Solution

Q1. (a)

$$
\begin{aligned}
&(1110101.101)_2\\
&=\left(1\times 2^6+1\times 2^5+1\times 2^4+0\times 2^3+1\times 2^2+0\times 2^1+1\times 2^0+1\times 2^{-1}+0\times 2^{-2}+1\times 2^{-3}\right)_{10}\\
&=(117.625)_{10}
\end{aligned}
$$

(b) We have

$$
\begin{aligned}
&2023=2\times 1011+1 \quad\Rightarrow\quad b_0=1\\
&1011=2\times 505+1 \quad\Rightarrow\quad b_1=1\\
&505=2\times 252+1 \quad\Rightarrow\quad b_2=1\\
&252=2\times 126+0 \quad\Rightarrow\quad b_3=0\\
&126=2\times 63+0 \quad\Rightarrow\quad b_4=0\\
&63=2\times 31+1 \quad\Rightarrow\quad b_5=1\\
&31=2\times 15+1 \quad\Rightarrow\quad b_6=1\\
&15=2\times 7+1 \quad\Rightarrow\quad b_7=1\\
&7=2\times 3+1 \quad\Rightarrow\quad b_8=1\\
&3=2\times 1+1 \quad\Rightarrow\quad b_9=1\\
&1=2\times 0+1 \quad\Rightarrow\quad b_{10}=1
\end{aligned}
$$

Then $(2023)_{10}=(11111100111)_2$.

Q2. (a) Since $\beta=2$, $n=5$ and $M=4$, the floating number is represented as

$$x=\pm(0.a_1a_2a_3a_4a_5)_2\cdot 2^e$$

Because $a_1\neq 0$ and a_i can only be 0 or 1, $i=1,2,3,4,5$, for positive number, $0.a_1a_2a_3a_4a_5$ is at least 0.10000 and at most 0.11111, 2^e is at least $2^{-4}=0.0625$ and at most $2^4=16$, then we know that the smallest positive number in decimal form is $(0.10000)_2\times 0.0625=2^{-1}\times 0.0625=0.03125$, and the largest number in decimal form is $(0.11111)_2\times 16=\left(2^{-1}+2^{-2}+2^{-3}+2^{-4}+2^{-5}\right)\times 16=0.9688\times 16=15.5008$.

(b) Since $n=5$ and we have $(3)_{10}=(11.000)_2$, we can obtain the maximum number smaller than π which can be represented by the floating number is

$(3.125)_{10} = (11.001)_2 = (0.11001)_2 \cdot 2^2$, and the minimum number larger than π which can be represented by the floating number is $(3.25)_{10} = (11.010)_2 = (0.11010)_2 \cdot 2^2$, obviously, the closest floating number to π is $(3.125)_{10} = (11.001)_2 = (0.11001)_2 \cdot 2^2$.

Q3. (a) Cancellation error happens as $\cos x$ is close to -1, that is, $x = (2k+1)\pi$, $k \in \mathrm{Z}$.

To remedy this problem, the function can be transferred to

$$f(x) = 1 + \cos x = 1 + \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}.$$

(b) Cancellation error happens as $\sqrt{x^2+1}$ is close to $\sqrt{x^2+4}$, that is, x is large positive or small negative.

To remedy this problem, the function can be transferred to

$$f(x) = \sqrt{x^2+1} - \sqrt{x^2+4} = \frac{-3}{\sqrt{x^2+1} + \sqrt{x^2+4}}.$$

(c) Cancellation error happens as $\ln x$ is close to $\ln(1/x)$, that is, $x = 1$.

To remedy this problem, the function can be transferred to

$$f(x) = \ln x - \ln(1/x) = 2\ln x.$$

(d) Cancellation error happens as x is close to $\sin x$, that is, $x = 0$.

To remedy this problem, the function can be transferred to

$$f(x) = x - \sin x = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^{2k+1}}{(2k+1)!}.$$

(e) Cancellation error happens as $2\sin^2 x$ is close to 1, that is, $x = \frac{1+2k}{4}\pi$, $k \in \mathrm{Z}$.

To remedy this problem, the function can be transferred to

$$f(x) = 1 - 2\sin^2 x = \cos(2x).$$

(f) Cancellation error happens as $\ln x$ is close to 1, that is, $x = e$.

To remedy this problem, the function can be transferred to

$$f(x) = \ln x - 1 = \ln\frac{x}{e}.$$

Q4. (a) Using Taylor polynomials, we have

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \frac{1}{6}f'''(x)h^3 + O(h^4)$$

$$f(x-h) = f(x) - f'(x)h + \frac{1}{2}f''(x)h^2 - \frac{1}{6}f'''(x)h^3 + O(h^4)$$

By adding them, we obtain

$$f(x+h) + f(x-h) = 2f(x) + f''(x)h^2 + O(h^4)$$

That is

$$f(x-h) - 2f(x) + f(x+h) + O(h^4) = f''(x)h^2$$

Then we finally have

$$f''(x) = \frac{f(x-h) - 2f(x) + f(x+h) + O(h^4)}{h^2} = \frac{f(x-h) - 2f(x) + f(x+h)}{h^2} + O(h^2)$$

(b) With $h = 2^{-n}$, $n = 1, 2, 3, \ldots, 10$, the curve of error against h is plotted as below.

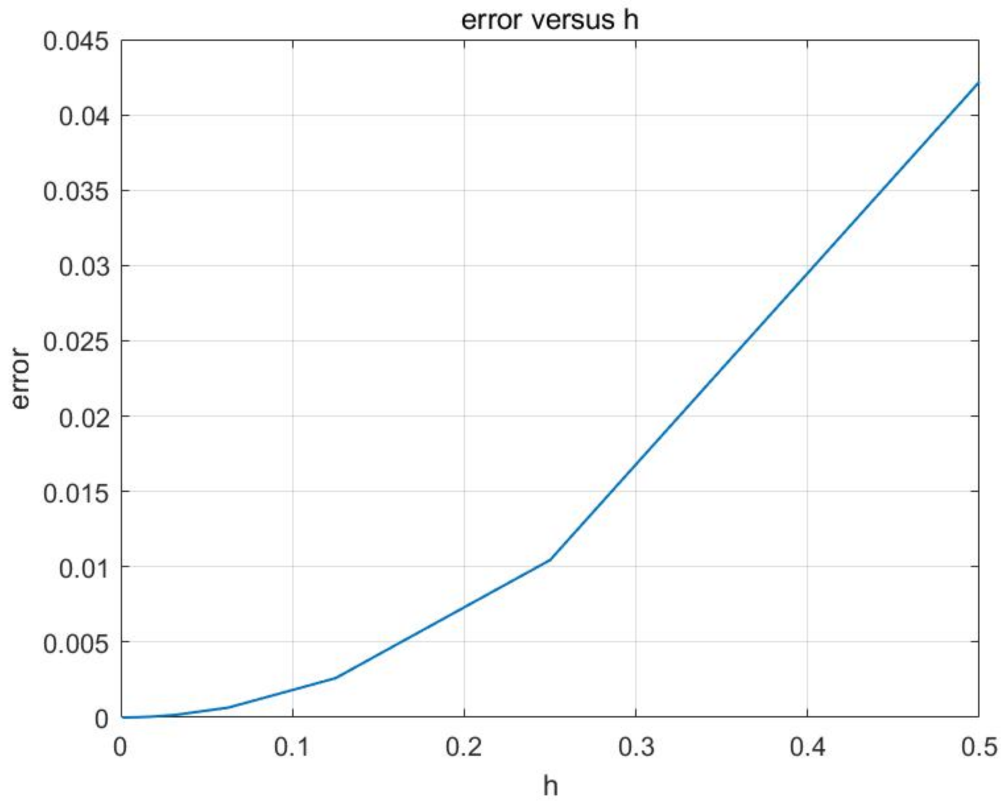

Figure.1 Curve of error against h

The table is listed as shown in Table.1.

Table.1 Relevant information

h	Df	E	E/h	E/h^2	E/h^3
2^{-1}	1.04219061098749	0.0421906109874948	0.0843812219749895	0.168762443949979	0.337524887899958
2^{-2}	1.01044926723267	0.0104492672326730	0.0417970689306921	0.167188275722769	0.668753102891074
2^{-3}	1.00260620192892	0.00260620192892347	0.0208496154313877	0.166796923451102	1.33437538760882
2^{-4}	1.00065116883507	0.000651168835069882	0.0104187013611181	0.166699221777890	2.66718754844624
2^{-5}	1.00016276836414	0.000162768364138088	0.00520858765241883	0.166674804877403	5.33359375607688
2^{-6}	1.000040690	4.069060087275	0.002604198455	0.1666687011	10.666796875

	60087	03e-05	85602	74785	1863
2^{-7}	1.000010172 55709	1.017255709001 57e-05	0.001302087307 52202	0.1666671753 62818	21.333398446 4407
2^{-8}	1.000002543 13343	2.543133433619 01e-06	0.000651042159 006465	0.1666667927 05655	42.666698932 6477
2^{-9}	1.000000635 78298	6.357829818171 01e-07	0.000325520886 690356	0.1666666939 85462	85.333347320 5566
2^{-10}	1.000000158 94568	1.589456815054 29e-07	0.000162760377 861559	0.1666666269 30237	170.66662597 6563

From the curve, it can be observed that the smaller h is, the smaller error is.

From the table, the same conclusion can be drawn, besides, since E/h^2 is relative the same, the rate of convergence is 2.

Q5. The program we write is as below

```
function d2b_function(decimal)
    if decimal == 0
        binary = '0';
    else
        binary = '';
        while decimal ~= 0
            binary = [num2str(mod(decimal, 2)) binary];
            decimal = floor(decimal / 2);
        end
    end
    disp(binary);
end
```

Using the program, we obtain the results as

(a) $(471)_{10} = (111010111)_2$;

(b) $(2016)_{10} = (11111100000)_2$.