MATH66: Stochastic and Numerical Methods **Fall 2023**

Homework 2

David Yang

Problems from Numerical Analysis (Sauer), Chapter 0.

Section 0.3 (Floating Point Representation of Real Numbers), Problem 3

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

Solution. First, note that the number 5 is represented in double precision floating point as

$$5 = 1.01 \times 2^2.$$

When this number (5) is summed with a number of the form 2^{-k}, rounding error will occur if and only if 1 is added after the 52^{nd} bit of the mantissa, i.e. any number smaller than

$$0.[0000000000000000000000000000000000000000000000000001] \times 2^2$$

(where the bracketed mantissa contains 52 *bits).* Thus, no rounding occur will occur as long as

$$\begin{aligned} 2^k &\leq 2^{-52} \times 2^2 \\ &= 2^{-50}. \end{aligned}$$

Thus, the number $5 + 2^{-k}$ will be represented exactly in double precision floating arithmetic for any positive integer k from $\boxed{\text{1 to 50}}$. ■

Section 0.3 (Floating Point Representation of Real Numbers), Problem 5(a)

Do the following sums by hand in IEEE double precision computer arithmetic, using the Rounding to Nearest Rule.

a) $(1 + (2^{-51} + 2^{-53}) - 1)$

Solution. Note that

$$2^{-51} = 0.[0000000000000000000000000000000000000000000000000010],$$

$$2^{-53} = 0.[0000000000000000000000000000000000000000000000000000]1,$$

and so

$$2^{-51} + 2^{-53} = 0.[0000000000000000000000000000000000000000000000000010]1.$$

By the Rounding to the Nearest Rule, since the 53rd bit is 1, the 52nd bit is 0, so $2^{-51} + 2^{-53}$ rounds down to

$$0.[0000000000000000000000000000000000000000000000000010] = 2^{-51}.$$

Adding 1 yields $1.[0000000000000000000000000000000000000000000000000010]$ and subtracting 1 afterwards yields

$$\boxed{0.[0000000000000000000000000000000000000000000000000010] = 2^{-51}}.$$

■

Section 0.3 (Floating Point Representation of Real Numbers), Problem 6(a)

Do the following sums by hand in IEEE double precision computer arithmetic, using the Rounding to Nearest Rule.

a) $(1 + (2^{-51} + 2^{-52} + 2^{-54}) - 1)$

Solution. Note that

$$2^{-51} = 0.[0000000000000000000000000000000000000000000000000010],$$

$$2^{-52} = 0.[0000000000000000000000000000000000000000000000000001], \text{ and}$$

$$2^{-54} = 0.[0000000000000000000000000000000000000000000000000000]01.$$

This means that

$$2^{-51} + 2^{-52} + 2^{-54} = 0.[0000000000000000000000000000000000000000000000000011]01.$$

By the Rounding to the Nearest Rule, since the 53rd bit is 0, so $2^{-51} + 2^{-52} + 2^{-54}$ rounds down to

$$0.[0000000000000000000000000000000000000000000000000011] = 2^{-51} + 2^{-52}.$$

Adding 1 yields 1.[0000000000000000000000000000000000000000000000000011] and subtracting 1 afterwards yields

$$\boxed{0.[0000000000000000000000000000000000000000000000000011] = 2^{-51} + 2^{-52}}.$$

■

Section 0.3 (Floating Point Representation of Real Numbers), Problem 11

Does the associative law hold for IEEE computer addition?

Solution. No. Consider $\epsilon_{\text{mach}} = 2^{-52}$. Note that

$$\left(1 + \frac{\epsilon_{\text{mach}}}{2}\right) + \frac{\epsilon_{\text{mach}}}{2} \neq 1 + \left(\frac{\epsilon_{\text{mach}}}{2} + \frac{\epsilon_{\text{mach}}}{2}\right).$$

This follows since $1 + \frac{\epsilon_{\text{mach}}}{2}$ rounds down to 1, so the left-hand side evaluates to 1 whereas the right-hand side evaluates to $1 + \epsilon_{\text{mach}}$. Thus, the associative law does not hold for IEEE computer addition. ■

Section 0.4 (Loss of Significance), Problem 1(a)

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

a) $\frac{1-\sec x}{\tan^2(x)}$

Solution. Subtraction of nearly equal numbers occurs when

$$1 \approx \sec(x) = \frac{1}{\cos(x)}$$

which occurs when $\cos(x)$ is very close to 1. Values of x at which this occur include x which are close to $2\pi n$ for integer n.

We want to find an alternate form of the expression

$$\frac{1-\sec x}{\tan^2 x}.$$

We can multiply both the numerator and denominator of the fraction by $1+\sec x$, which gives

$$\begin{aligned}\frac{1-\sec x}{\tan^2 x} &= \frac{1-\sec x}{\tan^2 x} \cdot \frac{1+\sec x}{1+\sec x} \\ &= \frac{1-\sec^2(x)}{\tan^2(x)(1+\sec x)}\end{aligned}$$

Using the identity $\tan^2(x) = \sec^2(x) - 1$, we can rewrite the numerator as $-\tan^2(x)$. Thus, we get that

$$\begin{aligned}\frac{1-\sec x}{\tan^2 x} &= \frac{1-\sec^2(x)}{\tan^2(x)(1+\sec x)} \\ &= \frac{-\tan^2(x)}{\tan^2(x)(1+\sec x)} \\ &= -\frac{1}{1+\sec x}.\end{aligned}$$

Thus, an alternate form of the expression $\frac{1-\sec x}{\tan^2(x)}$ that avoids the potential problem of subtraction of nearly equal numbers is

$$\boxed{-\frac{1}{1+\sec x}}.$$

■

Section 0.4 (Loss of Significance), Problem 3

Explain how to most accurately compute the two roots of the equation $x^2+bx-10^{-12}=0$, where b is a number greater than 100.

Solution. The quadratic formula tells us that the roots of equation

$$x = \frac{-b \pm \sqrt{b^2-4ac}}{2a} = \frac{-b \pm \sqrt{b^2+4\cdot 10^{-12}}}{2}.$$

Since $\sqrt{b^2+4\cdot 10^{-12}} \approx b$ as $b \gg 10^{-12}$, the root

$$x = \frac{-b + \sqrt{b^2+4\cdot 10^{-12}}}{2}$$

may lead to rounding error caused by the subtraction of nearly equal numbers.

As derived in Example 0.6, since b is positive, we can use the alternate form of the quadratic formula, which gives two roots

$$x_1 = -\frac{b+\sqrt{b^2-4ac}}{2a} \text{ and } x_2 = -\frac{2c}{(b+\sqrt{b^2-4ac})}.$$

Note that these forms avoid the rounding errors discussed above. Thus, in this instance, we can most accurately compute the two roots of the given equation by using the formulas

$$\boxed{x_1 = -\frac{b+\sqrt{b^2+4\cdot 10^{-12}}}{2} \text{ and } x_2 = \frac{2\cdot 10^{-12}}{(b+\sqrt{b^2+4\cdot 10^{-12}})}}$$

■