Computación Numérica

Primer Parcial A - Mayo 2015

1. Representamos un entero utilizando 7 bits:

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

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

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

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

y el máximo entero no negativo es 127.

b) Los enteros con signo que podemos representar con m bits serían los mismos que en el caso anterior menos el $sesgo = 2^{m-1} = 2^6 = 64$, es decir

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

y por lo tanto el valor máximo es 63.

2. Una máquina almacena números en punto flotante en 11 bits. El primer bit se usa para el signo del número, los cinco siguientes para el exponente sesgado y los últimos cinco bits para la mantisa. Si se sigue un criterio similar al de la norma IEEE 754:

   a) ¿Cual sería el ϵ de máquina?

   b) ¿Cual sería el menor real desnormalizado?

   c) Almacenar el número $(11{,}4)_{10}$ en este formato utilizando el redondeo al par más cercano.

a) El número de dígitos de la mantisa es 5. El número 1 se representa

$$1 \;=\; 1{,}00000 \times 2^0$$

y el siguiente número representable es

$$1+\epsilon \;=\; 1{,}00001 \times 2^0$$

Por lo que

$$\epsilon \;=\; 0{,}00001 \times 2^0 \;=\; 2^{-5}$$

Y en base 10, el número es

$$\epsilon = 0{,}03125.$$

b) Como seguimos un criterio similar a la norma IEEE 754

$$sesgo = 2^{m-1} - 1 = 2^{5-1} - 1 = 2^4 - 1 = 16 - 1 = 15.$$

el exponente mínimo es $1 - sesgo = 1 - 15 = -14$. Si el número es desnormalizado el exponente es $e = 00000$ pero se le atribuye el valor del menor exponente posible, $e = 00001$, es decir, -14. Además, el bit escondido es cero. La representación binaria del número más pequeño desnormalizado es

signo	exponente	mantisa
0	00000	00001

Que representa

$$0{,}00001 \times 2^{-14} = 2^{-5} \times 2^{-14} = 2^{-19} \approx 1{,}90735 \times 10^{-6}$$

c) Convertimos a binario la parte entera tomando los restos de dividir por dos y el último cociente, empezando por este:

Cociente	11	5	2	1
Resto	1	1	0	↙

$$(11)_{10} = (1011)_2$$

Convertimos a binario la parte fraccionaria multiplicando por dos y quedándonos con la parte entera:

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

Y tenemos que

$$(11{,}4)_{10} = (1011{,}01100110011\ldots)_2$$

es decir,

$$(11{,}4)_{10} = 1.\overbrace{01101}\,1001100\ldots \times 2^3$$

que redondeamos al para más cercano sumando $0{,}00001 \times 2^3$ a $1.\overbrace{01101} \times 2^3$. La aproximación es

$$(0{,}2)_{10} \approx 1{,}01110 \times 2^3$$

El sesgo es 15, y el exponente sesgado es $e = 3 + sesgo = 3 + 15 = 18$. Pasándolo a binario

Cocientes	18	9	4	2	1
Restos	0	1	0	0	↙

Y el exponente es, $(e)_2 = 10010$. Y teniendo en cuenta el bit escondido

signo	exponente	mantisa
0	10010	01110

3. Demostrar que en la representación binaria de precisión simple de la norma IEEE 754 el número de dígitos decimales significativos es aproximadamente 7.

Podemos escribir cualquier número binario, con $b_0 = 1$,

$$x = \pm\Big(1.b_1b_2\ldots b_{23}b_{24}b_{25}\ldots\Big)\times 2^e.$$

Si lo redondeamos hacia cero,

$$x^* = \pm\Big(1.b_1b_2\ldots b_{23}\Big)\times 2^e,$$

y el error relativo

$$\frac{|x-x^*|}{|x|} = \frac{\Big(0.\overbrace{00\ldots0}\,b_{24}b_{25}\ldots\Big)\times 2^e}{\Big(1.b_1b_2\ldots b_{23}b_{24}b_{25}\ldots\Big)\times 2^e} \leq \frac{0.\overbrace{00\ldots0}\,b_{24}b_{25}\ldots}{1.b_1b_2\ldots b_{23}b_{24}b_{25}\ldots} \leq$$

$$\leq \frac{0.\overbrace{00\ldots0}\,b_{24}b_{25}\ldots}{1} \leq \frac{0.\overbrace{00\ldots0}\,11\ldots}{1} \leq 0.\overbrace{00\ldots1} = 2^{-23} \approx 1{,}1921\times 10^{-7}$$

Computación Numérica

Primer Parcial B - Mayo 2015

1. Sea la ecuación

$$te^{-\frac{t}{5}} - 1 = 0$$

a) Demostrar que en $[1, 2]$ existe una única raíz.

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

c) Realizar tres iteraciones con el método de bisección.

d) Dar una cota del error cometido al calcular la raíz con tres iteraciones.

e) Demostrar que la ecuación $te^{-\frac{t}{5}} - 1 = 0$ tiene la misma raíz que $g_i(t) = t$ con $i = 1, 2$ siendo

$$g_1(t) = e^{\frac{t}{5}} \quad g_2(x) = 5 \ln t$$

f) Enunciar las 2 condiciones del Teorema de la Aplicación Contractiva.

g) Utilizándolo, escoger una de las dos funciones para aproximar la solución por el método de iteración de punto fijo comenzando en $x_0 = 1$.

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

- 1. f continua: f es continua porque es la suma, producto y composición de funciones continuas. El polinomio lo es siempre y la función exponencial también es continua en todo su dominio.
- 2. f tiene distinto signo en los extremos del intervalo: $f(1) = -0{,}18$ y $f(2) = 0{,}34$
- 3. $f' > 0$ o $f' < 0$ en $(1, 2)$:

$$f'(t) = e^{-\frac{t}{5}} - \frac{1}{5}te^{-\frac{t}{5}} = -\frac{1}{5}e^{-\frac{t}{5}}(t - 5).$$

Como $-\frac{1}{5}$ es negativo, $e^{-\frac{t}{5}}$ es positivo y $t - 5 < 0$ en ,$(1, 2)$ el producto es negativo. Por lo tanto $f' < 0$ en $(1, 2)$.

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

c)

k	a	c	b	$f(a)$	$f(c)$	$f(b)$	*cota de error*
0	1	(a+b)/2=(1+2)/2=1.5	2	-0.18	0.11	0.34	2-1=1
1	1	(a+b)/2=(1+1.5)/2=1.25	1.5	-0.18	-0.026	0.11	1.5-1=0.5
2	1.25	(a+b)/2=(0+0.25)/2=1.375	1.5	-0.026	0.44	0.11	1.5-1.25=0.25
3	1.25		1.375	-0.026		0.44	1.375-1.25=0.125

Y podemos dar como raíz aproximada 1,375 (la raíz exacta es 1,29586 $\in$ [1,25, 1,375]).

d) La raíz está en el intervalo [1,25, 1,375] por lo que el máximo error será $1{,}375-1{,}25 = 0{,}125$, que podemos dar como cota de error.

e) Podemos transformar la ecuación

$$f(t) = 0 \Longleftrightarrow te^{-\frac{t}{5}} - 1 = 0 \Longleftrightarrow te^{-\frac{t}{5}} = 1 \Longleftrightarrow t = e^{\frac{t}{5}} \Longleftrightarrow g_1(t) = t \text{ con } g_1(t) = e^{\frac{t}{5}}.$$

Lo mismo con

$$f(t) = 0 \Longleftrightarrow t = e^{\frac{t}{5}} \Longleftrightarrow \ln t = \frac{t}{5} \Longleftrightarrow 5\ln t = t \Longleftrightarrow g_2(t) = t \text{ con } g_2(x) = 5\ln t.$$

f) Sea g una función definida en el intervalo $[a, b] \subset \mathbb{R}$ y $t_0 \in [a, b]$ una aproximación inicial de la iteración de punto fijo dada por $t_{k+1} = g(t_k)$, con $k \geq 0$.

1. $g(t) \in [a, b]$ para todo $t \in [a, b]$,

2. g es diferenciable en $[a, b]$ y existe una constante $k < 1$ tal que $g'(t) \leq k$ para todo $t \in [a, b]$.

g) Función g_1: las funciónes

$$g_1(t) = e^{\frac{t}{5}} \text{ y } g_1'(x) = \frac{1}{5}e^{\frac{t}{5}}$$

son crecientes y para $t \in [1, 2]$

$$g_1(1) < g_1(x) < g_1(2) \Longleftrightarrow 1{,}22 < g_1(x) < 1{,}49 \Rightarrow g_1(t) \in [1, 2]$$

y se cumple la condición 1.

$$g_1'(1) < g_1'(x) < g_1'(2) \Longleftrightarrow 0{,}24 < g_1'(x) < 0{,}3 \Rightarrow |g_2'(t)| < 1$$

y se cumple la condición 2.

Función g_2: sin embargo, para $t \in [1, 2]$

$$|g_2'(t)| = \frac{5}{t} \geq g_2'(2) = \frac{5}{2} > 1 \Longleftrightarrow |g_2'(t)| > 1$$

y no se cumple la condición 2.

Así que para aproximar el punto fijo escogeríamos la función g_1.

2. En el método de bisección, usar la expresión

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

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

Una condición suficiente es que se cumpla

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

Y se tiene que

$$\frac{b_0 - a_0}{2^n} < \epsilon \Longleftrightarrow \frac{b_0 - a_0}{\epsilon} < 2^n \tag{1}$$

Como $f(x) = \log(x)$ es una función estrictamente creciente se tiene que si $0 < x < y \Longrightarrow \log(x) < \log(y)$ y a partir de (1)

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

y teniendo en cuenta las propiedades de los logaritmos

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

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

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

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

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

Computación Numérica

Segundo Parcial - Mayo 2015

1. Si el error de interpolación viene dado por

$$E(x) = f(x) - P_n(x) = f^{(n+1)}(c)\frac{(x-x_0)\dots(x-x_n)}{(n+1)!},$$

consideremos la función $f(x) = \ln(x)$ y su polinomio interpolador utilizando los nodos x_0 y x_1.

   (a) Demostrar que el error cometido al aproximar $f(x)$ mediante tal polinomio en cualquier punto de $[x_0, x_1]$ está acotado por $\frac{(x_1-x_0)^2}{8x_0^2}$.

   (b) Construir el polinomio interpolante, utilizando el método de Newton para $x_0 = 2$ y $x_1 = 4$ y dar una cota del error cometido.

a) El error de interpolación viene dado por

$$E(x) = f(x) - P_n(x) = f^{(n+1)}(c)\frac{(x-x_0)\dots(x-x_n)}{(n+1)!},$$

donde las x_i son los puntos de interpolación y c un punto del intervalo de interpolación. En este caso, como tenemos dos nodos de interpolación, la interpolación es lineal y el error es

$$|E(x)| = |f(x) - P_1(x)| = \left|f^{(2)}(c)\right| \frac{|(x-x_0)(x-x_1)|}{2!}.$$

Por una parte tenemos que $\left|f^{(2)}(c)\right| = \dfrac{1}{x^2} \leq \dfrac{1}{x_0^2}$ suponiendo $x_0 < x_1$.
Por otra parte $g(x) = |(x-x_0)(x-x_1)| = (x-x_0)(x_1-x) = -x^2 + x_0x + x_1x + x_0x_1$ como $g'(x) = -2x + x_0 + x_1 = 0$ para $x = \dfrac{x_0+x_1}{2}$. Y como $g''(x) = -2$ en este punto tenemos un máximo. El valor de la función g en este máximo es $g\left(\dfrac{x_0+x_1}{2}\right) = \dfrac{1}{4}(x_0-x_1)^2$. Por lo tanto se verifica que

$$|E(x)| = |f(x) - P_1(x)| = \left|f^{(2)}(c)\right| \frac{|(x-x_0)(x-x_1)|}{2!} < \frac{1}{x_0^2}\frac{1}{4}(x_0-x_1)^2\frac{1}{2!} = \frac{(x_1-x_0)^2}{8x_0^2}.$$

b) Para construir el polinomio

x	$f(x)$
2	$\ln 2$
3	$\ln 3$

$$\frac{\ln 4 - \ln 2}{4-2} = \frac{1}{2}(\ln 4 - \ln 2) = \frac{1}{2}\ln\frac{4}{2} = \frac{\ln 2}{2}$$

Ya tenemos

$$f[x_0] = \ln 2, \quad f[x_0, x_1] = \frac{\ln 2}{2}.$$

Construímos el polinomio interpolante en la forma de Newton

$$\begin{aligned} P_1(x) &= f[x_0] + f[x_0, x_1](x - x_0) \\ P_1(x) &= \ln 2 + \frac{\ln 2}{2}(x-2) \\ P_1(x) &= \frac{\ln 2}{2}x \end{aligned}$$

Y la cota del error es

$$\frac{(x_1 - x_0)^2}{8x_0^2} = \frac{(4-2)^2}{8(2)^2} = \frac{1}{8} = 0.125$$

2. Dada la función $f(x) = x^4$ en $[-1, 1]$ calcular la parábola que aproxima de forma continua a la función utilizando la base de polinomios $\{1, x^2\}$.

Para una base de polinomios $\{P_0, P_2\}$ el polinomio de aproximación es de la forma $P(x) = a_0 P_0(x) + a_2 P_2(x)$ donde los coeficientes a_0y a_2 son la solución del sistema lineal:

$$\begin{pmatrix} \langle P_0, P_0\rangle & \langle P_0, P_2\rangle \\ \langle P_2, P_0\rangle & \langle P_2, P_2\rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \end{pmatrix} = \begin{pmatrix} \langle P_0, f(x)\rangle \\ \langle P_2, f(x)\rangle \end{pmatrix}.$$

En este caso el producto escalar es

$$\langle g(x), h(x)\rangle = \int_{-1}^{1} g(x)h(x)dx.$$

Por lo tanto

$$\begin{pmatrix} \int_{-1}^{1} P_0P_0dx & \int_{-1}^{1} P_0P_2dx \\ \int_{-1}^{1} P_2P_0dx & \int_{-1}^{1} P_2P_2dx \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \end{pmatrix} = \begin{pmatrix} \int_{-1}^{1} P_0f(x)dx \\ \int_{-1}^{1} P_2f(x)dx \end{pmatrix},$$

es decir

$$\begin{pmatrix} \int_{-1}^{1} 1dx & \int_{-1}^{1} x^2dx \\ \int_{-1}^{1} x^2dx & \int_{-1}^{1} x^4dx \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \end{pmatrix} = \begin{pmatrix} \int_{-1}^{1} f(x)dx \\ \int_{-1}^{1} x^2f(x)dx \end{pmatrix},$$

Por lo tanto

$$\begin{pmatrix} 2 & 2/3 \\ 2/3 & 2/5 \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2/5 \\ 2/7 \end{pmatrix}.$$

Tenemos dos ecuaciones

$$\begin{cases} 2a_0 + \frac{2}{3}a_2 = \frac{2}{5} \\ \frac{2}{3}a_0 + \frac{2}{5}a_2 = \frac{2}{7} \end{cases}.$$

Despejando a_0 en la primera ecuación

$$a_0 = \frac{1}{2}\left(\frac{2}{5} - \frac{2}{3}a_2\right)$$

Y sustituyéndo en la segunda ecuación:

$$\frac{2}{3}(\frac{1}{5} - \frac{1}{3}a_2) + \frac{2}{5}a_2 = \frac{2}{7} \Longrightarrow \frac{2}{15} + \frac{8}{45}a_2 = \frac{2}{7} \Longrightarrow a_2 = \frac{6}{7} \approx 0.86.$$

Entonces

$$a_0 = \frac{1}{2}\left(\frac{2}{5} - \frac{2}{3}a_2\right) = \frac{1}{2}\left(\frac{2}{5} - \frac{2}{3}\left(\frac{6}{7}\right)\right) = -\frac{3}{35} \approx -0.086.$$

Y el polinomio que aproxima los puntos es

$$P(x) = a_0P_0(x) + a_2P_2(x) = -\frac{3}{35} \times 1 + \frac{6}{7} \times x^2 \approx -0.086 + 0.86x^2.$$

3. Calcular utilizando 4 subintervalos y la regla del Trapecio compuesta la integral

$$I = \int_1^2 \ln x \, dx.$$

Determinar el número de subintervalos suficientes para que la fórmula del Trapecio compuesta proporcione un valor aproximado de I con un error menor que 10^{-2}.

Dividimos el intervalo $[1, 2]$ en n subintervalos, cada uno de ellos de longitud h. Se verifica

$$h = \frac{b-a}{n} = \frac{2-1}{4} = \frac{1}{4} = 0.25.$$

Los nodos serán

$$\begin{aligned}
x_0 =& a = 1 \\
x_1 =& x_0 + h = 1 + 0.25 = 1.25 \\
x_2 =& x_1 + h = 1.25 + 0.25 = 1.5 \\
x_3 =& x_2 + h = 1.5 + 0.25 = 1.75 \\
x_4 =& x_3 + h = 1.75 + 0.25 = 2 = b
\end{aligned}$$

Aplicamos la fórmula simple del Trapecio 4 veces:

$$\begin{aligned}
\int_1^2 \ln x dx \approx& \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \int_{x_2}^{x_3} f(x)dx + \int_{x_3}^{x_4} f(x)dx = \\
=& \frac{x_1 - x_0}{2} \left(f(x_0) + f(x_2)\right) + \frac{x_2 - x_1}{2} \left(f(x_1) + f(x_2)\right) + \\
+& \frac{x_3 - x_2}{2} \left(f(x_2) + f(x_3)\right) + \frac{x_4 - x_3}{2} \left(f(x_3) + f(x_4)\right) \\
=& \frac{h}{2} \left(f(x_0) + 2\left(f(x_1) + f(x_2) + f(x_3)\right) + f(x_4)\right) = \\
=& \frac{0.25}{2} \left(\ln 1 + 2\left(\ln 1.25 + \ln 1.5 + \ln 1.75\right) + \ln 2\right) = 0.3837
\end{aligned}$$

La fórmula del error de la regla de los Trapecios compuesta es:

$$E_h^T = -(b-a)\frac{h^2}{12}f''(c), \quad c \in (a,b)$$

En este caso, tenemos

$$a = 1,\ b = 2,\ h = \frac{b-a}{n} = \frac{1}{n}$$

Y además

$$f(c) = \ln c,\ f'(c) = \frac{1}{c},\ f''(c) = -\frac{1}{c^2}.$$

Y como $\left|f''(c)\right|$ es una función estrictamente decreciente en $[1,2]$ el valor en 1 es una cota superior del valor de la función en el intervalo

$$|f''(c)| = \frac{1}{c^2} < \frac{1}{1^2} = 1.$$

Utilizando la fórmula del error

$$\left|E_h^T\right| = (b-a)\frac{h^2}{12}\left|f''(c)\right| \leq \frac{h^2}{12} \times 1 = \frac{1}{12}\frac{1}{(n)^2} < 0.01,$$

$$\frac{1}{12 \times n^2} < 0.01$$

$$\frac{100}{12} < n^2$$

$$\sqrt{\frac{100}{12}} < n \Longrightarrow 2.88 < n$$

Y si tomamos $n = 3$ (aplicamos tres veces la regla de los trapecios simple) el número de nodos será $n+1 = 4$ y podemos garantizar que el error al aproximar la integral con la regla de los Trapecios compuesta va a ser menor que 10^{-2}.

Computación Numérica

Tercer Parcial - Mayo 2015

1. Sea el sistema $A\mathbf{x} = \mathbf{b}$ donde

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 1 & 2 & 5 \\ 2 & 1 & 0 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 4 \\ 21 \\ 2 \end{pmatrix}$$

   a) Resolver el sistema por descomposición LU con pivote parcial. Indicar, en cada paso, las operaciones por fila realizadas.

   b) ¿Se puede resolver este sistema por Jacobi?¿Por qué?¿Y por Gauss-Seidel?¿Por qué?

   c) Reordenar las ecuaciones (filas) del sistema para que la matriz A' del sistema equivalente $A'x = b'$ sea diagonal dominante por filas. (Una vez construída A' justificar que es diagonal dominante)

   d) Si resolvemos el sistema $A'x = b'$ ¿Converge por Jacobi?¿Por qué? ¿Y por Gauss-Seidel? ¿Por qué?

a) La matriz P será inicialmente

$$P_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Aplicamos la estrategia del pivote: en el primer paso buscamos el elemento de mayor valor absoluto por debajo del pivote a_{11}. Este resulta ser a_{31} por lo que intercambiamos las filas 3 y 1 tanto en P como en A.

$$A_1 = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 5 \\ 0 & 2 & 1 \end{pmatrix} \quad P_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Ahora, hacemos ceros por debajo del elemento a_{11} restando la primera fila multiplicada por el real construído con el pivote (a_{11}) en el denominador y el elemento de esa fila por debajo del pivote (a_{21} y a_{23} respectivamente) en el numerador.

$$\begin{matrix} f_1 \\ f_2 \\ f_3 \end{matrix} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 5 \\ 0 & 2 & 1 \end{pmatrix} \quad \begin{matrix} f_1 \to f_1 \\ f_2 \to f_2 - \boxed{1/2}\ f_1 \\ f_3 \to f_3 - \boxed{0/2}\ f_1 \end{matrix}$$

Los multiplicadores, que aparecen en rojo, son los elementos con los que construímos la matriz L. Los insertamos en la matriz, en lugar de los ceros creados. La matriz transformada es

$$\begin{pmatrix} 2 & 1 & 0 \\ \boxed{1/2} & 3/2 & 5 \\ \boxed{0} & 2 & 1 \end{pmatrix}$$

Volvemos a aplicar la estrategia del pivote: buscamos el elemento de mayor valor absoluto por debajo del pivote a_{22}. Este resulta ser a_{32} por lo que intercambiamos las filas 2 y 3 en P, A y L.

$$A_2|L_2 = \begin{pmatrix} 2 & 1 & 0 \\ \boxed{0} & 2 & 1 \\ \boxed{1/2} & 3/2 & 5 \end{pmatrix} \quad P_2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Ahora, hacemos ceros por debajo del elemento a_{22} restando la segunda fila multiplicada por el real construído con el pivote (a_{22}) en el denominador y el elemento de esa fila por debajo del pivote (a_{32}) en el numerador.

$$\begin{pmatrix} 2 & 1 & 0 \\ \boxed{0} & 2 & 1 \\ \boxed{1/2} & 3/2 & 5 \end{pmatrix} \quad \begin{array}{lcl} f_1 & \to & f_1 \\ f_2 & \to & f_2 \\ f_3 & \to & f_3 - \boxed{(3/2)/2}\ f_2 \end{array}$$

Y llegamos a la matriz que almacena simultáneamente L y U.

$$\begin{pmatrix} 2 & 1 & 0 \\ \boxed{0} & 2 & 1 \\ \boxed{1/2} & \boxed{3/4} & 17/4 \end{pmatrix}$$

Y las matrices L, U y P son:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \boxed{0} & 1 & 0 \\ \boxed{1/2} & \boxed{3/4} & 1 \end{pmatrix} \quad U = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 17/4 \end{pmatrix} \quad P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Ahora, teniendo en cuenta que queremos resolver $Ax = b$, $PAx = Pb$ y que $PA = LU$ podemos escribir $LUx = Pb$ y si llamamos $Ux = y$:

1. Resolvemos el sistema triangular inferior $Ly = Pb$ $\left(Pb = (2, 4, 21)^T\right)$ y obtenemos y.

$$\begin{array}{rrrcl} y_1 & & & = & 2 \\ & y_2 & & = & 4 \\ (1/2)y_1 & +(3/4)y_2 & +y_3 & = & 21 \end{array} \qquad \begin{array}{lcl} y_1 & = & 2 \\ y_2 & = & 4 \\ y_3 & = & 21 - (1/2)y_1 - (3/4)y_2 = 17 \end{array}$$

2. Resolvemos el sistema triangular superior $Ux = y$ y obtenemos x, que era lo que buscábamos.

$$\begin{array}{rrrcl} 2x_1 & +x_2 & & = & 2 \\ & 2x_2 & +x_3 & = & 4 \\ & & (17/4)x_3 & = & 17 \end{array} \qquad \begin{array}{lcl} x_3 & = & 4 \\ x_2 & = & (4 - x_3)/2 = 0 \\ x_1 & = & (2 - x_2)/2 = 1 \end{array}$$

Y la solución es

$$\mathbf{x} = (1, 0, 4)^T$$

b) El sistema no se puede resolver por Jacobi ni por Gauss-Seidel porque la matriz de coeficientes tiene ceros en la diagonal principal y al construir el algoritmo, tanto de Jacobi como de Gauss-Seidel, utilizamos los elementos de la diagonal $a_{ii} \quad i = 1, \ldots, n$ como divisores y por lo tanto no pueden ser cero.

c) Reordenando las filas tanto de A como de b tenemos:

$$A' = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 2 & 5 \end{pmatrix} \quad \mathbf{b}' = \begin{pmatrix} 2 \\ 4 \\ 21 \end{pmatrix}$$

y la matriz de coeficientes de este sistema es diagonal dominante por filas porque

$$A' = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 2 & 5 \end{pmatrix} \begin{matrix} |2| > |1| + |0| \\ |2| > |0| + |1| \\ |5| > |1| + |2| \end{matrix}$$

d) Ahora el sistema convergería tanto por Jacobi como por Gauss-Seidel porque una condición suficiente de convergencia del método en ambos casos, es que la matriz de coeficientes sea diagonal dominante.

2. Sea $f(x,y) = x^2 + xy + y^2 - 3y$.

   a) Hallar un mínimo local de f.

   b) Probar que dicho mínimo es, de hecho, global.

   c) Aproximarlo con una iteración por el método de Newton tomando como punto inicial $(0,0)$.

   d) Hallar el mínimo sujeto a las restricciones $x \geq 0$ e $y \geq 0$.

   e) Proponer una función objetivo utilizando el método de penalización para aproximar el mínimo del apartado anterior.

a) La condición necesaria de mínimo para un punto (x_m, y_m) es que $\nabla f\,(x_m, y_m) = \left(f'_x, f'_y\right)(0,0)$. Es decir

$$\begin{cases} f'_x = 0 \\ f'_y = 0 \end{cases} \quad \begin{cases} 2x + y = 0 \\ x + 2y - 3 = 0 \end{cases} \quad \rightarrow \quad \begin{cases} & 2(3-2y) + y = 0 \rightarrow \\ x = 3 - 2y & \nearrow \end{cases} \quad \begin{cases} y = 2 & \searrow \\ & x = -1 \end{cases}$$

Y el punto $(x_m, y_m) = (-1, 2)$ cumple las condiciones necesarias de mínimo. Veamos si también cumple la condición suficiente que es que la matriz Hessiana en el punto (x_m, y_m) sea definida positiva. La matriz Hessiana para la función f en cualquier punto (x, y) es

$$H_f = \begin{pmatrix} f''_{xx} & f''_{xy} \\ f''_{yx} & f''_{yy} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Y en particular, en el punto $(x_m, y_m) = (-1, 2)$

$$H_f(-1,2) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

Veamos si es definida positiva. Para ello los determinantes de los menores principales han de ser estrictamente positivos. Es decir

$$|H_f(-1,2)| = \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} = 3 > 0 \quad \text{y} \quad |2| = 2 > 0.$$

Por lo tanto la matriz es definida positiva y se cumple la condición suficiente de mínimo.

b) Se tiene que:

- Cualquier mínimo local de una función convexa es también un mínimo absoluto.
- Si para todo elemento del dominio de f, la matriz hessiana $H_f(a)$ es definida positiva, entonces, la función f es estrictamente convexa.

En este caso particular, la matriz Hessiana f en cualquier punto coincide con la matriz Hessiana en $(-1, 2)$ y ya hemos demostrado que es definida positiva, por lo que la función f es convexa y el mínimo local $(-1, 2)$ es también mínimo global.

c) Realizaremos una iteración por el método de Newton usando la fórmula

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

Si consideramos que

$$H_{(x_0,y_0)} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \nabla f_{(x_0,y_0)} \tag{1}$$

entonces

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = H^{-1}_{(x_0,y_0)} \cdot \nabla f_{(x_0,y_0)}$$

y escribiremos

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{2}$$

donde $(c_1, c_2)^T$ es la solución del sistema (1).
Del apartado anterior tenemos que

$$\nabla f = (2x + y, x + 2y - 3)$$

y

$$\nabla f_{(x_0,y_0)} = \nabla f_{(0,0)} = (0, -3)\,.$$

Además

$$H_f = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

y

$$H_{(x_0,y_0)} = H_{(0,0)} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

El sistema (1) es

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -3 \end{pmatrix}$$

y resolviéndolo

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

Por lo tanto (2) es

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

Hemos mejorado porque $f\,(x_0, y_0) = f(0, 0) = 0$ y $f\,(x_1, y_1) = f(-1, 2) = -3$. Además $(-1, 2)$ es el mínimo global.

c) Necesitamos calcular los extremos en

- Todo el dominio de la función.
- Las fronteras de la zona donde buscamos el mínimo.
- La intersección de las fronteras.

Todo el dominio de la función. Ya está calculado en el apartado a) y es el punto $(-1, 2)$.
Las fronteras de la zona donde buscamos el mínimo. Las fronteras son $y = 0$ y $x = 0$.

$$\begin{cases} y = 0 \rightarrow & g_1(x) = f(x, 0) = x^2 \\ x = 0 \rightarrow & g_2(y) = f(0, y) = y^2 - 3y \end{cases}$$

Si evaluamos las condiciones necesarias para cada función

$$\begin{cases} y = 0 \rightarrow & g_1'(x) = 2x = 0 \rightarrow x = 0 \\ x = 0 \rightarrow & g_2'(y) = 2y - 3 = 0 \rightarrow y = 3/2 \end{cases}$$

Y las condiciones suficientes

$$\begin{cases} y = 0 \rightarrow & g_1''(x) = 2 \rightarrow g_1''(0) = 2 > 0 \rightarrow min \\ x = 0 \rightarrow & g_2''(y) = 2 \rightarrow g_2''(3/2) = 2 > 0 \rightarrow min \end{cases}$$

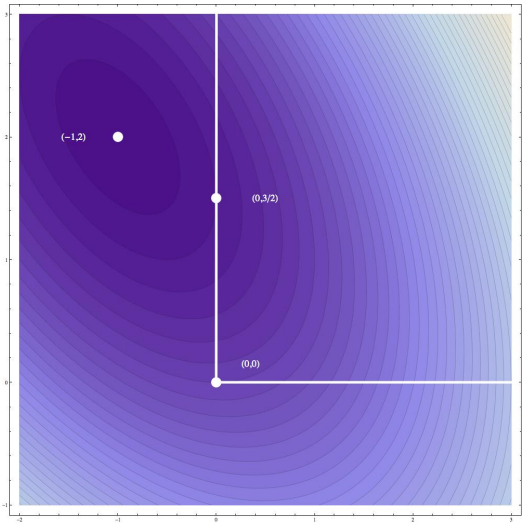
Por lo tanto

$$\begin{cases} y = 0 \rightarrow & (0, 0) \rightarrow min \\ x = 0 \rightarrow & (0, 3/2) \rightarrow min \end{cases}$$

La intersección de las fronteras. Es el punto $(0, 0)$ que también es un mínimo para la frontera $y = 0$.
Resumiendo, hemos obtenido los puntos

(x, y)	¿Pertenece a la región?	$f(x, y)$
$(-1, 2)$	No	$-$
$(0, 3/2)$	Si	$-9/4$
$(0, 0)$	Si	0

Como de los puntos que pertenecen a la región el mínimo valor le corresponde a $(0, 3/2)$ este es el mínimo global para la región que cumple las restricciones $x \geq 0,\ y \geq 0$.
La representación gráfica de la función f, la frontera y los puntos calculados es



d) La idea del método de penalización es reemplazar la función objetivo f por otra función

$$F(x, y) = f(x, y) + cP(x, y)$$

y resolver el problema sin restricciones. Para ello tomamos c como una constante positiva y P satisfaciendo:

- P es continua en el dominio de f.
- $P(x, y) \geq 0$ para todo punto del domino de f, y
- $P(x, y) = 0$ si y solo si el punto (x, y) satisface las restricciones.

De acuerdo con ello, una función F podria ser

$$F(x, y) = f(x, y) + 10\phi_1(x, y) + 10\phi_2(x, y)$$

con

$$\phi_1(x, y) = \begin{cases} x^2 & \text{si} \quad x < 0 \\ 0 & \text{si} \quad x \geq 0 \end{cases} \qquad \phi_2(x, y) = \begin{cases} y^2 & \text{si} \quad y < 0 \\ 0 & \text{si} \quad y \geq 0 \end{cases}$$

Su representación gráfica es