Chapitre 1 : Représentation en virgule flottante, stabilité et conditionnement

Accidents dus aux erreurs numériques

Missile Patriote :

- système américain d'interception des missiles ;
- pendant la guerre du Golf (1991), les erreurs d'arrondi dans l'estimation du temps de 0.34 secondes dans un des systèmes anti-missile ont causé une erreur sur la position du missile irakien d'environ un demi-kilomètre ;
- cela a couté la vie à 28 soldats ;

Fusée Ariane 5 :

- premier lancement (1996), 30 secondes après le décollage la fusée devient incontrôlable et est détruite ;
- la perte de contrôle est due au dépassement de la valeur maximale du registre qui contenait la vitesse horizontale ;
- coût : environ 500 millions de dollars.

Plus d'information sur `http://ta.twi.tudelft.nl/users/vuik/wi211/disasters.html`

Représentation des nombres réels

Dans la représentation en virgule flottante, les nombres réels ont la forme :

$$x \;=\; \pm\,\overline{0.d_1 d_2 \cdots d_t}\cdot\beta^e \;=\; \pm\beta^e \sum_{i=1}^{t} \frac{d_i}{\beta^i}$$

où

- β est la base ($\beta = 2$ – représentation binaire, $\beta = 10$ – décimale) ;
- t est le nombre de chiffres significatifs ;
- d_i est le $i$ème chiffre significatif ($0 \le d_i \le \beta - 1$) ;

  l'ensemble des chiffres significatifs $\overline{d_1 d_2 \cdots d_t}$ forment la mantisse ;
- e est l'exposant.

Exemples : avec $t = 3$ chiffres significatifs

- 2 est $\overline{0.200}\cdot 10^1$ en décimale (mais aussi $\overline{0.020}\cdot 10^2$ et $\overline{0.002}\cdot 10^3$)
  - vérifiez : $10^1(\frac{2}{10} + \frac{0}{10^2} + \frac{0}{10^3}) \;=\; 2$
- $1/2$ est $\overline{0.500}\cdot 10^0$ en décimale et $\overline{0.100}\cdot 2^0$ en binaire
  - vérifiez : $10^0(\frac{5}{10} + \frac{0}{10^2} + \frac{0}{10^3}) \;=\; \frac{1}{2} \;=\; 2^0(\frac{1}{2} + \frac{0}{2^2} + \frac{0}{2^3})$

Représentation des nombres réels (suite)

Dans la représentation en virgule flottante, les nombres réels ont la forme :

$$x = \pm \overline{0.d_1 d_2 \cdots d_t} \cdot \beta^e = \pm \beta^e \sum_{i=1}^{t} \frac{d_i}{\beta^i}$$

Certains réels ont de multiples représentations (ex : $\overline{0.200} \cdot 10^1$ et $\overline{0.020} \cdot 10^2$).
La représentation avec

$$d_1 \neq 0$$

est normalisée. Dans une base binaire cette représentation implique $d_1 = 1$.
L'ensemble des réels possédant une représentation normalisée est noté

$$\mathbb{F} = \left\{ x \mid x = \pm \overline{0.d_1 d_2 \cdots d_t} \cdot \beta^e \text{ , } e \in [e_{\min}, e_{\max}] \right\}.$$

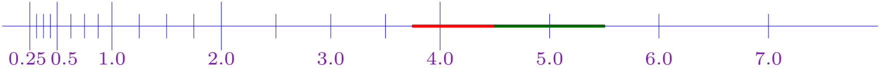
Exemple : Représentation de la partie positive de $\mathbb{F}$ pour $\beta = 2$, $t = 3$, $e_{\min} = -1$ et $e_{\max} = 3$. Notez que la distance entre deux réels consécutifs ne dépasse pas 2^{-2} fois leur valeur absolue.

Erreurs d'arrondi : Exemple

Comme l'ensemble $\mathbb{F}$ des réels représentables est fini alors que $\mathbb{R}$ est infini, les erreurs d'arrondi sont inévitables. Pour comprendre leurs effets, définissons

$$\mathrm{fl}(x) \;=\; \text{le réel dans } \mathbb{F} \text{ le plus proche de } x \in \mathbb{R}\,.$$

Exemple (suite) : $\mathrm{fl}(x) \;= 4.0$ dans la région rouge et $\mathrm{fl}(x) \;= 5.0$ dans la région verte



- La différence absolue $|\mathrm{fl}(x) - x|$ peut être d'autant plus importante que x est grand ; elle vaut au plus la moitié de la distance entre deux éléments de $\mathbb{F}$ qui entourent x .

  Pour cet exemple cela donne

  - pour $x \in]4, 7[$ on a $|\mathrm{fl}(x) - x| \;\leq\; 0.5$
  - pour $x \in]2, 4[$ on a $|\mathrm{fl}(x) - x| \;\leq\; 0.25$

    $\ldots$
  - pour $x \in]0.25, 0.5[$ on a $|\mathrm{fl}(x) - x| \;\leq\; 2^{-5}$

- On constate par contre que différence relative est bornée ; pour cet exemple

$$\frac{|\mathrm{fl}(x) - x|}{|x|} \;\leq\; \frac{1}{8}\,.$$

Erreurs d'arrondi : cas général

Analyse : Considérons un $x \in \mathbb{R}$ tel que $|x|$ se trouve entre les valeurs positives minimale et maximale de $\mathbb{F}$. Soit

$$x \;=\; \pm\, \overline{0.d_1 d_2 \cdots} \cdot \beta^e$$

son expansion (potentiellement infinie) normalisée ($d_1 \neq 0$) en base β ; on a en particulier

$$|x| \;\geq\; \overline{0.1} \cdot \beta^e \;=\; \beta^{e-1}\,. \tag{1}$$

Par ailleurs, la différence $|\mathrm{fl}(x) - x|$ vaut au plus la moitié de la distance entre deux éléments consécutifs x_+, x_- de $\mathbb{F}$ qui entourent x, et donc

$$|\mathrm{fl}(x) - x| \;\leq\; \frac{1}{2} \cdot |x_+ - x_-| \;=\; \frac{1}{2} \cdot \overline{0.00\cdots 1} \cdot \beta^e \;=\; \frac{1}{2}\, \beta^{e-t}\,. \tag{2}$$

Inégalités (1) et (2) donnent une relation importante

$$\frac{|\mathrm{fl}(x) - x|}{|x|} \;\leq\; \frac{\frac{1}{2}\beta^{e-t}}{\beta^{e-1}} \;=\; \frac{1}{2}\, \beta^{1-t} \;=:\; u\,, \tag{3}$$

où u est l'unité d'arrondi.

Erreurs d'arrondi : cas général (suite)

On a donc l'erreur relative

$$\frac{|\mathrm{fl}(x) - x|}{|x|} \leq u\,. \tag{3}$$

Rappel sur les erreurs : Soient x et son approximation $\hat{x}$. Alors

- l'erreur absolue est $\epsilon_{\mathrm{abs}} = |x - \hat{x}|$;
- l'erreur relative (pour $x \neq 0$) est $\epsilon_{\mathrm{rel}} = \dfrac{|x - \hat{x}|}{|x|}$; en particulier

$$\hat{x} = x(1+\epsilon)\,, \quad |\epsilon| = \epsilon_{\mathrm{rel}}$$

L'inégalité (3) est donc équivalente à

$$\boxed{\mathrm{fl}(x) = x(1+\epsilon)\,, \quad |\epsilon| \leq u\,.} \tag{4}$$

De manière similaire à (4) on montre (cf. travaux pratiques) que

$$\mathrm{fl}(x) = \frac{x}{1+\epsilon}\,, \quad |\epsilon| \leq u\,. \tag{5}$$

Standard IEEE 754 (1985)

Standard IEEE :

- universellement accepté aujourd'hui
- deux principaux formats en virgule flottante ($\beta = 2$) :
  `single` (simple précision) et `double` (double précision)
- représentation en mémoire

signe	exposant	mantisse

- spécifications principales :

`single`

1 bit	8 bits	23 bits

$$e_{\min} = -125$$
$$e_{\max} = 128$$
$$x_{\min} = 1.2 \cdot 10^{-38}$$
$$x_{\max} = 3.4 \cdot 10^{38}$$
$$u = \mathbf{6.0 \cdot 10^{-8}}$$

`double`

1 bit	11 bits	52 bits

$$e_{\min} = -1021$$
$$e_{\max} = 1024$$
$$x_{\min} = 2.2 \cdot 10^{-308}$$
$$x_{\max} = 1.8 \cdot 10^{308}$$
$$u = \mathbf{1.1 \cdot 10^{-16}}$$

- double précision est utilisée par défaut en Octave ;
  regardez les commandes `realmax`, `realmin` et `eps` ($= 2u$).

Modèle d'arithmétique

Modèle standard d'arithmétique en virgule flottante (satisfaite avec IEEE) :

soient

- $x,\ y \in \mathbb{F}$
- $\circ \ = \ +\,,\, -\,,\, \cdot\,,\, /$ les opérations habituelles dans $\mathbb{R}$;
  notez que $x \circ y$ n'est pas nécessairement dans $\mathbb{F}$
- $x \circ y \in [-x_{\max}, -x_{\min}] \cup [x_{\min}, x_{\max}]$
- $\circledcirc \ = \ \oplus, \ominus, \odot, \oslash$ les opérations en virgule flottante dans $\mathbb{F}$;
  avec donc $x \circledcirc y \in \mathbb{F}$

alors

$$\boxed{x \circledcirc y \ = \ \mathrm{fl}(x \circ y)\,.} \tag{6}$$

Interprétation : L'opération $\circledcirc$ est effectuée en arithmétique exacte, son résultat est ensuite converti dans $\mathbb{F}$

Ce modèle forme la base pour comprendre et prédire les effets des erreurs d'arrondi !

Analyse d'erreurs : règles de base

Modèle standard d'arithmétique pour $x\ , y \in \mathbb{F}$:

$$\boxed{x \circledcirc y \;=\; \mathrm{fl}(x \circ y)\,.} \tag{6}$$

Règles de base :

Pour l'ensemble des règles, on suppose $|\epsilon|, |\epsilon'|, |\epsilon_i|, |\epsilon_i'| \;\leq\; u$ et α, $\beta \;\in\; \mathbb{R}$

1. $x \circledcirc y \;=\; (x \circ y)(1+\epsilon)$

   (conséquence de la relation importante (4) pour les erreurs d'arrondi)

2. $x \circledcirc y \;=\; \dfrac{x \circ y}{1+\epsilon'}$

   (conséquence de la variante (5) de la relation importante)

3. $\alpha\epsilon_1 \pm \beta\epsilon_2 \;=\; (|\alpha| + |\beta|)\epsilon_3$

4. $(1+\alpha\epsilon_1)(1+\beta\epsilon_2) \;=\; 1 + (|\alpha| + |\beta|)\epsilon_3 + \mathcal{O}(u^2)$

5. $\dfrac{1}{1+\alpha\epsilon+\mathcal{O}(u^2)} \;=\; 1 + \alpha\epsilon' + \mathcal{O}(u^2)$

   (par développement en série de Taylor de $1/(1+x)$, avec $\epsilon' \;=\; -\epsilon$)

Modèle d'arithmétique : exemple 1

Exemple 1 : Estimer l'erreur d'arrondi sur le résultat de

$$(1 \oslash 3) \odot 3$$

Analyse : (tous les ϵ_i satisfont $|\epsilon_i| \leq u$)

$$\begin{aligned}
(1 \oslash 3) \odot 3 &= 1/3(1+\epsilon_1) \odot 3 && \text{(par ❶)} \\
&= (1/3(1+\epsilon_1) \cdot 3)\,(1+\epsilon_2) && \text{(par ❶)} \\
&= 1 \cdot (1+\epsilon_1)(1+\epsilon_2) \\
&= 1 \cdot (1+2\epsilon_3) + \mathcal{O}(u^2) && \text{(par ❹)}
\end{aligned}$$

Conclusion : erreur relative en double précision est au plus $2u = 2.2 \cdot 10^{-16}$.

Modèle d'arithmétique : exemple 2

Exemple 2 : soient x, y contaminés avec des erreurs d'arrondis :

- $\widetilde{x} = x(1+\epsilon_1) \in \mathbb{F}\,, \qquad |\epsilon_1| \;\leq\; u\,,$
- $\widetilde{y} = y(1+\epsilon_2) \in \mathbb{F}\,, \qquad |\epsilon_2| \;\leq\; u\,.$

Quels sont les erreurs d'arrondis de $\widetilde{x} \ominus \widetilde{y}$ comme approximation de $x - y$?

Analyse :(avec comme avant $|\epsilon_i| \leq u$)

- $\widetilde{x} \ominus \widetilde{y} \;=\; (\widetilde{x} - \widetilde{y})(1+\epsilon_3)$ (par 1)
- $\widetilde{x} - \widetilde{y} \;=\; (x-y) + (x\epsilon_1 - y\epsilon_2) \;=\; (x-y)\,(1+\delta u)$ avec

$$|\delta| \;=\; \left|\frac{x\epsilon_1 - y\epsilon_2}{x-y}\right| \cdot \frac{1}{u} \;\leq\; \frac{|x|+|y|}{|x-y|}$$

- et donc

$$\widetilde{x} \ominus \widetilde{y} \;=\; (x-y)(1+\delta u+\epsilon_3) + \mathcal{O}(u^2)$$

Si $x \approx y$, on a potentiellement $\delta \gg 1$ et donc un risque de perte de précision ; le phénomène est connu sous le nom d'annulation.

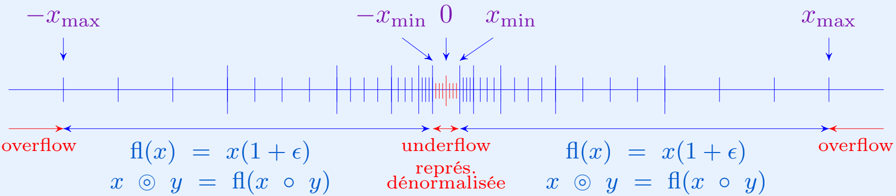
Note : c'est le phénomène qui s'est produit pour le calcul de la variance dans Exemple 6 du chapitre Motivation.

Standard IEEE : compléments et résumé

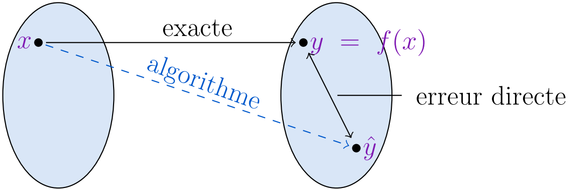
- Pour remédier à un changement brusque entre $x_{\min}$ et 0 on utilise une représentation dénormalisée si $|x| < x_{\min}$ (pour 0 y compris) ; le dépassement de $x_{\min}$ porte le nom d'underflow.
  En particulier, les relations (4), (6) ne sont plus valables et l'erreur relative peut (facilement) dépasser u.

- La représentation est aussi complétée avec
  - $\pm\infty$: provient du dépassement de $x_{\max}$ ou $1/0$; c'est un overflow ;
  - `NaN` : résulte de $0/0$, $0 \cdot \infty$ ou (dans certains cas, mais pas en Octave) $\sqrt{-1}$

Résumé (avec $t = 3$, $e_{\min} = -1$, $e_{\max} = 3$) :



Erreur directe



Soit un problème dont la solution (exacte) y est une fonction f des données x :

$$y = f(x) .$$

Attention : x, y peuvent être des scalaires, vecteurs, matrices, etc...

Exemples :

- évaluation d'une racine carrée : $f(x) = \sqrt{x}$
- soustraction de deux nombres : $f(x_1, x_2) = x_1 - x_2$

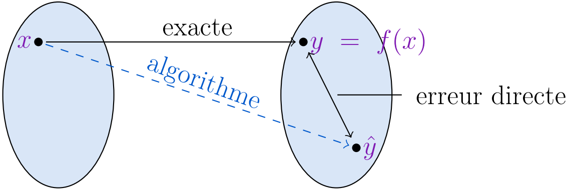
En arithmétique finie, un algorithme ne peut fournir qu'une solution approchée $\hat{y}$ (qui dépend de cet algorithme !)

La différence $\hat{y} - y$ est appelée erreur directe.

Intuition : Si cette erreur est petite (en norme), l'algorithme est stable.

Mais que veut dire «petite» ?

Stabilité directe



Mais que veut dire une «petite» erreur $\|\hat{y} - y\|/\|y\|$?

On considère qu'un algorithme a la stabilité directe en x si la norme de l'erreur directe $\|\hat{y} - y\| = \|\hat{y} - f(x)\|$ est comparable à la norme de l'erreur due aux effets d'arrondi.

En d'autres termes, si il existe $C_1, C_2 \geq 1$ (petits) tels que

$$\|\hat{y} - y\| \leq C_1 \|f(x + \delta x) - f(x)\|$$

pour au moins un δx tel que $||\delta x||/||x|| \leq C_2\, u$.

Motivation :

- souvent x est déjà entaché d'erreurs d'arrondi (au moins) ;
- les premières opérations sur chaque élément de x introduisent des erreurs relatives au moins aussi grandes que u (cf. Exemple 1)

Conditionnement

Un algorithme a la stabilité directe en x s'il existe C_1, $C_2 \geq 1$ tels que

$$\|\hat{y} - y\| \leq C_1 \|f(x + \delta x) - f(x)\|$$

pour au moins un δx tel que $\|\delta x\| / \|x\| \leq C_2\, u$.

Conditionnement

Un algorithme a la stabilité directe en x s'il existe $C_1, C_2 \geq 1$ tels que

$$\frac{\|\hat{y} - y\|}{\|y\|} \leq C_1 \frac{\|f(x + \delta x) - f(x)\|}{\|f(x)\|}$$

pour au moins un δx tel que $\|\delta x\| / \|x\| \leq C_2 u$.

Où on a utilisé $y = f(x)$.

Conditionnement

Un algorithme a la stabilité directe en x s'il existe $C_1, C_2 \geq 1$ tels que

$$\frac{\|\hat{y} - y\|}{\|y\|} \leq C_1 \frac{\|f(x+\delta x) - f(x)\|}{\|f(x)\|}$$

pour au moins un δx tel que $\|\delta x\| / \|x\| \leq C_2 u$.

En particulier, nous nous intéresserons à

$$\frac{\|f(x+\delta x) - f(x)\|}{\|f(x)\|} = \underbrace{\frac{\|f(x+\delta x) - f(x)\|\|x\|}{\|f(x)\|\|\delta x\|}}_{\text{facteur d'amplification}} \cdot \underbrace{\frac{\|\delta x\|}{\|x\|}}_{\leq C_2 u}$$

Le pire des facteur pour $\|\delta x\|$ petit, à savoir

$$\boxed{\kappa(x) := \lim_{\epsilon \to 0} \sup_{\|\delta x\| \leq \epsilon} \frac{\|f(x+\delta x) - f(x)\| \,/\, \|f(x)\|}{\|\delta x\| \,/\, \|x\|}}$$

est le conditionnement. En d'autres termes, le conditionnement est le pire des facteurs par lequel il faut multiplier les erreurs relatives dans les données x pour obtenir les erreurs relatives dans $f(x)$ (avec erreurs $\to 0$).

Conditionnement (suite)

Un algorithme a la stabilité directe en x s'il existe un C_1, $C_2 \geq 1$ tel que

$$\frac{\|\hat{y} - y\|}{\|y\|} \leq C_1 \frac{\|f(x + \delta x) - f(x)\|}{\|f(x)\|}$$

pour tout $\|\delta x\|/\|x\| \leq C_2\, u\,.$

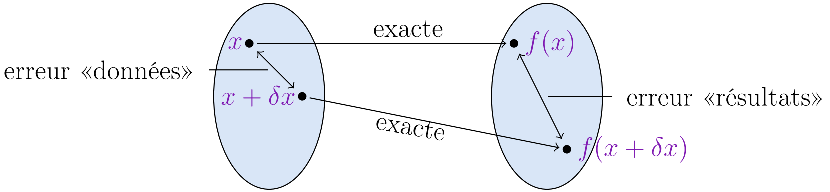
Pour revenir à la stabilité, on note (dans la limite de u infinitésimal) que

$$\frac{\|f(x + \delta x) - f(x)\|}{\|f(x)\|} \leq C_2 \kappa(x) u\,.$$

Par conséquent, un algorithme possédant la stabilité directe satisfait aussi

$$\boxed{\frac{\|\hat{y} - y\|}{\|y\|} \leq C_1 C_2 \kappa(x) u}$$

Conditionnement (suite)



$$\kappa(x) := \lim_{\epsilon \to 0} \sup_{\|\delta x\| \leq \epsilon} \frac{\|f(x + \delta x) - f(x)\| \;/\; \|f(x)\|}{\|\delta x\| \;/\; \|x\|} = \sup \frac{\text{erreur relative résultat}}{\text{erreur relative données}}$$

Commentaires :

- le conditionnement ne dépend pas d'un algorithme particulier (c.a.d $\hat{y}$) ; il ne dépend que du problème considéré (via $y = f(x)$).
- si $\kappa(x) \gg 1$ on parle d'un problème mal conditionné ; dans le cas contraire, il est bien conditionné.
- si $f(x)$ est différentiable (et $f'(x)$ est la matrice Jacobienne), on a

$$\kappa(x) = \frac{\|f'(x)\| \|x\|}{\|f(x)\|}$$

Conditionnement : exemples

$$\boxed{\kappa(x) \; = \; \frac{\|f'(x)\| \|x\|}{\|f(x)\|}}$$

Exemple 3 : Conditionnement de l'opération racine carrée $\sqrt{x}$ (pour $x > 0$).

$$f(x) \; = \; \sqrt{x}$$

et donc

$$\kappa(x) \; = \; \frac{|1/(2\sqrt{x})| \; |x|}{|\sqrt{x}|} \; = \; \frac{1}{2}\,.$$

Conclusion : C'est un problème bien conditionné ; un algorithme qui a la stabilité directe doit fournir une solution presque sans perte de précision.

Exemple 4 : Conditionnement de la soustraction $x_1 - x_2$.

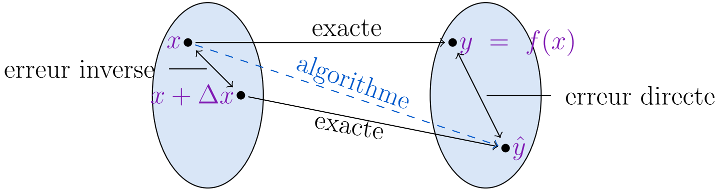
$$f(x_1, x_2) \; = \; x_1 - x_2\,, \qquad f'(x_1, x_2) \; = \; (1, -1)$$

et donc (en utilisant la norme euclidienne pour $\| \; \|$)

$$\kappa(x) \; = \; \frac{\sqrt{2}\;\sqrt{x_1^2 + x_2^2}}{|x_1 - x_2|}\,.$$

Conclusion : Risque de perte de précision si $x_1 \approx x_2$. (on le savait déjà !)

Erreur inverse, stabilité inverse



- L'erreur inverse d'un algorithme $\hat{y}$ est un Δx tel que

$$f(x + \Delta x) = \hat{y}\,; \tag{7}$$

(Un tel Δx n'existe pas nécessairement !)

- Un algorithme a la stabilité inverse si Δx satisfaisant (7) existe toujours et satisfait (pour un $C \geq 1$ petit)

$$\frac{\|\Delta x\|}{\|x\|} \leq Cu \tag{8}$$

- Comme la stabilité inverse implique

$$\|\hat{y} - y\| = \|f(x + \Delta x) - f(x)\|$$

pour un Δx de l'ordre de grandeur des erreurs d'arrondi (car (8)), elle implique la stabilité directe (avec $C_1 = 1$ et $C_2 = C$) .