Problem 1 (6 points) For each of the following expressions, explain when cancellations can occur and how to avoid them.

(a) $\sqrt{x+1}-1$

(b) $(e^x - e^{-x})/2$

(c) $(1-\cos x)/\sin x$

(a) For $x \approx 0$, $\sqrt{x+1} \approx 1$.

$$\sqrt{x+1}-1 = (\sqrt{x+1}-1)\frac{\sqrt{x+1}+1}{\sqrt{x+1}+1} = \frac{x}{\sqrt{x+1}+1}$$

(b) For $x \approx 0$.

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\ e^{-x} &= 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} - \cdots \\ (e^x - e^{-x}) &\approx x + \frac{x^3}{3!} \end{aligned}$$

(c) For $x \approx 2k\pi$. You can rewrite as

$$\frac{1-\cos x}{\sin x} = \tan(x/2).$$

You can also derive this using $\cos x = \cos^2(x/2) - \sin^2(x/2) = 1 - 2\sin^2(x/2)$ and $\sin x = 2\sin(x/2)\cos(x/2)$. Then

$$\frac{1-\cos x}{\sin x} = \frac{2\sin^2(x/2)}{2\sin(x/2)\cos(x/2)} = \tan(x/2).$$

Problem 2 (4 points) The following Matlab program

```
x = 1;
while  (x+1)-x == 1
    x = 2*x;
end
x
```

outputs `9.007199254740992e+15`. Explain why this loop terminates and explain how this value is produced.

On iteration k of this loop, $x = 2^k$. It terminates when $2^k + 1$ in double precision equals 2^k. In binary

$$2^k = 1.\underbrace{0\cdots0}_{\text{52 zeros}} \times 2^k.$$

When $k = 52$,

$$x + 1 = 2^{52} + 1 = 1.\underbrace{0\cdots0}_{51 \text{ zeros}}1 \times 2^{52} \neq x.$$

When $k = 53$,

$$x + 1 = 2^{53} + 1 = 1.\underbrace{0\cdots0}_{52 \text{ zeros}}1 \times 2^{52}$$

This number is in the middle of $1.\underbrace{0\cdots0}_{52 \text{ zeros}} \times 2^{53}$ and $1.\underbrace{0\cdots0}_{51 \text{ zeros}}1 \times 2^{53}$ and rounds to the even, which is $1.\underbrace{0\cdots0}_{53 \text{ zeros}} = x$, and the loop terminates with $2^{53} = \texttt{9.007199254740992e+15}$.

Problem 3 (4 points) Consider $f(x) = x\sin(x)$. Assume that you are given values for $f(x)$ at $x = 0, \pi/8, \pi/4, 3\pi/8$. Denote by $p(x)$ the polynomial interpolating these values. Derive a bound for $|f(x) - p(x)|$ for any $x \in [0, 3\pi/8]$.

We have $n = 3$ equally spaced subintervals with $h = \pi/8$. $f^{(4)}(x) = x\sin(x) - 4\cos(x)$ and

$$|f(x)| \leq |x\sin(x) - 4\cos(x)| \leq x\sin(x) + 4\cos(x) \leq \frac{3\pi}{8}\sin(3\pi/8) + 4 \approx 5.0884.$$

The from the formula for equally spaced points

$$|f(x) - p(x)| \leq \frac{M}{4(n+1)}h^{n+1} \approx \frac{5.0884}{4(3+1)}(\pi/8)^{3+1} \approx 7.5631 \times 10^{-3}.$$

A sharper bound is obtained by finding the maximum of $f^{(4)}(x)$ over $[0, 3\pi/8]$. Since

$$f^{(5)}(x) = 5\sin(x) + x\cos(x) \geq 0$$

on this interval, $f^{(4)}(x)$ is increasing on it. $f(0) = -4$, $f(3\pi/8) \approx -0.4423$ and hence

$$|f^{(4)}(x)| \leq 4, \quad \text{for all } x \in [0, 3\pi/8].$$

Then

$$|f(x) - p(x)| \leq \frac{4}{4(3+1)}(\pi/8)^{3+1} \approx 5.9454e - 03.$$

Problem 4 (4 points) Let A be an $n \times n$ nonsingular matrix and let B be an $n \times m$ matrix, where $m \geq 1$. How can you compute efficiently an $n \times m$ matrix X such that

$$AX = B$$

What is the complexity of your approach in big-O notation?

Compute the LU factorization of $A = LU$. This is done in $O(n^3)$. Denote the ith column of X by x_i and the ith column of B by b_i.

From $LUx_i = b_i$, solve for each $i = 1 : m$,

$$\begin{aligned} Ly &= b_i, \quad O(n^2) \\ Ux_i &= y \quad O(n^2) \end{aligned}$$

We have $O(mn^2)$ for this work. The overall complexity is $O(n^3 + mn^2)$.

Problem 5 (4 points) Let x and y be floating-point numbers. Assume that you have the log and exp functions available and you want to compute x^y using them. That is, you compute x^y by evaluating the expression $e^{y \ln x}$ using exp(y*log(x)), which is x^y in exact arithmetic.

Assume that $\mathrm{fl}(\mathsf{log}(\mathsf{x})) = (\ln x)(1+\epsilon)$, where $|\epsilon| \leq \eta$ for some η. Ignore the errors in the multiplication and the exp function, that is, assume they produce exact results.

What is the relative error in $\mathsf{exp}(\mathsf{y} * \mathsf{log}(\mathsf{x}))$. Can this error be large and why?

We have

$$\begin{aligned}
y \cdot \ln x \cdot (1+\epsilon) &= y \cdot \ln x + y \cdot \ln x \cdot \epsilon \\
e^{y \cdot \ln x \cdot (1+\epsilon)} &= e^{y \cdot \ln x + y \cdot \ln x \cdot \epsilon} = e^{y \cdot \ln x} e^{y \cdot \ln x \cdot \epsilon} \\
&= x^y (x^y)^\epsilon. \\
\mathrm{fl}[\mathsf{exp}(\mathsf{y} * \mathsf{log}(\mathsf{x}))] &= e^{y \cdot \ln x \cdot (1+\epsilon)} = x^y (x^y)^\epsilon = x^y (1 + \underbrace{(x^y)^\epsilon - 1}_{\delta}) \\
&= x^y (1+\delta).
\end{aligned}$$

This

$$\delta = (x^y)^\epsilon - 1$$

can be large when x^y is very large.