MATH 434/534

Theoretical Assignment 2 Solution

Chapter 3
No. 3.1
(a) Prove the expression
\[
\frac{|fl(x) - x|}{|x|} \leq \mu
\]
where \( \mu = \frac{1}{2} \beta^{1-t} \) for rounding and \( \mu = \beta^{1-t} \) for chopping.

(b) Show that (a) can be written in the form \( fl(x) = x(1 + \delta) \) for \( |\delta| \leq \mu \).

-Solution-

(a) First consider the bound for rounding. Let \( x \) be written as \( x = (d_1d_2\cdots d_t d_{t+1} \cdots) \times \beta^e \), where \( d_1 \neq 0 \) and \( d_i \in [0, \beta) \) are integers. When we round off \( x \), we have one of the following floating point numbers,

\[
x' = (d_1d_2\cdots d_t) \times \beta^e \quad x'' = [(d_1d_2\cdots d_t) + \beta^{-t}] \times \beta^e
\]

Since \( x \in (x', x'') \), without the loss of generality, we assume that \( x \) is close to \( x' \). Then we have

\[
|x - fl(x)| \leq \frac{1}{2} \beta^{e-t}
\]

So, the relative error is

\[
\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2} \frac{\beta^{e-t}}{(d_1d_2d_3\cdots) \times \beta^e} = \frac{1}{2} \frac{\beta^{-t}}{(d_1d_2d_3\cdots)}
\]

Since \( .d_1d_2d_3\cdots \geq .d_1 \) and \( d_i \) are integers such that \( 0 \leq d_i < \beta \), we have \( \frac{1}{.d_1d_2d_3\cdots} \leq \frac{1}{d_1} \). Now we have

\[
\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2} \frac{\beta^{-t}}{(d_1d_2d_3\cdots)} \leq \frac{1}{2} \frac{\beta^{-t}}{d_1} = \frac{1}{2} \frac{\beta^{-t}}{\beta-1}
\]

Therefore for rounding, we have

\[
\frac{|fl(x) - x|}{|x|} \leq \frac{1}{2} \beta^{1-t}
\]
Next consider the bound for chopping. Again let \( x \) be written as \( x = (d_1d_2 \cdots d_{t+1}) \times \beta^e \), where \( d_1 \neq 0 \) and \( d_i \in [0, \beta) \) are integers. Using the chopping method, we have the following floating point numbers,

\[
fl(x) = (d_1d_2 \cdots d_t) \times \beta^e
\]

Now consider the absolute error \( |x - x'| \).

\[
|x - fl(x)| = |(d_1d_2 \cdots d_{t+1}) \times \beta^e - (d_1d_2 \cdots d_t) \times \beta^e| = \left| (\underbrace{.00\cdots0}_{t \text{ zeros}} d_{t+2} \cdots) \times \beta^e \right|
\]

The relative error is

\[
\frac{|x - fl(x)|}{|x|} = \frac{|(d_{t+1}d_{t+2} \cdots) \times \beta^{e-t}|}{|(d_1d_2 \cdots) \times \beta^e|} \leq \beta^{1-t},
\]

because \( |d_1| \geq 1 \) and \( |d_{t+1}| \leq \beta \). Therefore for chopping, we have

\[
\frac{|fl(x) - x|}{|x|} \leq \beta^{1-t}.
\]

(b) Let \( \delta = \frac{fl(x) - x}{x} \). Then we have

\[
fl(x) = x + x\delta = x(1 + \delta)
\]

Also, from part (a) we have

\[
|\delta| = \left| \frac{fl(x) - x}{x} \right| \leq \mu
\]

Therefore, \( fl(x) = x(1 + \delta) \) for \( |\delta| \leq \mu \).

**No. 3.6**

(a) Construct an example to show that, when adding a list of floating point numbers, the rounding error will generally be less if the numbers are added in ordered of increasing magnitude.

(b) Find another example to show that this is not always necessarily true.

-Solution-

(a) Consider \( S_{1000} = \sum_{n=1}^{1000} \frac{1}{n} \). Now suppose that we round each term to 4 decimal places and compute \( S_{1000} = \sum_{n=1}^{1000} \frac{1}{n} \) in ordered of increasing magnitude (an ascending order) and in ordered of decreasing magnitude (a descending order).

<table>
<thead>
<tr>
<th></th>
<th>( S_{1000} = \sum_{n=1}^{1000} \frac{1}{n} )</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascending Order</td>
<td>7.486</td>
<td>0.0</td>
</tr>
<tr>
<td>Descending Order</td>
<td>7.449</td>
<td>0.037</td>
</tr>
</tbody>
</table>
In this case, the relative error of an ascending order is smaller than that of a descending order.

(b) Compute \( S_{1000} = \sum_{n=1}^{1000} \frac{1}{n} \) with rounding each term to 4 decimal places. Do the same as we did in the part (a).

<table>
<thead>
<tr>
<th>Ascending Order</th>
<th>( S_{1000} = \sum_{n=1}^{1000} \frac{1}{n} )</th>
<th>relative error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6.7928</td>
<td>0.0</td>
</tr>
</tbody>
</table>

In this case, the relative error of a descending order is smaller than that of an ascending order.

**No. 3.12**

Let \( \beta = 10, \ t = 4 \). Compute \( fl(A^T A) \).

where

\[
A = \begin{pmatrix}
1 & 10^{-4} \\
0 & 0 \\
10^{-4} & 0 \\
0 & 10^{-4}
\end{pmatrix}
\]

Repeat your computation with \( t = 9 \). Compare the results.

-Solution-

In the exact arithmetic,

\[
A^T A = \begin{pmatrix}
1 & 10^{-4} & 0 & 10^{-4} \\
1 & 0 & 10^{-4} & 0 \\
10^{-4} & 0 & 10^{-4} & 0 \\
0 & 10^{-4} & 0 & 10^{-4}
\end{pmatrix} \begin{pmatrix}
1 & 1 \\
10^{-4} & 0 \\
0 & 10^{-4} \\
0 & 10^{-4}
\end{pmatrix}
\]

\[
= \begin{pmatrix}
1 + 10^{-8} & 1 \\
1 & 1 + 10^{-8}
\end{pmatrix}
\]

\[
= \begin{pmatrix}
1.00000001 & 1 \\
1 & 1.00000001
\end{pmatrix}
\]

For \( \beta = 10 \) and \( t = 4 \), \( fl(1.00000001) = 1.0 \). Therefore, we have

\[
fl(A^T A) = \begin{pmatrix}
1 & 1 \\
1 & 1
\end{pmatrix}
\]

For \( \beta = 10 \) and \( t = 9 \), \( fl(1.00000001) = 1.000000001 = 1 + 10^{-8} \). Therefore, we have

\[
fl(A^T A) = \begin{pmatrix}
1 + 10^{-8} & 1 \\
1 & 1 + 10^{-8}
\end{pmatrix}
\]

**No. 3.13**

Show how to arrange computation in each of the following, so that the loss of significant digits can be avoided. Do one numerical example in each case to support your answer.

(a) \( e^x - x - 1 \) for negative values of \( x \).
(b) \( \frac{1}{x} - \frac{1}{x+1} \) for large values of \( x \).
(c) \( 1 - \cos x \) for values of \( x \) near zero.

-Solution-
(a) Let \( x = -y \) for negative values of \( x \), that is, for \( y > 0 \). Then

\[
e^x - x - 1 = e^{-y} + y - 1
= \frac{1}{e^{y}} + y - 1
= \frac{1}{\sum_{n=0}^{\infty} \frac{y^n}{n}} + y - 1
\]

Take \( y = 0.01 \). Then using the above formula with 4-digit arithmetic, we have

\[
fl \left( \frac{1}{\left(1 + \frac{1}{y} + \frac{1}{y^2} + \frac{1}{y^3} + \cdots \right)} + y - 1 \right) = \frac{1}{1.01} + 0.01 - 1 = -0.0089.
\]

However, substituting \( y = 0.01 \) into the above formulation,

\[
fl \left( e^{-0.01} + 0.01 - 1 \right) = 0.9900 + 0.01 - 1 = 0.
\]

(c) Let \( f(x) = \frac{1}{x} - \frac{1}{x+1} \). In order to avoid the loss of significant digits, use

\[
f(x) = \frac{1}{x(x+1)}
\]

Take \( x = 10^3 \). Then in 8-digit arithmetic, we have

\[
fl \left( \frac{1}{x} - \frac{1}{x+1} \right) = 0. \tag{1}
\]

However, using the above formulation, we have

\[
fl \left( \frac{1}{x(x+1)} \right) = 0.0000001
\]

(e) For \( x \) near zero, \( \cos(x) \) is close to 1. Thus, computing \( 1 - \cos(x) \) is floating point arithmetic. Cancellation will take place. To avoid this use the Taylor series expansion of

\[
\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} + \cdots.
\]

Let \( x = 10^{-4} \). Then we have

\[
fl \left( 1 - \cos(x) \right) = fl(1 - \cos(10^{-4})) = 0.
\]

Using the Taylor series of \( \cos(x) \), we have

\[
fl \left( 1 - \cos(x) \right) \approx \frac{1}{2} \times 10^{-8}.
\]
Therefore, \( f(10^{-4}) \approx \frac{10^{-8}}{2!} - \frac{10^{-16}}{4!} + \frac{10^{-24}}{6} - \frac{10^{-32}}{8!} = 4.99999996 \times 10^{-9}. \)

**No. 3.15**
Let \( \beta = 10, t = 4 \). Consider computing

\[
a = \left( \frac{1}{6} - 0.1666 \right) / 0.1666.
\]

How many correct digits of the exact answer will you get?

-Solution-
The exact value of \( a \) is

\[
a = \left( \frac{1}{6} - 0.1666 \right) / 0.1666
= 4.00160064 \times 10^{-4}
= 0.000400160064
\]

For \( \beta = 10 \) and \( t = 4, \frac{1}{6} = 0.166666 = 0.1667 \). Thus,

\[
fl(a) = (0.1667 - 0.1666) / 0.1666
= 0.0001 / 0.1666
= 6.00240096 \times 10^{-4}
= 0.000600240096
\]

Therefore, there are no correct significant digits.

**No. 3.16**
Consider evaluating \( e = \sqrt{a^2 + b^2} \). How can the computation be organized so that overflow in computing \( a^2 + b^2 \) for large values of \( a \) or \( b \) can be avoided?

-Solution-
If either \( a \) or \( b \) is too large or too small, we can get overflow or underflow. In order to avoid overflow, we normalize both \( a \) and \( b \). Let \( m = \max(|a|, |b|) \). Also let \( y_1 = \frac{a}{m} \) and \( y_2 = \frac{b}{m} \).

Then set \( e = m \sqrt{y_1^2 + y_2^2} \).

**No. 3.18**
What problem do you foresee in solving the quadratic equations
(a) \( x^2 - 10^6x + 1 = 0 \).
(b) \( 10^{-10}x^2 - 10^{10}x + 10^{10} = 0 \)
using the well-known formula

\[
x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}
\]

What remedy do you suggest? Now solve the equations using your suggested remedy, with \( t = 4 \). -Solution-
(a) By using the quadratic formula, we have

\[
x = \frac{10^6 \pm \sqrt{10^{12} - 4}}{2} = \frac{10^6 \pm \sqrt{10^{12}}}{2}
\]
because in four-digit arithmetic, \( \sqrt{10^{12} - 4} = 10^6 \). Then the roots of this equation are \( x_1 = 10^6 \) and \( x_2 = 0 \). However, since \(-b\) and \( \sqrt{b^2 - 4ac} \) are of the same order, the catastrophic cancellation takes place in calculating \( x_2 \). In order to avoid this cancellation, use the following formulas:

\[
x_1 = -\frac{b + \text{sign}(b)\sqrt{b^2 - 4ac}}{2a},
\]
\[
x_2 = \frac{c}{ax_1}
\]

Then we have \( x_1 = 10^6 \) and \( x_2 = 10^{-6} \).

cf. By using MATLAB, you can obtain \( x_1 = 9.99999999999 \times 10^5 \) and \( x_2 = 1.00000761449337 \times 10^{-6} \).

(b)

Similarly, first find the roots of the equation by using the quadratic formula. Then we have

\[
x = \frac{10^{10} \pm \sqrt{10^{20} - 4}}{2(10^{-10})} = \frac{10^{10} \pm \sqrt{10^{20}}}{2(10^{-10})}
\]

because in four-digit arithmetic, \( \sqrt{10^{20} - 4} = 10^{10} \). Then the roots of this equation are \( x_1 = 10^{20} \) and \( x_2 = 0 \). The catastrophic cancellation takes place when \( x_2 \) is computed. In order to avoid this cancellation, use the following formulas:

\[
x_1 = -\frac{b + \text{sign}(b)\sqrt{b^2 - 4ac}}{2a},
\]
\[
x_2 = \frac{c}{ax_1}
\]

Then we have \( x_1 = 10^{20} \) and \( x_2 = 1 \).

cf. By using MATLAB, you can obtain \( x_1 = 1.0 \times 10^{20} \) and \( x_2 = 0 \).

**No. 3.19**

Show that the integral

\[
y_i = \int_0^1 \frac{x^i}{x + 5} \, dx
\]

can be computed by using the recursion formula:

\[
y_i = \frac{1}{i} - 5y_{i-1}.
\]

Compute \( y_1, y_2, \ldots, y_{10} \) using this formula, taking

\[
y_0 = [\ln(x + 5)]_0^1 = \ln 6 - \ln 5 = \ln(1.2).
\]

What abnormalities do you observe in this computation? Explain what happened. Now rearrange the recursion so that the values of \( y_i \) can be computed more accurately.

-Solution-

For \( i = 1 \),

\[
y_1 = \int_0^1 \frac{x}{x + 5} \, dx
\]
\[
= \int_0^1 1 + \frac{-5}{x + 5} \, dx
\]
\[
= 1 - 5 \int_0^1 \frac{1}{x + 5} \, dx = 1 - 5y_0
\]
For $i = k$, assume that we can express the integral

$$y_k = \int_0^1 \frac{x^k}{x+5} \, dx$$

as

$$y_k = \frac{1}{k} - 5y_{k-1}.$$

With this assumption, we will show that the integral $y_{k+1} = \int_0^1 \frac{x^{k+1}}{x+5} \, dx$ expressed as the recursion formula $y_{k+1} = \frac{1}{k+1} - 5y_k$.

$$y_{k+1} = \int_0^1 \frac{x^{k+1}}{x+5} \, dx$$

$$= \int_0^1 \frac{x^k - 5x^k}{x+5} \, dx$$

$$= \int_0^1 x^k \, dx - 5 \int_0^1 \frac{x^k}{x+5} \, dx$$

$$= \frac{1}{k+1} - 5y_k$$

By the mathematical induction, for $i = 1, 2, 3 \ldots$ we showed that the integral $y_i = \int_0^1 \frac{x^i}{x+5} \, dx$ can be expressed as the formula $y_i = \frac{1}{i} - 5y_{i-1}$. Now computing the integral by using the recursion formula, you will have

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>0.1823</td>
<td>0.0884</td>
<td>0.0580</td>
<td>0.0431</td>
<td>0.0343</td>
<td>0.0285</td>
<td>0.0243</td>
<td>0.0212</td>
<td>0.0188</td>
<td>0.0169</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$i$</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>0.0154</td>
<td>0.0141</td>
<td>0.0130</td>
<td>0.0120</td>
<td>0.0112</td>
<td>0.0105</td>
<td>0.0099</td>
<td>0.0094</td>
<td>0.0087</td>
<td>0.0092</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$i$</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>0.0042</td>
<td>0.0264</td>
<td>-0.0866</td>
<td>0.4764</td>
<td>-2.3401</td>
<td>11.7405</td>
</tr>
</tbody>
</table>

Or with four decimal digits floating point rounded arithmetic, we have

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>0.1823</td>
<td>0.0885</td>
<td>0.0575</td>
<td>0.0458</td>
<td>0.0208</td>
<td>0.0958</td>
<td>-0.3125</td>
<td>1.7054</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$i$</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>-8.4018</td>
<td>42.1200</td>
<td>-210.5002</td>
</tr>
</tbody>
</table>

Since $x \in [0, 1]$, the integrand $\frac{x^i}{x+5}$ is positive for $i = 1, 2, \ldots, n$. Therefore, the integral $\int_0^1 \frac{x^i}{x+5} \, dx \geq 0$. However, we have negative values at $i = 22$ and $i = 24$. Now rewrite the recursion formula as

$$y_{i-1} = -\frac{y_i}{5} + \frac{1}{5i}$$

for $i = 2, 3, \ldots, n$. Also since $x \in [0, 1]$, we have $\frac{1}{6} \leq \frac{1}{x+5} \leq \frac{1}{5}$. Therefore, for $i = 1, 2, \ldots, n$ we have

$$\frac{x^i}{6} \leq \frac{x^i}{x+5} \leq \frac{x^i}{5} \leq x^i$$
This inequality tells us that $\int_0^1 \frac{x^i}{x+5}dx \leq \int_0^1 x^i dx = \frac{1}{i+1}$. For $i = 25$ we have

$$y_{25} = \int_0^1 \frac{x^i}{x+5}dx \leq \frac{1}{26}$$

Now assuming $y_{25} = 0$ and using this rearranged recursion formula, we will obtain

\[
\begin{array}{cccccccccc}
\hline
i & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
\hline
y_i & 0.0884 & 0.0580 & 0.0431 & 0.0343 & 0.0285 & 0.0243 & 0.0212 & 0.0188 & 0.0169 & 0.0154 \\
\hline
i & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 \\
\hline
y_i & 0.0141 & 0.0130 & 0.0120 & 0.0112 & 0.0105 & 0.0099 & 0.0093 & 0.0088 & 0.0084 & 0.0080 \\
\hline
i & 21 & 22 & 23 & 24 \\
\hline
y_i & 0.0076 & 0.0073 & 0.0067 & 0.0080 \\
\hline
\end{array}
\]