

CSE/Math 455
Lecture # 4

IEEE Arithmetic

Book

M.L. Overton, Numerical Computing with IEEE Floating Point Arithmetic, SIAM Publications, Philadelphia, PA, 2001.

Good book for general computer scientists.

Supports the following “special numbers.”

Inf

-Inf

0

-0

NaN

$$\begin{aligned} 1/0 &= \mathbf{Inf}, & 1/\mathbf{Inf} &= 0 \\ 1/(-0) &= \mathbf{-Inf} & 1/(\mathbf{-Inf}) &= -0 \end{aligned}$$

NaN (not a number) can be generated by

Inf + (**-Inf**)

0 * **Inf**

Inf/Inf

0/0

$\sqrt{-3}$

The last is for a “reals only” environment. IEEE Arithmetic is a closed system.

$$\begin{aligned} \{\text{floating point numbers, } \mathbf{Inf}, \mathbf{-Inf}, \mathbf{NaN}\} &\rightarrow \\ \{\text{floating point numbers, } \mathbf{Inf}, \mathbf{-Inf}, \mathbf{NaN}\} & \end{aligned}$$

from any operations.

You can take advantage of its clever features.

Example 1 *Computing*

$$f(x) = \sec x - \tan x, \quad x \text{ near } \pi/2$$

Plug straight into MATLAB and

$$f(\pi/2) = \infty - \infty = \mathbf{NaN}.$$

*MATLAB is smart enough not to give you **Inf** here. But*

$$\begin{aligned} f(x) &= (\sec x - \tan x)(\sec x + \tan x)/(\sec x + \tan x) \\ &= (\sec^2 x - \tan^2 x)/(\sec x + \tan x) = 1/(\sec x + \tan x) \end{aligned}$$

Thus

$$f(\pi/2) = 1/(\mathbf{Inf} + \mathbf{Inf}) = 0.$$

Sophisticated codes take advantage of this.

The following class of problems gives rise to two separate types of issues, one we have already discussed, one we have not.

Linear Systems of Equations

$$\begin{aligned} \frac{\varepsilon_M}{10}x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned}$$

A good approximate answer is $x_1 = x_2 = 1$. Use the augmented system approach.

$$\begin{aligned} &\left(\begin{array}{cc|c} \frac{\varepsilon_M}{10} & 1 & 1 \\ 1 & 1 & 2 \end{array} \right) \\ &\left(\begin{array}{cc|c} \frac{\varepsilon_M}{10} & 1 & 1 \\ 0 & 1 - 10/\varepsilon_M & 2 - 10/\varepsilon_M \end{array} \right) \\ \approx &\left(\begin{array}{cc|c} \frac{\varepsilon_M}{10} & 1 & 1 \\ 0 & -10/\varepsilon_M & -10/\varepsilon_M \end{array} \right), \quad \text{to machine precision} \end{aligned}$$

Back solve to get $x_1 = 0, x_2 = 1$. If you put these back in the original system, note that $x_1 + x_2 = 1$, so this is “way off.”

Again there is a “fix,” it is called partial pivoting. Put largest uneliminated entry in the column in pivot or diagonal position

$$\begin{aligned} & \left(\begin{array}{cc|c} 1 & 1 & 2 \\ \frac{\varepsilon_M}{10} & 1 & 1 \end{array} \right) \\ & \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 - \frac{\varepsilon_M}{10} & 1 - \frac{\varepsilon_M}{10} \end{array} \right) \\ \approx & \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 & 1 \end{array} \right) \end{aligned}$$

This yields the solution $x_1 = x_2 = 1$.

Sometimes changing the algorithm does NO GOOD AT ALL! Here ε_M is the value generated by MATLAB’s eps command.

$$\begin{aligned} (1 + 2\varepsilon_M)x_1 + (1 + 2\varepsilon_M)x_2 &= 2 \\ (1 + \varepsilon_M)x_1 + x_2 &= 2 \end{aligned}$$

Using the augmented matrix approach

$$\left(\begin{array}{cc|c} (1 + 2\varepsilon_M) & (1 + 2\varepsilon_M) & 2 \\ (1 + \varepsilon_M) & 1 & 2 \end{array} \right)$$

Multiply the first row by $\alpha = (1 + \varepsilon_M)/(1 + 2\varepsilon_M) = 1 - \varepsilon_M + O(\varepsilon_M^2)$ and you get

$$\left(\begin{array}{cc|c} (1 + 2\varepsilon_M) & (1 + 2\varepsilon_M) & 2 \\ 0 & -\varepsilon_M & 2\varepsilon_M \end{array} \right)$$

The solution is $x_1 = 4$ and $x_2 = -2$. This is correct to machine precision.

A small change in the right hand side yields

$$\left(\begin{array}{cc|c} (1 + 2\varepsilon_M) & (1 + 2\varepsilon_M) & 2 + 4\varepsilon_M \\ (1 + \varepsilon_M) & 1 & 2 + \varepsilon_M \end{array} \right)$$

The correct answer is $x_1 = x_2 = 1$, but with rounding I get $x_1 = 0$, and $x_2 = 2$. Why? Every trick I know except increasing the precision, yields similar wrong answers.

Reason MATLAB rounds this system to

$$\left(\begin{array}{cc|c} (1 + 2\varepsilon_M) & (1 + 2\varepsilon_M) & 2 + 4\varepsilon_M \\ (1 + \varepsilon_M) & 1 & 2 \end{array} \right)$$

which has the solution $x_1 = 0$ and $x_2 = 2$. This problem is very close to the singular system

$$\left(\begin{array}{cc|c} 1 & 1 & 2 \\ 1 & 1 & 2 \end{array} \right)$$

for which has the solutions

$$\mathbf{x} = (x_1, x_2)^T = (1, 1)^T + \beta * (-1, 1)^T, \quad \beta \in \mathbf{R},$$

Note that $(x_1, x_2) = (1, 1)$ and $(x_1, x_2) = (0, 2)^T$ are two of those solutions.