1.2

Convert to Floating-number

To calculate c, take the base as 2

10 9 8 7 6 5 4 3 2 1 0
10000000011

$1\times 2^{10} + 1\times 2^{1} + 1\times 2^{0}$

$= 2^{10} + 2^{1} + 2^{0} = 1027$

to calculate f, take the base as $\frac{1}{2}$

1 2 3 4 5 6 7 8 9 10 11 12
101110010001 00...

$f = (\frac{1}{2})^1 + (\frac{1}{2})^3 + (\frac{1}{2})^4 + (\frac{1}{2})^5 + (\frac{1}{2})^8 + (\frac{1}{2})^{12} = \frac{2961}{4096}$

$2^{c-1023}(1+f) = 2^{1027-1023}(1+\frac{2961}{4096}) = 27.56640$

$(-1)^0 = 1$ (0) pos $\quad (-1)^1 = -1$ (1) neg

$$\cancel{(-1)^s}\, 2^{c-1023}(1+f).$$

Right to left $\quad$ Left to right

Sign (s) ← 0 | 10000000011 | 1011100100010000000000000000000000000000000000000000. (52 bits)

0(+) $\quad$ 1(-) $\quad$ 1 bit $\quad$ 11 bits (c) $\quad$ f

The leftmost bit is $s = 0$, which indicates that the number is positive. The next 11 bits, 10000000011, give the characteristic and are equivalent to the decimal number

$$c = 1\cdot 2^{10} + 0\cdot 2^9 + \cdots + 0\cdot 2^2 + 1\cdot 2^1 + 1\cdot 2^0 = 1024 + 2 + 1 = 1027.$$

$$f = 1\cdot\left(\frac{1}{2}\right)^1 + 1\cdot\left(\frac{1}{2}\right)^3 + 1\cdot\left(\frac{1}{2}\right)^4 + 1\cdot\left(\frac{1}{2}\right)^5 + 1\cdot\left(\frac{1}{2}\right)^8 + 1\cdot\left(\frac{1}{2}\right)^{12}.$$

As a consequence, this machine number precisely represents the decimal number

$$(-1)^s 2^{c-1023}(1+f) = (-1)^0\cdot 2^{1027-1023}\left(1 + \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{256} + \frac{1}{4096}\right)\right)$$

$$= 27.56640625.$$

Actual value = p Approximate value = p^* $p - p^*$ = actual error (A.E.) Absolute error = $|p - p^*|$ Relative error = $\dfrac{\text{Absolute error}}{|p|}$

Suppose that p^* is an approximation to p. The **actual error** is $p - p^*$, the **absolute error** is $|p - p^*|$, and the **relative error** is $\dfrac{|p - p^*|}{|p|}$, provided that $p \neq 0$. ■

at most $\leq$
at least $\geq$
more than $>$
less than $<$

EXERCISE SET 1.2

① ⓐ $A.E. = |p - p^*| = \pi - \frac{22}{7} = 0.0012644$

$R.E. = \dfrac{|\pi - \frac{22}{7}|}{|\pi|} = 0.0004$

1. Compute the absolute error and relative error in approximations of p by p^*.

a. $p = \pi$, $p^* = 22/7$ **b.** $p = \pi$, $p^* = 3.1416$

c. $p = e$, $p^* = 2.718$ **d.** $p = \sqrt{2}$, $p^* = 1.414$

2. Compute the absolute error and relative error in approximations of p by p^*.

a. $p = e^{10}$, $p^* = 22000$ **b.** $p = 10^{\pi}$, $p^* = 1400$

c. $p = 8!$, $p^* = 39900$ **d.** $p = 9!$, $p^* = \sqrt{18\pi}(9/e)^9$

3. Suppose p^* must approximate p with relative error at most 10^{-3}. Find the largest interval in which p^* must lie for each value of p.
less than, greater to

a. 150 $= p$ **b.** 900 $= p$

c. 1500 $= p$ **d.** 90 $= p$

4. Find the largest interval in which p^* must lie to approximate p with relative error at most 10^{-4} for each value of p.

a. π **b.** e

c. $\sqrt{2}$ **d.** $\sqrt[3]{7}$

5. Perform the following computations (i) exactly, (ii) using three-digit chopping arithmetic, and (iii) using three-digit rounding arithmetic. (iv) Compute the relative errors in parts (ii) and (iii).

a. $\dfrac{4}{5} + \dfrac{1}{3}$ **b.** $\dfrac{4}{5} \cdot \dfrac{1}{3}$

c. $\left(\dfrac{1}{3} - \dfrac{3}{11}\right) + \dfrac{3}{20}$ **d.** $\left(\dfrac{1}{3} + \dfrac{3}{11}\right) - \dfrac{3}{20}$

③ ⓐ $R.E. \leq 10^{-3}$

$\dfrac{|p - p^*|}{|p|} \leq 10^{-3}$

$\dfrac{|150 - p^*|}{|150|} \leq 10^{-3}$

$\dfrac{|150 - p^*|}{150} \leq 10^{-3}$

$|150 - p^*| \leq 10^{-3} \times 150$

$-10^{-3} \times 150 \leq 150 - p^* \leq 10^{-3} \times 150$
$-150 \qquad -150 \qquad -150$

$-150 - 150 \times 10^{-3} \leq -p^* \leq -150 + 150 \times 10^{-3}$ $\times -1$

$\times -1$ $150 - 150 \times 10^{-3} \leq p^* \leq 150 + 150 \times 10^{-3}$

or $-150.15 \leq -p^* \leq -149.85$

$149.85 \leq p^* \leq 150.15$ belongs to $p^* \in [149.85, 150.15]$

④ ⓑ $RE \leq 10^{-4}$

$\dfrac{|p - p^*|}{|p|} \leq 10^{-4}$

$\dfrac{|e - p^*|}{|e|} \leq 10^{-4}$

$\dfrac{|e - p^*|}{e} \leq 10^{-4}$

$|e - p^*| \leq 10^{-4} \cdot e$

$-e \times 10^{-4} \leq e - p^* \leq e \cdot 10^{-4}$
$-e \qquad -e \qquad -e$

$-e - 10^{-4}e \leq -p^* \leq -e + 10^{-4}e$

$-2.71855 \leq -p^* \leq -2.71801$

$2.71801 \leq p^* \leq 2.71855$

$p \in [2.71801, 2.71855]$

⑤ a) i) $\frac{4}{5} + \frac{1}{3} = \frac{17}{15}$

ii) $\frac{4}{5} = 0.\underline{800}$

↳ not a nonzero, just add 2 zeros

If the decimal place is 0.08

then the answer will be $0.\underline{0}800$ (for 3 digits)

leading zero isn't counted

$\frac{1}{3} = 0.333$

$\frac{4}{5} + \frac{1}{3} = 0.800 + 0.333 = 1.13$ (3 digits)

iii) $\frac{4}{5} = 0.800$

$\frac{1}{3} = 0.333$

$\frac{4}{5} + \frac{1}{3} = 0.800 + 0.333 = 1.13$ (3 digits)

19. Use the 64-bit-long real format to find the decimal equivalent of the following floating-point machine numbers.

a. (pos) 0 / 10000001010 / 1001001100000000000000000000000000000000000000000000

10 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8

b. (neg) 1 / 10000001010 / 1001001100000000000000000000000000000000000000000000

10 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8

c. (pos) 0 / 01111111111 / 0101001100000000000000000000000000000000000000000000

1 2 3 4 5 6 7 8 52

d. (pos) 0 / 01111111111 / 0101001100000000000000000000000000000000000000000001

(b) $C = 2^1 + 2^3 + 2^{10} = 1034$

$f = \left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^7 + \left(\frac{1}{2}\right)^8 = \frac{147}{256}$

floating num: $-\left(2^{1034-1023}\right)\left(1+\frac{147}{256}\right) = -3224$

(a) exactly like (b) but pos as s is 0

so floating num: $\left(2^{1034-1023}\right)\left(1+\frac{147}{256}\right) = 3224$

(c) $C = 2^0+2^1+2^2+2^3+2^4+2^5+2^6+2^7+2^8+2^9 = 1023$

$f = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^7 + \left(\frac{1}{2}\right)^8 = \frac{83}{256}$

floating num: $2^{1023-1023}\left(1+\frac{83}{256}\right) = \frac{339}{256}$

(d) $C = 2^0+2^1+2^2+2^3+2^4+2^5+2^6+2^7+2^8+2^9 = 1023$

$f = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^7 + \left(\frac{1}{2}\right)^8 + \left(\frac{1}{2}\right)^{52} = \frac{83}{256}$

floating num: $2^{1023-1023}\left(1+\frac{83}{256}\right) = \frac{339}{256} = 1.324$

1.3

☆ You have to apply Taylor's theory

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots \qquad \cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots$$

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

6. Find the rates of convergence of the following sequences as $n \to \infty$.

a. $\lim_{n\to\infty} \sin \frac{1}{n} = 0$

b. $\lim_{n\to\infty} \sin \frac{1}{n^2} = 0$

c. $\lim_{n\to\infty} \left(\sin \frac{1}{n}\right)^2 = 0$

d. $\lim_{n\to\infty} [\ln(n+1) - \ln(n)] = 0$

7. Find the rates of convergence of the following functions as $h \to 0$.

a. $\lim_{h\to 0} \frac{\sin h}{h} = 1$

b. $\lim_{h\to 0} \frac{1-\cos h}{h} = 0$

c. $\lim_{h\to 0} \frac{\sin h - h\cos h}{h} = 0$

d. $\lim_{h\to 0} \frac{1-e^h}{h} = -1$

7 a $\frac{\sin 0}{0} = \frac{0}{0}$ (indeterminate form)

$$\lim_{h\to 0} \frac{\left(h - \frac{h^3}{3!} + \frac{h^5}{5!} - \dots\right)}{h} = \lim_{h\to 0} \frac{\cancel{h}\left(1 - \frac{h^2}{3!} + \frac{h^4}{5!} - \dots\right)}{\cancel{h}} = \lim_{h\to 0} \left(1 - \frac{h^2}{3!} + \frac{h^4}{5!} - \dots\right) = 1$$

$O(h^2)$ ↑ order of convergence

b $\lim_{h\to 0} \frac{1-\cos h}{h} = \frac{1-\cos 0}{0} = \frac{0}{0}$

$$\lim_{h\to 0} \frac{\cancel{1} - \left(\cancel{1} - \frac{h^2}{2!} + \frac{h^4}{4!} - \dots\right)}{h} = \lim_{h\to 0} \frac{\left(\frac{h^2}{2!} + \frac{h^4}{4!} - \dots\right)}{h}$$

$$= \lim_{h\to 0} \frac{\cancel{h}\left(\frac{-h}{2!} + \frac{h^3}{3!} - \dots\right)}{\cancel{h}} = \lim_{h\to 0}\left(\frac{-h}{2!} + \frac{h^3}{3!} - \dots\right) = 0$$

$O(h)$ order of convergence

d $\lim_{h\to 0} \frac{1-e^h}{h} = \frac{1-e^0}{0} = \frac{0}{0}$

$$= \lim_{h\to 0} \frac{\cancel{1} - \left(\cancel{1} + h + \frac{h^2}{2!} + \frac{h^3}{3!}\right)}{h}$$

$$= \lim_{h\to 0} \frac{-\left(h + \frac{h^2}{2!} + \frac{h^3}{3!} + \dots\right)}{h}$$

$$= \lim_{h\to 0} \frac{\cancel{h} - \left(1 + \cancel{\frac{h}{2!} + \frac{h^2}{3!} + \dots}^{0}\right)}{\cancel{h}} = -1 \qquad O(h)$$

6. Find the rates of convergence of the following sequences as $n \to \infty$. if the limit goes to infinity, the limit does not exist

a. $\lim_{n\to\infty} \sin\frac{1}{n} = 0$ $\frac{1}{\infty} = 0$

b. $\lim_{n\to\infty} \sin\frac{1}{n^2} = 0$

c. $\lim_{n\to\infty} \left(\sin\frac{1}{n}\right)^2 = 0$

d. $\lim_{n\to\infty} [\ln(n+1) - \ln(n)] = 0$

$\star$ $\sin n \le n$

$-\frac{1}{n} \le \sin\frac{1}{n} \le \frac{1}{n}$

$\star$ $\sin\frac{1}{n} \le \frac{1}{n}$

(a) $\lim_{n\to\infty} \left(\sin\frac{1}{n}\right) = 0$

$\sin\frac{1}{n} \le \frac{1}{n}$

$\lim_{n\to\infty} \sin\frac{1}{n} \le \lim_{n\to\infty} \frac{1}{n} = 0$

$\Rightarrow \lim_{n\to\infty} \left(\sin\frac{1}{n}\right) = 0$

$O\left(\frac{1}{n}\right)$

(c) $\lim_{n\to\infty} \left(\sin\frac{1}{n}\right)^2 = 0$

$\sin\frac{1}{n} \le \frac{1}{n}$

$\left(\sin\frac{1}{n}\right)^2 \le \left(\frac{1}{n}\right)^2$

$\left(\sin\frac{1}{n}\right)^2 \le \frac{1}{n^2}$

$\lim_{n\to\infty} \left(\sin\frac{1}{n}\right)^2 \le \lim_{n\to\infty} \frac{1}{n^2} = 0$

$\Rightarrow \lim_{n\to\infty} \left(\sin\frac{1}{n}\right)^2 = 0$

$O\left(\frac{1}{n^2}\right)$

(b) $\lim_{n\to\infty} \sin\frac{1}{n^2} = 0$