1st Seminar Exercise

Review Questions

1.42. Assume a decimal (base 10) floating-point system having machine precision $\epsilon_{mach} = 10^{-5}$ and an exponent range of ± 20. What is the result of each of the following floating-point arithmetic operations?
$L=-20 \quad U=20$
(a) $1 + 10^{-7}$
(b) $1 + 10^{3}$
(c) $1 + 10^{7}$
(d) $10^{10} + 10^{3}$
(e) $10^{10}/10^{-15}$
(f) $10^{-10} \times 10^{-15}$

Solution: We can derive that $\epsilon_{mach} = 10^{1-p_m} = 10^{-5} \Rightarrow p_m = b$, indicating this system consists of floating numbers with b digits. ($L=-20$, $U=20$)

(a)
For 10^{7}, it includes 7 digits in fraction parts, thus $1+10^{7} \approx 1$

(b)
For 10^{3}, $10^{3} = 1.00000 \times 10^{3}$ $P=b=p_m$ $E=3 \in [-20,20]$
For 1, $1 = 0.00100 \times 10^{3}$ $P=b=p_m$ $E=3 \in [-20,20]$
$\Rightarrow 1+10^{3} = 1.00100 \times 10^{3}$

(c)
For $10^{7} = 1.00000 \times 10^{7}$ $P=b=p_m$ $E=7 \in [-20,20]$
For $1 = 0.0000001 \times 10^{7}$, $P=8>p_m$ $E=7 \in [-20,20]$
reduce the precision of mantissa to 6 digits.
which is $0.00000 \Rightarrow 1+10^{7} = 1.00000 \times 10^{7}$

(d)
For $10^{10} = 1.00000 \times 10^{10}$ $P=b=p_m$ $E=10 \in [-20,20]$
For $10^{3} = 0.0000001 \times 10^{10}$, $P=8>p_m$, $E=10 \in [-20,20]$
reduce the mantissa precision to 6 digits,
which is $0.00000 \Rightarrow 10^{3}+10^{10} = 1.00000 \times 10^{10}$

(e)
For $10^{10} = 1.00000 \times 10^{10}$, $E=10 \in [-20,20]$
For $10^{-15} = 1.00000 \times 10^{-15}$, $E=-15 \in [-20,20]$
$\Rightarrow 10^{10}/10^{-15} = 10^{25}$, $E=25 \notin [-20,20]$, indicating that truncation to 20 digits are needed
$\Rightarrow 10^{10}/10^{-15} = 10^{20}$

(f)
For $10^{-10} = 1.00000 \times 10^{-10}$, $E=-10 \in [-20,20]$
For $10^{-15} = 1.00000 \times 10^{-15}$, $E=-15 \in [-20,20]$
$\Rightarrow 10^{-10} \times 10^{-15} = 10^{-25}$, $E=-25 \notin [-20,20]$, indicating that truncation to 20 digits are needed
$\Rightarrow 10^{-10} \times 10^{-15} = 0$

1.51. List at least two ways in which evaluation of the quadratic formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

may suffer numerical difficulties in floating-point arithmetic.

Solution:
① If the coefficients a, b and c are very large or very small, then b^2 or $4ac$ will overflow or underflow
② When b and $\sqrt{b^2-4ac}$ are vastly different magnitudes, significant digits in the smaller term can be lost due to the limited precision of floating-point arithmetic.

1.6. The sine function is given by the infinite series

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots.$$

(a) What are the forward and backward errors if we approximate the sine function by using only the first term in the series, i.e., $\sin(x) \approx x$, for $x = 0.1$, 0.5, and 1.0?
(b) What are the forward and backward errors if we approximate the sine function by using the first two terms in the series, i.e., $\sin(x) \approx x - x^3/6$, for $x = 0.1$, 0.5, and 1.0?

Solution:
(a) For $x=0.1$, $FE = \Delta y = \hat{f}(x) - f(x) = 0.1 - \sin(0.1) = 0.1 - 0.09983 = 0.00017$
Let $\hat{f}(\hat{x}) = f(\hat{x}) \Rightarrow f(\hat{x}) = \sin(\hat{x}) = 0.1 \Rightarrow \hat{x} = \arcsin(0.1) = 0.10017$
$BE = \Delta x = \hat{x} - x = 0.00017$

For $x=0.5$, $\Delta y = \hat{f}(x) - f(x) = 0.5 - \sin(0.5) \approx 0.5 - 0.47943 \approx 0.02057$
Let $\hat{f}(x) = f(\hat{x}) \Rightarrow f(\hat{x}) = \sin(\hat{x}) = 0.5 \Rightarrow \hat{x} = \arcsin(0.5) = \frac{\pi}{6} \approx 0.52360 \Rightarrow \Delta x = \hat{x} - x = 0.02360$

For $x=1.0$, $\Delta y = f(\hat{x}) - f(x) = 1.0 \sin(1) \approx 1.0 - 0.8414 = 0.15853$
Let $\hat{f}(x) = f(\hat{x}) \Rightarrow f(\hat{x}) = \sin(\hat{x}) = 1 \Rightarrow \hat{x} = \arcsin(1) \approx 1.57080$
$\Rightarrow \Delta x = \hat{x} - x \approx 0.57080$

(b) For $x=0.1$
$\Delta y = \hat{f}(x) - f(x) = 0.1 - \frac{0.1^3}{6} - \sin(0.1) \approx 9.98333333333 \times 10^{-2} - 9.98334166333 \times 10^{-2} \approx -8.332 \times 10^{-8}$
Let $\hat{f}(x) = f(\hat{x}) \Rightarrow f(\hat{x}) = \sin(\hat{x}) = 9.98333333333 \times 10^{-2} \Rightarrow \hat{x} = 9.99999 1626 \times 10^{-2}$
$\Rightarrow \Delta x = \hat{x} - x = -8.3735 \times 10^{-8}$

For $x=0.5$
$\Delta y = \hat{f}(x) - f(x) = 0.5 - \frac{0.5^3}{6} - \sin(0.5) \approx 4.79166666 7 \times 10^{-1} - 4.79425538 6 \times 10^{-1} \approx -2.5887 \times 10^{-4}$
Let $\hat{f}(x) = f(\hat{x}) \Rightarrow f(\hat{x}) = \sin(\hat{x}) = 4.79166666 7 \times 10^{-1} \Rightarrow \hat{x} = 0.4997050408$
$\Rightarrow \Delta x = \hat{x} - x = -2.9496 \times 10^{-4}$

For $x=1.0$
$\Delta y = \hat{f}(x) - f(x) = 1 - \frac{1^3}{6} - \sin(1) \approx 8.333333333 \times 10^{-1} - 8.41470984 8 \times 10^{-1} \approx -8.1377 \times 10^{-3}$
Let $\hat{f}(x) = f(\hat{x}) \Rightarrow f(\hat{x}) = \sin(\hat{x}) = 0.8333333333 \Rightarrow \hat{x} = 0.9851107833$
$\Rightarrow \Delta x = \hat{x} - x = -1.4889 \times 10^{-2}$

Exercises

1.2. What are the approximate absolute and relative errors in approximating π by each of the following quantities?
(a) 3
(b) 3.14
(c) 22/7

Solution:
Estimate $\pi \approx 3.1416$
(a) $ABE = 3 - \pi \approx -0.1416$
$RE = \frac{3-\pi}{\pi} \approx -4.5071 \times 10^{-2}$
(b) $ABE = 3.14 - \pi \approx -0.0016$
$RE = \frac{3.14-\pi}{\pi} \approx -5.07 \times 10^{-4}$
(c) $ABE = \frac{22}{7} - \pi = \frac{22-7\pi}{7} \approx 1.257 \times 10^{-3}$
$RE = \frac{\frac{22}{7}-\pi}{\pi} \approx 4.002 \times 10^{-4}$

1.4. Consider the problem of evaluating the function $\sin(x)$, in particular, the propagated data error, i.e., the error in the function value due to a perturbation h in the argument x.
(a) Estimate the absolute error in evaluating $\sin(x)$.
(b) Estimate the relative error in evaluating $\sin(x)$.
(c) Estimate the condition number for this problem.
(d) For what values of the argument x is this problem highly sensitive?

Solution:
Let $y = \sin(x)$
(a) Absolute rate $= \frac{\Delta y}{\Delta x} = \frac{dy}{dx} = \cos x$
$\Rightarrow ABRE = \Delta y = \cos(x) \cdot \Delta x = \cos(x)\, h$
(b) $RE = \frac{\cos(x) h}{\sin(x)} = \cot(x)\, h$
(c) $cond = \left| \frac{\cot(x) h}{\frac{h}{x}} \right| = |x \cot(x)|$
(d) This problem is highly sensitive for $x \to k\pi$ ($k \in \mathbb{Z}$)

1.11. If $x \approx y$, then we would expect some cancellation in computing $\log(x) - \log(y)$. On the other hand, $\log(x) - \log(y) = \log(x/y)$, and the latter involves no cancellation. Does this mean that computing $\log(x/y)$ is likely to give a better result? (*Hint*: For what value is the log function sensitive?)

Solution:

Notice that the log function is sensitive when x is near 1.

We know that $\log_2(x) - \log_2(y) = \log_2\left(\frac{x}{y}\right)$

if $x \simeq y \Leftrightarrow \frac{x}{y} \to 1 \Rightarrow \lim_{\frac{x}{y}\to 1} \log_2\left(\frac{x}{y}\right) = 0$.

which causes cancellation.

Thus, computing $\log_2\left(\frac{x}{y}\right)$ cannot make a better result.

1.13. The Euclidean norm of an n-dimensional vector $\boldsymbol{x}$ is defined by

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}.$$

How would you avoid overflow and harmful underflow in this computation?

Solution:

Divide each entry in vector $\vec{x}$ by the largest entry in magnitude.

Thus, it can necessarily ovoid overflow.

Multiply the scaling factor back after computing and obtaining the scaled vector.