Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática

Semestre Otoño 2007
Cálculo Numérico
MA33A-02

Profesor: Gonzalo Hernández.
Auxiliar: Gonzalo Ríos.
Fecha: 21 de Marzo

Auxiliar 2: Aproximación y Errores

Resumen Materia

1. **Aproximación en aritmética finita:** Sea $x = \pm 0.m_1 m_2 ... m_t m_{t+1} \cdot b^z$.

   (a) Al redondear x para una aritmetica finita de t cifras significativas se obtiene:

      i. $x^* = \pm 0.m_1 m_2 ... m_t \cdot b^z$ si $m_{t+1} < \frac{b}{2}$

      ii. $x^* = \pm 0.m_1 m_2 ... (m_t + 1) \cdot b^z$ si $m_{t+1} \geq \frac{b}{2}$

   (b) Al truncar x para una aritmetica finita de t cifras significativas se obtiene:

      i. $x^* = \pm 0.m_1 m_2 ... m_t \cdot b^z$

2. **Error Absoluto:** $E_A(x, x^*) = |x - x^*|$

3. **Error Relativo:** $E_R(x, x^*) = \frac{|x - x^*|}{|x|}$

4. **Relación entre punto flotante y un real:** $fl(x) = x(1 + \delta)$ donde δ puede ser cero, negativo o positivo. δ es el error relativo y se cumple que $|\delta| \leq \frac{b}{2} \times b^{-t}$ donde t es la cantidad de dígitos significativos. El valor $eps = \frac{b}{2} \times b^{-t}$ es la precisión de la máquina .

5. **Propagación de errores:** $fl(x \otimes y) = (x \otimes y)(1 + \varepsilon)$ para $\otimes = +, -, *, /$

   (a) Error suma: $\varepsilon_s = \frac{x}{x+y}\varepsilon_x + \frac{y}{x+y}\varepsilon_y$

   (b) Error resta: $\varepsilon_r = \frac{x}{x-y}\varepsilon_x - \frac{y}{x-y}\varepsilon_y$

   (c) Error multiplicación: $\varepsilon_m = \varepsilon_x + \varepsilon_y$

   (d) Error división: $\varepsilon_d = \varepsilon_x - \varepsilon_y$

   (e) Error general: $\varepsilon_\Phi = \frac{x}{\Phi(x,y)}\frac{\partial \Phi}{\partial x}\varepsilon_x + \frac{y}{\Phi(x,y)}\frac{\partial \Phi}{\partial y}\varepsilon_y$

6. **Estabilidad:** si los llamados números de condicionamiento $\mu_x = \frac{x}{\Phi(x,y)}\frac{\partial \Phi}{\partial x}$ y $\mu_y = \frac{y}{\Phi(x,y)}\frac{\partial \Phi}{\partial y}$ se pueden acotar por constantes, entonces la secuencia de operaciones matemáticas es estable. En caso contrario, es inestable.

Ejercicios

1. Dado que $fl(x) = x(1+\delta)$,analice el error entre un real x y su aproximación por redondeo y demuestre que $|\delta| \leq \frac{b}{2} \times b^{-t}$ donde t es la cantidad de dígitos significativos.

2. Los cientificos empezaron desde ya a prepararse para combatir la invación extraterreste. Para esto, compraron un misil atómico nuclear radioactivo, pero su costo fue de 10^{∞},por lo cual el gobierno de los Estados Unidos solo pudo comprar uno. Como tienen una única oportunidad de destruir al enemigo, necesitan dispararlo con el menor error posible. Para esto construyen una máquina que, al ingresar la posición del enemigo en coordenadas $x, y,$ esta dispara automáticamente el misil en esa coordenada. El equipo está discutiendo sobre el tipo de función que tenga esta máquina para disminuir el error:

   (a) El fïsico, por lo que sabe de matemáticas, siempre ha sabido que la exponencial siempre se porta bien, por lo que propone: $\phi(x, y) = e^{xy}$

   (b) Uno de los ingenieros, al ver el resultado anterior, propone: $\phi(x, y) = \ln(x^2 + y^2)$

   (c) Después de ver esto, el ingeniero con el mejor promedio de Cálculo Numérico de su generación, propone dos funciones que asegura que funcionaran de forma similar:

      i. $\phi(x, y) = x^2 y^2$

      ii. $\phi(x, y) = \sqrt{\dfrac{x}{y}}$

   (d) ¿Cuál de las dos funciones anteriores usaría usted? Explique

3. Sea $\phi(x, y) = \frac{(x+y)^2 - 2xy - y^2}{x^2}$. Usando una aritmética finita de 6 cifras significativas con redondeo, calcule $\phi(x, y)$ en $x = 0.000001,\ y = 1000$

4. Un ingeniero de la Universidad del Barrio desea encontrar una aproximación de $\pi^2,$ ya que se olvidó del valor de π, solo recuerda que es un poco mayor que 3.Un Beauchefiano amigo le desea ayudar, y le propone lo siguiente. Sea la serie $S = \sum\limits_{k=1}^{\infty} \frac{1}{k^2}$. Se sabe teoricamente que $S = \frac{\pi^2}{6}$.Denotemos por $S_n = \sum\limits_{k=1}^{n} \frac{1}{k^2}$ a la suma parcial n-ésima exacta, y por Z_n la suma parcial n-ésima calculada por el computador según el algoritmo $Z_k = Z_{k-1} \oplus (1 \oslash (k \odot k))$. El error absoluto está dado por $E_{abs} = |S - Z_n|$

   (a) Utilizando la identidad $S - Z_n = S - Z_n + S_n - S_n$, indique como el error se puede dividir en un error de truncación y en un error de redondeo. Determine su comportamiento cuando $n \to \infty$

   (b) Se puede demostrar que $Z_n = S_n + E_n$ donde $E_n \leq cn$, c constante. También se puede demostrar que:

   $$\sum_{k=n+1}^{\infty} f(k) \leq \int_{n}^{\infty} f(x)dx$$

   A partir de estos propuestos, determine una función $g(n)$ que acote el error $E_{abs} = |S - Z_n|$.

   (c) Encuentre un valor de n óptimo que minimice la suma de las cotas encontradas anteriormente. Estime cuantos términos de la serie se deben sumar y cuantos dígitos decimales correctos tendrá el resultado calculado si $c = 4 \times 10^{-16}$.

Respuestas

1. Sea $x = \pm 0.m_1 m_2 ... m_t m_{t+1} \cdot b^z$. Luego $fl(x) = \begin{cases} a_i = \pm 0.m_1 m_2 ... m_t \cdot b^z & \text{si} \quad m_{t+1} < \frac{b}{2} \\ a_{i+1} = \pm 0.m_1 m_2 ...(m_t + 1) \cdot b^z & \text{si} \quad m_{t+1} \geq \frac{b}{2} \end{cases}$

   $E_R(x, a_i) = \frac{|x - a_i|}{|x|} = 1 - \frac{a_i}{x}$

   $E_R(x, a_{i+1}) = \frac{|x - a_{i+1}|}{|x|} = \frac{a_{i+1}}{x} - 1$

   $E_R(x, a_i) = E_R(x, a_{i+1}) \implies x = \frac{a_i + a_{i+1}}{2} = \pm 0.m_1 m_2 ... m_t \cdot b^z + \frac{b^{-t}}{2} \cdot b^z = \pm 0.m_1 m_2 ... m_t (\frac{b}{2})_{t+1} \cdot b^z$

   $E_R(x, fl(x)) = \frac{|x - x(1+\delta)|}{|x|} = |\delta| \leq \frac{\frac{b}{2} \times b^{-(t+1)} \cdot b^z}{0.m_1 m_2 ... m_t (\frac{b}{2})_{t+1} \cdot b^z} \leq \frac{\frac{b}{2} \times b^{-(t+1)}}{b^{-1}} = \frac{b}{2} \times b^{-t}$

2. Propagación de errores

   (a) $\varepsilon_\Phi = \frac{x}{\Phi(x,y)} \frac{\partial \Phi}{\partial x} \varepsilon_x + \frac{y}{\Phi(x,y)} \frac{\partial \Phi}{\partial y} \varepsilon_y = \frac{x}{e^{xy}} y e^{xy} \varepsilon_x + \frac{y}{e^{xy}} x e^{xy} \varepsilon_y = xy\,(\varepsilon_x + \varepsilon_y)$ inestable

   (b) $\varepsilon_\Phi = \frac{x}{\ln(x^2+y^2)} \frac{2x}{x^2+y^2} \varepsilon_x + \frac{y}{\ln(x^2+y^2)} \frac{2y}{x^2+y^2} \varepsilon_y = 2 \frac{x^2 \varepsilon_x + y^2 \varepsilon_y}{(x^2+y^2)\ln(x^2+y^2)}$ inestable

   (c) Funciones estables

      i. $\varepsilon_\Phi = \frac{x}{x^2 y^2} 2xy^2 \varepsilon_x + \frac{y}{x^2 y^2} 2x^2 y \varepsilon_y = 2\,(\varepsilon_x + \varepsilon_y)$ estable

      ii. $\varepsilon_\Phi = \frac{x}{\sqrt{\frac{x}{y}}} \frac{1}{2y\sqrt{\frac{x}{y}}} \varepsilon_x + \frac{y}{\sqrt{\frac{x}{y}}} \left(-\frac{1}{2} \frac{x}{y^2 \sqrt{\frac{x}{y}}} \right) \varepsilon_y = \frac{1}{2}\,(\varepsilon_x - \varepsilon_y)$ estable

   (d) Si se puede asegurar que x, y tengan el mismo signo se puede ocupar la segunda, siempre y cuando la coordenada y no sea muy cercana a cero. En cambio, la primera fórmula siempre andará bien, por lo que la mejor elección sería $\phi(x,y) = x^2 y^2$

3. $\phi(x,y) = \frac{(x+y)^2 - y^2 - 2xy}{x^2}$. $x = 0.000001,\ y = 1000$

   (a) $x + y = 1000.000001 \Longrightarrow fl(x+y) = 1000$

   (b) $(x+y)^2 = 1000000$

   (c) $xy = 0.001 \Longrightarrow fl(xy) = 0.001$

   (d) $y^2 = 1000000$

   (e) $x^2 = 0.000000000001$

   (f) Primera forma

      i. $\left((x+y)^2 - y^2\right) - 2xy = (1000000 - 1000000) - 0.002 = -0.002$

      ii. $\frac{\left((x+y)^2 - y^2\right) - 2xy}{x^2} = \frac{-0.002}{0.000000000001} = -2000000000.0$

   (g) Segunda forma

      i. $\left((x+y)^2 - 2xy\right) - y^2 = (1000000 - 0.002) - 1000000 = 1000000 - 1000000 = 0$

      ii. $\frac{\left((x+y)^2 - y^2\right) - 2xy}{x^2} = \frac{0}{0.000000000001} = 0$

4. $S = \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{1}{6}\pi^2$

   (a) $E_{abs} = |S - Z_n| = |S - Z_n + S_n - S_n| = |S - S_n + S_n - Z_n| \leq |S - S_n| + |S_n - Z_n|$
   El primer término corresponde al error de truncacion que tiende a cero cuando $n \to \infty$ ya que la suma parcial tiende a la serie. El segundo término corresponde al error de redondeo, que aumenta cuando $n \to \infty$, ya que al aumentar el número de operaciones, se cometen más errores de redondeo.

   (b) $E_{abs} = |S - Z_n| \leq |S - S_n| + |S_n - Z_n| \leq \int_n^\infty \frac{1}{x^2} dx + C \cdot n = -\left.\frac{1}{x}\right|_n^\infty + C \cdot n = \frac{1}{n} + C \cdot n = g(n)$

   (c) $\min g(n) = \frac{1}{n} + C \cdot n \Longleftrightarrow \frac{dg}{dn}(\bar{n}) = 0 \wedge \frac{d^2 g}{dn^2}(\bar{n}) > 0$
   $\frac{dg}{dn} = -\frac{1}{n^2} + C = 0 \Longrightarrow \bar{n}_1 = \frac{1}{\sqrt{C}} \wedge \bar{n}_2 = -\frac{1}{\sqrt{C}}$
   $\frac{d^2 g}{dn^2} = 2\frac{1}{n^3} > 0 \Longrightarrow \boxed{\bar{n} = \frac{1}{\sqrt{C}}}$
   $n = \frac{1}{\sqrt{4 \times 10^{-16}}} = 5 \times 10^7$
   $E_{abs} \leq \frac{1}{n} + C \cdot n = \sqrt{C} + \frac{C}{\sqrt{C}} = 2\sqrt{C} = 2\sqrt{4 \times 10^{-16}} = 4 \times 10^{-8}$
   $E_{rel} \leq \frac{4 \times 10^{-8}}{\frac{\pi^2}{6}} \leq \frac{4 \times 10^{-8}}{\frac{3^2}{6}} = \frac{24}{9} \times 10^{-8} \leq \frac{10}{2} \times 10^{-8}$
   Lo que implica que tendrá 8 dígitos decimales correctos.