Numerical Linear Algebra: An Introduction - Notes

Enki A. Barra Melendrez

March 23, 2025

2 Error, Stability and Conditioning

Definition 2.1. *A B-adic, normalized floating point number of precision m is either $x = 0$ or:*

$$x = B^e \sum_{k=-m}^{-1} x_k B^{-k},\ x_{-1} \neq 0,\ x_k \in \{0, 1, \dots, B-1\}$$

Where:

- *$B \geq 2$ is the base of the number system.*
- *$e_{min} \leq e \leq e_{max}$ is the exponent.*
- *$\sum_{k=-m}^{-1} x_k B^{-k}$ is the mantissa.*

Many programming languages use the IEEE 754 standard for floating point arithmetic. In this standard, for a double precision number, the base is $B = 2$, the mantissa has $m = 52$ bits and the exponent has 11 bits.

Two additional numbers are added to the set of floating point numbers: $\pm\infty$ and NaN (Not a Number) which is used to represent undefined or unrepresentable values.

Definition 2.2. *The machine epsilon eps is the smallest positive number which satisfies:*

$$|x - rd(x)| \leq eps|x|$$

Where $rd(x)$ is the floating point representation of x. Usually this rounding function is taken to be the nearest machine number.

Theorem 2.3. *For a floating point number system with base B and precision m, the machine epsilon is given by: The machine epsilon is given by eps $= B^{1-m}$, i.e we have:*

$$|x - rd(x)| \leq B^{1-m}|x|$$

Theorem 2.4. *Let $\star$ be one of the operations $+, -, \times, /$ and let $\circledast$ be the equivalent floating point operation, then $\forall x, y$ in the floating point system, there exists an ϵ such that:*

$$x \star y = (x \circledast y)(1 + \epsilon).$$

Definition 2.12. *Given the norms $\|\cdot\|_{(n)}$ and $\|\cdot\|_{(m)}$ on $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, we say that a matrix norm $\|\cdot\|_\star$ is* ***compatible*** *with these norms if:*

$$\|Ax\|_{(m)} \leq \|A\|_\star \|x\|_{(n)}, \quad \forall x \in \mathbb{R}^n.$$