle P_0, P_2 \rangle \\ \langle P_1, P_0 \rangle & \langle P_1, P_1 \rangle & \langle P_1, P_2 \rangle \\ \langle P_2, P_0 \rangle & \langle P_2, P_1 \rangle & \langle P_2, P_2 \rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f(x) \rangle \\ \langle P_1, f(x) \rangle \\ \langle P_2, f(x) \rangle \end{pmatrix}$$

Que para el caso anterior nos daría el sistema

$$\begin{pmatrix} \sum_{k=1}^5 1 \cdot 1 & \sum_{k=1}^5 1 \cdot x_k & \sum_{k=1}^5 1 \cdot x_k^2 \\ \sum_{k=1}^5 x_k \cdot 1 & \sum_{k=1}^5 x_k \cdot x_k & \sum_{k=1}^5 x_k \cdot x_k^2 \\ \sum_{k=1}^5 x_k^2 \cdot 1 & \sum_{k=1}^5 x_k^2 \cdot x_k & \sum_{k=1}^5 x_k^2 \cdot x_k^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 1 \cdot f(x_k) \\ \sum_{k=1}^5 x_k \cdot f(x_k) \\ \sum_{k=1}^5 x_k^2 \cdot f(x_k) \end{pmatrix}$$

O lo que es lo mismo

$$\begin{pmatrix} \sum_{k=1}^5 1 & \sum_{k=1}^5 x_k & \sum_{k=1}^5 x_k^2 \\ \sum_{k=1}^5 x_k & \sum_{k=1}^5 x_k^2 & \sum_{k=1}^5 x_k^3 \\ \sum_{k=1}^5 x_k^2 & \sum_{k=1}^5 x_k^3 & \sum_{k=1}^5 x_k^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^5 1 \cdot x_k^4 \\ \sum_{k=1}^5 x_k \cdot x_k^4 \\ \sum_{k=1}^5 x_k^2 \cdot x_k^4 \end{pmatrix}$$

Calculamos las sumas utilizando la tabla siguiente

| k | 1 | x_k | x_k^2 | x_k^3 | x_k^4 | x_k^5 | x_k^6 |
|----------|-----|-------|---------|---------|---------|----------|----------|
| 1 | 1.0 | -1.0 | 1.00 | -1.000 | 1.0000 | -1.00000 | 1.000000 |
| 2 | 1.0 | -0.5 | 0.25 | -0.125 | 0.0625 | -0.03125 | 0.015625 |
| 3 | 1.0 | 0.0 | 0.00 | 0.000 | 0.0000 | 0.00000 | 0.000000 |
| 4 | 1.0 | 0.5 | 0.25 | 0.125 | 0.0625 | 0.03125 | 0.015625 |
| 5 | 1.0 | 1.0 | 1.00 | 1.000 | 1.0000 | 1.00000 | 1.000000 |
| Σ | 5.0 | 0.0 | 2.50 | 0.000 | 2.1250 | 0.00000 | 2.031250 |

Sustituyendo los resultados

$$\begin{pmatrix} 5 & 0 & 2.5 \\ 0 & 2.5 & 0 \\ 2.5 & 0 & 2.125 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2.125 \\ 0 \\ 2.03125 \end{pmatrix}$$

Si multiplicamos la primera fila por $-\frac{1}{2}$ y la sumamos a la tercera, es decir

$$e_3 \leftarrow e_3 - \frac{e_1}{2}$$

$$\begin{pmatrix} 5 & 0 & 2.5 \\ 0 & 2.5 & 0 \\ 0 & 0 & 0.875 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2.125 \\ 0 \\ 0.96875 \end{pmatrix}$$

Que es el sistema triangular

$$\begin{array}{rcl} 5a_0 & + 2.5a_2 & = 2.125 \\ 2.5a_1 & & = 0 \\ 0.875a_2 & & = 0.96875 \end{array}$$

Y si resolvemos empezando por la última ecuación

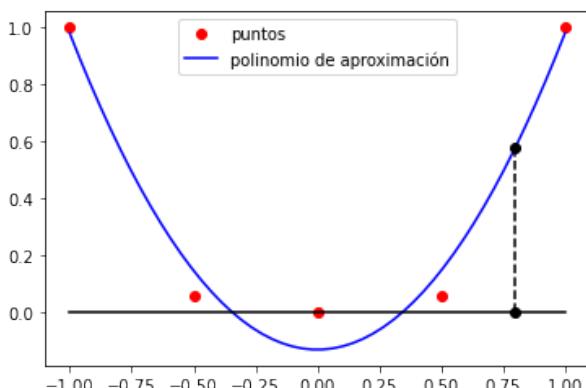
$$\begin{aligned} a_2 &= \frac{0.96875}{0.875} = 1.10714 \\ a_1 &= 0 \\ a_0 &= \frac{2.125 - (2.5)(1.10714)}{5} = -0.12856 \end{aligned}$$

Y como el polinomio de grado dos que aproxima a esta función era

$$P(x) = a_0 \cdot P_0(x) + a_1 \cdot P_1(x) + a_2 \cdot P_2(x) = a_0 \cdot 1 + a_1 \cdot x + a_2 \cdot x^2$$

sustituyendo los valores de a_0 , a_1 y a_2 el polinomio que aproxima los puntos por mínimos cuadrados es

$$P(x) = -0.12856 + 1.10714x^2$$



2. Aproximar el valor en el punto $x=0.8$ y calcular el error relativo y absoluto al usar el polinomio de aproximación en lugar de la función

Si x es el valor real y x^* el aproximado:

- Error absoluto

$$e_a = |x - x^*|$$

- Error relativo

$$e_r = \frac{e_a}{|x|}$$

El error relativo está dado en tanto por uno. Si lo multiplicamos por cien vendrá dado en porcentaje.

$$x = f(0.8) = 0.8^4 = 0.4096 \quad x^* = P(0.8) = -0.12856 + 1.10714(0.8)^2 = 0.58$$

Por lo tanto

- Error absoluto

$$e_a = |0.4096 - 0.5800| = 0.1704$$

- Error relativo

$$e_r = \frac{e_a}{|x|} = \frac{0.1704}{0.4096} = 0.41 = 41\%$$

que es un error bastante grande. Pero estamos approximando una función con solo 5 puntos y un polinomio de grado 2 no es necesariamente la mejor approximación en este caso.

TEMA 4

DERIVACIÓN E INTEGRACIÓN NUMÉRICA

4.1 Fórmulas en diferencias finitas

Ejercicio 4.1.1

Si $f(x) = e^x$

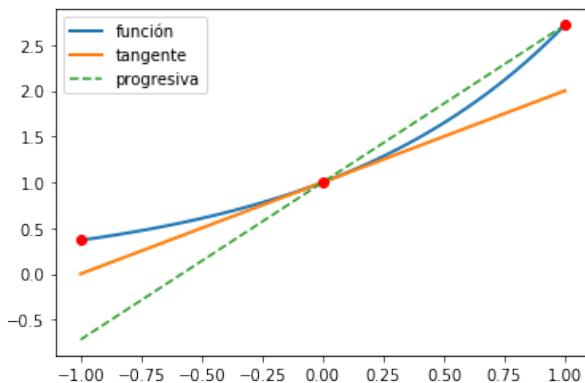
1. Aproximar la derivada en $x_0 = 0$ con $h = 1$ y $h = 0.1$ usando las fórmulas centrada, progresiva y regresiva.
2. Calcular el error absoluto y relativo en cada caso.
3. Estudiar el orden de cada una de las tres fórmulas.

A partir de la definición de derivada

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

podemos obtener tres aproximaciones o *fórmulas en diferencias finitas*

1.a Cálculo de la derivada aproximada con la fórmula progresiva



Si $x = x_0 + h$ con $h > 0$ entonces $h = x - x_0$ y

$$f'(x_0) \approx \frac{f(x_0 + h) - f(x_0)}{h}$$

Para $x_0 = 0$ y $h = 1$

$$f'(0) \approx \frac{f(0 + 1) - f(0)}{1} = \frac{2.72 - 1}{1} = 1.72$$

Y para $x_0 = 0$ con $h = 0.1$

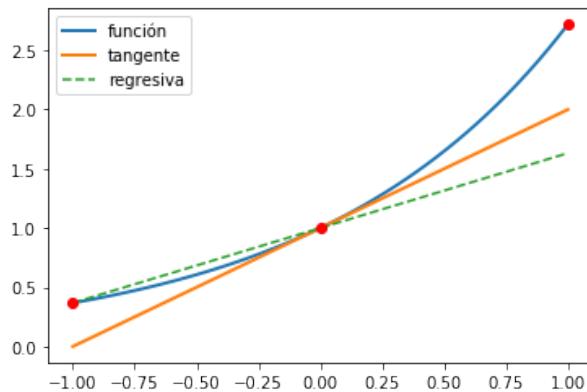
$$f'(0) \approx \frac{f(0 + 0.1) - f(0)}{h} = \frac{1.11 - 1}{0.1} = 1.05$$

Para estas fórmulas, geométricamente

- La derivada exacta de una función f en un punto x_0 nos da la pendiente de la recta *tangente* a la curva en el punto.
- La derivada aproximada de una función f en un punto x_0 nos da la pendiente de la recta *secante* a la curva en los dos puntos que usamos en la fórmula.

Por lo tanto, cuanto más parecidas sean estas dos rectas, más parecidos serán el valor exacto y la aproximación. En el dibujo, por ejemplo, son notablemente diferentes porque h es muy grande. Si los cálculos no tuvieran error (con aritmética finita sí que lo tienen), cuanto menor sean h menor es el error de la aproximación.

1.b Cálculo de la derivada aproximada con la fórmula regresiva



Si $x = x_0 - h$ con $h > 0$ entonces $-h = x - x_0$ y

$$f'(x_0) \approx \frac{f(x_0 - h) - f(x_0)}{-h} = \frac{f(x_0) - f(x_0 - h)}{h}$$

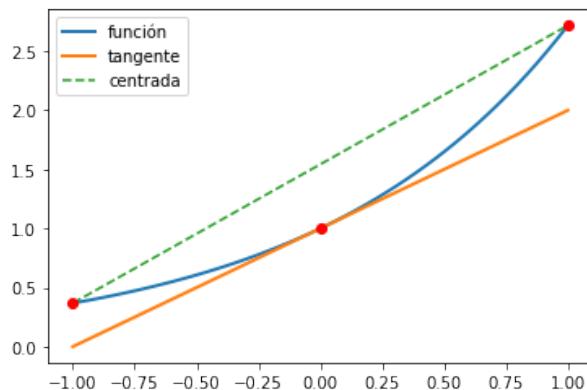
Para $x_0 = 0$ y $h = 1$

$$f'(0) \approx \frac{f(0) - f(0 - 1)}{1} = \frac{1. - 0.37}{1} = 0.63$$

Para $x_0 = 0$ y $h = 0.1$

$$f'(0) \approx \frac{f(0) - f(0 - 0.1)}{0.1} = \frac{1. - 0.905}{0.1} = 0.95$$

1.c Cálculo de la derivada aproximada con la fórmula centrada



Si hacemos un promedio de la fórmula progresiva y regresiva

$$f'(x_0) \approx \frac{1}{2} \left(\frac{f(x_0 + h) - f(x_0)}{h} + \frac{f(x_0) - f(x_0 - h)}{h} \right) = \frac{f(x_0 + h) - f(x_0 - h)}{2h}$$

Para $x_0 = 0$ y $h = 1$

$$f'(0) \approx \frac{f(0+1) - f(0-1)}{2(1)} = \frac{2.72 - 0.37}{2(1)} = 1.18$$

Para $x_0 = 0$ y $h = 0.1$

$$f'(0) \approx \frac{f(0+0.1) - f(0-0.1)}{2(0.1)} = \frac{1.1052 - 0.9048}{2(0.1)} = 1.00$$

En este caso la pendiente de la recta tangente y la recta secante, a pesar de ser h grande, es bastante parecida, por lo que cabe esperar que el valor exacto y el aproximado se parezcan más que en los casos anteriores. Esto se debe a que esta es una fórmula de *orden dos*, mientras que las dos fórmulas anteriores son de *orden uno*.

2. Cálculo del error

Si y es el valor real y y^* el aproximado:

- Para $h = 1$

- Fórmula progresiva
 - * Error absoluto

$$e_a = |y - y^*| = |1.00 - 1.72| = 0.72$$

- * Error relativo

$$e_r = \frac{e_a}{|y|} = \frac{0.72}{1} = 0.72 = 72\%$$

- Fórmula regresiva
 - * Error absoluto

$$e_a = |y - y^*| = |1.00 - 0.63| = 0.37$$

- * Error relativo

$$e_r = \frac{e_a}{|y|} = \frac{0.37}{1} = 0.37 = 37\%$$

- Fórmula centrada
 - * Error absoluto

$$e_a = |y - y^*| = |1.00 - 1.18| = 0.18$$

- * Error relativo

$$e_r = \frac{e_a}{|y|} = \frac{0.18}{1} = 0.18 = 18\%$$

- Para $h = 0.1$
 - Fórmula progresiva
 - * Error absoluto

$$e_a = |y - y^*| = |1.00 - 1.05| = 0.05$$

- * Error relativo

$$e_r = \frac{e_a}{|y|} = \frac{0.05}{1} = 0.05 = 5\%$$

- Fórmula regresiva
 - * Error absoluto

$$e_a = |y - y^*| = |1.00 - 0.95| = 0.05$$

- * Error relativo

$$e_r = \frac{e_a}{|y|} = \frac{0.37}{1} = 0.05 = 5\%$$

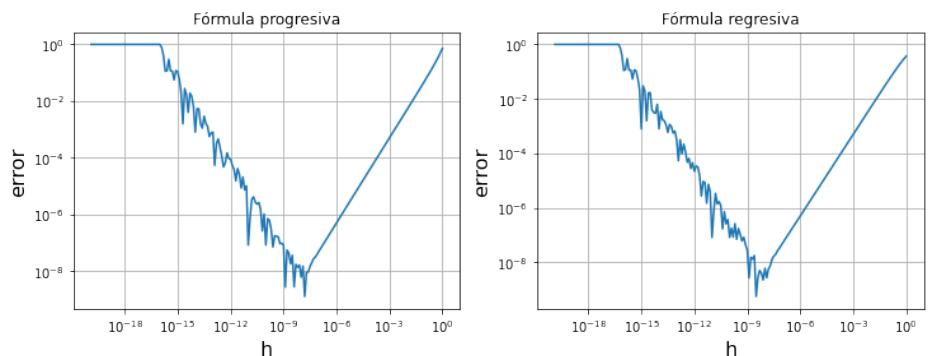
- Fórmula centrada
 - * Error absoluto

$$e_a = |y - y^*| = |1.000 - 1.002| = 0.002$$

- * Error relativo

$$e_r = \frac{e_a}{|y|} = \frac{0.002}{1} = 0.002 = 0.2\%$$

Vemos que para un mismo h la fórmula centrada da un error menor. Y también que cuanto más pequeño el h menor el error (hasta un cierto punto, porque si h es demasiado pequeño dominan los errores de redondeo y el error vuelve a aumentar).



3.a Estudio del orden de las fórmula progresiva

Decimos que una fórmula de derivación numérica es de orden n si el error absoluto

$$E_h = K h^n$$

Como h es un valor pequeño, cuanto mayor sea n más pequeño será el error de la fórmula (si $h = 0.1$ entonces $h^2 = 0.01$ y $h^3 = 0.001$) La fórmula progresiva es

$$f'(a) \approx \frac{f(a+h) - f(a)}{h}$$

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

$$f(a+h) - f(a) = f'(a) \frac{h}{1!} + f''(c) \frac{h^2}{2!}$$

y dividiendo por h

$$\frac{f(a+h) - f(a)}{h} = f'(a) + f''(c) \frac{h}{2}$$

o también

$$-f''(c) \frac{h}{2} = f'(a) - \frac{f(a+h) - f(a)}{h}$$

Y como el error es el valor exacto menos el aproximado

$$E_h = f'(a) - \frac{f(a+h) - f(a)}{h}$$

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.b Estudio del orden de las 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

$$\frac{f(a) - f(a-h)}{h} = f'(a) - f''(c) \frac{h}{2}$$

o también

$$f''(c) \frac{h}{2} = f'(a) - \frac{f(a) - f(a-h)}{h}$$

y como el error es el valor exacto menos el aproximado

$$E_h = f'(a) - \frac{f(a) - f(a-h)}{h}$$

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.c Estudio del orden de las 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 $2h$

$$\frac{f(a + h) - f(a - h)}{2h} = f'(a) + f'''(c_3) \frac{h^2}{6}$$

o también

$$-f'''(c_3) \frac{h^2}{6} = f'(a) - \frac{f(a + h) - f(a - h)}{2h}$$

Y como el error es el valor exacto menos el aproximado

$$E_h = f'(a) - \frac{f(a + h) - f(a - h)}{2h}$$

y 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.

4.2 Fórmulas interpolatorias de derivación numérica

Ejercicio 4.2.1

Supongamos que tenemos tres puntos

$$(x_0, y_0) \quad (x_1, y_1) \quad (x_2, y_2)$$

con

$$x_1 = x_0 + h \quad x_2 = x_0 + 2h \quad 0 < h < 1$$

1. Construir fórmulas que aproximen $f'(x_0)$, $f'(x_1)$, $f'(x_2)$ que utilicen sólo estos tres puntos y que sean de orden 2.
2. Construir también una fórmula que aproxime la derivada segunda.
3. Utilizarlas para aproximar la derivada primera y segunda de $f(x) = \ln x$ en 1.5 con $h = 0.1$.

1. Construir fórmulas que aproximen $f'(x_0)$, $f'(x_1)$, $f'(x_2)$

1.a Con números

Veamos primero un ejemplo con números. Luego generalizaremos a tres nodos cualesquiera.

| k | 0 | 1 | 2 |
|-----------------|--------|--------|--------|
| x_k | 1.4 | 1.5 | 1.6 |
| $y_k = \ln x_k$ | 0.3365 | 0.4055 | 0.4700 |

tenemos que

$$(x_0, y_0) = (1.4, 0.3365) \quad (x_1, y_1) = (1.5, 0.4055) \quad y \quad (x_2, y_2) = (1.6, 0.4700)$$

Por lo tanto $h = x_1 - x_0 = x_2 - x_1 = 0.1$. Calculemos la parábola que pasa por estos tres puntos usando el polinomio interpolante de Newton

$$P_2(x) = [y_0] + [y_0, y_1](x - x_0) + [y_0, y_1, y_2](x - x_0)(x - x_1)$$

Necesitamos calcular los coeficientes con la tabla de diferencias divididas

$$\begin{array}{cc} x_0 & y_0 \end{array}$$

$$[y_0, y_1] = \frac{y_1 - y_0}{x_1 - x_0}$$

$$\begin{array}{cc} x_1 & y_1 \end{array}$$

$$[y_0, y_1, y_2] = \frac{[y_1, y_2] - [y_0, y_1]}{x_2 - x_0}$$

$$[y_1, y_2] = \frac{y_2 - y_1}{x_2 - x_1}$$

$$\begin{array}{cc} x_2 & y_2 \end{array}$$

Cambiando las variables por sus valores

$$1.4 \quad [0.3365]^{c_0}$$

$$\frac{0.4055 - 0.3365}{0.1} = [0.6900]^{c_1}$$

$$1.5 \quad 0.4055$$

$$\frac{0.6450 - 0.6900}{0.2} = [-0.2250]^{c_2}$$

$$\frac{0.4700 - 0.4055}{0.1} = 0.6450$$

$$1.6 \quad 0.4700$$

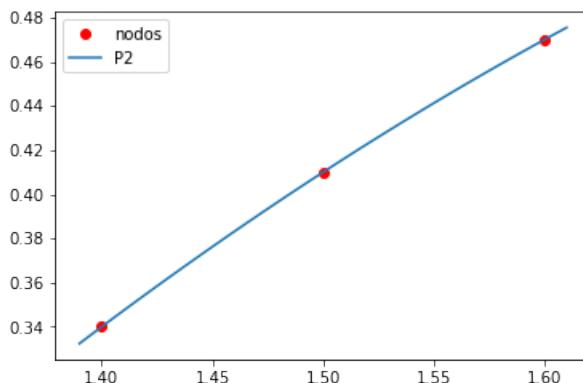
Y el polinomio de interpolación es

$$P_2(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1)$$

y sustituyendo los valores de la tabla

$$P_2(x) = 0.3365 + 0.6900(x - 1.4) - 0.2250(x - 1.4)(x - 1.5)$$

Y este es el polinomio de interpolación que pasa por los tres puntos



Ahora calculamos la derivada de P₂

$$P'_2(x) = 0.6900 - 0.2250[(x - 1.4) + (x - 1.5)]$$

y su derivada segunda

$$P''_2(x) = -0.2250[2] = -0.4500$$

Y ya podemos usarlo para estimar derivadas en los nodos

- Si usamos $x_0 = 1.4$ estaremos usando una **fórmula progresiva**

$$f'(1.4) \approx P'_2(1.4) = 0.6900 - 0.2250[(1.4 - 1.4) + (1.4 - 1.5)] = 0.7125$$

- Si usamos $x_1 = 1.5$ estaremos usando una **fórmula centrada**

$$f'(1.5) \approx P'_2(1.5) = 0.6900 - 0.2250[(1.5 - 1.4) + (1.5 - 1.5)] = 0.6675$$

- Si usamos $x_2 = 1.6$ estaremos usando una **fórmula regresiva**

$$f'(1.6) \approx P'_2(1.6) = 0.6900 - 0.2250[(1.6 - 1.4) + (1.6 - 1.5)] = 0.6225$$

Derivando otra vez, tendremos una aproximación de la derivada segunda

$$P''_2(x) = -0.2250(2) = -0.4500$$

- La usaremos para el punto medio $x_1 = 1.5$ y será una **fórmula centrada**

$$f''(1.5) \approx P''_2(1.5) = -0.4500$$

Comparemos con los valores exactos. Tomando

$$f'(x) = \frac{1}{x} \quad f''(x) = -\frac{1}{x^2}$$

| | $f'(1.4)$ | $f'(1.5)$ | $f'(1.6)$ | $f''(1.5)$ |
|------------|-----------|-----------|-----------|------------|
| exacta | 0.7142 | 0.6667 | 0.6250 | -0.4444 |
| aproximada | 0.7125 | 0.6675 | 0.6225 | -0.4500 |
| error | 0.0017 | 0.0008 | 0.0025 | 0.0056 |

Aunque el menor error lo vuelve a dar la fórmula centrada, todos los errores son, aproximadamente, del mismo orden de magnitud ya que todas las fórmulas son de orden dos.

1.b Con variables

Ahora generalizamos a tres nodos cualesquiera

Si usamos un polinomio de interpolación de f de 2º grado en x_0, x_1, x_2 , siendo $y_j = f(x_j)$, el polinomio de interpolación en la forma de Newton es:

$$p(x) = [y_0] + [y_0, y_1](x - x_0) + [y_0, y_1, y_2](x - x_0)(x - x_1)$$

Necesitamos calcular los coeficientes con la tabla de diferencias divididas

$$\begin{array}{ccccc}
x_0 & \boxed{y_0} & c_0 & & \\
& & & & \\
& \boxed{[y_0, y_1]} & c_1 & & \\
& \boxed{[y_0, y_1, y_2]} & c_2 & & \\
x_1 & y_1 & & & \\
& & & & \\
& \boxed{[y_1, y_2]} & & & \\
& \boxed{x_2 - x_1} & & & \\
x_2 & y_2 & & &
\end{array}$$

con

- El primer coeficiente del polinomio

$$c_0 = y_0$$

- El segundo coeficiente del polinomio

$$c_1 = \frac{y_1 - y_0}{x_1 - x_0} = \frac{y_1 - y_0}{h},$$

- Y el tercer coeficiente

$$c_2 = \frac{[y_1, y_2] - [y_0, y_1]}{x_2 - x_0} = \frac{1}{2h} \left(\frac{y_2 - y_1}{h} - \frac{y_1 - y_0}{h} \right) = \frac{1}{2h^2} (y_0 - 2y_1 + y_2).$$

Resumiendo

$$c_0 = y_0, \quad c_1 = \frac{y_1 - y_0}{h}, \quad c_2 = \frac{1}{2h^2} (y_0 - 2y_1 + y_2).$$

Como

$$p(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1)$$

Si derivamos $p(x)$

$$f'(x) \approx p'(x) = c_1 + c_2((x - x_0) + (x - x_1))$$

Fórmula progresiva

y para el caso particular del punto x_0

$$f'(x_0) \approx p'(x_0) = c_1 + c_2((x_0 - x_0) + (x_0 - x_1)),$$

Por lo tanto

$$f'(x_0) \approx \frac{y_1 - y_0}{h} + \frac{1}{2h^2} (y_0 - 2y_1 + y_2)(-h),$$

Y finalmente

$$f'(x_0) \approx \frac{-3y_0 + 4y_1 - y_2}{2h}$$

Fórmula centrada

Y si aproximamos la derivada en el punto x_1

$$f'(x_1) \approx p'(x_1) = c_1 + c_2((x_1 - x_0) + (x_1 - x_1)),$$

Por lo tanto

$$f'(x_1) \approx \frac{y_1 - y_0}{h} + \frac{1}{2h^2} (y_0 - 2y_1 + y_2)h,$$

Y finalmente

$$f'(x_1) \approx \frac{-y_0 + y_2}{2h}$$

Fórmula regresiva

Y si aproximamos la derivada en el punto x_2

$$f'(x_2) \approx p'(x_2) = c_1 + c_2((x_2 - x_0) + (x_2 - x_1)),$$

Por lo tanto

$$f'(x_2) \approx \frac{y_1 - y_0}{h} + \frac{1}{2h^2}(y_0 - 2y_1 + y_2)(2h + h),$$

Y finalmente

$$f'(x_2) \approx \frac{y_0 - 4y_1 + 3y_2}{2h}$$

2. Construir una fórmula que aproxime la derivada segunda

Si derivamos $p'(x)$

$$f'(x) \approx p'(x) = [y_0, y_1] + [y_0, y_1, y_2]((x - x_0) + (x - x_1)),$$

o también

$$f'(x) \approx p'(x) = c_1 + c_2((x - x_0) + (x - x_1)),$$

Tenemos que

$$f''(x) \approx p''(x) = 2c_2,$$

Por lo tanto

$$f''(x_1) \approx p''(x_1) = 2c_2 = \frac{1}{h^2}(y_0 - 2y_1 + y_2),$$

es decir

$$f''(x_1) \approx \frac{1}{h^2}(y_0 - 2y_1 + y_2)$$

O si llamamos $x_1 = a$ entonces $x_0 = a - h$ y $x_2 = a + h$ y la fórmula se puede escribir también como $f''(a) \approx p''(a)$

$$f''(a) \approx \frac{f(a - h) - 2f(a) + f(a + h)}{h^2}$$

3. Aproximar la derivada primera y segunda de $f(x) = \ln x$ en 1.5 con $h = 0.1$

- La fórmula progresiva es

$$f'(x_0) \approx \frac{-3y_0 + 4y_1 - y_2}{2h}.$$

Si el punto es $x_0 = 1.5$ entonces $x_1 = 1.6$ y $x_2 = 1.7$ y tomamos los valores

| k | 0 | 1 | 2 |
|-----------------|--------|--------|--------|
| x_k | 1.5 | 1.6 | 1.7 |
| $y_k = \ln x_k$ | 0.4055 | 0.4700 | 0.5306 |

que sustituyendo en la fórmula es

$$f'(1.5) \approx \frac{-3(0.4055) + 4(0.4700) - (0.5306)}{2(0.1)} = 0.6645$$

- La fórmula centrada es

$$f'(x_1) \approx \frac{-y_0 + y_2}{2h}.$$

Si el punto es $x_1 = 1.5$ entonces $x_0 = 1.4$ y $x_2 = 1.6$ y tomamos los valores

| k | 0 | 1 | 2 |
|-----------------|--------|--------|--------|
| x_k | 1.4 | 1.5 | 1.6 |
| $y_k = \ln x_k$ | 0.3365 | 0.4055 | 0.4700 |

que sustituyendo en la fórmula es

$$f'(1.5) \approx \frac{-0.3365 + 0.4700}{2(0.1)} = 0.6675$$

- La fórmula regresiva es

$$f'(x_2) \approx \frac{y_0 - 4y_1 + 3y_2}{2h}.$$

Si el punto es $x_0 = 1.5$ entonces $x_1 = 1.4$ y $x_0 = 1.3$ y tomamos los valores

| k | 0 | 1 | 2 |
|-----------------|--------|--------|--------|
| x_k | 1.3 | 1.4 | 1.5 |
| $y_k = \ln x_k$ | 0.2624 | 0.3365 | 0.4055 |

que sustituyendo en la fórmula es

$$f'(1.5) \approx \frac{0.2624 - 4(0.3365) + 3(0.4055)}{2(0.1)} = 0.6645$$

Y el valor exacto de la primera derivada es

$$f(x) = \ln x \quad f'(x) = \frac{1}{x} \quad f'(1.5) = \frac{1}{1.5} = 0.6667$$

- Para aproximar la **derivada segunda**, la fórmula que construimos es **centrada**, por lo tanto si el punto es $x_1 = 1.5$ entonces $x_0 = 1.4$ y $x_2 = 1.6$ y tomamos los valores

| k | 0 | 1 | 2 |
|-----------------|--------|--------|--------|
| x_k | 1.4 | 1.5 | 1.6 |
| $y_k = \ln x_k$ | 0.3365 | 0.4055 | 0.4700 |

Y la aproximación es

$$f''(x_1) \approx \frac{1}{h^2} (y_0 - 2y_1 + y_2) = \frac{1}{0.1^2} (0.3365 - 2(0.4055) + 0.4700) = -0.4500$$

Y el valor exacto de la derivada segunda es

$$f(x) = \ln x \quad f'(x) = \frac{1}{x} \quad f''(x) = -\frac{1}{x^2} \quad f''(1.5) = -\frac{1}{1.5^2} = -0.4444$$

Comparemos los resultados

| | $f'(1.5)$ progresiva | $f'(1.5)$ centrada | $f'(1.5)$ regresiva | $f''(1.5)$ centrada |
|------------|-------------------------|-----------------------|------------------------|------------------------|
| aproximada | 0.6645 | 0.6675 | 0.6644 | -0.4500 |
| exacta | | 0.6667 | | -0.4444 |
| error | 0.0022 | 0.0008 | 0.0023 | 0.0056 |

Aunque el menor error lo vuelve a dar la fórmula centrada, todos los errores son, aproximadamente, del mismo orden de magnitud, las milésimas, ya que todas las fórmulas son de orden dos.

4.3 Derivación numérica de funciones de dos variables

Ejercicio 4.3.1

Dada la función $f(x, y) = x^3 + y^3$, calcular una aproximación de

1. El gradiente $\nabla f(x, y)$
 2. El laplaciano $\Delta f(x, y)$
- para $(x_0, y_0) = (1, 2)$ con $h_x = h_y = 0.1$.

1. Gradiente $\nabla f(x, y)$

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$

Calculamos las derivadas parciales

$$\frac{\partial f}{\partial x} = 3x^2 \quad \frac{\partial f}{\partial y} = 3y^2$$

que para el valor $(x_0, y_0) = (1, 2)$

$$\left(\frac{\partial f}{\partial x} \right)_{(1,2)} = 3(1)^2 = 3 \quad \left(\frac{\partial f}{\partial y} \right)_{(1,2)} = 3(2)^2 = 12 \quad \boxed{\nabla f(1, 2) = (3, 12)}$$

En el **gradiente aproximado** la primera componente del vector será:

$$\begin{aligned} \frac{\partial f}{\partial x} &\approx \frac{f(x_m + h_x, y_n) - f(x_m - h_x, y_n)}{2h_x} = \frac{f(1 + 0.1, 2) - f(1 - 0.1, 2)}{2(0.1)} = \\ &= \frac{(1 + 0.1)^3 + (2)^3 - ((1 - 0.1)^3 + (2)^3)}{2(0.1)} = \\ &= \frac{9.331 - 8.729}{0.2} = \frac{0.602}{0.2} = 3.01 \end{aligned}$$

Y la segunda componente

$$\begin{aligned} \frac{\partial f}{\partial y} &\approx \frac{f(x_m, y_n + h_y) - f(x_m, y_n - h_y)}{2h_y} = \frac{f(1, 2 + 0.1) - f(1, 2 - 0.1)}{2(0.1)} = \\ &= \frac{(1)^3 + (2 + 0.1)^3 - ((1)^3 + (2 - 0.1)^3)}{2(0.1)} = \\ &= \frac{10.261 - 7.859}{0.2} = \frac{2.401}{0.2} = 12.01 \end{aligned}$$

Por lo tanto

$$\boxed{\nabla f(1, 2) \approx (3.01, 12.01)}$$

2. Laplaciano $\Delta f(x, y)$

El laplaciano exacto viene dado por

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

Como

$$\frac{\partial f}{\partial x} = 3x^2 \quad \frac{\partial f}{\partial y} = 3y^2$$

Se tiene que

$$\frac{\partial^2 f}{\partial x^2} = 6x \quad \frac{\partial^2 f}{\partial y^2} = 6y$$

Que para $(x_0, y_0) = (1, 2)$

$$\left(\frac{\partial^2 f}{\partial x^2} \right)_{(1,2)} = 6(1) = 6 \quad \left(\frac{\partial^2 f}{\partial y^2} \right)_{(1,2)} = 6(2) = 12$$

Y por lo tanto

$$\Delta f_{(1,2)} = \left(\frac{\partial^2 f}{\partial x^2} \right)_{(1,2)} + \left(\frac{\partial^2 f}{\partial y^2} \right)_{(1,2)} = 6 + 12 = 18 \quad \boxed{\Delta f(1, 2) = 18}$$

El laplaciano aproximado tendrá de primera componente:

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2} &\approx \frac{f(x_m + h_x, y_n) - 2f(x_m, y_n) + f(x_m - h_x, y_n)}{h_x^2} = \\ &= \frac{f(1 + 0.1, 2) - 2f(1, 2) + f(1 - 0.1, 2)}{(0.1)^2} = \\ &= \frac{\left((1 + 0.1)^3 + (2)^3\right) - 2\left((1)^3 + (2)^3\right) + \left((1 - 0.1)^3 + (2)^3\right)}{(0.1)^3} = \\ &= \frac{9.331 - 18 + 8.729}{0.01} = \frac{0.06}{0.01} = 6. \end{aligned}$$

Y la segunda componente

$$\begin{aligned} \frac{\partial^2 f}{\partial y^2} &\approx \frac{f(x_m, y_n + h_y) - 2f(x_m, y_n) + f(x_m, y_n - h_y)}{h_y^2} = \\ &= \frac{f(1, 2 + 0.1) - 2f(1, 2) + f(1, 2 - 0.1)}{(0.1)^2} = \\ &= \frac{\left((1)^3 + (2 + 0.1)^3\right) - 2\left((1)^3 + (2)^3\right) + \left((1)^3 + (2 - 0.1)^3\right)}{(0.1)^2} = \\ &= \frac{10.261 - 18 + 7.859}{0.01} = \frac{0.12}{0.01} = 12 \end{aligned}$$

Por lo tanto

$$\boxed{\Delta f(1, 2) \approx 6 + 12 = 18}$$

4.4 Fórmulas de cuadratura

Fórmulas de Newton-Cotes

Ejercicio 4.4.1

Deducir y estudiar la precisión de:

1. La regla del punto medio.
2. La regla del trapezio.

INTRODUCCIÓN

Fórmulas de cuadratura

Las fórmulas de integración numérica o de cuadratura son de la forma:

$$\int_a^b f(x) dx \approx \omega_0 f(x_0) + \omega_1 f(x_1) + \cdots + \omega_n f(x_N)$$

donde x_0, x_1, \dots, x_N (nodos) son $N + 1$ puntos distintos pertenecientes al intervalo $[a, b]$ y $\omega_0, \omega_1, \dots, \omega_N$ (pesos) son números reales.

Fórmulas interpolatorias

Si P_N es el polinomio que interpola a f en los puntos distintos $x_0, x_1, \dots, x_N \in [a, b]$ y

$$\int_a^b f(x) dx \approx \int_a^b P_N(x) dx = \omega_0 f(x_0) + \omega_1 f(x_1) + \cdots + \omega_n f(x_N)$$

decimos que la fórmula de cuadratura es de tipo interpolatorio.

Fórmulas de cuadratura simples y compuestas

Las fórmulas de cuadratura se llaman simples si la aproximación se hace en el intervalo completo (a, b) , y compuestas si, antes de aplicar la fórmula, dividimos el intervalo (a, b) , en n subintervalos.

Grado de precisión

Una fórmula de cuadratura tiene grado de precisión r si es exacta para

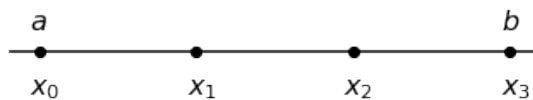
$$f(x) = 1, \quad f(x) = x, \quad f(x) = x^2, \dots, \quad f(x) = x^r$$

pero no es exacta para $f(x) = x^{r+1}$

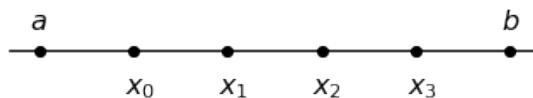
Fórmulas de cuadratura de Newton-Cotes simples

Son fórmulas de cuadratura de tipo interpolatorio, eligiendo los puntos de interpolación (nodos de la fórmula) igualmente separados de una de las dos formas siguientes:

- *Fórmulas cerradas.* Los límites de integración a y b son nodos de la fórmula. Por ejemplo, los nodos para una fórmula cerrada de cuatro nodos serían



- *Fórmulas abiertas.* Ninguno de los límites de integración es nodo de la fórmula. Por ejemplo, los nodos para una fórmula abierta de cuatro nodos serían



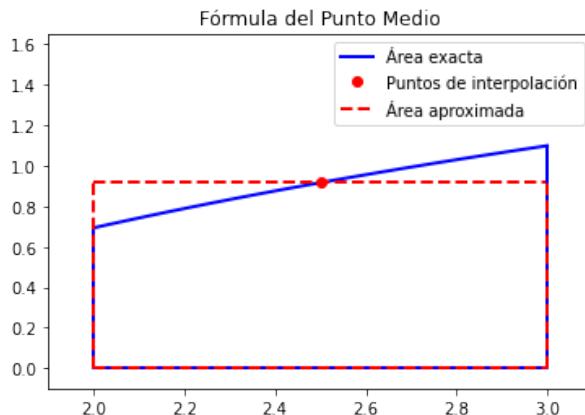
1. La regla del punto medio

La fórmula del punto medio es

$$\int_a^b f(x) dx \approx (b - a) f\left(\frac{a + b}{2}\right)$$

Usar la fórmula del punto medio para integrar una función en un intervalo, equivale a sustituir, dentro de la integral, la función a integrar por el polinomio de interpolación de grado cero, es decir, una recta horizontal que pasa por el punto medio de la función. Sustituimos la función por una recta e integramos. Estamos entonces calculando el área de un rectángulo. Por ejemplo, dada la integral

$$\int_2^3 \ln x dx$$



Construcción de la regla del punto medio

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

$$\begin{aligned} \int_a^b f(x) dx &\approx \int_a^b P_0(x) dx = \int_a^b f\left(\frac{a+b}{2}\right) dx = f\left(\frac{a+b}{2}\right) \int_a^b 1 dx = \\ &= f\left(\frac{a+b}{2}\right) [x]_a^b = f\left(\frac{a+b}{2}\right) (b-a) \end{aligned}$$

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

$$\int_a^b f(x) dx \approx h f\left(\frac{a+b}{2}\right)$$

que es el área del rectángulo de base h y altura $f\left(\frac{a+b}{2}\right)$.

Precisión de la regla del punto medio

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 $1, x, x^2, \dots, x^n$ 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 del punto medio 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 del punto medio 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.

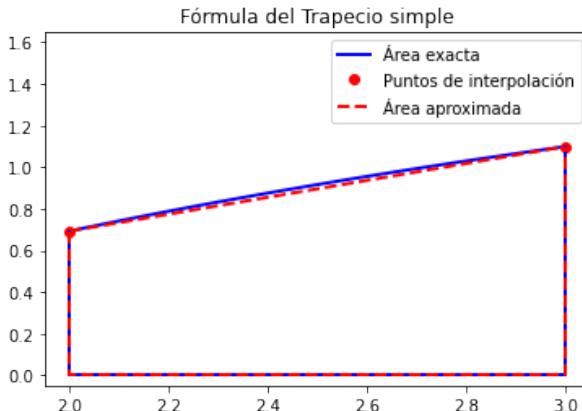
2. La regla del trapecio

La fórmula de los trapecios simple es

$$\int_a^b f(x) dx \approx \frac{b-a}{2} (f(a) + f(b))$$

Usar la fórmula de los trapecios para integrar una función en un intervalo, equivale a sustituir, dentro de la integral, la función a integrar por el polinomio de interpolación de grado uno, que pasa por los puntos de la función de los extremos del intervalo. Es decir, sustituimos la función por una recta e integramos. Estamos entonces calculando el área de un trapecio. Por ejemplo, dada la integral

$$\int_2^3 \ln x dx$$



Construcción de la regla del trapecio

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 que pasa por los puntos $(a, f(a))$ y $(b, f(b))$

$$P_1(x) = f(a) \frac{x-b}{a-b} + f(b) \frac{x-a}{b-a}$$

se tiene que

$$\begin{aligned} \int_a^b f(x) dx &\approx \int_a^b P_1(x) dx = \frac{f(a)}{a-b} \int_a^b (x-b) dx + \frac{f(b)}{b-a} \int_a^b (x-a) dx = \\ &= \frac{f(a)}{a-b} \left[\frac{(x-b)^2}{2} \right]_a^b + \frac{f(b)}{b-a} \left[\frac{(x-a)^2}{2} \right]_a^b = \frac{f(a)}{a-b} \frac{-(a-b)^2}{2} + \frac{f(b)}{b-a} \frac{(b-a)^2}{2} = \\ &= f(a) \frac{-(a-b)}{2} + f(b) \frac{(b-a)}{2} = \frac{b-a}{2} (f(a) + f(b)) \end{aligned}$$

Por lo tanto, la regla del trapecio simple es

$$\int_a^b f(x) dx \approx \frac{b-a}{2} (f(a) + f(b))$$

Y si llamamos $h = b - a$ a la longitud del intervalo de integración, podemos escribir la fórmula

$$\int_a^b f(x) dx \approx \frac{h}{2} (f(a) + f(b))$$

que es el área del trapecio de base h y alturas $f(a)$ y $f(b)$ (el área del trapecio es la base por el promedio de las alturas).

Precisión de la regla del trapecio

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 $1, x, x^2, \dots, x^n$ y no serlo para x^{n+1} . Veamos la fórmula de los Trapecios:

¿Es exacta para $f(x) = 1$? Sí, porque

$$\int_a^b 1 dx = b - a$$

y para este intervalo y esta función la fórmula de los trapecios es

$$\frac{b-a}{2} (f(a) + f(b)) = \frac{b-a}{2} (1 + 1) = b - a$$

¿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

$$\frac{b-a}{2} (f(a) + f(b)) = \frac{b-a}{2} (a + b) = \frac{(b+a)(b-a)}{2} = \frac{b^2 - a^2}{2}$$

¿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

$$\frac{b-a}{2} (f(a) + f(b)) = \frac{b-a}{2} (a^2 + b^2) = \frac{b^3 - ab^2 + a^2b - a^3}{2}$$

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.

Ejercicio 4.4.2

Calcular la integral

$$I = \int_0^3 e^x dx$$

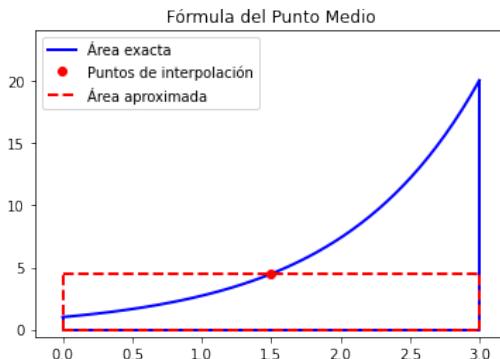
1. Usando la regla del punto medio simple.
2. Usando la regla del trapecio simple.
3. Usando la regla de Simpson simple.
4. ¿Cuál es el error en cada caso?

El valor exacto de la integral es

$$I = \int_0^3 e^x dx = (e^x)_0^3 = e^3 - e^0 = 20.0855 - 1 = 19.0855$$

Veamos los valores obtenidos con las distintas fórmulas de aproximación.

1. Fórmula del punto medio



La fórmula del punto medio es

$$\int_a^b f(x) dx \approx (b - a) f\left(\frac{a + b}{2}\right)$$

Usar la fórmula del punto medio para integrar una función en un intervalo, equivale a sustituir, dentro de la integral, la función a integrar por el polinomio de interpolación de grado cero, es decir, una recta horizontal que pasa por el punto medio de la función. Sustituimos la función por una recta e integramos. Estamos entonces calculando el área de un rectángulo.

Esta fórmula es de la forma

$$\int_a^b f(x) dx \approx \omega_0 f(x_0)$$

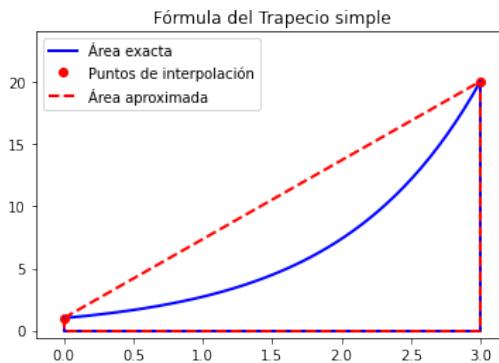
En este caso

$$I = \int_0^3 e^x dx \approx (3 - 0) e^{1.5} = 3 (4.4816) = 13.4451$$

y el error absoluto es

$$\text{Error} = |I - I_{\text{aprox}}| = 19.0855 - 13.4451 = 5.6404$$

2. Fórmula del trapecio



La fórmula de los trapecios simple es

$$\int_a^b f(x) dx \approx (b - a) \frac{f(a) + f(b)}{2}$$

Usar la fórmula de los trapecios para integrar una función en un intervalo, equivale a sustituir, dentro de la integral, la función a integrar por el polinomio de interpolación de grado uno, que pasa por los puntos de la función de los extremos del intervalo. Es decir, sustituimos la función por una recta e integramos. Estamos entonces calculando el área de un trapecio.

Veamos como encaja esta fórmula con la definición inicial

$$\int_a^b f(x) dx \approx \omega_0 f(x_0) + \omega_1 f(x_1)$$

Si escribimos la fórmula en este formato

$$\int_a^b f(x) dx \approx \frac{h}{2} f(a) + \frac{h}{2} f(b) = (b - a) (0.5 f(a) + 0.5 f(b))$$

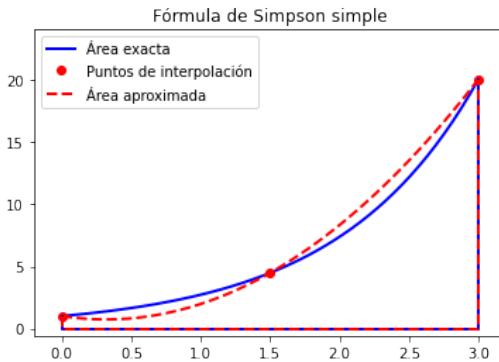
Calculemos la integral pedida

$$I = \int_0^3 e^x dx \approx \frac{3 - 0}{2} (e^0 + e^3) = 1.5 (1 + 20.0855) = 31.6283$$

Y el error es

$$\text{Error} = |I - I_{\text{aprox}}| = |19.0855 - 19.5061| = 0.4206$$

3. Regla de Simpson



La fórmula de Simpson simple es

$$\int_a^b f(x)dx \approx \frac{b-a}{6} \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right)$$

Usar la fórmula de Simpson para integrar una función en un intervalo, equivale a sustituir, dentro de la integral, la función a integrar por el polinomio de interpolación de grado dos, que pasa por los puntos de la función de los extremos y el punto medio del intervalo. Es decir, sustituimos la función por una parábola e integramos.

Veamos como encaja esta fórmula con la definición inicial

$$\int_a^b f(x) dx \approx \omega_0 f(x_0) + \omega_1 f(x_1) + \omega_2 f(x_2)$$

Si escribimos la fórmula en este formato

$$\int_a^b f(x)dx \approx \frac{b-a}{6}f(a) + \frac{4(b-a)}{6}f\left(\frac{a+b}{2}\right) + \frac{b-a}{6}f(b)$$

O también

$$\int_a^b f(x)dx \approx (b-a) \left(\frac{1}{6}f(a) + \frac{4}{6}f\left(\frac{a+b}{2}\right) + \frac{1}{6}f(b) \right)$$

Y la integral aproximada es

$$I = \int_0^3 e^x dx \approx \frac{3}{6}(e^0 + 4e^{1.5} + e^3) = 0.5(1 + 4.4817 + 20.0855) = 19.5061$$

Y el error absoluto es

$$\text{Error} = |I - I_{\text{aprox}}| = |19.0855 - 19.5061| = 0.4206$$

Ejercicio 4.4.3

Calcular la integral

$$\int_0^3 e^x dx$$

usando cinco nodos con

1. La regla del punto medio compuesta.
2. La regla del trapecio compuesta.
3. La regla de Simpson compuesta.

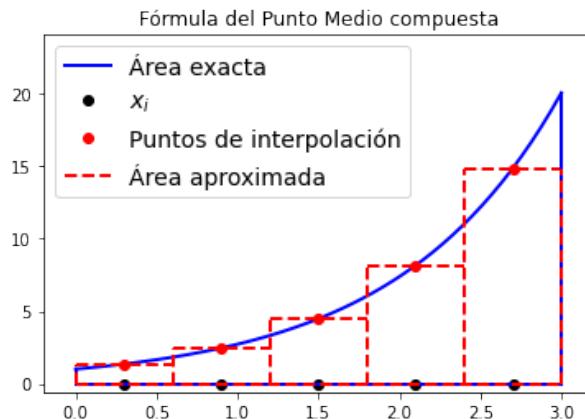
¿Cuál es el error en cada caso?

Como vamos a calcular el error, calculemos el valor exacto de la integral.

$$I = \int_0^3 e^x dx = (e^x)_0^3 = e^3 - e^0 = 20.0855 - 1 = 19.0855$$

Una forma de disminuir el error de las fórmulas simples es aumentar el número de nodos utilizando las **fórmulas compuestas**. Estas se obtienen dividiendo el intervalo $[a, b]$ en n subintervalos y se aplica en cada uno de estos subintervalos una fórmula de cuadratura simple.

1. Fórmula del punto medio compuesta



Para tener 5 nodos con la fórmula del Punto Medio hemos de dividir el intervalo en 5 subintervalos, es decir $n = 5$ y entonces, si $a = 0$ y $b = 3$, se tiene que

$$h = \frac{b-a}{n} = \frac{3-0}{5} = 0.6$$

y entonces los nodos serían

$$\begin{aligned}\bar{x}_1 &= a + \frac{h}{2} = 0 + 0.3 = 0.3 \\ \bar{x}_2 &= \bar{x}_1 + h = 0.3 + 0.6 = 0.9 \\ \bar{x}_3 &= \bar{x}_2 + h = 0.9 + 0.6 = 1.5 \\ \bar{x}_4 &= \bar{x}_3 + h = 1.5 + 0.6 = 2.1 \\ \bar{x}_5 &= \bar{x}_4 + h = 2.1 + 0.6 = 2.7\end{aligned}$$

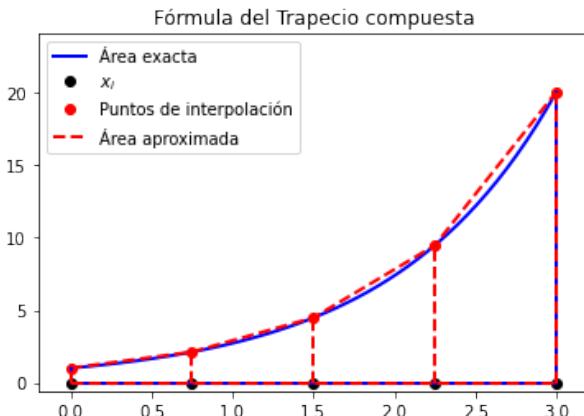
entonces

$$\begin{aligned}I &= \int_a^b f(x)dx = \\ &= \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \int_{x_2}^{x_3} f(x)dx + \int_{x_3}^{x_4} f(x)dx + \int_{x_4}^{x_5} f(x)dx \approx \\ &\approx hf(\bar{x}_1) + hf(\bar{x}_2) + hf(\bar{x}_3) + hf(\bar{x}_4) + hf(\bar{x}_5) = \\ &= h(f(\bar{x}_1) + f(\bar{x}_2) + f(\bar{x}_3) + f(\bar{x}_4) + f(\bar{x}_5)) = \\ &= h(f(0.3) + f(0.9) + f(1.5) + f(2.1) + f(2.7)) = \\ &= 0.6 \left(e^{0.3} + e^{0.9} + e^{1.5} + e^{2.1} + e^{2.7} \right) = \\ &= 0.6 (1.35 + 2.46 + 4.482 + 8.16614.88) = 18.8022\end{aligned}$$

Y el error es

$$Error = |I - I_{aprox}| = 19.0855 - 18.8022 = 0.2833$$

2. Regla del trapecio compuesta



Para tener 5 nodos con la fórmula del Trapecio Compuesta hemos de dividir el intervalo en 4 subintervalos, es decir $n = 4$ y entonces, si $a = 0$ y $b = 3$, se tiene que

$$h = \frac{b - a}{n} = \frac{3 - 0}{4} = 0.75$$

y entonces los nodos serían

$$\begin{aligned}x_0 &= a = 0 \\x_1 &= x_0 + h = 0 + 0.75 = 0.75 \\x_2 &= x_1 + h = 0.75 + 0.75 = 1.5 \\x_3 &= x_2 + h = 1.5 + 0.75 = 2.25 \\x_4 &= x_3 + h = 2.25 + 0.75 = 3 = b\end{aligned}$$

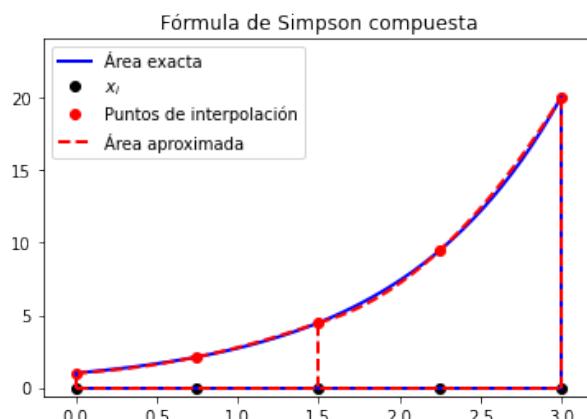
entonces

$$\begin{aligned}I &= \int_a^b f(x)dx = \\&= \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \int_{x_2}^{x_3} f(x)dx + \int_{x_3}^{x_4} f(x)dx \approx \\&\approx \frac{h}{2}(f(x_0) + f(x_1)) + \frac{h}{2}(f(x_1) + f(x_2)) + \frac{h}{2}(f(x_2) + f(x_3)) + \frac{h}{2}(f(x_3) + f(x_4)) = \\&= \frac{h}{2}(f(x_0) + 2(f(x_1) + f(x_2) + f(x_3)) + f(x_4)) = \\&= \frac{h}{2}(f(0) + 2(f(0.75) + f(1.5) + f(2.25)) + f(3)) = \\&= \frac{0.75}{2}(1 + 2(2.117 + 4.482 + 9.488) + 20.086) = 19.9719\end{aligned}$$

Y el error es

$$Error = |I - I_{aprox}| = |19.0855 - 19.9719| = 0.8864$$

3. Regla de Simpson compuesta



La fórmula de Simpson simple es, por ejemplo, para el intervalo $[x_0, x_2]$ con punto medio x_1 y si la longitud de los intervalos $[x_0, x_1]$ y $[x_1, x_2]$ es h

$$\int_{x_0}^{x_2} f(x)dx \approx \frac{x_2 - x_0}{6} (f(x_0) + 4f(x_1) + f(x_2))$$

entonces

$$\int_{x_0}^{x_2} f(x)dx \approx \frac{2h}{6} (f(x_0) + 4f(x_1) + f(x_2))$$

Para tener 5 nodos con la fórmula de Simpson Compuesta hemos de dividir el intervalo en 4 subintervalos, es decir $n = 4$ y entonces, si $a = 0$ y $b = 3$, se tiene que

$$h = \frac{b - a}{n} = \frac{3 - 0}{4} = 0.75$$

y entonces los nodos serían

$$\begin{aligned} x_0 &= a = 0 \\ x_1 &= x_0 + h = 0 + 0.75 = 0.75 \\ x_2 &= x_1 + h = 0.75 + 0.75 = 1.5 \\ x_3 &= x_2 + h = 1.5 + 0.75 = 2.25 \\ x_4 &= x_3 + h = 2.25 + 0.75 = 3 = b \end{aligned}$$

entonces

$$\begin{aligned} I &= \int_a^b f(x)dx = \\ &= \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx \approx \\ &\approx \frac{h}{3} (f(x_0) + 4f(x_1) + f(x_2)) + \frac{h}{3} (f(x_2) + 4f(x_3) + f(x_4)) = \\ &= \frac{h}{3} (f(x_0) + 4(f(x_1) + f(x_3)) + 2f(x_2) + f(x_4)) = \\ &= \frac{h}{3} (f(0) + 4(f(0.75) + f(2.25)) + 2f(1.5) + f(3)) = \\ &= \frac{0.75}{2} (e^0 + 4(e^{0.75} + e^{2.25}) + 2e^{1.5} + e^3) = \\ &= \frac{0.75}{2} (1 + 4(2.117 + 9.488) + 2(4.482) + 20.086) = 19.1170 \end{aligned}$$

Y el error es

$$Error = |I - I_{approx}| = |19.0855 - 19.1170| = 0.0314$$

Ejercicio 4.4.4

Dada la integral

$$I = \int_0^3 (x^3 + 1) dx$$

1. Aproximar su valor mediante la regla del trapecio simple.
2. Aproximar su valor mediante la regla de Simpson simple.
3. Comparar los valores aproximados con el valor exacto ¿Se podría haber predicho alguno de los errores?
4. Al utilizar la regla del trapecio compuesta para aproximar I ¿qué número de subintervalos será suficiente para que el error sea menor que 10^{-6} ?

La integral exacta es

$$I = \int_0^3 (x^3 + 1) dx = \left[\frac{x^4}{4} + x \right]_0^3 = \frac{3^4}{4} + 3 = 23.25$$

1. Regla del trapecio

La fórmula de los trapecios simple es

$$\int_a^b f(x) dx \approx \frac{b-a}{2} (f(a) + f(b))$$

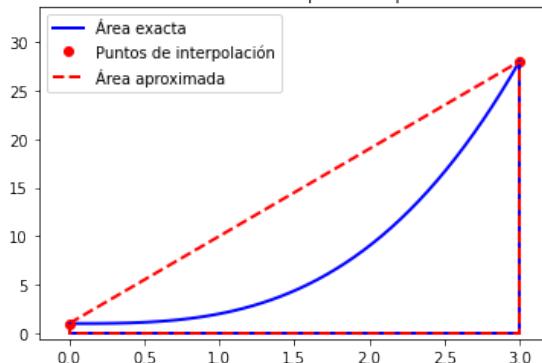
Aplicamos la fórmula

$$\int_0^3 f(x) dx \approx \frac{3-0}{2} (f(0) + f(3)) = \frac{3}{2} ((0^3 + 1) + (3^3 + 1)) = \frac{3}{2} \times 29 = 43.5$$

El error absoluto es

$$e_a = |I - I_{trap}| = |23.25 - 43.5| = 20.25$$

Fórmula del Trapecio simple



2. Regla de Simpson

La fórmula de Simpson simple es

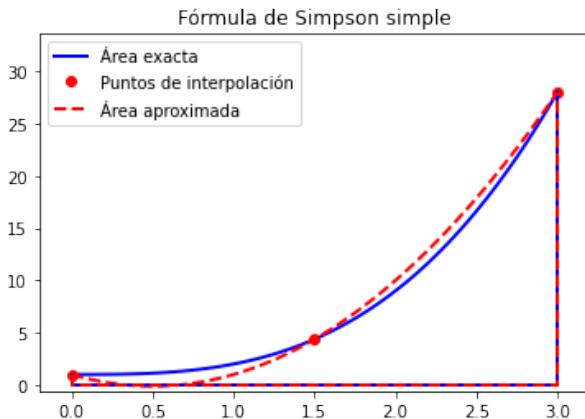
$$\int_a^b f(x)dx \approx \frac{b-a}{6} \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right)$$

$$\int_0^3 f(x)dx \approx \frac{3-0}{6} (f(0) + 4f(1.5) + f(3)) =$$

$$= \frac{3}{6} \left((0^3 + 1) + 4(1.5^3 + 1) + (3^3 + 1) \right) = 23.25$$

El error absoluto es

$$e_a = |I - I_{Simp}| = |23.25 - 23.25| = 0$$



3. Error

Se podía haber previsto que el error sería cero para la fórmula de Simpson porque esta fórmula es de precisión 3 y por tanto exacta para polinomios de grado 3.

4. Número de intervalos

FÓRMULA DEL ERROR DE LA REGLA DE LOS TRAPECIOS

Simple

El error de la regla del trapecio simple es

$$E^T = -f''(c_o) \frac{(b-a)^3}{12} \quad c_o \in (a, b)$$

Compuesta

Si subdividimos $[a, b]$ en n subintervalos iguales de longitud

$$h = \frac{b - a}{n} \quad (1)$$

y aplicamos la regla del trapecio simple en cada uno de ellos, el error será la suma de los errores para cada uno de estos n subintervalos

$$E_h^T = -f''(c_1) \frac{h^3}{12} - f''(c_2) \frac{h^3}{12} - \cdots - f''(c_n) \frac{h^3}{12}$$

sacando factor común

$$E_h^T = -\frac{h^3}{12} (f''(c_1) + f''(c_2) + \cdots + f''(c_n))$$

multiplicando y dividiendo por n

$$E_h^T = -\frac{n h^3}{12} \frac{f''(c_1) + f''(c_2) + \cdots + f''(c_n)}{n}$$

Del teorema del valor intermedio se puede deducir que existe un valor de $c \in (a, b)$ de forma que $f''(c)$ es igual este promedio que aparece en la fórmula anterior. Entonces

$$E_h^T = -\frac{n h^3}{12} f''(c)$$

y como, por (1), $nh = b - a$ la fórmula queda

$$E_h^T = -\frac{h^2}{12} (b - a) f''(c)$$

Si tomamos el error en valor absoluto

$$E_h^T = \frac{h^2}{12} (b - a) |f''(c)|$$

Si hacemos

$$\frac{h^2}{12} (b - a) |f''(c)| < 10^{-6}$$

entonces se verificará que

$$E_h^T < 10^{-6}$$

Si $f(x) = x^3 + 1$ y $[a, b] = [0, 3]$ entonces

$$f'(x) = 3x^2 \quad \text{y} \quad f''(x) = 6x$$

Y se tiene que

$$|f''(c)| = 6c < 6(3) = 18 \quad c \in (0, 3)$$

Por lo tanto, el error en valor absoluto es

$$E_h^T = |f''(c)| (b-a) \frac{h^2}{12} < \frac{18}{12} (b-a) h^2 = \frac{3}{2} (3-0) h^2 < 10^{-6}$$

Es decir

$$\frac{9}{2} h^2 < 10^{-6}$$

que con

$$h = \frac{b-a}{n} = \frac{3}{n}$$

da

$$\frac{9}{2} \left(\frac{3}{n} \right)^2 < 10^{-6}$$

o

$$\frac{81}{2n^2} < 10^{-6}$$

y entonces como

$$a < b, 0 < c \implies ac < bc$$

multiplicando ambos miembros de la desigualdad por n^2 y 10^6 tenemos

$$\frac{81}{2} 10^6 < n^2$$

Y teniendo en cuenta que si tenemos una función h creciente $x_1 < x_2 \implies h(x_1) < h(x_2)$ y como la función $h(x) = \sqrt{x}$ es creciente se tiene que

$$\sqrt{\frac{81}{2} 10^6} < n$$

o lo que es lo mismo

$$6363.96 < n$$

Y tomando

$$n = 6364$$

subintervalos podemos garantizar que el error es menor que 10^{-6} . De hecho, haciendo un programa que aproxime esta integral con la regla del trapecio compuesta y 6364 subintervalos los resultados han sido:

- El valor aproximado es 23.250000499993877
- El valor exacto es 23.25
- El error es $5 \times 10^{-7} = 0.5 \times 10^{-6} < 10^{-6}$

Ejercicio 4.4.5

Dada la integral

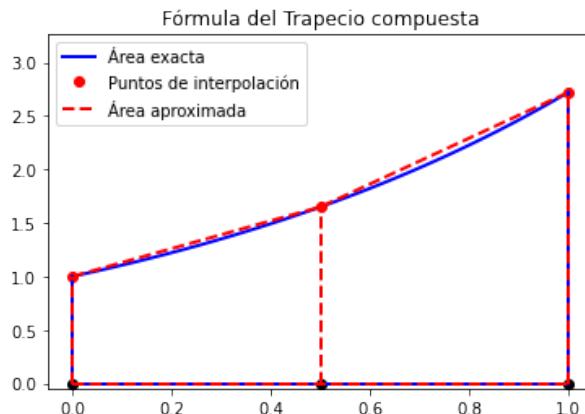
$$I = \int_0^1 e^x dx,$$

1. Obtener su valor aproximado mediante la regla del trapecio compuesta con dos subintervalos.
2. Acotar el error en valor absoluto.
3. Determinar el número n de subintervalos suficientes para que el error sea menor que 10^{-6} .

La integral exacta es

$$I = \int_0^1 e^x dx = [e^x]_0^1 = e^1 - e^0 = 2.7183 - 1 = 1.7183$$

1. Regla de los trapezios compuesta



La regla del trapecio simple es

$$\int_a^b f(x)dx \approx \frac{b-a}{2} (f(a) + f(b))$$

Para dos subintervalos

$$h = \frac{b-a}{n} = \frac{1-0}{2} = 0.5$$

Calculamos los nodos

$$\begin{aligned} x_0 &= a = 0 \\ x_1 &= x_0 + h = 0 + 0.5 = 0.5 \\ x_2 &= x_1 + h = 0.5 + 0.5 = 1 = b \end{aligned}$$

$$\begin{aligned}
 I &= \int_0^1 e^x dx = \int_0^{0.5} e^x dx + \int_{0.5}^1 e^x dx \approx \\
 &\approx \frac{h}{2} (f(x_0) + f(x_1)) + \frac{h}{2} (f(x_1) + f(x_2)) = \frac{h}{2} (f(x_0) + 2f(x_1) + f(x_2)) \\
 &= \frac{0.5}{2} (e^0 + 2e^{0.5} + e^1) = 0.25 (1 + 2(1.65) + 2.72) = 1.75
 \end{aligned}$$

Y el error es

$$e_a = |I - I_{trap}| = 1.75 - 1.72 = 0.03$$

2. Cota error absoluto

La fórmula del error de la regla de los trapecios compuesta es

$$E_h^T = |f''(c)| (b-a) \frac{h^2}{12} \quad \text{con} \quad h = \frac{b-a}{n}$$

$$\begin{aligned}
 \text{Si } f(x) = e^x \text{ y } [a, b] = [0, 1] \text{ entonces } f'(x) = e^x \quad \text{y} \quad f''(x) = e^x \\
 |f''(c)| = e^c < e^1 = e \quad c \in (0, 1)
 \end{aligned}$$

Por lo tanto

$$E_h^T = |f''(c)| (b-a) \frac{h^2}{12} < \frac{e^1}{12} (b-a) h^2 = \frac{e}{12} (1-0) 0.5^2 = 0.06$$

Y, como era de esperar, la cota de error, 0.06, es mayor que el error, 0.03.

3. Número de intervalos

Para aproximar $I = \int_0^1 e^x dx$ con un error menor que 10^{-6}

Si tomamos el error en valor absoluto

$$E_h^T = \frac{h^2}{12} (b-a) |f''(c)|$$

Si hacemos

$$\frac{h^2}{12} (b-a) |f''(c)| < 10^{-6}$$

entonces se verificará que

$$E_h^T < 10^{-6}$$

Como $f'(x) = e^x$ y $f''(x) = e^x$

$$|f''(c)| = e^c < e^1 = e \quad c \in (0, 1)$$

Por lo tanto

$$E_h^T = |f''(c)| (b-a) \frac{h^2}{12} < \frac{e}{12} (b-a) h^2 = \frac{e}{12} (1-0) h^2 < 10^{-6}$$

Es decir

$$\frac{e}{12}h^2 < 10^{-6}$$

Como

$$h = \frac{b-a}{n} = \frac{1}{n}$$

se tiene

$$\frac{e}{12} \frac{1}{n^2} < 10^{-6}$$

y entonces como

$$a < b, 0 < c \implies ac < bc$$

multiplicando ambos miembros de la desigualdad por n^2 y 10^6 tenemos

$$\frac{e}{12} 10^6 < n^2$$

Y teniendo en cuenta que si tenemos una función h creciente $x_1 < x_2 \implies h(x_1) < h(x_2)$ y como la función $h(x) = \sqrt{x}$ es creciente se tiene que

$$\sqrt{\frac{e}{12} 10^6} < n$$

o lo que es lo mismo

$$475.94 < n$$

Y una condición suficiente para que el error sea menor que 10^{-6} es que el número de subintervalos sea $n = 476$.

De hecho, haciendo un programa que aproxime esta integral con la regla del trapecio compuesta y 476 subintervalos los resultados han sido:

- El valor aproximado es 1.7182824604330484
- El valor exacto es 1.7182818284590453
- El error es aproximadamente $6 \times 10^{-7} = 0.6 \times 10^{-6} < 10^{-6}$

4.5 Fórmulas de cuadratura gaussianas

Ejercicio 4.5.1

1. Calcular

$$\int_0^3 e^{-\frac{x^2}{2}} dx$$

usando la fórmula de cuadratura gaussiana con tres nodos:

$$\int_{-1}^1 f(x) dx \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$$

2. ¿Cuál es el grado de precisión de esta fórmula?

INTRODUCCIÓN

Fórmulas de Gauss-Legendre

En la fórmula de cuadratura:

$$\int_a^b f(x) dx \approx \omega_0 f(x_0) + \omega_1 f(x_1) + \cdots + \omega_N f(x_N)$$

¿Es posible calcular los pesos ω_i y los nodos x_i de forma que el grado de precisión de la fórmula sea lo mayor posible? Sí, pero entonces los nodos no estarán equiespaciados y estas fórmulas se llaman de Gauss-Legendre. Los nodos para el intervalo $[-1, 1]$ serían los ceros de los llamados Polinomios de Legendre.

Por lo demás, una vez tenemos los nodos la forma de deducir la correspondiente fórmula de cuadratura es la misma que usando nodos equiespaciados: si P_N es el polinomio que interpola a f en los nodos de Legendre $x_0, x_1, \dots, x_N \in [-1, 1]$ y

$$\int_{-1}^1 f(x) dx \approx \int_{-1}^1 P_N(x) dx = \omega_0 f(x_0) + \omega_1 f(x_1) + \cdots + \omega_n f(x_N)$$

Esta fórmula se puede adaptar fácilmente para otro intervalo $[a, b]$.

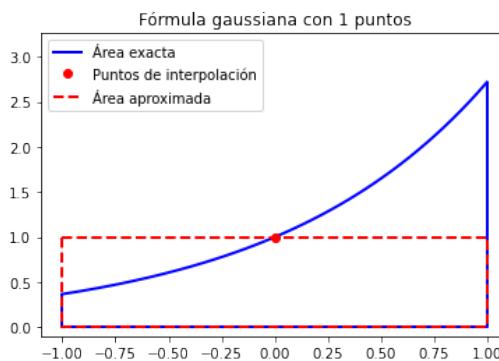
Para $[-1, 1]$ los pesos y nodos son

| n | w_i | x_i |
|-----|----------|----------------|
| 1 | 2.000000 | 0.000000 |
| 2 | 1.000000 | ± 0.577350 |
| 3 | 0.555556 | ± 0.774597 |
| | 0.888889 | 0.000000 |
| 4 | 0.347855 | ± 0.861136 |
| | 0.652145 | ± 0.339981 |
| 5 | 0.236927 | ± 0.906180 |
| | 0.478629 | ± 0.538469 |
| | 0.568889 | 0.000000 |

Así, por ejemplo

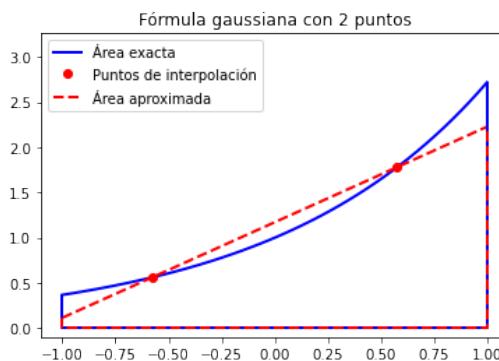
Fórmula gaussiana de un nodo

$$\int_{-1}^1 f(x) dx \approx 2 f(0)$$



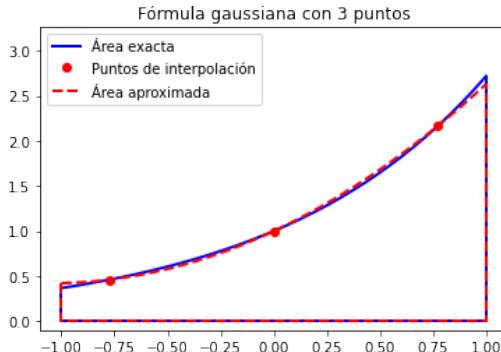
Fórmula gaussiana de dos nodos

$$\int_{-1}^1 f(x) dx \approx f(-0.5774) + f(0.5774)$$



Fórmula gaussiana de tres nodos

$$\int_{-1}^1 f(x) dx \approx 0.5556 f(-0.7746) + 0.8889 f(0) + 0.5556 f(0.7746)$$



Y así sucesivamente.

Estos resultados se pueden generalizar a cualquier intervalo $[a, b]$ cambiando los x_i por y_i de acuerdo con la fórmula

$$y_i = \frac{b-a}{2}x_i + \frac{a+b}{2}$$

Y entonces la fórmula de cuadratura es

$$\int_a^b f(x) dx \approx \frac{b-a}{2} (\omega_0 f(y_0) + \omega_1 f(y_1) + \dots + \omega_n f(y_n))$$

EJERCICIO

1. Calcular la integral

Queremos calcular

$$\int_0^3 e^{-\frac{x^2}{2}} dx$$

usando la fórmula de cuadratura gaussiana con tres nodos:

$$\int_{-1}^1 f(x) dx \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$$

Haremos un cambio de variable que nos lleve del intervalo $[0, 3]$ 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 $x = 0$, entonces $t = -1$ y

$$0 = -m + n.$$

Y si $x = 3$, entonces $t = 1$ y

$$3 = m + n.$$

Sumando estas dos últimas ecuaciones

$$3 = 2n \quad \boxed{n = \frac{3}{2}}$$

Si cambiamos de signo a la primera

$$0 = m - n \quad 3 = m + n$$

Y sumándolas

$$3 = 2m \quad \boxed{m = \frac{3}{2}}$$

Y la solución a este sistema es

$$m = \frac{3}{2} \quad n = \frac{3}{2}$$

Y por lo tanto, el cambio de variable es

$$x = \frac{3}{2}t + \frac{3}{2} \quad dx = \frac{3}{2}dt.$$

La fórmula de cuadratura era

$$\int_{-1}^1 f(x) dx \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$$

Y como hemos hecho el cambio de forma que el intervalo $[0, 3]$ se transforme en el intervalo $[-1, 1]$

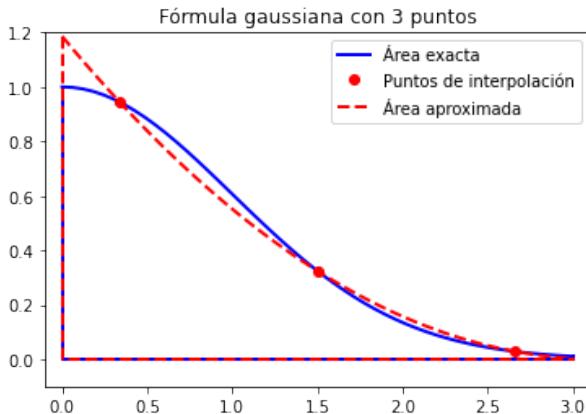
$$I = \int_0^3 f(x) dx = \int_{-1}^1 f\left(\frac{3}{2}t + \frac{3}{2}\right) \frac{3}{2} dt = \frac{3}{2} \int_{-1}^1 f\left(\frac{3}{2}t + \frac{3}{2}\right) dt \approx$$

$$\approx \frac{3}{2} \left[\frac{5}{9} f\left(-\frac{3}{2}\sqrt{\frac{3}{5}} + \frac{3}{2}\right) + \frac{8}{9} f\left(\frac{3}{2}\right) + \frac{5}{9} f\left(\frac{3}{2}\sqrt{\frac{3}{5}} + \frac{3}{2}\right) \right] =$$

$$= \frac{3}{2} \left[\frac{5}{9} f(0.3381) + \frac{8}{9} f(1.5) + \frac{5}{9} f(2.6619) \right] =$$

teniendo en cuenta que $f(x) = e^{-\frac{x^2}{2}}$

$$= \frac{3}{2} \left[\frac{5}{9} (0.9444) + \frac{8}{9} (0.3246) + \frac{5}{9} (0.02893) \right] = 1.24402$$



De hecho

- $I_e = 1.24993044474155$
- $I_a = 1.24401686205204$
- Error = 0.00591358268951

Grado de precisión de la fórmula

Las fórmulas gaussianas eligen los nodos de forma que se maximice la precisión de la fórmula.

En general, una cuadratura de Gauss-Legendre de n puntos será exacta para funciones polinomiales de grado menor o igual que $2n - 1$. Por lo tanto, con un nodo la precisión es 1, con dos nodos la precisión es 3 y con tres nodos la precisión es 5.

Si comparamos las fórmulas gaussianas con las fórmulas de Newton-Cotes con los mismos nodos

| Nodos | Grado de precisión Newton – Cotes | Grado de precisión Gauss – Legendre |
|-------|--------------------------------------|--|
| 1 | Punto Medio | 1 |
| 2 | Trapecios | 3 |
| 3 | Simpson | 5 |
| 4 | Simpson 3/8 | 7 |
| 5 | Boole | 9 |

TEMA 5

SISTEMAS DE ECUACIONES LINEALES

5.1 Método de Gauss

Ejercicio 5.1.1

Sea el sistema $Ax = \mathbf{b}$ donde

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} -2 \\ -1 \\ -3 \end{pmatrix}$$

Calcular x utilizando el método de Gauss.

INTRODUCCIÓN

En este tema vamos a resolver sistemas de ecuaciones lineales con el mismo número de ecuaciones que de incógnitas con matriz de coeficientes no singular. Por lo tanto, el sistema tiene solución única.

Dados los números a_{ij} y b_j para $i, j = 1, 2, \dots, n$ se trata de hallar los números x_1, x_2, \dots, x_n que verifican las n ecuaciones lineales

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n &= b_n. \end{aligned}$$

El sistema, expresado matricialmente, se escribiría

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n1} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

es decir

$$Ax = \mathbf{b}$$

Métodos de resolución de sistemas lineales

Los métodos pueden ser:

Métodos directos

- La solución se calcula en un número finito de pasos conocidos a priori.
- Sólo están sujetos a errores de redondeo.
- Son adecuados para resolver:
 - Sistemas pequeños.
 - Sistemas grandes con matriz llena. Veremos los métodos de Gauss, Gauss con pivote, Gauss-Jordan y Descomposición LU.

Métodos iterativos

- Construyen una sucesión que converge a la solución del sistema.
- Además de los errores de redondeo, existe un error de truncamiento. Veremos los métodos de Jacobi y Gauss-Seidel.

Método de Gauss

El método de Gauss consta de dos pasos:

1. **Triangularización.** Transformación del sistema de ecuaciones lineales original en un sistema de matriz triangular superior con las mismas soluciones. Para ello se realizan las operaciones:
 - Multiplicar una fila por un real y sumársela a otra:

$$f_i \rightarrow f_i + \lambda f_j, \quad j \neq i$$

- Intercambiar filas: $f_i \leftrightarrow f_j$ (si usamos la estrategia del pivote)

2. **Sustitución regresiva.** Resolución del sistema triangular superior mediante el algoritmo de sustitución regresiva. Consiste en despejar las incógnitas empezando por la última, siguiendo con la penúltima y siguiendo así hasta despejar la primera incógnita.

EJERCICIO

El sistema es $Ax = \mathbf{b}$ con

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} -2 \\ -1 \\ -3 \end{pmatrix}$$

Es decir, queremos resolver el sistema

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

1. Triangularización

Construimos la matriz extendida del sistema, que es la matriz A a la que se le ha añadido la matriz columna \mathbf{b} como cuarta columna.

El pivote es siempre un elemento de la diagonal principal. El primer pivote está en la primera fila y por tanto, en la primera columna.

Hacemos ceros por debajo del pivote, 1 , en la primera columna. Para ello:

- La fila del *pivote* es la primera fila. Y el *pivote* es ahora a_{11} .
- La fila del *pivote* se deja como está.
- A las demás filas se les suma la fila del *pivote* con un factor que se construye con el *pivote* en el denominador y el elemento de debajo del *pivote* en el numerador (3 para la fila 2 y 1 para la fila 3).

$$\left(\begin{array}{ccc|c} f_1 & 1 & 3 & -2 \\ f_2 & 3 & 0 & 1 \\ f_3 & 1 & -2 & 1 \end{array} \right) \quad \left(\begin{array}{l} f'_1 \\ f'_2 \\ f'_3 \end{array} \right) = \left(\begin{array}{l} f_1 \\ f_2 - \frac{3}{1}f_1 \\ f_3 - \frac{1}{1}f_1 \end{array} \right)$$

Hacemos ceros por debajo del segundo pivote -3 en la segunda columna. Para ello:

- La fila del *pivote* es la segunda fila. Y el *pivote* es ahora a_{22} .
- La fila del *pivote* y las filas por encima del pivote se dejan como están.
- A la fila restante se le suma la fila del *pivote* con un factor que se construye con el *pivote* en el denominador y el elemento de debajo del *pivote* en el numerador, -3 .

$$\left(\begin{array}{ccc|c} f'_1 & 1 & 3 & -2 \\ f'_2 & 0 & -3 & 5 \\ f'_3 & 0 & -3 & -1 \end{array} \right) \quad \left(\begin{array}{l} f''_1 \\ f''_2 \\ f''_3 \end{array} \right) = \left(\begin{array}{l} f'_1 \\ f'_2 \\ f'_3 - \frac{-3}{-3}f'_2 \end{array} \right)$$

Y ya tenemos una matriz triangular superior (con ceros por debajo de la diagonal principal).

$$\left(\begin{array}{ccc|c} f_1'' & 1 & 1 & -2 \\ f_2'' & 0 & -3 & 5 \\ f_3'' & 0 & 0 & -6 \end{array} \right)$$

2. Sustitución regresiva

Vamos a revertir a la notación del sistema con ecuaciones. El sistema triangular se escribe

$$\begin{array}{rclcl} x & + & y & + & 3z = -2 \\ & - & 3y & - & 8z = 5 \\ & & & & 6z = -6 \end{array}$$

Despejando z en la tercera ecuación

$$z = -1$$

Despejando y en la segunda ecuación

$$y = \frac{5 + 8z}{-3} = \frac{5 - 8}{-3} = 1$$

Y x en la primera

$$x = -2 - y - 3z = -2 - 1 + 3 = 0$$

tenemos la solución del sistema

| | | |
|---------|---------|----------|
| $x = 0$ | $y = 1$ | $z = -1$ |
|---------|---------|----------|

Ejercicio 5.1.2

Sea el sistema $Ax = \mathbf{b}$ donde

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} -2 \\ -1 \\ -3 \end{pmatrix}$$

Calcular x utilizando Gauss con pivote parcial.

INTRODUCCIÓN

La estrategia del pivote

Consiste en

- Intercambiar filas: pivote parcial.
- Intercambiar filas y columnas: pivote total.

Problemas de no usar la estrategia del pivote

- Puede que en algún paso nuestro pivote sea cero. Y como el pivote es el divisor en los factores que construimos en cada paso, no podríamos continuar.
- Como consecuencia de lo anterior, Gauss sin estrategia del pivote no funciona siempre con sistemas determinados.
- Si dividimos números muy diferentes entre sí, con un denominador comparativamente pequeño, podemos tener problemas con el error.

Los beneficios de usar la estrategia del pivote son:

- Evitar ceros en la posición del pivote.
- Evitar valores comparativamente pequeños en el denominador, que son perjudiciales desde el punto de vista numérico.

EJERCICIO

1. Triangularización

Construimos la matriz extendida del sistema.

De todos los elementos por debajo del pivote 1, nos quedamos con el que es mayor, en este caso 3. Intercambiamos esta fila con la del pivote.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|c} 1 & 1 & 3 & -2 \\ 3 & 0 & 1 & -1 \\ 1 & -2 & 1 & -3 \end{array} \right) \begin{array}{c} f'_1 \\ f'_2 \\ f'_3 \end{array} = \begin{array}{c} f_3 \\ f_1 \\ f_2 \end{array}$$

Hacemos ceros por debajo del pivote **3** realizando operaciones por filas. Sumamos a cada fila, la fila del pivote multiplicada por un factor que tiene como denominador el pivote, y como numerador el elemento de la fila que está debajo del pivote por **-1**.

$$\begin{array}{l} f'_1 \quad \left(\begin{array}{ccc|c} \textcolor{red}{3} & 0 & 1 & -1 \\ \textcolor{blue}{1} & 1 & 3 & -2 \\ \textcolor{green}{1} & -2 & 1 & -3 \end{array} \right) \quad f''_1 = f'_1 \\ f'_2 \quad f''_2 = f'_2 - \frac{\textcolor{blue}{1}}{\textcolor{red}{3}} f'_1 \\ f'_3 \quad f''_3 = f'_3 - \frac{\textcolor{green}{1}}{\textcolor{red}{3}} f'_1 \end{array}$$

Escogemos el pivote entre los elementos debajo del pivote **1** en la segunda columna. Nos quedamos con el mayor en valor absoluto **-2** e intercambiamos la del nuevo pivote por la del pivote.

$$\begin{array}{l} f''_1 \quad \left(\begin{array}{ccc|c} 3 & 0 & 1 & -1 \\ 0 & \textcolor{red}{1} & \frac{8}{3} & -\frac{5}{3} \\ 0 & -2 & \frac{2}{3} & -\frac{8}{3} \end{array} \right) \quad f'''_2 = f''_3 \\ f''_2 \quad f'''_2 = f''_2 \\ f''_3 \quad f'''_3 = f''_2 \end{array}$$

Hacemos ceros por debajo del nuevo pivote

$$\begin{array}{l} f'''_1 \quad \left(\begin{array}{ccc|c} 3 & 0 & 1 & -1 \\ 0 & \textcolor{red}{-2} & \frac{2}{3} & -\frac{8}{3} \\ 0 & \textcolor{blue}{1} & \frac{8}{3} & -\frac{5}{3} \end{array} \right) \quad f''''_1 = f'''_1 \\ f'''_2 \quad f''''_2 = f'''_2 \\ f'''_3 \quad f''''_3 = f'''_3 - \frac{\textcolor{blue}{+1}}{\textcolor{red}{-2}} f''''_2 \end{array}$$

Y ya tenemos una matriz triangular superior (con ceros por debajo de la diagonal principal).

$$\begin{array}{l} f''''_1 \quad \left(\begin{array}{ccc|c} 3 & 0 & 1 & -1 \\ 0 & -2 & \frac{2}{3} & -\frac{8}{3} \\ 0 & 0 & 3 & -3 \end{array} \right) \end{array}$$

2. Sustitución reversiva

Volviendo a la notación del sistema con ecuaciones. El sistema triangular se escribe

$$\begin{array}{rclcl} 3x & + & z & = & -1 \\ - 2y & + & (2/3)z & = & -8/3 \\ & & 3z & = & -3 \end{array}$$

De la última ecuación despejamos z . Luego con el valor de z obtenido obtenemos y de la segunda ecuación. Finalmente, de la primera ecuación despejamos x y sustituimos el valor de z .

$$z = -1 \quad y = \frac{-8/3 - 2/3(-1)}{-2} = \frac{-8/3 - 2/3(-1)}{-2} = 1 \quad x = \frac{-1 - z}{3} = \frac{-1 - (-1)}{3} = 0$$

Y la solución del sistema es $x = 0 \quad y = 1 \quad z = -1$

5.2 Método de Gauss-Jordan

Ejercicio 5.2.1

Sea la matriz

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

Calcular la matriz inversa de A utilizando Gauss-Jordan.

INTRODUCCIÓN

El método de Gauss-Jordan

Como el método de Gauss, el método de Gauss-Jordan realiza las operaciones por fila

- Multiplicar una fila por un real y sumársela a otra:

$$f_i \rightarrow f_i + \lambda f_j, \quad j \neq i$$

- Intercambiar filas: $f_i \leftrightarrow f_j$ (si usamos la estrategia del pivote).

Y además, realiza la operación por fila

- Multiplicar o dividir una fila por un real: $f_i \rightarrow \lambda f_i \quad \lambda \neq 0$

El método es parecido al de Gauss pero en cada paso no se hacen ceros por debajo del pivote sino *por encima y por debajo* del pivote.

Ahora el objetivo es pasar de un sistema cualquiera a un sistema diagonal (elementos distintos de cero solo en la diagonal) equivalente, que nos dé directamente la solución. Por ejemplo, el sistema diagonal equivalente

$$\begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \\ -3 \end{pmatrix} \iff \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

Y la solución es inmediata y es

$$x = 0 \quad y = 1 \quad z = -1$$

Para resolver un solo sistema es más rentable, desde el punto de vista del número de operaciones, el método de Gauss. Pero cuando se resuelven simultáneamente varios sistemas con la misma matriz de coeficientes es

más rentable Gauss-Jordan. Y este es el caso del cálculo de la matriz inversa.

Cálculo de la matriz inversa

La matriz inversa de una matriz A $n \times n$, caso de existir, es una matriz $n \times n$ que llamaremos A^{-1} que verifica

$$AA^{-1} = I = A^{-1}A.$$

Es decir

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Por la definición del producto de matrices, podemos escribir

$$A \begin{pmatrix} c_{11} \\ c_{21} \\ \vdots \\ c_{n1} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad A \begin{pmatrix} c_{12} \\ c_{22} \\ \vdots \\ c_{n2} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots, \quad A \begin{pmatrix} c_{1n} \\ c_{2n} \\ \vdots \\ c_{nn} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

Es decir, las columnas de la matriz inversa son las soluciones de estos n sistemas lineales que comparten matriz de coeficientes. Por lo tanto, si planteamos simultáneamente todos estos sistema y escribimos la matriz aumentada

$$\left(\begin{array}{cccc|ccc} a_{11} & a_{12} & \dots & a_{1n} & 1 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2n} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & 0 & 0 & \dots & 1 \end{array} \right)$$

Y hacemos transformaciones por filas de forma que la matriz equivalente sea la matriz diagonal identidad, las soluciones serán las columnas de la matriz inversa

$$\left(\begin{array}{cccc|ccc} 1 & 0 & \dots & 0 & c_{11} & c_{12} & \dots & c_{1n} \\ 0 & 1 & \dots & 0 & c_{21} & c_{22} & \dots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & c_{n1} & c_{n2} & \dots & c_{nn} \end{array} \right)$$

Es decir, podemos resumir el proceso como

- Escribir una matriz $n \times 2n$ que consiste en la matriz dada A a la izquierda y la matriz identidad I de dimensión $n \times n$ a la derecha $[A|I]$.
- Mediante operaciones por filas, transformar la matriz A en la matriz I , de forma que obtenemos a partir de $[A|I]$ obtenemos $[I|A^{-1}]$.

EJERCICIO

Escribimos la matriz $[A|I]$. Como el pivote es **1** lo dejamos como está. Hacemos ceros por debajo del pivote en la primera columna.

$$\begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} \textcolor{red}{1} & 1 & 3 & 1 & 0 & 0 \\ \textcolor{blue}{3} & 0 & 1 & 0 & 1 & 0 \\ \textcolor{green}{1} & -2 & 1 & 0 & 0 & 1 \end{array} \right) \begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} = \begin{array}{l} f_1 \\ f_2 - \textcolor{blue}{3}f_1 \\ f_3 - \textcolor{green}{1}f_1 \end{array}$$

Ahora el pivote es **-3** y para convertirlo a 1 dividimos la fila por su valor.

$$\begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & \textcolor{blue}{1} & 3 & 1 & 0 & 0 \\ 0 & \textcolor{red}{-3} & -8 & -3 & 1 & 0 \\ 0 & \textcolor{green}{-3} & -2 & -1 & 0 & 1 \end{array} \right) \begin{array}{l} f_2 \\ f_3 \end{array} = \begin{array}{l} f_2 / (-\textcolor{red}{3}) \\ f_3 \end{array}$$

Hacemos ceros por encima y por debajo del pivote **1**.

$$\begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & \textcolor{blue}{1} & 3 & 1 & 0 & 0 \\ 0 & \textcolor{red}{1} & 8/3 & 1 & -1/3 & 0 \\ 0 & \textcolor{green}{-3} & -2 & -1 & 0 & 1 \end{array} \right) \begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} = \begin{array}{l} f_1 - \textcolor{blue}{1}f_2 \\ f_3 - (\textcolor{green}{-3})f_2 \end{array}$$

Ahora el pivote es **6** y para convertirlo a 1 dividimos la fila por su valor.

$$\begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & \textcolor{blue}{1/3} & 0 & 1/3 & 0 \\ 0 & 1 & \textcolor{green}{8/3} & 1 & -1/3 & 0 \\ 0 & 0 & \textcolor{red}{6} & 2 & -1 & 1 \end{array} \right) \begin{array}{l} f_3 \\ f_1 \\ f_2 \end{array} = \begin{array}{l} f_3 / (\textcolor{red}{6}) \\ f_1 - (\textcolor{blue}{1/3})f_3 \\ f_2 - (\textcolor{green}{8/3})f_3 \end{array}$$

Hacemos ceros por encima del pivote **1**.

$$\begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & \textcolor{blue}{1/3} & 0 & 1/3 & 0 \\ 0 & 1 & \textcolor{green}{8/3} & 1 & -1/3 & 0 \\ 0 & 0 & \textcolor{red}{1} & 1/3 & -1/6 & 1/6 \end{array} \right) \begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} = \begin{array}{l} f_1 - (\textcolor{blue}{1/3})f_3 \\ f_2 - (\textcolor{green}{8/3})f_3 \\ f_3 \end{array}$$

Y ya tenemos la matriz identidad I a la izquierda

$$\begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -1/9 & 7/18 & -1/18 \\ 0 & 1 & 0 & 1/9 & 1/9 & -4/9 \\ 0 & 0 & 1 & 1/3 & -1/6 & 1/6 \end{array} \right)$$

Como a la izquierda ya tenemos la matriz I la matriz de la derecha será A^{-1} .

$$A^{-1} = \begin{pmatrix} -1/9 & 7/18 & -1/18 \\ 1/9 & 1/9 & -4/9 \\ 1/3 & -1/6 & 1/6 \end{pmatrix}$$

5.3 Factorización LU

Ejercicio 5.3.1

Sea el sistema $Ax = \mathbf{b}$ donde

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} -2 \\ -1 \\ -3 \end{pmatrix}$$

Calcular \mathbf{x} utilizando la factorización LU .

INTRODUCCIÓN

Resolución de sistemas por factorización LU

Queremos descomponer la matriz de coeficientes A en una matriz triangular superior U y una matriz triangular inferior L de forma que

$$A = LU$$

No todas las matrices admiten factorización LU . Si A es invertible, admite descomposición LU si y solo si todos sus menores principales son distintos de cero.

La factorización LU no es única. Por ejemplo:

- Método de Crout: los elementos de la diagonal de U son unos.
- Método de Doolittle: los elementos de la diagonal de L son unos.
- Método de Cholesky (matrices simétricas definidas positivas): los elementos de la diagonal son iguales en L y U . Es decir, $u_{ii} = l_{ii}$

Usaremos el método de Doolittle. En este caso:

- U es la matriz triangular superior que se obtiene al triangularizar A como en Gauss.
- Los elementos de la matriz L (por debajo de la diagonal principal) son los factores que usamos en las operaciones por filas en Gauss.

Consideramos el sistema de ecuaciones lineales $Ax = \mathbf{b}$ donde la matriz A admite la factorización LU . Para resolver el sistema, hay que

1. Descomponer $A = LU$. Como $Ax = \mathbf{b}$ se transforma en $LUx = \mathbf{b}$, si a Ux le llamamos \mathbf{y} podemos escribirlo como $Ly = \mathbf{b}$.
2. Resolver $Ly = \mathbf{b}$ por sustitución progresiva.
3. Resolver $Ux = \mathbf{y}$ por sustitución regresiva.

EJERCICIO

El sistema es $Ax = \mathbf{b}$ donde

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} -2 \\ -1 \\ -3 \end{pmatrix}$$

Los pasos para resolver el problema son

1. **Factorización:** $A = LU$
2. **Sustitución progresiva:** $Ax = b \implies LUx = b \implies Ly = b$.
Resolvemos $Ly = b$.
3. **Sustitución regresiva:** Resolvemos $Ux = y$.

1. Factorización A=LU

En el primer paso hacemos ceros por debajo del elemento a_{11} sumando la primera fila multiplicada por un real.

$$\begin{array}{ccc|c|c} f_1 & \left(\begin{array}{ccc} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{array} \right) & f_1 & \rightarrow & f_1 \\ f_2 & & f_2 & \rightarrow & f_2 \\ f_3 & & f_3 & \rightarrow & f_3 \end{array} \quad \begin{array}{l} - \\ - \\ - \end{array} \quad \begin{array}{c|c} 3/1 & f_1 \\ 1/1 & f_1 \end{array}$$

Los multiplicadores, que aparecen en recuadros, son los elementos con los que construimos la matriz L . Los insertamos en la matriz, en lugar de los ceros creados. La matriz transformada es

$$\left(\begin{array}{ccc|c|c} 1 & 1 & 3 \\ 3 & -3 & -8 \\ 1 & -3 & -2 \end{array} \right) \quad \begin{array}{ccc|c|c} f_1 & \rightarrow & f_1 \\ f_2 & \rightarrow & f_2 \\ f_3 & \rightarrow & f_3 \end{array} \quad \begin{array}{l} - \\ - \\ - \end{array} \quad \boxed{(-3)/(-3)} \quad f_2$$

Repetimos el proceso creando ceros por debajo de a'_{22} y llegamos a la matriz que almacena simultáneamente L y U .

$$\left(\begin{array}{ccc|c} 1 & 1 & 3 \\ 3 & -3 & -8 \\ 1 & 1 & 6 \end{array} \right)$$

Y las matrices L y U son:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad U = \begin{pmatrix} 1 & 1 & 3 \\ 0 & -3 & -8 \\ 0 & 0 & 6 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} -2 \\ -1 \\ -3 \end{pmatrix}$$

2. Sustitución progresiva Ly = b

$$\begin{array}{rcl} y_1 & = & -2 \\ 3y_1 + y_2 & = & -1 \\ y_1 + y_2 + y_3 & = & -3 \end{array}$$

De la primera ecuación

$$y_1 = -2$$

De la segunda

$$y_2 = -1 - 3y_1 = -1 - 3(-2) = 5$$

Y de la tercera

$$y_3 = -3 - y_1 - y_2 = -3 - (-2) - 5 = -6$$

Es decir

$$y_1 = -2 \quad y_2 = 5 \quad y_3 = -6$$

3. Sustitución regresiva Ux = y

$$U = \begin{pmatrix} 1 & 1 & 3 \\ 0 & -3 & -8 \\ 0 & 0 & 6 \end{pmatrix} \quad y = \begin{pmatrix} -2 \\ 5 \\ -6 \end{pmatrix}$$

$$\begin{array}{rcl} x_1 + x_2 + 3x_3 & = & -2 \\ - 3x_2 - 8x_3 & = & 5 \\ 6x_3 & = & -6 \end{array}$$

De la tercera ecuación

$$x_3 = -1$$

De la segunda

$$x_2 = \frac{5 + 8x_3}{-3} = \frac{5 + 8(-1)}{-3} = 1$$

Y de la primera

$$x_1 = -2 - x_2 - 3x_3 = -2 - 1 - 3(-1) = 0$$

Es decir

| | | |
|-----------|-----------|------------|
| $x_1 = 0$ | $x_2 = 1$ | $x_3 = -1$ |
|-----------|-----------|------------|

Diferencias entre Gauss y factorización LU

El método de factorización *LU* es, esencialmente, el método de Gauss. La matriz *U* es la misma matriz triangular que obtenemos por Gauss. ¿En qué son distintos entonces?

- En Gauss, incluimos a **b** en las transformaciones por fila, en factorización *LU* no.
- En *LU* es la matriz *L* la que almacena las transformaciones que estamos haciendo a **b** (calculamos **y** en $Ly = b$) para luego aplicárselas y obtener **y**, que sería el vector **b** transformado con las operaciones por filas de Gauss.

Entonces, si hacemos lo mismo que con Gauss ¿por qué no trabajar directamente con la matriz aumentada? ¿Por qué hacerlo en dos pasos?

El método de factorización *LU* tiene sentido (comparado con Gauss) si tenemos una serie de sistemas que hemos de resolver secuencialmente y donde la solución de uno determina el término independiente del siguiente **b**. Por ejemplo

1. $Ax_1 = x_0, \quad Ax_2 = x_1, \quad Ax_3 = x_2, \dots$
2. $Ly_1 = x_0, \quad Ly_2 = x_1, \quad Ly_3 = x_2, \dots$
3. $Ux_1 = y_1, \quad Ux_2 = y_2, \quad Ux_3 = y_3, \dots$

En este caso, con Gauss tendríamos que triangularizar una vez para cada sistema, cambiando el término independiente, mientras que con el método *LU* haríamos sólo una vez el paso 1 y repetiríamos para cada sistema los pasos 2 y 3.

Factorización LU como producto de matrices elementales

La secuencia de operaciones sobre las filas que se usa para transformar la matriz *A* en un matriz triangular superior equivalente es

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

$$f_2 \leftarrow f_2 - (3)f_1$$

$$\begin{pmatrix} 1 & 1 & 3 \\ 0 & -3 & -8 \\ 1 & -2 & 1 \end{pmatrix}$$

$$f_3 \leftarrow f_3 - (1)f_1$$

$$\begin{pmatrix} 1 & 1 & 3 \\ 0 & -3 & -8 \\ 0 & -3 & -2 \end{pmatrix}$$

$$f_3 \leftarrow f_3 - (1)f_2$$

$$\begin{pmatrix} 1 & 1 & 3 \\ 0 & -3 & -8 \\ 0 & 0 & 6 \end{pmatrix}$$

Estas transformaciones se consiguen con una serie de multiplicación de matrices por la izquierda que se corresponden con las matrices elementales

$$f_2 \leftarrow f_2 - (3)f_1$$

$$\begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 3 \\ 0 & -3 & -8 \\ 1 & -2 & 1 \end{pmatrix}$$

$$f_3 \leftarrow f_3 - (1)f_1$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 3 \\ 0 & -3 & -8 \\ 0 & -3 & -2 \end{pmatrix}$$

$$f_3 \leftarrow f_3 - (1)f_2$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} = \\ = \begin{pmatrix} 1 & 1 & 3 \\ 0 & -3 & -8 \\ 0 & 0 & 6 \end{pmatrix}$$

Cada matriz elemental G_1 , G_2 y G_3 se corresponde con una operación por fila

$$G_3 G_2 G_1 A = U$$

Y, por lo tanto

$$A = (G_3 G_2 G_1)^{-1} U = G_1^{-1} G_2^{-1} G_3^{-1} U$$

Si llamamos

$$L = G_1^{-1} G_2^{-1} G_3^{-1}$$

$$A = LU$$

donde L es una matriz triangular inferior. Calculemos esta matriz L

$$G_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$G_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

$$G_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

Y entonces

$$G_1^{-1} G_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

$$G_1^{-1} G_2^{-1} G_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

Y ya tenemos A factorizada como LU

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 3 \\ 0 & -3 & -8 \\ 0 & 0 & 6 \end{pmatrix} = LU$$

Ejercicio 5.3.2

Sea el sistema $Ax = \mathbf{b}$ donde

$$A = \begin{pmatrix} 0 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} -2 \\ -1 \\ -3 \end{pmatrix}$$

Calcular \mathbf{x} utilizando la factorización LU con pivote parcial.

En el método de LU sucede como en Gauss: no es totalmente funcional si no aplicamos la estrategia del pivote.

Para ir guardando las permutaciones de filas se usa la llamada *matriz de permutaciones* P que inicialmente es la matriz identidad y que se modifica cada vez que hay un intercambio de filas.

Los pasos son

1. **Factorización:** $PA = LU$
2. **Sustitución progresiva:** $PAx = Pb \implies LUX = Pb \implies Ly = Pb$.
Resolvemos $Ly = Pb$.
3. **Sustitución regresiva:** Resolvemos $UX = y$.

1. Factorización $PA = LU$

Las matrices A y P serán inicialmente

$$A = \begin{pmatrix} 0 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \quad P_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

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_{21} por lo que intercambiamos las filas 2 y 1 tanto en P como en A .

$$A_1 = \begin{pmatrix} 3 & 0 & 1 \\ 0 & 1 & 3 \\ 1 & -2 & 1 \end{pmatrix} \quad P_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Ahora, hacemos ceros por debajo del elemento a_{11} restando la primera fila multiplicada por el real construido con el pivote (a_{11}) en el denominador y el elemento de esa fila por debajo del pivote (a_{21} y a_{31} respectivamente) en el numerador.

$$\begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc} 3 & 0 & 1 \\ 0 & 1 & 3 \\ 1 & -2 & 1 \end{array} \right) \quad \begin{array}{l} f_1 \rightarrow f_1 \\ f_2 \rightarrow f_2 - \boxed{0/3} f_1 \\ f_3 \rightarrow f_3 - \boxed{1/3} f_1 \end{array}$$

Los multiplicadores, que aparecen en recuadros, son los elementos con los que construimos la matriz L . Los insertamos en la matriz, en lugar de los ceros creados. La matriz transformada es

$$\left(\begin{array}{ccc} 3 & 0 & 1 \\ \boxed{0} & \boxed{1} & 3 \\ \boxed{1/3} & -2 & 2/3 \end{array} \right)$$

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 .

$$\left(\begin{array}{ccc} 3 & 0 & 1 \\ \boxed{1/3} & -2 & 2/3 \\ \boxed{0} & 1 & 3 \end{array} \right) \quad P_2 = \left(\begin{array}{ccc} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{array} \right)$$

Ahora, hacemos ceros por debajo del elemento a_{22} restando la segunda fila multiplicada por el real construido con el pivote (a_{22}) en el denominador y el elemento de esa fila por debajo del pivote (a_{32}) en el numerador.

$$\left(\begin{array}{ccc} 3 & 0 & 1 \\ \boxed{1/3} & -2 & 2/3 \\ \boxed{0} & 1 & 3 \end{array} \right) \quad \begin{array}{l} f_1 \rightarrow f_1 \\ f_2 \rightarrow f_2 \\ f_3 \rightarrow f_3 - \boxed{1/(-2)} f_2 \end{array}$$

llegamos a la matriz que almacena simultáneamente L y U .

$$\left(\begin{array}{ccc} 3 & 0 & 1 \\ \boxed{1/3} & -2 & 2/3 \\ \boxed{0} & \boxed{-1/2} & \boxed{10/3} \end{array} \right)$$

Y las matrices L , U y P son:

$$L = \left(\begin{array}{ccc} 1 & 0 & 0 \\ \boxed{1/3} & 1 & 0 \\ \boxed{0} & \boxed{-1/2} & 1 \end{array} \right) \quad U = \left(\begin{array}{ccc} 3 & 0 & 1 \\ 0 & -2 & 2/3 \\ 0 & 0 & 10/3 \end{array} \right) \quad P_2 = \left(\begin{array}{ccc} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{array} \right)$$

Y por lo tanto \mathbf{b} y $P\mathbf{b}$ son

$$\mathbf{b} = \begin{pmatrix} -2 \\ -1 \\ -3 \end{pmatrix} \quad P\mathbf{b} = \begin{pmatrix} -1 \\ -3 \\ -2 \end{pmatrix}$$

2. Sustitución progresiva $\mathbf{Ly} = \mathbf{Pb}$

$$\begin{array}{rclclcl} y_1 & & & = & -1 \\ \frac{1}{3}y_1 + y_2 & = & -3 \\ -\frac{1}{2}y_2 + y_3 & = & -2 \end{array}$$

De la primera ecuación

$$y_1 = -1$$

De la segunda

$$y_2 = -3 - (1/3)y_1 = -3 - (1/3)(-1) = -8/3$$

Y de la tercera

$$y_3 = -2 + (1/2)y_2 = -2 + (1/2)(-8/3) = -10/3$$

Es decir

$$y_1 = -1 \quad y_2 = -8/3 \quad y_3 = -10/3$$

3. Sustitución regresiva $\mathbf{Ux} = \mathbf{y}$

$$\begin{array}{rclclcl} 3x_1 & + & x_3 & = & -1 \\ -2x_2 & + & (2/3)x_3 & = & -8/3 \\ (10/3)x_3 & = & -10/3 \end{array}$$

De la tercera ecuación

$$x_3 = -1$$

De la segunda

$$x_2 = \frac{-(8/3) - (2/3)x_3}{-2} = \frac{-(8/3) - (2/3)(-1)}{-2} = 1$$

Y de la primera

$$x_1 = \frac{-1 - x_3}{3} = \frac{-1 - (-1)}{3} = 0$$

Es decir

| | | |
|-----------|-----------|------------|
| $x_1 = 0$ | $x_2 = 1$ | $x_3 = -1$ |
|-----------|-----------|------------|

5.4 Determinantes

Ejercicio 5.4.1

Calcular los determinantes de las matrices

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix} \quad A_1 = \begin{pmatrix} 0 & 1 & 3 \\ 3 & 0 & 1 \\ 1 & -2 & 1 \end{pmatrix}$$

Teniendo en cuenta las siguientes propiedades de los determinantes:

1. Al **intercambiar dos filas** en una matriz, el determinante queda multiplicado por -1 .
2. Al **dividir una fila por una constante distinta de cero** el determinante de la matriz queda dividido por ese valor.
3. Si **multiplicamos una fila por un real no nulo y la sumamos a otra** el valor del determinante no cambia.
4. El **valor del determinante de una matriz triangular o diagonal** es el producto de los elementos de la diagonal.

1. Método de Gauss

En este caso, a partir de A , hemos llegado a la matriz

$$\begin{pmatrix} 1 & 1 & 3 \\ 0 & -3 & -8 \\ 0 & 0 & 6 \end{pmatrix}$$

utilizando filas sumadas a otras (propiedad 3) por lo que sólo nos resta multiplicar los elementos de la diagonal (propiedad 4) y

$$|A| = (1) \times (-3) \times (6) = -18.$$

2. Método de Gauss con pivote parcial

En este caso, a partir de A , hemos llegado a la matriz

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

Ahora, además de operaciones por filas, hemos intercambiado filas dos veces. Cada vez que intercambiamos filas el determinante de la matriz se

multiplica por -1 (propiedad 1). Por lo tanto tendremos que multiplicar el producto de los elementos de la diagonal por $(-1) \times (-1)$ y

$$|A| = (-1) \times (-1) \times (3) \times (-2) \times (3) = -18.$$

3. Método de Gauss-Jordan

La matriz final es la matriz identidad y hemos realizado operaciones por filas y hemos dividido las filas por el pivote. Por lo tanto hemos de multiplicar los factores por los que hemos dividido cada fila (propiedad 2) y

$$|A| = (1) \times (-3) \times (6) = -18.$$

4. Factorización LU (con o sin pivote)

Para la matriz A , la matriz triangular superior transformada es la misma que la obtenida por Gauss y el determinante se obtendría a partir de esta matriz como en el método de Gauss.

En el caso de la matriz A_1 llegamos a la matriz

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

Y como hemos intercambiado filas 2 veces, el determinante es

$$|A_1| = (-1) \times (-1) \times (3) \times (-2) \times (10/3) = -20.$$

Triangulación por Gauss y factorización LU son, si solo realizamos este primer paso, lo mismo.

En el caso de Gauss-Jordan, Gauss es un método más económico si sólo queremos calcular el determinante.

El método más adecuado para el cálculo del determinante sería Gauss con pivote parcial, porque existen matrices con determinantes distintos de cero donde el proceso por Gauss se puede atascar porque aparece un cero en la posición del pivote. Esto no sería inconveniente en Gauss con pivote parcial, porque intercambiaríamos la fila por otra donde el pivote no fuera cero. Y si esto no es posible, quiere decir que el determinante de la matriz es cero y problema solucionado.

5.5 Método de Jacobi

Ejercicio 5.5.1

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 3 \\ 0 & 3 & 4 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 3 \\ 3 \\ 5 \end{pmatrix}$$

1. ¿Es A diagonal dominante por filas?
2. Calcular la norma infinito, la norma uno y los autovalores de B_J .
3. A partir de cada uno de los apartados anteriores ¿qué podemos concluir acerca de la convergencia del método de Jacobi?
4. Realizar 3 iteraciones por Jacobi comenzando con

$$\mathbf{x}^{(0)} = (0, 0, 0)^T$$

INTRODUCCIÓN

Métodos iterativos de resolución de sistemas lineales

Dada una aproximación inicial $\mathbf{x}^{(0)}$, un método iterativo genera una sucesión de aproximaciones

$$\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(k)}, \dots \rightarrow \mathbf{x}$$

que converge a la solución.

Para generar esta sucesión, se repite el mismo esquema de operaciones hasta que se cumple un criterio de parada. Por ejemplo, hasta que hemos realizado un cierto número de iteraciones.

Los métodos iterativos clásicos (lineales) se basan en reescribir el problema

$$A\mathbf{x} = \mathbf{b} \iff \mathbf{x} = B\mathbf{x} + \mathbf{c}$$

donde B es una matriz $n \times n$ y \mathbf{c} es una matriz columna de n elementos.

Algoritmo

- Sea $\mathbf{x}^{(0)}$ una aproximación inicial a la solución
- Para $k = 1, 2, \dots$

$$\mathbf{x}^{(k)} = B\mathbf{x}^{(k-1)} + \mathbf{c}$$

La matriz B se llama *matriz de iteración* y el vector \mathbf{c} se llama *vector de iteración*.

Ventajas

- Los métodos iterativos son, en general, más eficientes que los métodos directos para resolver sistemas de ecuaciones lineales grandes y de matriz hueca.
- Si no se exige mucha precisión, se puede obtener una aproximación aceptable en un número pequeño de iteraciones.
- Son menos sensibles a los errores de redondeo que los métodos directos.

Inconvenientes

- En general, no es posible predecir el número de operaciones que se requieren para obtener una aproximación a la solución con una precisión determinada.
- El tiempo de cálculo y la precisión del resultado pueden depender de la elección de ciertos parámetros (método de sobrerrrelajación).
- Generalmente no se gana tiempo por iteración si la matriz de coeficientes es simétrica.

Método de Jacobi

Se basa en la descomposición

$$A = L + D + U$$

donde si

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

entonces

$$L = \begin{pmatrix} 0 & 0 & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & 0 \end{pmatrix} \quad D = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix}$$

$$U = \begin{pmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & 0 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

Hay que tener en cuenta que aunque los nombres de las matrices son los mismos, esta descomposición no tiene nada que ver con la factorización LU .

Tenemos que

$$A\mathbf{x} = \mathbf{b} \Rightarrow (L + D + U)\mathbf{x} = \mathbf{b} \Rightarrow$$

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

Y si

$$B_J = -D^{-1}(L + D) \quad \mathbf{c}_J = D^{-1}\mathbf{b}$$

podemos escribir

$$\mathbf{x} = B_J\mathbf{x} + \mathbf{c}_J$$

Y haciendo

$$\mathbf{x}^{(k+1)} = B_J\mathbf{x}^{(k)} + \mathbf{c}_J$$

Y si comenzamos con un $\mathbf{x}^{(0)}$

$$\mathbf{x}^{(1)} = B_J\mathbf{x}^{(0)} + \mathbf{c}_J, \quad \mathbf{x}^{(2)} = B_J\mathbf{x}^{(1)} + \mathbf{c}_J, \quad \mathbf{x}^{(3)} = B_J\mathbf{x}^{(2)} + \mathbf{c}_J, \quad \dots$$

EJERCICIO

Dado el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 3 \\ 0 & 3 & 4 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 3 \\ 3 \\ 5 \end{pmatrix}$$

1. ¿Es A diagonal dominante por filas?

Se dice que una matriz A de n filas y n columnas es diagonal dominante por filas si

$$\sum_{\substack{j=1 \\ i \neq j}}^n |a_{ij}| < |a_{ii}| \quad i = 1, \dots, n$$

Para nuestra matriz A

$$A = \begin{pmatrix} \boxed{4} & 1 & 0 \\ 1 & \boxed{4} & 3 \\ 0 & 3 & \boxed{4} \end{pmatrix} \quad \begin{array}{l} |1| + |0| < |4| \\ |1| + |3| \not< |4| \\ |0| + |3| < |4| \end{array}$$

Y no es diagonal dominante por filas puesto que en la segunda fila no se verifica la desigualdad ya que $|4| = |1| + |3|$

2. Calcular la norma infinito, la norma uno y los autovalores de la matriz de iteración de Jacobi

Se tiene que $B_J = -D^{-1}(L + U)$. O también:

- Dividimos cada fila por el correspondiente elemento de la diagonal.

$$\begin{pmatrix} 1 & 1/4 & 0 \\ 1/4 & 1 & 3/4 \\ 0 & 3/4 & 1 \end{pmatrix}$$

- Cambiamos todos los elementos de signo.

$$\begin{pmatrix} -1 & -1/4 & 0 \\ -1/4 & -1 & -3/4 \\ 0 & -3/4 & -1 \end{pmatrix}$$

- Ponemos ceros en la diagonal principal.

$$B_J = \begin{pmatrix} 0 & -1/4 & 0 \\ -1/4 & 0 & -3/4 \\ 0 & -3/4 & 0 \end{pmatrix}$$

Norma infinito

Si A es una matriz $m \times n$ su norma infinito viene dada por:

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

Y la norma infinito de B_J viene dada por

$$B_J = \begin{pmatrix} 0 & -1/4 & 0 \\ -1/4 & 0 & -3/4 \\ 0 & -3/4 & 0 \end{pmatrix} \quad \begin{matrix} 0 + 1/4 + 0 = 1/4 \\ 1/4 + 0 + 3/4 = 1 \\ 0 + 3/4 + 0 = 3/4 \end{matrix}$$

Por lo tanto

$$\|B_J\|_\infty = \text{Max}(1/4, 1, 3/4) = 1$$

Norma uno

Calculamos la norma uno. Si A es una matriz $m \times n$ su norma uno viene dada por:

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

$$B_J^T = \begin{pmatrix} 0 & -1/4 & 0 \\ -1/4 & 0 & -3/4 \\ 0 & -3/4 & 0 \end{pmatrix} \quad \begin{array}{l} 0 + 1/4 + 0 = 1/4 \\ 1/4 + 0 + 3/4 = 1 \\ 0 + 3/4 + 0 = 3/4 \end{array}$$

Por lo tanto

$$\|B_J\|_1 = \text{Max}(1/4, 1, 3/4) = 1$$

Autovalores

Calculamos los autovalores de B_J calculando las raíces de $|B_J - \lambda I| = 0$

$$\begin{vmatrix} 0 - \lambda & -1/4 & 0 \\ -1/4 & 0 - \lambda & -3/4 \\ 0 & -3/4 & 0 - \lambda \end{vmatrix} = -\lambda^3 - \left(-\frac{1}{16}\lambda - \frac{9}{16}\lambda \right) = \\ = -\lambda^3 + \frac{10}{16}\lambda = \frac{5}{8}\lambda - \lambda^3 = \lambda \left(\frac{5}{8} - \lambda^2 \right) = 0$$

Y los autovalores son

$$\lambda_1 = 0, \quad \lambda_{2,3} = \pm \sqrt{\frac{5}{8}} = \pm 0.79$$

3. A partir de cada uno de los apartados anteriores ¿qué podemos concluir acerca de la convergencia del método de Jacobi?

Es condición suficiente para la convergencia del método de Jacobi que la matriz de coeficientes A sea diagonal dominante por filas:

- Como A no es diagonal dominante, no podemos concluir nada.

Es condición suficiente para la convergencia del método de Jacobi que *alguna* norma de la matriz de iteración B_J sea menor que 1.

- Como $\|B_J\|_\infty \not< 1$ no podemos concluir nada.
- Como $\|B_J\|_1 \not< 1$ no podemos concluir nada.

Es condición necesaria y suficiente para la convergencia del método de Jacobi que todos los autovalores de la matriz de iteración B_J en valor absoluto sean menores que 1.

- 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.

4. Realizar 3 iteraciones con el método de Jacobi

El sistema es $A\mathbf{x} = \mathbf{b}$ con

$$A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 3 \\ 0 & 3 & 4 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 3 \\ 3 \\ 5 \end{pmatrix}$$

Si escribimos las ecuaciones

$$\begin{aligned} 4x + y &= 3 \\ x + 4y + 3z &= 3 \\ 3y + 4z &= 5 \end{aligned}$$

En la primera ecuación despejamos la primera incógnita, x , en la segunda, la segunda incógnita, y , y finalmente, z .

$$\begin{aligned} x &= (3 - y)/4 \\ y &= (3 - x - 3z)/4 \\ z &= (5 - 3y)/4 \end{aligned}$$

Realizamos las iteraciones asumiendo que todos los valores de las incógnitas a la derecha son los valores de la iteración anterior

$$\begin{aligned} x^{(1)} &= (3 - y^{(0)})/4 \\ y^{(1)} &= (3 - x^{(0)} - 3z^{(0)})/4 \\ z^{(1)} &= (5 - 3y^{(0)})/4 \end{aligned}$$

Realizamos una iteración, tomando como valor inicial, el vector nulo

$$\begin{aligned} x^{(1)} &= (3 - y^{(0)})/4 = (3 - 0)/4 = 3/4 = 0.75 \\ y^{(1)} &= (3 - x^{(0)} - 3z^{(0)})/4 = (3 - 0 - 0)/4 = 3/4 = 0.75 \\ z^{(1)} &= (5 - 3y^{(0)})/4 = (5 - 0)/4 = 5/4 = 1.25 \end{aligned}$$

Segunda iteración:

$$\begin{aligned} x^{(2)} &= (3 - y^{(1)})/4 = (3 - 0.75)/4 = 0.56 \\ y^{(2)} &= (3 - x^{(1)} - 3z^{(1)})/4 = (3 - 0.75 - 3(1.25))/4 = -0.38 \\ z^{(2)} &= (5 - 3y^{(1)})/4 = (5 - 3(0.75))/4 = 0.69 \end{aligned}$$

Tercera iteración:

$$\begin{aligned} x^{(3)} &= (3 - y^{(2)})/4 = (3 - (-0.38))/4 = 0.84 \\ y^{(3)} &= (3 - x^{(2)} - 3z^{(2)})/4 = (3 - 0.56 - 3(0.69))/4 = 0.09 \\ z^{(3)} &= (5 - 3y^{(2)})/4 = (5 - 3(-0.38))/4 = 1.53 \end{aligned}$$

Para comparar, la solución exacta es

$$x = 1 \quad y = -1 \quad z = 2$$

Algoritmo

El algoritmo que hemos aplicado se expresa formalmente

- Elegir una aproximación inicial $\mathbf{x}^{(0)}$
- Para $k = 1, 2, \dots, MaxIter$
 - Para $i = 1, 2, \dots, n$, calcular

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k-1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k-1)} \right)$$

- Si se cumple el criterio de parada, tomar $\mathbf{x}^{(k)}$ como aproximación a la solución.

Ejercicio 5.5.2

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 20 & 1 & 0 & 1 \\ 1 & 20 & 3 & 1 \\ 0 & 3 & 20 & 1 \\ 1 & 0 & 1 & 20 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 10 \\ 7 \\ 4 \\ 6 \end{pmatrix}$$

Resolver utilizando el método de Jacobi.

Utilizar como condición de parada $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty < 0.01$

Vamos a resolverlo usando python para hacer las cuentas

```
[1]: import numpy as np
np.set_printoptions(precision=3)
```

Escribimos los datos, que son la matriz de coeficientes y la de términos independientes

```
[2]: A = np.array([[20., 1, 0, 1],[1, 20, 3, 1],[0, 3, 20, 1],[1, 0, 1, 20]])
b = np.array([10., 7, 4, 6])
```

E inicializamos los dos vectores donde almacenaremos las dos últimas soluciones aproximadas

```
[3]: x0 = np.zeros(4)
x1 = np.zeros(4)
```

Si tenemos un sistema 4×4

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= b_3 \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 &= b_4 \end{aligned}$$

y despejamos la primera incógnita en la primera ecuación, la segunda incógnita en la segunda y así sucesivamente

$$\begin{aligned} x_1 &= \frac{b_1 - a_{12}x_2 - a_{13}x_3 - a_{14}x_4}{a_{11}} \\ x_2 &= \frac{b_2 - a_{21}x_1 - a_{23}x_3 - a_{24}x_4}{a_{22}} \\ x_3 &= \frac{b_3 - a_{31}x_1 - a_{32}x_2 - a_{34}x_4}{a_{33}} \\ x_4 &= \frac{b_4 - a_{41}x_1 - a_{42}x_2 - a_{43}x_3}{a_{44}} \end{aligned}$$

Y obtenemos los valores de la primera iteración a partir de los valores iniciales

$$\begin{aligned} x_1^{(1)} &= \frac{b_1 - a_{12}x_2^{(0)} - a_{13}x_3^{(0)} - a_{14}x_4^{(0)}}{a_{11}} \\ x_2^{(1)} &= \frac{b_2 - a_{21}x_1^{(0)} - a_{23}x_3^{(0)} - a_{24}x_4^{(0)}}{a_{22}} \\ x_3^{(1)} &= \frac{b_3 - a_{31}x_1^{(0)} - a_{32}x_2^{(0)} - a_{34}x_4^{(0)}}{a_{33}} \\ x_4^{(1)} &= \frac{b_4 - a_{41}x_1^{(0)} - a_{42}x_2^{(0)} - a_{43}x_3^{(0)}}{a_{44}} \end{aligned}$$

Algoritmo

El algoritmo que hemos aplicado se expresa formalmente

- Elegir una aproximación inicial $\mathbf{x}^{(0)}$
- Para $k = 1, 2, \dots, MaxIter$
 - Para $i = 1, 2, \dots, n$, calcular

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k-1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k-1)} \right)$$

- Si se cumple el criterio de parada, tomar $\mathbf{x}^{(k)}$ como aproximación a la solución.

Iteración 1

A partir de $x^{(0)}$ calculamos $x^{(1)}$

```
[4]: k = 1
x1[0] = (b[0] - A[0,1] * x0[1] - A[0,2] * x0[2] - A[0,3] * x0[3]) / A[0,0]
x1[1] = (b[1] - A[1,0] * x0[0] - A[1,2] * x0[2] - A[1,3] * x0[3]) / A[1,1]
x1[2] = (b[2] - A[2,0] * x0[0] - A[2,1] * x0[1] - A[2,3] * x0[3]) / A[2,2]
x1[3] = (b[3] - A[3,0] * x0[0] - A[3,1] * x0[1] - A[3,2] * x0[2]) / A[3,3]
```

Imprimimos la nueva iteración, el vector diferencia y la norma de dicho vector diferencia, que nos va a valer para controlar el error

```
[5]: print ('x1           = ', x1)
print ('x1-x0        = ', x1-x0)
print ('max(abs(x1-x0)) = ', max(abs(x1-x0)))
```

```
x1           = [ 0.5   0.35  0.2   0.3 ]
x1-x0        = [ 0.5   0.35  0.2   0.3 ]
max(abs(x1-x0)) =  0.5
```

Como la norma es mayor que la 0.01, seguimos iterando

Iteración 2

A partir de $x^{(1)}$ calculamos $x^{(2)}$ (aunque los guardamos respectivamente en $x0$ y $x1$)

```
[6]: k = 2
x0 = np.copy(x1)
x1[0] = (b[0] - A[0,1] * x0[1] - A[0,2] * x0[2] - A[0,3] * x0[3]) / A[0,0]
x1[1] = (b[1] - A[1,0] * x0[0] - A[1,2] * x0[2] - A[1,3] * x0[3]) / A[1,1]
x1[2] = (b[2] - A[2,0] * x0[0] - A[2,1] * x0[1] - A[2,3] * x0[3]) / A[2,2]
x1[3] = (b[3] - A[3,0] * x0[0] - A[3,1] * x0[1] - A[3,2] * x0[2]) / A[3,3]
```

Imprimimos la nueva iteración, el vector diferencia y la norma de dicho vector diferencia, que nos va a valer para controlar el error

```
[7]: print ('x1           = ', x1)
print ('x1-x0        = ', x1-x0)
print ('max(abs(x1-x0)) = ', max(abs(x1-x0)))
```

```
x1           = [ 0.467  0.28   0.133  0.265]
x1-x0        = [ -0.033 -0.07   -0.068 -0.035]
max(abs(x1-x0)) =  0.068
```

Como la norma es mayor que la 0.01, seguimos iterando

Iteración 3

A partir de $x^{(2)}$ calculamos $x^{(3)}$

```
[8]: k = 3
x0 = np.copy(x1)
x1[0] = (b[0] - A[0,1] * x0[1] - A[0,2] * x0[2] - A[0,3] * x0[3]) / A[0,0]
x1[1] = (b[1] - A[1,0] * x0[0] - A[1,2] * x0[2] - A[1,3] * x0[3]) / A[1,1]
x1[2] = (b[2] - A[2,0] * x0[0] - A[2,1] * x0[1] - A[2,3] * x0[3]) / A[2,2]
x1[3] = (b[3] - A[3,0] * x0[0] - A[3,1] * x0[1] - A[3,2] * x0[2]) / A[3,3]
```

Imprimimos el nuevo resultado y la norma del vector diferencia

```
[9]: print ('x1          = ', x1)
print ('x1-x0      = ', x1-x0)
print ('max(abs(x1-x0)) = ', max(abs(x1-x0)))
```

```
x1          = [0.473 0.293 0.145 0.27 ]
x1-x0      = [0.005 0.013 0.012 0.005]
max(abs(x1-x0)) = 0.013
```

Como la norma es mayor que la 0.01, seguimos iterando

Iteración 4

A partir de $x^{(3)}$ calculamos $x^{(4)}$

```
[10]: k = 4
x0 = np.copy(x1)
x1[0] = (b[0] - A[0,1] * x0[1] - A[0,2] * x0[2] - A[0,3] * x0[3]) / A[0,0]
x1[1] = (b[1] - A[1,0] * x0[0] - A[1,2] * x0[2] - A[1,3] * x0[3]) / A[1,1]
x1[2] = (b[2] - A[2,0] * x0[0] - A[2,1] * x0[1] - A[2,3] * x0[3]) / A[2,2]
x1[3] = (b[3] - A[3,0] * x0[0] - A[3,1] * x0[1] - A[3,2] * x0[2]) / A[3,3]
```

Imprimimos el resultado

```
[11]: print ('x1          = ', x1)
print ('x1-x0      = ', x1-x0)
print ('max(abs(x1-x0)) = ', max(abs(x1-x0)))
```

```
x1          = [0.472 0.291 0.142 0.269]
x1-x0      = [-0.001 -0.002 -0.002 -0.001]
max(abs(x1-x0)) = 0.002
```

Ahora ya se verifica que la norma del vector diferencia es menor que 0.01 y paramos. Nuestra solución aproximada es

```
[12]: print ('x1 = ', x1)
```

```
x1 = [0.472 0.291 0.142 0.269]
```

La solución exacta es

```
[13]: x = np.linalg.solve(A,b);
print(x)
```

```
[0.472 0.292 0.143 0.269]
```

5.6 Método de Gauss-Seidel

Ejercicio 5.6.1

Sea el sistema $Ax = \mathbf{b}$ donde

$$A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 3 \\ 0 & 3 & 4 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 3 \\ 3 \\ 5 \end{pmatrix}$$

1. ¿Es A diagonal dominante por filas?
2. Calcular la norma infinito, la norma uno y los autovalores de B_{G-S} .
3. A partir de cada uno de los apartados anteriores ¿qué podemos concluir acerca de la convergencia del método de Gauss-Seidel?
4. Realizar 3 iteraciones por Gauss-Seidel comenzando con $\mathbf{x}^{(0)} = (0, 0, 0)^T$.

INTRODUCCIÓN

Método de Gauss-Seidel

Se basa en la descomposición

$$A = L + D + U$$

donde si

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

entonces

$$L = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{pmatrix} \quad D = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}$$

$$U = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

Tenemos que

$$A\mathbf{x} = \mathbf{b} \Rightarrow (L + D + U)\mathbf{x} = \mathbf{b} \Rightarrow$$

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

con

$$B_{G-S} = -(L + D)^{-1}U \quad \mathbf{c}_{G-S} = (L + D)^{-1}\mathbf{b}$$

y podemos escribir

$$\mathbf{x} = B_{G-S}\mathbf{x} + \mathbf{c}_{G-S}$$

Y el método vendrá dado por

$$\mathbf{x}^{(k+1)} = B_{G-S}\mathbf{x}^{(k)} + \mathbf{c}_{G-S}$$

Y si comenzamos con un $\mathbf{x}^{(0)}$

$$\mathbf{x}^{(1)} = B_{G-S}\mathbf{x}^{(0)} + \mathbf{c}_{G-S},$$

$$\mathbf{x}^{(2)} = B_{G-S}\mathbf{x}^{(1)} + \mathbf{c}_{G-S},$$

$$\mathbf{x}^{(3)} = B_{G-S}\mathbf{x}^{(2)} + \mathbf{c}_{G-S},$$

...

EJERCICIO

1. ¿Es A diagonal dominante por filas?

Se dice que una matriz A de n filas y n columnas es diagonal dominante por filas si

$$\sum_{\substack{j=1 \\ i \neq j}}^n |a_{ij}| < |a_{ii}| \quad i = 1, \dots, n$$

Para nuestra matriz A

$$A = \begin{pmatrix} \boxed{4} & 1 & 0 \\ 1 & \boxed{4} & 3 \\ 0 & 3 & \boxed{4} \end{pmatrix} \quad \begin{array}{l} |1| + |0| < |4| \\ |1| + |3| \not< |4| \\ |0| + |3| < |4| \end{array}$$

Y no es diagonal dominante por filas puesto que en la segunda fila no se verifica la desigualdad ya que $|4| = |1| + |3|$

2. Calcular la norma infinito, la norma uno y los autovalores de la matriz de iteración de Gauss-Seidel

Se tiene que $B_{G-S} = -(L + D)^{-1}U$. Si

$$A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 3 \\ 0 & 3 & 4 \end{pmatrix}$$

se tiene que

$$L = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix} \quad D = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix} \quad U = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix}$$

$$L + D = \begin{pmatrix} 4 & 0 & 0 \\ 1 & 4 & 0 \\ 0 & 3 & 4 \end{pmatrix}$$

Calculemos su inversa por Gauss-Jordan:

Escribimos la matriz $[(L + D)|I]$. Como el pivote es 4 dividimos toda la fila por 4 para convertirlo en 1.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 4 & 0 & 0 & 1 & 0 & 0 \\ 1 & 4 & 0 & 0 & 1 & 0 \\ 0 & 3 & 4 & 0 & 0 & 1 \end{array} \right) \begin{array}{l} f_1 = f_1/(4) \\ f_2 \\ f_3 \end{array}$$

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/4 & 0 & 0 \\ 1 & 4 & 0 & 0 & 1 & 0 \\ 0 & 3 & 4 & 0 & 0 & 1 \end{array} \right) \begin{array}{l} f_1 \\ f_2 = f_2 - 1f_1 \\ f_3 = f_3 - 0f_1 \end{array}$$

Hacemos cero por debajo del pivote.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 4 & 0 & -1/4 & 1 & 0 \\ 0 & 3 & 4 & 0 & 0 & 1 \end{array} \right) \begin{array}{l} f_2 = f_2/(4) \\ f_3 \\ f_1 \end{array}$$

Como el pivote es 4 dividimos toda la fila por 4 para convertirlo en 1.

$$\begin{array}{c} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/4 & 0 & 0 \\ 0 & 1 & 0 & -1/16 & 1/4 & 0 \\ 0 & 3 & 4 & 0 & 0 & 1 \end{array} \right) \begin{array}{l} f_1 = f_1 - 0f_2 \\ f_2 \\ f_3 = f_3 - 3f_2 \end{array}$$

Hacemos ceros por encima y por debajo del pivote (aunque por encima ya hay un cero y no hay que hacer nada)

$$\left(\begin{array}{ccc|ccc} f_1 & 1 & 0 & \textcolor{blue}{0} & 1/4 & 0 & 0 \\ f_2 & 0 & 1 & \textcolor{green}{0} & -1/16 & 1/4 & 0 \\ f_3 & 0 & 0 & \textcolor{red}{4} & 3/16 & -3/4 & 1 \end{array} \right) \quad f_3 = f_3 / (\textcolor{red}{4})$$

Dividimos por el pivote la fila del pivote, y como por encima de él sólo hay ceros, ya hemos acabado.

$$\left(\begin{array}{ccc|ccc} f_1 & 1 & 0 & 0 & 1/4 & 0 & 0 \\ f_2 & 0 & 1 & 0 & -1/16 & 1/4 & 0 \\ f_3 & 0 & 0 & 1 & 3/64 & -3/16 & 1/4 \end{array} \right)$$

Como a la izquierda ya tenemos la matriz I la matriz de la derecha será $(L + D)^{-1}$.

$$(L + D)^{-1} = \begin{pmatrix} 1/4 & 0 & 0 \\ -1/16 & 1/4 & 0 \\ 3/64 & -3/16 & 1/4 \end{pmatrix}$$

La multiplicamos por U

$$B_{G-S} = -(L + D)^{-1} U = - \begin{pmatrix} 1/4 & 0 & 0 \\ -1/16 & 1/4 & 0 \\ 3/64 & -3/16 & 1/4 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix}$$

y

$$B_{G-S} = \begin{pmatrix} 0 & -1/4 & 0 \\ 0 & 1/16 & -3/4 \\ 0 & -3/64 & 9/16 \end{pmatrix}$$

Norma infinito

Si A es una matriz $m \times n$ su norma infinito viene dada por:

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

Pasamos a coma flotante para comparar mejor cantidades y estudiamos la norma infinito de B_{G-S}

$$B_{G-S} = \begin{pmatrix} 0 & -0.25 & 0 \\ 0 & 0.06 & -0.75 \\ 0 & -0.05 & 0.56 \end{pmatrix} \quad \begin{aligned} 0 + 0.25 + 0 &= 0.25 \\ 0 + 0.06 + 0.75 &= 0.81 \\ 0 + 0.05 + 0.56 &= 0.61 \end{aligned}$$

Por lo tanto

$$\|B_{G-S}\|_\infty = \text{Max}(0.25, 0.81, 0.61) = 0.81$$

Norma uno

Si A es una matriz $m \times n$ su norma uno viene dada por:

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

$$B_{G-S}^T = \begin{pmatrix} 0 & 0 & 0 \\ -0.25 & 0.06 & -0.05 \\ 0 & -0.75 & 0.56 \end{pmatrix} \quad \begin{array}{l} 0+0+0=0 \\ 0.25+0.06+0.05=0.36 \\ 0+0.75+0.56=1.31 \end{array}$$

Por lo tanto

$$\|B_{G-S}\|_1 = \text{Max}(0, 0.36, 1.31) = 1.31$$

Autovalores

Calculamos los autovalores de B_{G-S}

$$\begin{pmatrix} 0 & -1/4 & 0 \\ 0 & 1/16 & -3/4 \\ 0 & -3/64 & 9/16 \end{pmatrix}$$

Calculamos las raíces de $|B_{G-S} - \lambda I| = 0$

$$\left| \begin{array}{ccc} 0 - \lambda & -1/4 & 0 \\ 0 & (1/16) - \lambda & -3/4 \\ 0 & -3/64 & (9/16) - \lambda \end{array} \right| =$$

$$= -\lambda \left| \begin{array}{ccc} (1/16) - \lambda & -3/4 \\ -3/64 & (9/16) - \lambda \end{array} \right| =$$

$$= -\lambda \left[\left(\frac{1}{16} - \lambda \right) \left(\frac{9}{16} - \lambda \right) - \left(-\frac{3}{4} \right) \left(-\frac{3}{64} \right) \right] =$$

$$= -\lambda \left[\left(\frac{1}{16} - \lambda \right) \left(\frac{9}{16} - \lambda \right) - \frac{9}{4 \times 64} \right] =$$

$$= -\lambda \left[\frac{1}{16} \times \frac{9}{16} - \frac{1}{16} \lambda - \frac{9}{16} \lambda + \lambda^2 - \frac{9}{2^2 \times 2^6} \right] =$$

$$= -\lambda \left[\frac{9}{2^8} - \frac{10}{16} \lambda + \lambda^2 - \frac{9}{2^2 \times 2^6} \right] = -\lambda^2 \left(-\frac{10}{16} + \lambda \right) = -\lambda^2 \left(\lambda - \frac{5}{8} \right)$$

Y los autovalores son

$$\lambda_{1,2} = 0, \quad \lambda_3 = \frac{5}{8} = 0.625$$

3. A partir de cada uno de los apartados anteriores ¿qué podemos concluir acerca de la convergencia del método de Gauss-Seidel?

Es condición suficiente para la convergencia del método de Gauss-Seidel que la matriz de coeficientes A sea diagonal dominante por filas:

- Como A no es diagonal dominante, no podemos concluir nada.

Es condición suficiente para la convergencia del método de Gauss-Seidel que *alguna* norma de la matriz de iteración B_{G-S} sea menor que 1.

- Como $\|B_{G-S}\|_\infty < 1$ podemos concluir que el método de Gauss-Seidel converge para este sistema. (Y no haría falta seguir investigando y obtener la otra norma ni los autovalores, aunque lo hemos hecho para comprobar que los resultados son coherentes).
- De $\|B_{G-S}\|_1 > 1$ no podríamos concluir nada.

Es condición necesaria y suficiente para la convergencia del método de Gauss-Seidel que todos los autovalores de la matriz de iteración B_{G-S} en valor absoluto sean menores que 1.

- Como todos los autovalores de B_{G-S} son menores que uno en valor absoluto podemos concluir que el Método de Jacobi será convergente para cualquier valor inicial.

4. Realizar 3 iteraciones con Gauss-Seidel

El sistema es

$$\begin{array}{rcl} 4x + y & = & 3 \\ x + 4y + 3z & = & 3 \\ 3y + 4z & = & 5 \end{array}$$

En la primera ecuación despejamos la primera incógnita, x , en la segunda, la segunda incógnita, y , y finalmente, z .

$$\begin{aligned} x &= \frac{3-y}{4} \\ y &= \frac{3-x-3z}{4} \\ z &= \frac{5-3y}{4} \end{aligned}$$

Realizamos las iteraciones tomando los valores de la iteración anterior, pero los valores nuevos que vayamos calculando, los usamos ya

$$\begin{aligned} x^{(1)} &= \frac{3-y^{(0)}}{4} \\ y^{(1)} &= \frac{3-x^{(1)}-3z^{(0)}}{4} \\ z^{(1)} &= \frac{5-3y^{(1)}}{4} \end{aligned}$$

Realizamos una iteración, tomando como valor inicial, el vector nulo

$$\begin{aligned}x^{(1)} &= (3 - y^{(0)})/4 = (3 - 0)/4 = 3/4 = 0.75 \\y^{(1)} &= (3 - x^{(1)} - 3z^{(0)})/4 = (3 - 0.75 - 0)/4 = 0.56 \\z^{(1)} &= (5 - 3y^{(1)})/4 = (5 - 3(0.56))/4 = 0.83\end{aligned}$$

Realizamos la iteración 2

$$\begin{aligned}x^{(2)} &= (3 - y^{(1)})/4 = (3 - 0.56)/4 = 0.61 \\y^{(2)} &= (3 - x^{(2)} - 3z^{(1)})/4 = (3 - 0.61 - 3(0.83))/4 = -0.02 \\z^{(2)} &= (5 - 3y^{(2)})/4 = (5 - 3(-0.02))/4 = 1.27\end{aligned}$$

Realizamos la iteración 3

$$\begin{aligned}x^{(3)} &= (3 - y^{(2)})/4 = (3 - (-0.02))/4 = 0.76 \\y^{(3)} &= (3 - x^{(3)} - 3z^{(2)})/4 = (3 - 0.76 - 3(1.27))/4 = -0.39 \\z^{(3)} &= (5 - 3y^{(3)})/4 = (5 - 3(-0.39))/4 = 1.54\end{aligned}$$

La solución exacta es

$$x = 1 \quad y = -1 \quad z = 2$$

Algoritmo

El algoritmo que hemos aplicado se expresa formalmente

- Elegir una aproximación inicial $\mathbf{x}^{(0)}$
- Para $k = 1, 2, \dots, \text{MaxIter}$
 - Para $i = 1, 2, \dots, n$, calcular

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^n a_{ij} x_j^{(k-1)} \right)$$

- Si se cumple el criterio de parada, tomar $\mathbf{x}^{(k)}$ como aproximación a la solución.

Ejercicio 5.6.2

Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 20 & 1 & 0 & 1 \\ 1 & 20 & 3 & 1 \\ 0 & 3 & 20 & 1 \\ 1 & 0 & 1 & 20 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 10 \\ 7 \\ 4 \\ 6 \end{pmatrix}$$

Resolver utilizando el método de Gauss-Seidel.

Utilizar como condición de parada

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_{\infty} < 0.01$$

Vamos a resolverlo usando python para hacer las cuentas

```
[1]: import numpy as np
np.set_printoptions(precision=3)
```

Escribimos los datos, que son la matriz de coeficientes y la de términos independientes

```
[2]: A = np.array([[20., 1, 0, 1],[1, 20, 3, 1],[0, 3, 20, 1],[1, 0, 1, 20]])
b = np.array([10., 7, 4, 6])
```

E inicializamos los dos vectores donde almacenaremos las dos últimas soluciones aproximadas

```
[3]: x0 = np.zeros(4)
x1 = np.zeros(4)
```

Si tenemos un sistema 4×4

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= b_3 \\ a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 &= b_4 \end{aligned}$$

y despejamos la primera incógnita en la primera ecuación, la segunda incógnita en la segunda y así sucesivamente

$$\begin{aligned}x_1 &= \frac{b_1 - a_{12}x_2 - a_{13}x_3 - a_{14}x_4}{a_{11}} \\x_2 &= \frac{b_2 - a_{21}x_1 - a_{23}x_3 - a_{24}x_4}{a_{22}} \\x_3 &= \frac{b_3 - a_{31}x_1 - a_{32}x_2 - a_{34}x_4}{a_{33}} \\x_4 &= \frac{b_4 - a_{41}x_1 - a_{42}x_2 - a_{43}x_3}{a_{44}}\end{aligned}$$

Y obtenemos los valores de la primera iteración a partir de los valores iniciales, pero conforme vamos teniendo resultados nuevos de una variable, los actualizamos

$$\begin{aligned}x_1^{(1)} &= \frac{b_1 - a_{12}x_2^{(0)} - a_{13}x_3^{(0)} - a_{14}x_4^{(0)}}{a_{11}} \\x_2^{(1)} &= \frac{b_2 - a_{21}x_1^{(1)} - a_{23}x_3^{(0)} - a_{24}x_4^{(0)}}{a_{22}} \\x_3^{(1)} &= \frac{b_3 - a_{31}x_1^{(1)} - a_{32}x_2^{(1)} - a_{34}x_4^{(0)}}{a_{33}} \\x_4^{(1)} &= \frac{b_4 - a_{41}x_1^{(1)} - a_{42}x_2^{(1)} - a_{43}x_3^{(1)}}{a_{44}}\end{aligned}$$

Algoritmo

El algoritmo que hemos aplicado se expresa formalmente

- Elegir una aproximación inicial $\mathbf{x}^{(0)}$
- Para $k = 1, 2, \dots, \text{MaxIter}$
 - Para $i = 1, 2, \dots, n$, calcular

$$x_i^{(k)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^n a_{ij}x_j^{(k-1)} \right)$$

- Si se cumple el criterio de parada, tomar $\mathbf{x}^{(k)}$ como aproximación a la solución.

Iteración 1

A partir de $x^{(0)}$ calculamos $x^{(1)}$, pero para ir actualizando los valores de una coordenada en cuanto lo hemos actualizado, una estrategia sería usar un sólo vector durante la iteración. No obstante, necesitamos guardar la iteración anterior para comprobar el criterio de parada

```
[4]: k = 1
xant = np.copy(x0)
x0[0] = (b[0] - A[0,1] * x0[1] - A[0,2] * x0[2] - A[0,3] * x0[3]) / A[0,0]
x0[1] = (b[1] - A[1,0] * x0[0] - A[1,2] * x0[2] - A[1,3] * x0[3]) / A[1,1]
x0[2] = (b[2] - A[2,0] * x0[0] - A[2,1] * x0[1] - A[2,3] * x0[3]) / A[2,2]
x0[3] = (b[3] - A[3,0] * x0[0] - A[3,1] * x0[1] - A[3,2] * x0[2]) / A[3,3]
```

Imprimimos la nueva iteración, el vector diferencia y la norma de dicho vector diferencia, que nos va a valer para controlar el error

```
[5]: print ('x0          = ', x0)
print ('x0-xant      = ', x0-xant)
print ('max(abs(x0-xant)) = ', max(abs(x0-xant)))
```

```
x0          = [0.5    0.325  0.151  0.267]
x0-xant      = [0.5    0.325  0.151  0.267]
max(abs(x0-xant)) =  0.5
```

Como la norma es mayor que la 0.01, seguimos iterando

Iteración 2

A partir de $x^{(1)}$ calculamos $x^{(2)}$

```
[6]: k = 2
xant = np.copy(x0)
x0[0] = (b[0] - A[0,1] * x0[1] - A[0,2] * x0[2] - A[0,3] * x0[3]) / A[0,0]
x0[1] = (b[1] - A[1,0] * x0[0] - A[1,2] * x0[2] - A[1,3] * x0[3]) / A[1,1]
x0[2] = (b[2] - A[2,0] * x0[0] - A[2,1] * x0[1] - A[2,3] * x0[3]) / A[2,2]
x0[3] = (b[3] - A[3,0] * x0[0] - A[3,1] * x0[1] - A[3,2] * x0[2]) / A[3,3]
```

Imprimimos la nueva iteración, el vector diferencia y la norma de dicho vector diferencia, que nos va a valer para controlar el error

```
[7]: print ('x0          = ', x0)
print ('x0-xant      = ', x0-xant)
print ('max(abs(x0-xant)) = ', max(abs(x0-xant)))
```

```
x0          = [0.47   0.29   0.143  0.269]
x0-xant      = [-0.03  -0.035 -0.008  0.002]
max(abs(x0-xant)) =  0.035
```

Como la norma es mayor que la 0.01, seguimos iterando

Iteración 3

A partir de $x^{(2)}$ calculamos $x^{(3)}$ (aunque los guardamos respectivamente en x_0 y x_1)

```
[8]: k = 3
xant = np.copy(x0)
x0[0] = (b[0] - A[0,1] * x0[1] - A[0,2] * x0[2] - A[0,3] * x0[3]) / A[0,0]
x0[1] = (b[1] - A[1,0] * x0[0] - A[1,2] * x0[2] - A[1,3] * x0[3]) / A[1,1]
x0[2] = (b[2] - A[2,0] * x0[0] - A[2,1] * x0[1] - A[2,3] * x0[3]) / A[2,2]
x0[3] = (b[3] - A[3,0] * x0[0] - A[3,1] * x0[1] - A[3,2] * x0[2]) / A[3,3]
```

Imprimimos el nuevo resultado y la norma del vector diferencia

```
[9]: print ('x0          = ', x0)
print ('x0-xant      = ', x0-xant)
print ('max(abs(x0-xant)) = ', max(abs(x0-xant)))
```

```
x0      = [ 0.472  0.291  0.143  0.269]
x0-xant = [ 1.634e-03  1.052e-03 -2.522e-04 -6.911e-05]
max(abs(x0-xant)) = 0.001634
```

Ahora ya se verifica que la norma del vector diferencia es menor que 0.01 y paramos. Nuestra solución aproximada es

```
[10]: print ('x0 = ', x0)
```

```
x0 = [ 0.472  0.291  0.143  0.269]
```

La solución exacta es

```
[11]: x = np.linalg.solve(A,b);
print(x)
```

```
[ 0.472  0.292  0.143  0.269]
```

Y el error

```
[12]: print ('error = ', max(abs(x0-x)))
```

```
error = 5.146953503498697e-05
```

5.7 Condicionamiento de matrices

Ejercicio 5.7.1

Dados los puntos:

| | | | | | |
|-----|---|---|---|---|---|
| k | 0 | 1 | 2 | 3 | 4 |
| x | 1 | 2 | 3 | 4 | 5 |
| y | 1 | 2 | 4 | 3 | 5 |

1. Crea la matriz de Vandermonde para los nodos x y plantea el sistema que resuelve el problema de interpolación de estos puntos con dicha matriz y utilizando el comando de python, `numpy.linalg.solve`, resuélvelo.
2. Modifica algún elemento de la matriz de Vandermonde y vuelve a resolver el sistema.
3. Calcula el número de condición de la matriz utilizando `numpy.linalg.cond`. Explica por qué era de esperar este resultado.

1. Crea la matriz de Vandermonde y resuelve el sistema

Como vimos en el tema de interpolación, el sistema a resolver para calcular el polinomio de interpolación que pasa por 5 puntos sería

$$\begin{pmatrix} 1 & x_0 & x_0^2 & x_0^3 & x_0^4 \\ 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ 1 & x_3 & x_3^2 & x_3^3 & x_3^4 \\ 1 & x_4 & x_4^2 & x_4^3 & x_4^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$$

Y la matriz de coeficientes es la matriz de Vandermonde. Con los puntos dados, el sistema es

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 & 16 \\ 1 & 3 & 9 & 27 & 81 \\ 1 & 4 & 16 & 64 & 256 \\ 1 & 5 & 25 & 125 & 625 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \\ 3 \\ 5 \end{pmatrix}$$

Y resolviendo con python

```
[2]: A = np.array([[1.,1,1,1,1],
                 [1,2,4,8,16],
                 [1,3,9,27,81],
                 [1,4,16,64,256],
                 [1,5,25,125,625]])
b = np.array([1,2,4,3,5])
P = np.linalg.solve(A,b)
print('a = ', P)
```

a = [15. -28.66666667 19.08333333 -4.83333333 0.41666667]

2. Modifica algún elemento y vuelve a resolver el sistema

Si modificamos algún elemento

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 & \textcolor{red}{15} \\ 1 & 3 & 9 & 27 & 81 \\ 1 & 4 & 16 & 64 & 256 \\ 1 & 5 & \textcolor{red}{24} & 125 & 625 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 4 \\ 3 \\ 5 \end{pmatrix}$$

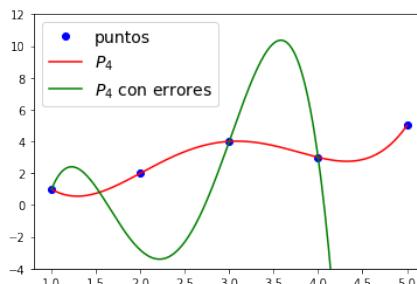
ahora la solución es

```
[3]: A1 = np.array([[1,1,1,1,1],
                  [1,2,4,8,15],
                  [1,3,9,27,81],
                  [1,4,16,64,256],
                  [1,5,24,125,625]])
b1 = np.array([1,2,4,3,5])
P1 = np.linalg.solve(A1,b1)
print('a = ', P1)
```

a = [-82. 187.91666667 -145.33333333 45.25 -4.83333333]

que, como vemos, es muy distinta de la solución anterior.

Estamos resolviendo el problema de interpolación, es decir, estamos calculando el polinomio que pasa por los puntos. Gráficamente, las soluciones son



3. Número de condición

¿Cómo se evalúa el buen o mal condicionamiento de la matriz de coeficientes? Con el *número de condición*

$$\text{cond}(A) = \det(A) \det(A^{-1})$$

El número de condición es siempre mayor que uno, pero cuanto más próximo a uno, mejor condicionada está la matriz y viceversa. En este ejemplo, si lo calculamos con el comando `numpy.linalg.cond` obtenemos

26169.68797063433

que es muy grande y por lo tanto la matriz de coeficientes del sistema está mal condicionada.

Dijimos que la matriz de Vandermonde está mal condicionada y que pequeños errores en los datos podían producir grandes errores en los resultados. Y, efectivamente, así es, como lo hemos comprobado.

TEMA 6

OPTIMIZACIÓN

6.1 Mínimo local y global. Cálculo simbólico

Ejercicio 6.1.1

1. Usando las condiciones necesarias de primer orden, encontrar un punto de mínimo de la función

$$f(x, y, z) = 2x^2 + xy + y^2 + yz + z^2 - 6x - 7y - 8z + 9.$$

2. Verificar que dicho punto es un mínimo local usando las condiciones suficientes de segundo orden.
3. Probar que el mínimo local encontrado es, de hecho, global.

INTRODUCCIÓN

Mínimo de una función de varias variables

Dado $f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$ decimos que $x^* \in \Omega$ es un mínimo de f en Ω si $f(x^*) \leq f(x)$ para todo $x \in \Omega$.

Casos principales

- Optimización sin restricciones: $\Omega = \mathbb{R}^n$.
- Optimización con restricciones: $\Omega \subsetneq \mathbb{R}^n$, habitualmente determinada por un conjunto de restricciones dadas por igualdades o desigualdades.

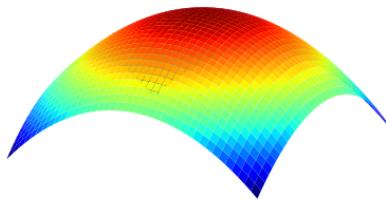
No hay técnicas que resuelvan, en general, el *problema de optimización global*. Por lo tanto, habitualmente se resuelve el problema de *optimización local*:

Encontrar $x^* \in \Omega$ de forma que $f(x^*) \leq f(x)$ para todo x tal que $\|x - x^*\| \leq R$.

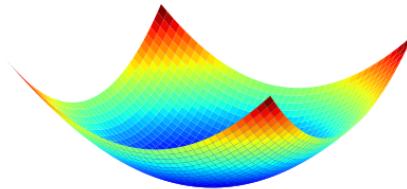
Si f es estrictamente convexa y Ω es un conjunto estrictamente convexo, entonces f tiene un mínimo global único en Ω .

El problema del encontrar un máximo se transforma fácilmente en el problema de encontrar un mínimo cambiando el signo de la función. Por ejemplo, la función de dos variables $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, $f(x, y) = -x^2 - y^2$ tiene un máximo en $(0, 0)$ y la función $-f(x, y) = x^2 + y^2$ tiene un mínimo en $(0, 0)$

$$f(x, y) = -x^2 - y^2 \text{ (máximo)}$$

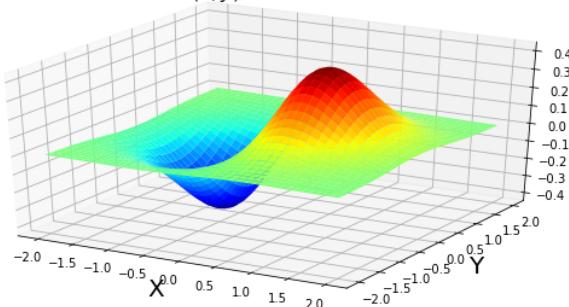


$$-f(x, y) = x^2 + y^2 \text{ (mínimo)}$$



Así que solo trataremos el problema de encontrar un mínimo. Empecemos repasando las condiciones necesarias y suficientes de mínimo de una función de varias variables suficientemente regular con el ejemplo de la función de dos variables $f(x, y) = xe^{-x^2-y^2}$. Si la representamos en $(x, y) \in [-2, 2] \times [-2, 2]$

$$f(x, y) = xe^{-x^2-y^2}$$



Vemos que tiene:

- Un mínimo relativo en $(x_0, y_0) = \left(-\frac{1}{\sqrt{2}}, 0\right) \approx (-0.71, 0)$
- Un máximo relativo en $(x_1, y_1) = \left(\frac{1}{\sqrt{2}}, 0\right) \approx (0.71, 0)$

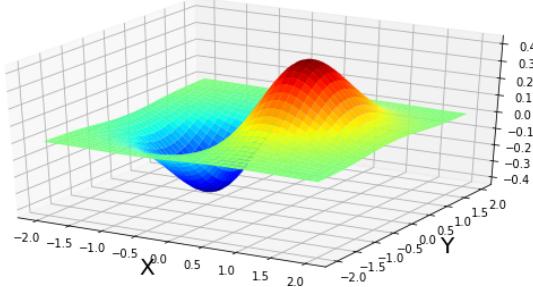
Condición necesaria de mínimo relativo

Si hacemos y constante e igual a cero, obtenemos la curva

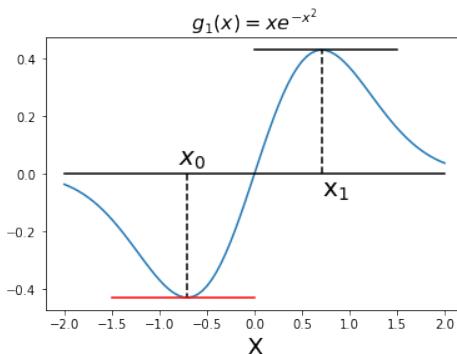
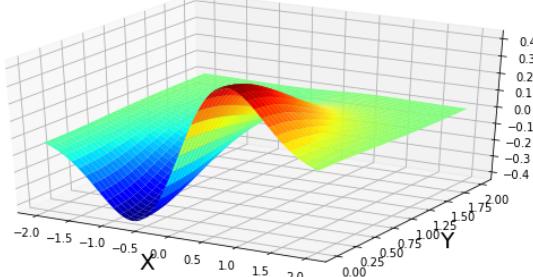
$$g_1(x) = f(x, 0) = x e^{-x^2-0^2} = x e^{-x^2}$$

que es la curva que obtenemos al cortar la superficie por el plano $y = 0$

$$f(x, y) = x e^{-x^2-y^2}$$

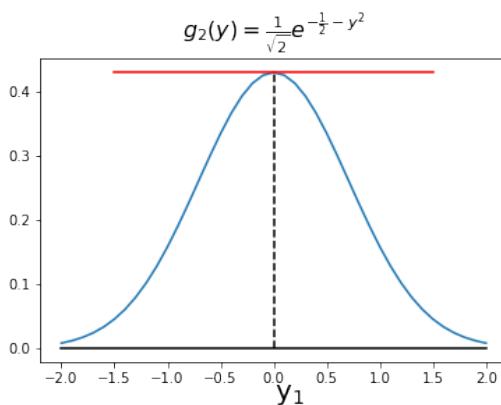
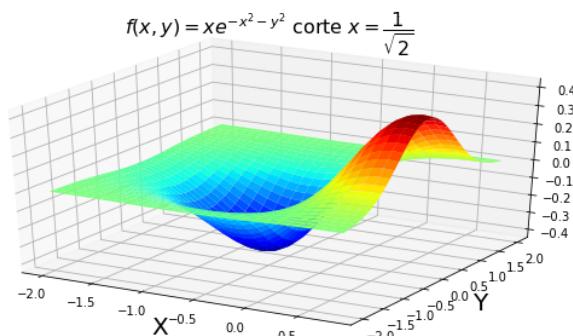
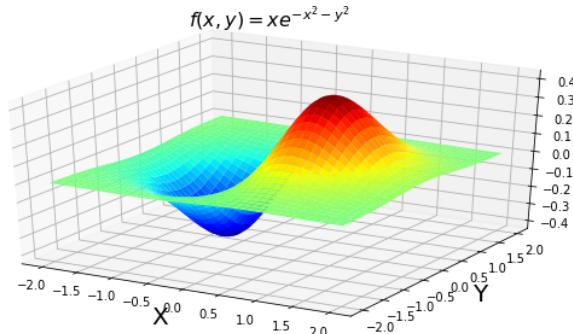


$$f(x, y) = x e^{-x^2-y^2} \text{ corte } y = 0$$



Y vemos que la derivada de esta curva en $x_0 = -\frac{1}{\sqrt{2}}$ y $x_1 = \frac{1}{\sqrt{2}}$ es igual a cero (tangente horizontal). Es decir, en ambos casos, la derivada parcial de la superficie f respecto de x es cero en estos puntos.

Si ahora hacemos x constante e igual a $\frac{1}{\sqrt{2}}$ equivale a cortar la superficie por el plano $x = \frac{1}{\sqrt{2}}$ y tenemos la curva $g_2(y) = f\left(\frac{1}{\sqrt{2}}, y\right) = \frac{1}{\sqrt{2}}e^{-\frac{1}{2}y^2}$

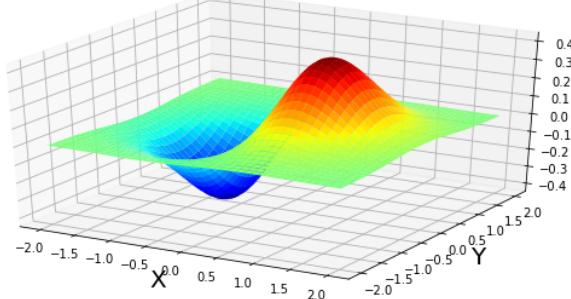


Y vemos que la derivada de esta curva en $y_1 = 0$ es igual a cero (tangente horizontal). Es decir, la derivada parcial de la superficie f respecto de la variable y es cero en este punto.

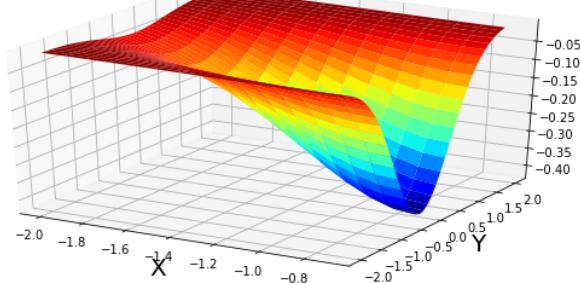
Y si ahora hacemos x constante e igual a $-\frac{1}{\sqrt{2}}$ equivale a cortar la superficie por el plano $x = -\frac{1}{\sqrt{2}}$ y tenemos la curva

$$g_2(y) = f\left(-\frac{1}{\sqrt{2}}, y\right) = -\frac{1}{\sqrt{2}}e^{-\frac{1}{2}-y^2}$$

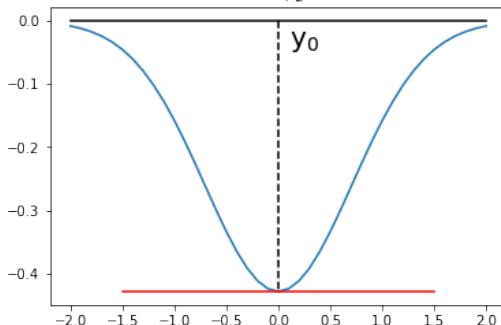
$$f(x, y) = xe^{-x^2 - y^2}$$



$$f(x, y) = xe^{-x^2 - y^2} \text{ corte } x = -\frac{1}{\sqrt{2}}$$



$$g_3(y) = -\frac{1}{\sqrt{2}}e^{-\frac{1}{2}-y^2}$$



Y vemos que la derivada de esta curva en $y_0 = 0$ es igual a cero (tangente horizontal). Es decir, la derivada parcial de la superficie f respecto de la variable y es cero en este punto.

Por lo tanto, si tenemos un máximo o un mínimo de una función de varias variables, la condición necesaria es que sus derivadas parciales sean cero. En particular, para una función de dos variables como f

$$\frac{\partial f}{\partial x} = 0 \quad \frac{\partial f}{\partial y} = 0$$

Y para la función del ejemplo $f(x, y) = x e^{-x^2-y^2}$

$$\begin{aligned}\frac{\partial f}{\partial x} &= 0 \quad e^{-x^2-y^2} + x e^{-x^2-y^2}(-2x) = 0 \quad e^{-x^2-y^2}(1 - 2x^2) = 0 \quad 1 - 2x^2 = 0 \\ \frac{\partial f}{\partial y} &= 0 \quad x e^{-x^2-y^2}(-2y) = 0 \quad e^{-x^2-y^2}(-2y) = 0 \quad -2y = 0\end{aligned}$$

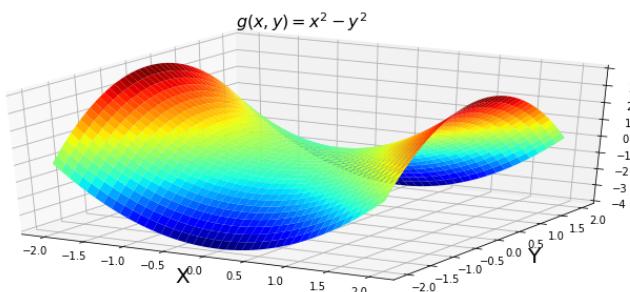
Y la solución de este sistema no lineal es

$$x = \pm \frac{1}{\sqrt{2}} \quad y = 0$$

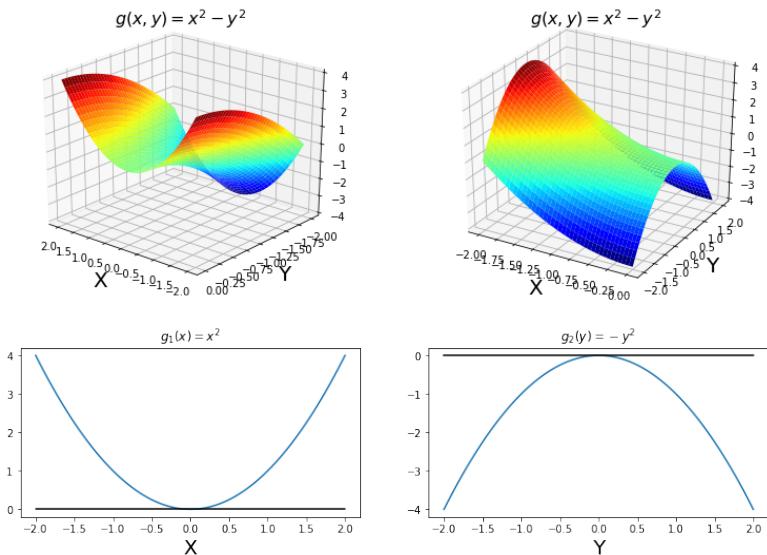
Y por lo tanto las soluciones son

$$(x_0, y_0) = \left(-\frac{1}{\sqrt{2}}, 0 \right) \approx (-0.71, 0) \quad (x_1, y_1) = \left(\frac{1}{\sqrt{2}}, 0 \right) \approx (0.71, 0)$$

Estas son condiciones necesarias, no suficientes, porque pueden darse y no haber ni un máximo ni un mínimo, y entonces se llama *punto de silla*. Un ejemplo es la función



Si en el punto $(0, 0)$ cortamos por el plano $y = 0$ tenemos un mínimo, pero si cortamos por el plano $x = 0$ tenemos un máximo. Por lo tanto en $(0, 0)$ la superficie no tiene ni un máximo ni un mínimo, aunque se cumplen las condiciones necesarias de extremo.



Condición suficiente de mínimo relativo

Una vez hemos establecido que en el punto se cumplen las condiciones necesarias la condición suficiente de mínimo es que la función sea convexa en el punto. En una dimensión, analíticamente era $f''(x_0) > 0$. En más de una dimensión, la matriz Hessiana en el punto ha de ser definida positiva. La matriz hessiana es

$$H(x, y) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix} = \begin{pmatrix} 2x(2x^2 - 3)e^{-x^2-y^2} & 2y(2x^2 - 1)e^{-x^2-y^2} \\ 2y(2x^2 - 1)e^{-x^2-y^2} & 2x(2y^2 - 1)e^{-x^2-y^2} \end{pmatrix}$$

$$H(x, y) = e^{-x^2-y^2} \begin{pmatrix} 2x(2x^2 - 3) & 2y(2x^2 - 1) \\ 2y(2x^2 - 1) & 2x(2y^2 - 1) \end{pmatrix}$$

$$\text{Y en el punto } (x_0, y_0) = \left(-\frac{1}{\sqrt{2}}, 0 \right) \approx (-0.71, 0)$$

$$H \left(-\frac{1}{\sqrt{2}}, 0 \right) = e^{-1/2} \begin{pmatrix} -\frac{2}{\sqrt{2}} \left(2 \frac{1}{2} - 3 \right) & 0 \\ 0 & \frac{2}{\sqrt{2}} \end{pmatrix}$$

$$H \left(-\frac{1}{\sqrt{2}}, 0 \right) = \frac{2}{\sqrt{2}} e^{-1/2} \begin{pmatrix} -(2 \frac{1}{2} - 3) & 0 \\ 0 & 1 \end{pmatrix} = \frac{2}{\sqrt{2}} e^{-1/2} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

Una forma de comprobar si una matriz es definida positiva es tomar todos los menores principales y si sus determinantes son positivos, la matriz es definida positiva.

En este caso, podemos no tener en cuenta el factor positivo que hemos sacado fuera de la matriz y

$$|2| = 2 > 0 \quad \left| \begin{array}{cc} 2 & 0 \\ 0 & 1 \end{array} \right| = 2 > 0$$

y la matriz hessiana es definida positiva en el punto (x_0, y_0) y podemos garantizar que hay un mínimo local.

EJERCICIO

1. Usando las condiciones necesarias de primer orden, encontrar un punto de mínimo de la función

$$f(x, y, z) = 2x^2 + xy + y^2 + yz + z^2 - 6x - 7y - 8z + 9.$$

La condición necesaria de mínimo para un punto (x_m, y_m, z_m) es que $\nabla f(x_m, y_m, z_m) = (f'_x, f'_y, f'_z) = (0, 0, 0)$. Es decir

$$\begin{cases} f'_x = 0 \\ f'_y = 0 \\ f'_z = 0 \end{cases} \quad \begin{cases} 4x + y - 6 = 0 \\ x + 2y + z - 7 = 0 \\ y + 2z - 8 = 0 \end{cases} \quad \begin{cases} x = 1.2 \\ y = 1.2 \\ z = 3.4 \end{cases}$$

Y el punto $(x_m, y_m) = (1.2, 1.2, 3.4)$ cumple las condiciones necesarias de mínimo.

Veamos cómo resolvemos el sistema por Gauss.

Queremos resolver el sistema

$$\begin{array}{rcl} 4x & + & y & + & = & 6 \\ x & + & 2y & + & z & = 7 \\ & & y & + & 2z & = 8 \end{array}$$

Hacemos ceros por debajo del pivote **4** en la primera columna.

$$\begin{array}{l} f_1 \\ f_2 \\ f_3 \end{array} \left(\begin{array}{ccc|c} \textcolor{red}{4} & 1 & 0 & 6 \\ 1 & 2 & 1 & 7 \\ 0 & 1 & 2 & 8 \end{array} \right) \quad \begin{array}{l} f'_1 \\ f'_2 \\ f'_3 \end{array} = \begin{array}{l} f_1 \\ f_2 - \frac{1}{4}f_1 \\ f_3 - \frac{1}{4}f_1 \end{array}$$

Hacemos ceros por debajo del pivote **-3** en la segunda columna.

$$\begin{array}{l} f'_1 \\ f'_2 \\ f'_3 \end{array} \left(\begin{array}{ccc|c} 4 & 1 & 0 & 6 \\ 0 & \textcolor{red}{7/4} & 1 & 11/2 \\ 0 & 1 & 2 & 8 \end{array} \right) \quad \begin{array}{l} f''_1 \\ f''_2 \\ f''_3 \end{array} = \begin{array}{l} f'_1 \\ f'_2 \\ f'_3 - \frac{1}{(7/4)}f'_2 \end{array}$$

Y ya tenemos una matriz triangular superior (con ceros por debajo de la diagonal principal).

$$\begin{array}{c} f_1'' \\ f_2'' \\ f_3'' \end{array} \left(\begin{array}{ccc|c} 4 & 1 & 0 & 6 \\ 0 & 7/4 & 1 & 11/2 \\ 0 & 0 & 10/7 & 34/7 \end{array} \right)$$

$$\begin{array}{rcl} 4x + y & = & 6 \\ \frac{7}{4}y + z & = & \frac{11}{2} \\ \frac{10}{7}z & = & \frac{34}{7} \end{array}$$

De la última ecuación obtenemos

$$z = \frac{\frac{34}{7}}{\frac{10}{7}} = \frac{34}{10} = 3.4$$

De la segunda ecuación

$$y = \frac{\frac{11}{2} - z}{\frac{7}{4}} = \frac{\frac{11}{2} - \frac{34}{10}}{\frac{7}{4}} = \frac{6}{5} = 1.2$$

Y finalmente, de la tercera

$$x = \frac{6 - y}{4} = \frac{6 - \frac{6}{5}}{4} = \frac{6}{5} = 1.2$$

Y por lo tanto

$$\boxed{x = 1.2 \quad y = 1.2 \quad z = 3.4}$$

2. Verificar que dicho punto es un mínimo local usando las condiciones suficientes de segundo orden.

Como

$$\begin{cases} f'_x = 4x + y - 6 \\ f'_y = x + 2y + z - 7 \\ f'_z = y + 2z - 8 \end{cases}$$

La condición suficiente es que la matriz Hessiana en el punto (x_m, y_m, z_m) sea definida positiva. La matriz Hessiana para la función f en cualquier punto (x, y) es

$$H_f = \begin{pmatrix} f''_{xx} & f''_{xy} & f''_{xz} \\ f''_{yx} & f''_{yy} & f''_{yz} \\ f''_{zx} & f''_{zy} & f''_{zz} \end{pmatrix} = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

Y en particular, en el punto $(x_m, y_m, z_m) = (1.2, 1.2, 3.4)$, como todos los elementos son constantes, el valor es el mismo.

La condición suficiente de mínimo es que la matriz hessiana sea definida positiva. Una condición para que una matriz sea definida positiva es que los determinantes de los menores principales han de ser estrictamente positivos. Es decir

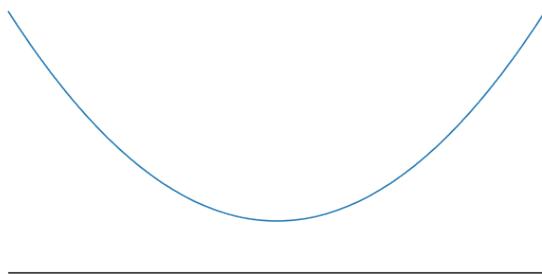
$$\begin{vmatrix} 4 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{vmatrix} = 10 > 0 \quad \begin{vmatrix} 4 & 1 \\ 1 & 2 \end{vmatrix} = 7 > 0 \quad y \quad |4| = 4 > 0.$$

Por lo tanto la matriz es definida positiva y se cumple la condición suficiente de mínimo.

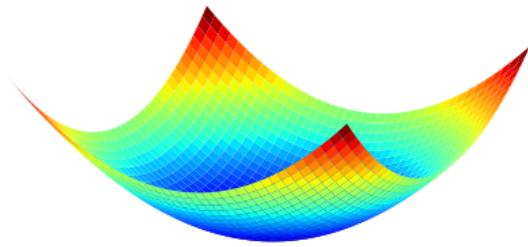
3. Probar que el mínimo local encontrado es, de hecho, global

Para demostrar que este *mínimo es único*, basta con demostrar que la función es convexa.

Función convexa de una variable
 $f'(x) > 0$



Función convexa de dos variables
 H definida positiva



Si el hessiano es definido positivo para todos los puntos de la función, esta es convexa. Como la matriz hessiana tiene elementos constantes y es definida positiva, lo es para toda la función y caso de existir un mínimo, es único.

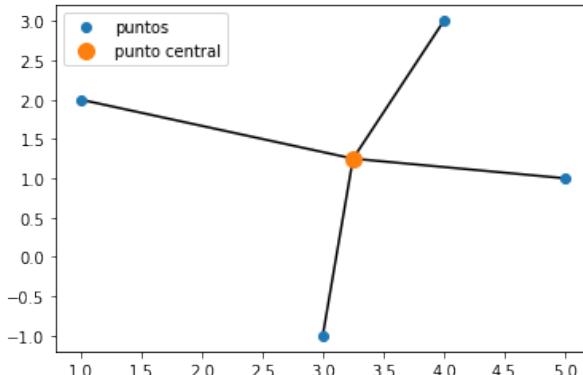
Ejercicio 6.1.2

Dado un conjunto de puntos (x_i, y_i) con $i = 1, 2, \dots, n$ encontrar el punto central (m_x, m_y) de forma que la suma de las distancias cuadráticas

$$D = \sum_{i=1}^n [(x_i - m_x)^2 + (y_i - m_y)^2]$$

a dicho punto sea mínima. El conjunto de puntos será

| | | | | |
|-----|---|----|---|---|
| x | 1 | 3 | 4 | 5 |
| y | 2 | -1 | 3 | 1 |



Empecemos aplicándolo para el caso particular

$$D(m_x, m_y) = d_1^2 + d_2^2 + d_3^2 + d_4^2$$

$$\begin{aligned} D(m_x, m_y) &= [(1 - m_x)^2 + (2 - m_y)^2] + [(3 - m_x)^2 + (-1 - m_y)^2] + \\ &\quad [(4 - m_x)^2 + (3 - m_y)^2] + [(5 - m_x)^2 + (1 - m_y)^2] \end{aligned}$$

La condición necesaria de óptimo es que las derivadas parciales sean cero

$$\frac{\partial D}{\partial m_x} = 0 \quad -2(1 - m_x) - 2(3 - m_x) - 2(4 - m_x) - 2(5 - m_x) = 0$$

$$\frac{\partial D}{\partial m_y} = 0 \quad -2(2 - m_y) - 2(-1 - m_y) - 2(3 - m_y) - 2(1 - m_y) = 0$$

Quitamos el factor -2 de las dos ecuaciones

$$\begin{aligned} (1 - m_x) + (3 - m_x) + (4 - m_x) + (5 - m_x) &= 0 \\ (2 - m_y) + (-1 - m_y) + (3 - m_y) + (1 - m_y) &= 0 \end{aligned}$$

Reorganizamos las ecuaciones

$$\begin{aligned}(1+3+4+5) &= 4m_x \\ (2-1+3+1) &= 4m_y\end{aligned}$$

La solución es

$$\begin{aligned}m_x &= \frac{1+3+4+5}{4} = 3.25 \\ m_y &= \frac{2-1+3+1}{4} = 1.25\end{aligned}$$

La condición suficiente es que la matriz Hessiana en el punto $(3.25, 1.25)$ sea definida positiva. La matriz Hessiana para la función D en cualquier punto (m_x, m_y) es

$$H(m_x, m_y) = \begin{pmatrix} \frac{\partial^2 D}{\partial m_x^2} & \frac{\partial^2 D}{\partial m_x \partial m_y} \\ \frac{\partial^2 D}{\partial m_y \partial m_x} & \frac{\partial^2 D}{\partial m_y^2} \end{pmatrix}$$

$$H(m_x, m_y) = \begin{pmatrix} 2+2+2+2 & 0 \\ 0 & 2+2+2+2 \end{pmatrix} = \begin{pmatrix} 8 & 0 \\ 0 & 8 \end{pmatrix}$$

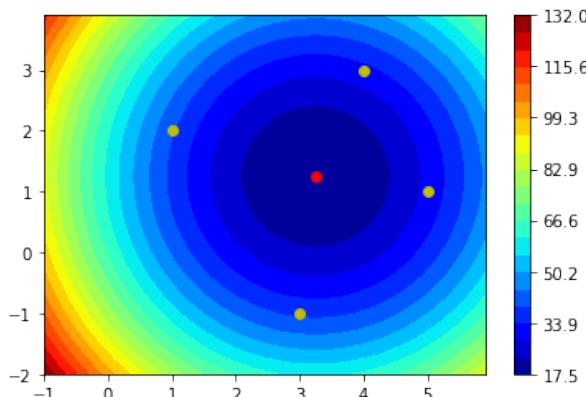
Y la matriz Hessiana es constante y la misma para todos los puntos del plano. En particular, es también el valor para el punto $(3.25, 1.25)$.

Una condición para que una matriz sea definida positiva es que los determinantes de los menores principales han de ser estrictamente positivos. Es decir

$$\begin{vmatrix} 8 & 0 \\ 0 & 8 \end{vmatrix} = 64 > 0 \quad \text{y} \quad |8| = 8 > 0.$$

Y se cumple la condición suficiente de mínimo.

En el dibujo siguiente vemos los valores que la función D toma para diferentes puntos y que, efectivamente, el punto $(3.25, 1.25)$ (en rojo) es el punto donde la suma de distancias a los otros cuatro puntos es mínima.



6.2 Mínimo local. Cálculo numérico. Métodos del gradiente, de Newton y de la sección áurea

Ejercicio 6.2.1

Aproximar con una iteración el mínimo de la función

$$f(x, y) = x^2 + 3y^2$$

comenzando por el punto inicial $(2, 1)$ y utilizando

1. El método del gradiente:
 - (a) Con el descenso más pronunciado.
 - (b) Con tasa de aprendizaje 0.1.
2. El método de Newton.
3. Utilizar el método de la sección áurea para optimizar el paso de la primera iteración del descenso más pronunciado.

INTRODUCCIÓN

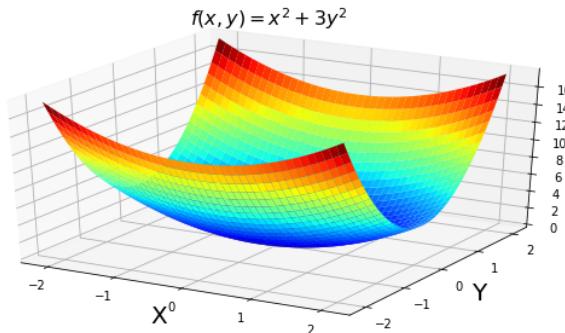
1. El método del gradiente

Los métodos de gradiente son un conjunto de métodos que, como su nombre indica, se basan en el uso del gradiente de la función en cada punto. Veamos un par de ejemplos

1.a Descenso más pronunciado

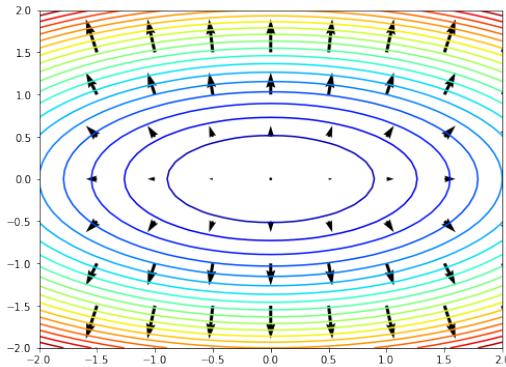
Busquemos el mínimo de una función. Pensemos en una función de dos variables y por lo tanto estamos sobre una superficie

En el siguiente gráfico vemos una superficie que tiene un mínimo en $(0, 0)$. Es un paraboloide.

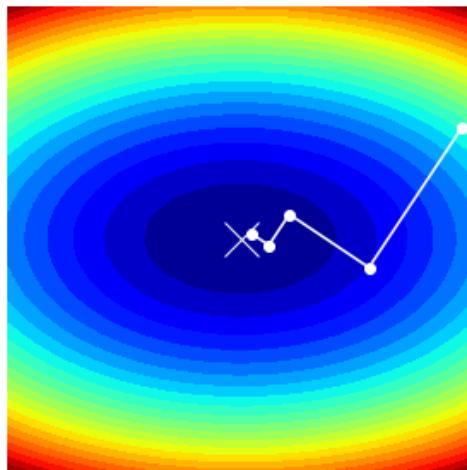


En el siguiente gráfico vemos los gradientes en algunos puntos de la superficie. Podemos observar que para cada punto

- Se orientan en la dirección de máximo crecimiento de la función.
- Su módulo depende de cuánto crece la función. A mayor crecimiento mayor módulo.

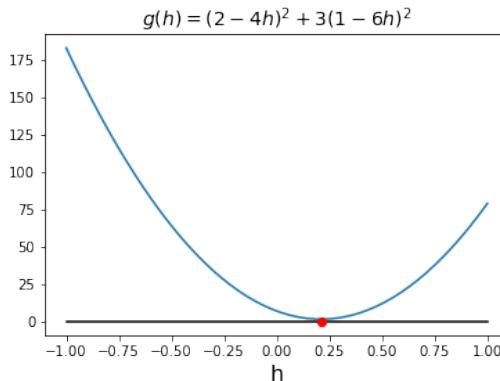


Busquemos el mínimo de esta función



- Nos situamos en un punto inicial (x_0, y_0) , que está situado en la esquina de arriba a la derecha.
- Buscamos la dirección del gradiente.
- Nos orientamos en sentido contrario al gradiente, puesto que el gradiente se orienta en sentido creciente de la función, hacia arriba, y nosotros buscamos un mínimo y queremos ir hacia abajo.
- Avanzamos en línea recta. Tenemos que ir bajando.
- En el momento que encontrremos el mínimo que está sobre nuestra trayectoria recta (cuando empezamos a subir otra vez), paramos. Estaremos en (x_1, y_1) .
- Buscamos de nuevo la dirección del gradiente y repetimos los pasos anteriores hasta que se cumpla una condición de parada.

El siguiente gráfico da la altura de la primera trayectoria recta en función de h . Descendemos hasta que llegamos al punto rojo, que es el punto de mínima altura dentro de nuestra trayectoria.



Utilizando el método del gradiente, vamos a aproximar el mínimo de la función $f(x, y) = x^2 + 3y^2$ utilizando el punto $(2,1)$ como punto inicial.

Realizaremos una iteración por el método del gradiente usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - h \nabla f(x_0, y_0) \quad (1)$$

y buscaremos h para que la función

$$g(h) = f(x_0 - h f_x(x_0, y_0), y_0 - h f_y(x_0, y_0))$$

tenga un valor mínimo.

Tenemos que

$$\nabla f = (f_x, f_y) = (2x, 6y)$$

y, sustituyendo el valor del punto inicial

$$\nabla f(x_0, y_0) = \nabla f(2, 1) = (4, 6).$$

y, sustituyendo el valor del punto inicial y del gradiente en el punto inicial en la función g

$$g(h) = f(x_0 - h f_x(x_0, y_0), y_0 - h f_y(x_0, y_0)) = f(2 - 4h, 1 - 6h)$$

$$g(h) = (2 - 4h)^2 + 3(1 - 6h)^2$$

Ahora tendríamos que calcular el mínimo de esta función de una variable numéricamente. Para simplificar el ejemplo, lo haremos analíticamente.

Calculemos el mínimo de esta función teniendo en cuenta que la condición necesaria de mínimo es que $g'(h) = 0$

$$g'(h) = 2(-4)(2 - 4h) + 6(-6)(1 - 6h) = 248h - 52 = 0 \implies h = 0.2097 \approx 0.21$$

y usando este valor en la ecuación (1)

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} - 0.21 \begin{pmatrix} 4 \\ 6 \end{pmatrix} = \begin{pmatrix} 1.16 \\ -0.26 \end{pmatrix}$$

Hemos mejorado porque

$$f(x_0, y_0) = f(2, 1) = 7 \quad y \quad f(x_1, y_1) = f(1.16, -0.26) = 1.55$$

que es un valor menor. La siguiente iteración se realizaría usando la fórmula

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - h \nabla f(x_1, y_1) \quad (2)$$

para lo cual habría que recalcular h de forma análoga a la primera iteración, calculándola para que

$$g(h) = f(x_1 - h f_x(x_1, y_1), y_1 - h f_y(x_1, y_1))$$

sea mínimo y sustituyendo entonces este h en (2).

1.b Método del descenso del gradiente con tasa de aprendizaje $\eta = 0.1$

En este caso no optimizamos el paso, es decir, no buscamos el punto mínimo de nuestra trayectoria recta, sino que usamos siempre el mismo h y entonces a este parámetro se le llama η o tasa de aprendizaje y hay que fijarlo. Si es demasiado pequeño, el algoritmo convergerá demasiado despacio y si es demasiado grande puede no converger.

Este método tiene la ventaja de su sencillez y el inconveniente de que hay que buscar un η adecuado.

En cada paso usamos la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \eta \nabla f(x_0, y_0)$$

Como

$$\nabla f = (f_x, f_y) = (2x, 6y)$$

tenemos

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \eta \begin{pmatrix} 2x_0 \\ 6y_0 \end{pmatrix}$$

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} - 0.1 \begin{pmatrix} 2(2) \\ 6(1) \end{pmatrix} = \begin{pmatrix} 1.6 \\ 0.4 \end{pmatrix}$$

$$f(1.6, 0.4) = 3.04$$

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \eta \begin{pmatrix} 2x_1 \\ 6y_1 \end{pmatrix}$$

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1.6 \\ 0.4 \end{pmatrix} - 0.1 \begin{pmatrix} 2(1.6) \\ 6(0.4) \end{pmatrix} = \begin{pmatrix} 1.28 \\ 0.16 \end{pmatrix}$$

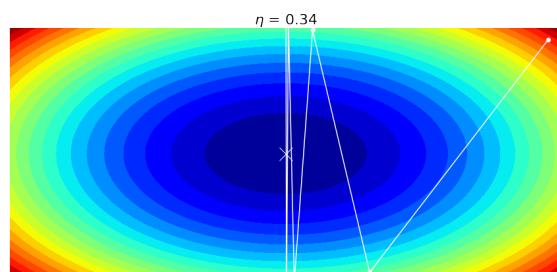
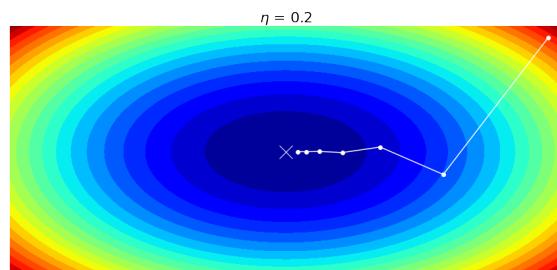
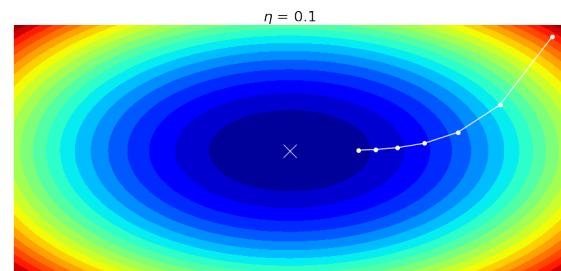
$$f(1.28, 0.16) = 1.7152$$

$$\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} - \eta \begin{pmatrix} 2x_2 \\ 6y_2 \end{pmatrix}$$

$$\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1.28 \\ 0.16 \end{pmatrix} - 0.1 \begin{pmatrix} 2(1.28) \\ 6(0.16) \end{pmatrix} = \begin{pmatrix} 1.024 \\ 0.064 \end{pmatrix}$$

$$f(1.024, 0.064) = 1.060864$$

En los gráficos siguientes vemos como afecta la tasa de aprendizaje a la convergencia



2. El método de Newton

Recordamos el método de Newton para encontrar una raíz de la ecuación no lineal $f(x) = 0$. Partiendo de un valor inicial x_0 , usando la fórmula

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

y obtenemos

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}, \quad x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}, \quad x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}, \quad \dots \rightarrow \alpha$$

Si quisieramos encontrar un máximo o un mínimo de la función f , si f es suficientemente regular, la condición necesaria de extremo es

$$f'(x) = 0$$

Es decir, buscamos una raíz de esta ecuación y, usando Newton, podríamos resolverlo, tomando un valor inicial x_0 , y usando la fórmula

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

De forma análoga, para funciones de n variables $f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$, teniendo en cuenta que las componentes del vector gradiente de una función $\nabla f(\mathbf{x})$ son sus derivadas parciales primeras, y las componentes de la matriz Hessiana $H(\mathbf{x})$ son las derivadas parciales segundas, podemos formular Newton partiendo de un valor inicial \mathbf{x}_0

$$\mathbf{x}_{k+1} = \mathbf{x}_k - H^{-1}(\mathbf{x}_k) \cdot \nabla f(\mathbf{x}_k)$$

Esto no es una demostración. Tanto en el caso unidimensional como en el multidimensional se puede construir el método de Newton a partir de la fórmula de Taylor correspondiente, y esa sería la demostración formal.

Como nuestra función es de dos variables, 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

$$H(x_0, y_0) \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \nabla f(x_0, y_0) \quad (1)$$

entonces

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = H^{-1}(x_0, y_0) \cdot \nabla f(x_0, y_0)$$

y escribiremos

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \quad (2)$$

donde $(c_1, c_2)^T$ es la solución del sistema (1). En general, tiene más sentido resolver el sistema (1) que calcular la matriz inversa de H y luego multiplicarla por el gradiente, porque calcular la inversa de una matriz equivale a resolver n sistemas (aunque con la misma matriz de coeficientes) y de esta forma estamos resolviendo solo un sistema.

La función a minimizar es

$$f(x, y) = x^2 + 3y^2$$

y empezaremos con el punto inicial $(2, 1)$

Tenemos que

$$\nabla f = \left(\frac{\partial f(x, y)}{\partial x}, \frac{\partial f(x, y)}{\partial y} \right) = (2x, 6y)$$

y

$$\nabla f(x_0, y_0) = \nabla f(2, 1) = (4, 6).$$

Además

$$H = \begin{pmatrix} \frac{\partial^2 f(x, y)}{\partial x \partial x} & \frac{\partial^2 f(x, y)}{\partial x \partial y} \\ \frac{\partial^2 f(x, y)}{\partial y \partial x} & \frac{\partial^2 f(x, y)}{\partial y \partial y} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix}$$

y

$$H(x_0, y_0) = H(2, 1) = \begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix}$$

El sistema (1) es

$$\begin{pmatrix} 2 & 0 \\ 0 & 6 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 6 \end{pmatrix}$$

y resolviéndolo

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

Por lo tanto (2) es

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Hemos mejorado porque

$$f(x_0, y_0) = f(2, 1) = 7 \text{ y } f(x_1, y_1) = f(0, 0) = 0$$

que es menor.

En este caso, hemos llegado al mínimo con una sola iteración.

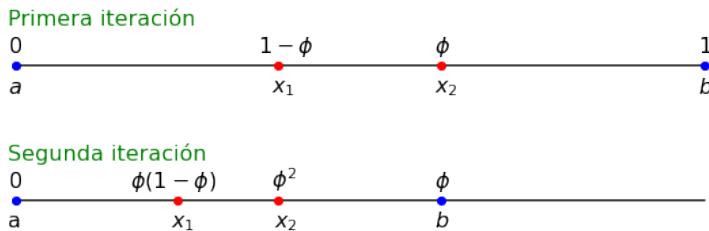
3. El método de la sección áurea

Vamos a optimizar una función g de una sola variable. El método de la sección áurea es un método de intervalo. Si el intervalo contiene un único mínimo, convergerá a dicho mínimo.

Si el intervalo contiene un único mínimo en $[a, b]$ y tenemos dos puntos interiores x_1 y x_2 de forma que $a < x_1 < x_2 < b$ teniendo en cuenta los valores de la función g en x_1 y x_2 , podemos descartar $[a, x_1)$ o $(x_2, b]$ manteniendo el mínimo dentro del intervalo.

Escogemos los puntos de forma que estén a una fracción ϕ y $1 - \phi$ de uno de los extremos, con $\phi \in (0.5, 1]$. Y con la condición de que, si descartamos $(x_2, b]$ el siguiente x_2 sea igual que el antiguo x_1 . Y si descartamos $[a, x_1)$ el siguiente x_1 sea el antiguo x_2 . Así, en cada iteración, solo tendremos que evaluar la función en un nuevo punto.

Sin perder generalidad, podemos suponer que el intervalo inicial $[a, b]$ es $[0, 1]$ y que en la primera iteración descartamos el intervalo de la derecha.

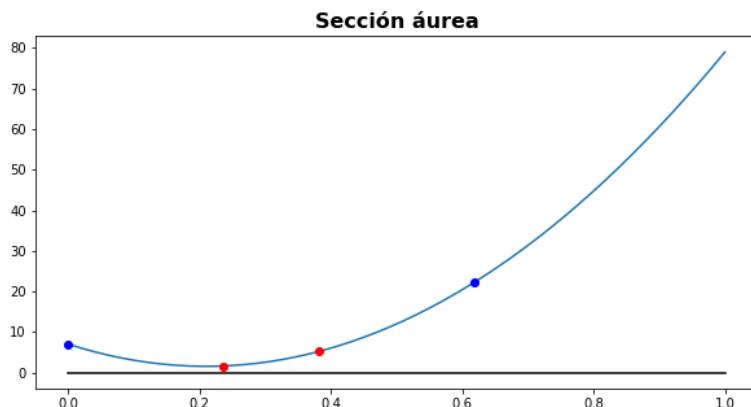
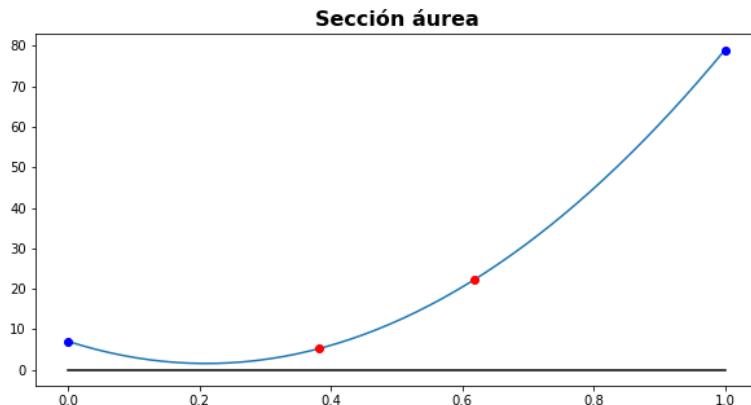


- En la primera iteración, y teniendo en cuenta que $\phi \in (0.5, 1]$, vemos la disposición de los puntos a , x_1 , x_2 y b y sus valores.
- Si descartamos $(x_2, b]$, la nueva disposición de los puntos es la de la segunda iteración, donde los puntos x_1 y x_2 son $1 - \phi$ y ϕ multiplicados por ϕ porque el intervalo ya no tiene longitud 1 sino longitud ϕ .
- Para que se cumpla que el nuevo x_2 es igual al antiguo x_1 se tendrá que verificar que

$$\phi^2 = 1 - \phi \quad \Rightarrow \quad \phi = \frac{\sqrt{5} - 1}{2} \approx 0.618034$$

Algoritmo

- Sea $a_1 = a$, $b_1 = b$ y $\phi = \frac{\sqrt{5} - 1}{2}$
- Para $k = 1, 2, \dots, \text{MaxNumIter}$
 - Calcular los puntos
 - * $x_1 = a + (1 - \phi)(b - a)$
 - * $x_2 = a + \phi(b - a)$
 - Si $g(x_1) > g(x_2)$ entonces:
 - * $a_{k+1} = x_1$
 - * $b_{k+1} = b_k$
 - En otro caso:
 - * $a_{k+1} = a_k$
 - * $b_{k+1} = x_2$
 - Si se satisface el criterio de parada, parar.



La idea es que el mínimo siempre esté dentro del intervalo.

Vamos a aplicarlo a calcular el mínimo de la función

$$g(h) = (2 - 4h)^2 + 3(1 - 6h)^2$$

| k | a | x_1 | x_2 | b | $g(x_1)$ | $g(x_2)$ |
|-----|--------|--------|--------|--------|----------|----------|
| 1 | 0.0000 | 0.3820 | 0.6180 | 1.0000 | 5.2291 | 22.2260 |
| 2 | 0.0000 | 0.2361 | 0.3820 | 0.6180 | 1.6347 | 5.2291 |
| 3 | 0.0000 | 0.1459 | 0.2361 | 0.3820 | 2.0528 | 1.6347 |
| 4 | 0.1459 | 0.2361 | 0.2918 | 0.3820 | 1.6347 | 2.3846 |
| 5 | 0.1459 | 0.2016 | 0.2361 | 0.2918 | 1.5564 | 1.6347 |
| 6 | 0.1459 | 0.1803 | 0.2016 | 0.2361 | 1.6551 | 1.5564 |
| 7 | 0.1803 | 0.2016 | 0.2148 | 0.2361 | 1.5564 | 1.5516 |
| 8 | 0.2016 | 0.2148 | 0.2229 | 0.2361 | 1.5516 | 1.5701 |
| 9 | 0.2016 | 0.2098 | 0.2148 | 0.2229 | 1.5484 | 1.5516 |
| 10 | 0.2016 | 0.2067 | 0.2098 | 0.2148 | 1.5495 | 1.5484 |

Iteración 1

- En la primera iteración calculamos x_1 y x_2 .
- Como $g(x_1) < g(x_2)$ quiere decir que el mínimo está a la izquierda:
 - Nos deshacemos del intervalo de la derecha.
 - El x_1 se convierte en el nuevo x_2 .
 - Calculamos el nuevo x_1 .

Iteración 2

- Como $g(x_1) < g(x_2)$ quiere decir que el mínimo está a la izquierda:
 - Nos deshacemos del intervalo de la derecha.
 - El x_1 se convierte en el nuevo x_2 .
 - Calculamos el nuevo x_1 .

Iteración 3

- Como $g(x_1) > g(x_2)$ quiere decir que el mínimo está a la derecha:
 - Nos deshacemos del intervalo de la izquierda.
 - El x_2 se convierte en el nuevo x_1 .
 - Calculamos el nuevo x_2 .

Y así sucesivamente.

De los dos últimos x_1 y x_2 tomaremos como aproximación aquel en el que la función tiene menor valor que es

0.2098

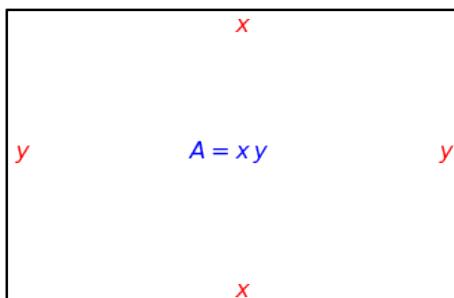
6.3 Mínimo con restricciones de igualdad. Cálculo simbólico: multiplicadores de Lagrange. Cálculo numérico: método de penalización

Ejercicio 6.3.1

Hallar el rectángulo de perímetro dado que maximiza el área mediante la resolución de las condiciones necesarias de primer orden. Verificar que las condiciones suficientes de segundo orden se cumplen.

INTRODUCCIÓN

Multiplicadores de Lagrange



Se puede plantear el problema como minimizar $f(x,y) = -xy$ con la restricción de que $2x + 2y = p$.

Veamos el ejemplo con $p = 1$. La representación de la función f es un paraboloida hiperbólico y podemos obtener en cada punto su gradiente ∇f , que será normal a la curva de nivel en dicho punto.

La restricción de igualdad $2x + 2y = 1$ puede representarse como una recta que hemos dibujado en negro. Podemos entender que esta recta es una curva de nivel del plano $g(x,y) = 2x + 2y - 1$ y entonces, el vector normal a la recta en cada punto viene dado por ∇g .

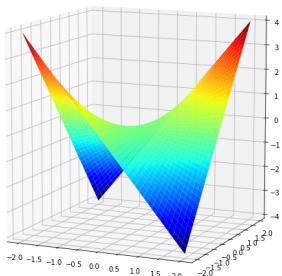
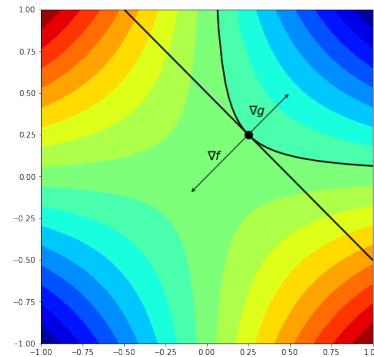
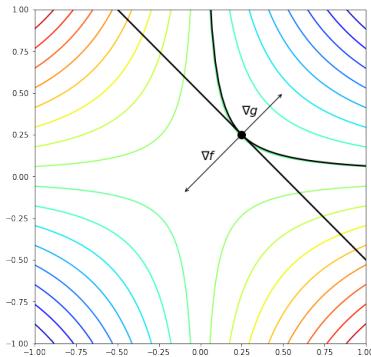
Estamos buscando el mínimo valor de la función f si tenemos en cuenta *solo* los puntos de la recta. Este vendrá dado por el punto donde una curva de nivel de f es tangente a la recta $2x + 2y - 1 = 0$ o, de otra manera, el punto donde las normales a las dos curvas son paralelos. Es decir, buscamos los puntos donde, para un λ distinto de cero

$$\nabla f = -\lambda \nabla g \implies \nabla f + \lambda \nabla g = \mathbf{0} \implies \nabla(f + \lambda g) = \mathbf{0}$$

La función

$$L(x, y, \lambda) = f(x, y) + \lambda g(x, y)$$

se llama *función lagrangiana* y la **condición necesaria** de mínimo con restricción de igualdad es que $\nabla L = 0$



EJERCICIO

1. Analíticamente: con multiplicadores de Lagrange

$$L(x, y, \lambda) = f(x, y) + \lambda g(x, y) \quad L(x, y, \lambda) = -xy + \lambda(2x + 2y - p)$$

y al calcular el gradiente de L e igualarlo a cero

$$\begin{aligned} L'_x &= -y + 2\lambda &= 0 \\ L'_y &= -x + 2\lambda &= 0 \\ L'_{\lambda} &= 2x + 2y - p &= 0 \end{aligned}$$

$$x = \frac{p}{4} \quad y = \frac{p}{4} \quad \lambda = \frac{p}{8}$$

Y el punto $x = \frac{p}{4}$, $y = \frac{p}{4}$ ($x = 0.25$, $y = 0.25$ para $p = 1$) cumple las **condiciones necesarias** de extremo.

Veamos ahora las **condiciones suficientes**. Para investigar si este punto es un máximo o un mínimo utilizaremos utilizando el determinante de la matriz hessiana de la Lagrangiana $L(x, y, \lambda)$ calculado en el punto crítico.

- Si es negativo, el punto es un mínimo.
- Si es positivo, el punto es un máximo.

La matriz hessiana de la función lagrangiana es

$$H_F = \begin{pmatrix} L''_{xx} & L''_{xy} & L''_{x\lambda} \\ L''_{yx} & L''_{yy} & L''_{y\lambda} \\ L''_{\lambda x} & L''_{\lambda y} & L''_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} 0 & -1 & 2 \\ -1 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix}.$$

Como $\det(H) = -8 < 0$ en el punto el punto $x = \frac{p}{4}$, $y = \frac{p}{4}$ hay un mínimo de la función $f(x, y) = -xy$, que se corresponde con un máximo de $-f(x, y) = xy$, que es el área del rectángulo. Por lo tanto el rectángulo de área máxima, fijado el perímetro, es un cuadrado puesto que los lados son iguales.

2. Numéricamente: función de penalización

Debemos fijar p para poder resolver el problema numéricamente. Elegimos $p = 1$.

La idea del método de penalización es reemplazar la función objetivo f por otra función

$$F(x, y) = f(x, y) + c P(x, y)$$

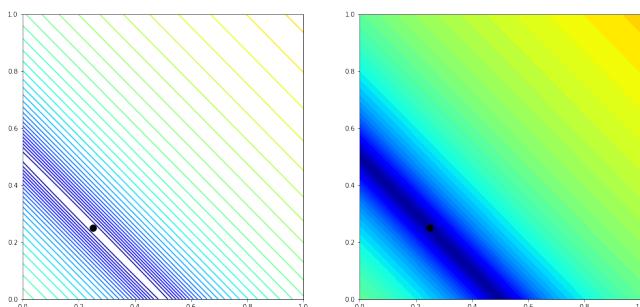
y resolver el problema sin restricciones. Para ello tomamos c como una constante positiva y P , la función de penalización, satisfaciendo:

- P es continua en el dominio de f .
- $P(x, y) \geq 0$ para todo punto del dominio 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

$$F(x, y) = -xy + 100(2x + 2y - 1)^2$$

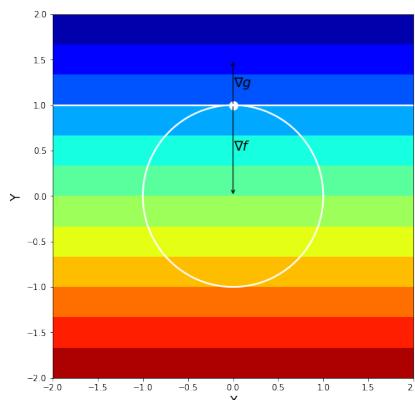
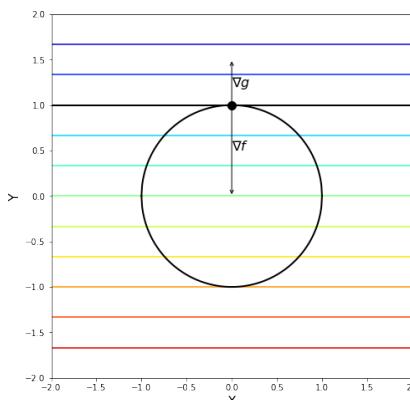
Ahora, el mínimo absoluto de esta función está cerca del mínimo con restricciones de f (tanto más cerca cuanto mayor sea c), como se ve en la gráfica siguiente, y podemos aproximarla con un método numérico, como Newton o el método del descenso del gradiente. Es una solución aproximada, como en general lo son las soluciones numéricas.



Ejercicio 6.3.2

Minimizar la función f sujeta a la condición c siendo

$$f(x, y) = -y \quad c : x^2 + y^2 = 1$$

1. Analíticamente: con multiplicadores de Lagrange

Construímos la función Lagrangiana

$$L(x, y, \lambda) = f(x, y) + \lambda g(x, y)$$

que en este caso es

$$L(x, y, \lambda) = -y + \lambda (x^2 + y^2 - 1)$$

y al calcular el gradiente de L e igualarlo a cero

$$\begin{aligned} L'_x &= 2x\lambda &= 0 \\ L'_y &= -1 + 2y\lambda &= 0 \\ L'_{\lambda} &= x^2 + y^2 - 1 &= 0 \end{aligned}$$

Como $\lambda \neq 0$ de la primera ecuación obtenemos $x = 0$. Sustituyendo este valor de x en la tercera ecuación tenemos que $y = \pm 1$ y sustituyendo estos valores en la segunda ecuación obtenemos los valores correspondientes de λ . Tenemos dos soluciones

$$x = 0 \quad y = -1 \quad \lambda = -\frac{1}{2} \quad \text{y} \quad x = 0 \quad y = 1 \quad \lambda = \frac{1}{2}$$

Para investigar si este punto es un máximo o un mínimo utilizaremos utilizando el determinante de la matriz hessiana del Lagrangiano $L(x, y, \lambda)$ calculado en el punto crítico.

- Si es negativo, el punto es un mínimo.
- Si es positivo, el punto es un máximo.

La matriz hessiana del lagrangiano es

$$H_L = \begin{pmatrix} L''_{xx} & L''_{xy} & L''_{x\lambda} \\ L''_{yx} & L''_{yy} & L''_{y\lambda} \\ L''_{\lambda x} & L''_{\lambda y} & L''_{\lambda\lambda} \end{pmatrix} \quad \text{que es} \quad H_L = \begin{pmatrix} 2\lambda & 0 & 2x \\ 0 & 2\lambda & 2y \\ 2x & 2y & 0 \end{pmatrix}.$$

Para

$$x = 0 \quad y = -1 \quad \lambda = -\frac{1}{2} \quad H_L = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & -2 \\ 0 & -2 & 0 \end{pmatrix} \quad \det(H_L) = 4 > 0$$

y es un máximo.

Para

$$x = 0 \quad y = 1 \quad \lambda = \frac{1}{2} \quad H_L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 0 \end{pmatrix} \quad \det(H_L) = -4 < 0$$

y es un mínimo.

2. Numéricamente: función de penalización

La idea del método de penalización es reemplazar la función objetivo f por otra función

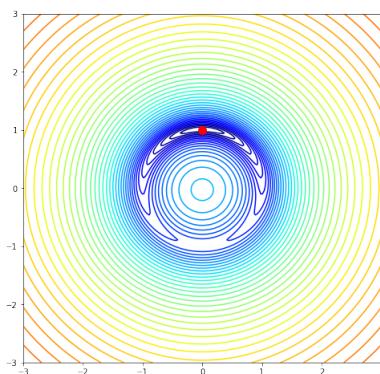
$$F(x, y) = f(x, y) + c P(x, y)$$

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 dominio 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

$$F(x, y) = -y + 10(x^2 + y^2 - 1)^2$$



6.4 Mínimo local con restricciones de desigualdad. Cálculo simbólico y numérico

Ejercicio 6.4.1

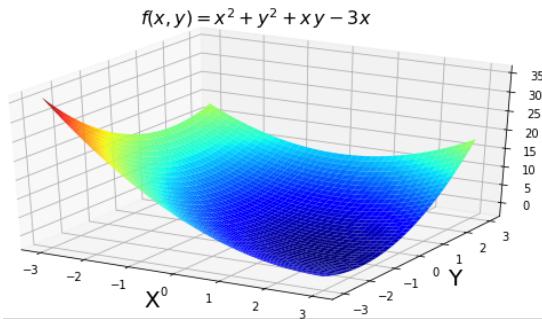
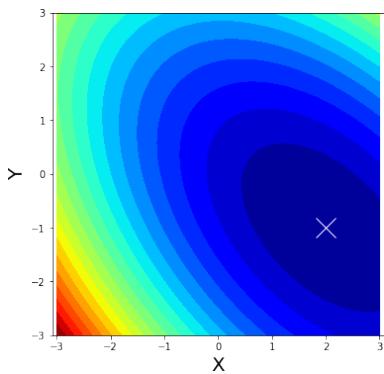
Sea

$$f(x, y) = x^2 + y^2 + xy - 3x$$

1. Hallar un mínimo local de f .
2. Probar que dicho mínimo es, de hecho, global.
3. Hallar el mínimo sujeto a las restricciones $x \geq 0$ e $y \geq 0$

1. Hallar un mínimo local

La función a minimizar se puede representar con curvas de nivel o como una superficie



$$f(x, y) = x^2 + y^2 + xy - 3x$$

La condición necesaria de mínimo para un punto (x_m, y_m) es que $\nabla f(x_m, y_m) = (f'_x, f'_y) = (0, 0)$. Es decir

$$\begin{cases} f'_x = 0 \\ f'_y = 0 \end{cases} \quad \begin{cases} 2x + y - 3 = 0 \\ x + 2y = 0 \end{cases} \quad \begin{cases} x = 2 \\ y = -1 \end{cases}$$

Y el punto $(x_m, y_m) = (2, -1)$ cumple las condiciones necesarias de mínimo.

La condición suficiente 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 $(x_m, y_m) = (2, -1)$, como todos los elementos son constantes, el valor es el mismo.

Como la condición suficiente de mínimo es que la matriz hessiana sea definida positiva, los determinantes de los menores principales han de ser estrictamente positivos. Es decir

$$\begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} = 3 > 0 \quad \text{y} \quad |2| = 2 > 0.$$

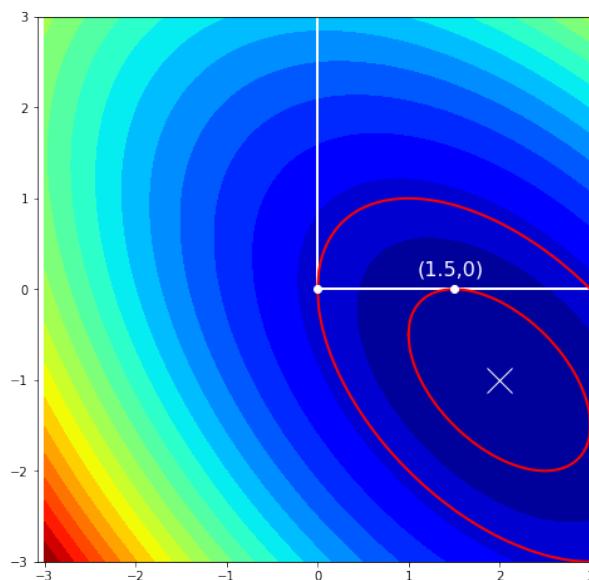
Por lo tanto la matriz es definida positiva y se cumple la condición suficiente de mínimo.

2. Probar que dicho mínimo es, de hecho, global

Para demostrar que este *mínimo es único*, basta con demostrar que la función es convexa.

Si el hessiano es definido positivo para todos los puntos de la función, esta es convexa. Como la matriz hessiana tiene elementos constantes y es definida positiva, lo es para toda la función y caso de existir un mínimo, es único.

3. Hallar el mínimo sujeto a restricciones de desigualdad



Analíticamente

Si consideramos las restricciones estamos buscando un mínimo en el primer cuadrante del plano. Para ello:

1. Buscamos un mínimo dentro del recinto.
2. Buscamos un mínimo en las fronteras.
3. Añadimos los puntos de intersección de las fronteras.
4. Consideraremos el valor de la función en todos los puntos anteriores, y el punto para el cual la función tenga el menor valor, ese es el mínimo.

1. Buscamos un mínimo dentro del recinto

Ya buscamos mínimos relativos en el apartado anterior, y el único que había era

$$(2, -1)$$

2. Buscamos un mínimo en las fronteras.

- (a) Función $f(x, y) = x^2 + y^2 + xy - 3x$ y restricción $x = 0$. Entonces la función Lagrangiana es $L(x, y, \lambda) = f(x, y) + \lambda g(x, y)$

$$L(x, y, \lambda) = x^2 + y^2 + xy - 3x + \lambda x$$

Calculamos las derivadas parciales y las igualamos a cero

$$\begin{aligned} L'_x &= 2x + y - 3 + \lambda = 0 \\ L'_y &= 2y + x = 0 \\ L'_{\lambda} &= x = 0 \end{aligned}$$

De la tercera ecuación $x = 0$. Teniendo en cuenta este valor y la segunda ecuación $x = 0$. Por lo tanto, posible mínimo el punto

$$(0, 0)$$

- (b) Función $f(x, y) = x^2 + y^2 + xy - 3x$ y restricción $y = 0$. Entonces la función Lagrangiana es

$$L(x, y, \lambda) = x^2 + y^2 + xy - 3x + \lambda y$$

Calculamos las derivadas parciales y las igualamos a cero

$$\begin{aligned} L'_x &= 2x + y - 3 = 0 \\ L'_y &= 2y + x + \lambda = 0 \\ L'_{\lambda} &= y = 0 \end{aligned}$$

De la tercera ecuación $y = 0$. Teniendo en cuenta este valor y la primera ecuación $x = 1.5$ Por lo tanto, posible mínimo el punto

$$(1.5, 0)$$

3. Añadimos los puntos de intersección de las fronteras.

La intersección de las dos fronteras es

$$(0,0)$$

4. Consideramos el valor de la función en todos los puntos anteriores, y el punto para el cual la función tenga el menor valor, ese es el mínimo.

| Punto | En dominio? | $f(x,y)$ | Mínimo |
|----------|-------------|----------|--------|
| (2, -1) | No | | |
| (0, 0) | Si | 0 | |
| (1.5, 0) | Si | -2.25 | Si |

Por lo tanto el mínimo está en (1.5, 0)

2. Numéricamente: función de penalización

La idea del método de penalización es reemplazar la función objetivo f por otra función

$$F(x,y) = f(x,y) + c_1 g_1(x,y) + c_2 g_2(x,y)$$

y resolver el problema sin restricciones. Para ello tomamos c_i como una constante positiva y g_i satisfaciendo:

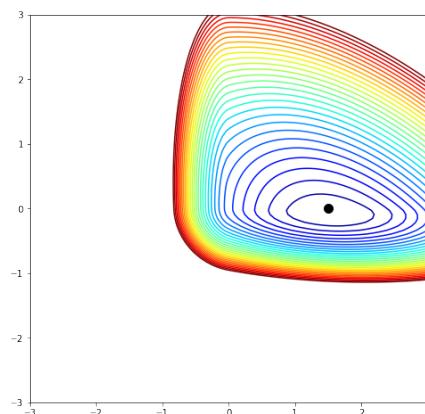
- g_i es continua en el dominio de f .
- $g_i(x,y) \geq 0$ para todo punto del dominio de f , y
- $g_i(x,y) = 0$ si y solo si el punto (x,y) satisface las constricciones.

Una posible función para aproximar el mínimo con restricciones sería

$$F(x,y) = f(x,y) + 10g_1(x,y) + 10g_2(x,y)$$

donde

$$g_1(x,y) = \begin{cases} 0 & \text{si } x \geq 0 \\ x^2 & \text{si } x < 0 \end{cases} \quad g_2(x,y) = \begin{cases} 0 & \text{si } y \geq 0 \\ y^2 & \text{si } y < 0 \end{cases}$$



Ejercicio 6.4.2

Maximizar la función

$$f(x, y) = 14x - x^2 + 6y - y^2 + 7$$

sujeta a las restricciones $x + y \leq 2$ y $x + 2y \leq 3$.

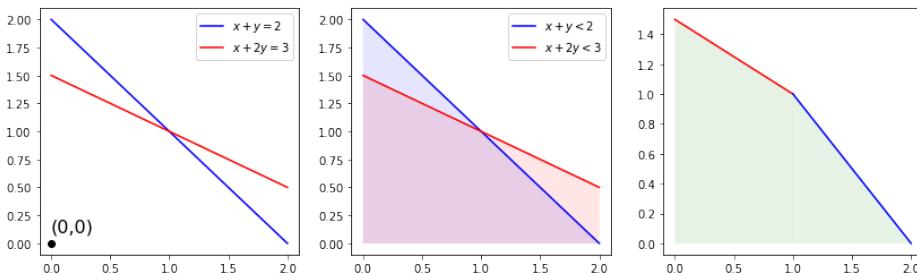
1. Analíticamente**Zona definida por las restricciones**

En lugar de buscar el máximo de f buscaremos el mínimo de $-f$. Para simplificar, le llamaremos f a partir de ahora

$$f(x, y) = x^2 + y^2 - 14x - 6y - 7$$

Los límites de la zona donde buscamos el mínimo son

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



Cada recta divide el plano en tres zonas. Por ejemplo, $x + y = 2$ divide el plano en las zonas donde

- $x + y > 2$ es un semiplano.
- $x + y = 2$ es la recta.
- $x + y < 2$ es el otro semiplano

Para ver cual de las dos zonas es la que contiene, por ejemplo, al punto $(0, 0)$ basta probar con este punto y la desigualdad $x + y \leq 2$

$$0 + 0 \leq 2$$

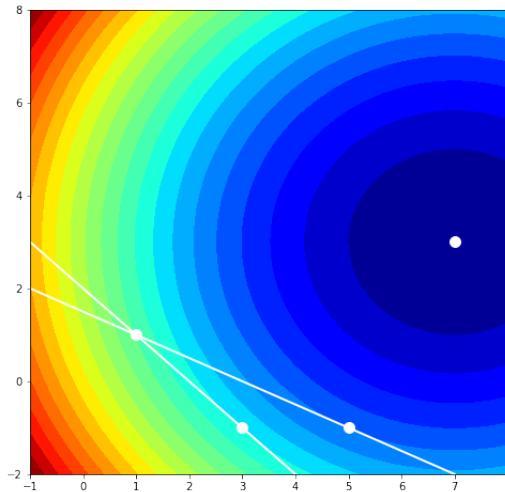
Por lo tanto $x + y \leq 2$ es la zona que comprende la recta y el semiplano por debajo de la recta azul.

Análogamente, probamos con $(0, 0)$ y $x + 2y \leq 3$ y

$$0 + 2(0) \leq 3$$

y es el semiplano por debajo de la recta roja.

Y la zona que cumple las dos condiciones es la intersección, que está dibujada en verde.



Por lo tanto, considerando las restricciones, estamos buscando un mínimo en la zona por debajo de las semirrectas roja y azul que está pintada de verde. Para ello:

1. Buscamos un mínimo dentro del recinto.
2. Buscamos un mínimo en las fronteras.
3. Añadimos los puntos de intersección de las fronteras.
4. Consideraremos el valor de la función en todos los puntos anteriores, y el punto para el cual la función tenga el menor valor, ese es el mínimo.

1. Buscamos un mínimo dentro del recinto

La condición necesaria de mínimo para un punto (x_m, y_m) es que $\nabla f(x_m, y_m) = (f'_x, f'_y) = (0, 0)$. Es decir para

$$f(x, y) = x^2 + y^2 - 14x - 6y - 7$$

se tiene que

$$\begin{cases} f'_x = 0 \\ f'_y = 0 \end{cases} \quad \begin{cases} 2x - 14 = 0 \\ 2y - 6 = 0 \end{cases} \quad \begin{cases} x = 7 \\ y = 3 \end{cases}$$

Y ya tenemos un punto a considerar

$$(7, 3)$$

2. Buscamos un mínimo en las fronteras

1. Función $f(x, y) = f(x, y) = x^2 + y^2 - 14x - 6y - 7$ y restricción $x + y = 2$. Entonces la función Lagrangiana es $L(x, y, \lambda) = f(x, y) + \lambda g(x, y)$

$$L(x, y, \lambda) = x^2 + y^2 - 14x - 6y - 7 + \lambda(x + y - 2)$$

Calculamos las derivadas parciales y las igualamos a cero

$$\begin{aligned} L'_x &= 2x - 14 + \lambda = 0 \\ L'_y &= 2y - 6 + \lambda = 0 \\ L'_{\lambda} &= x + y - 2 = 0 \end{aligned}$$

De la primera ecuación $x = \frac{14 - \lambda}{2}$, de la segunda ecuación $y = \frac{6 - \lambda}{2}$. Metiendo estos valores en la tercera ecuación

$$\frac{14 - \lambda}{2} + \frac{6 - \lambda}{2} - 2 = 0 \quad \rightarrow \quad 14 - \lambda + 6 - \lambda = 4 \quad \rightarrow \quad \lambda = 8$$

Y sustituyendo este valor de λ en x e y un posible es mínimo el punto

$$(3, -1)$$

2. Función $f(x, y) = f(x, y) = x^2 + y^2 - 14x - 6y - 7$ y restricción $x + 2y = 3$. Entonces la función Lagrangiana es

$$L(x, y, \lambda) = x^2 + y^2 - 14x - 6y - 7 + \lambda(x + 2y - 3)$$

Calculamos las derivadas parciales y las igualamos a cero

$$\begin{aligned} L'_x &= 2x - 14 + \lambda = 0 \\ L'_y &= 2y - 6 + 2\lambda = 0 \\ L'_{\lambda} &= x + 2y - 3 = 0 \end{aligned}$$

De la primera ecuación $x = \frac{14 - \lambda}{2}$, de la segunda ecuación $y = \frac{6 - 2\lambda}{2}$. Metiendo estos valores en la tercera ecuación

$$\frac{14 - \lambda}{2} + 2 \frac{6 - 2\lambda}{2} - 3 = 0 \quad \rightarrow \quad 14 - \lambda + 12 - 4\lambda = 6 \quad \rightarrow \quad \lambda = 4$$

Y sustituyendo este valor de λ en x e y un posible es mínimo el punto

$$(5, -1)$$

3. Añadimos los puntos de intersección de las fronteras

La intersección de las dos fronteras $x + y = 2$ y $x + 2y = 3$ es

$$\boxed{(1,1)}$$

4. Consideramos el valor de la función en todos los puntos anteriores, y el punto para el cual la función tenga el menor valor, ese es el mínimo.

Tenemos por lo tanto 4 puntos candidatos que colocamos en la primera columna. Lo primero es comprobar si cumplen las restricciones:

- $(7, 3)$ ¿Cumple $x + y \leq 2$? Como $7 + 3 > 2$, no pertenece al dominio.
- $(3, -1)$ ¿Cumple $x + y \leq 2$? Si, $3 + (-1) \leq 2$ ¿Cumple $x + 2y \leq 3$? Si, $3 + 2(-1) \leq 3$
- $(5, -1)$ ¿Cumple $x + y \leq 2$? Como $5 + (-1) > 2$, no pertenece al dominio.
- $(1, 1)$ ¿Cumple $x + y \leq 2$? Si, $1 + 1 \leq 2$ ¿Cumple $x + 2y \leq 3$? Si, $1 + 2(1) \leq 3$

Han pasado el primer filtro dos puntos. Ahora, evaluamos la función en estos dos puntos y de los posibles, nos quedamos con el punto en el que la función tiene menor valor.

| Punto | En dominio? | $f(x, y)$ | Mínimo |
|-----------|-------------|-----------|--------|
| $(7, 3)$ | No | | |
| $(3, -1)$ | Si | -33 | |
| $(5, -1)$ | No | | |
| $(1, 1)$ | Si | -25 | |

El menor valor es -33 y por lo tanto el mínimo está en $\boxed{(3, -1)}$

2. Numéricamente: función de penalización

La idea del método de penalización es reemplazar la función objetivo f por otra función

$$F(x, y) = f(x, y) + c_1 g_1(x, y) + c_2 g_2(x, y)$$

y resolver el problema sin restricciones. Para ello tomamos c_i como una constante positiva y g_i satisfaciendo:

- g_i es continua en el dominio de f .
- $g_i(x, y) \geq 0$ para todo punto del dominio de f , y
- $g_i(x, y) = 0$ si y solo si el punto (x, y) satisface las restricciones.

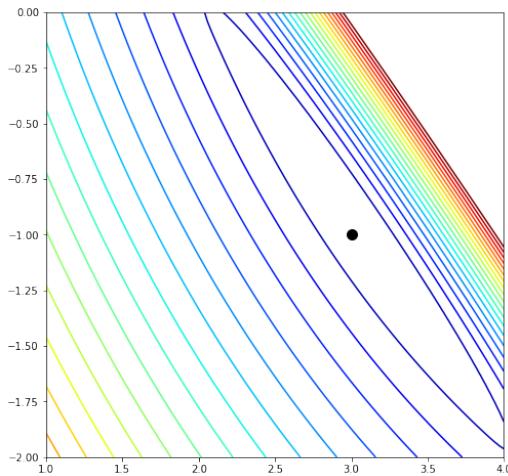
Una posible función para aproximar el mínimo con restricciones sería

$$F(x, y) = f(x, y) + 50g_1(x, y) + 50g_2(x, y)$$

donde

$$g_1(x, y) = \begin{cases} 0 & \text{si } x + y \leq 2 \\ (x + y - 2)^2 & \text{si } x + y > 2 \end{cases}$$

$$g_2(x, y) = \begin{cases} 0 & \text{si } x + 2y \leq 3 \\ (x + 2y - 3)^2 & \text{si } x + 2y > 3 \end{cases}$$



6.5 Extremo de función lineal con restricciones lineales

Ejercicio 6.5.1

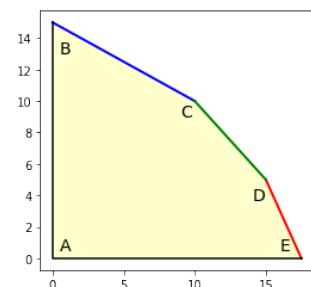
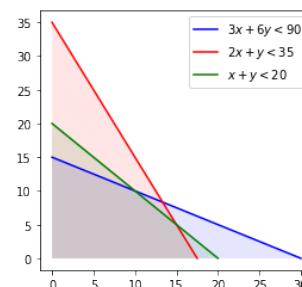
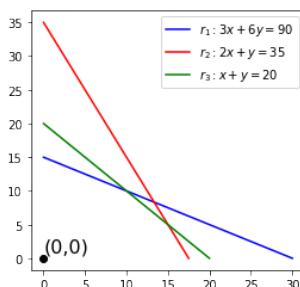
Maximizar la función $f(x, y) = 2.5x + 3y$ con las restricciones

$$\begin{cases} 3x + 6y \leq 90, \\ 2x + y \leq 35, \\ x + y \leq 20, \\ x, y \geq 0. \end{cases}$$

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 tres rectas, r_1 , r_2 y r_3 .

$$\begin{array}{lll} r_1 \quad 3x + 6y = 90 & \rightarrow & \frac{3x}{90} + \frac{6y}{90} = 1 & \rightarrow & \frac{x}{30} + \frac{y}{15} = 1 \\ r_2 \quad 2x + y = 35 & \rightarrow & \frac{2x}{35} + \frac{y}{35} = 1 & \rightarrow & \frac{x}{17.5} + \frac{y}{35} = 1 \\ r_3 \quad x + y = 20 & \rightarrow & \frac{x}{20} + \frac{y}{20} = 1 & & \end{array}$$

y entonces r_1 corta al eje OX en $x = 30$ y al eje OY en $y = 15$. Análogamente r_2 corta al eje OX en $x = 17.5$ y al eje OY en $y = 35$. Y r_3 corta al eje OX en $x = 20$ y al eje OY en $y = 20$.



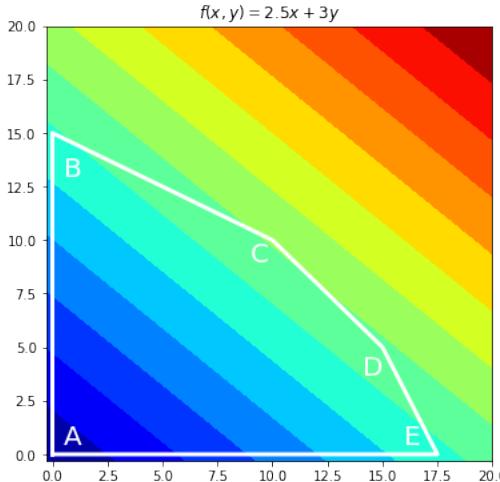
Cada una de estas rectas divide el plano en dos regiones, una que verifica la condición y otra que no. Basta por lo tanto con probar con un punto y si este punto cumple la condición, la región del plano donde se encuentra este punto es pertenece a la región.

Por simplicidad, probamos con el origen $(0, 0)$

- El origen cumple la condición $3x + 6y \leq 90$ porque $3(0) + 6(0) \leq 90$.
Pero no cumple la condición $3x + 6y \geq 90$ porque $3(0) + 6(0) \geq 90$.
- El origen también cumple las condiciones $2x + y \leq 35$ y $x + y \leq 20$.

Así que la región es la intersección de las tres regiones bajo las rectas y es la región representada en amarillo, que recordamos, está en el primer cuadrante porque además $x \geq 0$ y $y \geq 0$.

En cuanto a la función, como es lineal, para funciones de dos variables, su representación como superficie es un plano y sus isolíneas o curvas de nivel son rectas, como se puede ver en el siguiente gráfico.



El máximo estará en uno de los vértices (o de los lados) del polígono que rodea al área. Así que necesitamos calcular los vértices.

- A es el origen y $A = (0, 0)$.
- B es el corte de la recta r_1 con el eje Y . Por lo tanto $B = (0, 15)$
- C es la intersección de $3x + 6y = 90$ con $x + y = 20$. Si $x = 20 - y$ entonces $3(20 - y) + 6y = 90$ y entonces $60 + 3y = 90$ y $C = (10, 10)$.
- D es la intersección de $x + y = 20$ con $2x + y = 35$. Si $x = 20 - y$ entonces $2(20 - y) + y = 35$ y entonces $40 - y = 35$ y $D = (15, 5)$.
- E es el corte de la recta r_2 con el eje X . Por lo tanto $E = (17.5, 0)$.

Calculamos el valor de la función $f(x, y) = 2.5x + 3y$ en estos puntos

| Vértice | x | y | $f(x, y)$ |
|---------|------|-----|-----------|
| A | 0 | 0 | 0 |
| B | 0 | 15 | 45 |
| C | 10 | 10 | 55 |
| D | 15 | 5 | 52.5 |
| E | 17.5 | 0 | 43.75 |

Y el máximo está en el punto C , porque la función toma el máximo valor.

Ejercicio 6.5.2

Dos fábricas de cemento, F_1 y F_2 , producen respectivamente 3000 y 4000 sacos de cemento al día. Hay que enviar ese cemento a tres centros de ventas C_1 , C_2 y C_3 en cantidades de 3000, 2500 y 1500 sacos, respectivamente. Los costes de transporte de cada fábrica a los puntos de venta vienen dados, en euros por cada saco,

| Costes por unidad | Hasta C_1 | Hasta C_2 | Hasta C_3 |
|-------------------|-------------|-------------|-------------|
| Desde F_1 | 2 | 2.5 | 2 |
| Desde F_2 | 1.5 | 4 | 1 |

Determinar cómo hay que distribuir la producción para que el transporte resulte lo más económico posible.

| Unidades | Hasta C_1 | Hasta C_2 | Hasta C_3 | Σ |
|-------------|-------------|-------------|----------------|----------|
| Desde F_1 | x | y | $3000 - x - y$ | 3000 |
| Desde F_2 | $3000 - x$ | $2500 - y$ | $x + y - 1500$ | 4000 |
| Σ | 3000 | 2500 | 1500 | 7000 |

Asumimos x unidades a trasportar desde F_1 hasta C_1 e y unidades a trasportar desde F_1 hasta C_2 . Todas las demás las deducimos teniendo en cuenta x e y y las unidades totales a enviar desde las fábricas y a recibir en los centros de venta.

Coste

Obtenemos el coste para cada uno de los 6 itinerarios multiplicando el número de unidades (tabla 2) por el coste por unidad (tabla 1) para ese caso. Sumando los 6 casos distintos tenemos el coste total.

$$C = 2x + 2.5y + 2(3000 - x - y) + 1.5(3000 - x) + 4(2500 - y) + 1(x + y - 1500)$$

Que simplificando, queda

$$C = 19000 - 0.5x - 2.5y$$

Restricciones

Las restricciones son que el número de unidades transportadas de cada fábrica a cada centro han de ser cantidades positivas. Como hay 6 canales de distribución tenemos 6 restricciones

$$\begin{aligned}x &\geq 0 \\y &\geq 0 \\3000 - x - y &\geq 0 \\3000 - x &\geq 0 \\2500 - y &\geq 0 \\x + y - 1500 &\geq 0\end{aligned}$$

Problema

Minimizar

$$C = 19000 - 0.5x - 2.5y$$

con las restricciones (reordenamos las condiciones anteriores)

$$\begin{aligned}x + y &\leq 3000 \\x &\leq 3000 \\y &\leq 2500 \\x + y &\geq 1500 \\x, y &\geq 0\end{aligned}$$

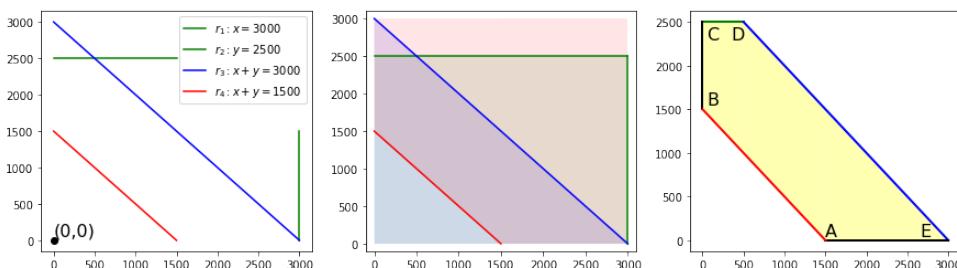
Las condiciones $x \geq 0$ e $y \geq 0$ significan que la región del plano que describen las condiciones está en el primer cuadrante.

La condición $x \leq 3000$ significa que x ha de estar a la izquierda de la recta vertical r_1 que pasa por 3000 y $y \leq 2500$ significa que y ha de estar por debajo de la recta horizontal r_2 que pasa por 2500

Dibujamos las demás fronteras de la región donde ha de encontrarse la solución. Tenemos dos rectas r_3 y r_4

$$\begin{array}{lll}r_3 & x + y = 3000 & \rightarrow \quad \frac{x}{3000} + \frac{y}{3000} = 1 \\r_4 & x + y = 1500 & \rightarrow \quad \frac{x}{1500} + \frac{y}{1500} = 1\end{array}$$

y entonces r_3 corta al eje OX en $x = 3000$ y al eje OY en $y = 3000$. Análogamente r_4 corta al eje OX en $x = 1500$ y al eje OY en $y = 1500$.



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.

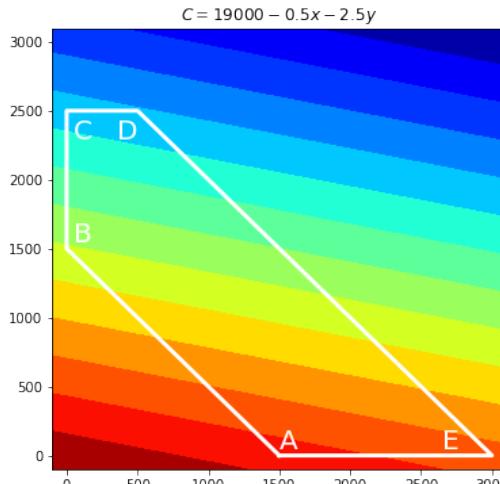
Por simplicidad, probamos con el origen $(0, 0)$

- El origen cumple la condición $x + y \leq 3000$ porque $0 + 0 \leq 3000$. Por lo tanto, la zona que cumple esta condición es la que está por debajo de la recta azul.
- El origen no cumple las condiciones $x + y > 1500$, por lo tanto, la región es la que está por encima de la recta r_4 (roja).

Además, habíamos dicho que

- La condición $x \leq 3000$ significa que x ha de estar a la izquierda de la recta vertical r_1 que pasa por 3000
- $y \leq 2500$ significa que y ha de estar por debajo de la recta horizontal r_2 que pasa por 2500

Así que la región es la intersección de las tres regiones (rosa, azul y verde) y es la región representada en amarillo, que recordamos, está en el primer cuadrante porque además $x \geq 0$ y $y \geq 0$.



El máximo estará en uno de los vértices (o de los lados) del polígono que rodea al área. Así que necesitamos calcular los vértices.

- A la intersección de $x + y = 1500$ con el eje OX , es decir, con $y = 0$. Por lo tanto $A = (1500, 0)$.
- B es el corte de la recta $y = 2500$ con el eje OY , que es $x = 0$. Por lo tanto $B = (0, 2500)$
- C es la intersección de $y = 2500$ con $x = 0$. $C = (0, 2500)$.

- D es la intersección de $x + y = 3000$ con $y = 2500$. Como $x = 3000 - 2500$, se tiene $D = (500, 2500)$.
- E es el corte de la recta $x + y = 3000$ con $y = 0$ (eje OX), y $x = 3000$. Por lo tanto $E = (3000, 0)$.

Calculamos en estos vértices el valor de la función $C = 19000 - 0.5x - 2.5y$

| Vértice | x | y | C |
|---------|------|------|-------|
| A | 1500 | 0 | 18250 |
| B | 0 | 1500 | 16750 |
| C | 0 | 2500 | 15250 |
| D | 500 | 2500 | 15000 |
| E | 3000 | 0 | 17500 |

Y el máximo está en el punto D donde $x = 500$ e $y = 2500$, porque la función toma el máximo valor. Así que la solución del problema vendría sustituyendo en la tabla

| Unidades | Hasta C_1 | Hasta C_2 | Hasta C_3 | Σ |
|-------------|-------------|-------------|----------------|----------|
| Desde F_1 | x | y | $3000 - x - y$ | 3000 |
| Desde F_2 | $3000 - x$ | $2500 - y$ | $x + y - 1500$ | 4000 |
| Σ | 3000 | 2500 | 1500 | 7000 |

los valores $x = 500$ e $y = 2500$

| Unidades | Hasta C_1 | Hasta C_2 | Hasta C_3 | Σ |
|-------------|--------------|---------------|---------------------|----------|
| Desde F_1 | 500 | 2500 | $3000 - 500 - 2500$ | 3000 |
| Desde F_2 | $3000 - 500$ | $2500 - 2500$ | $500 + 2500 - 1500$ | 4000 |
| Σ | 3000 | 2500 | 1500 | 7000 |

Por lo tanto, la solución es

| Unidades | Hasta C_1 | Hasta C_2 | Hasta C_3 | Σ |
|-------------|-------------|-------------|-------------|----------|
| Desde F_1 | 500 | 2500 | 0 | 3000 |
| Desde F_2 | 2500 | 0 | 1500 | 4000 |
| Σ | 3000 | 2500 | 1500 | 7000 |

Como se puede ver, estamos evitando transportes a precios altos para llevar cemento a C_2 y C_3 y lo compensamos distribuyendo el transporte hasta C_1 donde los precios son más parecidos

| Costes por unidad | Hasta C_1 | Hasta C_2 | Hasta C_3 |
|-------------------|-------------|-------------|-------------|
| Desde F_1 | 2 | 2.5 | 2 |
| Desde F_2 | 1.5 | 4 | 1 |

