Universidad de Buenos Aires - Facultad de Ciencias Exactas y Naturales - Depto. de Matemática
Segundo Cuatrimestre de 2003

Elementos de Cálculo Numérico

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

1. Utilizando el método de redondeo:

   (a) Halle el número de máquina más próximo a 125.6 si trabaja con
   - Base 10 y mantisa de 2 dígitos.
   - Base 2 y mantisa de 8 dígitos.

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

2. Utilizando el método de truncamiento:

   (a) Rehacer el Ejercicio 1, 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 $\beta + \epsilon$?

3. 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).

4. Pérdida de dígitos significativos:

   (a) Si $x, y \geq 0$ verifique que
   $$|\frac{x + y - fl(fl(x) + fl(y))}{x + y}| \leq 2\epsilon + \epsilon^2.$$
   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?

   (c) Calcule para $x = 126$ e $y = 125.6$, trabajando con base 2 y mantisa de 8 dígitos
   $$|\frac{x - y - fl(fl(x) - fl(y))}{x - y}|$$

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

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

   (b) Compute estos cálculos usando MATLAB y compruebe sus predicciones.

6. Hallar la raíz menor en módulo de la ecuación

$$x^2 + .4002x + .8\ 10^{-4} = 0$$

   Utilice aritmética de 4 dígitos, y compare con el resultado obtenido utilizando aritmética más exacta. Asegúrese de comprender de dónde viene la pérdida de dígitos significativos. ¿Se le ocurre cómo calcular con mayor precisión dicha raiz?

7. Hallar una forma de calcular sin pérdida de dígitos significativos las siguientes cantidades, para x tal como se indica en cada caso:

(a)	$(\alpha+x)^n-\alpha^n$	$x\sim 0$
(b)	$\alpha-\sqrt{\alpha^2-x}$	$x\sim 0$
(c)	$\cos x-1$	$x\sim 0$
(d)	$\sin(\alpha+x)-\sin(\alpha)$	$x\sim 0$

8. ¿Cuántos dígitos tiene su calculadora de mano? Calcular las raíces de

$$x^2 - 4x + 3.99999999.... = 0$$

   con tantos nueves como entren en su calculadora. Estime el resultado exacto y compare con lo obtenido. Analice las fuentes de error.
   (Sug.: escriba la ecuación en la forma

$$(x-2)^2 = 10^{-k}$$

   para algún k adecuado).

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

   (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.

10. Hacer un programa que pretenda estimar el valor de

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

Intentar una manera más astuta de hacer la suma usando que:

$$\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}$$

para probar que $\Phi(1) = 1$. Luego expresar $\Phi(x) - \Phi(1)$ como una serie infinita $\Psi(x)$, que converja más rápidamente que la que define $\Phi(x)$. Repetir este mecanismo para obtener una serie que converja aún con mayor velocidad que $\Psi(x)$ usando $\Psi(2)$.

Para $x = 0$, graficar las sumas parciales de los tres métodos.

11. Recordemos la fórmula para la suma de una serie geométrica:

$$G_N = \sum_{k=0}^{N} r^k = \frac{(1 - r^{N+1})}{1 - r} = Q_N$$

donde $r \neq 1$.

Escribir un programa que calcule G_N y Q_N para r, N cualesquiera. Tomar un r próximo a 1 (p.ej. $N = 1000$, $r = 1 - 10^{-12}$), y observar las diferencias entre ambos números. ¿Cuál de los dos valores obtenidos le parece más confiable? Analice. ¿Da lo mismo sumar G_N en un sentido o en el otro?¿por qué?.

12. La fórmula de Taylor proporciona una forma muy inestable de calcular e^x para x negativo. Calcular e^{-12} evaluando la serie de Taylor para $x = -12$, y comparar con el valor exacto: 0.000006144212354.... ¿Cuáles son las principales fuentes de error? ¿Se le ocurre otra forma de calcular e^{-12}, que sea mejor? (Sug.: Analice $e^{-x} = 1/e^x$).

13. Calcule los valores: $\mathrm{sen}(\pi/2 + 2\pi 10^j)$ con $1 \leq j \leq 18$. ¿Cuánto debería dar? ¿Qué está pasando?

14. Aproximación de la derivada de una función.

    (a) Utilice la conocida fórmula:

    $$f'(x) \sim \frac{f(x+h) - f(x)}{h}$$

    para aproximar la derivada de la función x^2 en el punto $x = 1$. Si llamamos $d_h(f) = \frac{f(1+h)-f(1)}{h}$ a la derivada discreta demuestre usando Taylor que entonces:

    $$|f'(1) - d_h(f)| \leq |f''(1)|h + o(h)$$

    siempre que la f sea suficientemente derivable. Emplear valores de h en el rango $10^{-20} < h < .1$. Grafique los valores obtenidos y su error en función de h. Discuta los resultados obtenidos. ¿Cuál es, en su máquina, el valor aproximado de h para el cual el error es menor?

(b) Repita todo lo anterior para la aproximación de f' conocida como diferencia centrada y compare:

$$f'(x) \sim \frac{f(x+h) - f(x-h)}{2h}$$

(c) Repita el item anterior con $f(x) = x^3$. Analice el error de la aproximación. ¿Cuál es ahora el mejor valor de h?

15. Implemente el siguiente algoritmo para calcular π. Iguale $a = 0$, $b = 1$, $c = 1/\sqrt{2}$, $d = 1/4$, $e = 1$. Luego itere (en el orden dado) las fórmulas:

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

Defina $f = b^2/d$, $g = (b+c)^2/(4d)$. Grafique $|f - \pi|$, $|g - \pi|$. ¿Cuál converge más rápido? ¿Cuán precisos son sus resultados? El valor de π correcto hasta 36 dígitos es

$$\pi = 3.14159265358979323846264338327950288$$

16. Las funciones de Bessel J_n se pueden definir:

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

Demuestre que $|J_n(x)| \leq 1$. Se sabe por otra parte que $J_{n+1}(x) = 2n/xJ_n(x) - J_{n-1}(x)$. Use el MATLAB y esta recurrencia para calcular $J_2(1)$, $J_3(1)$, $\ldots$, partiendo de los valores ya estimados $J_0(1) \sim 0.7651976865$, $J_1(1) \sim 0.4400505857$. Investigue si la condición $|J_n(x)| \leq 1$ deja de satisfacerse. ¿Qué esta sucediendo?