UECM3033 Numerical Methods

Tutorial : Mathematical Preliminaries

(Jan 2012)

1. The Maclaurin series expansion for $\sin x$ is

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

   Starting with the simplest $\sin x = x$, add a terms one at a time to estimate $\sin(\pi/4)$. Compute the percent relative errors (use your calculator for the true value) after each term is added.

2. Use zero- through third-order Taylor series expansions centred about $x = 1$ to predict $f(2)$ for

$$f(x) = 25x^3 - 6x^2 + 7x - 88.$$

   Compute the precent relative error ϵ for each approximation.

3. To calculate a planet's space coordinates, we have to solve the function $f(x) = x - 1 - 0.5 \sin x$. Let the centre of Taylor series expansion be at $x = \pi/2$ on the interval $[0, \pi]$. Determine the highest-order Taylor series expansion needed to obtain a maximum error of 0.015 on the specific interval.

4. (Familiarization with MatLab). Start up a MATLAB session. Find:
   a) the largest floating point number in MATLAB.
   b) the smallest machine tolerance in MATLAB.
   c) what is the possible round off error of a number in your computer?

5. Write down the IEEE single precision (32-bit) representation for each of the following numbers:
   a) $1 = 2^0 * (1.0 \ldots 0)_2$,
   b) $2^{-1} * (1.111111111\ 11111111111\ 111)_2$,
   c) $2^{127} * (1.111111111\ 11111111111\ 111)_2$.

6. Write down the IEEE double precision (64-bit) representations for the same numbers in the previous question.

7. Assume that 4-digits arithmetic with rounding is used in calculating $(\frac{1}{10} + \frac{1}{3}) + \frac{1}{5}$.
   a) Find the error and relative error of the calculation.
   b) Find also the error and relative error if the same calculation is done by a modern 64-bit computer.

8. Add $x_1 = 0.36789, x_2 = 2.5678, x_3 = 0.1234$ and $x_4 = 0.034567$ using 4-digit rounding floating-point arithmetic. Rearrange these numbers from the smallest to the largest and then add them using 4-digit rounding arithmetic. Compare the two sums with the exact sum.