Elementos de Cálculo Numérico - Cálculo Numérico
Primer Cuatrimestre de 2021
Entrega n°5

1. Se considera una máquina muy precaria que trabaja con una aritmética de punto flotante de 4 dígitos y redondeo en base 10. Sea $b = (2,1)^t$, se desea resolver por eliminación gaussiana sin pivoteo el sistema lineal $Ax = b$ donde:

$$A = \begin{pmatrix} 10^{-4} & 1 \\ -1 & 2 \end{pmatrix}$$

a) Analizar si el resultado difiere significativamente de la solución real o no.

b) Repetir la resolución pero utilizando pivoteo y estudiar si mejora significativamente o si esencialmente queda igual.

2. Considerar la siguiente ecuación diferencial:

$$\begin{cases} y'(t) &= & t\cos\left(y(t)^2\right) \\ y(0) &= & 1 \end{cases}$$

a) Probar que $0 \le y(t) \le 2$, $\forall t \in [0,1]$.

b) Escribir la iteración del método de Euler correspondiente a este problema.

c) Estudiar el error de truncado local, y hallar el valor del paso h que garantice que el error cometido al aproximar $y(1)$ sea menor que 10^{-5}. (Observación: también vale que $0 \le y_i \le 2 \; \forall \, i$)

3. Se tiene el siguiente problema de valores de contorno:

$$\begin{cases} u_{xx}(t,x) - u(t,x) = u_t(t,x) \quad x \in (0,1), \; t \in (0,1) \\ u(t,0) = u(t,1) = 0 \quad \forall t \\ u(0,x) = g(x) \end{cases}$$

a) Describir el esquema discreto explícito que utiliza la discretización usual de la derivada segunda y diferencias forward en la derivada temporal, y escribir el esquema matricial asociado.

b) Probar que si $2r + \delta t < 1$, para $r = \frac{\delta t}{(\delta x)^2}$ el método resulta estable en norma infinito.

4. Estimar la $\mathrm{cond}_\infty(A_\varepsilon)$ de la siguiente matriz en función de épsilon cuando $\varepsilon \to 0^+$:

$$A_\varepsilon = \begin{pmatrix} 1+\varepsilon & 1 & 1-\varepsilon \\ \varepsilon & \varepsilon & 0 \\ \varepsilon - 1 & 1 & -1 \end{pmatrix}$$

y concluir que está mal condicionada para $\varepsilon > 0$ suficientemente chico.

Elementos de Cálculo Numérico - Cálculo Numérico
Primer Cuatrimestre de 2021
Entrega n°5 - Resolución del modelo de parcial

1a) Calculemos primero la solución exacta:

$$\left(\begin{array}{cc|c} 10^{-4} & 1 & 2 \\ -1 & 2 & 1 \end{array}\right) \to \left(\begin{array}{cc|c} 1 & -2 & -1 \\ 0 & 1+2\times 10^{-4} & 2+10^{-4} \end{array}\right) \Rightarrow x_2 = \frac{2{,}0001}{1{,}0002},$$

y por lo tanto $x_1 = -1 + 2 \times \frac{2{,}0001}{1{,}0002} = \frac{3}{1{,}0002}$. De esta forma $x_1 \sim 2{,}9994001\ldots$, $x_2 \sim 1{,}9997\ldots$

Calculemos ahora sin pivoteo, con la aritmética de la máquina. Primero notemos que todas las entradas de la matriz y del vector b son números de máquina. Ahora calculamos $fl(F_2 - fl(fl(-10^4) \times F_1)) = fl(F_2 + fl(10000 \times F_1))$. Veamos coordenada a coordenada:

- $fl(-1 + fl(10^4 \times 10^{-4})) = fl(-1+1) = 0$.
- $fl(2 + fl(10000 \times 1)) = fl(2 + 10000) = fl(10002) = 10000$.
- $fl(1 + fl(10000 \times 2)) = fl(20001) = 20000$.

Así, nos queda: $\left(\begin{array}{cc|c} 10^{-4} & 1 & 2 \\ -1 & 2 & 1 \end{array}\right) \to \left(\begin{array}{cc|c} 10^{-4} & 1 & 2 \\ 0 & 10000 & 20000 \end{array}\right) \Rightarrow x_2 = 2, x_1 = 0.$

Veamos el error relativo, por ejemplo, en norma infinito:

$$\frac{\|(0,2) - (\frac{3}{1{,}0002}, \frac{2{,}0001}{1{,}0002})\|_\infty}{\|(\frac{3}{1{,}0002}, \frac{2{,}0001}{1{,}0002})\|_\infty} = \frac{\|(-\frac{3}{1{,}0002}, \frac{0{,}0003}{1{,}0002})\|_\infty}{\|(\frac{3}{1{,}0002}, \frac{2{,}0001}{1{,}0002})\|_\infty} = \frac{\frac{3}{1{,}0002}}{\frac{3}{1{,}0002}} = 1.$$

1b) Ahora lo calculamos con pivoteo. Como $|-1| > |10^{-4}|$, cambiamos las filas de lugar: $\left(\begin{array}{cc|c} -1 & 2 & 1 \\ 10^{-4} & 1 & 2 \end{array}\right) \underset{fl(F2+fl(fl(10^{-4})\times F_1))}{\longrightarrow} \left(\begin{array}{cc|c} -1 & 2 & 1 \\ 0 & 1 & 2 \end{array}\right) \Rightarrow x_2 = 2, x_1 = 3.$

Las cuentas en detalle son:

- $fl(10^{-4} + fl(10^{-4} \times (-1))) = fl(10^{-4} - 10^{-4}) = 0$.
- $fl(1 + fl(0,0001 \times 2)) = fl(1 + 0,0002) = fl(1,0002) = 1$.
- $fl(2 + fl(0,0001 \times 1)) = fl(2,0001) = 2$.
- $-fl(1 - fl(2 \times 2)) = -fl(-3) = 3$.

Veamos el error relativo, por ejemplo, en norma infinito:

$$\frac{\|(3,2) - (\frac{3}{1{,}0002}, \frac{2{,}0001}{1{,}0002})\|_\infty}{\|(\frac{3}{1{,}0002}, \frac{2{,}0001}{1{,}0002})\|_\infty} = \frac{\|(-\frac{0{,}0006}{1{,}0002}, \frac{0{,}0003}{1{,}0002})\|_\infty}{\|(\frac{3}{1{,}0002}, \frac{2{,}0001}{1{,}0002})\|_\infty} = \frac{\frac{0{,}0006}{1{,}0002}}{\frac{3}{1{,}0002}} = 0{,}0002,$$

que es notablemente mejor.

2a) Para probar que $0 \leq y(t) \leq 2$, $\forall t \in [0,1]$, vamos a utilizar el Teorema del Valor Medio:

$$y(t) - y(0) = y'(\xi)(t-0) \Rightarrow y(t) = 1 + t\xi \cos\left(y(\xi)^2\right)$$

para algún $\xi \in [0,1]$. Como $t \in [0,1]$ y $\cos\left(y(\xi)^2\right) \in [-1,1]$, vale entonces que $y(t) \in [0,2]$.

2b) La iteración del método de Euler con paso $h = \frac{1}{N}$ es:

$$\begin{cases} y_{i+1} = y_i + ht_i(\cos(y_i)^2), \text{ para } 0 \leq i \leq N-1, \\ y_0 = 1. \end{cases}$$

2c) El error de truncado local para el método de Euler es $\tau_i = \frac{h}{2} y''(\xi_i)$, para $\xi_i \in [t_i, t_{i+1}]$, $0 \leq i \leq N-1$. Como

$$\begin{aligned} y''(t) &= \frac{d}{dt} f(t, y(t)) = \frac{d}{dt}(t\cos((y(t)^2)) = \cos(y(t)^2) + t(-\operatorname{sen}(y(t)^2))2y(t)y'(t) \\ &= \cos(y(t)^2)[1 - 2t^2 y(t)\operatorname{sen}(y(t)^2)], \end{aligned}$$

tenemos:

$$\tau = \frac{h}{2} y''(\xi) = \frac{h}{2} \cos(y(\xi)^2)[1 - 2\xi^2 y(\xi)\operatorname{sen}(y(\xi)^2)].$$

Para acotar esta expresión (en módulo) debemos considerar que $0 \leq y(\xi) \leq 2$, $|\cos(z)| \leq 1$, $|\operatorname{sen}(z)| \leq 1$ para cualquier argumento z, que $\xi \in (0,1)$ y por lo tanto $|\xi| < 1$, más la desigualdad triangular. Luego:

$$|\tau| \leq \frac{h}{2}[1 + 2 \times 2] = \frac{5}{2} h.$$

De esta forma, el valor máximo que puede tomar $|\tau|$, τ_{MAX}, se puede acotar por $\tau_{MAX} \leq \frac{5}{2} h$.

Para encontrar un valor del paso h para que el error cometido al aproximar $y(1)$ sea menor que 10^{-5}, debemos trabajar con el error *global* del método. Esto es, $|e_N| = |y(t_f) - y_N|$, donde t_f es el tiempo final y $N = (t_f - t_0)/h$ para t_0 el tiempo inicial. Sabemos que

$$|e_N| \leq \frac{\tau_{MAX}}{K}(e^{K(t_f - t_0)} - 1),$$

donde, para el método de Euler, K es tal que es independiente de t y de h y vale que $|f(t,y) - f(t,z)| \leq K|y - z|$.

Veamos:

$$|f(t,y) - f(t,z)| = |t\cos(y^2) - t\cos(z^2)| \underset{t \in [0,1]}{\leq} |\cos(y^2) - \cos(z^2)|.$$

Para acotar $|\cos(y^2) - \cos(z^2)|$ podemos usar el Teorema del Valor Medio, para $\varphi(x) = \cos(x^2)$ y nos queda $|\cos(y^2) - \cos(z^2)| = |-\operatorname{sen}(\xi)2\xi||y - z|$:

$$|f(t,y) - f(t,z)| \leq 2|\xi||y - z|,$$

para ξ entre y y z. Si miramos la demostración del Teorema 8.10 del apunte de la materia, notaremos que los roles de y y z son los de las soluciones exacta y numérica, respectivamente. Como sabemos que cumplen $0 \leq y(t) \leq 2$ para todo $t \in [0,1]$ y $0 \leq y_i \leq 2$ para todo $0 \leq i \leq N$, tendremos $\xi \in [0,2]$, y por lo tanto podemos tomar $K = 4$.

De esta forma, nos queda:

$$|e_N| \leq \frac{\tau_{MAX}}{K}(e^{K(1-0)} - 1) \leq \frac{5h}{8}(e^4 - 1) \leq 50h.$$

Ahora busquemos un valor de h para que este error sea menor que 10^{-5}:

$$50h < 10^{-5} \Leftrightarrow h < \frac{1}{5000000} \sim 2 \times 10^{-7}.$$

Por lo tanto, cualquier $h < 10^{-7}$ sirve para tal propósito.

3a) Si usamos la discretización usual de la derivada segunda y diferencias forward en la derivada temporal obtenemos:

$$\begin{cases} \dfrac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} - u_j^n = \dfrac{u_j^{n+1} - u_j^n}{k}, \text{ para } 1 \leq j \leq J, 1 \leq n \leq N \\ u_0^n = u_{J+1}^n = 0, \text{ para } 0 \leq n \leq N+1, \\ u_j^0 = g(x_j), \text{ para } 0 \leq j \leq J+1, \end{cases}$$

donde $h = \frac{1}{J+1}$, $k = \frac{1}{N+1}$ y $x_j = jh$.

Agrupemos mejor la iteración. Llamemos $r = \frac{k}{h^2}$, entonces $u_j^{n+1} = ru_{j-1}^n + (1 - 2r - k)u_j^n + ru_{j+1}^n$. De esta forma:

- Si $j = 1$: $u_1^{n+1} = (1 - 2r - k)u_1^n + ru_2^n$
- Si $j = J$: $u_{J-1}^{n+1} = ru_{J-1}^n + (1 - 2r - k)u_J^n$

Luego, el sistema queda:

$$\begin{pmatrix} 1-2r-k & r & & & \\ r & 1-2r-k & r & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & r \\ & & & r & 1-2r-k \end{pmatrix} \begin{pmatrix} u_1^n \\ \vdots \\ \vdots \\ \vdots \\ u_J^n \end{pmatrix} = \begin{pmatrix} u_1^{n+1} \\ \vdots \\ \vdots \\ \vdots \\ u_J^{n+1} \end{pmatrix}$$

3b) Tenemos $\|u^{n+1}\|_\infty \leq \|A\|_\infty^{n+1} \|u^0\|_\infty$. Y $\|A\|_\infty = 2r + |1 - 2r - k|$. Si $2r + k < 1$, $|1 - 2r - k| = 1 - 2r - k \Rightarrow \|A\|_\infty = 1 - k < 1$ para todo k paso temporal en $(0,1)$, y por lo tanto $\|A\|_\infty^{n+1} \leq 1$ y el método resulta estable.

---

4) Si bien $\operatorname{cond}_\infty(A_\varepsilon) = \|A_\varepsilon\|_\infty \|A_\varepsilon^{-1}\|_\infty$, antes de calcular la inversa de una matriz conviene recordar el resultado de la Práctica 4 (Ejercicio 6) que nos indica que

$$\operatorname{cond}_\infty(A_\varepsilon) \geq \frac{\|A_\varepsilon\|_\infty}{\|A_\varepsilon - B\|_\infty},$$

para cualquier matriz B singular en $\mathbb{R}^{3\times 3}$.

Notemos que si ponemos $\varepsilon = 0$, obtenemos la matriz $B_\varepsilon = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & 1 & -1 \end{pmatrix}$, que es singular porque tiene una fila de ceros. Analicemos ahora el límite cuando $\varepsilon \to 0^+$, considerando que $\|A_\varepsilon\|_\infty = \text{máx}\{|1+\varepsilon|+1+|1-\varepsilon|, 2\varepsilon, |\varepsilon-1|+2\} \underset{0<\varepsilon<1}{=} \text{máx}\{3, 2\varepsilon, 3-\varepsilon\} = 3$ y que

$$A_\varepsilon - B_\varepsilon = \begin{pmatrix} \varepsilon & 0 & -\varepsilon \\ \varepsilon & \varepsilon & 0 \\ \varepsilon & 0 & 0 \end{pmatrix},$$

y por lo tanto $\|A_\varepsilon - B_\varepsilon\|_\infty = 2\varepsilon$. De esta forma,

$$\operatorname{cond}_\infty(A_\varepsilon) \geq \frac{3}{2\varepsilon} \underset{\varepsilon\to 0^+}{\longrightarrow} +\infty,$$

como queríamos probar.
