problem 1.

a.

Our algorithm is based on Horner's Method.
We have that $a = \sum_{j=0}^{n} a_j \beta^{j+e_a}$ and $b = \sum_{i=0}^{m} b_i \beta^{i+e_b}$, then we can compute ab.

$$
\begin{aligned}
ab &= \sum_{j=0}^{n} a_j \beta^{j+e_a} \cdot \sum_{i=0}^{m} b_i \beta^{i+e_b} \\
&= \beta^{e_a+e_b} \sum_{k=0}^{N} (\sum_{j=0}^{k} a_j b_{k-j}) \beta^k \\
&= \beta^{e_a+e_b} \sum_{k=0}^{N} p^* \beta^k,
\end{aligned}
$$

here we assume $N = \max\{n, m\}$ and we don' t need to pay much attention on $\beta^{e_a+e_b}$. Therefore, we only need an algorithm to compute $\sum_{k=0}^{N} p^* \beta^k$ which is based on Horner's Method, that is :

$$
\sum_{k=0}^{N} p^* \beta^k = ((p_N^* \beta + p_{N-1}^*)\beta + \cdots)\beta + p_0^*,
$$

the algorithm should be

1. Firstly we set $k = N$;

2. Then we let $b_k = p_k^*$ and $b_{k-1} = p_{k-1}^* + b_k \beta$;

3. Set $k = k - 1$;

4. If $k \geq 0$, we still do $b_{k-1} = p_{k-1}^* + b_k \beta$, if not the algorithm is ended.
In this algorithm we need N additions and N multiplication where $N = \max\{n, m\}$,and its storage requirements are only n times the number of bits of β. Although n and m could be large number our algorithm will terminate in a finite number of steps, and that it returns the correct answer.

b.

Proof. We want to show $0 \leq ab - \tilde{p} \leq ab \cdot \beta^{k^*+3-n-m}$ in this question, where

$$
\tilde{p} = \sum_{k=k^*}^{n+m} \Big(\sum_{j=0}^{k} a_j b_{k-j} \Big) \beta^k.
$$

Remark. $\tilde{p}$ contains all digits from k^* to $m + n$, thus $ab - \tilde{p}$ contains the digits from 0 to $k^* - 1$. This implies that $ab - \tilde{p} = 0$ if all digits from 0 to $k^* - 1$ are 0 then $ab - \tilde{p} \geq 0$ always holds.

Now we'll show $0 \le ab - \tilde{p} \le ab \cdot \beta^{k^*+3-n-m}$.

$$\begin{aligned}
ab - \tilde{p} &= \sum_{k=0}^{k^*-1} \Big(\sum_{j=0}^{k} a_j b_{k-j}\Big)\beta^k, a_j, b_{k-j} \le \beta^k \\
&\le \sum_{k=0}^{k^*-1} (\beta-1)^2 k\beta^k \\
&= \beta(\beta-1)^2 \sum_{k=0}^{k^*-1} k\beta^{k-1} \\
&= \beta(\beta-1)^2 (\sum_{k=0}^{k^*-1} k\beta^{k-1})' \\
&= \beta(\beta-1)^2 (\frac{\beta^{k^*}-1}{\beta-1})' \\
&= (k^*-1)\beta^{k^*+1} - k^*\beta^{k^*} + \beta
\end{aligned}$$

We will have $(k^*-1)\beta^{k^*+1} - k^*\beta^{k^*} + \beta \le \beta^{k^*+3} \iff k^* \le 9$ if we assume $\beta = 2$, which is the smallest β. For bigger β, k^* also increase. We set $k^* \le 9$ then

$$ab \cdot \beta^{k^*+3-n-m} = \sum_{k=0}^{n+m} \Big(\sum_{j=0}^{k} a_j b_{k-j}\Big)\beta^k \cdot \beta^{k^*+3-n-m}$$

Now we consider the last term $\sum_{j=0}^{n+m} a_j b_{n+m-j}\beta^{k^*+3}$ $(**)$. Since a_j, b_i and β are all positive integers, also we know $m + n \ge k^*$, thus

$$(**) \ge \sum_{j=0}^{k^*} a_j b_{k^*-j}\beta^{k^*+3} \ge \beta^{k^*+3}.$$

Hence,

$$0 \le ab - \tilde{p} \le ab \cdot \beta^{k^*+3-n-m} \text{ always holds.}$$

□

problem 2.

a).

Let $S_n := x_1 + x_2 + \cdots + x_n$ and $\widetilde{S_n} = x_1 \oplus \cdots \oplus x_n$.
Note that $x_i \in \mathbb{R}(i = 1, \cdots, n)$, $\widetilde{S_n}$ is the pairwise summation of x_i. We also assume n is a power of 2 s.t. $n = 2^k$.
Remark-Axiom 1: For each $\star \in \{+, -, \times, /\}$, $\exists$ a binary operation $\circledast : \widetilde{\mathbb{R}} \times \widetilde{\mathbb{R}} \to \mathbb{R}$ s.t.

$$|x \star y - x \circledast y| \le \epsilon |x * y|, x, y \in \widetilde{\mathbb{R}}.$$

We use this Axiom several times, and we denote $|\delta| \le \epsilon$ in order to avoid too many ϵ_i's, we should know that every pair of x_i's have different δ but this doesn't affect the result of our round-off error analysis.
Now we'll do the round-off error analysis,

$$\begin{aligned}
\widetilde{S_n} &= (1+\delta)(x_1 + x_2) \oplus \cdots \oplus (1+\delta)(x_{n-1} + x_n) \\
&= (1+\delta)^k (x_1 + x_2) + \cdots + (1+\delta)^k (x_{n-1} + x_n) \\
&= (1+\delta)^{log_2^n} (x_1 + x_2) + \cdots + (1+\delta)^{log_2^n} (x_{n-1} + x_n) \\
&= (1+\delta)^{log_2^n} (x_1 + \cdots + x_n)
\end{aligned}$$

and

$$\begin{aligned}|\widetilde{S_n} - S_n| &\leq |(1+\delta)^{log_2^n} - 1| * |x_1 + \cdots + x_n| \\ &\leq |(1+\epsilon)^{log_2^n} - 1| * |x_1| + \cdots + |(1+\delta)^{log_2^n} - 1| * |x_n| \\ &= [(1+\epsilon)^{log_2^n} - 1] \sum_{i=1}^{n} |x_i|\end{aligned}$$

Here we can estimate

$$(1+\epsilon)^{log_2^n} - 1 \leq \frac{\epsilon log_2^n}{1 - \epsilon log_2^n},$$

this estimation method is based on the lecture notes. So, we can bound the relative error as the from

$$\frac{|\widetilde{S_n} - S_n|}{|S_n|} \leq \frac{\epsilon log_2^n}{1 - \epsilon log_2^n} \frac{\sum_{i=1}^{n} |x_i|}{|\sum_{i=1}^{n} x_i|} = \rho(n, \epsilon)\kappa(x).$$

If n is not a power of 2, we can use 0 instead of some non-exist x_i to do the pairwise summation, hence $\rho(n, \epsilon) = O(\epsilon log_2^n)$.

b).

Let $M_n := x_1 + x_2 + \cdots + x_n$ and $\widetilde{M_n} = x_1 \otimes \cdots \otimes x_n$.
Note that $x_i \in \mathbb{R}(i = 1, \cdots, n)$, $\widetilde{M_n}$ is the pairwise multiplication of x_i. We also assume n is a power of 2 s.t. $n = 2^k$. We use the Axiom 1 several times, and we denote $|\delta| \leq \epsilon$ in order to avoid too many ϵ_i's, we should know that every pair of x_i's have different δ but this doesn't affect the result of our round-off error analysis.
Now we'll do the round-off error analysis,

$$\begin{aligned}\widetilde{M_n} &= (1+\delta)(x_1 x_2) \otimes \cdots \otimes (1+\delta)(x_{n-1} x_n) \\ &= (1+\delta)^{n-1} \prod_{i=1}^{n} x_i \\ &= (1+\delta)^{n-1} M_n\end{aligned}$$

and

$$\begin{aligned}|\widetilde{M_n} - M_n| &\leq |(1+\delta)^{n-1} - 1||M_n| \\ &\leq |(1+\epsilon)^{n-1} - 1||M_n| \\ &= [(1+\epsilon)^{n-1} - 1]|M_n|\end{aligned}$$

Here we can estimate

$$(1+\epsilon)^{n-1} - 1 \leq \frac{\epsilon(n-1)}{1 - \epsilon(n-1)}.$$

So, we can bound the relative error as the from

$$\frac{|\widetilde{M_n} - M_n|}{|M_n|} \leq \frac{\epsilon(n-1)}{1 - \epsilon(n-1)},$$

where $\rho(n, \epsilon) = \frac{\epsilon(n-1)}{1-\epsilon(n-1)}$ and $\kappa(x)$.
If n is not a power of 2, we can use 1 instead of some non-exist x_i to do the pairwise multiplication.

problem 3.

a).

$$\cos x = 1 - 2\sin^2(\frac{x}{2})$$
$$\tan x = \frac{\sin x}{\cos x} = \frac{\sin x}{1 - 2\sin^2(\frac{x}{2})}$$
$$\arcsin(x) = 2\arctan(\frac{x}{1+\sqrt{1-x^2}}), x \in (-1,1)$$
$$\arccos(x) = 2\arctan(\frac{\sqrt{1-x^2}}{1+x}), x \in (-1,1)$$
$$x^a = e^{a\log x}$$

b).

We assume that $y \in [0,1]$. We know that $x \in \mathbb{R}$.
So, we can rewrite x as

$$x = \begin{cases} -\frac{1}{y}, & \text{if } x < -1 \\ -y, & \text{if } x \in [-1,0) \\ y, & \text{if } x \in [0,1] \\ \frac{1}{y}, & \text{if } x > 1 \end{cases}, y \in [0,1]$$

Lemma: $\forall x \in \mathbb{R}, \arctan(x) + \arctan(\frac{1}{x}) = \frac{\pi}{2}$ if $x \geq 0$ and $-\frac{\pi}{2}$ otherwise.

Proof. Let $f(x) = \arctan(x) + \arctan(1/x)$ for all $x \in (0,\infty)$. Then

$$f'(x) = \frac{1}{1+x^2} + \frac{-x^{-2}}{1+x^{-2}} = \frac{1}{1+x^2} - \frac{1}{x^2+1} = 0.$$

Hence $f(x)$ is constant on $(0,\infty)$. Since $f(1) = \frac{\pi}{4} + \frac{\pi}{4} = \frac{\pi}{2}$, we conclude that $f(x) = \frac{\pi}{2}$ for all $x \in (0,\infty)$. And $f(-1) = -f(1) = -\frac{\pi}{2}$, it follows that

$$\arctan(x) + \arctan\left(\frac{1}{x}\right) = \begin{cases} \frac{\pi}{2}; & \text{if } x > 0 \\ \frac{\pi}{2}; & \text{if } x < 0 \end{cases}.$$

$\square$

Therefore, we can rewrite $arctan(x)$ as

$$\arctan(x) = \begin{cases} -\arctan(y) - \frac{\pi}{2}, & \text{if } x < -1 \\ -\arctan(y), & \text{if } x \in [-1,0) \\ \arctan(y), & \text{if } x \in [0,1] \\ \arctan(y) + \frac{\pi}{2}, & \text{if } x > 1 \end{cases}, y \in [0,1]$$

In this case, we can compute $\arctan(x)$ for any $x \in \mathbb{R}$ by $\arctan(y)$ with $y \in [0,1]$, thus we reduce the argument to $[0,1]$. Then we can do further reduction since

$$\arctan(\widetilde{x}) = 2\arctan(\frac{\widetilde{x}}{1+\sqrt{1+\widetilde{x}^2}}).$$

We have already reduced the $\arctan(x), x \in \mathbb{R}$ into $\arctan(y), y \in [0,1]$, then we reduce $y \in [0,1]$ to $\frac{y}{1+\sqrt{1+y^2}} \in [0, 0.414]$.
Hence we we can compute $\arctan(x)$ for any $x \in \mathbb{R}$ by $\arctan(y)$ with $y \in [0, 0.b]$ with $b = 0.414$.

problem 4.

a).

The round-off error analysis part.
Here we use the Gregory series to compute $\log y$, so let

$$p_n = \log \frac{1+x}{1-x} = 2x + \frac{2}{3}x^3 + \frac{2}{5}x^5 \cdots = b_0 + \cdots + b_n$$

such that $b_k = \frac{2x^{2k+1}}{2k+1}$. Then we assume $|\widetilde{b_k} - b_k| \leq k\epsilon |b_k|$. Thus we have

$$p_n' = \widetilde{b_0} + \cdots + \widetilde{b_n},$$

$$\widetilde{p}_n = \widetilde{b_0} \oplus \cdots \oplus \widetilde{b_n}.$$

Note y_n' is the exact sum of inexact values, $\widetilde{y}_n$ is the inexact sum of inexact value.
So,

$$|\widetilde{p}_n - p_n'| \leq \rho(n,\epsilon) \sum_{k=0}^{n} |\widetilde{b_k}| \ (*)$$

by $|\widetilde{b_k}| \leq (1+k\epsilon)|b_k|$,

$$(*) \leq \rho(n,\epsilon) \sum_{k=0}^{n} (1+k\epsilon)|b_k|.$$

$$|\widetilde{p_n} - p_n| \leq |\widetilde{p}_n - p_n'| + |p_n' - p_n| \leq \rho(n,\epsilon) \sum_{k=0}^{n} (1+k\epsilon)|b_k| + \sum_{k=0}^{n} k\epsilon |b_k| \quad (**).$$

Note that $\rho(n,\epsilon)$ is an algorithm dependent error.
Assume $|b_k| \leq cq^k$, for $c > 0, q < 1$ and since $b_k = \frac{2x^{2k+1}}{2k+1}$ and $|x| \leq \frac{1}{2}$. We can set $c = 1$ and $q = \frac{1}{2}$, $|b_k| \leq (\frac{1}{2})^k$ always holds.
So,

$$(**) \leq \rho(n,\epsilon) \sum_{k=0}^{n} (1+k\epsilon)(\frac{1}{2})^k + \epsilon \sum_{k=0}^{n} k(\frac{1}{2})^k$$

Using the derivative of geometric series,

$$(**) \leq \frac{\rho(n,\epsilon)}{(1-\frac{1}{2})} + \frac{\epsilon(1+\rho(n,\epsilon))\frac{1}{2}}{(1-\frac{1}{2})^2} = (2+2\epsilon)\rho(n,\epsilon) + 2\epsilon.$$

Hence the round-off error analysis ended.

b).

A procedure to reduce the argument into $-\frac{1}{2} \leq x \leq \frac{1}{2}$. Since $x \in [-\frac{1}{2}, \frac{1}{2}]$, so $\frac{1+x}{1-x} \in [\frac{1}{3}, 3]$, thus there $\exists \beta \in \mathbb{N}$ and $k \in \mathbb{Z}$ s.t.

$$y = \beta^k (\frac{1+x}{1-x}), |x| \leq \frac{1}{2}.$$

Therefore, we can rewrite $\log y$ as

$$\log y = \log[\beta^k (\frac{1+x}{1-x})] = k\log(\beta) + \log(\frac{1+x}{1-x}).$$

In this case, we can compute $\log y$ for any $y \in \mathbb{R}$ by the sum of a constant and $\log(\frac{1+x}{1-x})$ with $x \leq |\frac{1}{2}|$, thus we reduce the argument to $-\frac{1}{2} \leq x \leq \frac{1}{2}$.

problem 5.

We use the Maclaurin series to design an algorithm to compute $\sin(x)$. We have already known that any derivative of $\sin(x)$ is in $[-1, 1]$ s.t. the maximum value is 1. Also, we know the reminder of the series is

$$|r_n| \leq \frac{x^{2n+1}}{(2n+1)},$$

define S_n as the expansion of $\sin(x)$ up to n terms, then we can have :

$$|\sin(x) - S_n| \leq r_n$$

$$S_n - S_{n-1} = (-1)^n r_{n-1}.$$

Then we denote the absolute error by

$$\epsilon_a(n) = |\sin(x) - S_n|$$

where we don't need more floating point arithmetic and thus the relative error should be

$$\epsilon_r(n) = \frac{\epsilon_a(n)}{\sin(x)} \quad (*),$$

note that $x \in [0, \frac{\pi}{4}$, we can assume $\sin(x) \geq \frac{2x}{\pi}$ and thus

$$(*) \leq \frac{\epsilon_a(n)\pi}{2x} \leq \frac{r_{n+1}\pi}{2x}.$$

The relative error we desire is denoted by ϵ s.t.

$$r_{n+1} \leq \frac{2x}{\pi}\epsilon.$$

Our computational relative error $\epsilon_r(n) < \epsilon$.
By the above analysis part we can formulate the algorithm,

1. Firstly we input $k = r_n$ with $n = 0$;

2. Then we do a loop, the first step is checking that if $k > \frac{2x}{\pi}\epsilon$, if it's true we will end the algorithm, if not we do the next step.;

3. Let $n = n + 1$ and $k = y_n$, then check it again. We may do this loop many times until we find $k \geq \frac{2x}{\pi}\epsilon$ and then we end this algorithm.
Finally we reduce the argument of $\sin(x)$ into $[0, \frac{\pi}{4}]$, by the periodic and symmetric property of $\sin(x)$, we can reduce the function into $[0, \frac{pi}{2}$, further we use the formula

$$\sin(x) = 2\sin(\frac{x}{2})\sqrt{1 - \sin^2(\frac{x}{2})}$$

to reduce the argument into $[0, \frac{\pi}{4}]$.

Remark:

I have worked this assignment with David Knapik, Luke Steverango, Ralph Sarkis, Kabilan Sriranjan and Mathieu Rundström.