Chapitre I :

Analyse des erreurs.

I- Analyse des erreurs
II- Le standard IEEE 754
III- Propagation des erreurs
IV- Erreurs direct et inverse

(I)

* $\mathbb{R} \rightarrow$ infini
* Ordinateur utilise un nbre fini de bits pour représenter un nombre $\Rightarrow$ on ne représente qu'un nbre fini de nbres.

Soit $x \in \mathbb{R}$; On note $fl(x)$ sa représentat° machine.

* Dans un ordinateur, les nbres st stockés ss forme normalisée :

$$x, xxxxxx \times 10^{k}.$$

De plus on utilise le codage en forme flottante binaire : on normalise un nbre comme suit

$$1, xxxxxx\ldots \times 10^{k}$$

et on ne représente que sa partie décimale.
On ne stocke pas le chiffre le plus significatif: on parle de bit caché.

$$\underbrace{1,}_{\text{bit caché}}\ \underbrace{xxxx}_{\substack{\text{mantisse} \\ \text{taille } p}}\ldots \times \underbrace{\beta^{k}}_{\text{base}}$$

* on appelle précision eps d'un ordinateur le plus petit nbre

$$1 + eps \neq 1$$

* Les différents types d'erreurs :

1°) erreur de représentation :

$$\Delta x = fl(x) - x$$

2°) erreur relative de représentation :

$$\varepsilon(x) = \frac{fl(x) - x}{fl(x)}$$

3°) erreur relative de précision

$$\eta(x) = \frac{fl(x) - x}{x}$$

* Lorsque la machine manque de précision pour représenter un nombre, elle doit choisir par quel nombre l'approximer. Il existe deux stratégies : 1* troncature (x_-)
2* arrondi (plus proche entre x_+ / x_-)

Théorème de la précision relative :
Base β ; mantisse de taille p.

$$|\eta(x)| \leq \begin{cases} \beta^{1-p} & \text{si troncature} \\ \frac{1}{2} \cdot \beta^{1-p} & \text{si arrondi} \end{cases}$$

II La norme IEEE-754.

Un nbre est représenté sous la forme suivante :

X XXXXXXXX XXXXXXXXXXXXXXXXXXXXXXX

signe : S — exposant : E (q bits) — mantisse = M (p bits)

$$\Rightarrow \quad (-1)^S \times 1,M \times 2^{E - \text{biais}}$$

$$\text{avec} \quad \text{biais} = 2^{q-1} - 1$$

Ainsi, il n'y a pas de bit de signe pour l'exposant grâce au biais.

En simple précision, on obtient les résultats suivants :

Bit de signe + bit caché	Taille p Mantisse (↓ 23)	Taille q Exposant (↓ 8)	
			Plus Petit Nb normalisé : 3,4 E+38
			Plus Grand Nb normalisé : 1,18 E-38
			Valeur Min Denormalisé : 1,4 E-45.

$$\Rightarrow \text{eps} = 2^{-23} \sim 1,192 \cdot 10^{-7}$$

Il y a quatre types de nb flottants binaires :

1°) des nbres normalisés.

2°) des nbres ss-normalisés

3°) des nbres infinis

4°) des NaN (Not a Number)

Rq : Les exposants 0000 0000 et 1111 1111 sont à usage réservé.
↳ l'exposant est dc compris entre
$E_{min} = -\text{biais} + 1$ et $E_{max} = \text{biais}$.

- Les infinis st de la forme suivante :

0 1111 1111 0000 0000 0000 0000 0000 000 = $+\infty$
1 1111 1111 0000 0000 0000 0000 0000 000 = $-\infty$.

- Les NaN st de la forme suivante :
$E = 1111\ 1111$
$M \neq \{0\}^{23}$

- Les nbres ss normalisés (pas garantie des résultats de calcul) :
$E = 0000\ 0000$; $M \neq \{0\}^{23}$

- Il y a deux zéros :
$S = \{^{1}_{0}$ $E = 0000\ 0000$; $M = 0000\ 0000\ 0000\ 0000\ 0000\ 000$
$\Rightarrow \pm 0$.

III

* Sources d'erreurs lors de la modélisat° puis simulat° d'un pb. physique par ordinateur :

1°) Modélisat° mathématique du système physique
2°) Approximation des fct° analytiques du modèle mathématique
3°) Discrétisation des fct° obtenues par approximation des fct° analytiques
4°) Calcul numérique par ordinateur

* Voici les erreurs obtenues par une opérat° $\otimes$

$$a \otimes b \rightarrow fl(fl(a) \otimes fl(b))$$

On décompose cette erreur en deux parties :

améliorable → 1°) Une erreur d'entrée $\eta^I(a \otimes b)$

rien à faire → 2°) Une erreur de calcul $\eta^C(a \otimes b)$

* Erreur des sommes : $S = x_1 + x_2 \dots$

1°) Erreur absolue $\Delta^I S = \Delta^I x_1 + \Delta^I x_2 \dots$

2°) Erreur de précision $\eta^I(S) = \frac{x_1}{S}\eta^I(x_1) + \frac{x_2}{S}\eta^I(x_2)$

Démo : $m(S) = S + \Delta^I S + \Delta^C S$

$m(S) = (x_1 + \Delta^I x_1) + (x_2 + \Delta^I x_2) \dots + \Delta^C S$

$\Rightarrow \Delta^I S = \Delta^I x_1 + \Delta^I x_2 \dots$

or : $\Delta x = \eta(x) * x$

$\Rightarrow S \cdot \eta^I(S) = x_1 \eta^I(x_1) + x_2 \eta_2^I(x_2) \dots$

Borne Max : $|\eta^I(S)| \leq \frac{\sum |x_i|}{|\sum x_i|} * eps$

* Erreur du produit : $P = x_1 * x_2 * x_3 \ldots$

avec $x_i \neq 0 \ \forall i$

1°) erreur absolue : $\Delta^I P = P * \left(\frac{\Delta^I x_1}{x_1} + \frac{\Delta^I x_2}{x_2} \ldots \right)$

2°) erreur précision : $\eta^I(P) = \eta^I(x_1) + \eta^I(x_2) \ldots$

Démo : $m(P) = P + \Delta^I P + \Delta^C P$

$= m(x_1) * m(x_2) \ldots + \Delta^C P$

$= x_1 x_2 x_3 \ldots + \Delta^I x_1 (x_2 x_3 \ldots)$

$+ \Delta^I x_2 (x_1 x_3 \ldots) \ldots + \Delta^I x_1 \Delta^I x_2 (x_3 \ldots)$

$\ldots + \Delta^I x_1 \Delta^I x_2 \Delta^I x_3 \ldots + \Delta^C P$

On néglige les termes qui ont plus de deux facteurs d'erreur absolue ($\Delta^I x_1 \Delta^I x_2 (x_3 \ldots)$)

$\Rightarrow \Delta^I P = \Delta x_1 (x_2 x_3 \ldots) + \Delta x_2 (x_1 x_3 \ldots)$

$= P \left(\frac{\Delta x_1}{x_1} + \frac{\Delta x_2}{x_2} \ldots \right)$

De plus $\Delta^I x = x \, \eta(x) \Rightarrow \eta^I(P) = \sum \eta^I(x_i)$

Borne Max : $|\eta^I(P)| \leq N * eps$

* erreur de la division : $Q = \frac{a}{b}$; $b \neq 0$

1°) erreur absolue : $\Delta^I Q = \frac{b \Delta^I a - a \Delta^I b}{b^2}$

2°) erreur précision : $\eta^I(Q) = \eta^I(a) - \eta^I(b)$

Démo : $m(Q) = Q + \Delta^I Q + \Delta^C Q = \frac{m(a)}{m(b)} + \Delta^C Q$

$\Rightarrow m(Q) = \frac{(a + \Delta^I a)}{(b + \Delta^I b)} + \Delta^C Q = \frac{(a + \Delta^I a)(b - \Delta^I b)}{b^2 - \Delta^I b^2} + \Delta^C Q$

On néglige comme ci-dessus : $Q + \Delta^I Q = \frac{b \Delta^I a - a \Delta^I b + ab}{b^2}$

De plus : $\Delta^I x = x . \eta(x) \Rightarrow \eta^I(Q) = \eta^I(a) - \eta^I(b)$.

Borne Max : $|\eta^I(Q)| \leq 2 \, eps$

Récapitulatif

- addition / soustraction : $\eta^I(a \pm b) = \frac{a}{a \pm b}\eta^I(a) \pm \frac{b}{a \pm b}\eta^I(b)$
- multiplication : $\eta^I(a \times b) = \eta^I(a) + \eta^I(b)$
- division : $\eta^I\left(\frac{a}{b}\right) = \eta^I(a) - \eta^I(b)$
- racine carrée : $\eta^I(\sqrt{a}) = \frac{1}{2}\eta^I(a)$.

* Le nbre condit° d'une fct° décrit la sensibilité d'une fct° à des variations de la valeur d'entrée :

$$\kappa(x) = \max_{V(x')} = \frac{\left|\frac{f(x)-f(x')}{f(x)}\right|}{\left|\frac{x-x'}{x}\right|}$$

Si $|x - x'|$ très petit :

$$\kappa(x) \sim \left|\frac{f'(x)}{f(x)} \cdot x\right|$$

Lorsque $\kappa(x)$ est grand, on dit que la fct° est mal conditionnée.

* La stabilité d'un algorithme exprime le fait que pour des petites variations des valeurs d'entrées, les valeurs de sorties présentent des petites variations :
  ⇒ il faut décomposer l'algo. en opérat° élémentaires, calculer le nb. condit° pour chaque opérat°, et regarder la propagat° des erreurs (composit° des nbres conditions)

* L'erreur du résultat $Y = [y_1 \dots y_m]^T$ d'un algorithme avec comme entrées $X = [x_1 \dots x_m]^T$ est :

$$\Delta Y = \begin{bmatrix} \Delta y_1 \\ \vdots \\ \Delta y_m \end{bmatrix} \simeq \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1} & \dots & \frac{\partial \phi_1}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial \phi_m}{\partial x_1} & \dots & \frac{\partial \phi_m}{\partial x_m} \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \vdots \\ \Delta x_m \end{bmatrix} = J[\phi(x)] \cdot \Delta x$$

où $J[\phi(x)]$ est le Jacobien de ϕ.

* L'erreur relative du résultat est :

$$\eta(y) = \begin{bmatrix} \eta(y_1) \\ \vdots \\ \eta(y_m) \end{bmatrix} = \begin{bmatrix} K(\phi_1, x_1) & \dots & K(\phi_1, x_m) \\ \vdots & & \\ K(\phi_m, x_1) & \dots & K(\phi_m, x_m) \end{bmatrix} \begin{bmatrix} \eta(x_1) \\ \vdots \\ \eta(x_m) \end{bmatrix} = K(\phi, x) * \eta(x)$$

où $K(\phi_i, x_j)$ le nbre-condit relatif à la fct° ϕ_i relatif au point x_j.

* Déroulement d'un algo :

$$x = x^{(0)} \to x^{(1)} = \phi^1(x^{(0)}) \to x^{(2)} = \phi^2(x^{(1)}) \dots x^{(n)} = \phi^{(n)}(x^{(n-1)}) = y$$

Rappel : $J(f \circ g) = J(f(g(x))) * J(g(x))$

$$\Rightarrow J(\phi(x)) = J(\phi^n(x^{(n-1)})) * J(\phi^{n-1}(x^{(n-2)})) \dots J(\phi^{(1)}(x^{(0)}))$$

* Etude de propagat° d'erreur : $x^{(k+1)} = \phi^{(k+1)}(x^{(k)})$

erreur absolue $\Rightarrow \Delta x^{(k+1)} = fl(x^{(k+1)}) - x^{(k+1)}$

$$= fl(\phi^{(k+1)}[fl(x^{(k)})]) - \phi^{(k+1)}(x^{(k)})$$

$$= \underbrace{[fl(\phi^{(k+1)}[fl(x^{(k)})]) - \phi^{(k+1)}(fl(x^{(k)}))]}_{\phi^{(k+1)}[fl(x^{(k)})] \cdot (I + H_{k+1}) \,/\, \phi^{(k+1)}[fl(x^{(k)})] \cdot I} + \underbrace{[\phi^{(k+1)}(fl(x^{(k)})) - \phi^{(k+1)}(x^{(k)})]}_{J(\phi^{(k+1)}(x^{(k)})) \cdot \Delta x^{(k)}}$$

$H_{k+1} \cdot \phi^{(k+1)}[fl(x^{(k)})]$

$$\simeq J(\phi^{(k+1)}(x^{(k)})) \cdot \Delta x^{(k)} + H_{k+1} \cdot x^{k+1} \quad (**)$$

* $fl(\phi_i^{k+1}(fl(x^{(k)})))$
$= \phi_i^{k+1}(fl(x^{(k)}))(1 + \eta_i(fl(x^{(k)})))$

avec $H_{k+1} = \text{diag}(\eta_1(fl(x^{(k)})), \eta_2(fl(x^{(k)})) \dots)$

** $H_{k+1} \cdot \phi^{k+1}[fl(x^{(k)})] \simeq H_{k+1} \cdot \phi^{k+1}[x^{(k)}]$
$= H_{k+1} \cdot x^{k+1}$

* On en déduit l'erreur d'un algo :

$$\Delta y = J(\phi(x)) \cdot \Delta x + J\varphi^{1}(x^{(1)}) \cdot H_1 \cdot x^{(1)} + \dots + J\varphi^{n-1}(x^{n-1}) \cdot H_{n-1} \cdot x^{(n-1)} + H_n y$$

avec $\varphi^k = \phi^{(k)} \circ \phi^{(k-1)} \dots \phi^{(k+1)}$

* Pour deux algo A, A' qui font le même résultat, on introduit la notion d'erreur E_n afin de les comparer :

$$E_n(A, x) = \Delta y - J(\phi(x)) \cdot \Delta x$$
$$= J(\varphi^{1}(x^{1})) \cdot H_1 \cdot x^{(1)} + \dots + H_n \cdot y$$

$E_n(A, x) \leq E_n(A', x) \Leftrightarrow A$ plus crédible que A'

(IV) Supposons que nous essayions de résoudre $f(x)$ et que nous obtenions $\tilde{f}(y)$. IP y a donc une erreur $\Delta y = \tilde{f}(y) - y$. Deux hypothèses peuvent alors se poser :

1°) l'écart entre y et $\tilde{f}(y)$ est l'erreur de calcul.

2°) $\tilde{f}(y) = f(\tilde{f}(x))$, c-à-d que les calculs sont exacts mais que l'entrée était perturbée.
A partir de $\tilde{f}(y)$ on trouve $\tilde{f}(x)$ et on en déduit Δx.
Cette erreur sur les données en entrée ($|\Delta x|$) est appelée erreur inverse.

erreur directe ($|\Delta y|$) $\leq$ erreur inverse * nb. condit° de f