Tutorial 5 Solutions 11.3YC1 **J C Eilbeck, November 26, 1999**

Question 1

$x = \pm 0.d_1d_2d_3d_4d_5 \times 10^E$
Normalised $\Rightarrow 1 \le d_1 \le 9, \quad 1 \le d_j \le 9, j = 2, \ldots, 5$. $700 \le x < 900 \Rightarrow E = 2$.
Numbers are $0.70000 \times 10^2, 0.70001 \times 10^2, \ldots, 0.89998 \times 10^2, 0.89999 \times 10^2$, i.e. $d_1 = 7, 8, \quad d_j = 0, 1, \ldots, 9$.

digit	d_1	d_2	d_3	d_4	d_5	Total
choices	2	10	10	10	10	2×10^4

So 2×10^4 different numbers.

Question 2

$x = \pm 0.d_1d_2d_3 \times 10^{\pm e_1e_2}$
Normalised $\Rightarrow 1 \le d_1 \le 9, \quad 1 \le d_2, d_3 \le 9$
Closest to 0 $\Rightarrow$ smallest values of $d_1d_2d_3$ and most $-$ve exponent, i.e. $x = \pm 0.100 \times 10^{-99}$ (two choices) $= \pm 10^{-100}$
Biggest $\Rightarrow$ biggest $d_1d_2d_3$ and most +ve exponent, i.e. $x = \pm 0.999 \times 10^{99} \approx 10^{99}$

Question 3

(ii)(a) $14.1 \oplus 0.0981 = \text{fl}(\text{fl}(14.1) + \text{fl}(0.0981)) = \text{fl}(14.1 + 0.0981) = \text{fl}(14.1981) = 14.1$. (Exact = 14.1981).
(b) $0.0218 \otimes 179 = \text{fl}(\text{fl}(0.0218) \times \text{fl}(179)) = \text{fl}(0.0218 \times 179) = \text{fl}(3.9022) = 3.90$. (Exact = 3.9022).
(c)$(164 \oplus 0.913) \ominus (143 \oplus 21.0) = 164 \ominus 164 = 0$ (Exact$= 0.913$)
(d)$(164 \ominus 143) \oplus (0.913 \ominus 21.0) = 21.0 \oplus (-20.0) = 1.00$ (Exact$= 0.913$)
Parts (c) & (d) show that floating point arithmetic is not associative.

	exact	chop	round
a	14.1981	14.1	14.2
b	3.9022	3.9	3.9
c	0.913	0.00	1.00
d	0.913	1.00	0.900
e	-0.003198...	-0.004	-0.003

Each operation must be done in floating point!.

Question 4

$13.11 \otimes (31.69x + 14.31y) = 13.11 \otimes 45.0$
$\Rightarrow 415.4x + 187.6y = 589.9 \qquad$ (A)
$31.69 \otimes (13.11x + 5.89y) = 19.00 \otimes 31.69$
$\Rightarrow 415.4x + 186.6y = 602.1 \qquad$ (B)
(A)-(B) $\Rightarrow (187.6 \ominus 186.6)y = 589.9 \ominus 602.1$
$\Rightarrow 1.000y = -12.2$
$\Rightarrow y = -12.20$
(B) $\Rightarrow x = (602.1 - 186.6 \otimes (-12.20))/415.4$
$= 2878/415.4 = 6.928$
(Exact solution is $y = -12.8, x = 7.2$).

The problem is loss of accuracy on the subtraction $187.6 - 186.6$ and $589.9 - 602.1$, which are reduced to 2 and 3 significant digits respectively.

Question 5

(i) Set $y = e^x$: so $y = \exp(1.53) = 4.62$ (3 digit rounding).
Direct calculation gives $1.01y^4 - 4.62y^3 - 3.11y^2 + 12.2y - 1.99 = -6.79$.
(ii) Polynomial nesting (Horner's rule) gives -7.07.

(Remember to use floating point arithmetic at each step).

Question 6

Let $y = 0.d_1 d_2 \dots d_k d_{k+1} \dots \times 10^n$, normalised so that $1 \le d_1 \le 9$, $0 \le d_j \le 9$, $j \ge 2$. k digit rounding depends on the value of the d_{k+1} digit.
(i) If $d_{k+1} < 5$ then chop y at kth digit

$$fl(y) = 0.d_1 d_2 \dots d_k d_{k+1} \times 10^n$$

(ii) If $d_{k+1} \ge 5$ then

$$fl(y) = 0.d_1 d_2 \dots d_k d_{k+1} \times 10^n + 0.00 \dots 01 \times 10^n$$

Case (i):

$$\begin{aligned} |y - fl(y)| &= 0.00 \dots 0 \dots d_k d_{k+1} \dots \times 10^n \\ &= 0.d_k d_{k+1} \dots \times 10^{n-k} \end{aligned}$$

So

$$\begin{aligned} |y - fl(y)|/|y| &\le \text{biggest } |y - fl(y)|/\text{smallest } |y| \\ &< 0.5 \times 10^{n-k}/0.100 \dots \times 10^n = 5 \times 10^{-k} \end{aligned}$$

as required.
Case (ii):

$$\begin{aligned} |y - fl(y)| &= |0.d_k d_{k+1} \dots \times 10^{n-k} - 1 \times 10^{n-k}| \\ &\le 0.5 \times 10^{n-k} \end{aligned}$$

So

$$|y - fl(y)|/|y| \le 0.5 \times 10^{n-k}/0.1 \times 10^n = 5 \times 10^{-k}$$

as required.

Question 7

When $x \approx 0$, $\cos x \approx 1$ and $e^x \approx 1 + x$ so both the numerator and denominator are affected by cancellation rounding errors. When x is small enough, the computed version of $\cos x$ **is** 0 and the computed version of e^x **is** the same as $1 + x$ and so we get 0/something, something/0 or 0/0 - all are completely wrong.

To cure this, Taylor expand the numerator and denominator independently about $x = 0$ to get

$$f(x) = \frac{e^x - 1 - x}{1 - \cos x} = \frac{1 + x + x^2/2 + x^3/3! + \dots - 1 - x}{1 - 1 + x^2/2 - x^4/4! \cdots}$$

which reduces to

$$f(x) = \frac{x^2/2 + x^3/3! + x^4/4! \cdots}{x^2/2 - x^4/4! \cdots} = \frac{1 + x/3 + x^2/12 \cdots}{1 - x^2/12 \cdots} .$$

Only need to keep enough terms in the expansion to influence the result. e.g. if $x = 10^{-4}$ and we work to 6 significant figures, then the x^2 terms can only influence the 8th or 9th decimal place and can be ignored.