Tarea 1

Mateo Restrepo S.
Juan S. Cárdenas R.
Juan J. Jaramillo C.
David Plazas E.

Análisis Numérico
Universidad EAFIT
2019

1.

Sea $x = 0.d_1d_2...d_kd_{k+1}d_{k+2}... \times 2^n$ un número en notación punto flotante en base 2. Se desea hallar un valor máximo para el error relativo, dada una aproximación con k cifras significativas. Se sabe que el valor se ve afectado por el tipo de redondeo, así que se hará un análisis con redondeo simétrico y redondeo por corte.

1. Redondeo simétrico.
   Si $d_{k+1} = 0$:

$$\hat{x} = 0.d_1d_2...d_k \times 2^n$$

$$\begin{aligned}
|\varepsilon| &= \left|\frac{x - \hat{x}}{x}\right| \\
&= \left|\frac{0.d_1d_2...d_kd_{k+1}d_{k+2}... \times 2^n - 0.d_1d_2...d_k \times 2^n}{0.d_1d_2...d_kd_{k+1}d_{k+2}... \times 2^n}\right| \\
&= \left|\frac{0.d_1d_2...d_kd_{k+1}d_{k+2}... - 0.d_1d_2...d_k}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right| \\
&= \left|\frac{0.d_{k+1}d_{k+2}... \times 2^{-k}}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right| \\
&= \left|\frac{0.0d_{k+2}... \times 2^{-k}}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right| \\
&= \left|\frac{0.d_{k+2}... \times 2^{-k-1}}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right|
\end{aligned}$$

   Ahora, para escoger una cota superior para el error relativo, se busca el mayor numero posible, el cual se obtiene con el mayor numerador y el menor numerador. Para el numerador, ese numero es 0.11111... el cual es equivalente a 1. Para el denominador, ese numero es 0.1000... dado que al estar en notación punto flotante, $d_1 \neq 0$. Por lo tanto,

$$\begin{aligned}
|\varepsilon| &< \left|\frac{1 \times 2^{-k-1}}{0.1}\right| \\
|\varepsilon| &< 2^{-k}
\end{aligned}$$

   Si $d_{k+1} = 1$:

$$\hat{x} = 0.d_1d_2...(d_k + 1) \times 2^n$$

$$
\begin{aligned}
|\varepsilon| &= \left|\frac{x-\hat{x}}{x}\right| \\
&= \left|\frac{0.d_1d_2...d_kd_{k+1}d_{k+2}... \times 2^n - 0.d_1d_2...(d_k+1) \times 2^n}{0.d_1d_2...d_kd_{k+1}d_{k+2}... \times 2^n}\right| \\
&= \left|\frac{0.d_1d_2...d_kd_{k+1}d_{k+2}... - 0.d_1d_2...(d_k+1)}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right| \\
&= \left|\frac{0.d_1d_2...d_kd_{k+1}d_{k+2}... - 0.d_1d_2...d_k - 2^{-k}}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right| \\
&= \left|\frac{0.d_{k+1}d_{k+2}... \times 2^{-k} - 2^{-k}}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right| \\
&= \left|\frac{(0.d_{k+1}d_{k+2}... - 1) \times 2^{-k}}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right| \\
&= \left|\frac{(1 - 0.d_{k+1}d_{k+2}...) \times 2^{-k}}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right| \\
&= \left|\frac{(1 - 0.1d_{k+2}...) \times 2^{-k}}{0.d_1d_2...d_k1d_{k+2}...}\right|
\end{aligned}
$$

Nuevamente, buscamos una cota superior buscando el mayor numerador y el menor denominador. Para el numerador, se obtiene buscando el valor mínimo para $0.1d_{k+2}...$, el cual es $0.1000...$. Para el denominador, ese valor es $0.00...0100...$, por lo que nos sirve 0.1 para la cota, puesto que es menor. Por lo tanto,

$$
\begin{aligned}
|\varepsilon| &< \left|\frac{(1-0.1) \times 2^{-k}}{0.1}\right| \\
|\varepsilon| &< \left|\frac{0.1 \times 2^{-k}}{0.1}\right| \\
|\varepsilon| &< 2^{-k}
\end{aligned}
$$

2. Redondeo por corte.

$$\hat{x} = 0.d_1d_2...d_k \times 2^n$$

$$
\begin{aligned}
|\varepsilon| &= \left|\frac{x-\hat{x}}{x}\right| \\
&= \left|\frac{0.d_1d_2...d_kd_{k+1}d_{k+2}... \times 2^n - 0.d_1d_2...d_k \times 2^n}{0.d_1d_2...d_kd_{k+1}d_{k+2}... \times 2^n}\right| \\
&= \left|\frac{0.d_1d_2...d_kd_{k+1}d_{k+2}... - 0.d_1d_2...d_k}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right| \\
&= \left|\frac{0.d_{k+1}d_{k+2}... \times 2^{-k}}{0.d_1d_2...d_kd_{k+1}d_{k+2}...}\right|
\end{aligned}
$$

De la misma manera, buscamos una cota superior, hallando un valor máximo en el numerador y un valor mínimo en el denominador. Para el numerador, ese valor es $0.111...$, que es equivalente a 1. Para el denominador, ese valor es $0.1000...$, puesto que x esta en notación punto flotante.

$$
\begin{aligned}
|\varepsilon| &< \left|\frac{1 \times 2^{-k}}{0.1}\right| \\
|\varepsilon| &< 2^{-k+1}
\end{aligned}
$$

2.

$$x = \frac{4}{5} = 0.8$$

$$\begin{aligned} 0.8 \times 2 &= \boxed{1}.6 \\ 0.6 \times 2 &= \boxed{1}.2 \\ 0.2 \times 2 &= \boxed{0}.4 \\ 0.4 \times 2 &= \boxed{0}.8 \\ &\vdots \end{aligned}$$

$$x = \frac{4}{5} = 0.8_{10} = 0.\overline{1100}_2$$

Para un computador de 32 bits, supongamos que la distribución de los 32 bits se realiza de la siguiente forma (como se mostró en la presentación):

$$\pm 0.d_1 \underbrace{d_2 d_3 ... d_2 4}_{\text{23 bits físicos}} \pm \quad \underbrace{e_1 e_2 ... e_7}_{\text{7 bits exponente}}$$

Primero, escribamos el número en notación punto flotante normalizada:

$$x = \left(\frac{4}{5}\right)_{10} = \left(+0.11001100... \times 2^{+0}\right)_2$$

Por lo tanto, la representación de este número en el computador de 32 bits sería

$$\underbrace{1}_{\text{signo}} \underbrace{10011001100110011001100}_{\text{mantisa}} \quad \underbrace{1}_{\text{signo exponente}} \quad \underbrace{0000000}_{\text{exponente}} \rightarrow 11001100110011001100110010000000$$

Para calcular el error absoluto, se determinará el número representado por el computador y se calculará la diferencia entre éste y 0.8. Luego,

$$\begin{aligned} 11001100110011001100110010000000 &\longrightarrow 0.10011001100110011001100 \\ \hat{x} = (0.10011001100110011001100)_2 &= (1 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + ... + 0 \times 2^{-24})_{10} \\ &= 0.799999952316284 \end{aligned}$$

Ahora

$$E = x - \hat{x} = 0.8 - 0.799999952316284 = 4.768371608676603 \times 10^{-8}$$

$$\varepsilon = \frac{E}{x} = \frac{4.768371608676603 \times 10^{-8}}{0.8} = 5.960464510845753 \times 10^{-8}$$

3.

Junto a este documento está el código en Matlab para calcular el épsilon y el número más grande del computador, junto con un toolbox [1] que fue utilizado para simplificar los cálculos. Se puede instalar directamente.

Referencias

[1] J. D'Errico, "Variable precision integer arithmetic," 2015. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/22725-variable-precision-integer-arithmetic

[2] F. Correa Zabala, "Métodos numéricos," pp. 57–59.