Programmation et Méthodes Numériques

Représentation d'un nombre en machine, erreurs numériques, etc…

M. Casaletti (massimiliano.casaletti@upmc.fr)

Laboratoire d'Electronique et Electromagnétisme (L2E)
Université Pierre et Marie Curie

Présentation générale

- **Déroulement** sur 9 semaines:
  - 8 Cours magistraux (4 M.N.+ 4 P.)
  - 7 TD (8h M.N. + 6h P.)
  - 6 TP (6h M.N. + 6h P.)
  - Mini projet (4h TD + 10h TP + soutenance)

- **Évaluation:**
  - 2 Examens (60%)
  - Mini projet (40%)
    - Simulation des circuits RLC
    - Rayonnement des antennes
    - Propagation de la chaleur
    - …

Plan du cours

Points abordés:

- Représentation d'un nombre en machine, erreurs numériques, conditionnement d'un problème, etc…
- Interpolation polynomiale
- Intégration numérique
- Résolution numérique des équations différentielles ordinaires (EDO)

Ouvrages:

- **Introduction à l'analyse numérique**, J. Bastien et J.N. Martin, DUNOD
- **Numerical Recipes : The Art of Scientic Computing**, Cambridge University Press

Bagages nécessaires …

Mathématique :

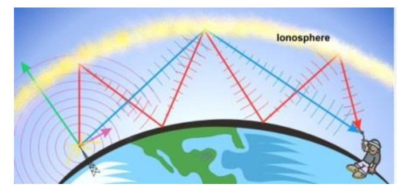
- Dérivation
- Intégration
- Développement de Taylor
- Espaces vectoriels (base, produit scalaire, distance,…)
- Équations différentielles ordinaires

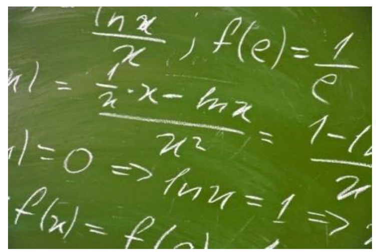
Physique (mini projet) :

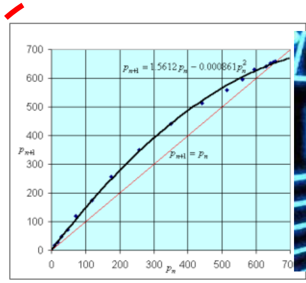
- Mécanique (cinématique, lois de Newton,…)
- Théorie des circuits électroniques
- Champ électrique et magnétique
- ……

Plan du cours

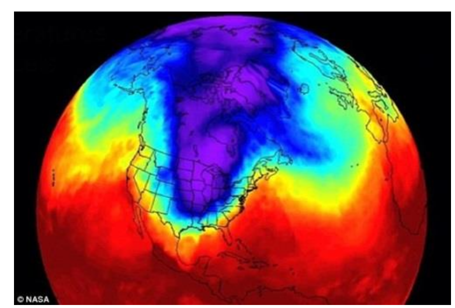
- **Représentation d'un nombre en machine, erreurs numériques, etc…**
  - **Introduction générale**
  - Représentation des nombres entiers et réels
  - Opérations élémentaires en virgule flottante
  - Conditionnement d'un problème
  - Opérations complexes
  - Conclusion: différentes sources d'erreur
- Interpolation polynomiale
- Intégration numérique
- Résolution numérique des équations différentielles ordinaires (EDO)


Ionosphere



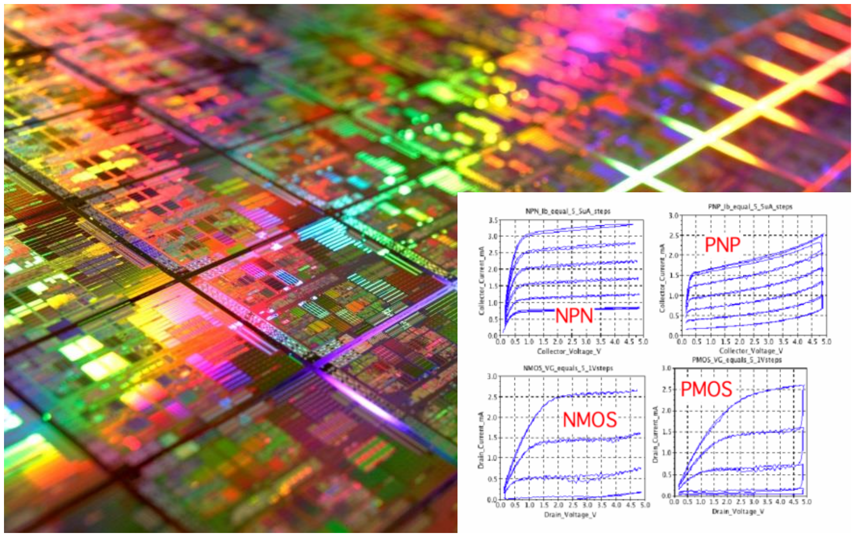

700
600
500
400
300
200
100
0
p_{n+1}
$p_{n+1} = 1.5612 p_n - 0.000861 p_n^2$
$p_{n+1} = p_n$
0
100
200
300
p_n
400
500
600


yes
no
yes
no

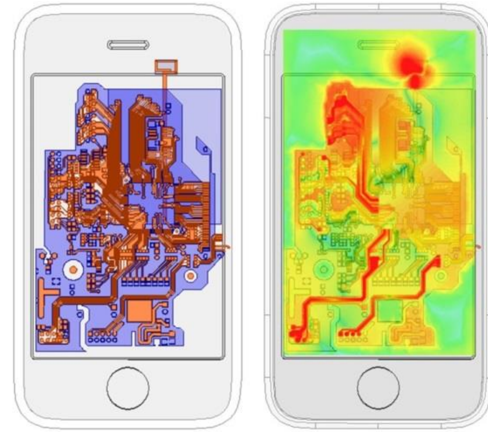

© NASA

Introduction générale: exemple 1

Circuits électroniques

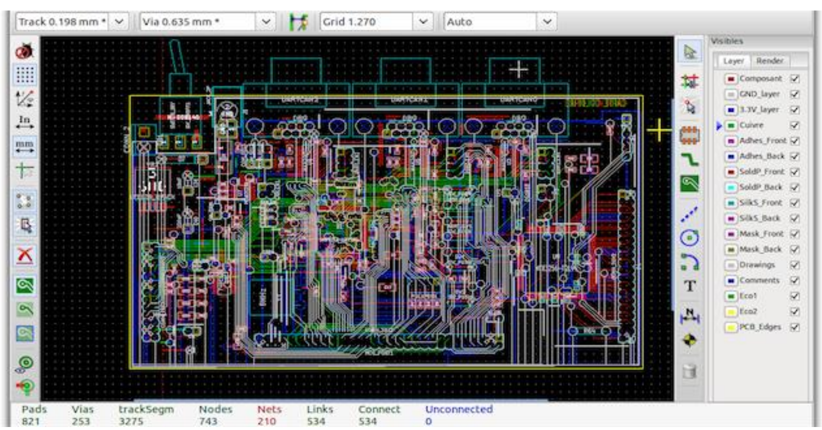


Analyse des circuits complexes et non linéaires



Dissipation de la chaleur

$$\frac{\partial u(\vec{\mathbf{r}})}{\partial t} - \alpha \nabla^2 u(\vec{\mathbf{r}}) = f(x,t)$$



Optimisation des connections circuits

Lois de Kirchhoff

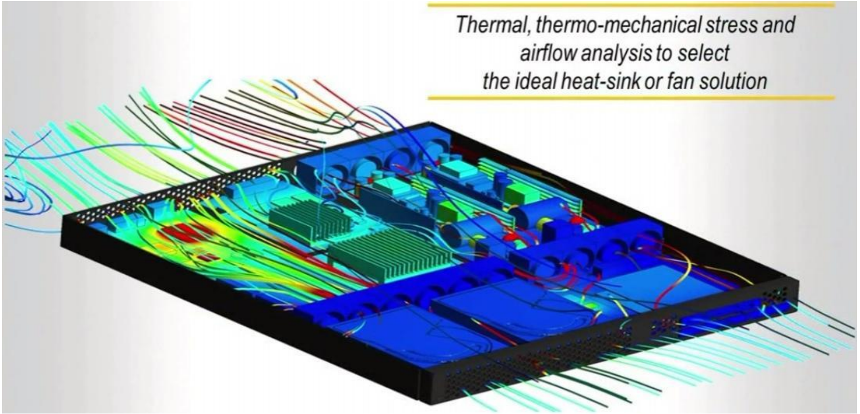
$$\sum_{i \in \sigma} i_k(t) = 0$$

Optimisation non linéaire

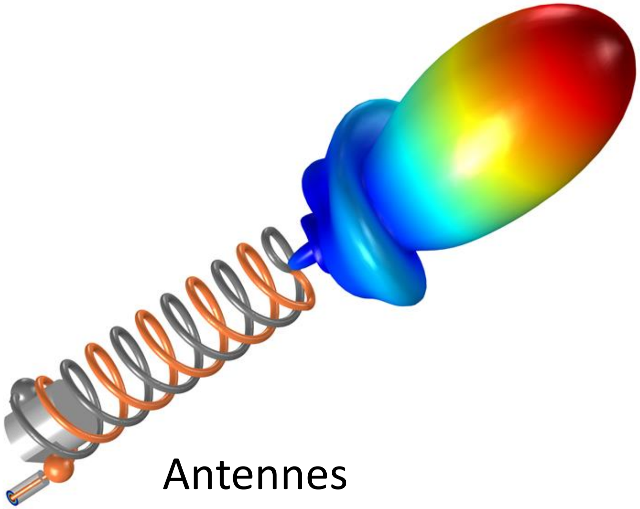
$$f(x) = \sum_{i=1}^{N} a_i \varphi\left(\left\| x - x_i \right\|, c_i\right) + R(x)$$

Introduction générale: exemple 2

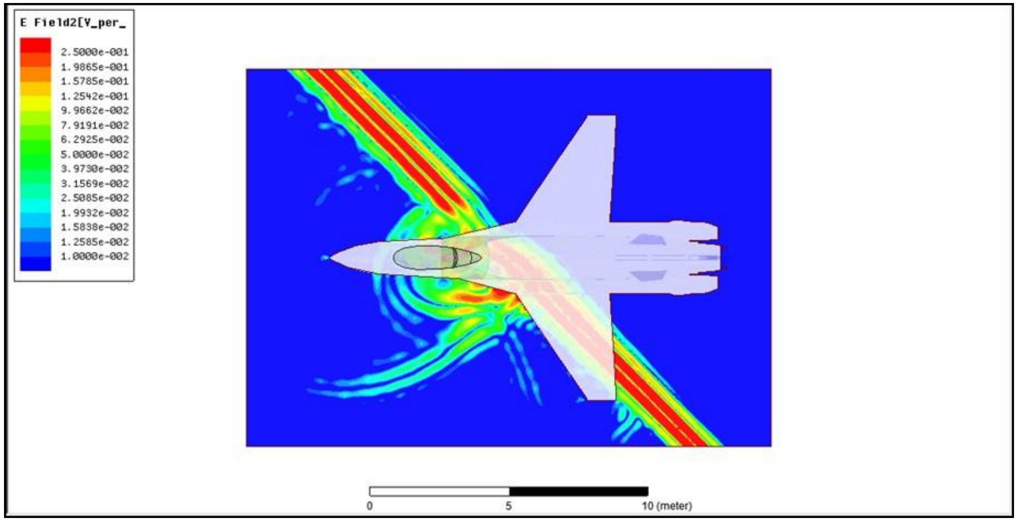
Electronique haute fréquence et propagation électromagnétique



Circuits haute fréquence



Antennes



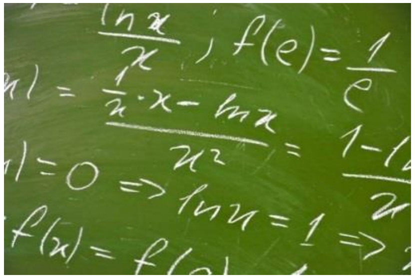
radar

$$
\begin{cases}
\vec{\nabla} \times \vec{\mathbf{E}}(\vec{\mathbf{r}}) = -\mu_0 \mu_r(\vec{\mathbf{r}}) \dfrac{\partial \vec{\mathbf{H}}(\vec{\mathbf{r}})}{\partial t} \\
\vec{\nabla} \times \vec{\mathbf{H}}(\vec{\mathbf{r}}) = \varepsilon_0 \varepsilon_r(\vec{\mathbf{r}}) \dfrac{\partial \vec{\mathbf{E}}(\vec{\mathbf{r}})}{\partial t} + \mathbf{J}(\vec{\mathbf{r}}) \\
\vec{\nabla} \cdot \vec{\mathbf{E}}(\vec{\mathbf{r}}) = -\dfrac{\rho(\vec{\mathbf{r}})}{\varepsilon_0 \varepsilon_r(\vec{\mathbf{r}})} \\
\vec{\nabla} \cdot \vec{\mathbf{H}}(\vec{\mathbf{r}}) = 0
\end{cases}
$$

Equations de Maxwell

Introduction générale

Bureautique, audio-vidéo,…

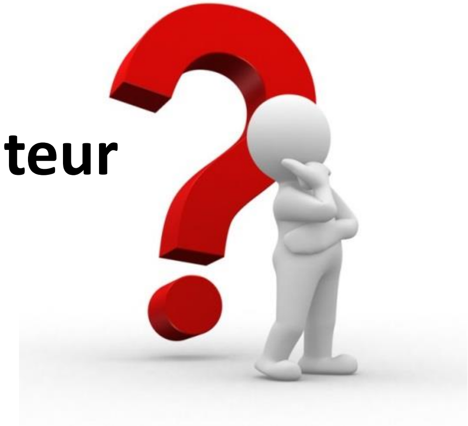


Calcul scientifique, Simulation, modélisation,…

Systèmes embarqués

les ordinateurs effectuent des **calculs**

Puis-je faire confiance aux résultats de mon ordinateur



Bourse de Vancouver (1982)

Création d'un nouvel indice de valeur initiale 1000
recalculé après chaque transaction et **tronqué après le 3e chiffre**

1000,000

22 mois

valeur calculée **524,881**

valeur exacte **1098,811**

Source du problème :
les erreurs de troncature ont le même signe

$$10,4562 \rightarrow 10,456 + \Delta$$
$$10,4568 \rightarrow 10,456 + \Delta$$
$$\Delta > 0$$

Source: Wikipedia

Ariane V (1996)

Source du problème : représentation de l'accélération horizontale.

- L'accélération horizontale maximum d'Ariane 4 était d'environ 64.

la variable a été **codée sur 8 bits**

- Ariane 5 était plus rapide: son accélération pouvait atteindre la valeur 300 (qui vaut 100101100 en binaire et nécessite 9 bits).

Overflow: 100000000 -> 00000000

04/06/1996 à Kourou
Le lanceur fut détruit après 37 secondes de vol.
European Space Agency (ESA)
(coût 500 million de dollars)

Le logiciel de contrôle face à des valeurs vraiment pas normales décida de l'autodestruction de la fusée.

Introduction générale: petites erreurs grandes conséquences

- **Dysfonctionnement dans l'unité de calcul en virgule flottante du *Pentium 5* (1994)**

valeur correcte
4 195 835,0 / 3 145 727,0 = 1,333 820 449 136 241 002

valeur erronée retournée par le processeur:
4 195 835,0 / 3 145 727,0 = 1,333 739 068 902 037 589

Source: Wikipedia

- **Bug dans Excel 2007**

Source du problème : Conversion des nombres binaires en caractères pour la visualisation

"That's because **0.1** has *no exact representation in binary*... it's a repeating binary number. It's sort of like how 1/3 has no representation in decimal. 1/3 is 0.33333333 and you have to keep writing 3's forever"

	A	B	C	D	E	F
1						
2		850.00	x	77.10	=	100,000.00
3		1700.00	x	38.55	=	100,000.00
4		3400.00	x	19.28	=	100,000.00
5		6800.00	x	9.64	=	100,000.00
6		13600.00	x	4.82	=	100,000.00
7		27200.00	x	2.41	=	100,000.00
8		425.00	x	154.20	=	100,000.00
9		212.50	x	308.40	=	100,000.00
10		106.25	x	616.80	=	100,000.00
11		53.13	x	1233.60	=	100,000.00
12		26.56	x	2467.20	=	100,000.00
13		13.28	x	4934.40	=	100,000.00
14						

Sheet1

Source: www.joelonsoftware.com

Plan du cours

- **Représentation d'un nombre en machine, erreurs numériques, etc…**
  - Introduction générale
  - **Représentation des nombres entiers et réels**
  - Opérations élémentaires en virgule flottante
  - Conditionnement d'un problème
  - Opérations complexes
  - Conclusion: différentes sources d'erreur
- Interpolation polynomiale
- Intégration numérique
- Résolution numérique des équations différentielles ordinaires (EDO)

- **On appelle base un entier β supérieur ou égal à 2**

- **Un chiffre sera un entier (symbole) compris entre 0 et β -1**

Exemples: β=2 0,1 (bit)
β=10 0,1,2,3,4,5,6,7,8,9
β=16 0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F

- **Un nombre x de n chiffres = une séquence**

X_{n-1}			X_1	X_0

telle que

$$x = \sum_{i=0}^{n-1} x_i \beta^i$$

- **Proposition**: Pour tout entier $y \geq 1$, il existe un entier unique *n* et des entiers $0 \leq x_i \leq \beta - 1$ avec $x_{n-1} \neq 0$ tels que

$$y = \sum_{i=0}^{n-1} x_i \beta^i$$

Exemples

1	1	0	1	0
x_4	x_3	x_2	x_1	x_0

$$\boxed{x = \sum_{i=0}^{n-1} x_i \beta^i}$$

$$\beta = 2, \qquad x = 0\cdot 2^0 + 1\cdot 2 + 0\cdot 2^2 + 1\cdot 2^3 + 1\cdot 2^4 = 2+8+16 = 26$$

$$\beta = 10, \qquad x = 0\cdot 10^0 + 1\cdot 10 + 0\cdot 10^2 + 1\cdot 10^3 + 1\cdot 10^4 = 11101$$

$$\beta = 16, \qquad x = 0\cdot 16^0 + 1\cdot 16 + 0\cdot 16^2 + 1\cdot 16^3 + 1\cdot 16^4 = 69648$$

$$x = 101 \Rightarrow \{x_i\} = ?$$

$\beta = 2$

$$\begin{array}{ll}
101:2 = 50 & r = 1 = x_0 \\
50:2 = 25 & r = 0 = x_1 \\
25:2 = 12 & r = 1 = x_2 \\
12:2 = 6 & r = 0 = x_3 \\
6:2 = 3 & r = 0 = x_4 \\
3:2 = 1 & r = 1 = x_5 \\
1:2 = 0 & r = 1 = x_6
\end{array}$$

1	1	0	0	1	0	1
x_6	x_5	x_4	x_3	x_2	x_1	x_0

Divisions successives module β

$$x = \sum_{i=0}^{n-1} x_i \beta^i = x_n \beta^n + \ldots + x_1 \beta^1 + x_0$$

$$\Rightarrow \frac{x}{\beta} = \sum_{i=0}^{n-1} x_i \beta^{i-1} = x_n \beta^{n-1} + \ldots + x_1 + \frac{x_0}{\beta}$$

x_0 reste de la division

$\beta = 16$

$$\begin{array}{ll}
101:16 = 6 & r = 5 = x_0 \\
6:16 = 0 & r = 6 = x_1
\end{array}$$

6	5
x_1	x_0

$x = -26$ ⇨ $\{x_i\} = ?$

$\beta = 2$

s	1	1	0	1	0

$+ \to s = 0$

$- \to s = 1$

$|x| = 26$

$n = 6$

$x = -26$

1	1	1	0	1	1

x=0 est représenté par 2 séquences:

0	0	0	0	0	0

et

1	0	0	0	0	0

$$-2^{n-1} + 1 \geq x \geq 2^{n-1} - 1$$

Complément à deux $n = 6$

si

$x > 0$ $\quad x \to \{x_i\}$

0	0	0	1	0	1

$\to x = 5$

$x < 0$ $\quad 2^{n-1} - x > 0 \to \{x_i\}$

1	0	0	1	0	1

$\to -(2^5 - 5) = -27$

$$-2^{n-1} \geq x \geq 2^{n-1} - 1$$

Nombres entiers avec signe: exemple

$$x = -131 \Rightarrow \{x_i\} = ?$$

- $|x| = 131 \qquad \beta = 2$

$$\begin{array}{ll}
131 : 2 = 65 & r = 1 = x_0 \\
65 : 2 = 32 & r = 1 = x_1 \\
32 : 2 = 16 & r = 0 = x_2 \\
16 : 2 = 8 & r = 0 = x_3 \\
8 : 2 = 4 & r = 0 = x_4 \\
4 : 2 = 2 & r = 0 = x_5 \\
2 : 2 = 1 & r = 0 = x_6 \\
1 : 2 = 0 & r = 1 = x_7
\end{array}$$

$$|x| = 131 \quad \begin{array}{cccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ x_7 & x_6 & x_5 & x_4 & x_3 & x_2 & x_1 & x_0 \end{array}$$

- $x = -131$

$$\overline{0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 1} + 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1$$

$$= 1\ 0\ 1\ 1\ 1\ 1\ 1\ 0\ 0 + 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1$$

$$= \begin{array}{ccccccccc} 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ x_8 & x_7 & x_6 & x_5 & x_4 & x_3 & x_2 & x_1 & x_0 \end{array}$$

Nombres réels

$$x = s\left(\underbrace{\sum_{i=0}^{n-1} x_i\beta^i}_{x_e} + \underbrace{\sum_{j=-1}^{-q} x_j\beta^j}_{x_f}\right)$$

Signe

Partie entière

Partie fractionnelle

Remarque: possibilité que $q = \infty$

Exemple: β=10   1/3=0,3333333…

Un ordinateur possède une **mémoire limitée**

Nombre limité de chiffres non nuls!!!

Solution

Ecriture à virgule flottante normalisée

On fixe un nombre r de chiffre significatifs

$$x = s\left(\sum_{i=-1}^{-r} x_i\beta^{-i}\right)\cdot\beta^k \rightarrow s\ 0,x_{-1}x_{-2}\ldots x_{-r}\cdot\beta^k$$

On place la virgule juste avant le premier chiffre non nul du développement (virgule flottante)

Nombres réels

s	x_{-1}	x_{-2}			x_{-r}	y_1	y_2			y_k

Signe 1 chiffre — r chiffres significatifs — k chiffres pour représenter l'exposant ⇒ K+r+1 chiffres

Mantisse — Exposant

Développement illimité (exact) de x

$$x = 0, x_{-1}x_{-2}\ldots x_{-r}x_{-r-1}\ldots\ldots\beta^{j}$$

Hypothèse: j représentable avec k chiffres

$$m' = \begin{cases} 0, x_{-1}x_{-2}\ldots x_{-r} & \text{si } x_{-r-1} < \beta/2 \\ 0, x_{-1}x_{-2}\ldots x_{-r} + \beta^{-r} & \text{si } x_{-r-1} \geq \beta/2 \end{cases}$$

On définit **l'arrondi de x**, noté $fl(x) = m'\beta^{j}$

Les r-1 premiers chiffres sont conservés et peuvent être considérés comme exacts !!!

Nombres réels: erreur relatif

$$e_r \triangleq \frac{|fl(x)-x|}{|x|} = \frac{|m'\beta^j - 0,x_{-1}x_{-2}\dots x_{-r}x_{-r-1}\dots\dots\beta^j|}{|0,x_{-1}x_{-2}\dots x_{-r}x_{-r-1}\dots\dots\beta^j|} = \frac{|0,0\dots0(a_{-r}-x_{-r})(-x_{-r-1})\dots\dots|}{|0,x_{-1}x_{-2}\dots x_{-r}x_{-r-1}\dots\dots|}$$

Par la définition d'arrondi on a $(a_{-r}-x_{-r})(-x_{-r-1})\dots \le \frac{\beta}{2}\beta^{-r-1}$

$$\frac{|fl(x)-x|}{|x|} \le \frac{\frac{\beta}{2}\beta^{-r-1}}{|0,x_{-1}x_{-2}\dots x_{-r}x_{-r-1}\dots\dots|} \le \frac{\frac{\beta}{2}\beta^{-r-1}}{\beta^{-1}} = \frac{\beta}{2}\beta^{-r}$$

$$x_{-1} \ne 0 \Rightarrow |0,x_{-1}x_{-2}\dots x_{-r}x_{-r-1}\dots\dots| \ge \beta^{-1}$$

- $\varepsilon = \frac{\beta}{2}\beta^{-r}$ est appelée **précision machine**.

$$fl(x) = x(1+\alpha) \qquad |\alpha| \le \varepsilon$$

Nombres réels: exemples

Exemple 1: $x=-131.23411$ β=10 r=6 (chiffres mantisse)

$$x=-131.23411 \qquad x=-0.13123411\cdot 10^3 \qquad fl(x)=-0.131234\cdot 10^3 \qquad \varepsilon=\frac{\beta}{2}\beta^{-r}=5\cdot 10^{-6}$$

$$e_r=\frac{\left|-0.13123411\cdot 10^3+0.131234\cdot 10^3\right|}{\left|-0.13123411\cdot 10^3\right|}=8.382\cdot 10^{-7}<\varepsilon$$

Exemple 2: $x=12.23$ β=2 r=6 (chiffres mantisse) $\varepsilon=\frac{\beta}{2}\beta^{-r}=2^{-6}=1.56\cdot 10^{-2}$

Partie entière (x_e)

$$12:2=6 \quad r=0=x_0$$
$$6:2=3 \quad r=0=x_1$$
$$3:2=1 \quad r=1=x_2$$
$$1:2=0 \quad r=1=x_3$$

Partie fractionnelle (x_f)

Multiplications successives

$$x_f=\sum_{i=-1}^{-q}x_i\beta^i=x_{-1}\beta^{-1}+\ldots+x_{-q+1}\beta^{-q+1}+x_{-q}\beta^{-q}$$

$$\Rightarrow \beta x_f=x_{-1}+x_{-2}\beta^{-1}+\ldots+x_{-q}\beta^{-q+1}$$

x_{-1} partie entière βx_f

$$0,23\cdot 2=0,46 \Rightarrow x_{-1}=0$$
$$0,46\cdot 2=0,92 \Rightarrow x_{-2}=0$$
$$0,92\cdot 2=1,84 \Rightarrow x_{-3}=1$$
$$(1,84-1)\cdot 2=1,68 \Rightarrow x_{-4}=1$$
$$0,68\cdot 2=1,36 \Rightarrow x_{-5}=1$$
$$0,36\cdot 2=0,72 \Rightarrow x_{-6}=0$$
$$0,72\cdot 2=1,44 \Rightarrow x_{-7}=1$$
$$0,44\cdot 2=0,88 \Rightarrow x_{-8}=0$$

$$x=\underset{x_3x_2x_1x_0}{1100},\underset{x_{-1}x_{-2}x_{-3}x_{-4}x_{-5}x_{-6}x_{-7}x_{-8}}{00111010}1\ldots=0,1100001110101\ldots\cdot 2^4 \Rightarrow fl(x)=0,11000\cdot 2^4$$

Standard IEEE 754-1985

IEEE

Institute of Electrical and Electronics Engineers

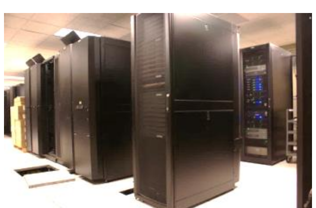
Elle est la norme la plus employée actuellement pour le calcul des nombres à virgule flottante dans le domaine informatique (PC, système embarqué, tablette, smartphone, …).

1) simple precision (32 bits) : 1 bit de signe, 8 bits d'exposant, 23 bits de mantisse
   0,1 + 23 bits-> m=24
2) double precision (64 bits) : 1 bit de signe, 11 bits d'exposant, 52 bits de mantisse
   0,1 + 53 bits-> m=53

$$s \quad \overbrace{x_{-1}\; x_{-2}\; \quad \quad x_{-m}}^{m} \cdot \beta^{k} \qquad L \le k \le U$$

1) en base 2, L = -126, U = 127, eps = 2^{-24};
   en base 10, L≈–38, U≈38, eps≈10^{-8}

2) en base 2, L = -1023, U = 1024, eps≈2^{-53};
   en base 10, L≈-308, U≈308, eps≈10^{-16}

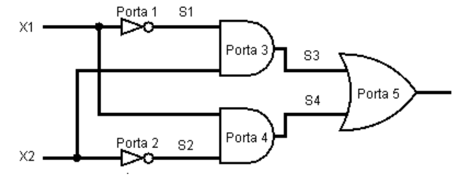




Plan du cours

- **Représentation d'un nombre en machine, erreurs numériques, etc…**
  - Introduction générale
  - Représentation des nombres entiers et réels
  - **Opérations élémentaires en virgule flottante**
  - Conditionnement d'un problème
  - Opérations complexes
  - Conclusion: différentes sources d'erreur
- Interpolation polynomiale
- Intégration numérique
- Résolution numérique des équations différentielles ordinaires (EDO)

Opérations élémentaires en virgule flottante



Opérations théoriques

- Addition $x+y$
- Soustraction $x-y$
- Multiplication $x\cdot y$
- Division x/y

Opérations en virgule flottante

$$x\oplus y = fl\left(fl\left(x\right)+ fl\left(y\right)\right)$$

$$x\ominus y = fl\left(fl\left(x\right)- fl\left(y\right)\right)$$

$$x\odot y = fl\left(fl\left(x\right)\cdot fl\left(y\right)\right)$$

$$x\oslash y = fl\left(fl\left(x\right)/ fl\left(y\right)\right)$$

Exemple: propagation des erreurs

$$fl\left(x\right)=x\left(1+\alpha\right),\quad \left|\alpha\right|<\varepsilon$$

$$fl\left(y\right)=y\left(1+\beta\right),\quad \left|\beta\right|<\varepsilon$$

$$x\oplus y = fl\left(x\left(1+\alpha\right)+y\left(1+\beta\right)\right)$$

$$=\left(x\left(1+\alpha\right)+y\left(1+\beta\right)\right)\left(1+\delta\right),\quad \left|\delta\right|<\varepsilon$$

$$=\left(x+x\alpha+y+y\beta\right)\left(1+\delta\right)$$

$$=\left(x+y\right)\left(1+\gamma\right)$$

$$\gamma = x\alpha + y\beta + \left(x+x\alpha+y+y\beta\right)\delta \Rightarrow \left|\gamma\right| \le \left|x\right|\left|\alpha\right| + \left|y\right|\left|\beta\right| + \left|x\right|\left|\delta\right| + \left|x\right|\left|\alpha\right|\left|\delta\right| + \left|y\right|\left|\delta\right| + \left|y\right|\left|\beta\right|\left|\delta\right|$$

$$\left|\gamma\right| \le \left|x\right|\varepsilon + \left|y\right|\varepsilon + \left|x\right|\varepsilon + \left|x\right|\varepsilon^2 + \left|y\right|\varepsilon + \left|y\right|\varepsilon^2 \simeq \left(2\left|x\right| + \left|2y\right|\right)\varepsilon$$

Opérations élémentaires en virgule flottante: propriétés

- **propriété de commutativité**

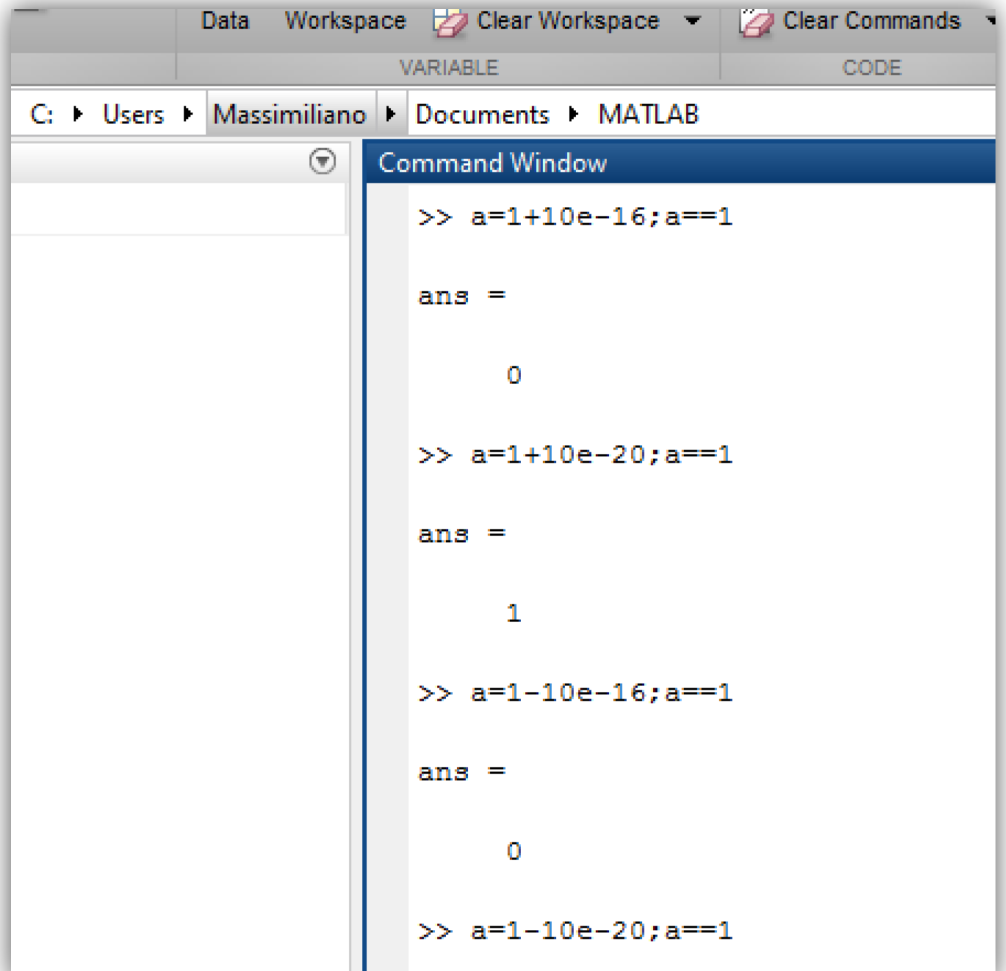
$$a \oplus b = b \oplus a \qquad a \odot b = b \odot a$$

- **propriétés d'associativité et de distributivité**

$$a \oplus (b \oplus c) \neq (a \oplus b) \oplus c$$
$$a \odot (b \odot c) \neq (a \odot b) \odot c$$
$$a \odot (b \oplus c) \neq a \odot b \oplus a \odot c$$
$$a \odot (b / c) \neq a$$
$$(a / b) \odot b \neq a$$
$$(a \odot b) / c \neq (a / c) \odot b$$

Exemple 1: précision machine

$$1 \pm b = 1 \quad \text{si} \quad b \in \mathbb{R} \wedge |b| < \varepsilon$$



MATLAB (double précision)

Exemple 2: somme de deux nombres voisins

$$x_1 = 0{,}191019721 \cdot 10^3 \qquad x_2 = 0{,}191017083 \cdot 10^3$$

r=6 chiffres significatifs $\beta = 10, \quad x_1 \oplus x_2 = ?$

$$\varepsilon = \frac{\beta}{2}\beta^{-r} = 5 \cdot 10^{-6}$$

$fl(x_1)$	**1**	**9**	**1**	**0**	**2**	**0**	10^3
$fl(x_2)$	**1**	**9**	**1**	**0**	**1**	**7**	10^3
$fl(x_1)+fl(x_2)$	**3**	**8**	**2**	**0**	**3**	**7**	10^3

$$e_{x_1} = \frac{|fl(x_1) - x_1|}{|x_1|} = 1.4606 \cdot 10^{-6} < \varepsilon$$

$$e_{x_2} = \frac{|fl(x_2) - x_2|}{|x_2|} = 4.3452 \cdot 10^{-7} < \varepsilon$$

6 chiffres significatifs

$$x_1 + x_2 = 3.82036 \cdot 10^2 \qquad x_1 \oplus x_2 = 3.82037 \cdot 10^2 \qquad e_{x_1 - x_2} = \frac{|3.82036 - 3.82037|}{|3.82036|} = -5.13 \cdot 10^{-7}$$

Même ordre de précision ($< \varepsilon$)

Exemple 3: différence de deux nombres voisins

$$x_1 = 0{,}191019721 \cdot 10^3 \qquad x_2 = 0{,}191017083 \cdot 10^3$$

6 chiffres significatifs $\beta = 10, \quad x_1 \ominus x_2 = ?$

$$\varepsilon = \frac{\beta}{2}\beta^{-r} = 5 \cdot 10^{-6}$$

$fl(x_1)$	**1**	**9**	**1**	**0**	**2**	**0**	10^3
$fl(x_2)$	**1**	**9**	**1**	**0**	**1**	**7**	10^3
$fl(x_1) - fl(x_2)$	**0**	**0**	**0**	**0**	**0**	**3**	10^3

$$e_{x_1} = \frac{|fl(x_1) - x_1|}{|x_1|} = 1.4606 \cdot 10^{-6} < \varepsilon$$

$$e_{x_2} = \frac{|fl(x_2) - x_2|}{|x_2|} = 4.3452 \cdot 10^{-7} < \varepsilon$$

1 chiffre significatif

$$x_1 - x_2 = 2.63799 \cdot 10^{-3} \qquad x_1 \ominus x_2 = 3 \cdot 10^{-3}$$

!!!

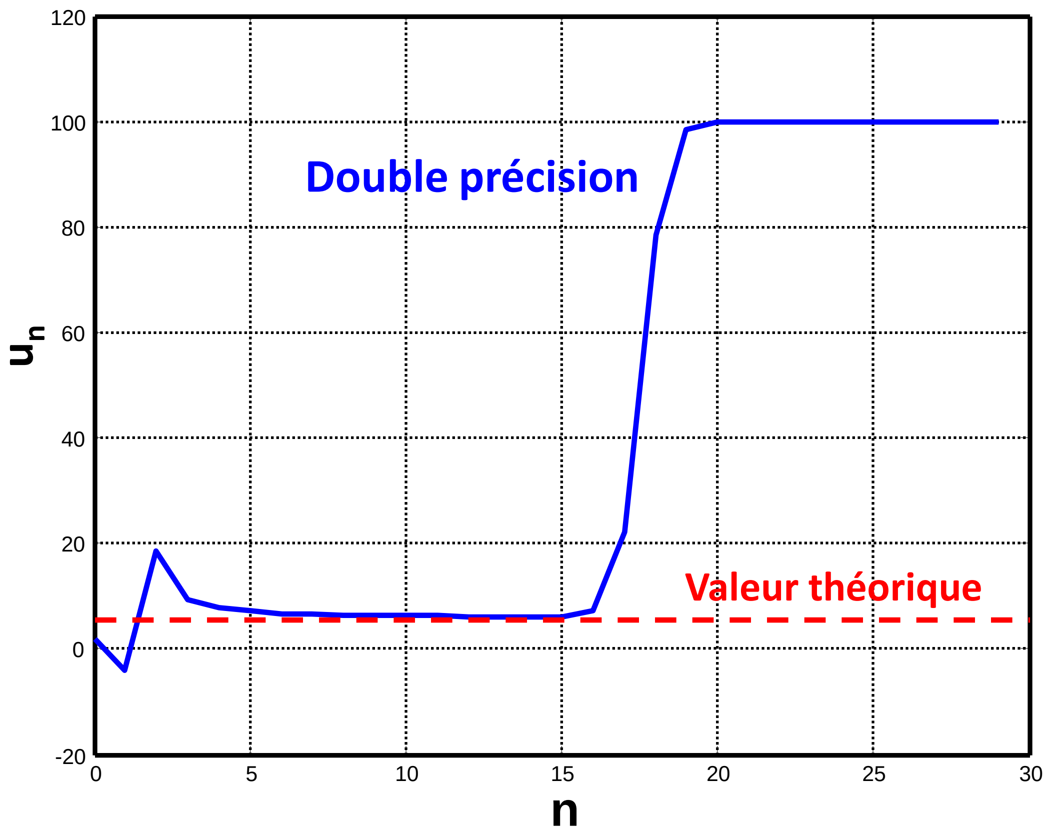
$$e_{x_1 - x_2} = \frac{|2.6380 - 3|}{|2.6380|} = 0.1372 \simeq 13.7\%$$

<u>Phénomène de cancellation</u>

$$e_{x_1 - x_2} >> \varepsilon$$

Exemple 3: Suite numérique de Muller

$$u_n = \begin{cases} u_0 = 2 \\ u_1 = -4 \\ u_{n+1} = 111 - \dfrac{1130}{u_n} + \dfrac{3000}{u_n \cdot u_{n-1}} \end{cases}$$



Exemple 4: racines d'une équation du second degré

$$x^2 - 2ax + \varepsilon = 0 \qquad \begin{cases} x_1 = a + \sqrt{a^2 - \varepsilon} \\ x_2 = a - \sqrt{a^2 - \varepsilon} \end{cases}$$

- Si $\varepsilon \ll a \Rightarrow \sqrt{a^2 - \varepsilon} \simeq |a| \Rightarrow \begin{cases} x_1 = a + |a| \\ x_2 = a - |a| \end{cases} \Rightarrow \begin{matrix} x_2 = 0 \quad \text{si} \quad a > 0 \\ x_1 = 0 \quad \text{si} \quad a < 0 \end{matrix}$ ***Phénomène de cancellation***

Solution: $\varepsilon = x_1 x_2$ ⇨

$$\boxed{\begin{cases} x_1 = a + \operatorname{sgn}(a)\sqrt{a^2 - \varepsilon} \\ x_2 = \dfrac{\varepsilon}{x_1} \end{cases}}$$

- Si $\varepsilon \simeq a^2 \Rightarrow \sqrt{a^2 - \varepsilon} \simeq 0 \Rightarrow \begin{cases} x_1 = a \\ x_2 = a \end{cases}$ ***Phénomène de cancellation***

aucune solution possible

Plan du cours

- **Représentation d'un nombre en machine, erreurs numériques, etc…**
  - Introduction générale
  - Représentation des nombres entiers et réels
  - Opérations élémentaires en virgule flottante
  - **Conditionnement d'un problème**
  - Opérations complexes
  - Conclusion: différentes sources d'erreur
- Interpolation polynomiale
- Intégration numérique
- Résolution numérique des équations différentielles ordinaires (EDO)

Conditionnement d'un problème

P

Entrée $x(t)$ ⇒  ⇒ $y(t)$ **Sortie**

Forme explicite

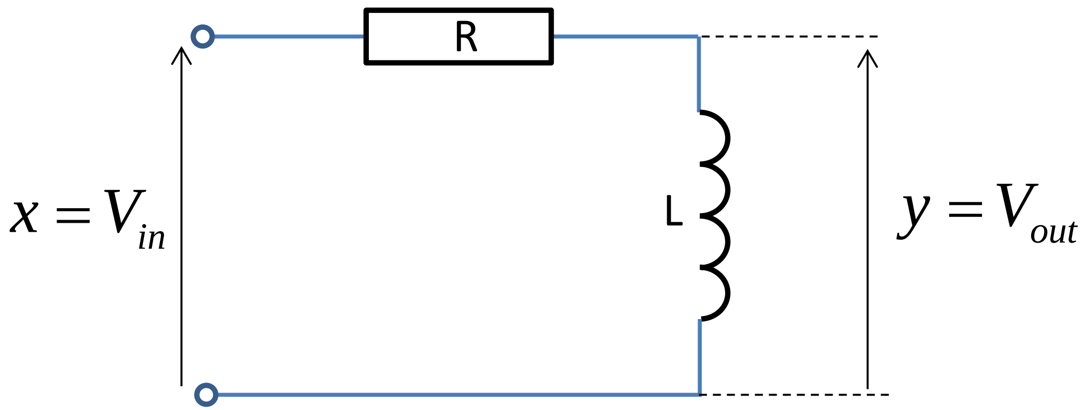
$$y(t) = f\left[x(t)\right]$$

Forme implicite

$$g\left[x(t), y(t)\right] = 0$$

Conditionnement d'un problème

Exemple



Forme explicite

$$V_{out} = V_{in} \frac{j\omega L}{R + j\omega L} \qquad f(x) = \left[\frac{j\omega L}{R + j\omega L} \right] x$$

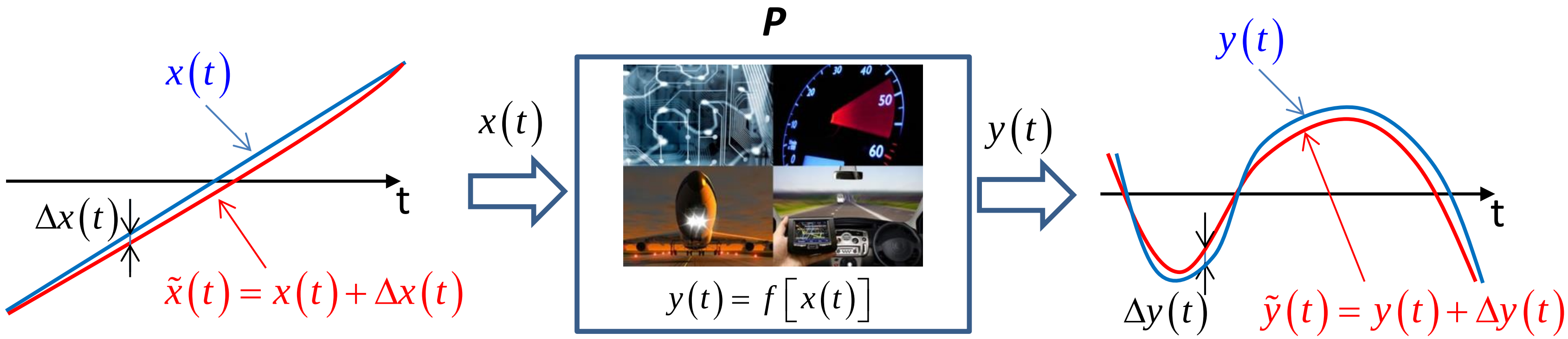
Forme implicite

Loi des mailles

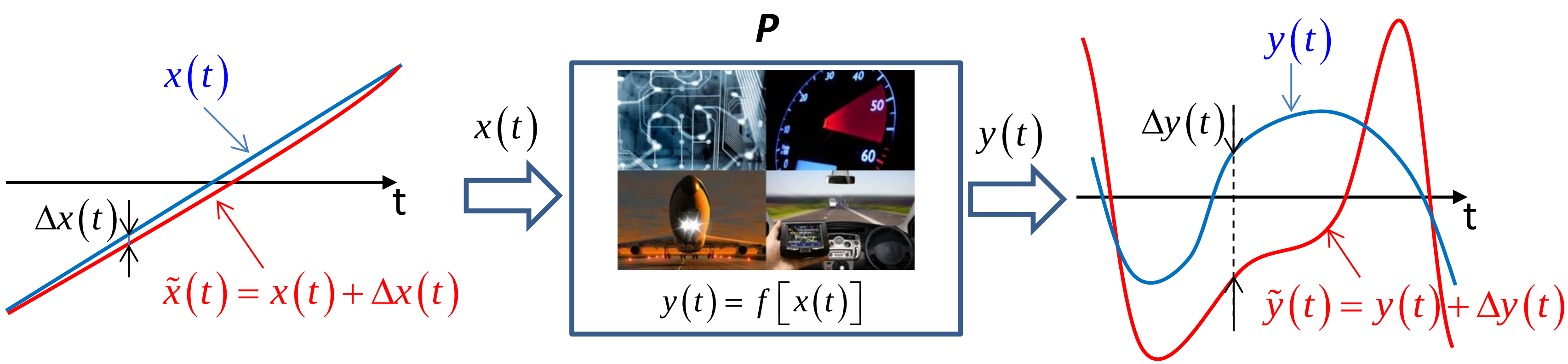
$$V_{in} + \frac{R}{R + j\omega L} + V_{out} = 0 \qquad g(x, y) = x + \frac{R}{R + j\omega L} + y$$

Conditionnement d'un problème

- le **conditionnement** mesure la dépendance de la solution d'un **problème numérique *P*** par rapport aux données du problème (mesure de la difficulté du calcul numérique)

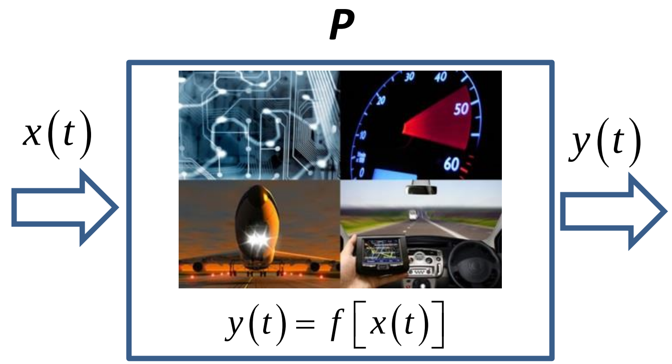


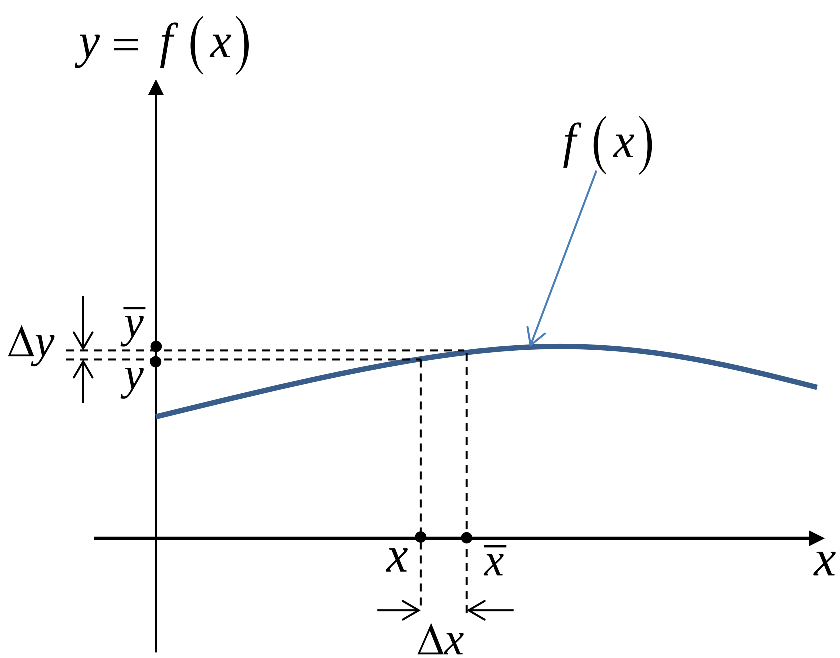
On dit que *P* est **bien conditionné** si un petit changement Δx de x entraîne un petit changement Δ*Y* de *Y*.



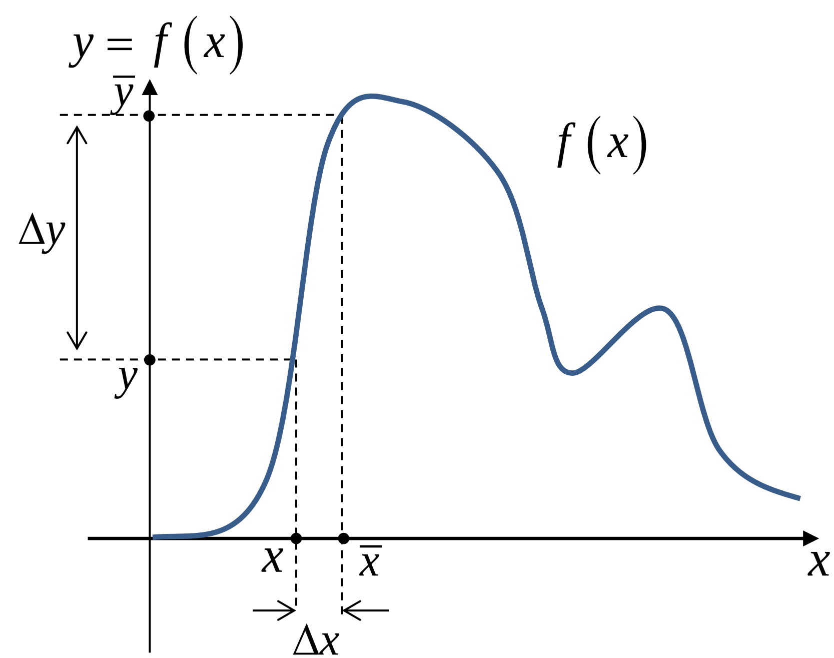
On dit que *P* est **mal conditionné** si un petit changement Δx de x entraîne un grand changement Δ*Y* de *Y*.

Conditionnement d'un problème





Problème **bien conditionné**



Problème **mal conditionné**

Conditionnement d'un problème

- Le **conditionnement d'un problème par rapport à l'erreur absolue** est donnée par le nombre de condition absolu *k*

$$k = \frac{\|\Delta y\|}{\|\Delta x\|}$$

rapport d'amplification des erreurs

- Le **conditionnement d'un problème par rapport à l'erreur relative** est donnée par le nombre de condition relatif k_r

$$k_r = \frac{\dfrac{\|\Delta y\|}{\|y\|}}{\dfrac{\|\Delta x\|}{\|x\|}} = \frac{\|\Delta y\|\|x\|}{\|\Delta x\|\|y\|}$$

rapport d'amplification des erreurs relatives

Le problème ***P*** est bien conditionné si *k* n'est pas *très grand*. Sinon, ce problème ***P*** est mal conditionné.

Normes vectorielles

$$\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$$

$$\|\mathbf{x}\| \geq 0 \qquad \|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = 0$$

(séparation)

$$\|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\|$$

(homogénéité)

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

(inégalité triangulaire)

Exemples

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{N} x_i^2}$$

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{N} |x_i|$$

$$\|\mathbf{x}\|_\infty = \max_{i=1,\dots,N} |x_i|$$

Normes matricielles

(subordonnée à une norme vectorielle)

$$\underline{\underline{\mathbf{A}}}, \underline{\underline{\mathbf{B}}} \in \mathbb{R}^{N \times N}$$

$$\|\underline{\underline{\mathbf{A}}}\| \geq 0 \qquad \|\underline{\underline{\mathbf{A}}}\| = 0 \Leftrightarrow \underline{\underline{\mathbf{A}}} = 0$$

$$\|\alpha \underline{\underline{\mathbf{A}}}\| = |\alpha| \|\underline{\underline{\mathbf{A}}}\|$$

$$\|\underline{\underline{\mathbf{A}}} + \underline{\underline{\mathbf{B}}}\| \leq \|\underline{\underline{\mathbf{A}}}\| + \|\underline{\underline{\mathbf{B}}}\|$$

$$\|\underline{\underline{\mathbf{A}}}\underline{\underline{\mathbf{B}}}\| = \|\underline{\underline{\mathbf{A}}}\| \|\underline{\underline{\mathbf{B}}}\|$$

Exemples

$$\|\underline{\underline{\mathbf{A}}}\|_p = \max_{\|\mathbf{x}\| \neq 0} \frac{\|\underline{\underline{\mathbf{A}}}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

$$\|\underline{\underline{\mathbf{A}}}\|_2 = \sqrt{\rho\left\{\underline{\underline{\mathbf{A}}}^T \underline{\underline{\mathbf{A}}}\right\}}$$

En dimension finie, toutes les normes sont équivalentes

Exemple 1

Soit ***P*** définit par $y=x_1+x_2$

$$\tilde{x}_1 = x_1 + \Delta x_1 \qquad \tilde{x}_2 = x_2 + \Delta x_2$$

$$\tilde{y} = f\left(\tilde{x}_1, \tilde{x}_2\right) = \tilde{x}_1 + \tilde{x}_2 = \left(x_1 + \Delta x_1\right) + \left(x_2 + \Delta x_2\right) = \left(x_1 + x_2\right) + \left(\Delta x_1 + \Delta x_2\right) = y + \Delta y$$

$$\left\|y\right\| = \left|x_1 + x_2\right| \qquad \left\|\Delta y\right\| = \left|\Delta x_1 + \Delta x_2\right|$$

$$\mathbf{\Delta x} = \left(\Delta x_1, \Delta x_2\right)$$

$$\left|\Delta x_1\right|, \left|\Delta x_2\right| \le \varepsilon$$

$$\Rightarrow$$

$$\left\|\mathbf{x}\right\|_1 = \left\|\left(x_1, x_2\right)\right\|_1 = \left|x_1\right| + \left|x_2\right|$$

$$\left\|\mathbf{\Delta x}\right\|_1 = \left\|\left(\Delta x_1, \Delta x_2\right)\right\|_1 = \left|\Delta x_1\right| + \left|\Delta x_2\right| \le 2\varepsilon$$

$$\Downarrow$$

$$k = \frac{\left\|\Delta y\right\|}{\left\|\mathbf{\Delta x}\right\|_1} = \frac{\left|\Delta x_1 + \Delta x_2\right|}{\left|\Delta x_1\right| + \left|\Delta x_2\right|} \le \frac{\left|\Delta x_1\right| + \left|\Delta x_2\right|}{\left|\Delta x_1\right| + \left|\Delta x_2\right|} = 1$$

$$k_r = \frac{\left\|\Delta y\right\| \left\|\mathbf{x}\right\|_1}{\left\|\mathbf{\Delta x}\right\|_1 \left\|y\right\|} = \frac{\left|\Delta x_1 + \Delta x_2\right| \left(\left|x_1\right| + \left|x_2\right|\right)}{\left(\left|\Delta x_1\right| + \left|\Delta x_2\right|\right)\left|x_1 + x_2\right|} \le \frac{\left(\left|\Delta x_1\right| + \left|\Delta x_2\right|\right)\left(\left|x_1\right| + \left|x_2\right|\right)}{\left(\left|\Delta x_1\right| + \left|\Delta x_2\right|\right)\left|x_1 + x_2\right|} = \frac{\left|x_1\right| + \left|x_2\right|}{\left|x_1 + x_2\right|}$$

Exemple 2

Soit ***P*** définit par $y=x_1-x_2$

$$\tilde{y} = f\left(\tilde{x}_1,\tilde{x}_2\right) = \tilde{x}_1-\tilde{x}_2 = \left(x_1+\Delta x_1\right)-\left(x_2+\Delta x_2\right) = \left(x_1-x_2\right)+\left(\Delta x_1-\Delta x_2\right) = y+\Delta y$$

$$\|y\| = |x_1-x_2| \qquad \|\Delta y\| = |\Delta x_1-\Delta x_2| \le 2\max\left\{|\Delta x_1|,|\Delta x_2|\right\}$$

$$\|\mathbf{\Delta x}\|_\infty = \left\|\left(\Delta x_1,\Delta x_2\right)\right\|_\infty = \max\left\{|\Delta x_1|,|\Delta x_2|\right\} \qquad \|\mathbf{x}\|_\infty = \left\|\left(x_1,x_2\right)\right\|_\infty = \max\left\{|x_1|,|x_2|\right\}$$

$$k = \frac{\|\Delta y\|}{\|\mathbf{\Delta x}\|_\infty} = \frac{|\Delta x_1-\Delta x_2|}{\max\left\{|\Delta x_1|,|\Delta x_2|\right\}} \le \frac{2\max\left\{|\Delta x_1|,|\Delta x_2|\right\}}{\max\left\{|\Delta x_1|,|\Delta x_2|\right\}} = 2$$

$$k_r = \frac{\|\Delta y\|\|\mathbf{x}\|_\infty}{\|\mathbf{\Delta x}\|_\infty\|y\|} = \frac{|\Delta x_1-\Delta x_2|\max\left\{|x_1|,|x_2|\right\}}{\max\left\{|\Delta x_1|,|\Delta x_2|\right\}|x_1-x_2|} \le \frac{2\max\left\{|\Delta x_1|,|\Delta x_2|\right\}\max\left\{|x_1|,|x_2|\right\}}{\max\left\{|\Delta x_1|,|\Delta x_2|\right\}|x_1-x_2|} = \frac{2\max\left\{|x_1|,|x_2|\right\}}{|x_1-x_2|}$$

Si $x_1 \simeq x_2$ le conditionnement relatif est très grand. La soustraction est donc mal conditionnée. (phénomène d'annulation).

Soit ***P*** définit par y=x_1*x_2

$$\tilde{y} = \tilde{x}_1\tilde{x}_2 = \left(x_1 + \Delta x_1\right)\left(x_2 + \Delta x_2\right) = \left(x_1x_2\right) + \left(x_1\Delta x_2 + x_2\Delta x_1 + \Delta x_1\Delta x_2\right) = y + \Delta y$$

$$\|\mathbf{x}\|_2 = \left\|\left(x_1, x_2\right)\right\|_2$$

$$\|\Delta\mathbf{x}\|_2 = \left\|\left(\Delta x_1, \Delta x_2\right)\right\|_2$$

$$\|y\| = |x_1x_2|$$

$$\|\Delta y\| = |x_1\Delta x_2 + x_2\Delta x_1 + \Delta x_1\Delta x_2| \simeq |x_1\Delta x_2 + x_2\Delta x_1| = \left|\left(x_1, x_2\right)\cdot\left(\Delta x_2, \Delta x_1\right)\right|$$

$$= |\mathbf{x}\cdot\Delta\mathbf{x}| \leq \|\mathbf{x}\|_2\,\|\Delta\mathbf{x}\|_2 = \|x_1, x_2\|_2\,\|\Delta x_1, \Delta x_2\|_2$$

$$\Downarrow$$

$$k = \frac{\|\Delta y\|}{\|\Delta\mathbf{x}\|_2} \leq \frac{\|x_1, x_2\|_2\,\|\Delta x_1, \Delta x_2\|_2}{\|\Delta x_1, \Delta x_2\|_2} = \|x_1, x_2\|_2 = \|\mathbf{x}\|_2$$

$$k_r = \frac{\|\Delta y\|\|\mathbf{x}\|_2}{\|\Delta\mathbf{x}\|_2\,\|y\|} \leq \|\mathbf{x}\|_2\frac{\|\mathbf{x}\|_2}{\|y\|} = \frac{\left(\|\mathbf{x}\|_2\right)^2}{|x_1x_2|} = \frac{{x_1}^2}{|x_1x_2|} + \frac{{x_2}^2}{|x_1x_2|} = \frac{|x_1||x_1|}{|x_1||x_2|} + \frac{|x_2||x_2|}{|x_1||x_2|} = \frac{|x_1|}{|x_2|} + \frac{|x_2|}{|x_1|}$$

Conditionnement d'un problème: système linéaire

Soit ***P*** définit par le système linéaire $\underline{\underline{\mathbf{A}}}\mathbf{y} = \mathbf{x}$ de solution exacte et unique.

Forme implicite

$$\underline{\underline{\mathbf{A}}}\mathbf{y} - \mathbf{x} = 0$$

Forme explicite

$$\mathbf{y} = \underline{\underline{\mathbf{A}}}^{-1}\mathbf{x} = f\left(\mathbf{x}\right)$$

$$\left\|\mathbf{x}\right\| = \left\|\underline{\underline{\mathbf{A}}}\mathbf{y}\right\| \le \left\|\underline{\underline{\mathbf{A}}}\right\|\left\|\mathbf{y}\right\| \Rightarrow \left\|\mathbf{y}\right\| \ge \frac{\left\|\mathbf{x}\right\|}{\left\|\underline{\underline{\mathbf{A}}}\right\|}$$

$$\tilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$$

$$\underline{\underline{\mathbf{A}}}\tilde{\mathbf{y}} = \tilde{\mathbf{x}} \Leftrightarrow \underline{\underline{\mathbf{A}}}\left(\mathbf{y} + \Delta\mathbf{y}\right) = \mathbf{x} + \Delta\mathbf{x} \Rightarrow \underline{\underline{\mathbf{A}}}\Delta\mathbf{y} = \Delta\mathbf{x} \Rightarrow \Delta\mathbf{y} = \underline{\underline{\mathbf{A}}}^{-1}\Delta\mathbf{x}$$

$$\left\|\Delta\mathbf{y}\right\| = \left\|\underline{\underline{\mathbf{A}}}^{-1}\Delta\mathbf{x}\right\| \le \left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|\left\|\Delta\mathbf{x}\right\|$$

$$\frac{\left\|\Delta\mathbf{y}\right\|}{\left\|\mathbf{y}\right\|} \le \frac{\left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|\left\|\Delta\mathbf{x}\right\|}{\frac{\left\|\mathbf{x}\right\|}{\left\|\underline{\underline{\mathbf{A}}}\right\|}} = \left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|\left\|\underline{\underline{\mathbf{A}}}\right\|\frac{\left\|\Delta\mathbf{x}\right\|}{\left\|\mathbf{x}\right\|}$$

$$k_r \le \left\|\underline{\underline{\mathbf{A}}}^{-1}\right\|\left\|\underline{\underline{\mathbf{A}}}\right\|$$

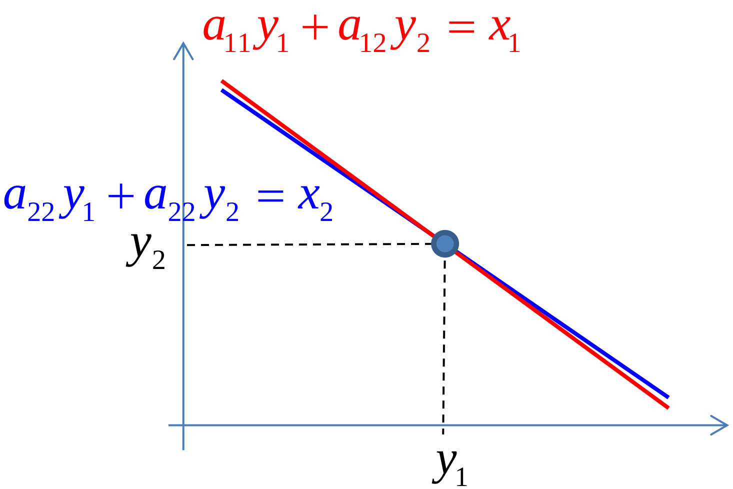
Conditionnement d'un problème: système linéaire

$$\underline{\underline{\mathbf{A}}}\mathbf{y} = \mathbf{x} \qquad \mathbf{y} = \underline{\underline{\mathbf{A}}}^{-1}\mathbf{x} = f(\mathbf{x})$$

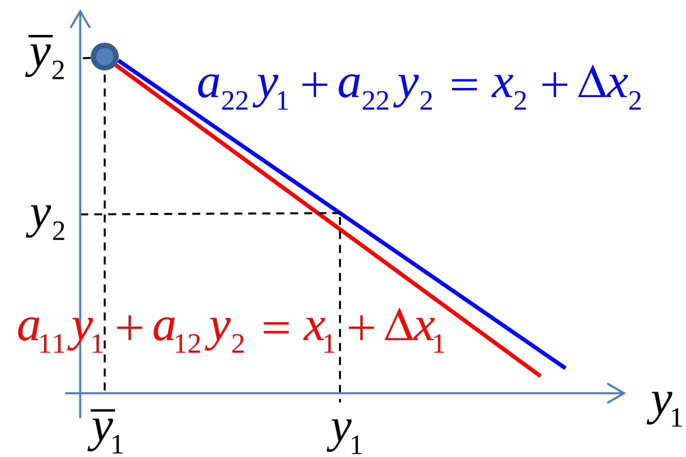
$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Leftrightarrow \begin{cases} a_{11}y_1 + a_{12}y_2 = x_1 \\ a_{21}y_1 + a_{22}y_2 = x_2 \end{cases}$$

Système linéaire mal conditionnée: ***droites presque parallèles***

Solution exacte $\underline{\underline{\mathbf{A}}}\mathbf{y} = \mathbf{x}$



Solution système perturbé $\underline{\underline{\mathbf{A}}}\mathbf{y} = \tilde{\mathbf{x}} = \mathbf{x} + \Delta\mathbf{x}$

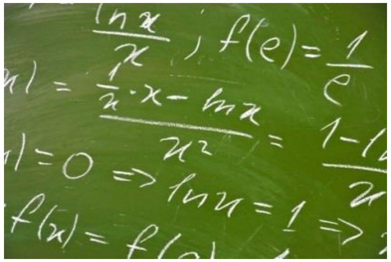


Exemple:
$$\begin{cases} 2y_1 + 2y_2 = 3 \\ 0.499y_1 + 1.001y_2 = 1.5 \end{cases}$$

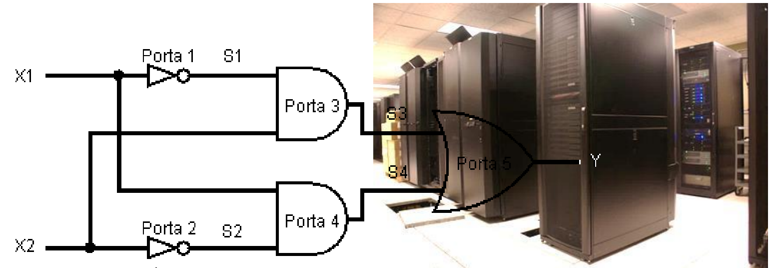
Plan du cours

- **Représentation d’un nombre en machine, erreurs numériques, etc…**
  - Introduction générale
  - Représentation des nombres entiers et réels
  - Opérations élémentaires en virgule flottante
  - Conditionnement d’un problème
  - **Opérations complexes**
  - Conclusion: différentes sources d’erreur
- Interpolation polynomiale
- Intégration numérique
- Résolution numérique des équations différentielles ordinaires (EDO)

Opérations complexes



$$f\left(x\right)$$



$$\oplus \quad \ominus \quad \odot \quad \oslash$$

Solution

Représentation de f comme combinaison des fonctions élémentaires

$$f\left(x\right)=\sum_{n=0}^{\infty} f_n\left(x\right) \qquad f\left(x\right)=\lim_{n\to\infty} u_n$$

f limite d'une série ou d'une suite

Valeur exacte pour un nombre d'itérations $n \to \infty$

Opérations complexes

Série de Taylor

$$f\left(x\right)=\sum_{k=0}^{n}\frac{f^{\left(k\right)}\left(x_0\right)}{k!}\left(x-x_0\right)^k+e\left(x,x_0\right), \qquad e\left(x,x_0\right)=o\left(\left(x-x_0\right)^k\right)$$

Si $f^{\left(n\right)}\left(x_0\right)$ est représentable comme une combinaison des opérations de machine élémentaires

OK, mais de combien de termes a-t-on besoin

Formule de Taylor-Lagrange

Si f est de classe C^n sur $\left[x_0,x\right]$ et admet une dérivée d'ordre n+1, alors il existe $\xi\in\left]x_0,x\right[$ tel que

$$e\left(x,x_0\right)=\frac{f^{\left(n+1\right)}\left(\zeta\right)}{\left(n+1\right)!}\left(x-x_0\right)^{n+1}$$

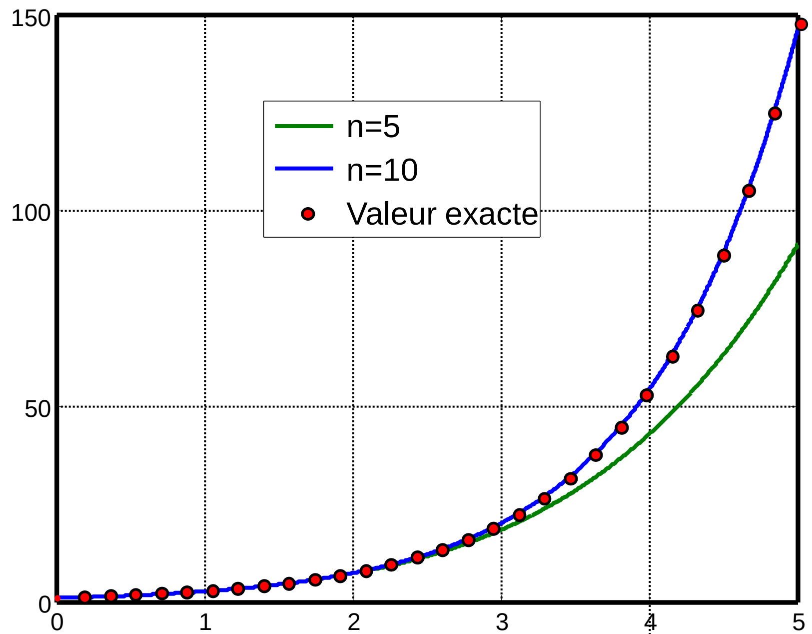
L'erreur dépend de la "*variabilité*" de la fonction et de la distance entre x et le centre de la série x_0

Exemple 1: fonction exponentielle

$f^{(k)}(x) = e^x \quad \forall k \geq 0$ Si $x_0 = 0$ on a $f(x) = f^{(k)}(x) = 1$

$$f(x) = \sum_{k=0}^{n} \frac{x^k}{k!} + err(x, x_0) \qquad err(x, x_0) = \frac{e^{\zeta} x^{n+1}}{(n+1)!} \leq \frac{e^{x_{\max}} x^{n+1}}{(n+1)!} \qquad \lim_{n\to\infty} e(x, x_0) = 0$$

On appel α l'erreur relative maximale désirée, on détermine n par la relation: $\frac{e^x x^{n+1}}{(n+1)!} \leq \alpha e^x$



$$\alpha \geq \frac{x^{n+1}}{(n+1)!}$$

Exemple 1: fonction exponentielle

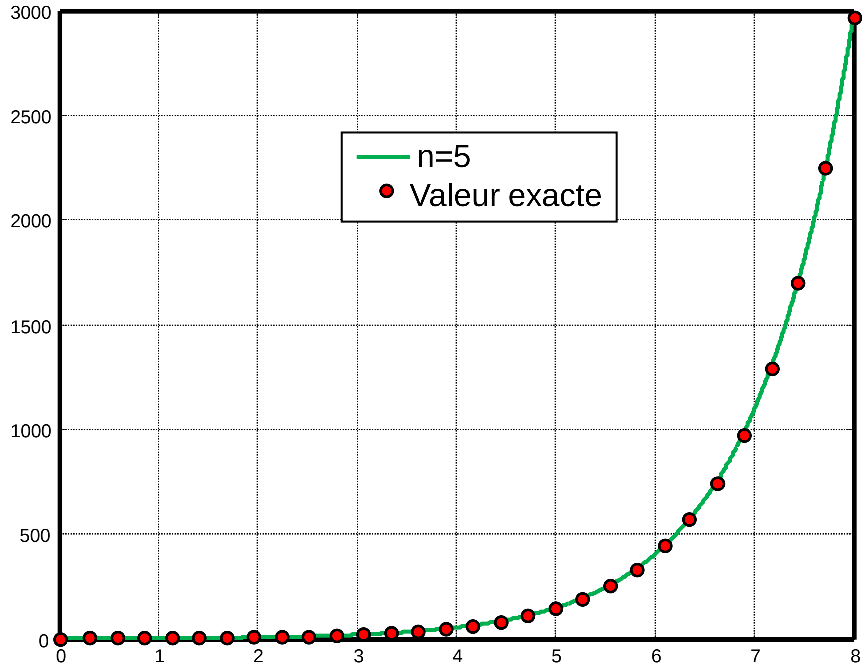
$$f(x) = e^{x} \qquad x = 12.3451 = 12 + 0.3451 = x_I + x_F \qquad f(x) = e^{x_I + x_F} = e^{x_I} \cdot e^{x_F}$$

- $$e^{x_I} = \prod_{i=1}^{x_I} e$$
- $$e^{x_F} = \sum_{k=0}^{n} \frac{x^k}{k!} + er(x_F), \quad \text{avec } x_F \leq 1$$

$$err(x_F) = \frac{e^{\zeta} x_F^{\,n+1}}{(n+1)!} \leq \frac{1}{(n+1)!} = err(1)$$

$$\Rightarrow \quad \alpha \geq \frac{1}{(n+1)!}$$

$$n = 5 \Rightarrow \alpha = 0.0083$$



Exemple 2: fonction sinus

Si $x_0 = 0$ on a

$$f(x) = \sin(x)$$

$$f^{(n)}(x) = \begin{cases} (-1)^{\frac{n-1}{2}} \cos(x) & \text{n impair} \\ (-1)^{\frac{n}{2}} \sin(x) & \text{n pair} \end{cases}$$

$$f(a) = 0$$

$$f^{(n)}(a) = \begin{cases} (-1)^{\frac{n-1}{2}} & \text{n impair} \\ 0 & \text{n pair} \end{cases}$$

$$f(x) = \sin(x) = x - \frac{x^3}{6} + \frac{x^5}{5!} \ldots + \frac{(-1)^n x^{2n+1}}{(2n+1)!} + o\left(x^{2n+2}\right) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!} + err(x, x_0)$$

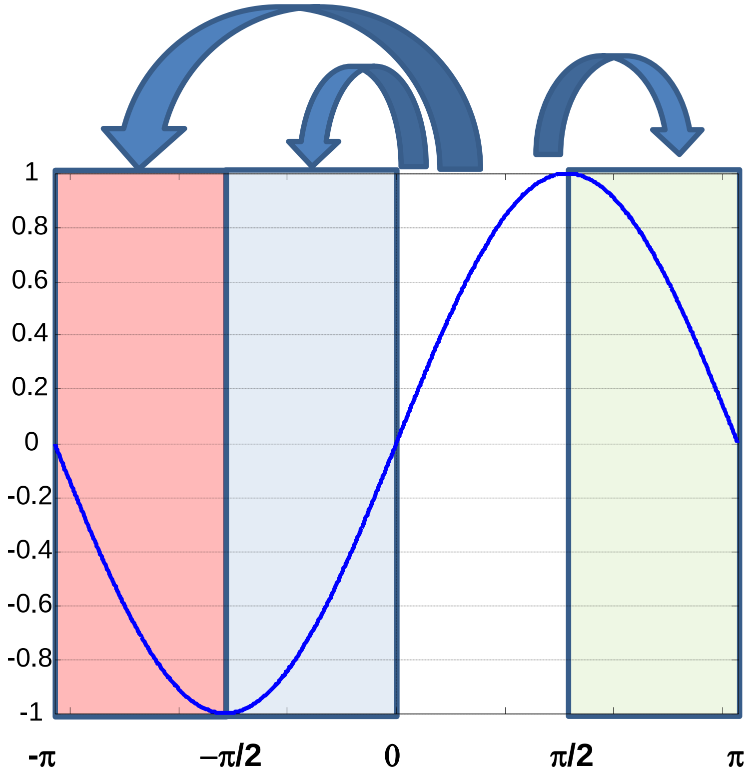
$$|e(x,0)| = \left| \frac{f^{(n+1)}(\zeta)}{(n+1)!} x^{n+1} \right| \leq \frac{\left| f^{(n+1)}(\zeta) \right|}{(n+1)!} \left| x^{n+1} \right| \leq \frac{\left| x^{n+1} \right|}{(n+1)!}$$

Exemple 2: fonction sinus

$$f(x) = \sin(x) = \sin(x+2\pi) = f(x+2\pi)$$ *fonction périodique de période* 2π

$$f(-x) = \sin(-x) = -\sin(x) = -f(x)$$ *fonction impaire*

$$f(x) = f(\pi - x) \quad \text{pour} \quad 0 < x \le \frac{\pi}{2}$$



Rappels de notation

x Donnée exacte	$\bar{x} = x + \Delta x$ Donnée perturbée
$\bar{y} = f(\bar{x}) = f(x + \Delta x)$	Fonction exacte calculée en précision infinie
$y_1 = \hat{f}(\bar{x}) = \hat{f}(x + \Delta x)$	Fonction approchée calculée en précision infinie
$y^* = fl\left(\hat{f}(\bar{x})\right) = fl\left(\hat{f}(x + \Delta x)\right)$	Fonction approchée calculée en précision finie

Soit f une fonction et $\hat{f}$ sa version approchée.

- Un algorithme est dit **stable** si pour tout

$$\frac{\|\bar{y} - y^*\|}{\|\bar{y}\|} = \frac{\left\|\tilde{f}(\bar{x}) - f(\bar{x})\right\|}{\left\|f(\bar{x})\right\|} = O(\varepsilon) \qquad \text{avec} \qquad \frac{\|x - \bar{x}\|}{\|x\|} = O(\varepsilon)$$

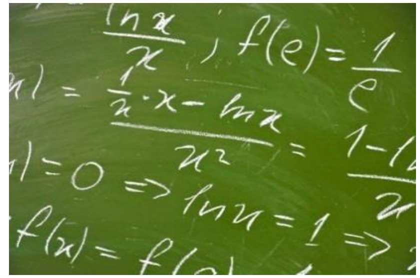
Un problème est stable, si pour une donnée $\tilde{x}$ pas très loin de x on obtient une solution $\tilde{f}(\bar{x})$ pas très loin de $f(\bar{x})$.

Plan du cours

- Représentation d'un nombre en machine, erreurs numériques, etc…
  - Introduction générale
  - Représentation des nombres entiers et réels
  - Opérations élémentaires en virgule flottante
  - Conditionnement d'un problème
  - Opérations complexes
  - Conclusion: différentes sources d'erreur
- Interpolation polynomiale
- Intégration numérique
- Résolution numérique des équations différentielles ordinaires (EDO)

Conclusion

Bureautique, audio-vidéo,…

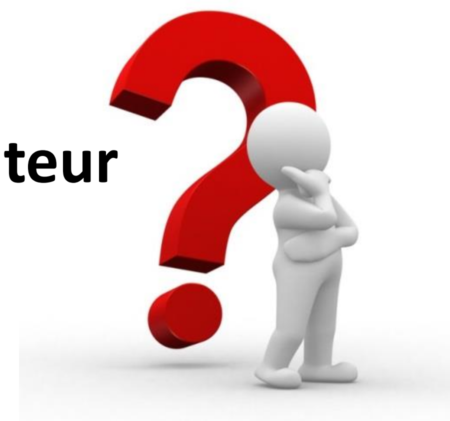


Calcul scientifique, Simulation, modélisation,…

Systèmes embarqués

les ordinateurs effectuent des **calculs**

Puis-je faire confiance aux résultats de mon ordinateur



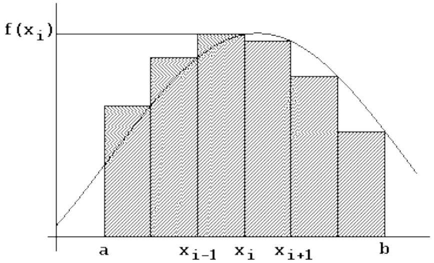
Conclusion: différentes sources d'erreur

- **Erreurs sur les données**: imprécision des mesures physiques, etc.

On peut étudier l'influence de ces erreurs sur le résultat final (conditionnement, etc.)

- **Erreurs de méthode**: elles sont dues à l'algorithme utilisé.

(approximation d'une intégrale, d'une somme infinie, …)



- **Erreurs de calcul en machine**: elles sont liées à l'arrondi de calcul pour les nombre flottants.

Conclusion: différentes sources d'erreur

Rappels de notation

x Donnée exacte	$\bar{x} = x + \Delta x$ Donnée perturbée
$y = f(x)$ $\bar{y} = f(\bar{x}) = f(x + \Delta x)$	Fonction exacte calculée en précision infinie
$y_1 = \hat{f}(\bar{x}) = \hat{f}(x + \Delta x)$	Fonction approchée calculée en précision infinie
$y^* = fl\left(\hat{f}(\bar{x})\right) = fl\left(\hat{f}(x + \Delta x)\right)$	Fonction approchée calculée en précision finie

Erreur absolue

$$\left|y - y^*\right| = \left|y - y^* + \bar{y}_1 - \bar{y}_1 + \bar{y} - \bar{y}\right|$$

$$\le \underbrace{\left|y - \bar{y}\right|}_{\text{Conditionnement du problème}} + \underbrace{\left|\bar{y} - \bar{y}_1\right|}_{\text{Erreur d'approximation de la fonction}} + \underbrace{\left|\bar{y}_1 - y^*\right|}_{\text{erreur de l'algorithme (arithmétique en précision finie)}}$$

Conditionnement du problème

Erreur d 'approximation de la fonction

erreur de l'algorithme (arithmétique en précision finie)