MAT 128A - Assignment 1

Karry Wong

September 27, 2018

Sauer's book, Chapter 0.3 - 3,4,11; Chapter 0.4 - 1,3; Chapter 0.5 - 1a, 2a, 4a

Chapter 0.3 Problem 3

For which positive integer k can the number $5 + 2^{-k}$ be represented exactly (with no rounding error) in double precision floating point arithmetic?

Ans: Since $5 = (101)_2 = 1.01 \times 2^2$,

$$\begin{aligned} 5 + 2^{-k} &= 1.01 \times 2^2 + 1 \times 2^{-k} = 1.01 \times 2^2 + 1 \times 2^2 \times 2^{-k-2} \\ &= (1.01 + 1 \times 2^{-k-2}) \times 2^2 \end{aligned}$$

To make an exact representation possible, i.e. $fl(5 + 2^{-k}) = 5 + 2^{-k}$, we need in the above expression $2^{-k-2} \geqslant \epsilon_{\text{machine}} := 2^{-52}$.

Therefore, $k + 2 \leqslant 52 \quad \Rightarrow \quad k \leqslant 50$. Since k needs to be a positive integer, the final answer is that k can be any integers between 1 and 50.

Chapter 0.3 Problem 4

Find the largest integer k for which $fl(19 + 2^{-k}) > fl(19)$ in double precision floating point arithmetic.

Ans: Similar to the question above,

$$\begin{aligned} 19 + 2^{-k} &= 1.0011 \times 2^4 + 1 \times 2^{-k} \\ &= (1.0011 + 2^{-k-4}) \times 2^4 \end{aligned}$$

For $fl(19 + 2^{-k}) > fl(19)$, we need $2^{-k-4} \geqslant \epsilon_{\text{machine}}$, so $k \leqslant 48$. Since k can be any integers (even negative), the largest integer is 48.

Chapter 0.3 Problem 11

Does the associative law hold for IEEE computer addition?

Ans: No, $a + (b + c) = (a + b) + c$ is not always true in floating point arithmetic. For example, try to perform the following computation in MATLAB:

```
>> x = (0.1 + 0.2) + 0.3;
>> y = 0.1 + (0.2 + 0.3);
>> x - y

ans =

    1.110223024625157e-16
```

Notice that $1.110223024625157e - 16 = 1 \times 2^{-53}$. Can you explain why?

Hint: The logics is similar to the example in section 0.3.3 in Sauer's book.

For those of you who are interested, the floating point multiplication is not distributive over addition. Try to compare in MATLAB $10 \times (0.1 + 0.2)$ and $10 \times 0.1 + 10 \times 0.2$. Are they the same?

Chapter 0.4 Problem 1

Identify for which values of x there is subtraction of nearly equal numbers, and find an alternate form that avoids the problem.

$$(a).\frac{1 - secx}{tan^2 x} \qquad (b).\frac{1 - (1 - x)^3}{x} \qquad (c).\frac{1}{1 + x} - \frac{1}{1 - x}$$

1a - *Ans:* The loss of significance occurs when $1 - secx \approx 0$. Setting $1 - secx = 0$, we have $x = 2k\pi$ where k is an arbitrary integer. That is the answer for the values of x.

The reformulation makes use of the trigonometric identity $sec^2x = tan^2x + 1$, i.e.

$$\frac{1 - secx}{tan^2x} \cdot \frac{1 + secx}{1 + secx} = \frac{1 - sec^2x}{tan^2x(1 + secx)} = \frac{-1}{1 + secx}$$

Now there will not be any subtraction of nearly equal numbers.

Extra Info.: Using MATLAB, you will also find discrepancy in evaluation between the original form and the alternate form at $x = (2k+1)\pi$. This is again related to floating point arithmetrics.

1b - *Ans:* Similarly, the loss of significance occurs when $1 - (1 - x)^3 \approx 0$, i.e. when x is near 0. Expanding the term $(1 - x)^3 = 1 - 3x + 3x^2 - x^3$, we can simplify the expression to $x^2 - 3x^2 + 3$ in which subtraction of nearly equal numbers will not occur.

1c - *Ans:* The loss of significance occurs when $\frac{1}{1+x} \approx \frac{1}{1-x}$, i.e. when x is near 0. The book suggests that we should reformulate it by reducing the given form into a single fraction, i.e.

$$\frac{1}{1 + x} - \frac{1}{1 - x} = \frac{1 - x - (1 + x)}{1 - x^2} = \frac{-2x}{1 - x^2}$$

But we can do better than this! The alternate form above will have no subtraction of nearly equal numbers at $x = 0$ but at $x = 1$. The better solution here is to **apply Taylor expansion on both fractions**, i.e.

$$\frac{1}{1 + x} - \frac{1}{1 - x} = (1 - x + x^2 - x^3 \cdots) - (1 + x + x^2 + x^3 + \cdots) = -2x - 2x^3 - \cdots$$

For example, by using $\frac{1}{1+x} \approx \frac{1}{1-x} \approx -2x - 2x^3 - 2x^5$, we have error of order x^7 with no problem of subtracting nearly equal numbers at any x.

Chapter 0.4 Problem 3

Explain how to most accurately compute the two roots of the equation $x^2 + bx10^{12} = 0$, where $b > 100$

Ans: Similar to examples presented in class, using the quadratice formula, we have

$$x_1 = \frac{-b - \sqrt{b^2 + 4 \times 10^{-12}}}{2}, x_2 = \frac{-b + \sqrt{b^2 + 4 \times 10^{-12}}}{2}$$

When b is large enough, there is a danger of subtracting two nearly equal numbers in computing x_2. So multiplying both numerator and denominator with $-b - \sqrt{b^2 + 4 \times 10^{-12}}$, it becomes

$$x_2 = \frac{2 \times 10^{-12}}{b + \sqrt{b^2 + 4 \times 10^{-12}}}$$

which is numerically stable.

Chapter 0.5 Problem 1a

Use the Intermediate Value Theorem (IVT) to prove that $f(c) = 0$ for some $0 < c < 1$.

$$(a). f(x) = x^3 - 4x + 1$$

Ans: First, $f(x)$ is clearly continuous because it is a polynomial in x.

Second, $f(0) = 1$ and $f(1) = 1 - 4 + 1 = -2 < 0$. So by the IVT, there exists at least one root $c \in (0, 1)$ such that $f(c) = 0$.

Chapter 0.5 Problem 3a

Find c satisfying the Mean Value Theorem for Integrals (MVTI) with $f(x), g(x)$ in the interval $[0, 1]$.

$$(a).\ f(x) = x,\ g(x) = x \text{ in } [0, 1]$$

Ans: First, we compute the integral on the R.H.S. of MVTI, $\int_0^1 f(x)g(x)\,dx = \int_0^1 x^2\,dx = \frac{1}{3}$.

Second, we compute the integral on the L.H.S. of MVTI, $\int_0^1 g(x)\,dx = \frac{1}{2}$.

Hence,

$$c = \frac{\int_0^1 f(x)g(x)\,dx}{\int_0^1 g(x)\,dx} = \frac{2}{3}$$

Chapter 0.5 Problem 5a

Find the Taylor polynomial of degree 5 about the point $x = 0$ for:

$$(a).\ f(x) = e^{x^2}$$

The standard way of solving this problem is to calculate all the derivatives of the given function up to degree 5 and evaluate them at $x = 0$, i.e. $f(0), f'(0), f''(0), f'''(0), f^{(4)}(0)$, and $f^{(5)}(0)$. So the Taylor's expansion is $f(x) \approx 1 + x^2 + \frac{x^4}{2}$ around $x = 0$

A faster way is to make use of the Taylor's expansion of $g(x) = e^x$ around $x = 0$:

$$g(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Replacing x with x^2, we have

$$f(x) = g(x^2) = 1 + x^2 + \frac{(x^2)^2}{2} + \cdots \quad \Rightarrow \quad f(x) \approx 1 + x^2 + \frac{x^4}{2}$$

And the result follows.