Examples

CS/SE 4X03

Ned Nedialkov

McMaster University

October 25, 2021

Example 1. Let x be a real number. Derive the error in the floating-point evaluation of

(a) `x*x`

(b) `sqrt(x)`

Assume that the square root is correcty rounded. That is, for a FP number y, $\mathrm{fl}\left(\sqrt{y}\right) = \sqrt{y}(1+\delta)$, where $|\delta| \le u$, and u is the unit roundoff.

(a) Let $\mathrm{fl}\,(x) = x(1+\epsilon)$, where $|\epsilon| \le u$. We have

$$\begin{aligned}
\mathrm{fl}\,(x * x) &= \mathrm{fl}\,(x)\,\mathrm{fl}\,(x)\,(1+\delta) \\
&= x^2(1+\epsilon)^2(1+\delta) = x^2(1+2\epsilon+\epsilon^2)(1+\delta) \\
&\approx x^2(1+2\epsilon)(1+\delta), \quad \text{since } \epsilon^2 \le u^2 \ll u \\
&= x^2(1+2\epsilon+\delta+2\epsilon\delta) \\
&\approx x^2(1+2\epsilon+\delta), \quad \text{since } |\epsilon\delta| \le u^2 \ll u \\
&= x^2(1+\gamma), \quad \text{where } \gamma = 2\epsilon+\delta.
\end{aligned}$$

Example 1. cont.

Then $|\gamma| \le 2|\epsilon| + |\delta| \le 3u$. Note the error in x is $\approx$ doubled.

(b) Let $f(x+h) = \sqrt{x+h}$. Then for a sufficiently small small h,

$$f(x+h) \approx f(x) + f'(x)h, \quad \sqrt{1+h} \approx \sqrt{1} + \frac{1}{2\sqrt{1}}h = 1 + h/2$$

Then

$$\begin{aligned}
\text{fl}\left(\sqrt{x}\right) &= \sqrt{\text{fl}\,(x)}(1+\delta) \\
&= \sqrt{x(1+\epsilon)}(1+\delta) = \sqrt{x}\sqrt{1+\epsilon}(1+\delta) \\
&\approx \sqrt{x}(1+\epsilon/2)(1+\delta), \quad \text{since } \sqrt{1+\epsilon} \approx 1+\epsilon/2 \\
&= \sqrt{x}(1+\epsilon/2+\delta+\epsilon\delta/2) \\
&\approx \sqrt{x}(1+\epsilon/2+\delta), \quad \text{since } |\epsilon\delta/2| \le u^2/2 \ll u \\
&= \sqrt{x}(1+\Delta), \quad \text{where } \Delta = \epsilon/2+\delta,
\end{aligned}$$

and $|\Delta| = |\epsilon/2| + |\delta| \le 1.5|u|$. Note the error in x is $\approx$ halved.

Example 2. Assume that a and b are normalized IEEE floating-point numbers; a and b are in the same precision, single or double. Which of the following statements is true in IEEE arithmetic:

(a) $\mathsf{fl}(a \circ b) = fl(b \circ a)$, where $\circ = +, *$

(b) $\mathsf{fl}(0.5 * a) = \mathsf{fl}(a/2)$

(c) $\mathsf{fl}\left(a \circ (b \circ c)\right) = \mathsf{fl}\left((a \circ b) \circ c\right)$

(d) $a \leq \mathsf{fl}\big((a + b)/2\big) \leq b$, where $a \leq b$.

Assume that no exceptions occur in the above operations.

(a) True.

(b) True. Multiplication by 0.5 and division by 2 result in decreasing the exponent by 1.

(c) False in general. Addition and multiplication are not associative.

Example 2. cont.

(d) For $x \leq y$, $\mathsf{fl}\,(x) \leq \mathsf{fl}\,(y)$.

We have

$$
\begin{aligned}
2a &\leq a + b \leq 2b \\
\mathsf{fl}\,(2a) &\leq \mathsf{fl}(a+b) \leq \mathsf{fl}\,(2b) \\
2a &\leq \mathsf{fl}(a+b) \leq 2b, \quad \text{since multiplication by 2 is exact} \\
a &\leq \mathsf{fl}\big((a+b)/2\big) \leq b \quad \text{since division by 2 is exact}
\end{aligned}
$$

Example 3.

1. Let A, B, and C be $n \times n$ matrices, where B and C are nonsingular. For an n-vector b, describe how you would implement the formula

$$x = B^{-1}(2A + I)(C^{-1} + A)b$$

without computing any inverses. Here, I is the $n \times n$ identity matrix.

2. What is the complexity of your approach in terms of big-O notation?

Example 3. cont.

We compute first

$$F = 2A + I.$$

Then we write

$$Bx = F(C^{-1} + A)b = FC^{-1}b + FAb.$$

We set $C^{-1}b = y$ and determine y by solving the linear system $Cy = b$.

Therefore, the overall computation can be written as

1. $F = 2A + I$
2. Solve $Cy = b$ for y
3. $f = Fy + FAb$
4. Solve $Bx = f$ for x

The complexity is $O(n^3)$.

Example 4.

Suppose we want to approximate e^x on $[0,1]$ using polynomial approximation with $x_0 = 0$, $x_1 = 1/2$, and $x_2 = 1$. Let p_2 be the interpolating polynomial. Find an upper bound for the error magnitude

$$\max_{0 \le x \le 1} |e^x - p_2(x)|.$$

$f'''(x) = e^x \le e$ on $[0,1]$. Then

$$|e^x - p(x)| \le \frac{e}{4(2+1)} \left(\frac{1}{2}\right)^3 \approx 0.028315.$$

Example 5. Given the data points

x_i	-1	0	1	2
y_i	1	1	2	0

write the interpolating polynomials using (a) monomial, (b) Newton and (c) Lagrange basis.

We have 4 points, so the degree of the interpolation polynomial is at most 3. (a) The polynomial is of the form

$$p(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3.$$

Then,

$$\begin{aligned} p(-1) &= c_0 - c_1 + c_2 - c_3 &&= 1 \\ p(0) &= c_0 &&= 1 \\ p(1) &= c_0 + c_1 + c_2 + c_3 &&= 2 \\ p(2) &= c_0 + 2c_1 + 4c_2 + 8c_3 &&= 0 \end{aligned}$$

Example 5. cont.

Since $c_0 = 1$, we have the system

$$\begin{aligned} -c_1 + c_2 - c_3 &= 0 \\ c_1 + c_2 + c_3 &= 1 \\ 2c_1 + 4c_2 + 8c_3 &= -1 \end{aligned}$$

From the first two equations, $c_2 = 1/2$. Using it in equations two and three,

$$\begin{aligned} c_1 + c_3 &= \frac{1}{2} \\ 2c_1 + 8c_3 &= -3 \end{aligned}$$

from which we determine $c_3 = -2/3$ and $c_1 = 7/6$.

Hence the interpolating polynomial is

$$p(x) = 1 + \frac{7}{6}x + \frac{1}{2}x^2 - \frac{2}{3}x^3.$$

Example 5. cont.

(b) The divided differences are

x_i	y_i	$f[\cdot,\cdot]$	$f[\cdot,\cdot,\cdot]$	$f[\cdot,\cdot,\cdot,\cdot]$
-1	1			
0	1	0		
1	2	1	$\frac{1}{2}$	
2	0	-2	$-\frac{3}{2}$	$-\frac{2}{3}$

The polynomial in Newton's form is

$$
\begin{aligned}
p(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\
+ f[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2) \\
= 1 + \frac{1}{2}(x + 1)x - \frac{2}{3}(x + 1)x(x - 1).
\end{aligned}
$$

If we simplify the above expression, we obtain the same polynomial as in the monomial basis.

Example 5. cont.

(c) In Lagrange form

$n+1 = 4 \Rightarrow n = 3$

$$p_3(x) = \sum_{j=0}^{3} y_j L_j(x) = L_0(x) + L_1(x) + 2L_2(x).$$

We have $x_0 = -1$, $x_1 = 0$, $x_2 = 1$, $x_3 = 2$. Then

$$L_0(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = \frac{(x-0)(x-1)(x-2)}{(-1-0)(-1-1)(-1-2)}$$
$$= -\frac{1}{6}x(x-1)(x-2)$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{(x+1)(x-1)(x-2)}{(0+1)(0-1)(0-2)}$$
$$= \frac{1}{2}(x+1)(x-1)(x-2)$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = \frac{(x+1)(x-0)(x-2)}{(1+1)(1-0)(1-2)}$$
$$= -\frac{1}{2}x(x+1)(x-2)$$

Example 5. cont. The polynomial is

$$p_3(x) = -\frac{1}{6}x(x-1)(x-2) + \frac{1}{2}(x+1)(x-1)(x-2) - x(x+1)(x-2).$$

If we simplify it, we obtain

$$\begin{aligned} p_3(x) &= -\frac{1}{6}x(x-1)(x-2) + \frac{1}{2}(x+1)(x-1)(x-2) - x(x+1)(x-2) \\ &= -\frac{1}{6}(x^3 - 3x^2 + 2x) + \frac{1}{2}(x^3 - 2x^2 - x + 2) - (x^3 - x^2 - 2x) \\ &= 1 - \frac{1}{3}x - \frac{1}{2}x + 2x + \frac{1}{2}x^2 - x^2 + x^2 - \frac{1}{6}x^3 + \frac{1}{2}x^3 - x^3 \\ &= 1 + \frac{7}{6}x + \frac{1}{2}x^2 - \frac{2}{3}x^3. \end{aligned}$$