1 Floating Point Numbers

Let $F$ be the set of floating point numbers and let each number $X$ in $F$ be represented in terms of the base, $\beta$ which is always even, by

$$X = \pm \left( d_1 \frac{1}{\beta} + d_2 \frac{1}{\beta^2} + \cdots + d_n \frac{1}{\beta^n} \right) \beta^E$$  \hspace{1cm} (1a)

where

$$0 \leq d_i \leq \beta - 1 \quad (i = 1, \ldots, n)$$  \hspace{1cm} (1b)

$$L \leq E \leq U$$  \hspace{1cm} (1c)

If $d_1 \neq 0$, then the floating point number system is said to be normalized and we define the different quantities as:

\[ n = \text{precision} \]  \hspace{1cm} (2a)

\[ E = \text{exponent} \]  \hspace{1cm} (2b)

\[ f = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \cdots + \frac{d_n}{\beta^n} = \text{fraction, mantissa, significand} \]  \hspace{1cm} (2c)

and note that $\beta^{1-n}$ is an estimate of the relative accuracy of the arithmetic.

$F$ contains exactly $2(\beta - 1)\beta^{n-1}(U - L + 1) + 1$ numbers, not all equally spaced.

For $\beta = 8, E = 0, n = 1$ we have

\[ 0, \frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{4}{8}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8} \]

and if normalized, $F$ does not contain 0.

For $\beta = 2, n = 3, L = 0, U = 0$ we have $0 \leq d_i \leq \beta - 1 = 1, f = \pm (\frac{d_1}{2} + \frac{d_2}{4} + \frac{d_3}{8})$ where

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$f$</th>
<th>Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4/8</td>
<td>yes</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6/8</td>
<td>yes</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5/8</td>
<td>yes</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7/8</td>
<td>yes</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2/8</td>
<td>no</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1/8</td>
<td>no</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3/8</td>
<td>no</td>
</tr>
</tbody>
</table>

Let $fl(X)$ be the number in $F$ closest to $X$, then

$$\left| \frac{fl(X) - X}{X} \right| \leq \frac{1}{2} \beta^{1-n}$$  \hspace{1cm} (3)
If $U = 1$, we have for $E = 0$ the numbers

$$0, \frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{4}{8}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}$$

and for $E = 1$ the numbers

$$0, \frac{2}{8}, \frac{4}{8}, \frac{6}{8}, \frac{8}{8}, \frac{10}{8}, \frac{12}{8}, \frac{14}{8}$$

or the unique numbers

$$0, \frac{1}{8}, \frac{2}{8}, \frac{3}{8}, \frac{4}{8}, \frac{5}{8}, \frac{6}{8}, \frac{7}{8}, \frac{8}{8}, \frac{10}{8}, \frac{12}{8}, \frac{14}{8}$$

We note that $\frac{5}{8} + \frac{6}{8} = \frac{11}{8}$ does not exist in this set of numbers. Thus $X + Y$ may not be in $F$. If $X > F$ we have overflow and often $XY$ is not in $F$. On the other hand $XY$ may be smaller than the smallest number in $F$ and we have underflow.

We must recognize that some usual operations may not exist

<table>
<thead>
<tr>
<th>Operation</th>
<th>Commute</th>
<th>Distribute</th>
<th>Commute yes $A \cdot B = B \cdot A$</th>
<th>Distribute no $(A + B) + C \neq A + (B + C)$</th>
<th>Distribute no $(A \cdot B) \cdot C \neq A \cdot (B \cdot C)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Many numbers in common use are not in $F$, for example

$$0.1 = \frac{1}{10} = \frac{0}{2} + \frac{0}{4} + \frac{0}{8} + \frac{1}{16} + \frac{1}{32} + \frac{0}{64} + \cdots$$

$$= (0.0001100110011 \cdots)_2$$

$$= (0.063146314 \cdots)_8$$

$$= (0.19999 \cdots)_16$$

and terminating after $d_n$ does not yield $0.1$. It does fall between two floating point numbers, but is not represented by either of them.

The relative error can vary by as much as a factor of $\beta$. This factor is called the wobble. The relative error is bounded by the machine epsilon defined to be $(\beta/2)\beta^{-n}$.

## 2 Machine Epsilon

The machine epsilon is also a number $\epsilon$ such that

$$1 + \epsilon \rightarrow 1$$

$\epsilon$ reflects the accuracy of floating point arithmetic. You can access $\epsilon$ from the Matlab command window with the command "eps"
3 Matlab Formats

From the Matlab command window you can set different formats for displaying results:

<table>
<thead>
<tr>
<th>Format</th>
<th>Digits after dec pt</th>
<th>Exponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>format long</td>
<td>15</td>
<td>actual</td>
</tr>
<tr>
<td>format short</td>
<td>4</td>
<td>actual</td>
</tr>
<tr>
<td>shortEng</td>
<td>4</td>
<td>power of 3</td>
</tr>
<tr>
<td>longEng</td>
<td>15</td>
<td>power of 3</td>
</tr>
<tr>
<td>compact</td>
<td></td>
<td>compact line spacing</td>
</tr>
</tbody>
</table>

4 Roundoff Error

Consider the Taylor series
\[ e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \] (6)
and let \( \beta = 10, n = 3, x = -3.5 \). Then the terms in the series are

\[
e^{-3.5} = \begin{array}{c}
1 \\
-3.5 \\
6.13 \text{ rounded up from } 6.125 \\
-7.15 \\
6.25 \\
-4.38 \\
2.55 \\
-1.28 \\
0.55 \\
-0.22 \\
0.08 \\
-0.02 \\
0.01 \\
0.00 \ (0.001899)
\end{array}
\]

and there is no agreement between the series and the correct result. The term 6.125 when rounded has an error of 0.005 or 1/6 of the final result.

Taking the difference between numbers may lead to catastrophic cancellation.

4.1 solution

a) increase the precision \( n \)
b) different algorithm.
Consider $e^{-x} = 1/e^x$, then

\begin{align*}
e^{3.5} &= 1 + 3.5 + 6.13 + 7.15 \cdots \quad (7) \\
33.3 \quad (8) \\
e^{-3.5} &= 0.0300 \quad (9) \\
exact &= 0.0302 \quad (n = 3) \quad (10) \\
diff &= 0.002 \cdot 10^{-1} \quad (11)
\end{align*}
5 Unstable algorithms--sensitivity

Consider the usual quadratic equation

\[ ax^2 + bx + c = 0 \]  

(12a)

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

(12b)

if \( b^2 \approx 4ac \) catastrophic cancellation  

(12c)

if \( 4ac << b^2 \) catastrophic cancellation in \( -b \pm \sqrt{ } \)  

(12d)

(12e)

the solution is to recognize that

\[ ax^2 + bx + c = a(x - x_1)(x - x_2) \]  

(13a)

where \( c = ax_1x_2 \)  

(13b)

first evaluate

\[ q = -\frac{1}{2} \left[ b + \text{sign}(b) \sqrt{b^2 - 4ac} \right] \]  

(13c)

then the roots are

\[ x_1 = q/a, \; x_2 = c/q \]  

(13d)

\( x^2 - y^2 \) is even more susceptible to catastrophic cancellation. It is better to write this as \( (x - y)(x + y) \) although this is less accurate if \( x >> y \) or \( x << y \).

6 Instability

Consider evaluating

\[ E_n = \int_0^1 x^n e^{x-1} dx \]  

(14a)

integrating by parts gives

\[ E_n = 1 - n E_{n-1}, \; \; n = 2, 3, \cdots \]  

(14b)

with \( E_1 = 1/e \)  

(14c)

Using

\[ E_n = 1 - n E_{n-1} \]  

(15)

we see that each error is multiplied by \( n \) as we progress.

Instead use

\[ E_{n-1} = \frac{1 - E_n}{n} \]  

(16)
and then each error will be reduced by \( n \). The question is how do we start? Since

\[
E_n = \int_0^1 x^n e^{x-1} dx \leq \int_0^1 x^n dx = \frac{x^{n+1}}{n+1} = \frac{1}{n+1}
\]

so \( E_n \rightarrow 0 \) as \( n \) increases

then the initial error for large \( n \) will decrease as Eq. 16 is used. Eq.15 is an unstable algorithm.

## 7 Taylor Series

For \( f(x) \) the Taylor series is

\[
f(x) = f(a) + \frac{df}{dx}|_a (x-a) + \frac{1}{2!} \frac{d^2 f}{dx^2}|_a (x-a)^2 + \cdots + \frac{1}{n!} \frac{d^n f}{dx^n}|_a (x-a)^n
\]

For \( f(x, y) \) we have

\[
f(x + h, y + k) = f(x, y) + (h \frac{\partial f}{\partial x} + k \frac{\partial f}{\partial y})|_{x,y} + \frac{1}{2!}(h^2 \frac{\partial^2 f}{\partial x^2} + 2hk \frac{\partial^2 f}{\partial x \partial y} + k^2 \frac{\partial^2 f}{\partial y^2})|_{xy} + \frac{1}{3!}(h^3 \frac{\partial^3 f}{\partial x^3} + 3h^2k \frac{\partial^3 f}{\partial x^2 \partial y} + 3hk^2 \frac{\partial^3 f}{\partial x \partial y^2} + k^3 \frac{\partial^3 f}{\partial y^3})|_{xy} + \cdots
\]