HW1

Textbook #19.

$(-1)^s \times 2^{c-1023} \times (1+f)$

- #a. $s = 0$, $c = 2^1 + 2^3 + 2^{10} = 1034, f = 2^{-1} + 2^{-4} + 2^{-7} + 2^{-8}, x = 3224$.
- #b. $s = 1$, c, f the same as above, $x = -3224$.
- #c. $s = 0$, $c = 2^0 + 2^1 + \ldots + 2^9 = 1023, f = 2^{-2} + 2^{-4} + 2^{-7} + 2^{-8}, x = 1.32421875$.
- #d. s, c the same as above, $f = 2^{-2} + 2^{-4} + 2^{-7} + 2^{-8} + 2^{-52}$,
  $x = 1 + 2^{-2} + 2^{-4} + 2^{-7} + 2^{-8} + 2^{-52}$.

Remark.

To see what 2^{-52} looks like, try

```
%Matlab only
vpa(2^-52,50)
```

or

```
%Matlab only
digits(50)
vpa(2^-52)
```

The answer is

$$2^{-52} = .2220446049250313080847263336181640625000000000000e-15.$$

To see what x looks like directly, try

```
%Matlab only
digits(55)
x=vpa(1)+vpa(2^-2)+vpa(2^-4)+vpa(2^-7)+vpa(2^-8)+vpa(2^-52)
```

The answer is

$$x = 1.324218750000000222044604925031308084726333618164062500.$$

Textbook #22.

```
%Method 1
format long
m = 5;
N = 11;
sum1 = 1;
sum2 = 1;
for i=1:N
  sum1 = sum1 + (-m)^i/factorial(i);
end
for i=1:N
  sum2 = sum2 + (m)^i/factorial(i);
end
sum1
sum2 = 1/sum2
exp(-m)
```

```
%Method 2
function [a, absa, rela, b, absb, relb] = p1_22(x,n)
format long
f = 1;
t = 1;
s = 1;
for i = 1:n
  f = f*i;
  t = t*x;
  s += t/f;
end
%#22(a). Input x = -5
a = s;
absa = abs(e^-5-a);
rela = abs((e^-5-a)/e^-5);
%#22(b). Input x = 5, then compute ans = chopk(1/s,3)
b = 1/s;
absb = abs(e^5-s);
relb = abs((e^-5-b)/e^-5);
end
```

For $n = 9$,

$$\begin{aligned} a &= -1.82710537918871 \\ rela &= 272.166481338708 \\ b &= 0.00695945286364954 \\ relb &= 0.0328743851197008. \end{aligned}$$

For $n = 10$,

$$\begin{aligned} a &= 0.864039076278661 \\ rela &= 127.234768898588 \\ b &= 0.00683150631297318 \\ relb &= 0.0138854333375455. \end{aligned}$$

For $n = 11$,

$$\begin{aligned} a &= -0.359208403479235 \\ rela &= 54.3112539365463 \\ b &= 0.00677489110297060 \\ relb &= 0.00548299116780556. \end{aligned}$$

Method (b) is more accurate. The reason why it produces smaller relative error is mentioned as follows.

Let

$$\begin{aligned} \text{err1} &= \text{absolute error of } e^{-5} \text{ by Taylor series} \\ \text{err2} &= \text{absolute error of } e^{5} \text{ by Taylor series} \\ A &= \frac{5^{n+1}}{(n+1)!} + \frac{5^{n+3}}{(n+3)!} + \frac{5^{n+5}}{(n+5)!} + \dots \\ B &= \frac{5^{n+2}}{(n+2)!} + \frac{5^{n+4}}{(n+4)!} + \frac{5^{n+6}}{(n+6)!} + \dots \end{aligned}$$

W.l.o.g. n is odd. The other case is similar. Then

$$\text{err1} = A - B$$

$$\text{err2} = A + B.$$

Note that

$$B \le A \cdot \frac{5}{n+2}.$$

That implies $(n > 3)$

$$\text{err1} = A\left(1 - \frac{B}{A}\right) \ge A\left(1 - \frac{A \cdot \frac{5}{n+2}}{A}\right) = A \cdot \frac{n-3}{n+2},$$

$$\text{err2} = A\left(1+\frac{B}{A}\right) \le A\left(1+\frac{A\cdot\frac{5}{n+2}}{A}\right) = A\cdot\frac{n+7}{n+2}.$$

And so,

$$\begin{aligned}
\frac{\text{err1}}{\text{err2}} &= \frac{A\left(1-\frac{B}{A}\right)}{A\left(1+\frac{B}{A}\right)} \le 1,\\
\frac{\text{err1}}{\text{err2}} &= \frac{A\left(1-\frac{B}{A}\right)}{A\left(1+\frac{B}{A}\right)} \ge \frac{n-3}{n+7} = 1-\frac{10}{n+7} > \delta \text{ for some } \delta > 0 \text{ if } n \text{ is large enough.}
\end{aligned}$$

Hence, we can assume that

$$\text{err1} \approx \text{err2}.$$

Furthermore,

$$e^{-5} << e^{5}.$$

Now consider the relative erros

$$\begin{aligned}
\text{RelErr of (a)} &= \left|\frac{(e^{-5})_h - e^{-5}}{e^{-5}}\right| = \left|\frac{(e^{-5}+\text{err1}) - e^{-5}}{e^{-5}}\right| = \left|\frac{\text{err1}}{e^{-5}}\right|,\\
\text{RelErr of (b)} &= \left|\frac{\frac{1}{(e^5)_h} - e^{-5}}{e^{-5}}\right| = \left|\frac{\frac{1}{e^5+\text{err2}} - e^{-5}}{e^{-5}}\right| = \left|\frac{\text{err2}}{e^5+\text{err2}}\right| \approx \left|\frac{\text{err2}}{e^5}\right|.
\end{aligned}$$

Therefore, the second one is much smaller than the previous one.

Remark.

The main reason for the better accuray of Method (b) is not due to "no subtractions" since you can get almost the same values of Methods (a) (b) even if you apply `vpa` function in Matlab. The following results were derived from `digits(24)`.

For $n = 9$,

$$\begin{aligned}
a &= -1.82710537918871273978268\\
rela &= 272.166481338707982599396\\
b &= .695945286364953569030809e-2\\
relb &= .328743851197009117883939e-1.
\end{aligned}$$

For $n = 10$,

$$\begin{aligned} a &= .8640390762786595031245 5\\ rela &= 127.23476889858801663493 5\\ b &= .68315063129731835268477 5e-2\\ relb &= .13885433337545577754973 6e-1. \end{aligned}$$

For $n = 11$,

$$\begin{aligned} a &= -.35920840347923699111006\\ rela &= 54.311253936546531467427 4\\ b &= .67748911029705950894826 3e-2\\ relb &= .54829911678056335252493 2e-2. \end{aligned}$$

Textbook #28.

If $d_{k+1} < 5$, then

$$\begin{aligned}
\left|\frac{y - fl(y)}{y}\right| &= \left|\frac{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n - 0.d_1 \ldots d_k \times 10^n}{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n}\right| \\
&= \left|\frac{0.d_{k+1} \ldots}{0.d_1 \ldots d_k d_{k+1} \ldots}\right| \times 10^{-k} \\
&\leq \frac{0.5}{0.1} \times 10^{-k} = 0.5 \times 10^{-k+1}.
\end{aligned}$$

If $d_{k+1} \geq 5$, then

$$\begin{aligned}
\left|\frac{y - fl(y)}{y}\right| &= \left|\frac{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n - 0.d_1 \ldots d_k \times 10^n - 10^{n-k}}{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n}\right| \\
&= \left|\frac{0.d_{k+1} \ldots - 1}{0.d_1 \ldots d_k d_{k+1} \ldots}\right| \times 10^{-k} \\
&\leq \frac{0.5}{0.1} \times 10^{-k} = 0.5 \times 10^{-k+1}.
\end{aligned}$$

HW #2.

W.l.o.g. $x, y \geq 0$.

$$\left|\frac{(x+y)-(x(1+\delta_1)+y(1+\delta_2))(1+\delta_3)}{x+y}\right| \approx \left|\frac{-x(\delta_1+\delta_3)-y(\delta_2+\delta_3)}{x+y}\right| \leq 2\epsilon_M.$$

HW #3.

```
b = 1.e4;
c=  1.e-4;

x1 = ( -b+sqrt(b^2-4*c) )/2
x2 = ( -b-sqrt(b^2-4*c) )/2
x1p = c/x2
(x1-x1p)/x1p
x1pp = -2*c/(b+sqrt(b^2-4*c))
(x1-x1pp)/x1pp
```

- $x_1 = \frac{-10^4+\sqrt{10^8-4\times10^{-4}}}{2} = -\frac{2\times10^{-4}}{10^4+\sqrt{10^8-4\times10^{-4}}}$ or $x_1 = 10^{-4}/x_2$.

$$x_1 = -1.000000000001000e - 08$$

- $x_2 = \frac{-10^4-\sqrt{10^8-4\times10^{-4}}}{2}$.

$$x_2 = -9.999999999999000e + 03 \text{ (Octave)} = -9.999999999989999e + 03 \text{ (Matlab)}$$

Matlab and Octave have different outputs due to the difference of chopping and rounding.

HW #4.

```
% What is the purpose of this program?
% Which one of f1, f2 is more correct? why?
% Most of the newer matlab contains vpa.
% If yours doesn't, try google 'matlab vpa'

a = 3.e-8;
h = a/16;
x = -a:h:a;
f1 = exp(x) - cos(x) - x;
f2 = x.^2 + x.^3/6;
f3 = x.^2;
figure(1), plot(x,f1,'x',x,f2,'o')

x_longer = vpa(x,24);
f_longer = exp(x_longer)-cos(x_longer) - x_longer;
f0 = double(f_longer);

figure(2), plot(x,f0,x,f1,'x',x,f2,'o',x,f3,'*')
```

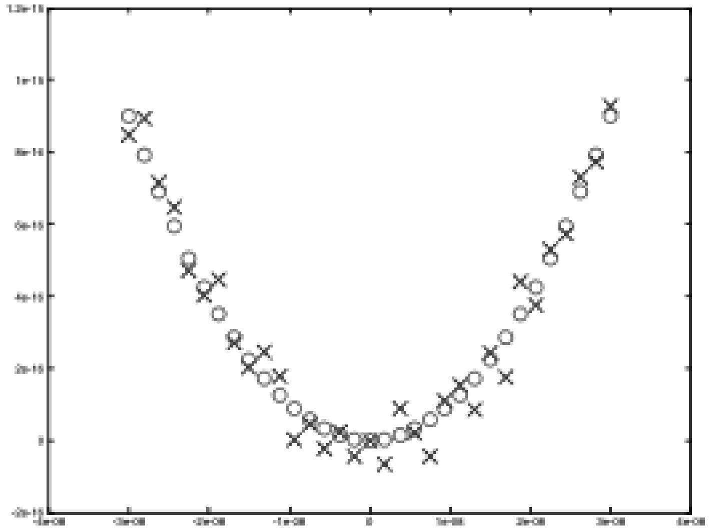
- The purpose of the program is to find a method to compute $e^x - \cos x - x$ correctly.
- f_2 is more correct than f_1 due to the reason:
  - $f_1 = e^x - (\cos x + x)$ may delete significant digits.
  - $f_2 = x^2 + \frac{x^3}{6} + O(x^5)$ with $x^5 \approx 10^{-40}$ over the range $|x| \leq 10^{-8}$.

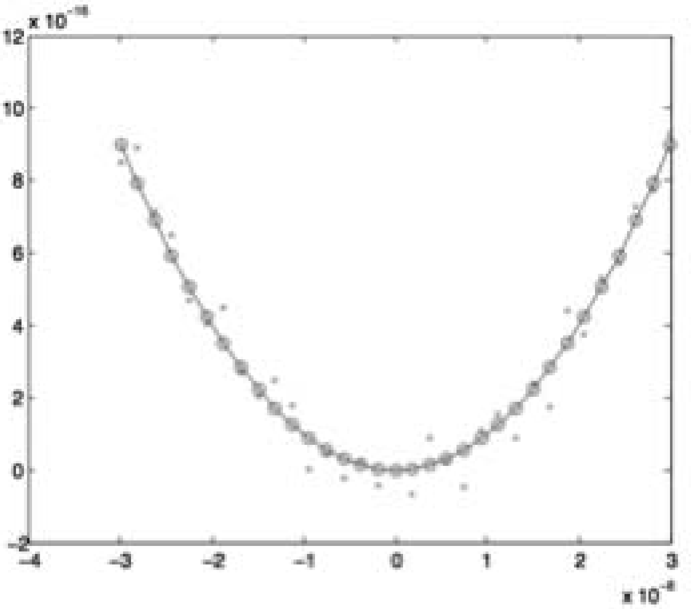
Remark.

f_1 is a typical example of subtraction of two positive numbers of nearly the same magnitute, resulting loss of siginificance. There is no general method of avoiding this type of loss of signigicance. It is usually cured on a case-by-case basis.

One popular cure is to find equivalent formula better suited for numerical computation. In this problem, the first few terms of Taylor expansion serves as a good approximation of the original formula.

In case of a better alternative (equivalent formula) is not available, one can try to verify with high precision (meaning: more accurate than double precision) calculations.


1.2e-15
1e-15
8e-16
6e-16
4e-16
2e-16
0
-2e-16
-4e-08
-3e-08
-2e-08
-1e-08
0
1e-08
2e-08
3e-08
4e-08


$\times 10^{-16}$
12
10
8
6
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
$\times 10^{-8}$