Analisi Numerica
Error analysis

Università di **Perugia**, **Italia**

Approximating real numbers

Approximating real numbers

Real numbers: limits of sequences in $\mathbb{Q}$ $\rightarrow$ can be approximated by rational numbers.

Approximating real numbers

Real numbers: limits of sequences in $\mathbb{Q}$ $\rightarrow$ can be approximated by rational numbers.

Trivia: diophantine numbers are $x \in \mathbb{R}$ for which there exists $c > 0$ such that for any rational number p/q,

$$\left| x - \frac{p}{q} \right| \geqslant \frac{c}{q^2}.$$

Approximating real numbers

Real numbers: limits of sequences in $\mathbb{Q}$ $\to$ can be approximated by rational numbers.

Trivia: diophantine numbers are $x \in \mathbb{R}$ for which there exists $c > 0$ such that for any rational number p/q,

$$\left| x - \frac{p}{q} \right| \geqslant \frac{c}{q^2}.$$

Rational numbers are infinitely many $\Rightarrow$ we can only deal with finite sets of numbers.

Problem

Select a finite number of representatives to approximate real numbers.

Naive attempt for numbers: **uniform distribution**. Choose $\varepsilon > 0$, N even and

$$\mathcal{E} = \left\{ \varepsilon k \,:\, -\frac{N}{2} < k \leqslant \frac{N}{2},\, k \in \mathbb{Z} \right\}.$$

Naive attempt for numbers: **uniform distribution**. Choose $\varepsilon > 0$, N even and

$$\mathcal{E} = \left\{ \varepsilon k \,:\, -\frac{N}{2} < k \leqslant \frac{N}{2},\, k \in \mathbb{Z} \right\}.$$

We can represent real numbers in the set

$$S = \left\{ -\varepsilon \frac{N}{2} < x \leqslant \varepsilon \frac{N}{2} \right\} \subset \mathbb{R}.$$

Naive attempt for numbers: **uniform distribution**. Choose $\varepsilon > 0$, N even and

$$\mathcal{E} = \left\{ \varepsilon k \ : \ -\frac{N}{2} < k \leqslant \frac{N}{2},\ k \in \mathbb{Z} \right\}.$$

We can represent real numbers in the set

$$S = \left\{ -\varepsilon \frac{N}{2} < x \leqslant \varepsilon \frac{N}{2} \right\} \subset \mathbb{R}.$$

Representation function (from real numbers to the chosen set)

$$e : S \to \mathcal{E}, \quad x \to k\varepsilon, \quad (k-1)\varepsilon < x \leqslant k\varepsilon.$$

Naive attempt for numbers: **uniform distribution**. Choose $\varepsilon > 0$, N even and

$$\mathcal{E} = \left\{ \varepsilon k \ : \ -\frac{N}{2} < k \leqslant \frac{N}{2},\ k \in \mathbb{Z} \right\}.$$

We can represent real numbers in the set

$$S = \left\{ -\varepsilon \frac{N}{2} < x \leqslant \varepsilon \frac{N}{2} \right\} \subset \mathbb{R}.$$

Representation function (from real numbers to the chosen set)

$$e : S \to \mathcal{E}, \quad x \to k\varepsilon, \quad (k-1)\varepsilon < x \leqslant k\varepsilon.$$

Representation absolute error

$$e(x) - x, \qquad \max_{x \in S} |e(x) - x| < \varepsilon$$

$$\mathcal{E} = \left\{ \varepsilon k \,:\, -\frac{N}{2} < k \leqslant \frac{N}{2},\, k \in \mathbb{Z} \right\}, \quad S = \left\{ -\varepsilon \frac{N}{2} < x \leqslant \varepsilon \frac{N}{2} \right\}$$

Representation function (from real numbers to the chosen set)

$$e : S \to \mathcal{E}, \quad x \to k\varepsilon, \quad (k-1)\varepsilon < x \leqslant k\varepsilon.$$

Representation error

$$e(x) - x, \qquad \max_{x \in S} |e(x) - x| \leqslant \varepsilon$$

This choice makes the absolute error uniformly small.

Absolute vs. Relative Error

- Absolute error $\widetilde{x} - x$.
- Relative error $\dfrac{\widetilde{x} - x}{x}$, for $x \neq 0$.

Absolute vs. Relative Error

- Absolute error $\widetilde{x} - x$.
- Relative error $\dfrac{\widetilde{x} - x}{x}$, for $x \neq 0$.

Troubles with absolute error:

- large numbers $\to$ correct digits
- small numbers $\to$ wrong digits

Absolute vs. Relative Error

- Absolute error $\widetilde{x} - x$.
- Relative error $\dfrac{\widetilde{x} - x}{x}$, for $x \neq 0$.

Troubles with absolute error:

- large numbers $\rightarrow$ correct digits
- small numbers $\rightarrow$ wrong digits

x	$\widetilde{x}$	absolute	relative
1.234	1.235	$1 \cdot 10^{-3}$	$8.1 \cdot 10^{-4}$
1234	1235	1	$8.1 \cdot 10^{-4}$
0.001234	0.001235	$1 \cdot 10^{-6}$	$8.1 \cdot 10^{-4}$

We wish to have a small relative error.

New problem

Select a finite number of representatives to approximate real numbers with a uniformly bounded relative error.

Idea: consider the digit representation of a real number.

Given a numeration basis $\beta \geqslant 2$, there is a correspondence between real numbers in $[0, 1)$ and sequences of digits

$$\mathbb{R} \Longleftrightarrow \{d_i\}_{i=1,2,\ldots}, \ d_i \in \{0, \ldots, \beta - 1\}.$$

Given a numeration basis $\beta \geqslant 2$, there is a correspondence between real numbers in $[0, 1)$ and sequences of digits

$$\mathbb{R} \Longleftrightarrow \{d_i\}_{i=1,2,\ldots},\ d_i \in \{0, \ldots, \beta - 1\}.$$

Not a bijection: two sequences may represent the same real number

Given a numeration basis $\beta \geqslant 2$, there is a correspondence between real numbers in $[0,1)$ and sequences of digits

$$\mathbb{R} \Longleftrightarrow \{d_i\}_{i=1,2,\dots},\ d_i \in \{0,\dots,\beta-1\}.$$

Not a bijection: two sequences may represent the same real number ($0.0\overline{9} = 0.1\overline{0}$).

Given a numeration basis $\beta \geqslant 2$, there is a correspondence between real numbers in $[0, 1)$ and sequences of digits

$$\mathbb{R} \Longleftrightarrow \{d_i\}_{i=1,2,\dots},\ d_i \in \{0, \dots, \beta - 1\}.$$

Not a bijection: two sequences may represent the same real number ($0.0\overline{9} = 0.1\overline{0}$).

- fixed point (natural): $x \to (\{d_i\}_{i=1,2,\dots}, M)$ (d_i is the ith digit and M says where the point is).
- floating point (scientific): $x \to (0.d_1 d_2 d_3 \cdots)\beta^p$, with $d_1 \neq 0$.

$$12.38 = 0.1238 \cdot 10^2.$$

Theorem (Basis representation theorem)

Let $x \in \mathbb{R} \setminus \{0\}$ and let $\beta \geqslant 2$ be a numeration basis, there exist unique $p \in \mathbb{Z}$ and a sequence $\{d_i\}_{i=1,2,\ldots}$ such that

Theorem (Basis representation theorem)

Let $x \in \mathbb{R} \setminus \{0\}$ and let $\beta \geqslant 2$ be a numeration basis, there exist unique $p \in \mathbb{Z}$ and a sequence $\{d_i\}_{i=1,2,\dots}$ such that

(i) $d_i \in \{0, 1, \dots, \beta - 1\}$;

Theorem (Basis representation theorem)

Let $x \in \mathbb{R} \setminus \{0\}$ and let $\beta \geqslant 2$ be a numeration basis, there exist unique $p \in \mathbb{Z}$ and a sequence $\{d_i\}_{i=1,2,\ldots}$ such that

(i) $d_i \in \{0, 1, \ldots, \beta - 1\}$;

(ii) $d_1 \neq 0$; *(To avoid ambiguity, e.g.,* $0.03 \cdot 10^2$ *and* $0.3 \cdot 10^1$*)*

Theorem (Basis representation theorem)

Let $x \in \mathbb{R} \setminus \{0\}$ and let $\beta \geqslant 2$ be a numeration basis, there exist unique $p \in \mathbb{Z}$ and a sequence $\{d_i\}_{i=1,2,\ldots}$ such that

(i) $d_i \in \{0, 1, \ldots, \beta - 1\}$;

(ii) $d_1 \neq 0$;

(iii) *d_i not definitely equal to $\beta - 1$; (There exists an infinite set of indices K such that $d_k \neq \beta - 1$ for every $k \in K$)*

Theorem (Basis representation theorem)

Let $x \in \mathbb{R} \setminus \{0\}$ and let $\beta \geqslant 2$ be a numeration basis, there exist unique $p \in \mathbb{Z}$ and a sequence $\{d_i\}_{i=1,2,\dots}$ such that

(i) $d_i \in \{0, 1, \dots, \beta - 1\}$;

(ii) $d_1 \neq 0$;

(iii) d_i *not definitely equal to* $\beta - 1$;

such that

$$x = sign(x)\beta^p \sum_{i=1}^{\infty} \beta^{-i} d_i.$$

The number $\sum_{i=1}^{\infty} \beta^{-i} d_i$ is said to be **mantissa**.

Theorem (Basis representation theorem)

Let $x \in \mathbb{R} \setminus \{0\}$ and let $\beta \geqslant 2$ be a numeration basis, there exist unique $p \in \mathbb{Z}$ and a sequence $\{d_i\}_{i=1,2,\dots}$ such that

(i) $d_i \in \{0, 1, \dots, \beta - 1\}$;

(ii) $d_1 \neq 0$;

(iii) d_i *not definitely equal to* $\beta - 1$;

such that

$$x = sign(x)\beta^p \sum_{i=1}^{\infty} \beta^{-i} d_i.$$

The number $\sum_{i=1}^{\infty} \beta^{-i} d_i$ is said to be **mantissa**.

Idea: consider numbers with finite mantissa

Given a numeration basis $\beta \geqslant 2$, the number $t > 0$ of digits of the mantissa, and m, M positive integers, we define a set of **floating point numbers**

$$\mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i, \\ -m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

These number have zero digits from the $(t+1)$-st on.

Given a numeration basis $\beta \geqslant 2$, the number $t > 0$ of digits of the mantissa, and m, M positive integers, we define a set of **floating point numbers**

$$\mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i, \\ -m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

These number have zero digits from the $(t+1)$-st on.

We will use this set to represent real numbers.

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$?

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$
- $0.03 \in \mathcal{F}$?

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$
- $0.03 \in \mathcal{F}$? yes $\Longleftarrow 0.03 = 0.30 \cdot 10^{-1}$

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$
- $0.03 \in \mathcal{F}$? yes $\Longleftarrow 0.03 = 0.30 \cdot 10^{-1}$
- $\pi \in \mathcal{F}$?

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$
- $0.03 \in \mathcal{F}$? yes $\Longleftarrow 0.03 = 0.30 \cdot 10^{-1}$
- $\pi \in \mathcal{F}$? no $\Longleftarrow \pi = 0.31415 \ldots \cdot 10^1$ (more than two digits)

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$
- $0.03 \in \mathcal{F}$? yes $\Longleftarrow 0.03 = 0.30 \cdot 10^{-1}$
- $\pi \in \mathcal{F}$? no $\Longleftarrow \pi = 0.31415\ldots \cdot 10^1$ (more than two digits)
- $12.345 \in \mathcal{F}$?

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$
- $0.03 \in \mathcal{F}$? yes $\Longleftarrow 0.03 = 0.30 \cdot 10^{-1}$
- $\pi \in \mathcal{F}$? no $\Longleftarrow \pi = 0.31415\ldots \cdot 10^1$ (more than two digits)
- $12.345 \in \mathcal{F}$? no $\Longleftarrow 12.345 = 0.12345 \cdot 10^2$ (more than two digits)

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$
- $0.03 \in \mathcal{F}$? yes $\Longleftarrow 0.03 = 0.30 \cdot 10^{-1}$
- $\pi \in \mathcal{F}$? no $\Longleftarrow \pi = 0.31415\ldots \cdot 10^1$ (more than two digits)
- $12.345 \in \mathcal{F}$? no $\Longleftarrow 12.345 = 0.12345 \cdot 10^2$ (more than two digits)
- $1000 \in \mathcal{F}$?

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$
- $0.03 \in \mathcal{F}$? yes $\Longleftarrow 0.03 = 0.30 \cdot 10^{-1}$
- $\pi \in \mathcal{F}$? no $\Longleftarrow \pi = 0.31415 \ldots \cdot 10^1$ (more than two digits)
- $12.345 \in \mathcal{F}$? no $\Longleftarrow 12.345 = 0.12345 \cdot 10^2$ (more than two digits)
- $1000 \in \mathcal{F}$? no $\Longleftarrow 1000 = 0.10 \cdot 10^4$ (too big)

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$
- $0.03 \in \mathcal{F}$? yes $\Longleftarrow 0.03 = 0.30 \cdot 10^{-1}$
- $\pi \in \mathcal{F}$? no $\Longleftarrow \pi = 0.31415\ldots \cdot 10^1$ (more than two digits)
- $12.345 \in \mathcal{F}$? no $\Longleftarrow 12.345 = 0.12345 \cdot 10^2$ (more than two digits)
- $1000 \in \mathcal{F}$? no $\Longleftarrow 1000 = 0.10 \cdot 10^4$ (too big)
- $1/10000 \in \mathcal{F}$?

As an example

$$\mathcal{F} := \mathcal{F}(10, 2, 2, 3)$$

contains numbers with 2 digits

- $12 \in \mathcal{F}$? yes $\Longleftarrow 12 = 0.12 \cdot 10^2$
- $0.03 \in \mathcal{F}$? yes $\Longleftarrow 0.03 = 0.30 \cdot 10^{-1}$
- $\pi \in \mathcal{F}$? no $\Longleftarrow \pi = 0.31415\ldots \cdot 10^1$ (more than two digits)
- $12.345 \in \mathcal{F}$? no $\Longleftarrow 12.345 = 0.12345 \cdot 10^2$ (more than two digits)
- $1000 \in \mathcal{F}$? no $\Longleftarrow 1000 = 0.10 \cdot 10^4$ (too big)
- $1/10000 \in \mathcal{F}$? no $\Longleftarrow 0.001 = 0.10 \cdot 10^{-3}$ (too small in modulus)

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$
$$-m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \dots, t,\ d_1 \neq 0 \Big\}.$$

$\mathcal{F} \subset \mathbb{R}$ is a **finite set** with cardinality

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$
$$- m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

$\mathcal{F} \subset \mathbb{R}$ is a **finite set** with cardinality

$$1 + 2(m + M + 1)(\beta - 1)\beta^{t-1}$$

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$

$$-\, m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \dots, t,\ d_1 \neq 0 \Big\}.$$

$\mathcal{F} \subset \mathbb{R}$ is a **finite set** with cardinality

$$1 + 2(m + M + 1)(\beta - 1)\beta^{t-1}$$

The **largest** number in $\mathcal{F}$ is

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$

$$- m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

$\mathcal{F} \subset \mathbb{R}$ is a **finite set** with cardinality

$$1 + 2(m + M + 1)(\beta - 1)\beta^{t-1}$$

The **largest** number in $\mathcal{F}$ is

$$\Omega = \beta^M \sum_{i=1}^{t} \beta^{-i}(\beta - 1)$$

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$
$$-\, m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

$\mathcal{F} \subset \mathbb{R}$ is a **finite set** with cardinality

$$1 + 2(m + M + 1)(\beta - 1)\beta^{t-1}$$

The **largest** number in $\mathcal{F}$ is

$$\Omega = \beta^M \sum_{i=1}^{t} \beta^{-i}(\beta - 1) = \beta^M (1 - \beta^{-t}).$$

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm\beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$

$$-\, m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

$\mathcal{F} \subset \mathbb{R}$ is a **finite set** with cardinality

$$1 + 2(m + M + 1)(\beta - 1)\beta^{t-1}$$

The **largest** number in $\mathcal{F}$ is

$$\Omega = \beta^M \sum_{i=1}^{t} \beta^{-i}(\beta - 1) = \beta^M(1 - \beta^{-t}).$$

The **smallest positive** number in $\mathcal{F}$ is

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$
$$-m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

$\mathcal{F} \subset \mathbb{R}$ is a **finite set** with cardinality

$$1 + 2(m + M + 1)(\beta - 1)\beta^{t-1}$$

The **largest** number in $\mathcal{F}$ is

$$\Omega = \beta^M \sum_{i=1}^{t} \beta^{-i}(\beta - 1) = \beta^M(1 - \beta^{-t}).$$

The **smallest positive** number in $\mathcal{F}$ is

$$\omega = \beta^{-m}\beta^{-1} = \beta^{-m-1}.$$

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$

$$-m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

$\mathcal{F} \subset \mathbb{R}$ is a **finite set** with cardinality

$$1 + 2(m + M + 1)(\beta - 1)\beta^{t-1}$$

The **largest** number in $\mathcal{F}$ is

$$\Omega = \beta^M \sum_{i=1}^{t} \beta^{-i}(\beta - 1) = \beta^M(1 - \beta^{-t}).$$

The **smallest positive** number in $\mathcal{F}$ is

$$\omega = \beta^{-m}\beta^{-1} = \beta^{-m-1}.$$

For $x \geqslant \Omega$ we cannot represent numbers, for $0 < x < \omega$ we make a large relative error.

$\mathcal{F}(2,2,1,1)$ is made of

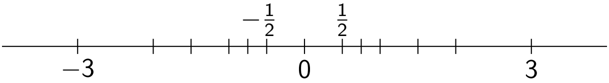
$\mathcal{F}(2,2,1,1)$ is made of 13 numbers

$$\begin{array}{ccc} 0 & & \\ \pm 0.10 \cdot 2^{-1} & \pm 0.10 \cdot 2^{0} & \pm 0.10 \cdot 2^{1} \\ \pm 0.11 \cdot 2^{-1} & \pm 0.11 \cdot 2^{0} & \pm 0.11 \cdot 2^{1} \end{array}$$

$\mathcal{F}(2,2,1,1)$ is made of 13 numbers

$$\begin{array}{ccc} 0 & & \\ \pm 0.10 \cdot 2^{-1} & \pm 0.10 \cdot 2^{0} & \pm 0.10 \cdot 2^{1} \\ \pm 0.11 \cdot 2^{-1} & \pm 0.11 \cdot 2^{0} & \pm 0.11 \cdot 2^{1} \end{array}$$

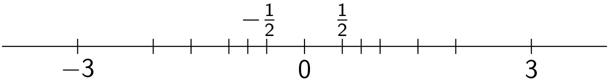
They are **not uniform**: between $1/2$ and $1/4$ and between $1/2$ and 1 there is the same number of elements of $\mathcal{F}$.



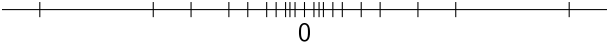
$\mathcal{F}(2,2,1,1)$ is made of 13 numbers

$$\begin{array}{ccc} 0 & & \\ \pm 0.10 \cdot 2^{-1} & \pm 0.10 \cdot 2^{0} & \pm 0.10 \cdot 2^{1} \\ \pm 0.11 \cdot 2^{-1} & \pm 0.11 \cdot 2^{0} & \pm 0.11 \cdot 2^{1} \end{array}$$

They are **not uniform**: between $1/2$ and $1/4$ and between $1/2$ and 1 there is the same number of elements of $\mathcal{F}$.



$\mathcal{F}(2,2,2,2)$ has 21 numbers



$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm\beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$

$$- m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

Given $S = \{x \in \mathbb{R} \,:\, \omega \leqslant x \leqslant \Omega\}$, we construct a **representation function**

$$\mathrm{fl} : \mathbb{R} \longrightarrow \mathcal{F} \cup \{\pm\infty\}$$

with one of the two rules, where $x = \beta^p \sum_{i=1}^{\infty} \beta^{-i} d_i \in S$,

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$

$$-m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

Given $S = \{x \in \mathbb{R} : \omega \leqslant x \leqslant \Omega\}$, we construct a **representation function**

$$\mathrm{fl} : \mathbb{R} \longrightarrow \mathcal{F} \cup \{\pm\infty\}$$

with one of the two rules, where $x = \beta^p \sum_{i=1}^{\infty} \beta^{-i} d_i \in S$,

- **truncation**: $\widetilde{x} = \mathrm{fl}(x) = \beta^p \sum_{i=1}^{t} \beta^{-i} d_i$.

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm \beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$
$$-m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \dots, t,\ d_1 \neq 0 \Big\}.$$

Given $S = \{x \in \mathbb{R} \,:\, \omega \leqslant x \leqslant \Omega\}$, we construct a **representation function**

$$\mathrm{fl} : \mathbb{R} \ \longrightarrow\ \mathcal{F} \cup \{\pm\infty\}$$

with one of the two rules, where $x = \beta^p \sum_{i=1}^{\infty} \beta^{-i} d_i \in S$,

- **truncation**: $\widetilde{x} = \mathrm{fl}(x) = \beta^p \sum_{i=1}^{t} \beta^{-i} d_i$.
- **rounding**: $\widetilde{x} = \mathrm{fl}(x)$ the truncation of $x + \beta^{p-t-1}/2$.

$$\mathcal{F} := \mathcal{F}(\beta, t, m, M) = \{0\} \cup \Big\{ \pm\beta^p \sum_{i=1}^{t} \beta^{-i} d_i,$$

$$-m \leqslant p \leqslant M,\ 0 \leqslant d_i < \beta \text{ integer for } i = 1, \ldots, t,\ d_1 \neq 0 \Big\}.$$

Given $S = \{x \in \mathbb{R} : \omega \leqslant x \leqslant \Omega\}$, we construct a **representation function**

$$\mathrm{fl} : \mathbb{R} \longrightarrow \mathcal{F} \cup \{\pm\infty\}$$

with one of the two rules, where $x = \beta^p \sum_{i=1}^{\infty} \beta^{-i} d_i \in S$,

- **truncation**: $\widetilde{x} = \mathrm{fl}(x) = \beta^p \sum_{i=1}^{t} \beta^{-i} d_i$.
- **rounding**: $\widetilde{x} = \mathrm{fl}(x)$ the truncation of $x + \beta^{p-t-1}/2$.

If $x > \Omega$ we set $\mathrm{fl}(x) = \infty$ (**overflow**), if $0 \leqslant x < \omega$ we set $\mathrm{fl}(x) = 0$ (**underflow**) and for negative numbers the extension is straightforward.

We define the **machine precision** as $u = \beta^{-t+1}$ in the case of truncation and $u = \beta^{-t+1}/2$ in the case of rounding.

Theorem

Let $x \in S$ then we have the following bound for the relative error

$$\left| \frac{\mathrm{fl}(x) - x}{x} \right| < u.$$

Proof. For the truncation

We define the **machine precision** as $u = \beta^{-t+1}$ in the case of truncation and $u = \beta^{-t+1}/2$ in the case of rounding.

Theorem

Let $x \in S$ then we have the following bound for the relative error

$$\left|\frac{\mathrm{fl}(x) - x}{x}\right| < u.$$

Proof. For the truncation

$$\frac{|\,\mathrm{fl}(x) - x|}{|x|} = \frac{x - \mathrm{fl}(x)}{x}$$

Since $x \geqslant \mathrm{fl}(x)$.

We define the **machine precision** as $u = \beta^{-t+1}$ in the case of truncation and $u = \beta^{-t+1}/2$ in the case of rounding.

Theorem

Let $x \in S$ then we have the following bound for the relative error

$$\left| \frac{\mathrm{fl}(x) - x}{x} \right| < u.$$

Proof. For the truncation

$$\frac{|\,\mathrm{fl}(x) - x|}{|x|} = \frac{x - \mathrm{fl}(x)}{x} = \frac{\beta^p \sum_{i>t} \beta^{-i} d_i}{\beta^p \sum_{i=1}^{\infty} \beta^{-i} d_i}$$

By definition: $\sum_{i=1}^{\infty} \beta^{-i} d_i - \sum_{i=1}^{t} \beta^{-i} d_i = \sum_{i=t+1}^{\infty} \beta^{-i} d_i$.

We define the **machine precision** as $u = \beta^{-t+1}$ in the case of truncation and $u = \beta^{-t+1}/2$ in the case of rounding.

Theorem

Let $x \in S$ then we have the following bound for the relative error

$$\left| \frac{\mathrm{fl}(x) - x}{x} \right| < u.$$

Proof. For the truncation

$$\frac{|\,\mathrm{fl}(x) - x|}{|x|} = \frac{x - \mathrm{fl}(x)}{x} = \frac{\cancel{\beta^p} \sum_{i>t} \beta^{-i} d_i}{\cancel{\beta^p} \sum_{i=1}^{\infty} \beta^{-i} d_i}$$

We define the **machine precision** as $u = \beta^{-t+1}$ in the case of truncation and $u = \beta^{-t+1}/2$ in the case of rounding.

Theorem

Let $x \in S$ then we have the following bound for the relative error

$$\left|\frac{\mathrm{fl}(x) - x}{x}\right| < u.$$

Proof. For the truncation

$$\frac{|\,\mathrm{fl}(x) - x|}{|x|} = \frac{x - \mathrm{fl}(x)}{x} = \frac{\sum_{i>t} \beta^{-i} d_i}{\sum_{\mathbf{i=1}}^{\infty} \beta^{-\mathbf{i}} \mathbf{d_i}} \leqslant \frac{\beta^{-t}}{\beta^{-1}} = u.$$

$$\sum_{i=1}^{\infty} \beta^{-i} d_i \geqslant \beta^{-1}.$$

We define the **machine precision** as $u = \beta^{-t+1}$ in the case of truncation and $u = \beta^{-t+1}/2$ in the case of rounding.

Theorem

Let $x \in S$ then we have the following bound for the relative error

$$\left| \frac{\mathrm{fl}(x) - x}{x} \right| < u.$$

Proof. For the truncation

$$\frac{|\,\mathrm{fl}(x) - x|}{|x|} = \frac{x - \mathrm{fl}(x)}{x} = \frac{\sum_{i>t} \beta^{-i} d_i}{\sum_{i=1}^{\infty} \beta^{-i} d_i} \leqslant \frac{\beta^{-t}}{\beta^{-1}} = u.$$

$$\sum_{i=t+1}^{\infty} \beta^{-i} d_i \leqslant (\beta - 1) \sum_{i=t+1}^{\infty} \beta^{-i} = (\beta - 1)\Big(\sum_{i=0}^{\infty} \beta^{-i} - \sum_{i=0}^{t} \beta^{-i}\Big)$$

$$= (\beta - 1)\Big(\frac{1}{1 - 1/\beta} - \frac{1 - \beta^{-t-1}}{1 - 1/\beta}\Big) = \frac{\beta - 1}{\frac{\beta-1}{\beta}} \beta^{-t-1} = \beta^{-t}.$$

We define the **machine precision** as $u = \beta^{-t+1}$ in the case of truncation and $u = \beta^{-t+1}/2$ in the case of rounding.

Theorem

Let $x \in S$ then we have the following bound for the relative error

$$\left| \frac{\mathrm{fl}(x) - x}{x} \right| < u.$$

Proof. For the truncation

$$\frac{|\,\mathrm{fl}(x) - x|}{|x|} = \frac{x - \mathrm{fl}(x)}{x} = \frac{\sum_{i>t} \beta^{-i} d_i}{\sum_{i=1}^{\infty} \beta^{-i} d_i} \leqslant \frac{\beta^{-t}}{\beta^{-1}} = u.$$

Floating point on a computer

For the single precision floating point (`float` 32 bits)

$$\mathcal{F}(2, 24, 126, 127), \qquad u = 2^{-24}.$$

Floating point on a computer

For the single precision floating point (`float` 32 bits)

$$\mathcal{F}(2, 24, 126, 127), \qquad u = 2^{-24}.$$

For the double precision floating point (`double` 64 bits)

$$\mathcal{F}(2, 52, 1022, 1023), \qquad u = 2^{-52} \approx 2.2 \cdot 10^{-16}.$$

Bad news: $\mathcal{F}$ has **few algebraic properties**, e.g.,

$$x, y \in \mathcal{F} \not\Rightarrow x + y \in \mathcal{F}$$

We must define floating point operations.

Bad news: $\mathcal{F}$ has **few algebraic properties**, e.g.,

$$x, y \in \mathcal{F} \not\Rightarrow x + y \in \mathcal{F}$$

We must define floating point operations.

We may assume that there exists a floating point sum $\oplus$ such that, if $x, y \in \mathcal{F}$, then (if overflow does not occur)

$$x \oplus y \in \mathcal{F}, \qquad x \oplus y = (x + y)(1 + \varepsilon), \qquad |\varepsilon| < u$$

Similarly we define $\otimes$, $\ominus$, $\oslash$.

Bad news: $\mathcal{F}$ has **few algebraic properties**, e.g.,

$$x, y \in \mathcal{F} \not\Rightarrow x + y \in \mathcal{F}$$

We must define floating point operations.

We may assume that there exists a floating point sum $\oplus$ such that, if $x, y \in \mathcal{F}$, then (if overflow does not occur)

$$x \oplus y \in \mathcal{F}, \qquad x \oplus y = (x + y)(1 + \varepsilon), \qquad |\varepsilon| < u$$

Similarly we define $\otimes$, $\ominus$, $\oslash$.

A simple idea is to define, for instance, $x \otimes y = \mathrm{fl}(x + y)$, but details are more complicate.

Floating point operations verify just some properties

- $x \oplus y = y \oplus x$ (commutativity of the sum);
- $x \otimes y = y \otimes x$ (commutativity of the product);
- $x \oslash x = 1$.

Floating point operations verify just some properties

- $x \oplus y = y \oplus x$ (commutativity of the sum);
- $x \otimes y = y \otimes x$ (commutativity of the product);
- $x \oslash x = 1$.

They do not verify others

- associativity of the sum and product;
- distributive law;
- simplification (it may happen $x \otimes (y \oslash x) \neq y$);
- no null factor law (it may happen $x \otimes y = z \otimes y$ with $y \neq 0$ and $x \neq z$)

Well-posed problems

Which problems are we interested in?

Well-posed problems

Which problems are we interested in?

But first. . .

Well-posed problems

What is a problem?

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 1: the **linear system** problem.

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 1: the **linear system** problem.

Given a matrix A (coefficient) and a vector b (right hand side),

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 1: the **linear system** problem.

Given a matrix A (coefficient) and a vector b (right hand side), find all vectors x (unknown)

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 1: the **linear system** problem.

Given a matrix A (coefficient) and a vector b (right hand side), find all vectors x (unknown) such that $Ax = b$.

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 2: the **polynomial equation** problem.

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 2: the **polynomial equation** problem.

Given a polynomial $p(x)$,

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 2: the **polynomial equation** problem.

Given a polynomial $p(x)$, find all complex numbers x (unknowns)

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 2: the **polynomial equation** problem.

Given a polynomial $p(x)$, find all complex numbers x (unknowns) such that $p(x) = 0$.

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 3: the **integration**.

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 3: the **integration**.

Given a continuous function $f(x)$ (integrand) and two real numbers $a \leqslant b$ (extremes),

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 3: the **integration**.

Given a continuous function $f(x)$ (integrand) and two real numbers $a \leqslant b$ (extremes), find a real number I (unknown)

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

Example 3: the **integration**.

Given a continuous function $f(x)$ (integrand) and two real numbers $a \leqslant b$ (extremes), find a real number I (unknown) such that $I = \int_a^b f(x)dx$.

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

In all cases the unknowns are functions of the data.

Well-posed problems

What is a problem?

A problem is made of three parts

- Data
- Unknowns
- Conditions

In all cases the unknowns are functions of the data.

Solutions of problems $\Longleftrightarrow$ Evaluation of a function

In practical cases, only an **approximation** can be provided.

Well-posed problems

Which problems are we interested in?

Well-posed problems

-
-
-

Well-posed problems

Which problems are we interested in?

Well-posed problems

- They have a solution
-
-

Well-posed problems

Which problems are we interested in?

Well-posed problems

- They have a solution (or else what are we computing?)
-
-

Well-posed problems

Which problems are we interested in?

Well-posed problems

- They have a solution
- The solution is unique
-

Well-posed problems

Which problems are we interested in?

Well-posed problems

- They have a solution
- The solution is unique (or else what is the correct answer?)
-

Well-posed problems

Which problems are we interested in?

Well-posed problems

- They have a solution
- The solution is unique
- The solution depends continuously from the data

Well-posed problems

Which problems are we interested in?

Well-posed problems

- They have a solution
- The solution is unique
- The solution depends continuously from the data

Continuous dependence from data is (less obviously) important

- Data in real problems are affected by errors
- Computation is made on finite arithmetic and there are rounding errors

Continuous dependence from data

Consider a problem with data and unknowns in the Banach spaces V and W, respectively.

The problem **strongly** continuously depends from data at $A \in V$ if there exists a neighborhood $\mathcal{U} \subset V$ of A such that the problem has a unique solution $X(B)$ for $B \in \mathcal{U}$ and $\lim_{B \to A} X(B) = X(A)$.

The problem **weakly** continuously depends from data at $A \in V$ if for $B \in V$, there exists $t_0 > 0$ such that the problem with data $A + tB$ has a unique solution $X(t)$ for $t \in [0, t_0)$ and $\lim_{t \to 0^+} X(t) = X(0)$.

Strong dependence implies weak dependence, but not the contrary.

Weak dependence is more frequent and sometimes it is enough in applications.

Implicit functions theorem

A problem is often stated as an implicit equation (e. g. systems of equations, zeros of polynomials, ODEs, PDEs).

Theorem (Implicit functions)

Let $F : \Omega \to \mathbb{R}^n$, *with* $F \in C^1(\Omega)$, $\Omega \in \mathbb{R}^n \times \mathbb{R}^m$ *such that* $F_x(x_0, y_0) : \mathbb{R}^n \to \mathbb{R}^n$ *is invertible for* $(x_0, y_0) \in \Omega$[a]*.*
There exist neighborhoods $\mathcal{U} \ni (x_0, y_0)$ *and* $\mathcal{V} \ni y_0$ *and* $g : \mathcal{V} \to \mathbb{R}^n$, *with* $g \in C^1(\mathcal{V})$ *such that for* $(x, y) \in \mathcal{U}$

$$F(x, y) = F(x_0, y_0) \Longleftrightarrow x = g(y).$$

[a] we use the notation $F(x, y)$ with $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$

An analogous results holds for $F : \Omega \to \mathbb{C}^n$ and F analytic in $\Omega \in \mathbb{C}^n \times \mathbb{C}^m$. In this case g is analytic as well.

Example: square linear system

Theorem

Let $A \in \mathbb{C}^{n \times n}$. The linear system $Ax = b$ is **strongly** *well posed for $b \in \mathbb{C}^n$ if and only A is invertible.*

Proof.

$\square$

Example: square linear system

Theorem

Let $A \in \mathbb{C}^{n\times n}$. The linear system $Ax = b$ is **strongly** *well posed for $b \in \mathbb{C}^n$ if and only A is invertible.*

Proof.

Since $(A, b) \in \mathbb{C}^{n\times n} \times \mathbb{C}^n \cong \mathbb{C}^{n^2+n}$, we can see data as belonging to a vector space.

Let $F : \mathbb{C}^{n^2+n} \times \mathbb{C}^n \to \mathbb{C}^n$ be such that $F((A, b), x) = Ax - b$.
Writing

$$F((A,b),x) = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n - b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n - b_2 \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n - b_n \end{bmatrix},$$

we see that F is differentiable, since its components are polynomials.

Example: square linear system

Theorem

Let $A \in \mathbb{C}^{n\times n}$. The linear system $Ax = b$ is **strongly** *well posed for $b \in \mathbb{C}^n$ if and only A is invertible.*

Proof.

The derivative with respect to x_i is $\frac{\partial F_i}{\partial F_j} = a_{ij}$ and thus the Jacobian matrix is A and we can say that $F_x((A, b), x)[h] = Ah$, with $h \in \mathbb{C}^n$.

The function F is differetiable with F_x invertible, the implicit functions theorem implies that there exists a neighborhood of $((A, b), x)$ such that $Ax = b$, where

$$\widetilde{A}\widetilde{x} - \widetilde{b} = F((\widetilde{A}, \widetilde{b}), \widetilde{x}) = F((A, b), x) = Ax - b = 0$$

if and only if $\widetilde{x} = g(\widetilde{A}, \widetilde{b})$ with g differentiable. (Note that $g(\widetilde{A}, \widetilde{b}) = \widetilde{A}^{-1}\widetilde{b}$.) $\square$

Example: square linear system

Theorem

Let $A \in \mathbb{C}^{n\times n}$. The linear system $Ax = b$ is **strongly** *well posed for $b \in \mathbb{C}^n$ if and only A is invertible.*

Proof.

The converse is left as an exercise. $\square$

The theorem can be proved using only linear algebra.

Example: square linear system

The linear system $Ax = b$ with A square and invertible.

Example: square linear system

The linear system $Ax = b$ with A square and invertible.

Is the problem well posed? The solution exists and is unique (Cramer's theorem).

Example: square linear system

The linear system $Ax = b$ with A square and invertible.

Is the problem well posed? The solution exists and is unique (Cramer's theorem).

The solution is $x = A^{-1}b$. Let $(A^{-1})_{ij} = \widetilde{a}_{ij}$, for $i, j = 1, \ldots, n$, be the elements of the inverse

$$x_i = \sum_{j=1}^{n} \widetilde{a}_{ij} b_j$$

Example: square linear system

The linear system $Ax = b$ with A square and invertible.

Is the problem well posed? The solution exists and is unique (Cramer's theorem).

The solution is $x = A^{-1}b$. Let $(A^{-1})_{ij} = \widetilde{a}_{ij}$, for $i, j = 1, \ldots, n$, be the elements of the inverse

$$x_i = \sum_{j=1}^{n} \widetilde{a}_{ij} b_j$$

with (adjoint's formula, **inefficient!**)

$$\widetilde{a}_{ij} = \frac{1}{\det(A)} (-1)^{i+j} \det(M^{(ij)}),$$

where $M^{(ij)}$ is obtained removing the i-th row and the j-th column from A^T.

Example: square linear system

The linear system $Ax = b$ with A square and invertible.

The solution is $x = A^{-1}b$.

$$x_i = \sum_{k=1}^{n} \widetilde{a}_{ij} b_j, \qquad \widetilde{a}_{ij} = \frac{1}{\det(A)}(-1)^{i+j} \det(M^{(ij)}),$$

Example: square linear system

The linear system $Ax = b$ with A square and invertible.

The solution is $x = A^{-1}b$.

$$x_i = \sum_{k=1}^{n} \widetilde{a}_{ij} b_j, \qquad \widetilde{a}_{ij} = \frac{1}{\det(A)} (-1)^{i+j} \det(M^{(ij)}),$$

Note that

- the first formula is a polynomial;
- $\det(X)$ is a polynomial of the entries of X;
- the second formula is a rational function.

Example: square linear system

The linear system $Ax = b$ with A square and invertible.

The solution is $x = A^{-1}b$.

$$x_i = \sum_{k=1}^{n} \widetilde{a}_{ij} b_j, \qquad \widetilde{a}_{ij} = \frac{1}{\det(A)} (-1)^{i+j} \det(M^{(ij)}),$$

Note that

- the first formula is a polynomial;
- $\det(X)$ is a polynomial of the entries of X;
- the second formula is a rational function.

Thus, x_i is a rational function (continuous) of the data. We can write

$$x = f(a_{11}, \dots, a_{nn}, b_1, \dots, b_n), \qquad f : \mathbb{K}^{n^2+n} \to \mathbb{K}^n.$$

We have translated a problem to a function.

Example: roots of polynomials

The roots of the polynomial

$$az^2 + bz + c = 0,$$

can be written as

$$z_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \qquad z_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a},$$

where the complex square root is the one with positive real part (or positive imaginary part if the real part is 0).

They are analytic functions of the coefficients if $b^2 - 4ac \neq 0$ in a neighborhood of the polynomial, while for $b^2 - 4ac = 0$ they loose analyticy (and continuity in some sense).

Example: roots of polynomials

Theorem

Let $p(z) \in \mathbb{C}[z]$. The polynomial equations $p(z) = 0$ is a **strongly** *well posed problem if p has distinct roots.*

Proof.

$\square$

Example: roots of polynomials

Theorem

Let $p(z) \in \mathbb{C}[z]$. The polynomial equations $p(z) = 0$ is a **strongly** *well posed problem if p has distinct roots.*

Proof.

We prove the theorem by induction on the degree n of p.

If $n = 1$ then $p(z) = a_1 z + a_0$ has one distinct root α. If $\widetilde{p}(z) = \widetilde{a}_1 z + \widetilde{a}_0$ lies in a neighboor of $p(z)$ with $\widetilde{a}_1 \neq 0$ then $\widetilde{p}(z)$ has a unique root $\widetilde{\alpha} = -\widetilde{a}_0/\widetilde{a}_1$. Note that $\widetilde{\alpha}$ is a continuous function of the coefficients of $\widetilde{p}$ and $\lim_{\widetilde{p} \to p} \widetilde{\alpha} = -a_0/a_1 = \alpha$.[a]

$\square$

[a] We have that $\widetilde{p} \to p$, if and only if $\widetilde{a}_0 \to a_0, \ldots, \widetilde{a}_n \to a_n$.

Example: roots of polynomials

Theorem

Let $p(z) \in \mathbb{C}[z]$. The polynomial equations $p(z) = 0$ is a **strongly** *well posed problem if p has distinct roots.*

Proof.

If $p(z) = a_0 + a_1 z + \cdots + a_n z^n$, and α is a root of p, then

$$p(z) = (z - \alpha) \underbrace{\left(b_0 + b_1 z + \cdots + b_{n-1} z^{n-1}\right)}_{q(z)},$$

for a unique $q(z)$, from which we get the system of equations

$$\begin{cases} a_0 + \alpha b_0 = 0, \\ a_1 + \alpha b_1 - b_0 = 0, \\ \quad \vdots \\ a_{n-1} + \alpha b_{n-1} - b_{n-2} = 0, \\ a_n - b_{n-1} = 0. \end{cases}$$

Example: roots of polynomials

Theorem

Let $p(z) \in \mathbb{C}[z]$. The polynomial equations $p(z) = 0$ is a **strongly** *well posed problem if p has distinct roots.*

Proof.

$$\begin{cases} a_0 + \alpha b_0 = 0, \\ a_1 + \alpha b_1 - b_0 = 0, \\ \quad \vdots \\ a_{n-1} + \alpha b_{n-1} - b_{n-2} = 0, \\ a_n - b_{n-1} = 0. \end{cases}$$

can be written as $F(a_0, \ldots, a_n, \alpha, b_0, \ldots, b_{n-1}) = 0$. We have that $F : \mathbb{C}^{n+1} \times \mathbb{C}^{n+1} \to \mathbb{C}^{n+1}$ is differentiable (it is a polynomial) and

$$\frac{\partial F}{\partial \alpha} = \begin{bmatrix} b_0 \\ \vdots \\ b_{n-1} \\ 0 \end{bmatrix} \in \mathbb{C}^{n+1}, \qquad \frac{\partial F}{\partial b} = \begin{bmatrix} \alpha & 0 & \cdots & 0 \\ -1 & \alpha & \ddots & \vdots \\ 0 & -1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \alpha \\ 0 & \cdots & 0 & -1 \end{bmatrix} \in \mathbb{C}^{(n+1)\times n}.$$

Example: roots of polynomials

Theorem

Let $p(z) \in \mathbb{C}[z]$. The polynomial equations $p(z) = 0$ is a **strongly** *well posed problem if p has distinct roots.*

Proof.

The partial Fréchet derivative is associated with the matrix

$$F_{\alpha, b_0, \dots, b_{n-1}} = \begin{bmatrix} \frac{\partial F}{\partial \alpha} & \frac{\partial F}{\partial b_{n-1}} \end{bmatrix} \in \mathbb{C}^{(n+1)\times(n+1)}.$$

We claim that $F_{\alpha, b_0, \dots, b_{n-1}}$ is invertible if and only if $q(\alpha) \neq 0$, that is α is not a root of q.

Since the last row of $F_{\alpha, b_0, \dots, b_{n-1}}$ contains just a -1 in the last position, it is sufficient to consider

$$M := \begin{bmatrix} b_0 & \alpha & 0 & \cdots & 0 \\ b_1 & -1 & \alpha & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & 0 \\ b_{n-1} & \vdots & \ddots & \ddots & \alpha \\ 0 & 0 & \cdots & 0 & -1 \end{bmatrix}$$

Example: roots of polynomials

Theorem

Let $p(z) \in \mathbb{C}[z]$. The polynomial equations $p(z) = 0$ is a **strongly** *well posed problem if p has distinct roots.*

Proof.

$$M := \begin{bmatrix} b_0 & \alpha & 0 & \cdots & 0 \\ b_1 & -1 & \alpha & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \alpha \\ b_{n-1} & 0 & \cdots & 0 & -1 \end{bmatrix}$$

We prove that M is singular if and only if $q(\alpha) = 0$. If $q(\alpha) = 0$, then

$$\begin{aligned} &\begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} \end{bmatrix} M \\ &= \begin{bmatrix} b_0 + b_1\alpha + b_2\alpha^2 + \cdots + b_{n-1}\alpha^{n-1} & \alpha - \alpha & \cdots & \alpha^{n-1} - \alpha^{n-1} \end{bmatrix} \\ &= \begin{bmatrix} q(\alpha) & 0 & \cdots & 0 \end{bmatrix} = 0. \end{aligned}$$

We have proved that M^T has a nonzero kernel, this implies that M^T is singular and thus M is singular. □

Example: roots of polynomials

Theorem

Let $p(z) \in \mathbb{C}[z]$. The polynomial equations $p(z) = 0$ is a **strongly** *well posed problem if p has distinct roots.*

Proof.

On the contrary, if M is singular, then there exists a nonzero vector $v = \begin{bmatrix} y_0 & y_1 & \cdots & y_{n-1} \end{bmatrix}$ such that $vM = 0$.
The equation $vM = 0$ is equivalent to

$$\begin{cases} b_0 y_0 + b_1 y_1 + \cdots + b_{n-1} y_{n-1} = 0 \\ \alpha y_0 - y_1 = 0 \\ \alpha y_1 - y_2 = 0 \\ \vdots \\ \alpha y_{n-2} - y_{n-1} = 0, \end{cases}$$

that gives $y_1 = \alpha y_0, \ldots, y_{n-1} = \alpha^{n-1} y_0$ and

$$(b_0 + b_1 \alpha + \cdots + b_{n-1} \alpha^{n-1}) y_0 = q(\alpha) y_0 = 0$$

Since $y_0 \neq 0$ (or else $v = 0$) we obtain $q(\alpha) = 0$. $\square$

Example: roots of polynomials

Theorem

Let $p(z) \in \mathbb{C}[z]$. The polynomial equations $p(z) = 0$ is a **strongly** *well posed problem if p has distinct roots.*

Proof.

The hypotheses of the implicit functions theorem are fulfilled if $q(\alpha) \neq 0$.

For any $\widetilde{p}(z)$ in a neighborhood of $p(z)$ there exists $\widetilde{\alpha}_1$ and $\widetilde{q}(z)$ such that $\widetilde{p}(z) = (z - \widetilde{\alpha})\widetilde{q}(z)$ and $\widetilde{\alpha}_1$ and $\widetilde{b}_0, \ldots, \widetilde{b}_{n-1}$ are analytic functions of the coefficients of $\widetilde{p}(z)$.

The polynomial $q(z)$ has distinct roots. By induction, there exists $\widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_n$ that are solutions of $\widetilde{q}(z)$ (in a smaller neighborhood if necessary) and that are analytic with respect to the coefficients of $\widetilde{p}(z)$. Since the composition of analytic functions is analytic the roots are analytic functions of the coefficients of $\widetilde{p}(z)$. $\square$

Example: roots of polynomials

The roots of the polynomial

$$x^2 - \alpha = 0, \qquad \alpha = \rho e^{i\theta} \in \mathbb{C},$$

are $\varphi_1(\alpha) = \sqrt{\rho} e^{i\theta/2}$ and $\varphi_2(\alpha) = -\sqrt{\rho} e^{i\theta/2}$.

Theorem

There exists no square root function continuous in a neighborhood of $0 \in \mathbb{C}$.

Proof.

$\square$

Example: roots of polynomials

The roots of the polynomial

$$x^2 - \alpha = 0, \qquad \alpha = \rho e^{i\theta} \in \mathbb{C},$$

are $\varphi_1(\alpha) = \sqrt{\rho} e^{i\theta/2}$ and $\varphi_2(\alpha) = -\sqrt{\rho} e^{i\theta/2}$.

Theorem

There exists no square root function continuous in a neighborhood of $0 \in \mathbb{C}$.

Proof.

If there exist a continuous function $\lambda : \mathcal{U} \to \mathbb{C}$ such that $\lambda(z)^2 = z$, where $\mathcal{U}$ is a neighborhood of 0.
We have $\lambda(z)^2 = z$ on a circle $S^1 = \{z \in \mathbb{C} \, : \, |z| = \rho\}$, with $\rho > 0$.
Define the two sets

$$\mathcal{U}_1 := \{z \in S^1 \, : \, \lambda(z) = \varphi_1(z)\}, \quad \mathcal{U}_2 := \{z \in S^1 \, : \, \lambda(z) = \varphi_2(z)\}.$$

We claim that $\mathcal{U}_1$ is open in S^1. $\square$

Example: roots of polynomials

Theorem

There exists no square root function continuous in a neighborhood of $0 \in \mathbb{C}$.

Proof.

For $z \in \mathcal{U}_1$ there exists a neighborhood $\mathcal{V}$ of z such that $|\lambda(w) - \lambda(z)| < \sqrt{\rho}$ because λ is continuous, but $|\varphi_1 - \varphi_2| = 2\sqrt{\rho}$ and thus $\lambda(z) \equiv \varphi_1(z)$ on $\mathcal{V}$.
Analogously $\mathcal{U}_2$ is open.

The two open subset of S^1 are disjoint and their union is S^1, but since S^1 is a connected set, we have that $S^1 = \mathcal{U}_1$ or $S^1 = \mathcal{U}_2$.

In both cases $\lim_{\theta \to 0^+} \lambda(z) \neq \lim_{\theta \to 2\pi^-} \lambda(z)$, while they should coincide if λ is continuous. (For instance if $S^1 = \mathcal{U}_1$ we get $\sqrt{\rho}$ and $-\sqrt{\rho}$ for the two limits.) $\square$

Example: roots of polynomials

Find all solutions of the polynomial equations
$p(x) = a_0 + a_1 x + \cdots + a_n x^n = 0$, where $p \in \mathbb{C}_n[x]$ with $a_n \neq 0$.

Is the problem well posed? The equation has n solutions $\xi_1, \ldots, \xi_n$ counted with multiplicity (fundamental theorem of algebra).

Example: roots of polynomials

Find all solutions of the polynomial equations $p(x) = a_0 + a_1 x + \cdots + a_n x^n = 0$, where $p \in \mathbb{C}_n[x]$ with $a_n \neq 0$.

Is the problem well posed? The equation has n solutions $\xi_1, \ldots, \xi_n$ counted with multiplicity (fundamental theorem of algebra).

If the roots are distinct, let $p(x) + tq(x)$, for $t \in [-a, a]$ and $q(x)$ polynomial of the same degree as $p(x)$, with $a > 0$, be a polynomial. We have $p_t \rightarrow_{t \rightarrow 0} p(x)$, then

Example: roots of polynomials

Find all solutions of the polynomial equations
$p(x) = a_0 + a_1 x + \cdots + a_n x^n = 0$, where $p \in \mathbb{C}_n[x]$ with $a_n \neq 0$.

Is the problem well posed? The equation has n solutions $\xi_1, \ldots, \xi_n$ counted with multiplicity (fundamental theorem of algebra).

If the roots are distinct, let $p(x) + tq(x)$, for $t \in [-a, a]$ and $q(x)$ polynomial of the same degree as $p(x)$, with $a > 0$, be a polynomial. We have $p_t \to_{t \to 0} p(x)$, then

for a sufficiently small t the polynomial $p_t(x)$ has distinct roots $\zeta_1(t), \ldots, \zeta_n(t)$ which are analytic (continuous) functions of t and $\zeta_i(0) = \xi_i$ (a theorem in complex analysis).

The problem is well-posed: we have a function $f : \Omega \to \mathbb{C}^n$ ($\Omega \subset \mathbb{C}^n$ open)

Example: integrals

Compute $\int_a^b f(x)dx$, with f continuous on $[a, b]$.

Riemann (Lebesgue) integrability guarantees the existence and uniqueness (fundamental theorem of calculus).

Example: integrals

Compute $\int_a^b f(x)dx$, with f continuous on $[a, b]$.

Riemann (Lebesgue) integrability guarantees the existence and uniqueness (fundamental theorem of calculus).

The problem can be stated as a function evaluation

$$f : C[a, b] \to \mathbb{R}$$

where $C[a, b]$ are continuous functions on $[a, b]$.

Notice that the space $C[a, b]$ as a vector space has infinite dimension!

Problems to be solved

What do we need to solve?

Problems to be solved

What do we need to solve?

A problem can be stated as a function evaluation.

In the general case we have a function

$$f : V \to W,$$

where V, W are **infinite-dimensional** vector spaces (Banach spaces).

Problems to be solved

What do we need to solve?

A problem can be stated as a function evaluation.

In the general case we have a function

$$f : V \to W,$$

where V, W are **infinite-dimensional** vector spaces (Banach spaces).

What can we solve?

Problems to be solved

What do we need to solve?

A problem can be stated as a function evaluation.

In the general case we have a function

$$f : V \to W,$$

where V, W are **infinite-dimensional** vector spaces (Banach spaces).

What can we solve?

By hand or with a computer, we can evaluate only **rational functions**

$$f : \mathbb{R}^n \to \mathbb{R}, \qquad f = \frac{p(x_1, \dots, x_n)}{q(x_1, \dots, x_n)}$$

with p, q polynomials (in n variables).
Moreover we cannot use all real numbers but just a fistful of them.

Problems to be solved

What do we need to solve?

A problem can be stated as a function evaluation.

In the general case we have a function

$$f : V \to W,$$

where V, W are **infinite-dimensional** vector spaces (Banach spaces).

What can we solve?

By hand or with a computer, we can evaluate only **rational functions**

$$f : \mathbb{R}^n \to \mathbb{R}, \qquad f = \frac{p(x_1, \ldots, x_n)}{q(x_1, \ldots, x_n)}$$

with p, q polynomials (in n variables).
Moreover we cannot use all real numbers but just a fistful of them.

This is awkward!

Approximation

We accept an approximated solution.

It is perhaps surprising that most of continuous problems can be well approximated using only elementary operations.

Approximation

We accept an approximated solution.

It is perhaps surprising that most of continuous problems can be well approximated using only elementary operations.

First steps of the approximation

- Go from infinite dimensional spaces to finite dimensional (discretization);

Approximation

We accept an approximated solution.

It is perhaps surprising that most of continuous problems can be well approximated using only elementary operations.

First steps of the approximation

- Go from infinite dimensional spaces to finite dimensional (discretization);
- Go from a generic function to a rational function.

We make the following approximation

$$\widehat{f} : V \to W \quad \Longrightarrow \quad \varphi : \mathbb{R}^n \to \mathbb{R} \quad \Longrightarrow \quad f : \mathbb{R}^n \to \mathbb{R}$$

with f rational.

Approximation

We accept an approximated solution.

It is perhaps surprising that most of continuous problems can be well approximated using only elementary operations.

First steps of the approximation

- Go from infinite dimensional spaces to finite dimensional (discretization);
- Go from a generic function to a rational function.

We make the following approximation

$$\widehat{f} : V \to W \quad \Longrightarrow \quad \varphi : \mathbb{R}^n \to \mathbb{R} \quad \Longrightarrow \quad f : \mathbb{R}^n \to \mathbb{R}$$

with f rational. The function f evaluated on finite arithmetic is said to be a **numerical algorithm**.

We can define the relative analytic error, when $\varphi(x) \neq 0$, as

$$\varepsilon_{AN} = \frac{f(x) - \varphi(x)}{\varphi(x)}$$

Analytic error

Let $\operatorname{conv}(a, b) = \operatorname{conv}(\{a, b\})$ be the interval $[a, b]$ if $a < b$; the interval $[b, a]$ if $a > b$; or the point a if $a = b$.

Analytic error

Let $\operatorname{conv}(a, b) = \operatorname{conv}(\{a, b\})$ be the interval $[a, b]$ if $a < b$; the interval $[b, a]$ if $a > b$; or the point a if $a = b$.

Theorem (Taylor's formula)

Let $f \in C^{r+1}(\mathbb{R})$ and let $x_0 \in \mathbb{R}$. Then for $x \in \mathbb{R}$, there exists $\xi(x) \in$conv(x_0, x), such that

$$f(x) = p(x) + r(x)$$

where

$$p(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(r)}(x_0)}{r!}(x - x_0)^r$$

and

$$r(x) = \frac{f^{(r+1)}(\xi(x))}{(r+1)!}(x - x_0)^{r+1}$$

Analytic error

Theorem (Taylor's formula)

Let $f \in C^{r+1}(\mathbb{R})$ and let $x_0 \in \mathbb{R}$. Then for $x \in \mathbb{R}$, there exists $\xi(x) \in conv(x_0, x)$, such that

$$f(x) = p(x) + r(x)$$

where

$$p(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(r)}(x_0)}{r!}(x - x_0)^r$$

and

$$r(x) = \frac{f^{(r+1)}(\xi(x))}{(r+1)!}(x - x_0)^{r+1}$$

Analytic error

Theorem (Taylor's formula)

Let $f \in C^{r+1}(\mathbb{R})$ and let $x_0 \in \mathbb{R}$. Then for $x \in \mathbb{R}$, there exists $\xi(x) \in$conv(x_0, x), such that

$$f(x) = p(x) + r(x)$$

where

$$p(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(r)}(x_0)}{r!}(x - x_0)^r$$

and

$$r(x) = \frac{f^{(r+1)}(\xi(x))}{(r+1)!}(x - x_0)^{r+1}$$

$p(x)$ is the Taylor polynomial of degree r at x_0.

Analytic error

Theorem (Taylor's formula)

Let $f \in C^{r+1}(\mathbb{R})$ and let $x_0 \in \mathbb{R}$. Then for $x \in \mathbb{R}$, there exists $\xi(x) \in conv(x_0, x)$, such that

$$f(x) = p(x) + r(x)$$

where

$$p(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(r)}(x_0)}{r!}(x - x_0)^r$$

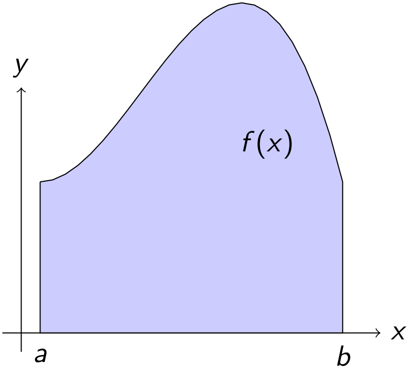
and

$$r(x) = \frac{f^{(r+1)}(\xi(x))}{(r+1)!}(x - x_0)^{r+1}$$

$r(x)$ is the remainder (Lagrange's remainder) and it is an **absolute** analytic error (that is small for x near to x_0).

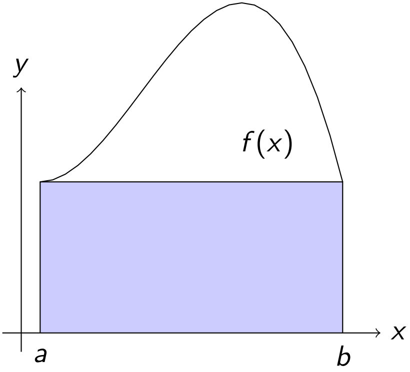
Analytic error

The integral of $f \in C[a, b]$, with $f > 0$ can be approximated by a trapezoid (trapezoidal rule).



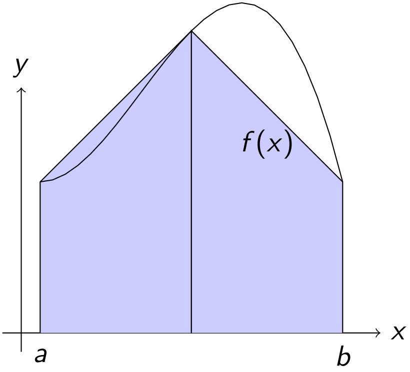
Analytic error

The integral of $f \in C[a, b]$, with $f > 0$ can be approximated by a trapezoid (trapezoidal rule).



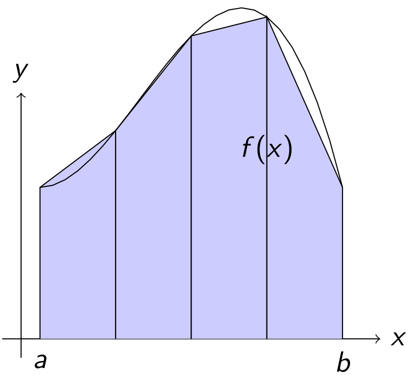
Analytic error

The integral of $f \in C[a, b]$, with $f > 0$ can be approximated by a trapezoid (trapezoidal rule).

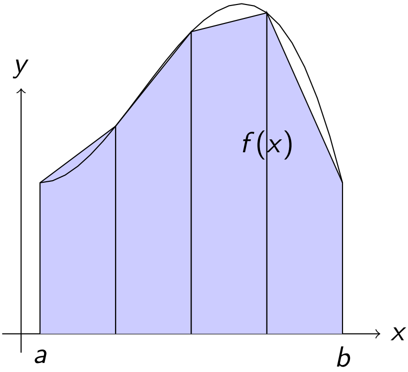


Analytic error

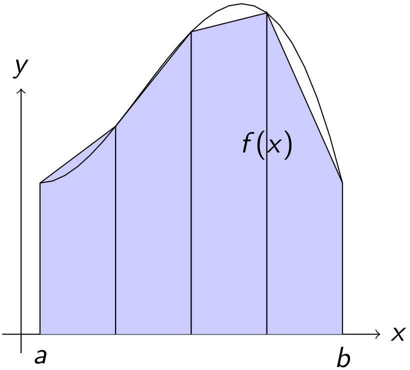
The integral of $f \in C[a, b]$, with $f > 0$ can be approximated by a trapezoid (trapezoidal rule).



Analytic error



Analytic error



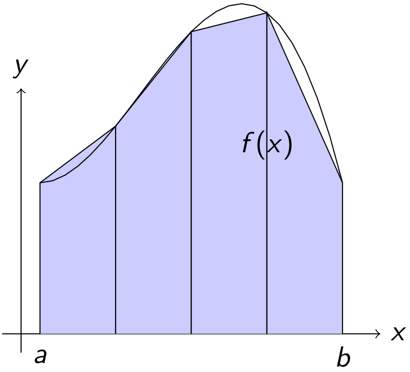
Let $h = (b-a)/n$, and let $x_i = a + \frac{b-a}{n} i$, then

$$\int_a^b f(x)dx = h\Big(\frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n)\Big) + r(x)$$

It can be proved that, if $f \in C^2[a,b]$, there exists $\xi \in (a,b)$, such that

$$r(x) = \frac{(b-a)^3}{12n^2} f''(\xi).$$

Analytic error



Let $h = (b-a)/n$, and let $x_i = a + \frac{b-a}{n} i$, then

$$\int_a^b f(x)dx = h\Big(\frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n)\Big) + r(x)$$

It can be proved that, if $f \in C^2[a,b]$, there exists $\xi \in (a,b)$, such that

$$r(x) = \frac{(b-a)^3}{12n^2} f''(\xi).$$

Disclaimer

We will not discuss anymore of the analytic error, and we will only consider rational functions (or elementary functions).

Error for rational functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ rational, that is $f = p/q$ with p, q polynomials.

From mathematical analysis we know that f is defined and differentiable for $q \neq 0$. (We may assume that p and q are prime).

Examples of rational functions:

$$f(x) = \frac{x^2 + 2x}{x + 1}, \qquad f(x) = \frac{xy + x^2 + 5x^2yz}{z + \pi y^2}$$

But also

Error for rational functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ rational, that is $f = p/q$ with p, q polynomials.

From mathematical analysis we know that f is defined and differentiable for $q \neq 0$. (We may assume that p and q are prime).

Examples of rational functions:

$$f(x) = \frac{x^2 + 2x}{x + 1}, \qquad f(x) = \frac{xy + x^2 + 5x^2yz}{z + \pi y^2}$$

But also

- the Taylor polynomial (one variable);
- the trapezoidal rule, if the integrand function is a rational function;
- the determinant of A.

Error for rational functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ rational, that is $f = p/q$ with p, q polynomials.

From mathematical analysis we know that f is defined and differentiable for $q \neq 0$. (We may assume that p and q are prime).

Examples of rational functions:

$$f(x) = \frac{x^2 + 2x}{x + 1}, \qquad f(x) = \frac{xy + x^2 + 5x^2yz}{z + \pi y^2}$$

But also

- the Taylor polynomial (one variable);
- the trapezoidal rule, if the integrand function is a rational function;
- the determinant of A.

There are two types of error in the evaluation of f at x on floating point numbers.

Error for rational functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ rational, that is $f = p/q$ with p, q polynomials.

- **Inherent error**. We do not evaluate $f(x)$ but rather $f(\widetilde{x})$ where $\widetilde{x} = \mathrm{fl}(x)$

$$\varepsilon_{IN} = \frac{f(\widetilde{x}) - f(x)}{f(x)}, \qquad f(x) \neq 0.$$

Error for rational functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ rational, that is $f = p/q$ with p, q polynomials.

- **Inherent error**. We do not evaluate $f(x)$ but rather $f(\widetilde{x})$ where $\widetilde{x} = \mathrm{fl}(x)$

$$\varepsilon_{IN} = \frac{f(\widetilde{x}) - f(x)}{f(x)}, \qquad f(x) \neq 0.$$

- **Algorithmic error**. We do not evaluate $f(\widetilde{x})$ but another function $\widetilde{f}(\widetilde{x})$, since the arithmetic operations of f are computed as machine operations.

$$\varepsilon_{ALG} = \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(\widetilde{x})}, \qquad f(\widetilde{x}) \neq 0.$$

Error for rational functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ rational, that is $f = p/q$ with p, q polynomials.

- **Inherent error**. We do not evaluate $f(x)$ but rather $f(\widetilde{x})$ where $\widetilde{x} = \mathrm{fl}(x)$

$$\varepsilon_{IN} = \frac{f(\widetilde{x}) - f(x)}{f(x)}, \qquad f(x) \neq 0.$$

- **Algorithmic error**. We do not evaluate $f(\widetilde{x})$ but another function $\widetilde{f}(\widetilde{x})$, since the arithmetic operations of f are computed as machine operations.

$$\varepsilon_{ALG} = \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(\widetilde{x})}, \qquad f(\widetilde{x}) \neq 0.$$

Inherent error can be defined also for non-rational functions.

Algorithmic error can be defined also for elementary functions, which are treated as operations.

Error for rational functions

- **Inherent error**. We do not evaluate $f(x)$ but rather $f(\widetilde{x})$ where $\widetilde{x} = \mathrm{fl}(x)$

$$\varepsilon_{IN} = \frac{f(\widetilde{x}) - f(x)}{f(x)}, \qquad f(x) \neq 0.$$

  If the inherent error is relatively small we say that the problem is **well-conditioned** or else **ill-conditioned**.

- **Algorithmic error**. We do not evaluate $f(\widetilde{x})$ but another function $\widetilde{f}(\widetilde{x})$, since the arithmetic operations of f are computed as machine operations.

$$\varepsilon_{ALG} = \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(\widetilde{x})}, \qquad f(\widetilde{x}) \neq 0.$$

  If the algorithmic error is relatively small we say that algorithm f is **numerically stable** or else **numerically unstable**.

A bit imprecise. Can be made more precise assuming $u \to 0$ and asking $\lim_{u\to 0} \varepsilon_{IN/ALG} = 0$.

Error for rational functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ rational, that is $f = p/q$ with p, q polynomials.

The total error (or forward error) is

$$\varepsilon_{TOT} = \frac{\widetilde{f}(\widetilde{x}) - f(x)}{f(x)}, \qquad f(x) \neq 0,$$

which gives a genuine measure of the error in the evaluation.

The ideal situation is $|\varepsilon_{TOT}| < u$, where u is the machine precision.
But, in practice, it is sufficient $|\varepsilon_{TOT}| \leqslant Mu$, with M constant.

Error for rational functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ rational, that is $f = p/q$ with p, q polynomials.

The total error (or forward error) is

$$\varepsilon_{TOT} = \frac{\widetilde{f}(\widetilde{x}) - f(x)}{f(x)}, \qquad f(x) \neq 0,$$

which gives a genuine measure of the error in the evaluation.

The ideal situation is $|\varepsilon_{TOT}| < u$, where u is the machine precision.
But, in practice, it is sufficient $|\varepsilon_{TOT}| \leqslant Mu$, with M constant.

What is the relationship between these errors?

Error for rational functions

Theorem

Let $x \in \mathbb{R}^n \setminus \{0\}$, *and* $f : \mathbb{R}^n \to \mathbb{R}$ *rational with* $f(x) \neq 0$ *and* $f(\widetilde{x}) \neq 0$, *where* $\widetilde{x} = \mathrm{fl}(x)$, *then*

$$\varepsilon_{TOT} = \varepsilon_{IN} + \varepsilon_{ALG} + \varepsilon_{IN}\varepsilon_{ALG}.$$

Error for rational functions

Theorem

Let $x \in \mathbb{R}^n \setminus \{0\}$, and $f : \mathbb{R}^n \to \mathbb{R}$ rational with $f(x) \neq 0$ and $f(\widetilde{x}) \neq 0$, where $\widetilde{x} = \mathrm{fl}(x)$, then

$$\varepsilon_{TOT} = \varepsilon_{IN} + \varepsilon_{ALG} + \varepsilon_{IN}\varepsilon_{ALG}.$$

Proof.

$$\begin{aligned}
\varepsilon_{TOT} &= \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x}) + f(\widetilde{x}) - f(x)}{f(x)} = \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(x)} + \frac{f(\widetilde{x}) - f(x)}{f(x)} \\
&= \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(\widetilde{x})} \frac{f(\widetilde{x})}{f(x)} + \varepsilon_{IN} = \varepsilon_{ALG}\left(\frac{f(\widetilde{x})}{f(x)} - 1 + 1\right) + \varepsilon_{IN} \\
&= \varepsilon_{ALG}\left(\frac{f(\widetilde{x}) - f(x)}{f(x)} + 1\right) + \varepsilon_{IN} = \varepsilon_{ALG}(\varepsilon_{IN} + 1) + \varepsilon_{IN} \\
&= \varepsilon_{IN} + \varepsilon_{ALG} + \varepsilon_{IN}\varepsilon_{ALG}.
\end{aligned}$$

□

Error for rational functions

Theorem

Let $x \in \mathbb{R}^n \setminus \{0\}$, and $f : \mathbb{R}^n \to \mathbb{R}$ rational with $f(x) \neq 0$ and $f(\widetilde{x}) \neq 0$, where $\widetilde{x} = \mathrm{fl}(x)$, then

$$\varepsilon_{TOT} = \varepsilon_{IN} + \varepsilon_{ALG} + \varepsilon_{IN}\varepsilon_{ALG}.$$

If ε_{IN} and ε_{ALG} are small (tends to zero as $u \to 0$, we have)

$$\varepsilon_{TOT} = \varepsilon_{IN} + \varepsilon_{ALG} + o(u) \doteq \varepsilon_{IN} + \varepsilon_{ALG}$$

Error analysis

Error analysis consists in relating the forward error with the error in the representation (and thus on the machine precision)

Formulae cannot be used directly

- Inherent error, we use the derivative (if the function is differentiable).
- Algorithmic error, we use a diagram analysis

Inherent error

Let $f : \mathbb{R}^n \to \mathbb{R}$, rational,

$$\varepsilon_{IN} = \frac{f(\widetilde{x}) - f(x)}{f(x)}, \qquad f(x) \neq 0.$$

where $\widetilde{x} = \mathrm{fl}(x)$.
If $x_i \neq 0$, for $i = 1, \ldots, n$, we want to relate ε_{IN} with the representation errors

$$\varepsilon_i = \frac{\widetilde{x}_i - x_i}{x_i},$$

and we make the usual assumption

$$|\varepsilon_i| < u,$$

where u is the machine precision. Ignoring overflow and underflow for simplicity.

Inherent error

We try to use the definition.

Example 1.

$$f(x) = x^2,$$

Let $\widetilde{x} = \mathrm{fl}(x) = x(1+\varepsilon_1)$, with $|\varepsilon_1| < u$, obtained as

$$\frac{\widetilde{x} - x}{x} = \varepsilon_1 \Longleftrightarrow \widetilde{x} - x = x\varepsilon_1 \Longleftrightarrow \widetilde{x} = x(1+\varepsilon_1).$$

We have, for $x \neq 0$,

$$\varepsilon_{IN} = \frac{\widetilde{x}^2 - x^2}{x^2} = \frac{x^2(1+\varepsilon_1)^2 - x^2}{x^2} = \frac{x^2(1+2\varepsilon+\varepsilon_1^2-1)}{x^2} = 2\varepsilon_1 + \varepsilon_1^2.$$

Since we are mostly interested on what happens as $u \to 0$, we can consider only first order terms

$$\varepsilon_{IN} \doteq 2\varepsilon_1$$

or $\varepsilon_{IN} = 2\varepsilon_1 + o(u)$. **Well-conditioned!**

Inherent error

In general using the definition is complicated. A simpler technique is obtained using the following.

Theorem

Let $x, \widetilde{x} \in \mathbb{R} \setminus \{0\}$ and let $f \in C^1(conv(x, \widetilde{x}))$, with $f(x) \neq 0$. There exists $\xi \in conv(x, \widetilde{x})$, such that

$$\varepsilon_{IN} = \frac{x}{f(x)} f'(\xi) \varepsilon_x,$$

where $\varepsilon_x = \varepsilon_1$ is the representation error on x. If, moreover $f \in C^2(conv(x, \widetilde{x}))$, then

$$\varepsilon_{IN} = \frac{x}{f(x)} f'(x) \varepsilon_x + o(u),$$

that is $\varepsilon_{IN} \doteq \frac{x}{f(x)} f'(x) \varepsilon_x$.

The formula above is very useful.

Inherent error

Proof.

$$\frac{f(\widetilde{x})-f(x)}{f(x)}=\frac{f'(\xi)(\widetilde{x}-x)}{f(x)}\frac{x}{x}=\frac{xf'(\xi)}{f(x)}\frac{\widetilde{x}-x}{x}=\frac{x}{f(x)}f'(\xi)\varepsilon_1.$$

The mean value theorem has been used. $\square$

Inherent error

$$\varepsilon_{IN} = \frac{f(\widetilde{x}) - f(x)}{f(x)} \doteq \frac{x}{f(x)} f'(x)\varepsilon_x.$$

The term $c_x = \frac{x}{f(x)} f'(x)$ is said to be **amplification factor** and measure the amplification of the error.

We use the formula with the function $f(x) = x^2$.

$$\varepsilon_{IN} \doteq \frac{x}{x^2} 2x\varepsilon_1 = 2\varepsilon_1$$

Much easier!

Inherent error

$$\varepsilon_{IN} = \frac{f(\widetilde{x}) - f(x)}{f(x)} \doteq \frac{x}{f(x)} f'(x)\varepsilon_x.$$

The term $c_x = \frac{x}{f(x)} f'(x)$ is said to be **amplification factor** and measure the amplification of the error.

We use the formula with the function $f(x) = x^2$.

$$\varepsilon_{IN} \doteq \frac{x}{x^2} 2x\varepsilon_1 = 2\varepsilon_1$$

Much easier!

What happens for more than one variable?

Inherent error

If $x = (x_1, \dots, x_n)$, and consider $f(x)$, we have the formula (with $x_i \neq 0$, $f(x) \neq 0$)

$$\varepsilon_{IN} \doteq \frac{x_1}{f(x)}\frac{\partial f}{\partial x_1}(x)\varepsilon_1 + \frac{x_2}{f(x)}\frac{\partial f}{\partial x_2}(x)\varepsilon_2 + \cdots + \frac{x_n}{f(x)}\frac{\partial f}{\partial x_n}(x)\varepsilon_n$$

where

$$\varepsilon_i = \frac{\widetilde{x}_i - x_i}{x_i}, \qquad c_i = \frac{x_i}{f(x)}\frac{\partial f}{\partial x_i}(x),$$

are the representation error and the amplification coefficients, respectively.

Inherent error on operations

$$\varepsilon_{IN} \doteq c_1\varepsilon_1 + \cdots c_n\varepsilon_n, \qquad c_i = \frac{x}{f(x)}\frac{\partial f}{\partial x_i}(x).$$

Let $f(x, y) = xy$, the multiplication. Then $\varepsilon_{IN} \doteq c_x\varepsilon_x + c_y\varepsilon_y$, where

Inherent error on operations

$$\varepsilon_{IN} \doteq c_1\varepsilon_1 + \cdots c_n\varepsilon_n, \qquad c_i = \frac{x}{f(x)}\frac{\partial f}{\partial x_i}(x).$$

Let $f(x,y) = xy$, the multiplication. Then $\varepsilon_{IN} \doteq c_x\varepsilon_x + c_y\varepsilon_y$, where

$$c_x = \frac{x}{f(x,y)}\frac{\partial f}{\partial x}(x,y) = \frac{x}{xy}y = 1,$$

Inherent error on operations

$$\varepsilon_{IN} \doteq c_1\varepsilon_1 + \cdots c_n\varepsilon_n, \qquad c_i = \frac{x}{f(x)}\frac{\partial f}{\partial x_i}(x).$$

Let $f(x,y) = xy$, the multiplication. Then $\varepsilon_{IN} \doteq c_x\varepsilon_x + c_y\varepsilon_y$, where

$$c_x = \frac{x}{f(x,y)}\frac{\partial f}{\partial x}(x,y) = \frac{x}{xy}y = 1, c_y = \frac{y}{f(x,y)}\frac{\partial f}{\partial y}(x,y) = \frac{y}{xy}x = 1.$$

Inherent error on operations

$$\varepsilon_{IN} \doteq c_1\varepsilon_1 + \cdots c_n\varepsilon_n, \qquad c_i = \frac{x}{f(x)}\frac{\partial f}{\partial x_i}(x).$$

Let $f(x, y) = xy$, the multiplication. Then $\varepsilon_{IN} \doteq c_x\varepsilon_x + c_y\varepsilon_y$, where

$$c_x = \frac{x}{f(x,y)}\frac{\partial f}{\partial x}(x,y) = \frac{x}{xy}y = 1, c_y = \frac{y}{f(x,y)}\frac{\partial f}{\partial y}(x,y) = \frac{y}{xy}x = 1.$$

Thus we have

$$c_x = 1, \qquad c_y = 1,$$

and the problem is well-conditioned. Obvious, really!

Inherent error on operations

$$\varepsilon_{IN} \doteq c_1\varepsilon_1 + \cdots c_n\varepsilon_n, \qquad c_i = \frac{x}{f(x)}\frac{\partial f}{\partial x_i}(x).$$

Let $f(x, y) = x + y$, the sum. Then $\varepsilon_{IN} \doteq c_x\varepsilon_x + c_y\varepsilon_y$, where

Inherent error on operations

$$\varepsilon_{IN} \doteq c_1\varepsilon_1 + \cdots c_n\varepsilon_n, \qquad c_i = \frac{x}{f(x)}\frac{\partial f}{\partial x_i}(x).$$

Let $f(x, y) = x + y$, the sum. Then $\varepsilon_{IN} \doteq c_x\varepsilon_x + c_y\varepsilon_y$, where

$$c_x = \frac{x}{f(x, y)}\frac{\partial f}{\partial x}(x, y) = \frac{x}{x + y},$$

Inherent error on operations

$$\varepsilon_{IN} \doteq c_1\varepsilon_1 + \cdots c_n\varepsilon_n, \qquad c_i = \frac{x}{f(x)}\frac{\partial f}{\partial x_i}(x).$$

Let $f(x, y) = x + y$, the sum. Then $\varepsilon_{IN} \doteq c_x\varepsilon_x + c_y\varepsilon_y$, where

$$c_x = \frac{x}{f(x,y)}\frac{\partial f}{\partial x}(x,y) = \frac{x}{x+y}, \quad c_y = \frac{y}{f(x,y)}\frac{\partial f}{\partial y}(x,y) = \frac{y}{x+y}.$$

Inherent error on operations

$$\varepsilon_{IN} \doteq c_1\varepsilon_1 + \cdots c_n\varepsilon_n, \qquad c_i = \frac{x}{f(x)}\frac{\partial f}{\partial x_i}(x).$$

Let $f(x, y) = x + y$, the sum. Then $\varepsilon_{IN} \doteq c_x\varepsilon_x + c_y\varepsilon_y$, where

$$c_x = \frac{x}{f(x,y)}\frac{\partial f}{\partial x}(x,y) = \frac{x}{x+y}, \quad c_y = \frac{y}{f(x,y)}\frac{\partial f}{\partial y}(x,y) = \frac{y}{x+y}.$$

Thus we have

$$c_x = x/(x+y), \qquad c_y = y/(x+y),$$

and the problem becomes ill-conditioned when $x \approx -y$. Perhaps surprising!

Inherent error on operations

Let $f(x, y) = x + y$, the sum. Then $\varepsilon_{IN} \doteq c_x \varepsilon_x + c_y \varepsilon_y$, where

$$c_x = x/(x+y), \qquad c_y = y/(x+y),$$

and the problem becomes ill-conditioned when $x \approx -y$. Perhaps surprising!

This phenomenon is said to be **numerical cancellation** and can be understood intuitively.

If we subtract two very near numbers we loose significant digits.

Example. In $\mathbb{R}$, $\pi - 3.1 \approx 0.0415926$. In $\mathcal{F}(10, 3, *, *)$, $\mathrm{fl}(\pi) = 3.14$ and $\mathrm{fl}(3.1) = 3.1$, and thus $\mathrm{fl}(\pi) - \mathrm{fl}(3.1) = 0.04$ (while $\mathrm{fl}(\pi - 3.1) = 0.0416$) and we have lost two significant digits.

Inherent error for some functions

operation	c_x	c_y
$x+y$	$\frac{x}{x+y}$	$\frac{y}{x+y}$

Inherent error for some functions

operation	c_x	c_y
$x+y$	$\frac{x}{x+y}$	$\frac{y}{x+y}$
$x-y$	$\frac{x}{x-y}$	$\frac{-y}{x-y}$

$$c_x = \frac{x}{x-y}\frac{\partial f}{\partial x} = \frac{x}{x-y}, \qquad c_y = \frac{y}{x-y}\frac{\partial f}{\partial y} = \frac{y}{x-y}\cdot(-1)$$

Inherent error for some functions

operation	c_x	c_y
$x + y$	$\frac{x}{x+y}$	$\frac{y}{x+y}$
$x - y$	$\frac{x}{x-y}$	$\frac{-y}{x-y}$
xy	1	1

Inherent error for some functions

operation	c_x	c_y
$x+y$	$\frac{x}{x+y}$	$\frac{y}{x+y}$
$x-y$	$\frac{x}{x-y}$	$\frac{-y}{x-y}$
xy	1	1
x/y	1	-1

$$c_x = \frac{x}{f}\frac{\partial f}{\partial x} = \frac{x}{x/y}\frac{1}{y} = x\frac{y}{x}\frac{1}{y} = 1, \qquad c_y = \frac{y}{x/y}\frac{-x}{y^2} = -1,$$

Inherent error for some functions

operation	c_x	c_y
$x+y$	$\frac{x}{x+y}$	$\frac{y}{x+y}$
$x-y$	$\frac{x}{x-y}$	$\frac{-y}{x-y}$
xy	1	1
x/y	1	-1
$\exp(x)$	x	

$$c_x = \frac{x}{\exp(x)} \exp(x) = x,$$

Inherent error for some functions

operation	c_x	c_y
$x+y$	$\frac{x}{x+y}$	$\frac{y}{x+y}$
$x-y$	$\frac{x}{x-y}$	$\frac{-y}{x-y}$
xy	1	1
x/y	1	-1
$\exp(x)$	x	
$\sqrt{x}$	$\frac{1}{2}$	

$$c_x = \frac{x}{\sqrt{x}} \frac{1}{2\sqrt{x}} = \frac{1}{2},$$

Inherent error for some functions

operation	c_x	c_y
$x+y$	$\frac{x}{x+y}$	$\frac{y}{x+y}$
$x-y$	$\frac{x}{x-y}$	$\frac{-y}{x-y}$
xy	1	1
x/y	1	-1
$\exp(x)$	x	
$\sqrt{x}$	$\frac{1}{2}$	
x^α	α	

$$c_x = \frac{x}{x^\alpha} \alpha x^{\alpha-1} = \alpha.$$

Exercise

Study the inherent error (conditioning) of $f(x, y) = x^2 - y^2$.

Exercise

Study the inherent error (conditioning) of $f(x, y) = x^2 - y^2$.

Rule of thumb: one must compute the **inherent error** or the **amplification coefficients** and find out when these quantities go to infinity.

Exercise

Study the inherent error (conditioning) of $f(x,y) = x^2 - y^2$.

Rule of thumb: one must compute the **inherent error** or the **amplification coefficients** and find out when these quantities go to infinity.

$$c_x = \frac{x}{f(x,y)} \frac{\partial f}{\partial x}(x,y) =$$

Exercise

Study the inherent error (conditioning) of $f(x, y) = x^2 - y^2$.

Rule of thumb: one must compute the **inherent error** or the **amplification coefficients** and find out when these quantities go to infinity.

$$c_x = \frac{x}{f(x, y)} \frac{\partial f}{\partial x}(x, y) = \frac{x}{x^2 - y^2} 2x = \frac{2x^2}{x^2 - y^2},$$

Exercise

Study the inherent error (conditioning) of $f(x,y) = x^2 - y^2$.

Rule of thumb: one must compute the **inherent error** or the **amplification coefficients** and find out when these quantities go to infinity.

$$c_x = \frac{x}{f(x,y)} \frac{\partial f}{\partial x}(x,y) = \frac{x}{x^2 - y^2} 2x = \frac{2x^2}{x^2 - y^2},$$

$$c_y = \frac{y}{f(x,y)} \frac{\partial f}{\partial y}(x,y) =$$

Exercise

Study the inherent error (conditioning) of $f(x, y) = x^2 - y^2$.

Rule of thumb: one must compute the **inherent error** or the **amplification coefficients** and find out when these quantities go to infinity.

$$c_x = \frac{x}{f(x, y)} \frac{\partial f}{\partial x}(x, y) = \frac{x}{x^2 - y^2} 2x = \frac{2x^2}{x^2 - y^2},$$

$$c_y = \frac{y}{f(x, y)} \frac{\partial f}{\partial y}(x, y) = \frac{y}{x^2 - y^2}(-2y) = \frac{-2y^2}{x^2 - y^2},$$

Exercise

Study the inherent error (conditioning) of $f(x, y) = x^2 - y^2$.

Rule of thumb: one must compute the **inherent error** or the **amplification coefficients** and find out when these quantities go to infinity.

$$c_x = \frac{x}{f(x, y)} \frac{\partial f}{\partial x}(x, y) = \frac{x}{x^2 - y^2} 2x = \frac{2x^2}{x^2 - y^2},$$

$$c_y = \frac{y}{f(x, y)} \frac{\partial f}{\partial y}(x, y) = \frac{y}{x^2 - y^2}(-2y) = \frac{-2y^2}{x^2 - y^2},$$

$$\varepsilon_{IN} = \frac{2x^2}{x^2 - y^2}\varepsilon_x + \frac{-2y^2}{x^2 - y^2}\varepsilon_y,$$

where ε_x and ε_y are the representation errors for x and y, respectively.

Exercise

Study the inherent error (conditioning) of $f(x, y) = x^2 - y^2$.

Exercise

Study the inherent error (conditioning) of $f(x, y) = x^2 - y^2$.

Rule of thumb: one must compute the inherent error or the amplification coefficients and **find out when these quantities go to infinity**.

$$\varepsilon_{IN} = \frac{2x^2}{x^2 - y^2}\varepsilon_x + \frac{-2y^2}{x^2 - y^2}\varepsilon_y.$$

Exercise

Study the inherent error (conditioning) of $f(x, y) = x^2 - y^2$.

Rule of thumb: one must compute the inherent error or the amplification coefficients and **find out when these quantities go to infinity**.

$$\varepsilon_{IN} = \frac{2x^2}{x^2 - y^2}\varepsilon_x + \frac{-2y^2}{x^2 - y^2}\varepsilon_y.$$

It is apparent that the inherent error is large for $x^2 \approx y^2$, that is

Exercise

Study the inherent error (conditioning) of $f(x, y) = x^2 - y^2$.

Rule of thumb: one must compute the inherent error or the amplification coefficients and **find out when these quantities go to infinity**.

$$\varepsilon_{IN} = \frac{2x^2}{x^2 - y^2}\varepsilon_x + \frac{-2y^2}{x^2 - y^2}\varepsilon_y.$$

It is apparent that the inherent error is large for $x^2 \approx y^2$, that is

$$x \approx \pm y.$$

We expect a large error when x is near to y or $-y$.

Algorithmic error

$$\varepsilon_{ALG} = \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(\widetilde{x})}$$

Algorithmic error is very hard to be computed using the definition.

We drop the tilde for simplicity.

Algorithmic error

$$\varepsilon_{ALG} = \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(\widetilde{x})}$$

Algorithmic error is very hard to be computed using the definition.

We drop the tilde for simplicity.

Example. $f(x) = x^2$, $\widetilde{f}(x) = x \otimes x = x^2(1 + \eta)$,

Algorithmic error

$$\varepsilon_{ALG} = \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(\widetilde{x})}$$

Algorithmic error is very hard to be computed using the definition.

We drop the tilde for simplicity.

Example. $f(x) = x^2$, $\widetilde{f}(x) = x \otimes x = x^2(1+\eta)$, where $|\eta| \leqslant u$ is the error in the operation.

Algorithmic error

$$\varepsilon_{ALG} = \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(\widetilde{x})}$$

Algorithmic error is very hard to be computed using the definition.

We drop the tilde for simplicity.

Example. $f(x) = x^2$, $\widetilde{f}(x) = x \otimes x = x^2(1+\eta)$, where $|\eta| \leqslant u$ is the error in the operation.

$$\varepsilon_{ALG} = \frac{x^2(1+\eta) - x^2}{x^2} =$$

Algorithmic error

$$\varepsilon_{ALG} = \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(\widetilde{x})}$$

Algorithmic error is very hard to be computed using the definition.

We drop the tilde for simplicity.

Example. $f(x) = x^2$, $\widetilde{f}(x) = x \otimes x = x^2(1+\eta)$, where $|\eta| \leqslant u$ is the error in the operation.

$$\varepsilon_{ALG} = \frac{x^2(1+\eta) - x^2}{x^2} = \frac{x^2(1+\eta-1)}{x^2} =$$

Algorithmic error

$$\varepsilon_{ALG} = \frac{\widetilde{f}(\widetilde{x}) - f(\widetilde{x})}{f(\widetilde{x})}$$

Algorithmic error is very hard to be computed using the definition.

We drop the tilde for simplicity.

Example. $f(x) = x^2$, $\widetilde{f}(x) = x \otimes x = x^2(1+\eta)$, where $|\eta| \leqslant u$ is the error in the operation.

$$\varepsilon_{ALG} = \frac{x^2(1+\eta) - x^2}{x^2} = \frac{x^2(1+\eta-1)}{x^2} = \eta.$$

This is a special case: a single operation.

Algorithmic error

Consider the function $f(x, y) = x^2 - y^2$ and the two algorithms

$$\begin{cases} z^{(1)} = x^2, \\ z^{(2)} = y^2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}. \end{cases}$$

Algorithmic error

Consider the function $f(x, y) = x^2 - y^2$ and the two algorithms

$$\begin{cases} z^{(1)} = x^2, \\ z^{(2)} = y^2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}. \end{cases} \qquad \begin{cases} v^{(1)} = x - y, \\ v^{(2)} = x + y, \\ f = v^{(3)} = v^{(1)} v^{(2)}. \end{cases}$$

Algorithmic error

Consider the function $f(x, y) = x^2 - y^2$ and the two algorithms

$$\begin{cases} z^{(1)} = x^2, \\ z^{(2)} = y^2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}. \end{cases} \qquad \begin{cases} v^{(1)} = x - y, \\ v^{(2)} = x + y, \\ f = v^{(3)} = v^{(1)} v^{(2)}. \end{cases}$$

The two algorithms are different due to the finite arithmetic.

Algorithmic error

Consider the function $f(x, y) = x^2 - y^2$ and the two algorithms

$$\begin{cases} z^{(1)} = x^2, \\ z^{(2)} = y^2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}. \end{cases} \qquad \begin{cases} v^{(1)} = x - y, \\ v^{(2)} = x + y, \\ f = v^{(3)} = v^{(1)} v^{(2)}. \end{cases}$$

The two algorithms are different due to the finite arithmetic.

We may assume that there is a local error for any operation

$$\begin{cases} z^{(1)} = x^2, & \varepsilon_1, \\ z^{(2)} = y^2, & \varepsilon_2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}, & \varepsilon_3. \end{cases}$$

Algorithmic error

Consider the function $f(x, y) = x^2 - y^2$ and the two algorithms

$$\begin{cases} z^{(1)} = x^2, \\ z^{(2)} = y^2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}. \end{cases} \qquad \begin{cases} v^{(1)} = x - y, \\ v^{(2)} = x + y, \\ f = v^{(3)} = v^{(1)} v^{(2)}. \end{cases}$$

The two algorithms are different due to the finite arithmetic.

We may assume that there is a local error for any operation

$$\begin{cases} z^{(1)} = x^2, & \varepsilon_1, \\ z^{(2)} = y^2, & \varepsilon_2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}, & \varepsilon_3. \end{cases} \qquad \begin{cases} v^{(1)} = x - y, & \eta_1, \\ v^{(2)} = x + y, & \eta_2, \\ f = v^{(3)} = v^{(1)} v^{(2)} & \eta_3. \end{cases}$$

where $|\varepsilon_i| \leqslant u$, $|\eta_i| \leqslant u$.

Algorithmic error

Studying the algorithmic error using the definition is horrible

$$\begin{cases} z^{(1)} = x^2, & \varepsilon_1, \\ z^{(2)} = y^2, & \varepsilon_2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}, & \varepsilon_3. \end{cases}$$

$$\varepsilon_{ALG} = \frac{(x \otimes x) \ominus (y \otimes y) - (x^2 - y^2)}{x^2 - y^2}$$

Algorithmic error

Studying the algorithmic error using the definition is horrible

$$\begin{cases} z^{(1)} = x^2, & \varepsilon_1, \\ z^{(2)} = y^2, & \varepsilon_2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}, & \varepsilon_3. \end{cases}$$

$$\varepsilon_{ALG} = \frac{(x \otimes x) \ominus (y \otimes y) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{\big(x^2(1+\varepsilon_1) - y^2(1+\varepsilon_2)\big)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

Algorithmic error

Studying the algorithmic error using the definition is horrible

$$\begin{cases} z^{(1)} = x^2, & \varepsilon_1, \\ z^{(2)} = y^2, & \varepsilon_2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}, & \varepsilon_3. \end{cases}$$

$$\varepsilon_{ALG} = \frac{(x \otimes x) \ominus (y \otimes y) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{\big(x^2(1+\varepsilon_1) - y^2(1+\varepsilon_2)\big)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{(x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

Algorithmic error

Studying the algorithmic error using the definition is horrible

$$\begin{cases} z^{(1)} = x^2, & \varepsilon_1, \\ z^{(2)} = y^2, & \varepsilon_2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}, & \varepsilon_3. \end{cases}$$

$$\varepsilon_{ALG} = \frac{(x \otimes x) \ominus (y \otimes y) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{\big(x^2(1+\varepsilon_1) - y^2(1+\varepsilon_2)\big)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{(x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2 + x^2\varepsilon_3 - x^2\varepsilon_1\varepsilon_3 - y^2\varepsilon_3 - y^2\varepsilon_2\varepsilon_3 - x^2 - y^2}{x^2 - y^2}$$

Algorithmic error

$$\varepsilon_{ALG} = \frac{(x \otimes x) \ominus (y \otimes y) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{\big(x^2(1+\varepsilon_1) - y^2(1+\varepsilon_2)\big)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{(x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2 + x^2\varepsilon_3 - x^2\varepsilon_1\varepsilon_3 - y^2\varepsilon_3 - y^2\varepsilon_2\varepsilon_3 - x^2 + y^2}{x^2 - y^2}$$

Algorithmic error

$$\varepsilon_{ALG} = \frac{(x \otimes x) \ominus (y \otimes y) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{\big(x^2(1+\varepsilon_1) - y^2(1+\varepsilon_2)\big)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{(x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2 + x^2\varepsilon_3 - x^2\varepsilon_1\varepsilon_3 - y^2\varepsilon_3 - y^2\varepsilon_2\varepsilon_3 - x^2 + y^2}{x^2 - y^2}$$

$$= \frac{x^2\varepsilon_1 - y^2\varepsilon_2 + x^2\varepsilon_3 - y^2\varepsilon_3 - x^2\varepsilon_1\varepsilon_3 - y^2\varepsilon_2\varepsilon_3}{x^2 - y^2}$$

Algorithmic error

$$\begin{aligned}
\varepsilon_{ALG} &= \frac{(x \otimes x) \ominus (y \otimes y) - (x^2 - y^2)}{x^2 - y^2} \\
&= \frac{\big(x^2(1+\varepsilon_1) - y^2(1+\varepsilon_2)\big)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2} \\
&= \frac{(x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2} \\
&= \frac{x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2 + x^2\varepsilon_3 - x^2\varepsilon_1\varepsilon_3 - y^2\varepsilon_3 - y^2\varepsilon_2\varepsilon_3 - x^2 + y^2}{x^2 - y^2} \\
&= \frac{x^2\varepsilon_1 - y^2\varepsilon_2 + x^2\varepsilon_3 - y^2\varepsilon_3 - x^2\varepsilon_1\varepsilon_3 - y^2\varepsilon_2\varepsilon_3}{x^2 - y^2} \\
&\doteq \frac{x^2\varepsilon_1 - y^2\varepsilon_2 + (x^2 - y^2)\varepsilon_3}{x^2 - y^2} = \frac{(x^2 - y^2)\varepsilon_3}{x^2 - y^2} + \frac{x^2\varepsilon_1 - y^2\varepsilon_2}{x^2 - y^2}
\end{aligned}$$

Algorithmic error

$$\varepsilon_{ALG} = \frac{(x \otimes x) \ominus (y \otimes y) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{\big(x^2(1+\varepsilon_1) - y^2(1+\varepsilon_2)\big)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{(x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2)(1+\varepsilon_3) - (x^2 - y^2)}{x^2 - y^2}$$

$$= \frac{x^2 - x^2\varepsilon_1 - y^2 - y^2\varepsilon_2 + x^2\varepsilon_3 - x^2\varepsilon_1\varepsilon_3 - y^2\varepsilon_3 - y^2\varepsilon_2\varepsilon_3 - x^2 + y^2}{x^2 - y^2}$$

$$= \frac{x^2\varepsilon_1 - y^2\varepsilon_2 + x^2\varepsilon_3 - y^2\varepsilon_3 - x^2\varepsilon_1\varepsilon_3 - y^2\varepsilon_2\varepsilon_3}{x^2 - y^2}$$

$$\doteq \frac{x^2\varepsilon_1 - y^2\varepsilon_2 + (x^2 - y^2)\varepsilon_3}{x^2 - y^2} = \frac{(x^2 - y^2)\varepsilon_3}{x^2 - y^2} + \frac{x^2\varepsilon_1 - y^2\varepsilon_2}{x^2 - y^2}$$

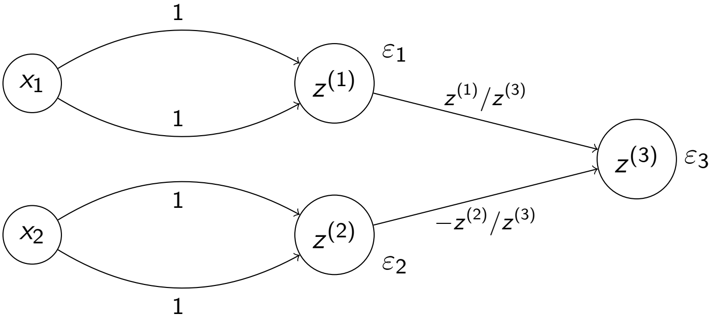
$$= \varepsilon_3 + \frac{x^2\varepsilon_1 - y^2\varepsilon_2}{x^2 - y^2}$$

Algorithmic error

$$\begin{cases} z^{(1)} = x^2, & \varepsilon_1, \\ z^{(2)} = y^2, & \varepsilon_2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}, & \varepsilon_3. \end{cases}$$

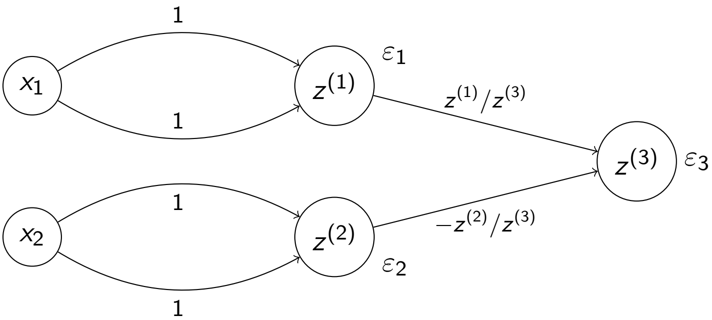
Algorithmic error

$$\begin{cases} z^{(1)} = x^2, & \varepsilon_1, \\ z^{(2)} = y^2, & \varepsilon_2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}, & \varepsilon_3. \end{cases}$$



Algorithmic error

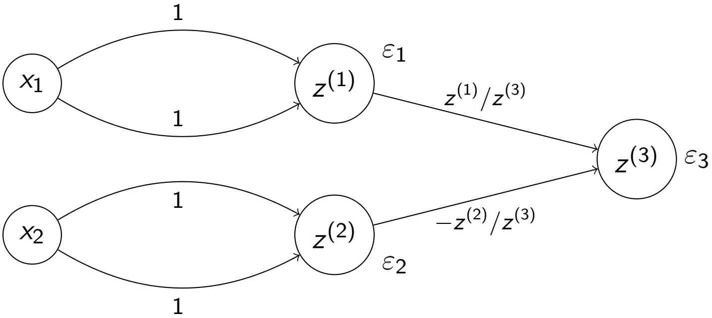
$$\begin{cases} z^{(1)} = x^2, & \varepsilon_1, \\ z^{(2)} = y^2, & \varepsilon_2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}, & \varepsilon_3. \end{cases}$$



$$\varepsilon_{ALG} \doteq \varepsilon_3 + \frac{z^{(1)}}{z^{(3)}}\varepsilon_1 - \frac{z^{(2)}}{z^{(3)}}\varepsilon_2 =$$

Algorithmic error

$$\begin{cases} z^{(1)} = x^2, & \varepsilon_1, \\ z^{(2)} = y^2, & \varepsilon_2, \\ f = z^{(3)} = z^{(1)} - z^{(2)}, & \varepsilon_3. \end{cases}$$



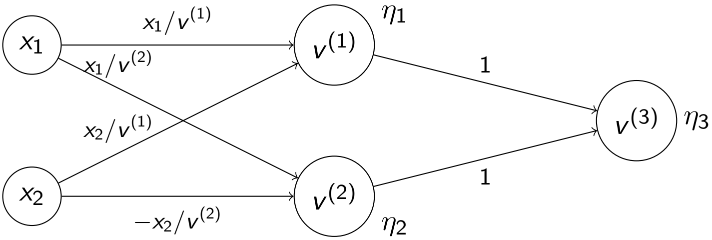
$$\varepsilon_{ALG} \doteq \varepsilon_3 + \frac{z^{(1)}}{z^{(3)}}\varepsilon_1 - \frac{z^{(2)}}{z^{(3)}}\varepsilon_2 = \varepsilon_3 + \frac{x^2\varepsilon_1 - y^2\varepsilon_2}{x^2 - y^2}$$

Algorithmic error

$$\begin{cases} v^{(1)} = x + y, & \eta_1, \\ v^{(2)} = x - y, & \eta_2, \\ f = v^{(3)} = v^{(1)} v^{(2)}, & \eta_3. \end{cases}$$

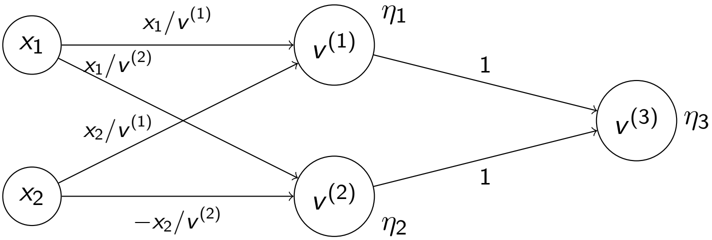
Algorithmic error

$$\begin{cases} v^{(1)} = x + y, & \eta_1, \\ v^{(2)} = x - y, & \eta_2, \\ f = v^{(3)} = v^{(1)} v^{(2)}, & \eta_3. \end{cases}$$



Algorithmic error

$$\begin{cases} v^{(1)} = x + y, & \eta_1, \\ v^{(2)} = x - y, & \eta_2, \\ f = v^{(3)} = v^{(1)} v^{(2)}, & \eta_3. \end{cases}$$



$$\varepsilon_{ALG} \doteq \eta_3 + \eta_1 + \eta_2$$

Algorithmic error

Which is the best?

$$\varepsilon_{ALG}^{(1)} \varepsilon_3 + \frac{x^2\varepsilon_1 - y^2\varepsilon_2}{x^2 - y^2} \qquad \varepsilon_{ALG}^{(2)} \doteq \eta_3 + \eta_1 + \eta_2.$$

Algorithmic error

Which is the best?

$$\varepsilon^{(1)}_{ALG}\varepsilon_3 + \frac{x^2\varepsilon_1 - y^2\varepsilon_2}{x^2 - y^2} \qquad \varepsilon^{(2)}_{ALG} \doteq \eta_3 + \eta_1 + \eta_2.$$

We can bound

$$|\varepsilon^{(1)}_{ALG}| \leqslant \left(1 + \frac{x^2 + y^2}{|x^2 - y^2|}\right)u, \qquad |\varepsilon^{(2)}_{ALG}| \leqslant 3u.$$

The second is the best in most cases.