Set 1 — Computer Arithmetic and Errors

Week 8

* — tutorial question; A — advanced

1. Give ways to evaluate the following functions accurately to the number of digits used for $x$ near the given values.

(a) $f(x) = \frac{x - \sin x}{\tan x}$, $x \approx 0$.

(b) $f(x) = \ln(x + 1) - \ln x$, $x \gg 1$.

Solution.

Assume that $UFL \ll \epsilon_{\text{mach}}$ and that $|x| \ll \epsilon_{\text{mach}}$. Then $x^2 \ll |x|$ and $x^2$ can be neglected compared to $x$.

(a) Using the Taylor expansions, $\tan x = x + O(x^3)$ and $x - \sin x = \frac{1}{6}x^3 + O(x^5)$,

$$\frac{x - \sin x}{\tan x} = \frac{1}{6}x^3 + O(x^5) = \frac{1}{6}x^2 + O(x^4).$$

(b) Using the Taylor expansion, $\ln(1 + z) = z + O(z^2)$,

$$\ln(x + 1) - \ln x = \ln \left(1 + \frac{1}{x}\right) = \frac{1}{x} + O \left(\frac{1}{x^2}\right).$$

2. Use Taylor’s theorem for functions of two variables to find a quadratic approximation to $\sqrt{1 + x - y}$ at $(x, y) = (0, 0)$.

Solution.

Let $z = \sqrt{1 + x - y}$. The derivatives of $z$ up to second order are

$$z_x = \frac{1}{2\sqrt{1 + x - y}} = -z_y,$$

$$z_{xx} = \frac{1}{4(1 + x - y)^{3/2}} = z_{yy} = -z_{xy}.$$ 

Evaluating $z$ and these derivatives at $(0, 0)$ gives

$$z = 1, \quad z_x = \frac{1}{2} = -z_y, \quad z_{xx} = \frac{1}{4} = z_{yy} = -z_{xy}.$$
Hence

\[
z = z(0,0) + z_x(0,0)x + z_y(0,0)y + \frac{1}{2}z_{xx}(0,0)x^2 + z_{xy}(0,0)xy + \frac{1}{2}z_{yy}(0,0)y^2 + \ldots
\]

\[
= 1 + \frac{1}{2}x - \frac{1}{2}y - \frac{1}{8}x^2 + \frac{1}{4}xy - \frac{1}{8}y^2 + \ldots.
\]

Alternatively, use the binomial expansion to expand \( \sqrt{1+z} \) in a Taylor series,

\[
\sqrt{1+z} = 1 + \frac{1}{2}z - \frac{1}{8}z^2 + \ldots,
\]

and then set \( z = x - y \).

3* Fortran 90 provides the user with intrinsic functions to determine the properties of the number system(s) which his programs use. Write a simple program which:

(a) uses \textit{epsilon(x)}, \textit{huge(x)} and \textit{tiny(x)} to determine the \( \epsilon_{\text{mach}} \), OFL and UFL of type \textit{real};
(b) uses \textit{precision(x)} and \textit{range(x)} to determine the decimal precision and largest decimal exponent of type \textit{real};
(c) uses \textit{digits(x)}, \textit{maxexponent(x)}, \textit{minexponent(x)} and \textit{radix(x)} to determine the number of bits in the mantissa, the largest and smallest binary exponents, and the base of type \textit{real};
(d) repeats Parts (a)–(c) for type \textit{double precision};
(e) calculates \( x = 4.**1/2 \) and \( x = (-8.)**2/3. \), and prints the answers; and
(f) uses a \texttt{DO}-loop to calculate \( 2, 2^2, 2^3, \ldots, 2^{32} \).

(Note: \( x \) need only have the required type, \textit{real} or \textit{double precision}, and not an actual numerical value in these particular intrinsic functions.)

\textit{Solution.}

See set1q3.ans.

4. Consider the floating point number system

\[
\left\{ \pm d_1.d_2d_3 \times 10^{\pm p} \mid d_2, d_3, p = 0, \ldots, 9; \ d_1 = 1, \ldots, 9, \text{ unless } d_1 = d_2 = d_3 = 0 \right\}
\]

What are the UFL, OFL and \( \epsilon_{\text{mach}} \) of this system?

\textit{Solution.}

The smallest and largest numbers are

\[
UFL = 1.00 \times 10^{-9}
\]

\[
\]

The smallest number greater than 1 is 1.01; thus \( 1 + \epsilon_{\text{mach}} = 1.01 \). With rounding, \( 1.00 + 5.00 \times 10^{-2} = 1.005 = 1.01 \), but \( 1.00 + 4.99 \times 10^{-2} = 1.00499 = 1.00 \). With chopping \( 1.00 + 1.00 \times 10^{-2} = 1.01 \), but \( 1.00 + 9.99 \times 10^{-3} = 1.00999 = 1.00 \). Hence

\[
\epsilon_{\text{mach}} = \begin{cases} \ 5.00 \times 10^{-3}, & \text{rounding;} \\ 1.00 \times 10^{-2}, & \text{chopping.} \end{cases}
\]
5. (a)* Use the central difference quotient of \( f(x) \) at \( x \) defined by
\[
\delta_h f(x) := \frac{f(x + h) - f(x - h)}{2h},
\]
with \( h = 10^k 3^{\sqrt{\epsilon_{\text{mach}}}} L \) for \( k = -2, -1, 0, 1, 2 \) to estimate the first derivatives of

(i) \( f_1(x) = \sin x \) at \( x = 1, L = 1 \);
(ii) \( f_2(x) = 10000 \sin x \) at \( x = 1, L = 1 \);
(iii) \( f_3(x) = \tan x \) at \( x = 1.59, L = 1 \);
(iv) \( f_4(x) = \sin 100x \) at \( x = 1, L = 1/100 \).

Which values of \( h \) give the best estimates?

(b)* By expanding \( f(x + h) \) in a Taylor series in \( h \) to six terms with a remainder find an expression for the truncation error if the central difference quotient in Part (a) is used to approximate the derivative \( f'(x) \).

(c)* Using the result that the maximum absolute relative roundoff error is bounded by the machine epsilon, show that the round-off error in approximating \( f'(x) \) by \( \delta_h f(x) \) is bounded by \( |f(x)| \epsilon_{\text{mach}} / h \).

(d)* Show that the total error in approximating \( f'(x) \) by \( \delta_h f(x) \) is bounded by
\[
\frac{h^2}{6} |f'''(x)| + \frac{1}{h} |f(x)| \epsilon_{\text{mach}} + \frac{h^4}{120} |f^{(5)}(\eta)|,
\]
where \( \eta \) is some number near \( x \). Show that the \( h \) that minimises the sum of the two dominant terms on the right hand side is given by,
\[
h \sim 3^{\sqrt{\epsilon_{\text{mach}}}} L,
\]
where \( L \) is the length scale of \( f \) such that \( |f'''(x)| \sim |f(x)| / L^3 \). What is the minimised error?

(e)* An example of (Richardson) extrapolation: Show that
\[
\delta_h f(x) = f'(x) + \frac{h^2}{6} f'''(x) + \frac{h^4}{120} f^{(5)}(x) + \ldots.
\]

Hence show that
\[
\frac{100\delta_{h/10} f(x) - \delta_h f(x)}{99} = f'(x) - \frac{h^4}{100 \cdot 120} f^{(5)}(x) + \ldots.
\]

Use this extrapolation formula to improve your estimates in Part (a).

Solution.

(a) See set1q5.ans.
(b) Expand \( f(x + h) \) and \( f(x - h) \) in Taylor series to 6 terms (not 3) about \( h = 0 \),

\[
\begin{align*}
    f(x + h) &= f(x) + h f'(x) + \frac{1}{2} h^2 f''(x) + \frac{1}{6} h^3 f'''(x) + \frac{1}{24} h^4 f^{(4)}(x) + \frac{1}{120} h^5 f^{(5)}(x) + \ldots \\
    f(x - h) &= f(x) - h f'(x) + \frac{1}{2} h^2 f''(x) - \frac{1}{6} h^3 f'''(x) + \frac{1}{24} h^4 f^{(4)}(x) - \frac{1}{120} h^5 f^{(5)}(x) + \ldots .
\end{align*}
\]

Subtracting and dividing by \( 2h \), the even order derivatives cancel, giving

\[
\delta_h f(x) = f'(x) + \frac{1}{6} h^2 f'''(x) + \frac{1}{120} h^4 f^{(5)}(x) + \ldots . \tag{1}
\]

Thus the truncation error in approximating \( f'(x) \) by \( \delta_h f(x) \) is

\[
f'(x) - \delta_h f(x) = -\frac{1}{6} h^2 f'''(x) - \frac{1}{120} h^4 f^{(5)}(\eta) .
\]

(c) Assume \( f \) is well-conditioned so that \( f(x - h) \) and \( f(x + h) \) can be evaluated with a relative error at most \( \epsilon_{\text{mach}} \). In the worst case the errors in \( f(x - h) \) and \( f(x + h) \) add absolutely, giving an absolute error in \( f(x + h) - f(x - h) \) of \( (|f(x - h)| + |f(x + h)|) \epsilon_{\text{mach}} \). The round-off error in \( h \) is at worst \( \epsilon_{\text{mach}} \). Since the relative error of the quotient of two numbers is at worst the sum of the relative errors of the numbers, the relative error in the central difference formula is at worst,

\[
\frac{|f(x - h)| + |f(x + h)|}{|f(x + h) - f(x - h)|} \epsilon_{\text{mach}} + \epsilon_{\text{mach}} .
\]

Hence the round-off error in the central difference formula is at worst

\[
\left\{ \frac{|f(x - h)| + |f(x + h)|}{h} + |\delta_h f(x)| \right\} \epsilon_{\text{mach}} .
\]

The last term in the brackets is negligible compared to the first and second terms, and \( f(x + h) \approx f(x - h) \approx f(x) \). Thus the absolute round-off error in \( \delta_h f(x) \) is at worst

\[
\frac{\epsilon_{\text{mach}} |f(x)|}{h} .
\]

(d) The total error in approximating \( f'(x) \) by the central difference formula \( \delta_h f(x) \) is bounded by

\[
\frac{1}{6} h^2 |f'''(x)| + \frac{\epsilon_{\text{mach}} |f(x)|}{h} + \frac{1}{120} h^4 |f^{(5)}(\eta)| .
\]

The minimum of \( w = ax^2 + bx^{-1} \) is \( 3^{\frac{3}{2}} ab^2 / 2 \) at \( x = \sqrt[3]{b / 2a} \). Hence the minimum of

\[
\frac{1}{6} h^2 |f'''(x)| + \frac{\epsilon_{\text{mach}} |f(x)|}{h}
\]
occurs at
\[ h = \sqrt[3]{3\epsilon_{\text{mach}} |f(x)| / |f'''(x)|}. \]

If \( L \) is the length scale of \( f(x) \) for which \( f'''(x) \sim f(x)/L^3 \), then \( h \sim \sqrt[3]{3\epsilon_{\text{mach}} L} \) and the minimum is \( \sim |f(x)|^{2/3} / L \).

(e) Replacing \( h \) by \( h/10 \) in equation (1),
\[ \delta_{h/10} f(x) = f'(x) + \frac{1}{6} \left( \frac{h}{10} \right)^2 f'''(x) + \frac{1}{120} \left( \frac{h}{10} \right)^4 f^{(5)}(x) + \ldots. \] (2)

Eliminating the terms in equations (1) and (2) which contain \( f'''(x) \), gives
\[ \frac{100\delta_{h/10} f(x) - \delta_h f(x)}{99} = f'(x) - \frac{h^4}{100.120} f^{(5)}(x) + \ldots. \]

See also set1q5.ans.

6. Give floating point examples in which the associative law of multiplication and the distributive law fail to hold.

Solution.

Let the floating-point arithmetic be one significant figure with rounding.

\[ (0.7 \times 0.2) \times 0.2 = (0.14) \times 0.2 = 0.1 \times 0.2 = 0.02 \]

but
\[ 0.7 \times (0.2 \times 0.2) = 0.7 \times 0.04 = 0.028 = 0.03 \neq 0.02. \]
\[ 0.4 \times (0.4 + 0.4) = 0.4 \times 0.8 = 0.32 = 0.3 \]

but
\[ 0.4 \times 0.4 + 0.4 \times 0.4 = 0.16 + 0.16 = 0.2 + 0.2 = 0.4 \neq 0.3. \]

7. Is the initial value problem
\[ y'' - y = 0, \quad y(0) = 1, \quad y'(0) = -1 \]

well-conditioned or ill-conditioned?

Solution.

This problem is ill-conditioned. The general solution is
\[ y = c_1 e^x + c_2 e^{-x}. \]

The initial conditions imply that \( c_1 = 0 \) and \( c_2 = 1 \), and hence \( y = e^{-x} \).

However, if the initial conditions are perturbed slightly to \( y(0) = 1 + \delta_1 \) and \( y'(0) = -1 + \delta_2 \), then \( c_1 = \frac{1}{2} (\delta_1 + \delta_2) \) and \( c_2 = 1 + \frac{1}{2} (\delta_1 - \delta_2) \). The term \( c_1 e^x \) eventually grows and swamps the true solution, no matter how small \( \delta_1 \) and \( \delta_2 \) are. Even if there were no errors in \( c_1 \) and \( c_2 \), if errors are introduced later a term depending on \( e^x \) will be introduced.
An ill-conditioned problem: The general solution of the three term linear recurrence relation

\[ \phi_{k+1}(x) = \frac{2k}{x} \phi_k(x) - \phi_{k-1}(x) \]

is

\[ \phi_k(x) = c_1 J_k(x) + c_2 Y_k(x), \]

where \( c_1 \) and \( c_2 \) are independent of \( k \), and \( J_k(x) \), \( Y_k(x) \) are \( k \)th order Bessel functions of the first and second kinds, respectively. If \( \phi_0(x) = J_0(x) \) and \( \phi_1(x) = J_1(x) \) the solution is \( \phi_k(x) = J_k(x) \).

(a) Write a program using the recurrence relation with \( J_0(1) = 0.7651976865 \) and \( J_1(1) = 0.4400505857 \) to calculate \( J_k(1) \) for \( k = 2, \ldots, 20 \).

(b) Use the behaviour of \( J_k(1) \) and \( Y_k(1) \) for large \( k \) (Abramowitz and Stegun, Handbook of Mathematical Functions), namely

\[ J_k(x) \sim \frac{1}{\sqrt{2\pi k}} \left( \frac{ex}{2k} \right)^k, \quad Y_k(x) \sim -\frac{\sqrt{2 \pi k}}{\sqrt{2k}} \left( \frac{ex}{2k} \right)^{-k}, \]

to explain your results in Part (a).

Solution.

(a) This is a linear three term recurrence relation. \( J_k(x) \) and \( Y_k(x) \) are linearly independent solutions. Hence the general solution is

\[ \phi_k(x) = c_1 J_k(x) + c_2 Y_k(x), \]

where \( c_1 \) and \( c_2 \) are constants.

(b) The conditions \( \phi_0(1) = J_0(1) \) and \( \phi_1(1) = J_1(1) \) determine \( c_1 \) and \( c_2 \), namely \( c_1 = 1 \) and \( c_2 = 0 \). See set1q8.ans.

(c) Thus \( Y_k(1) \) increases rapidly with \( k \), whereas \( J_k(1) \) decreases rapidly. The exact solution to Part (a) is \( J_k(1) \), but small errors in the values of \( J_0(1) \) and \( J_1(1) \) due to round-off mean that \( c_2 \) is not exactly zero. Eventually \( c_2 Y_k(1) \) swamps \( J_k(1) \). The problem of calculating \( J_k(x) \) using the recurrence relation and starting with \( J_0(x) \) and \( J_1(x) \) is ill-conditioned. The numerical method it provides for evaluating \( J_k(x) \) is unstable. The recurrence relation can actually be used backward starting with any non-zero values for \( J_n(x) \) and \( J_{n-1}(x) \), where \( n \) is sufficiently large (depending on the desired accuracy). This method gives the correct ratios \( J_k(x)/J_{k-1}(x) \) if \( k \) is sufficiently smaller than \( n \). If \( J_0(x) \) is known \( J_k(x) \) can then be found. The \( Y_k(x) \) can be found similarly by using the recurrence relation in the forward direction. These are commonly used methods to evaluate these functions.