Floating point. Exercises

CS/SE 4X03

Ned Nedialkov

McMaster University

September 21, 2021

FP addition and multiplication are not associative.

- $a + (b + c)$ may not be the same as $(a + b) + c$
- $a * (b * c)$ may not be the same as $(a * b) * c$

Adding FP numbers

Example 1. $0.59 + 0.24 \times 10^{-1} + 0.26 \times 10^{-2} + 0.64 \times 10^{-3} = 0.61724$

Add in 2-digit (after ".") arithmetic with rounding to nearest.

In decreasing magnitude:

$$
\begin{aligned}
0.59 + 0.24 \times 10^{-1} &= 0.614 && \to 0.61 \\
0.61 + 0.26 \times 10^{-2} &= 0.6126 && \to 0.61 \\
0.61 + 0.64 \times 10^{-3} &= 0.61064 && \to 0.61 \\
\text{error } |0.61724 - 0.61| &= 7.24 \times 10^{-3}
\end{aligned}
$$

In increasing magnitude:

$$
\begin{aligned}
0.64 \times 10^{-3} + 0.26 \times 10^{-2} &= 0.00324 && \to 0.32 \times 10^{-2} \\
0.32 \times 10^{-2} + 0.24 \times 10^{-1} &= 0.0272 && \to 0.27 \times 10^{-1} \\
0.27 \times 10^{-1} + 0.59 &= 0.617 && \to 0.62 \\
\text{error } |0.61724 - 0.62| &= 2.76 \times 10^{-3}
\end{aligned}
$$

The error can be smaller (but not always) if added in increasing magnitude.

Example 2. For what range of x is

$$\frac{e^x - 1}{2x} \approx 0.5 \qquad \text{correct to 15 decimal digits?}$$

From $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$,

$$\begin{aligned} e^x - 1 &= x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \\ \frac{e^x - 1}{2x} &= 0.5 + \frac{x}{4} + \frac{x^2}{2 \cdot 3!} + \frac{x^3}{2 \cdot 4!} + \cdots \\ \frac{e^x - 1}{2x} - 0.5 &= \frac{x}{4} + \frac{x^2}{2 \cdot 3!} + \frac{x^3}{2 \cdot 4!} \end{aligned}$$

If $|x| \le 4 \times 10^{-15}$,

$$\left| \frac{e^x - 1}{2x} - 0.5 \right| \lessapprox \frac{|x|}{4} \le 10^{-15}$$

The remaining terms are very small.

Example 3 (p. 64, exercise 12). Given $x \in [-\pi/2, \pi/2]$ and a value for $\cos x$, compute

$$\sin x = \pm\sqrt{1 - \cos^2 x}.$$

When $x \approx 0$, $\cos^2 x \approx 1$ and cancellations can occur in $1 - \cos^2 x$.

Compute using

$$\sin(x) \approx \begin{cases} x - \frac{x^3}{6} & |x| \le \epsilon \\ \sqrt{1 - \cos^2(x)} & \epsilon < x \le \pi/2 \\ -\sqrt{1 - \cos^2(x)} & -\pi/2 \le x < -\epsilon, \end{cases}$$

where take for example $\epsilon = 10^{-6}$.

Example 4 (p. 65, exercise 29). Consider solving $x^2 - 10^5 x + 1 = 0$ using 8 decimal digits.

$b^2 - 4ac = 10^{10} - 4 = 9.999\,999\,996 \times 10^9$ rounds to 10^{10}

Using the standard formula, $x_{1,2} = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$, the roots are

$$x_1 = 0, \quad x_2 = 10^5.$$

In double precision

$x_1 = 1.000\,000\,338\,535\,756 \times 10^{-5}$, $x_2 = 9.999\,999\,999 \times 10^4$

Using $x_1 x_2 = c/a$,

$$x_1 = c/(a x_2) = 10^{-5}.$$

Try also Quadratic Equation Calculator

Example 5 (p. 64, exercise 23). Let $x_0 > -1$ and consider $x_{n+1} = 2^{n+1}\left(\sqrt{1+2^{-n}x_n} - 1\right)$ for $n \geq 0$.

This sequence converges to $\ln(x_0 + 1)$.

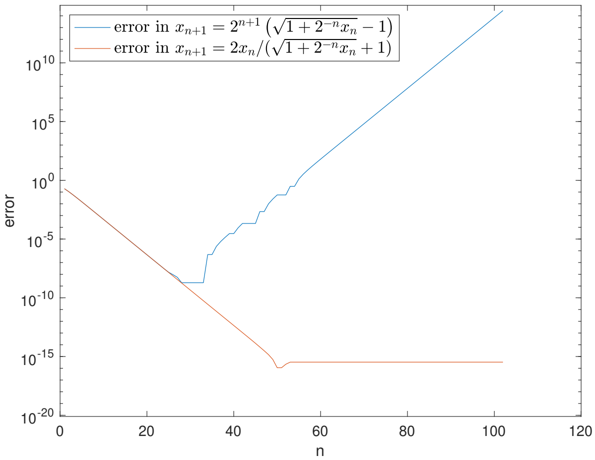
Rewrite this iteration as

$$\begin{aligned}
x_{n+1} &= 2^{n+1}\left(\sqrt{1+2^{-n}x_n} - 1\right)\frac{\sqrt{1+2^{-n}x_n}+1}{\sqrt{1+2^{-n}x_n}+1} \\
&= 2^{n+1}\frac{2^{-n}x_n}{\sqrt{1+2^{-n}x_n}+1} \\
&= 2\frac{x_n}{\sqrt{1+2^{-n}x_n}+1}
\end{aligned}$$

The following Matlab code

```
clear all; close all;
x0 = -0.5; n = 100;
x(1) = x0;
for n =0:n
    x(n+2) = 2^(n+1)*(sqrt(1+2^(-n)*x(n+1))-1);
end
semilogy([1:length(x)], abs(x-log(x0+1)))
xorig = x;
for n =0:n
    x(n+2) = 2*x(n+1)/(sqrt(1+2^(-n)*x(n+1))+1);
end
xnew = x;
hold on
semilogy([1:length(x)], abs(x-log(x0+1)));
legend('error in $x_{n+1} = 2^{n+1}\left(\sqrt{1+2^{-n}x_n}-1\right)$',...
    'error in $x_{n+1}  = 2 {x_n}/({\sqrt{1+2^{-n}x_n}+1})$',...
    'interpreter','latex', 'FontSize', 14, 'Location', 'NorthWest')
  xlabel('n')
  ylabel('error')
set(gca, 'FontSize', 12);
print('-depsc2', 'p64e23.eps')
```

produces


error in $x_{n+1} = 2^{n+1}\left(\sqrt{1+2^{-n}x_n} - 1\right)$
error in $x_{n+1} = 2x_n/(\sqrt{1+2^{-n}x_n} + 1)$
10^{10}
10^{5}
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
error
0
20
40
60
80
100
120
n

Example 6. For what values of x the expression $e^x - \sin(x) - \cos(x)$ can have cancellations and how to avoid them?

When $x \approx 0$, $e^x \approx 1$, $\sin(x) \approx 0$, $\cos(x) \approx 1$.

When $x \approx 0$

$$
\begin{aligned}
e^x &\approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6} \\
\sin(x) &\approx x - \frac{x^3}{6} \\
\cos(x) &\approx 1 - \frac{x^2}{2} \\
e^x - \sin(x) - \cos(x) &\approx x^2 + \frac{x^3}{3}
\end{aligned}
$$

For small x, use $x^2 + \frac{x^3}{3}$