Homework 2

Numerical Analysis (CMPS/MATH 305)

Hani Mehrpouyan

This homework is due by Thursday April 25th 2013

QUESTION 1 (5 POINTS)

Using MATLAB find the binary double precision IEEE floating-point expressions for the following numbers:

a) 8
   
   \[ \begin{array}{c}
   4020000000000000 \\
   3f00000000000000 \\
   \end{array} \]

b) 0.5
   
   \[ \begin{array}{c}
   4020000000000001 \\
   3f00000000000000 \\
   \end{array} \]

QUESTION 2 (5 POINTS)

Let \( x > 0 \) be a floating point number. Consider a computer using a positive binary floating-point representation with \( n \) bits of precision in the significand, e.g., \( n = 24 \). Assume that rounding is used in going from a number \( x \) outside the computer to its floating-point approximation \( f\ell(x) \) inside the computer.

a) Show that
   \[ -2^{e-n} \leq x - f\ell(x) \leq 2^{e-n} \]  
   \( (1) \)

b) Show that \( x \geq 2^e \) and use this to show
   \[ \frac{|x - f\ell(x)|}{x} \leq 2^{-n} \]  
   \( (2) \)

c) Let
   \[ \frac{|x - f\ell(x)|}{x} = -\epsilon \]  
   \( (3) \)

and then solve for \( f\ell(x) \). What are the bounds on \( E \)? (This result extends to \( x < 0 \), with the assumption of \( x > 0 \) being used to simplify the algebra.)
Solution: (a) As shown in the text, we can write $x$ and $\bar{f}(x)$ as

$$x = \sigma \cdot \xi \cdot 2^e, \quad \bar{f}(x) = \sigma \cdot \bar{\xi} \cdot 2^e$$

where $\sigma = \pm 1$, $e$ is an integer, and $1 \leq \xi, \bar{\xi} < 2$. The significand $\xi$ contains all the binary digits of $x$, whereas $\bar{\xi}$ is the machine significand and therefore contains a finite number of digits $n - 1$. We have

$$|x - \bar{f}(x)| = |\sigma \cdot \xi \cdot 2^e - \sigma \cdot \bar{\xi} \cdot 2^e| = 2^e \cdot |\xi - \bar{\xi}|$$

If $\xi$ has a zero in the $n$th bit, then $\bar{\xi}$ is obtained by simply chopping $\xi$ after the $(n-1)$th digit. Then

$$\xi - \bar{\xi} \leq 2^{-(n+1)} + 2^{-(n+2)} + 2^{-(n+3)} + \ldots \leq 2^{-(n+1)} \left(1 + \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \ldots\right)$$

$$\leq 2^{-(n+1)} \frac{1}{1 - \frac{1}{2}} = 2 \cdot 2^{-n-1} = 2^{-n}$$

If $\xi$ has a 1 in the $n$th bit, then $\bar{\xi}$ is obtained by chopping $\xi$ after the $(n-1)$th digit and adding $2^{-n+1}$ to the result. Then it is easy to see that

$$\xi - \bar{\xi} \leq 2^{-n+1} - 2^{-n} = 2^{-n+1} \left(1 - \frac{1}{2}\right) = 2^{-n}$$

Combining these last two results we have

$$|\xi - \bar{\xi}| \leq 2^{-n}$$

so that

$$|x - \bar{f}(x)| \leq 2^e \cdot 2^{-n} = 2^{-n}.$$

(b) Since $x = \sigma \cdot \xi \cdot 2^e$, $1 \leq \xi < 2$, it is true that

$$|x| = \xi \cdot 2^e \geq 1 \cdot 2^e = 2^e$$

or

$$\frac{1}{|x|} \leq 2^{-e}.$$ 

Hence, using part (a),

$$\frac{|x - \bar{f}(x)|}{|x|} \leq 2^{-e} \cdot 2^{-n} = 2^{-n}$$

(c) If

$$\frac{x - \bar{f}(x)}{x} = -\epsilon,$$

then $\bar{f}(x) = x(1 + \epsilon)$. From part (b),

$$\left|\frac{x - \bar{f}(x)}{x}\right| = |\epsilon| \leq 2^{-n}$$

QUESTION 3 (5 POINTS)

Calculate the error, relative error, and number of significant digits in the following approximations $x_A \approx x_T$:

a) $x_T = 28.254$, $x_A = 28.271$,

b) $x_T = 0.028254$, $x_A = 0.028271$, and

c) $x_T = \sqrt{2}$, $x_A = 1.414$. 
In some situations, loss-of-significance errors can be avoided by rearranging the function being evaluated, as was done in class. Do something similar for the following cases, in some cases using trigonometric identities. In all but case (b), assume \( x \) is near 0.

\( \text{a) } \frac{1 - \cos(x)}{x^2}, \)

\( \text{b) } \log(x + 1) - \log(x), \text{ } x \text{ large, and} \)

\( \text{c) } \sqrt{1 + x} - 1. \)


Solution: (a)

\[
\frac{1 - \cos x}{x^2} = \frac{1 - \cos^2 x}{x^2(1 + \cos x)} = \frac{\sin^2 x}{x^2(1 + \cos x)}
\]

or

\[
\frac{1 - \cos x}{x^2} = \frac{2\sin^2(x/2)}{x^2}
\]

which uses \(\cos(2a) = 1 - 2\sin^2 a\).

(b)

\[
\log(x + 1) - \log x = \log \left(\frac{x + 1}{x}\right) \quad \text{or} \quad \log \left(1 + \frac{1}{x}\right)
\]

(d)

\[
\sqrt[3]{1 + x} - 1 = \left[(1 + x)^{1/3} - 1\right] \frac{(1 + x)^{2/3} + (1 + x)^{1/3} + 1}{(1 + x)^{2/3} + (1 + x)^{1/3} + 1} = \frac{(1 + x) - 1}{(1 + x)^{2/3} + (1 + x)^{1/3} + 1}
\]

where we have used the identity

\[a^3 - b^3 = (a - b)(a^2 + ab + a^2)\]

**QUESTION 5 (5 POINTS)**

Use Taylor polynomial approximations to avoid the loss-of-significance errors in the following formulas when \(x\) is near 0:

a) \(\frac{e^x - 1}{x}\)

b) \(\frac{1 - e^{-x}}{x}\)

c) \(\frac{e^x - e^{-x}}{2x}\)

d) \(\frac{\log(1 - x) + xe^{x/2}}{x^3}\)
Consider the identity
\[ \int_0^x \sin(xt) \, dt = \frac{1 - \cos(x^2)}{x}. \]

Explain the difficulty in using the right-hand fraction to evaluate this expression when \( x \) is close to zero. Give a way to avoid this problem and be as precise as possible.
Solution: When $x$ is close to zero, $\cos(x^2)$ is close to 1. When this term is subtracted from 1, leading significant digits will be lost. So it is not good to use the right-hand side expression to compute the integral for $x$ close to 0.

To avoid the loss of significant digits, we use another formulation for $f(x)$, avoiding the subtraction of nearly quantities. By using the Taylor approximation in the right-hand fraction, we get

$$\cos(x^2) = 1 - \frac{x^4}{2!} + \frac{x^8}{4!} + R_4(x)$$

$$R_4(x) = -\frac{x^{12}}{6!} \cos(\xi^2)$$

with $\xi$ an unknown number between 0 and $x$. Therefore,

$$f(x) = \frac{1}{x} \{1 - [1 - \frac{x^4}{2!} + \frac{x^8}{4!} + R_4(x)]\}$$

$$= \frac{x^3}{2!} - \frac{x^7}{4!} + \frac{x^{11}}{6!} \cos \xi^2$$

For $|x| \leq 0.1$,

$$\left| \frac{x^{11}}{6!} \cos \xi^2 \right| \leq \frac{10^{-11}}{6!} \approx 1.4 \times 10^{-14}$$

Hence,

$$f(x) \approx \frac{x^3}{2!} - \frac{x^7}{4!}$$

with an error bound $1.4 \times 10^{-14}$. This gives a much better way of evaluating $f(x)$ for small values of $x$.

**Question 7 (5 points)**

Repeat Problem 7 with the identity

$$f(x) = \frac{1}{x} \int_{0}^{x} e^{-xt} \, dt = \frac{1 - e^{-x^2}}{x^2}, \quad x \neq 0.$$
\[ e^{-x^2} = 1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + R_3(x) \]
\[ R_3(x) = \frac{x^8}{4!} \cos (\xi^2) \]
with \( \xi \) an unknown number between 0 and \( x \). Therefore,
\[ f(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{3!} - \frac{x^6}{4!} \cos (\xi^2) \]
Thus, for \( |x| \leq 0.1, \)
\[ \left| \frac{x^6}{4!} \cos (\xi^2) \right| \leq \frac{10^{-6}}{4} \approx 4.2 \times 10^{-8} \]
Hence,
\[ f(x) \approx 1 - \frac{x^2}{2!} + \frac{x^4}{3!} \]
with an error bound \( 4.2 \times 10^{-8} \) for \( |x| \leq 0.1 \). This give a much better way of evaluating \( f(x) \) for small values of \( x \).

**Question 8 (5 Points)**

Find an accurate value of
\[ f(x) = \sqrt{1 + \frac{1}{x}} - 1. \]
for large values of \( x \). Calculate \( \lim_{x \to \infty} xf(x) \)

**Solution:** \( 1/x \) is a very small value if \( x \) is large, and the computation
\[ f(x) = \sqrt{1 + \frac{1}{x}} - 1 \]
has a loss-of-significance error. There is a simple way to reformulate it so as to avoid the loss-of-significance error.

\[ f(x) = \sqrt{1 + \frac{1}{x}} - 1 = \sqrt{\frac{x + 1}{x}} - 1 = \frac{1}{\sqrt{x}} (\sqrt{x + 1} - \sqrt{x}) \]
\[ = \frac{1}{\sqrt{x}} (\sqrt{x + 1} - \sqrt{x}) \frac{\sqrt{x + 1} + \sqrt{x}}{\sqrt{x + 1} + \sqrt{x}} = \frac{1}{\sqrt{x}(\sqrt{x + 1} + \sqrt{x})} \]
The latter expression will not have any loss-of-significance errors in its evaluation.
To compute the limit, we begin with the latter expression.

\[ \lim_{x \to \infty} xf(x) = \lim_{x \to \infty} x \left( \frac{1}{\sqrt{x}(\sqrt{x + 1} + \sqrt{x})} \right) = \lim_{x \to \infty} \frac{\sqrt{x}}{\sqrt{x + 1} + \sqrt{x}} = \lim_{x \to \infty} \frac{1}{\sqrt{1 + \frac{1}{x} + 1}} = \frac{1}{2} \]
QUESTION 9 (5 POINTS)

For what values of \( x \) does \( x^{10} \) underflow using IEEE double precision normalized floating-point arithmetic. When does it overflow?

**Solution:** (1) When using IEEE double precision arithmetic, the smallest nonzero positive number expressible in normalized floating-point format is

\[ m = 2^{-1022} \approx 2.225 \times 10^{-308} \]

\( x^{10} \) will be set to zero if

\[ x^{10} < m \]

\[ |x| < \sqrt[10]{m} \approx 1.717 \times 10^{-31} \]

\[ -1.717 \times 10^{-31} < x < 1.717 \times 10^{-31} \]

(2) \( x^{10} = [(2^{53} - 1) \cdot 2^{-52}] \cdot 2^{1023} \),

\[ 10 \log_{10} x = \log_{10} [(2^{53} - 1) \cdot 2^{-52}] + 1023 = \log_{10} ((2^{53} - 1) \cdot 2^{971}) \]

Then \( x \approx 6.6907 \times 10^{30} \).
**QUESTION 10 (5 POINTS)**

Find bounds for the error and relative error in approximating $\sin(\sqrt{2})$ by using the $\sin(1.414)$.

**Solution:** $x_A = 1.414$

$$|x_T - x_A| \leq 0.00022$$

$$f(x) = \sin x$$

$$|\sin x_T - \sin x_A| \leq (\cos 1.414)(0.00022) \leq 3.4354 \times 10^{-5}$$

$$|\text{Rel}(\sin x_A)| = \frac{|\sin x_T - \sin x_A|}{|\sin x_T|} \leq \frac{(\cos 1.414)(0.00022)}{\sin 1.414} \leq 3.48 \times 10^{-5}$$
QUESTION 11 (5 POINTS)

In the following function evaluations \( f(x_A) \), assume the numbers \( x_A \) are correctly rounded to the number of digits shown. Bound the error \( f(x_T) - f(x_A) \) and the relative error in \( f(x_A) \) using

\[
f(x_T) - f(x_A) \approx f'(x_T)(x_T - x_A) \approx f'(x_A)(x_T - x_A).
\]

a) \( \cos(1.473) \)

b) \( \tan^{-1}(2.62) \)

c) \( \sqrt{0.0425} \)

Solution:

(a) \( x_A = 1.473 \)

\[
| x_T - x_A | \leq 0.0005 \quad f(x) = \cos x
\]

\[
| \cos x_T - \cos x_A | \leq (\sin 1.4375)(0.0005) \leq 4.98 \times 10^{-4}
\]

\[
| \text{Rel}(\cos x_A) | \leq \frac{| \cos x_T - \cos x_A |}{| \cos x_T |} \leq \frac{(\sin 1.4375)(0.0005)}{\cos 1.4735} \leq 5.12 \times 10^{-3}
\]

(b) \( x_A = 2.62 \)

\[
| x_T - x_A | \leq 0.005 \quad f(x) = \tan^{-1} x
\]

\[
| \tan^{-1} x_T - \tan^{-1} x_A | \leq \frac{1}{1 + (2.615)^2}(0.005) \approx 6.38 \times 10^{-4}
\]

\[
| \text{Rel}(\tan^{-1} x_A) | \leq \frac{| \tan^{-1} x_T - \tan^{-1} x_A |}{\tan^{-1} x_A} \approx 5.29 \times 10^{-4}
\]

(c) \( x_A = 0.0425 \)

\[
| x_T - x_A | \leq 0.00005 \quad f(x) = \sqrt{x}
\]

\[
| \sqrt{x_T} - \sqrt{x_A} | \leq \frac{1}{2\sqrt{0.04245}}(0.00005) \approx 1.21 \times 10^{-4}
\]

\[
| \text{Rel}(\sqrt{x_A}) | \leq \frac{| \sqrt{x_T} - \sqrt{x_A} |}{\sqrt{x_A}} \approx 5.89 \times 10^{-4}
\]