# NLA II: Error analysis in linear equation solving

M. Sombra, 2020

## II.1 Floating point arithmetic

Ref: [D, §1.5]

The general form of floating point numbers is

$$f = \pm . d_1 d_2 \cdots d_t \times \beta^e$$

with $\beta \geq 2$ base (typically 2 or 10)

$$0 \leq d_i < \beta \qquad i = 1, \ldots, t \qquad \text{and} \qquad d_1 \neq 0$$

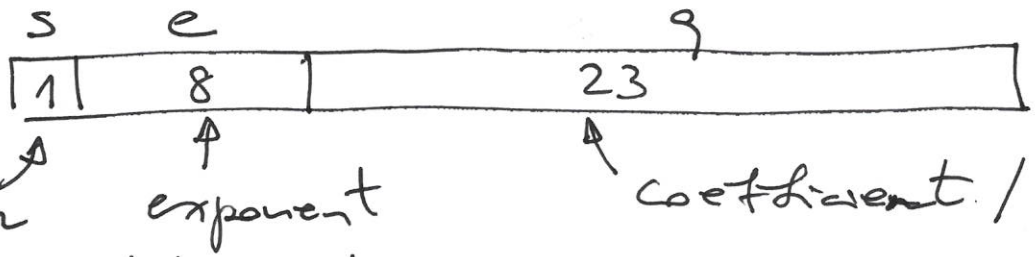
$$L \leq e \leq U$$

$L$: underflow, $U$: overflow

Hence

$$\beta^{L-1} \leq f \leq \beta^U (1 - \beta^{-t})$$

IEEE standards: single precision




The coded number is

$$f = (-1)^s (1+q)\, 2^{e-127}$$

Coresponds to $t = 24$ $\beta = 2$ $L = -126$ $U = 127$

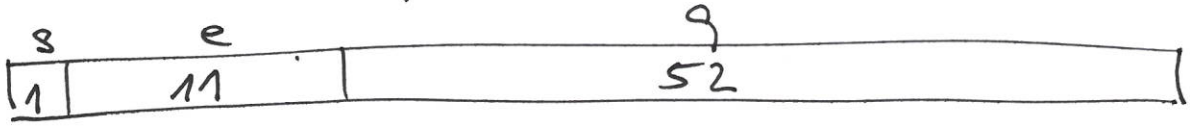
The relative maximal error is

$$2^{-24} \approx 6 \cdot 10^{-8}$$

and the range is

$$2^{-127} \approx 6 \cdot 10^{-38} \leq f \leq 2^{129}(1-2^{-24}) \approx 7 \cdot 10^{39}$$

IEEE double precision:




The coded number is

$$f = (-1)^s (1+q)\, 2^{e-1023}$$

Corresponds to $t=53$, $\beta=2$, $L=-1022$, $U=1026$

The maximal error is

$$2^{-53} \approx 6 \cdot 10^{-16}$$

and the range is

$$2^{-1022} \approx 2 \cdot 10^{308} \leq f \leq 2^{1029}(1-2^{-24}) \approx 7 \cdot 10^{309}$$

When the value of an operation cannot be represented exactly, it is defined as the closest floating point number:

$$a*b - fl(a*b) \quad \text{roundoff error}$$

~~The machine~~

The machine epsilon is a bound for the relative error of rounding:

$$\frac{|a - fl(a)|}{|a|} \leq \varepsilon = \beta^{1-t}/2$$

## II.2 Vector and matrix norms

Ref. [D, § 1.7]

Norms are used to measure errors in matrix computations

A norm on $F^n$ is a function

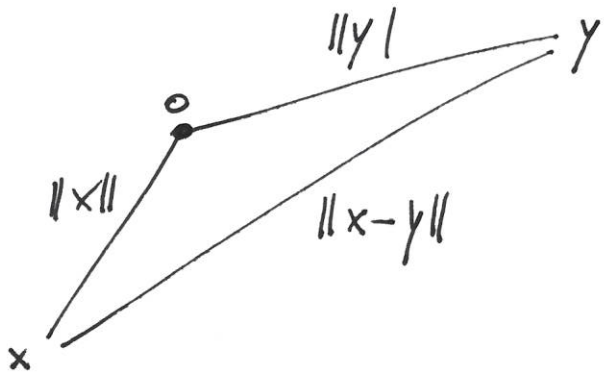
$$\|\cdot\| : F^n \to \mathbb{R}_{\geq 0}$$

s.t.

(1) $\|x\| \geq 0$ $\forall x \in F^n$ and $\|x\| = 0 \iff x = 0$

(2) $\|\alpha x\| = |\alpha| \cdot \|x\|$ $\forall \alpha \in F$ and $\forall x \in F^n$

(3) $\|x+y\| \leq \|x\| + \|y\|$ (triangle inequality)



Example: for $1 \le p \le +\infty$, the $p$-norm is

$$\|x\|_p = \begin{cases} \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} & \text{if } p < +\infty \\ \max_i |x_i| & \text{if } p = +\infty \end{cases}$$

Any two norms can be compared: $\exists c_1, c_2 > 0$ s.t.

$$\|\cdot\|_1 \le c_1 \|\cdot\|_2 \quad \text{and} \quad \|\cdot\|_2 \le c_2 \|\cdot\|_1$$

Examples

$$\|\cdot\|_2 \le \|\cdot\|_1 \le \sqrt{n}\, \|\cdot\|_2$$

$$\|\cdot\|_\infty \le \|\cdot\|_1 \le n\, \|\cdot\|_\infty$$

The space of $n \times n$ matrices is $\mathbb{F}^{n \times n}$

$\rightsquigarrow$ we can consider norms on it

A matrix norm is a norm on $\mathbb{F}^{n \times n}$ s.t.

$$\|A \cdot B\| \le \|A\| \cdot \|B\| \qquad \forall A, B$$

Example: The Frobenius norm

$$\|A\|_F = \left(\sum |a_{ij}|^2\right)^{1/2}$$

= Euclidean norm of $A$ as an element of $\mathbb{F}^{n \times n}$

**Definition**: given a vector norm on $F^n$, the associated **operator norm** is

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

It is a matrix norm

**Properties**

(1) $\|A\|_\infty = \max_i \sum_j |a_{ij}|$ (maximum row sum)

(2) $\|A\|_1 = \max_j \sum_i |a_{ij}|$ (maximum column sum)

(3) $\|A\|_2 = (\lambda_{max}(A^* A))^{1/2}$

$\lambda_{max}$: maximal eigenvalue

$A^* = \bar{A}^T$ conjugate transposed

(4) If $Q, Q'$ are **orthogonal** (if $F = \mathbb{R}$) or **unitary** (if $F = \mathbb{C}$)

$\|QAQ'\|_2 = \|A\|_2$ and $\|QAQ'\|_F = \|A\|_F$

In particular, $\|Q\|_2 = \|Q\|_F = 1$.

## II.3 Perturbation theory in linear equation solving

Ref: [D, § 2.2]

A matrix $A$ is well / badly (or: ill) conditioned if relatively small changes in $A$ cause relatively small / large changes in the solution of

$$Ax = b$$

This is measured by the condition number

Let $x$ & $\hat{x}$ be the solutions to

$$Ax = b \qquad \text{and} \qquad (A+\delta A)\hat{x} = b + \delta b$$

Fix a norm $\|\cdot\|$ on $F^n$. Our goal is to bound the norm of

$$\delta x = \hat{x} - x$$

We have

$$\begin{array}{rcl} (A+\delta A)(x+\delta x) & = & b+\delta b \\ - \quad Ax & = & b \\ \hline \delta A x + (A+\delta A)\delta x & = & \delta b \end{array}$$

Hence

$$\delta x = A^{-1}(-\delta A \hat{x} + \delta b)$$

Taking norms

$$\|\delta x\| \leq \|A^{-1}\| \left( \|\delta A\| \|\hat{x}\| + \|\delta b\| \right)$$

vector norms

operator norms

Can be rearranged to

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \kappa_{\|\cdot\|}(A) \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \cdot \|\hat{x}\|} \right) \qquad (*)$$

with $\kappa_{\|\cdot\|}(A) := \|A\| \cdot \|A^{-1}\|$ the <u>condition number</u> of $A$ with respect to $\|\cdot\|$.

The <u>relative error</u> $\frac{\|\delta x\|}{\|\hat{x}\|}$ can be controlled in terms of $\frac{\|\delta A\|}{\|A\|}$ and $\frac{\|\delta b\|}{\|A\| \cdot \|\hat{x}\|}$ via the condition number.

The relative error is closely connected with the precision of the approximation.

let $\lambda \in \mathbb{R}$ and $\hat{\lambda} = \lambda + \delta\lambda$ an approximation of $\lambda$ with $k$ correct digits in base $\beta \geq 2$.

Then

$$\lambda = \beta^e \times 0.d_1 \dots d_k d_{k+1} \dots$$

$$\hat{\lambda} = \beta^e \times 0.d_1 \dots d_k \tilde{d}_{k+1} \dots$$

and so $$\frac{|\delta\lambda|}{|\lambda|} \leq \beta^{-k}$$

or equivalently

$$\boxed{-\log_\beta \left( \frac{|\delta\lambda|}{|\lambda|} \right) \geq k}$$

For instance, round off with IEEE single precision and double precision give approximations with 24 and 53 correct bits.

Hence

single precision: $-\log_2 \frac{|\delta_s \lambda|}{|\lambda|} \geq 24$ and double precision: $-\log_2 \frac{|\delta_d \lambda|}{|\lambda|} \geq 53$

The inequality (*) translates into

$$-\log_\beta \frac{\|\delta x\|}{\|x\|} \geq -\log_\beta K(A) - \log_\beta \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \cdot \|x\|} \right)$$

Ill-conditioned matrices (= with large $K(A)$) destroy the quality of your approximations.
For instance, if you had exact data truncated with IEEE single or double precision, the computed result (with any algorithm) of $Ax = b$ will be meaningless if soon

$$K(A) > 2^{24} \approx 6 \cdot 10^8 \quad \text{(single precision)}$$

$$K(A) > 2^{53} \approx 10^{16} \quad \text{(double precision)}$$

The condition number of the 2-norm has a nice <u>geometrical interpretation</u>:

T(Eckart-Young)

$$\kappa_2(A) = \frac{1}{\text{distance}(A, \Sigma)}$$

with $\Sigma = \{A \mid \det A = 0\}$

"the condition number of A is the inverse of its distance to the set of ill-posed problems"