Introduction

– $1 + 1 = 0$ or “machine epsilon”?
– » eps = 2.220446049250313e-016

How does matlab produce its numbers?

• Where we learn about number formats, truncation errors and roundoff
Matlab real number formats

» format long (default for π)
   \[ \pi = 3.14159265358979 \]

» format short
   \[ \pi = 3.1416 \]

» format short e
   \[ \pi = 3.1416e+000 \]

» format long e
   \[ \pi = 3.141592653589793e+000 \]
Floating-point numbers

\[ x = \pm \left( \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \frac{d_3}{\beta^3} + \cdots + \frac{d_t}{\beta^t} \right) \beta^e \]

\( \beta \) Base or radix
\( t \) Precision
\([L,U]\) Exponent range

\[ 0 \leq d_i \leq \beta - 1, i = 1, \ldots, p; d_1 \neq 0 \]

\[ L \leq e \leq U \]

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>Base</th>
<th>Precision</th>
<th>L(ow Exp)</th>
<th>U(pperExp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE SP</td>
<td>2</td>
<td>24</td>
<td>-126</td>
<td>127</td>
</tr>
<tr>
<td>IEEE DP</td>
<td>2</td>
<td>53</td>
<td>-1022</td>
<td>1023</td>
</tr>
<tr>
<td>Cray</td>
<td>2</td>
<td>48</td>
<td>-16383</td>
<td>16384</td>
</tr>
<tr>
<td>HP Calc</td>
<td>10</td>
<td>12</td>
<td>-499</td>
<td>499</td>
</tr>
<tr>
<td>IBM mainfr</td>
<td>16</td>
<td>6</td>
<td>-64</td>
<td>63</td>
</tr>
</tbody>
</table>

\[ x = (-1)^s \cdot (0.d_1d_2 \cdots d_t) \cdot \beta^e = (-1)^s \cdot m \cdot \beta^{e-t} \]

\[ m = d_1d_2 \cdots d_t \quad d_1 \neq 0 \]
The set $F$ of f.p. numbers

- Basis $\beta$
- Significant digits $t$
- Range $(U, L)$

$F(\beta, t, U, L):$ $F(2, 53, -1021, 1024)$ is the IEEE standard
sign bit

mantissa (52 bits)
(it is assumed d1 = 1, so only need d2, ... d53)

exponent (11 bits)
IEEE double precision standard

If $E=2047$ and $F$ is nonzero, then $V=\text{NaN}$
\begin{center} ("Not a number") \end{center}

If $E=2047$ and $F=0$ and $S=0,(1)$, then $V=\text{Inf, (-Inf)}$

If $E=0$ and $F=0$ and $S=0,(1)$, then $V=0,(0)$

**If $0<E<2047$ then**

\begin{equation}
V=(-1)^*S * 2 ** (E-1023) * (1.F)
\end{equation}

where "1.F" denotes the binary number created by prefixing $F$ with an implicit leading 1 and a binary point.

If $E=0$ and $F$ is nonzero, then
\begin{equation}
V=(-1)^*S * 2 ** (-1022) * (0.F)
\end{equation}

These are "unnormalized" values.
\[ F(2,2,-2,2) \]

\[ \varepsilon = \beta^{1-t} = 2^{1-2} = \frac{1}{2} \]

gap contains “unnormalized” values, as allowed, e.g., in IEEE standard
Floating point numbers

\[ \text{Underflow level: } UFL = \beta^{L-1} \]

\[ \text{Overflow level: } OFL = \beta^U (1 - \beta^{-t}) \]

\[ \epsilon := \text{eps} = \beta^{1-t} \]

The machine precision is the smallest number \( \epsilon \) such that:

\[ fl(1 + \epsilon) > 1 \]

\[ \text{IEEE } _{sp} \epsilon = 2^{-23} \approx 10^{-7} \quad \text{IEEE } _{dp} \epsilon = 2^{-52} \approx 10^{-16} \]
Machine epsilon

• The distance from 1 to the next larger float

\[ \varepsilon := \text{eps} = \beta^{1-t} \]

• Gives the relative error in representing a real number in the system F:

\[ \left| \frac{x - \text{fl}(x)}{|x|} \right| \leq \frac{1}{2} \varepsilon \]

(RELATIVE) ROUNDOFF ERROR
Machine epsilon computed

\begin{verbatim}
a = 1; b = 1;
while a+b ~= a
    b = b/2;
end
b
\end{verbatim}

\begin{verbatim}
% b = 1.110223024625157e-016
% shows that a+b = a is satisfied by
% numbers b not equal to 0
% here b = eps/2 is the largest such
% number for a = 1
\end{verbatim}
Overflow does not only cause programs to crash!
Arianne V’s short maiden flight on 7/4/96 was due to a floating exception.
FLOAT $\rightarrow$ INTEGER

- During the conversion of a 64-bit floating-point number to a 16-bit signed integer
- Caused by the float being outside the range representable by such integers
- The programming philosophy employed did not guard against software errors—a fatal assumption!
COMPLEX NUMBERS

- $z = x + iy$
- $x = \text{Re}(z)$ is the real part
- $y = \text{Im}(z)$ is the imaginary part
- $i^2 = -1$ is the imaginary unit

- polar form $z = \rho e^{i\vartheta} = \rho (\cos \vartheta + i \sin \vartheta)$

- complex conjugate $\overline{z} = x - iy$
• Matlab commands:
  >> z = 3+i*4
  >>% Cartesian form:
     \[ x = \text{real}(z); \ y = \text{imag}(z) \]
  >>% Polar form:
     \[ \text{theta} = \text{angle}(z); \ \rho = \text{abs}(z) \]
  So: \[ z = \text{abs}(z)*(\cos(\text{angle}(z)) + i*\sin(\text{angle}(z))) \]

     \[ x-i*y = \text{conj}(z) \]
z=[1+2*i, 3,-1+i, -i];

compass(z,’r’)

defines an array of complex numbers which are plotted as vectors in the x-y plane
Roots of complex numbers

```matlab
>> x = -1 ; x^(1/3)
ans =

    0.5000 + 0.8660i

Matlab assumes complex arithmetic, and returns automatically the root with the smallest phase angle
( the other two roots are
    -1
and
    0.5000 - 0.8660i
```
Types of errors in numerical computation

- Roundoff errors
  \[ \pi = 3.14159 \]
  \[ \pi = 3.1415926535897932384626 \]

- Truncation errors
  \[ \cos x = 1 - \frac{x^2}{2} \]
  \[ \cos x = 1 - \frac{x^2}{2} + \frac{x^4}{4!} \]

  Errors usually accumulate randomly (random walk)

  But they can also be systematic, and the reasons may be subtle!
x = linspace(1-2*10^-8,1+2*10^-8,401);
f = x.^7-7*x.^6+21*x.^5-35*x.^4+35*x.^3-21*x.^2+7*x-1;
plot(x,f)

CANCELLATION ERRORS
\[
g = -1 + x \cdot (7 + x \cdot (-21 + x \cdot (35 + x \cdot (-35 + x \cdot (21 + x \cdot (-7 + x))))));
\]

plot(x,g)
h = (x-1).^7;
plot(x,h)
plot(x,h,x,g,x,f)
\[ z(1) = 0; \quad z(2) = 2; \]
\[ \text{for } k = 2:29 \]
\[ z(k+1) = 2^{(k-1/2)} \cdot (1 - (1 - 4^{(1-k)} \cdot z(k)^2)^{1/2})^{1/2}; \]
\[ \text{end} \]
\[ \text{semilogy}(1:30, \text{abs}(z - \pi)/\pi) \]
I. ARITHMETIC OPERATIONS and symbols

(1) format long, longe; format short, short e

(2) +, *, ^, ~, /, –

(3) suppress output: ending commands with “;”

(4) Complex:
   real, imag, conj, i, j, angle, abs, compass

(5) Machine constants and special variables:
   eps, realmin, realtime, pi

(6) loops: loop until condition
   while (condition true)
   end
Summary

- Roundoff and other errors
- Formats and floating point numbers
- Complex numbers
References

• Higham & Higham, Matlab Guide, SIAM
• SIAM News, 29(8), 10/98 (Arianne V failure)
• B5 Trailer; http://www.scifi.com/b5rangers/