Algorítmica Numérica **Enero 2021**

Ejercicio 1 (4 puntos): Sea la familia de funciones $u(x)=a+b(x^2-4)$ y la tabla de datos:

x_i	0	1	3
u_i	-2	1	3

a) Dar la matriz del sistema lineal sobredeterminado que resulta al ajustar los datos de la tabla (en el sentido de mínimos cuadrados) con funciones del tipo u(x) dado inicialmente ¿Dimensión de la matriz de coeficientes?

b) Se desea ahora ajustar los datos de la tabla por funciones del tipo de u(x) con la restricción adicional de que u(2)=1. Dar las ecuaciones normales resultantes. Resolver y dar la expresión de la función solución.

c) Si se calcula el error global de los ajustes anteriores en los datos de la tabla se obtienen los valores de 1.8221 y 1.8207. Sin hacer cálculos justificad cuál de esos errores correspondería al ajuste realizado en el apartado a).

Ejercicio 2 (4 puntos): Dada la ecuación $xe^x=1$, se pide:

a) Demostrar que la ecuación tiene al menos una solución en el intervalo [0,1].

b) Con el método de la bisección, ¿cuántas iteraciones serían necesarias para estimar la solución con un error menor que 10^{-2}? No es necesario realizar las iteraciones.

c) Al aplicar Newton-Raphson se obtienen las estimaciones de error: e1~|x1-x2|=0.0254 y e2~|x2-x3|=0.0006. A partir de dichos valores dar una estimación del error e3 para la tercera iteración x3. Aproximadamente, ¿cuántas iteraciones adicionales (a partir de x3) tendría sentido hacer si trabajamos en doble precisión? Justificar vuestra respuesta.

Ejercicio 3 (2 puntos): Sea una representación en coma flotante en base 2 que usa redondeo al más próximo. Cada palabra utiliza en memoria 7 bits, de los cuales 4 bits son para la mantisa $m=(b_1b_2b_3b_4)_2$ y 3 bits para el exponente $e=(e_1e_2e_3)_2$. Los números máquina representados son los siguientes

$$\hat{x}=\left(1.b_1\,b_2\,b_3\,b_4\right)_2\times 2^{(e_1\,e_2\,e_3)_2-4}$$

a) ¿Cuántos números máquina hay en la representación? ¿Cuántos en el intervalo [2,4)?

b) Dad el valor en decimal y el contenido de los 7 bits en memoria de los números máquina correspondientes a:

- El valor máximo de la representación
- El número real x=5.6

c) Sin realizar la operación justificad cuál sería el resultado de hacer (5.6+0.1) en una máquina que usase esta representación.

SOLUCIONES

Ejercicio 1 (4 puntos):

a) Con el tipo de función aproximante dado se plantea el sistema lineal sobredeterminado en los puntos de la tabla: $u(x_i)= a + b\cdot(x_i^2-4) \approx u_i$ para i=1,2,3

Escrito en formato matricial Hc=B $\begin{pmatrix} 1 & x_1^2-4 \\ 1 & x_2^2-4 \\ 1 & x_3^2-4 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} \approx \begin{pmatrix} -2 \\ 1 \\ 3 \end{pmatrix}$. Por lo tanto H= $\begin{pmatrix} 1 & -4 \\ 1 & -3 \\ 1 & 5 \end{pmatrix}$

H es una matriz de dimensión 3x2 (se ajustan tres datos con una función que depende de dos parámetros).

b) Estudiemos primero las funciones aproximantes: funciones del tipo $u(x)=a+b(x^2-4)$ que verifiquen la restricción $u(2)=1$. Imponiendo esta restricción se obtiene a=1. Por lo tanto, la familia de funciones aproximantes a considerar en este apartado tienen la forma $u(x)=1+b(x^2-4)$ (1 parámetro libre para ajustar los datos). Imponiendo que aproxime los datos de la tabla, se llega al sistema lineal:

$$u(x_i)= 1 + b\cdot(x_i^2-4) \approx u_i \rightarrow b\cdot(x_i^2-4) \approx (u_i - 1) \quad \text{para } i=1,2,3$$

Escribimos el sistema lineal resultante en formato matricial Hb=B: $\begin{pmatrix} -4 \\ -3 \\ 5 \end{pmatrix} b \approx \begin{pmatrix} -3 \\ 0 \\ 2 \end{pmatrix}$

En este caso la matriz H tiene dimensión 3x1 (1 parámetro para ajustar 3 datos).

Las ecuaciones normales del sistema vienen dadas por $H^T\cdot H\cdot b= H^T B$, en nuestro caso:

$$\begin{pmatrix} -4 & -3 & 5 \end{pmatrix}\begin{pmatrix} -4 \\ -3 \\ 5 \end{pmatrix} b = \begin{pmatrix} -4 & -3 & 5 \end{pmatrix}\begin{pmatrix} -3 \\ 0 \\ 2 \end{pmatrix} \rightarrow 50\cdot b=22 \rightarrow b=0.44$$

Por lo tanto, la función pedida tiene la expresión $u(x)=1+0.44(x^2-4)$

c) La función del Apartado (b) es un subtipo de las funciones aproximantes del Apartado (a), por ello proporcionará un error mayor o igual que la función del Apartado (a). De los valores dados el menor error 1.8207 correspondería por lo tanto al ajuste del Apartado (a).

Ejercicio 2 (4 puntos): La ecuación $xe^x=1$ es equivalente a $f(x) = xe^x-1 =0$

a) $f(0) = -1 < 0$ y $f(1) = e-1 = 1.7\ldots >0$. Al ser la función continua debe tener al menos un cruce del eje Y=0 y por lo tanto al menos una solución.

b) En una primera estimación sabiendo que la bisección reduce el error a la mitad en cada paso y que log2(10)~3.3, se requieren algo más de 3 iteraciones para ganar cada factor de 10 en la precisión, por lo que necesitaremos del orden de 2·3.33 ~ 7 iteraciones.

Formalmente, aplicando la cota de bisección: $(b-a)/2^n < 10^{-2} \rightarrow 10^2 < 2^n \rightarrow 2 < n \cdot \log_{10}(2)$

Luego $n > 2/\log_{10}(2) = 6.64$. <u>Se precisan al menos 7 iteraciones.</u>

c) Como NR tiene convergencia cuadrática se verificará que $e_2 \sim K \cdot e_1^2$, de donde podemos estimar el valor de $K \sim 0.0006/0.0254^2 = 0.93$.

Conocida K, podemos estimar el error de x3, $e_3 \sim K \cdot e_2^2 \sim 0.93 \cdot 0.0006^2 = 3.35 \cdot 10^{-7}$.

En la siguiente iteración tendríamos un error del orden de $K \cdot e_3^2 \sim 1.04 \cdot 10^{-13}$, muy cerca ya del límite en doble precisión que es del orden de 10^{-15}. Por consiguiente, según nuestra aplicación podría valernos con una iteración adicional o podríamos hacer dos, dependiendo de si deseamos hacer el trabajo adicional para ganar esas 2 cifras decimales extras.

<u>Ejercicio 3 (2 puntos):</u>

a) ¿Cuántos números máquina hay en la representación? Los 7 bits nos permiten 2^7 combinaciones por lo que tendremos 128 números máquina.

¿Cuántos en el intervalo [2,4)? El intervalo [2, 4) corresponde a [1,2) x 2, los números con exponente=1 y todas las mantisas posibles. Como hay 4 bits dedicados a la mantisa tenemos $2^4 = 16$ mantisas posibles → **16 números máquina en el intervalo [2, 4).**

b) Dad el valor en decimal y el contenido de los 7 bits en memoria de:

El valor máximo de la representación: la máxima mantisa y el máximo exponente son **$m=(1111)_2$** y **$e=(111)_2$** y el número máquina representado es:

$$\hat{x} = (1.1111)_2 \times 2^{(111)_2 - 4} = (1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}) \times 2^{7-4} = 15.5$$

El número real x = 5.6 = 4 + 1 + 1/2 + 1/16 + 1/32 + …
= 4 · (1 + 1/4 +1/8 + 1/64 + 1/128 + …)
= 2^2 · (1.0110011…)

Redondeando la mantisa a 4 bits obtenemos el número máquina:

$$\hat{x} = (1.0110)_2 \times 2^{(110)_2 - 4} = (1 + \frac{1}{4} + \frac{1}{8}) \times 2^{6-4} = 5.5$$

Los bits guardados en memoria son **$m=(0110)_2$** y **$e=(110)_2$**

c) Hemos visto que 5.6 se guarda como el número máquina 5.5. En el intervalo [4,8) en el que estamos el salto entre números máquina es de $(1/16) \cdot 2^2 = 0.25$.

Como 0.1 < eps/2=0.125, dicho número no alcanza la mitad del salto entre los números máquina (eps/2) por lo que no pasamos al siguiente número máquina. **El resultado de la operación (5.6+0.1) es 5.5**.