Elementos de Cálculo Numérico

Práctica N°1: Aritmética de punto fijo y flotante. Error de redondeo.

1. Decidir cuales de las siguientes afirmaciones son ciertas:
   (a) $\frac{n+1}{n^2} = O(\frac{1}{n})$
   (b) $\frac{1}{n} = o(\frac{1}{n \ln n})$
   (c) $\frac{1}{n \ln n} = o(\frac{1}{n})$
   (d) $\sqrt{x^2+1} + \sin(x) = O(x) \qquad (x \to \infty)$
   (e) $\frac{1}{x^2} = O(\frac{1}{x}) \qquad (x \to 0)$
   (f) $\frac{1}{x} = o(\frac{1}{x^2}) \qquad (x \to 0)$

2. Sea $\{x_n\}_{n\in\mathbb{N}}$ una sucesión de números reales. Demostrar que $x_n = x + o(1)$ si y sólo si $\lim_{n\to\infty} x_n = x$.

3. Demostrar que:
   (a) $\cos(x) = 1 - \frac{x^2}{2} + O(x^4) \qquad (x \to 0)$
   (b) $\sin(x) = x - \frac{x^3}{6} + O(x^5) \qquad (x \to 0)$

4. Sea $(a_n)_{n\in\mathbb{N}}$ una sucesión de números reales con límite l. Demostrar que si existen $p, c \in \mathbb{R},\ p, c > 0$ tal que
$$\lim_{n\to\infty} \frac{|a_{n+1} - l|}{|a_n - l|^p} = c,$$
entonces el orden de convergencia de la sucesión es exactamente p.

5. (a) Hallar el límite y calcular el orden de convergencia de las siguientes sucesiones:
$$\text{(i)} \quad a_n = \left(\frac{1}{2}\right)^n \qquad \text{(ii)} \quad b_n = \left(\frac{1}{2}\right)^{n^2} \qquad \text{(iii)} \quad c_n = \left(\frac{1}{2}\right)^{2^n}$$
   (b) Hacer un programa en `Matlab` que calcule los primeros 25 términos de las sucesiones del ítem anterior y observar la velocidad con la que cada una de ellas tiende al límite.

6. Utilizando el método de redondeo:
   (a) Hallar el número de máquina más próximo a 125.6 y a= 126 si trabaja con
   - Base 10 y mantisa de 2 dígitos.
   - Base 2 y mantisa de 8 dígitos.

   (b) Verificar para $x = 125.6$, la conocida cota para el error relativo
$$\left|\frac{x - fl(x)}{x}\right| \leq \epsilon$$
   si $\epsilon = 1/2\beta^{1-d}$ donde β es la base y d la longitud de la mantisa.

(c) ¿Cuál es, en cada caso, el valor que da la máquina como resultado de las operaciones $126 + 125.6$ y $126 - 125.6$? ¿Cuál es el error relativo de estos resultados?

7. Utilizando el método de truncamiento:

   (a) Rehacer el Ejercicio 6, con el ϵ correspondiente, es decir: $\epsilon = \beta^{-d+1}$, donde β y d son como antes.

   (b) Demostrar que, en este caso, ϵ es el menor número de máquina tal que $1+\epsilon \neq 1$. ¿Cuánto da $\beta + \epsilon$?

8. Mostrar que $fl(x)$ tiene (para ambos métodos) una escritura de la forma

$$fl(x) = x(1 + \delta_x)$$

   donde $|\delta_x| \leq \epsilon$. (Usar la cota para el error relativo).

9. Pérdida de dígitos significativos:

   (a) Si $x, y \geq 0$ demostrar que

$$\left| \frac{x + y - fl(fl(x) + fl(y))}{x + y} \right| \leq 2\epsilon + \epsilon^2.$$

   Observar que en la expresión $2\epsilon + \epsilon^2$ el valor de ϵ^2 es despreciable dado que ϵ es pequeño.

   (b) Si x e y no poseen el mismo signo, ¿puede repetir la misma cuenta? (Sugerencia: recordar el error relativo de $126 - 125.6$ en el ejercicio 6, item (c), utilizando la computadora binaria con mantisa de 8 dígitos.)

10. Un ejemplo que muestra que algunas de las reglas de la aritmética no son válidas para operaciones de punto flotante.

    (a) Intentar anticipar el resultado de los siguientes cálculos:
    (i) $(1 + \frac{\epsilon}{2}) + \frac{\epsilon}{2}$ (ii) $1 + (\frac{\epsilon}{2} + \frac{\epsilon}{2})$
    (iii) $\left((1 + \frac{\epsilon}{2}) + \frac{\epsilon}{2}\right) - 1$ (iv) $\left(1 + (\frac{\epsilon}{2} + \frac{\epsilon}{2})\right) - 1$

    (b) Efectuar estos cálculos usando `Matlab` y comprobar las predicciones hechas.

11. Hallar la raíz menor en módulo de la ecuación

$$x^2 - 40x + 0.25 = 0,$$

    utilizando aritmética de 4 dígitos y comparar con el resultado obtenido utilizando aritmética exacta. Calcular el error relativo y asegurarse de comprender de dónde viene la pérdida de dígitos significativos. ¿Se le ocurre cómo calcular con mayor precisión dicha raíz? ¿Cuál es el error relativo con el nuevo método?

12. Hallar una forma de calcular sin pérdida de dígitos significativos las siguientes cantidades, para $x \sim 0$:

    (a) $(\alpha + x)^n - \alpha^n$
    (b) $\alpha - \sqrt{\alpha^2 - x}$
    (c) $\cos x - 1$
    (d) $\sin(\alpha + x) - \sin(\alpha)$

13. Se pretende calcular las sumas $S_N = \sum_{k=1}^{N} a_k$ con $N \in \mathbb{N}$. Llamemos $\widehat{S_N}$ al valor calculado que se logra de hacer $fl(\widehat{S_{N-1}} + a_N)$.

    (a) $S_N = \displaystyle\sum_{k=1}^{N} \frac{1}{k}$. Mostrar que $\widehat{S_N}$ se estaciona a partir de algún N suficientemente grande. Deducir que a partir de entonces $S_N \neq \widehat{S_N}$.

    (b) Idem (a) para la suma $S_N = \displaystyle\sum_{k=1}^{N} \frac{2^{-k+100}+1}{k}$. Encontrar, haciendo un programa en `Matlab`, el valor de N para el cual $\widehat{S_N}$ se estaciona.

14. El desarrollo de Taylor de la función e^x proporciona una forma muy inestable de calcular este valor cuando x es negativo. Hacer un programa en `Matlab` que estime e^{-12} evaluando el desarrollo de Taylor hasta grado n de la función e^x en $x = -12$, para $n = 1, \ldots, 100$. Comparar con el valor exacto: $0.000006144212353328210\ldots$ ¿Cuáles son las principales fuentes de error? Realizar otra estimación= de e^{-12} con algún otro método que evite los problemas del método anterior (Sugerencia: Considerar $e^{-x} = 1/e^x$).

15. Calcular en `Matlab` los valores: $\text{sen}(\pi/2 + 2\pi 10^j)$ c= on $1 \leq j \leq 18$. ¿Cuánto debería dar? ¿Qué está pasando?

16. Aproximación de la derivada de una función.

    (a) Llamamos derivada discreta de f en $x = 1$ al valor

    $$d_h f(1) = \frac{f(1+h) - f(1)}{h}.$$

    Utilizando el desarrollo de Taylor, demostrar que

    $$|f'(1) - d_h f(1)| \leq |f''(1)|\frac{h}{2} + o(h) \qquad (h \to 0)$$

    siempre que f sea suficientemente derivable.

    (b) Considerar la función $f(x) = x^2$. Hacer un programa en `Matlab` que calcule los valores de $d_h f(1)$ para aproximar $f'(1)$, dándole a h los valores 10^{-18}, $10^{-17.9}$, $10^{-17.8}, \ldots,$ 10^{-1} y grafique los resultados obtenidos. Decidir si éstos se contradicen con el resultado del ítem anterior. Hacer un análisis de los cálculos efectuados para calcular $d_h f(1)$, teniendo en cuenta que la máquina utiliza aritmética de punto flotante.

    (c) Repetir el ítem anterior, dándole otros valores a h, de modo que el resultado resulte más confiable.

17. Las funciones de Bessel J_n se pueden definir del siguiente modo:

    $$J_n(x) = \frac{1}{\pi}\int_0^{\pi} \cos(x \sin\theta - n\theta) d\theta.$$

    y verifican que $|J_n(x)| \leq 1$. Se sabe además que $J_{n+1}(x) = 2n/x J_n(x) - J_{n-1}(x)$. Con los valores estimados $J_0(1) \sim 0.7651976865$, $J_1(1) \sim 0.4400505857$ y la recurrencia dada, hacer un programa en `Matlab` para calcular $J_2(1)$, $J_3(1)$, $\ldots$, $J_{10}(1)$. Decidir si la condición $|J_n(x)| \leq 1$ deja de satisfacerse. ¿Qué esta sucediendo?

18. Dada la función $\Phi : \mathbb{R} \to \mathbb{R}$ definida por

$$\Phi(x) = \sum_{k=1}^{\infty} \frac{1}{k(k+x)},$$

consideramos las siguiente dos maneras de estimar numéricamente el valor de $\Phi(x)$ para un x fijo:

- sumar los primeros n términos de la serie $\Phi(x)$,
- teniendo en cuenta que $\Phi(1) = 1$, definir

$$\Psi(x) = \Phi(x) - \Phi(1) = \sum_{k=1}^{\infty} \left(\frac{1}{k(k+x)} - \frac{1}{k(k+1)}\right) = \sum_{k=1}^{\infty} \frac{1-x}{k(k+1)(k+x)},$$

luego expresar $\Phi(x) = 1 + \Psi(x)$ y, de este modo, estimar $\Phi(x)$ como 1 más la suma de los primeros n términos de la serie $\Psi(x)$.

Predecir cuál de las dos maneras converge más rápidamente. Luego, hacer un programa que calcule y grafique el resultado obtenido con los dos métodos propuestos para calcular $\Phi(0)$, con $n = 1, \ldots, 100$. Comparar con el resultado exacto, que es $\frac{\pi^2}{6}$.

19. Algoritmo para calcular π.

Comenzar inicializando las variables a, b, c, d y e del siguiente modo: $a = 0$, $b = 1$, $c = 1/\sqrt{2}$, $d = 1/4$, $e = 1$. Luego, iterar n veces en el orden dado las siguientes fórmulas:

$$a = b, \quad b = \frac{b+c}{2}, \quad c = \sqrt{ca}, \quad d = d - e(b-a)^2, \quad e = 2e.$$

Finalmente, el valor de π puede estimarse como $f = b^2/d$, o como $g = (b+c)^2/(4d)$.

Hacer un programa que calcule los valores de π estimados por f y g cuando $n = 1, 2, \ldots, 10$. ¿Qué estimación converge más rápido? ¿Cuán precisos son sus resultados? El valor de π correcto hasta 36 dígitos es

$$\pi = 3.14159265358979323846264338327950288$$