

Chapter 2

Number System and Errors

Exercise set 2.1

1. (a) $(3.4375)_{10} = (11.0111)_2 = (0.110111) \times 2^2$
(b) $(12.875)_{10} = (1100.111)_2 = (0.1100111) \times 2^4$.

2.

- (a) $(5.25)_{10} = (101.01)_2 = (0.10101)_2 \times 2^3$, since $(3) = (11)_2$ the internal representation is

$$(5.25)_{10} = (000000111010100...0)_2$$

- (b) $(-3.84375)_{10} = (11.11011)_2 = (0.1111011)_2 \times 2^2$, since $(2) = (10)_2$ the internal representation is

$$(-3.84375)_{10} = (1000001011110110....0)_2$$

3. (a) The largest number is $(0111011111)_2 = 2^7 \times (2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} + 2^{-5})_{10} = 2^7 \times 0.96875 = (124)_{10}$
(b) The exponent is $+(101)_2 = 5$. The mantissa is positive and is $= (10110)_2 = 2^{-1} + 2^{-3} + 2^{-4} = (0.6875)_{10}$. Hence,

$$(0101010110)_2 = 2^5 \times 0.6875 = (22)_{10}$$

4. (a) $351.78125, (492)_{10} = (101011111)_2, (0.78125)_{10} = (0.110010)_2$
 $(351.78125)_{10} = (101011111.110010)_2 = (1.01011111110010)_2 \times 2^8$
exponent $= 127 + 8 = (135)_{10} = (10000111)_2$
 $(351.78125)_{10} = (0\ 10\ 000111\ 01011111110010\0)_2$

$$\begin{aligned}
\text{(b)} \quad (0.5625)_{10} &= (0.1001)_2 = (1.001 \times 2^{-1}) \\
\text{exponent} &= 127 - 1 = (126)_{10} = (01111110)_2 \\
-(0.5625)_{10} &= (1 \ 01111110 \ 0010000...0)_2
\end{aligned}$$

5.

$$\text{(a)} \quad \text{The exponent } (1010)_2 = 10 \text{ and the mantissa } (110101)_2 = 2^{-1} + 2^{-2} + 2^{-4} + 2^{-6} = (0.828 \ 125)_{10}$$

$$a = -0.828 \ 125 \times 2^{10} = (-848)_{10}$$

$$\text{(b)} \quad \text{The exponent } (110)_2 = 6 \text{ and the mantissa } (1000101)_2 = 2^{-1} + 2^{-5} + 2^{-7} = (0.539 \ 062 \ 5)_{10}$$

$$b = 0.539 \ 062 \ 5 \times 2^6 = (34.5)_{10}$$

$$\text{(c)} \quad \text{The exponent } (1110)_2 = 14 \text{ and the mantissa } (11011)_2 = 2^{-1} + 2^{-2} + 2^{-4} + 2^{-5} = (0.843 \ 75)_{10}$$

$$c = -0.843 \ 75 \times 2^{14} = (-13824)_{10}$$

6.

$$\text{(a)} \quad 2^{-1022} \approx 2.2 \times 10^{-308}$$

$$\text{(b)} \quad (2 - 2^{-52})2^{1023} \approx -1.8 \times 10^{308}$$

$$\text{(c)} \quad 2 \times 2046 \times 2^{52} + 1 \approx 1.8 \times 10^{19}$$

Exercise set 2.2

1. (a) i) $\frac{1}{3} + \frac{7}{4} \approx 0.333 \ 3 + 1.750 = 2.083$ ii) $\frac{1}{3} + \frac{7}{4} \approx 0.333 \ 3 + 1.750 = 2.083$
(b) i) $\left(\frac{2}{3} - \frac{4}{7}\right) + \left(\frac{8}{5} - \frac{13}{7}\right) \approx (0.6667 - 0.5714) + (1.600 - 1.857) = 0.0953 - 0.2570 = -0.161 \ 7$
ii) $\approx (0.6666 - 0.5714) + (1.600 - 1.857) = -0.161 \ 8$
(c) i) $\left(\frac{5}{3} \times \frac{2}{7}\right) + \frac{2}{3} \approx (1.667 \times 0.285 \ 7) + 0.666 \ 7 = 0.476 \ 3 + 0.666 \ 7 = 1.143$
ii) $\approx (1.666 \times 0.285 \ 7) + 0.6666 = 0.475 \ 9 + 0.6666 = 1.142$

2. The roots of the quadratic equation are given by

$$\frac{-6 \pm \sqrt{36 - 4(1)(2)}}{2(4)} = \frac{-6 \pm \sqrt{28}}{6} \approx \frac{-6 \pm 5.292}{6}$$

Hence $x_1 = -1.882$ and $x_2 = -0.1180$.

3.

$$\begin{aligned}
 \text{(a) Error} &= \frac{0.86429168+0.86431221-[fl(0.86439868)+fl(0.86433221)]}{0.86439868+0.86433221} \\
 &= \frac{0.86429168+0.86431221-[0.86429+0.86431]}{0.86439868+0.86433221} \\
 &= \frac{1.72873089-1.7286}{1.72873089} = 7.571450291 \times 10^{-5}
 \end{aligned}$$

$$\begin{aligned}
 \text{(b) Error} &= \frac{0.86429168 \times 0.86431221 - [fl(0.86439868) \times fl(0.86433221)]}{0.86439868 + 0.86433221} \\
 &= \frac{0.86429168 \times 0.86431221 - [0.86429 \times 0.86431]}{0.86439868 + 0.86433221} \\
 &= \frac{0.747017852025413 - 0.74701}{0.747017852025413} = 1.051116167 \times 10^{-5}
 \end{aligned}$$

4. We have $x_1 = X_1 - \epsilon_1$ and $x_2 = X_2 - \epsilon_2$, substitute to get

$$\frac{X_1 X_2 - (X_1 - \epsilon_1)(X_2 - \epsilon_2)}{X_1 X_2} = \frac{\epsilon_1}{X_1} + \frac{\epsilon_2}{X_2} - \frac{\epsilon_1 \epsilon_2}{X_1 X_2} \approx \frac{\epsilon_1}{X_1} + \frac{\epsilon_2}{X_2}.$$

Using the fact that $\epsilon_1 \epsilon_2 / X_1 X_2 \ll 1$. The symbol " \ll " means "much less than".

5. We have

$$\begin{aligned}
 \frac{0.67323 \times (12.751 + 12.687)}{12.751^2 - 12.687^2} &= 10.51921875 \\
 \frac{0.673 \times (12.8 + 12.7)}{12.8^2 - 12.7^2} &= \frac{0.673 \times (25.5)}{163 - 161} = \frac{17.2}{2} = 8.6
 \end{aligned}$$

6. error = 0.00159265358979... $\approx 0.15926 \times 10^{-2}$

7.

(a) $p(1.07) = -1.31$, The true solution is $-1.297...$

(b) $p(1.07) = -1.33$, The true solution is $-1.297...$

8.

(a) $0.001 \times 150 = 0.15 \implies 149.85 < p < 150.15$

(b) $0.001 \times 1500 = 1.5 \implies 1498.5 < p < 1501.5$

9.

(a) $N = 6$ the correct value is 2.45

(b) absolute = 0.01

(c) relative = 0.00408

10.

(a) 2.70 rel. error = 0.0067

(b) 2.71 rel. error = 0.0030

11.

(a) $a_2 = 50,000 + a_0$. and for all values of a_0 the result will always be $a_2 = 50,000$.

(b) The final value is independent of a_0 because of the large multiplier on a_1 compared to the precision of the computation.

12.

(a) $x = 2^e(1 + g) = 2^{2\bar{e}}(1 + \bar{g})$. If e is even then $e = 2\bar{e}$ and so $1 \leq 1 + \bar{g} = 1 + g < 2$.

(b) If e is odd then $2\bar{e} = (e - 1)$ and so $2 \leq 1 + \bar{g} = 2(1 + g) < 4$. altogether $1 \leq 1 + \bar{g} < 4$.

Exercise set 2.3

1. $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$

$$e^{0.5} \approx 1 + 0.5 + \frac{0.5^2}{2!} = 1.625, \text{ error} = e^{0.5} - 1.625 = 0.0237213.$$

2. $\ln(1 - x) = -x - \frac{1}{2}x^2 - \frac{1}{3}x^3 + O(x^4)$ and $\ln(1 - 0.5) = -0.5 - \frac{1}{2}0.5^2 - \frac{1}{3}0.5^3 \approx -0.66666667$

3. We have $\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 - \frac{1}{720}x^6 + \frac{1}{40320}x^8 + O(x^{10})$

$$\cos x = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 - \frac{1}{720}x^6 + \dots$$

To get an accuracy of 10^{-4} when $x = 1.7$, we must have

$$\left| \frac{1.7^{2n+2}}{(2n+2)!} \right| \leq 10^{-4}$$

Exercise set 2.4

1. (a) $\text{Bound} = [0, 1] \cdot [0, 1] + 2[0, 1] + [1, 1] = [0, 1] + [0, 2] + [1, 1] = [1, 4]$
 (b) $\text{Bound} = [-1, 1] \cdot [-1, 1] + [1, 2] = [-1, 1] + [1, 2] = [0, 2]$
 (c) $\text{Bound} = [0, 1]/([1, 1] + [0, 1]) = [0, 1]/[1, 2] = [0, 1] \cdot [1/2, 1] = [1/2, 1]$
2. we have $\chi[2, 4] = 1/2 > \chi[-2, 1] = -1/2$. Thus, the solution set of the equation is

$$S = [-2, 1]/[2, 4] = [-2, 1] \cdot [1/4, 1/2] = [-1, 1/2]$$

| case | condition | result |
|------|------------------------|--------------------|
| 1 | $a > 0, c > 0$ | $[ac, bd]$ |
| 2 | $a > 0, d < 0$ | $[bc, ad]$ |
| 3 | $b < 0, c > 0$ | $[ad, bc]$ |
| 4 | $b < 0, d < 0$ | $[bd, ac]$ |
| 5 | $a < 0 < b, c > 0$ | $[ad, bd]$ |
| 6 | $a < 0 < b, d < 0$ | $[bc, ac]$ |
| 7 | $a > 0, c < 0 < d$ | $[bc, bd]$ |
| 8 | $b < 0, c < 0 < d$ | $[ad, ac]$ |
| 9 | $a < 0 < b, c < 0 < d$ | $[ad, bc, ac, bd]$ |

3. (a) $[(1)(3), (2)(4)] = [3, 8]$
 (b) $[(2)(-4), (-1)(-4)] = [-8, 4]$
5. (a) $[0, 2] + [2, 3] = [0 + 2, 2 + 3] = [2, 5]$,
 (b) $[2.0, 2.2] - [1.0, 1.1] = [2.0 - 1.1, 2.2 - 1.0] = [0.9, 1.1]$,
 (c) $[-5, -3] \cdot [-7, 4] = [\min\{35, 21, -12, -20\}, \max\{35, 21, -12, -20\}] = [-20, 35]$,
 (d) $[-2, 1] \cdot [-4.4, 2.1] = [\min\{8.8, -4.2, -4.4, 2.1\}, \max\{8.8, -4.2, -4.4, 2.1\}] = [-4.4, 8.8]$
 (e) $[-2, 2] \div [-3/2, -1] = [-2, 2] \cdot [-1, -2/3] = [-2, 2]$,
 (f) $[3, 3] \cdot [-2, 4] = [\min\{-6, 12\}, \max\{-6, 12\}] = [-6, 12]$,
 (g) $[-1, 0]/[1, 2] + [-2, 3] \cdot [4, 5] = [-1, 0] \cdot [1/2, 1] + [-10, 15] = [-1, 0] + [-10, 15] = [-11, 15]$