POLYTECH INFORMATIQUE
3ÈME ANNÉE
Alexandra Bac

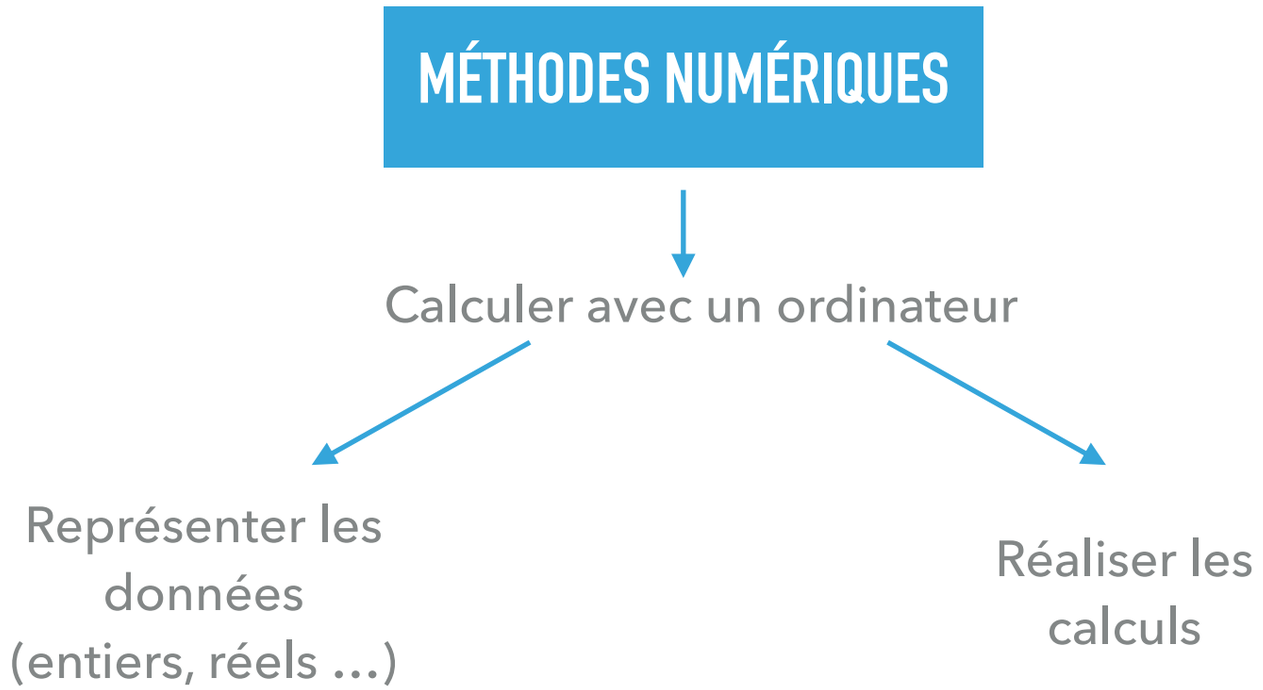
NUMERICAL METHODS

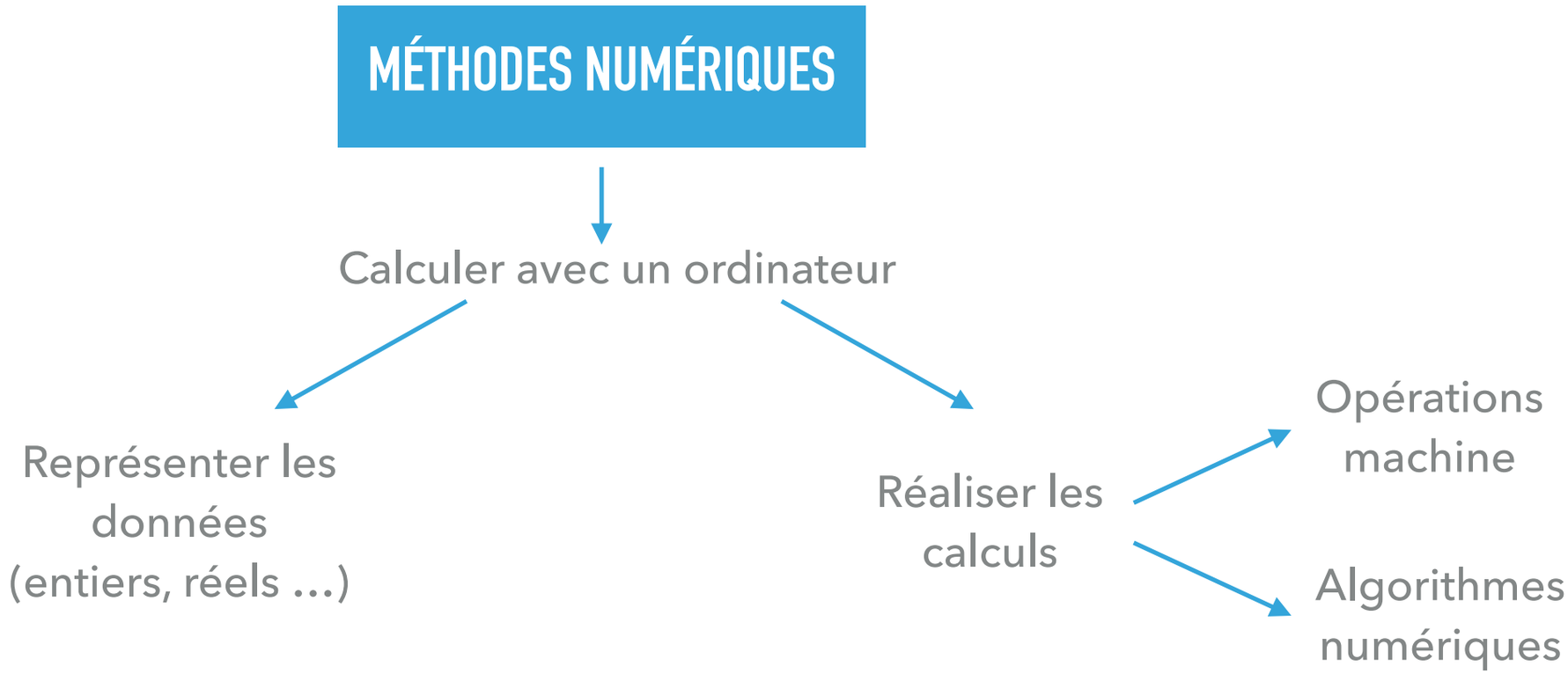
COURS 1
INTRODUCTION, NOMBRES ET MACHINES

MÉTHODES NUMÉRIQUES

↓

Calculer avec un ordinateur


MÉTHODES NUMÉRIQUES
Calculer avec un ordinateur
Représenter les données (entiers, réels ...)
Réaliser les calculs


MÉTHODES NUMÉRIQUES
Calculer avec un ordinateur
Représenter les données (entiers, réels ...)
Réaliser les calculs
Opérations machine
Algorithmes numériques

MÉTHODES NUMÉRIQUES

Calculer avec un ordinateur

Représenter les données (entiers, réels ...)

Réaliser les calculs

Opérations machine

Algorithmes numériques

Erreurs de représentation

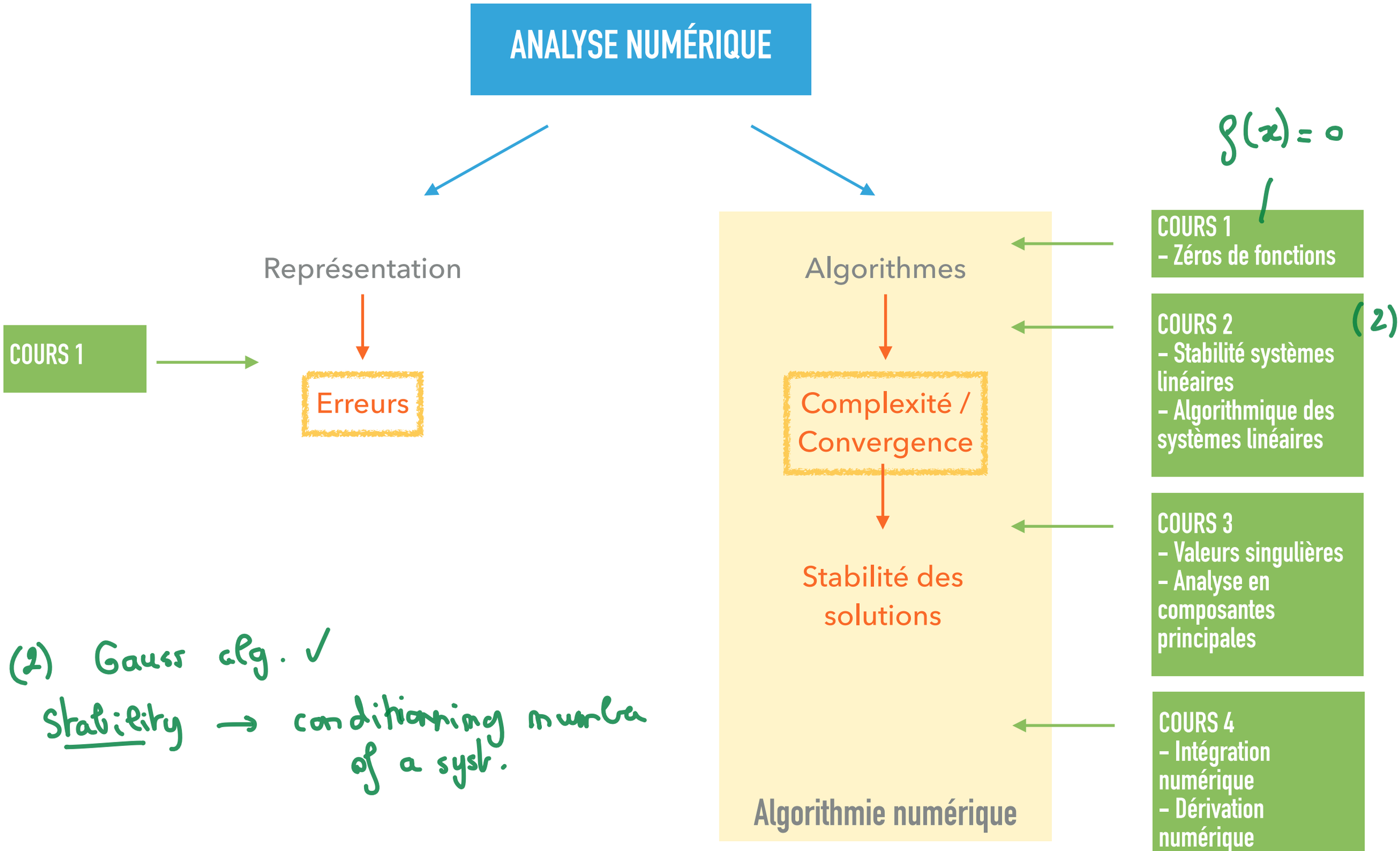
Erreurs de calcul

Arrondi

Convergence

π, $\sqrt{2}$ ...

0.1 → cannot be rep. exactly on a computer ...



(2) Gauss alg. ✓

Stability → conditionning number of a syst.

REPRÉSENTATION DES ENTIERS

PROBLÈME DE LA REPRÉSENTATION

Représentation en base b $\quad (b \in \mathbb{N}, b > 0)$

$$n = \sum_{i=1}^{N} c_i \cdot b^i \quad \text{avec } c_i < b$$

Représenté par

$$(c_0, \ldots, c_N)$$

$b = 2$

$$n = \sum_{i=1}^{N} c_i \cdot b^i \quad \text{avec } c_i \in \{0,1\}$$

bits

Représenté par

$$(c_0, \ldots, c_N)$$

N bits

PROBLÈME DE LA REPRÉSENTATION

Représentation en base b $\quad (b \in \mathbb{N}, b > 0)$

$$n = \sum_{i=1}^{N} c_i \cdot b^i \quad \text{avec } c_i < b$$

Représenté par

$$(c_0, \ldots, c_N)$$

$b = 2$

$$n = \sum_{i=1}^{N} c_i \cdot b^i \quad \text{avec } c_i \in \{0,1\}$$

bits

Représenté par

$$(c_0, \ldots, c_N)$$

N bits

Ordinateur → objets finis

integer ~ 32 bits / 4 bytes

Entiers négatifs ?

PROBLÈME DE LA REPRÉSENTATION

$$n = \sum_{i=1}^{N} c_i \cdot b^i \quad \text{avec } c_i \in \{0,1\}$$

ENTIERS POSITIFS

$$N \leftrightarrow 0 \leqslant n \leqslant 2^N - 1$$

- N=8 (unsigned char)
  $2^n - 1 = 255$
- N=16 (unsigned short)
  $2^n - 1 = 65535$
- N=32 (unsigned int)
  $2^n - 1 = 4.294.967.295$
- N=64 (unsigned long int)
  $2^n - 1 = 1{,}844 \cdot 10^{19}$

ENTIERS RELATIFS

1 bit utilisé pour coder le signe

$$N \leftrightarrow -2^{N-1} \leqslant n \leqslant 2^{N-1} - 1$$

- N=8 (char)
  $-128 \leqslant n \leqslant 127$
- N=16 (short)
  $-32768 \leqslant n \leqslant 32767$
- N=32 (int)
  $-2.147.483.648 \leqslant n \leqslant 2.147.483.647$
- N=64 (long int)
  $-9{,}2 \cdot 10^{18} \leqslant n \leqslant 9{,}2 \cdot 10^{18}$

PROBLÈME DE LA REPRÉSENTATION DES ENTIERS RELATIFS

$$n = \sum_{i=1}^{N} c_i \cdot b^i \quad \text{avec } c_i \in \{0,1\}$$

Codage du signe

CODAGE NAÏF PAR UN BIT

$$c_{N-1}c_{N-2}\dots c_1c_0 \;\rightarrow\; n = (-1)^{c_{N-1}}\left(\sum_{i=0}^{n-2} c_i \cdot 2^i\right)$$

Bit de signe

-11 sur 8 bits :

1	0	0	0	1	0	1	1

COMPLEMENT A UN

Le plus utilisé actuellement

$$c_{N-1}c_{N-2}\dots c_1c_0 \;\rightarrow\; n = -c_{N-1} \cdot 2^{N-1} + \sum_{i=0}^{n-2} c_i \cdot 2^i$$

-11 sur 8 bits :

-11 = -128 + 117

1	1	1	1	0	1	0	1

PROBLÈME DE LA REPRÉSENTATION DES ENTIERS RELATIFS

$$n = \sum_{i=1}^{N} c_i \cdot b^i \quad \text{avec } c_i \in \{0,1\}$$

Codage du signe

CODAGE NAÏF PAR UN BIT

- 0 a deux représentations
  00000000 / 10000000

- Addition → cas à traiter
  -11 + (-5)
  -11 + 5

COMPLEMENT A UN

Le plus utilisé actuellement

- 0 a une seule représentation
  00000000

- L'addition reste valable

```
 -11 : 11110101
  -5 : 11111011
+ ----------------
   1 1110000   →   112 (déc)
```

Résultat : -128 + 112 = -16 (déc)

PROBLÈME DE LA REPRÉSENTATION DES ENTIERS RELATIFS

$$n = \sum_{i=1}^{N} c_i \cdot b^i \quad \text{avec } c_i \in \{0,1\}$$

Codage du signe

CODAGE NAÏF PAR UN BIT

- 0 a deux représentations
  00000000 / 10000000

- Addition → cas à traiter
  -11 + (-5)
  -11 + 5

COMPLEMENT A UN

Le plus utilisé actuellement

- 0 a une seule représentation
  00000000

- L'addition reste valable

```
 -11 : 11110101
   5 : 00000101
+ ----------------
   1 1111010   ──→  122 (déc)
```

Résultat : -128 + 122 = -6 (déc)

REPRÉSENTATION DES FLOTTANTS . . .

Mériterait un cours entier ... cf TD ...

UN MOT SUR LES ERREURS

EVALUATION DE L'ERREUR COMMISE

Erreur commise entre x et x' ?

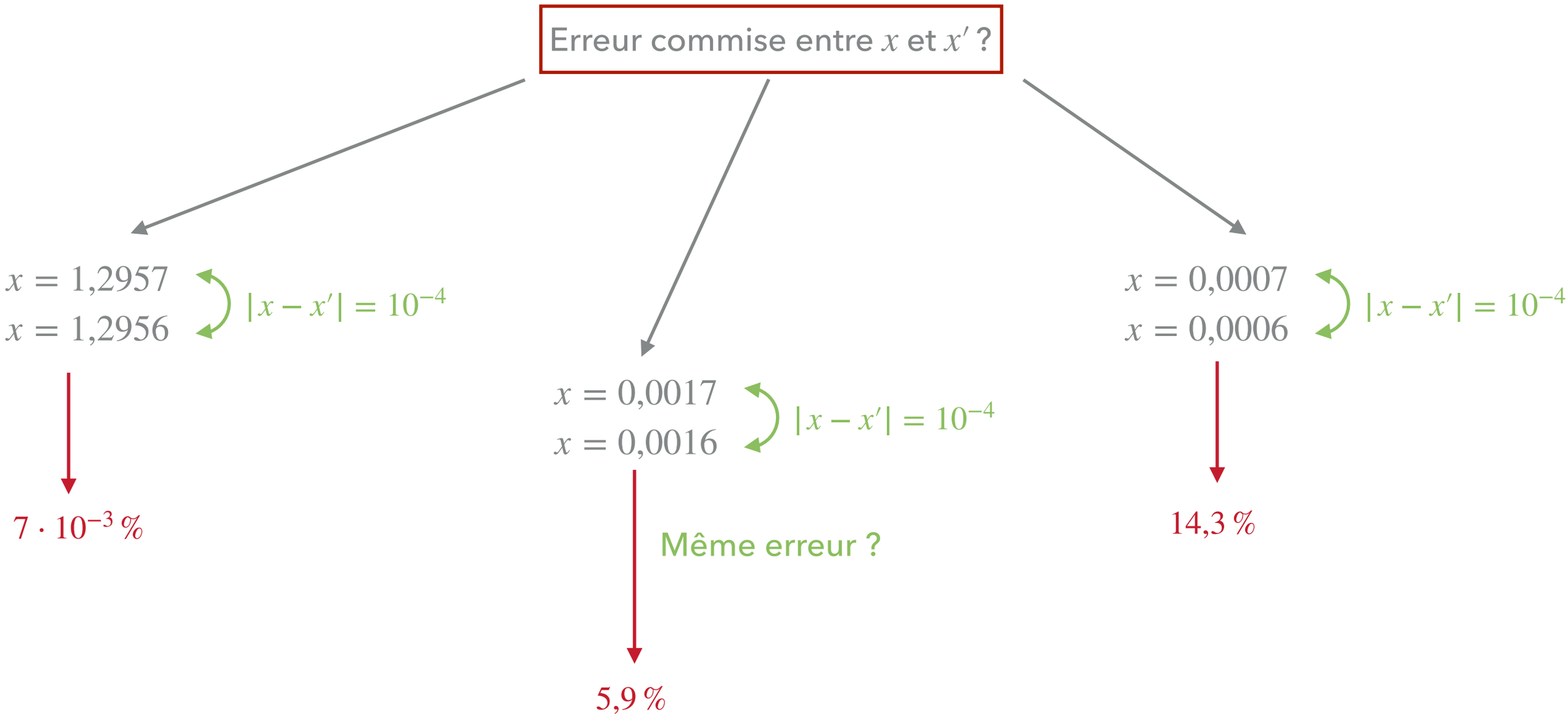
$x = 1{,}2957$
$x = 1{,}2956$
$|x - x'| = 10^{-4}$

$x = 0{,}0007$
$x = 0{,}0006$
$|x - x'| = 10^{-4}$

$x = 0{,}0017$
$x = 0{,}0016$
$|x - x'| = 10^{-4}$

Même erreur ?

EVALUATION DE L'ERREUR COMMISE



EVALUATION DE L'ERREUR COMMISE

Erreur commise entre x et x' ?

ERREUR ABSOLUE

$$|x - x'|$$

- Naturel sur les petits nombres
  $|x| < 1$
- Peut être inatteignable sur de grands nombres

ERREUR RELATIVE

$$\frac{|x - x'|}{|x|}$$

- Naturel sur les grands nombres
  $|x| \gg 1$
- L'erreur dépend de l'ordre de grandeur de x