

INFOF-205 : Calcul formel et numérique

R. Petit

Année académique 2015 - 2016

Contents

1	Introduction	1
2	Analyse et mesure d'erreurs	6
2.1	Définitions	6
2.2	les erreurs numériques	6
2.3	Représentation interne des nombres par l'ordinateur	7
2.3.1	Représentation en virgule fixe	7
2.3.2	Représentation en virgule flottante	8
2.3.3	Distance relative dans \mathbb{F}	9
2.4	Erreurs de propagation et de génération	10

1 Introduction

Définition 1.1. Le *calcul numérique* est une discipline qui traite de la définition, l'analyse, et l'implémentation d'algorithmes pour la résolution numérique des problèmes mathématiques qui proviennent de la modélisation des phénomènes réels.¹

Le calcul numérique représente donc le triplet (modèle analytique, solution théorique, résolution algorithmique).

Remarque. Le calcul numérique étant une application informatique pour la résolution de problèmes mathématiques continus, plusieurs types de problèmes peuvent arriver (section suivante). Afin de limiter les problèmes de précision sur des fonctions continues compliquées à évaluer précisément informatiquement, plusieurs outils d'approximation peuvent être utilisés. Le théorème de Taylor permet d'approcher certaines fonctions de manière plus facilement calculatoires de manière informatique.

Théorème 1.2 (Théorème de Taylor). Soient $f : I \rightarrow \mathbb{R}$ où I est un intervalle, $a \in I$ et $k \geq 1$. Alors $T(f, a, k)(x)$ est le seul polynôme de degré inférieur ou égal à k tel que :

$$\lim_{x \rightarrow a} \frac{f(x) - T(f, a, k)}{(x - a)^k} = 0,$$

où $T(f, a, k)(x)$, appelé polynôme de Taylor est défini par :

$$T(f, a, k)(x) = \sum_{i=0}^k \frac{f^{(i)}(a)(x - a)^i}{i!}.$$

Le calcul numérique est défini comme étant une application informatique à des problèmes relatifs à des phénomènes réels. Bien souvent, ces phénomènes sont modélisés de manière *continue*. Les ordinateurs quant à eux sont *discrets*. Cette différence majeure force des approximations par le matériel informatique. Les transformations discrètes (*discrétisations*) des phénomènes réels doivent alors être rigoureusement analysés afin de déterminer leur robustesse et leur fiabilité.

Les problèmes peuvent être classifiés de plusieurs manières selon leurs propriétés et ce qui en est attendu.

Définition 1.3. Un problème est dit :

- *qualitatif* lorsque le comportement des solutions est étudié afin d'en déterminer sa stabilité, son comportement asymptotique ;
- *quantitatif* lorsque la solution (numérique) précise est étudiée afin de tirer des conclusions spécifiques.

Un problème qualitatif cherche des conclusions globales à un problème (étude d'une famille de problèmes) et un problème quantitatif est une application précise appliquée à un problème donné.

Définition 1.4. Un problème est dit sous forme :

- *explicite* quand la solution x d'un problème est une fonction donnée des données d ;
- *implicite* quand la solution ne peut être extraite explicitement des données.

¹Nick Trefehen, Oxford.

Exemple 1. Trouver la racine carrée d'un nombre $d \in \mathbb{R}$ est un problème explicite que l'on peut écrire $x = f(d) = \sqrt{d}$.

Déterminer l'ensemble des solutions x à l'équation $ax^2 + bx + c$ est un problème implicite car la solution n'est pas séparée explicitement des données. On peut tout de même exprimer ce problème implicite comme :

$$X = \left\{ \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \right\},$$

qui est un problème explicite.

Remarque. La forme générale d'un problème explicite est $x = F(d)$, alors que la forme générale d'un problème implicite est $F(x, d) = 0$.

Définition 1.5. Le problème explicite $x = F(d)$ est dit *bien posé* si la solution x :

1. existe ;
2. est unique ;
3. dépend **continument** de d .

Si un problème n'est pas bien posé, on dit qu'il est *mal posé*.

Définition 1.6. Un problème est dit *problème inverse* s'il correspond à la détermination des données d'entrée x d'une application G en connaissant les données de sortie $G(x) = d$.

Remarque. Les problèmes inverses sont de manière générale mal posés, entre autre parce qu'il est fréquent que l'application G ne soit pas injective et donc que plusieurs données x peuvent donner la même donnée $d = G(x)$.

Remarque. Ce cours ne traitera que de problèmes bien posés.

Définition 1.7.

- Le *conditionnement* d'un problème $x = F(d)$ bien posé est la mesure de la *sensibilité* de x à des petits changements sur d .
- On définit par δd une faible perturbation (également dite *perturbation admissible*) sur la donnée d . On définit alors la perturbation induite sur x par $\delta x = F(d + \delta d) - F(d)$.
- Le *conditionnement relatif* du problème est la quantité :

$$\kappa(d) := \lim_{|D| \rightarrow 0} \sup_{\delta d \in D} \frac{\|\delta x\| \|d\|}{\|x\| \|\delta d\|} = \lim_{|D| \rightarrow 0} \sup_{\delta d \in D} \frac{\|F(d + \delta d) - F(d)\| \|d\|}{\|F(d)\| \|\delta d\|},$$

où D est un voisinage de l'origine, et \sup désigne la borne supérieure.

Remarque.

- si $x = 0$, alors le conditionnement relatif est forcément infini ;
- si $d = 0$, alors le conditionnement relatif est forcément nul ;
- quand $0 \in \{x, d\}$, il faut définir une autre notion, le conditionnement absolu :

$$\kappa_{\text{abs}}(d) := \lim_{|D| \rightarrow 0} \sup_{\delta d \in D} \frac{\|\delta x\|}{\|\delta d\|}.$$

Proposition 1.8. Si $x = F(d)$ est un problème explicite, et F est une fonction différentiable en d , alors, par définition :

$$\lim_{\delta d \rightarrow 0} \frac{\delta x}{\delta d}$$

ne dépend pas de la trajectoire de δd vers 0. On note alors :

$$F'(d) = \lim_{\delta d \rightarrow 0} \frac{\delta x}{\delta d},$$

et on réécrit le conditionnement par :

$$\kappa(d) = \lim_{\delta d \rightarrow 0} \frac{\|\delta x\|}{\|\delta d\|} \frac{\|d\|}{\|x\|} = \frac{\|d\|}{\|x\|} \lim_{\delta d \rightarrow 0} \frac{\|\delta x\|}{\|\delta d\|} = \|F'(d)\| \frac{\|d\|}{\|x\|}.$$

Remarque. Si $d \in \mathbb{R}^m$, et $x \in \mathbb{R}^n$, alors $F'(d)$ est la matrice jacobienne $(\text{Jac } F)(d)$. Et si $q = 1$, la matrice jacobienne n'a qu'une seule ligne et on parle du gradient $(\nabla F)(d)$.

Exemple 2 (Conditionnement d'une différence). Soit le problème $x = F(d) = d - a$. La fonction F est différentiable, on peut donc déterminer le conditionnement :

$$\kappa(d) = \|F'(d)\| \frac{\|d\|}{\|x\|} = 1 \frac{\|d\|}{\|d - a\|}.$$

On voit alors que le conditionnement est petit pour $\|d - a\| \gg 0$. Cependant, pour d proche de a , le conditionnement devient très grand. Le problème est donc *mal conditionné* pour d proche de a .

Exemple 3 (Conditionnement d'une exponentiation). Soit le problème $x = F(d) = d^r$ avec $r \in \mathbb{R}$. À nouveau, F est différentiable, et donc on peut exprimer le conditionnement comme :

$$\kappa(d) = \|F'(d)\| \frac{\|d\|}{\|x\|} = \|r\| \|d^{r-1}\| \frac{\|d\|}{\|d^r\|} = \|r\|.$$

Le conditionnement dépend donc de l'exposant r . Si r est grand, le problème est mal conditionné, peu importe la valeur de d .

Le conditionnement d'un problème est une caractéristique qui lui est intrinsèque. Si un problème est mal conditionné, peu importe l'algorithme utilisé pour le résoudre, la solution restera fortement influencé par des perturbations sur les données d . De plus, plus un problème a un grand conditionnement, plus il sera difficile à résoudre numériquement.

Remarque. les exemples ci-dessus montrent qu'un problème peut avoir un conditionnement faible pour certaines valeurs de d et un conditionnement élevé pour d'autres valeurs.

Soit $F(x, d) = g(x) - d = 0$, un problème implicite. Prenons $g : \mathbb{R} \rightarrow \mathbb{R}$ une fonction réelle différentiable. On peut alors exprimer $x = g^{-1}(d)$. Donc le conditionnement vaut :

$$\kappa(d) = \left| \frac{(g^{-1})'(d)d}{g^{-1}(d)} \right| = \left| \frac{1}{g'(g^{-1}(d))} \right| \left| \frac{d}{g^{-1}(d)} \right| = \left| \frac{1}{g'(x)} \right| \left| \frac{d}{x} \right|.$$

On remarque donc que pour $g'(x)$ petit le conditionnement devient grand. C'est intuitivement lié au fait qu'une fonction à faible dérivée (pente) en d va croître lentement et donc pour une faible perturbation sur d (axe vertical), une grande perturbation sur x (axe horizontal) va être induite.

Définition 1.9. Soit $F(x, d) = 0$ un problème bien posé. On définit l'*algorithme numérique* pour la résolution du problème F par la suite de problèmes approchés $F_1(x^{(1)}, d^{(1)}), F_2(x^{(2)}, d^{(2)}), \dots, F_n(x^{(n)}, d^{(n)})$ dépendant d'un paramètre n .

L'idée derrière la division en sous-problèmes est d'avoir des F_k plus simples à résoudre que F , ce qui permet une résolution numérique.

Définition 1.10. Un algorithme numérique est dit :

- *direct* si le nombre n de sous-problèmes est fixé (du moins majoré), habituellement par une fonction de la taille du problème ;
- *itératif* si n n'est pas borné et s'il existe une routine $f(u)$ admissible, indépendante de n telle que $x^{(n)} = f(x^{(n-1)})$.

Définition 1.11. Un algorithme numérique $\{F_i(x^{(i)}, d^{(i)})\}_{1 \leq i \leq n}$ est consistant si :

$$\lim_{n \rightarrow +\infty} F_n(x^{(n)}, d^{(n)}) = F(x, d^{(n)}),$$

et donc si x , la solution précise, est une solution ses sous-problèmes pour $n \rightarrow +\infty$.

Remarque. Un algorithme itératif $x^{(n)} = f(x^{(n-1)})$ est consistant si $x^{(n)} = x$ implique que $x^{(n+1)} = x$. Ce qui revient à dire qu'une fois la solution exacte atteinte, l'algorithme itératif la conserve et ne diverge pas.

Définition 1.12. Un algorithme $F_n(x^{(n)}, d^{(n)})$ est dit *fortement* consistant si :

$$\forall n \geq 1 : F_n(x, d^{(n)}) = 0,$$

c'est-à-dire si x est une solution admissible de tous les sous-problèmes $F_n(x^{(n)}, d^{(n)})$.

Intéressons-nous au problème $F(x, d)$ et à $F_n(x^{(n)}, d^{(n)})$, un algorithme numérique de résolution du problème. Supposons que cet algorithme numérique soit composé d'étapes telles que :

$$x^{(n)} = F_n^{(m)}(d_{m-1}^{(n)}),$$

où :

$$d_k^{(n)} = F_n^{(k)}(d_{k-1}^{(n)}) \quad \text{et} \quad d_0^{(n)} = d^{(n)}.$$

On considère que la résolution numérique est l'application de cet algorithme sur des valeurs perturbées :

$$\begin{cases} \bar{x}^{(n)} &= F_m^{(n)}(\bar{d}_{m-1}^{(n)}) + \delta d_m^{(n)} \\ \bar{d}_k^{(n)} &= F_k^{(n)}(\bar{d}_{k-1}^{(n)}) + \delta d_k^{(n)} \\ \bar{d}_0^{(n)} &= d^{(n)} \end{cases}$$

Remarque. On voit bien que l'algorithme à l'étape n commence avec $\bar{d}_0^{(n)} = d^{(n)}$ qui est une donnée initiale non perturbée.

Définition 1.13. Soit $F_n(x^{(n)}, d^{(n)})$ un algorithme numérique. On dit qu'il est *stable* (ou *bien posé*) si :

- $\forall n : \exists ! x^{(n)}, \text{ solution ;}$
-

$$\forall \epsilon > 0 : \exists \eta > 0, n_0 \in \mathbb{N}^* \text{ t. q. } \forall n > n_0 : \left(\forall i \in \{1, \dots, m\} : \|\delta d_i^{(n)}\| < \eta \right) \Rightarrow \left(\|\bar{x}^{(n)} - x^{(n)}\| < \epsilon \right).$$

Définition 1.14. Si l'opération $F_m^{(n)}(d)$ ne dépend ni de m , ni de n , on parle d'algorithme *itératif*.

Remarque. Un algorithme est donc dit itératif si c'est constamment la même opération qui est appliquée sur des données consécutives. On a alors :

$$x^{(n)} = f(x^{(n-1)}) = (f \circ f)(x^{(n-2)}) = \dots = (f^n)(x^{(0)}) = (f^n)(d).$$

On peut traduire la définition de stabilité de manière plus informelle par le fait qu'un algorithme est stable s'il existe un voisinage de x , la solution tel que pour tout x_1, x_2 dans ce voisinage, on a un $R \in (0, 1)$ tel que :

$$\|f(x_2) - f(x_1)\| \leq R \|x_2 - x_1\|.$$

Si f est une fonction différentiable, alors on peut dire que :

$$\|f'(x_2)\| \leq R.$$

2 Analyse et mesure d'erreurs

2.1 Définitions

Définition 2.1. \mathbb{F} est l'ensemble des nombres représentés exactement par un ordinateur.

Remarque. On remarque aisément que $\mathbb{F} \subset \mathbb{R}$ car $\mathbb{F} \not\subset \mathbb{C} \setminus \mathbb{R}$. On peut tout de même raffiner car il semble évident que $\mathbb{F} \subset \mathbb{Q}$: un nombre irrationnel n'a pas de représentation physique finie, et ne peut donc pas être représentés par un ordinateur.

Définition 2.2. Soit x un nombre dans \mathbb{Q} . Une approximation de x est une valeur \hat{x} légèrement différente de x et qui remplace x dans les calculs effectués. On note également $x \simeq \hat{x}$ pour montrer que \hat{x} est une approximation de x .

Si $\hat{x} < x$, on parle d'approximation par défaut, et si $\hat{x} > x$, on parle d'approximation par excès.

Définition 2.3. On définit l'*erreur absolue* entre x et son approximation \hat{x} par :

$$\delta_x = |\hat{x} - x|.$$

On définit également l'*écart relatif* entre x et \hat{x} par :

$$\rho_x = \frac{\hat{x} - x}{x}.$$

On définit alors l'écart absolue défini comme $\epsilon_x = |\rho_x|$.

Remarque. L'écart relatif permet d'écrire l'approximation sous la forme $\hat{x} = x(1 + \rho_x)$ où ρ_x est censé être faible (car l'écart entre x et son approximation doit rester relativement faible). Cette notation est fortement utilisée.

Les écarts ne sont cependant définis qu'en $x \neq 0$, ce dont il faut tenir compte.

Définition 2.4. La *borne supérieure d'erreur relative* d'une approximation \hat{x} , notée u est nombre positif quelconque tel que $u \geq \epsilon_x$.

Mis à part les erreurs de calcul (qui consistent à déterminer un résultat erroné sur base de données bien définies) et les erreurs d'implémentation (qui consiste en une implémentation erronée amenant à des erreurs de calcul), il existe cinq catégories d'erreurs que l'on peut regrouper en deux familles :

- les erreurs de modèle qui consistent en simplifier les phénomènes en omettant des éléments afin de le rendre plus facile à étudier ;
- les erreurs numériques qui consistent en l'approximation due à l'observation et non à la résolution théorique.

2.2 Les erreurs numériques²

Définition 2.5. Les *erreurs de troncature* sont les erreurs liées aux processus théoriques mathématiques infinis (série infinie par exemple).

²Les erreurs de modélisation ne sont pas traitées ici.

Définition 2.6. Les *erreurs d'arrondi* sont les erreurs liées au système numérique de la machine venant du fait qu'un ordinateur ne peut représenter qu'un sous-ensemble fini \mathbb{F} des nombres.

Définition 2.7. Les *erreurs de génération et de propagation* sont les erreurs liées à l'évaluation d'une opération sur des opérandes pas exactes.

Remarque. Les erreurs de troncature sont compliquées à éviter : il est nécessaire d'avoir un procédé fini pour y échapper. Cette catégorie d'erreurs ne seront donc pas étudiées dans ce cours.

Les différents types d'erreurs concernent différentes propriétés des problèmes ou des algorithmes numériques utilisés.

Proposition 2.8.

- Les erreurs d'approximation ou de troncature se mesurent par la **consistance** de l'algorithme ;
- les erreurs de propagation se mesurent par le **conditionnement** de l'algorithme ;
- les erreurs de génération se mesurent par la **stabilité de l'algorithme** ;
- les erreurs d'arrondi sont liées à la représentation interne des nombres.

2.3 Représentation interne des nombres par l'ordinateur

Il existe deux manières de représenter les nombres décimaux par un ordinateur : la *virgule fixe* et la *virgule flottante*.

2.3.1 Représentation en virgule fixe

Soit $x \in \mathbb{F}$ un nombre. Sa représentation en virgule fixe est donnée par le triplet $(\{a_i\}_{-m \leq i \leq n}, b, s)$ où :

- $b \geq 2$ est la base de représentation ;
- $s \in \{0, 1\}$ est le signe ;
- $\{a_{-m}, \dots, a_n\}$ est l'ensemble de chiffres $0 \leq a_i \leq b-1$ pour tout i .

Remarque. m représente le nombre de chiffres suivant la virgule, et $n + 1$ caractérise le nombre de chiffres précédant la virgule. On en déduit que la virgule est (implicitement) placée entre a_0 et a_{-1} .

On peut donc exprimer x par :

$$x = (-1)^s \sum_{k=-m}^n a_k b^k.$$

Remarque. La base choisie pour les ordinateurs est souvent 2, 8 ou 16 car 2 représente le binaire, et les bases octale et hexadécimale représentent chacune une écriture *compactée* du binaire par paquets de 3 ou 4 bits.

La base 10 est par fois également choisie.

Proposition 2.9 (Propriétés des nombres en virgule fixe).

- Les nombres représentés par virgule fixe sont équirépartis sur la droite des réels ;
- l'écart entre deux nombres consécutifs représentés en virgule flottante est b^{-m} , qui est le plus petit nombre représentable (en valeur absolue) ;
- et la plus grande valeur représentable est $b^{n+1} - 1$, où $n + 1$ est le nombre de chiffres avant la virgule.

Remarque. La représentation par virgule flottante limite fortement les valeurs extrêmes que peut prendre un nombre de par son équirépartition. Les nombres à virgule fixe (comme vu juste après) ont une densité de répartition différente, ce qui leur permet de prendre des valeurs maximum et minimum bien plus haute (en terme d'ordre de grandeur).

2.3.2 Représentation en virgule flottante

Soit $x \in \mathbb{F}$ un nombre. Sa représentation en virgule flottante est donnée par le quadruplet $(\{a_i\}_{1 \leq i \leq t}, b, s, e)$ où :

- b et s représentent les mêmes quantités que pour la virgule fixe ;
- t est le nombre de chiffres significatifs ;
- $\{a_i\}_{1 \leq i \leq t}$ est l'ensemble des chiffres $0 \leq a_1 \leq b$ et $0 \leq a_i \leq b$ pour $i > 1$.

Remarque. On impose $a_1 \neq 0$ afin que la représentation d'un nombre soit unique.

Définition 2.10. La *mantisse* est la quantité définie par $\sum_{i=1}^t a_i b^{t-i}$. On la note $m \in \mathbb{N}$, ou encore $m[x] \in \mathbb{N}$ quand il est nécessaire de montrer que m est la mantisse de la valeur x .

Remarque. Dû à la normalisation des a_i , on sait borner m par :

$$b^{t-1} \leq m \leq b^t - 1.$$

On peut donc au final exprimer x comme :

$$x = (-1)^s b^e \sum_{i=1}^t a_i b^{-i} = (-1)^s b^e b^t b^{-t} \sum_{i=1}^t a_i b^{-i} = (-1)^s b^{e-t} \sum_{i=1}^t a_i b^{t-i} = (-1)^s b^{e-t} m.$$

Remarque. On note L et U les valeurs respectivement plus petite et plus grande que peut prendre la valeur e de l'exposant.

Proposition 2.11. On paramétrise l'ensemble \mathbb{F} par les quantités b , t , L , et U . Quand la paramétrisation doit être explicite, on note l'ensemble :

$$\mathbb{F}(b, t, L, U).$$

Remarque. Dans une représentation normalisée, $0 \notin \mathbb{F}$.

Théorème 2.12. Si x est un élément de $\mathbb{F}(b, t, L, U)$, alors on peut borner x tel que :

$$b^{L-1} \leq |x| \leq b^U (1 - b^t).$$

Proposition 2.13. Le cardinal de l'ensemble $\mathbb{F}(b, t, L, U)$ est donné par :

$$|\mathbb{F}(b, t, L, U)| = 2(b-1)b^{t-1}(U-L+1).$$

Remarque. Étant donné que l'exposant e est codé sur un nombre n_e de bits (et représenté en complément à 2) dans une représentation machine de représentation flottante, on peut trouver que $L = -2^{n_e-1} - 1 \leq e \leq 2^{n_e} = U$. On voit alors que les valeurs maximum prises par un nombre sont beaucoup plus grandes (en valeur absolue de l'ordre de grandeur) qu'en virgule fixe.

2.3.3 Distance relative dans \mathbb{F}

Prenons x_i et x_{i+1} dans \mathbb{F} . On note $\eta(x_i)$ la distance relative entre x_i et x_{i+1} que l'on définit :

$$\eta(x_i) : \left| \frac{x_{i+1} - x_i}{x_i} \right|.$$

Dans un système à virgule fixe, $\eta(x_i)$ est fixe pour tout i . Pour un système à virgule flottante, on trouve :

$$\eta(x_i) = \left| \frac{x_{i+1} - x_i}{x_i} \right| = \left| \frac{m[x_{i+1}]b^{e-t} - m[x_i]b^{e-t}}{m[x_i]b^{e-t}} \right| = \frac{1}{m[x_i]}.$$

Si x_{i+1} a un exposant plus grand que x_i , on a :

$$\eta(x_i) = \left| \frac{b^{t-1}b^{e-t+1} - (b^t - 1)b^{e-t}}{(b^t - 1)b^{e-t}} \right| = \left| \frac{b^{e-t}}{(b^t - 1)b^{e-t}} \right| = \frac{1}{m[x_i]}.$$

On peut donc, en virgule flottante, également déterminer une distance générale (formule) mais qui n'est pas constante.

En sachant que :

$$b^{t-1} \leq m < b^t,$$

on peut trouver :

$$b^{-t} < \frac{1}{m} = \eta \leq b^{1-t}.$$

En divisant de part et d'autre par b^{1-t} , on trouve :

$$\frac{\epsilon}{b} < \eta(x_i) \leq \epsilon.$$

Définition 2.14. On appelle $\epsilon := b^{1-t}$ l'*epsilon machine* qui représente la distance maximale entre deux nombres consécutifs en virgule flottante.

Remarque. L'*epsilon machine* permet de déterminer $1 + \epsilon$ qui est le plus petit nombre $x \in \mathbb{F}$ tel que $x \geq 1$.

Remarque. Il ne faut pas confondre ϵ qui est la distance maximale entre deux nombres et b^{L-1} qui est le plus petit nombre que l'on peut représenter. En pratique, $\epsilon \simeq 10^{-16}$ alors que $b^{L-1} \simeq 10^{-308}$.

Définition 2.15. Le phénomène d'oscillation de $\eta(x_i)$ est appelé *wobbling precision* en anglais.

Remarque. Le phénomène de *wobbling precision* est d'autant plus grand que la base b est grande. C'est entre autres pour cela que les petites bases ($b = 2$) sont utilisées en pratique.

Définition 2.16. Soit $x \in \mathbb{R} \setminus \mathbb{F}$ un réel. x n'a pas de représentation exacte par l'ordinateur puisque $x \notin \mathbb{F}$.

- Si $|x| > b^U(1 - b^{-t})$, alors x est plus grand que le plus grand nombre que l'on peut représenter. On parle alors d'*overflow*. En cas d'*overflow*, habituellement, le système interrompt le programme.
- Si $|x| < b^{L-1}$, alors x est plus petit que le plus petit nombre possible à représenter. On parle alors d'*underflow*. En cas d'*underflow*, habituellement, le nombre x est arrondi à 0.
- Si x ne dépasse pas des bornes de représentation, on considère a_{t+1} , le chiffre suivant a_t dans la représentation théorique à précision infinie de x . On peut alors déterminer \hat{x} de deux manières :
 - par arrondi ;

– par troncature.

Un arrondi se fait par l'application $x \mapsto \hat{x} = \text{fl}(x)$, qui remplace a_t par a_t^* défini par :

$$a_t^* = \begin{cases} a_t & \text{si } a_{t+1} < \frac{b}{2}, \\ a_t + 1 & \text{si } a_{t+1} \geq \frac{b}{2}. \end{cases}$$

Si $a_{t+1} \geq \frac{b}{2}$ et $a_t = b - 1$, on réitère le processus pour a_{t-1} selon a_t , etc.

Une troncature se fait par l'application $x \mapsto \hat{x} = \text{tr}(x)$, qui conserve uniquement les t premiers chiffres de la notation théorique.

Définition 2.17. On définit ϵ_x l'erreur relative de x par :

$$\epsilon_x = \left| \frac{\text{fl}(x) - x}{x} \right|.$$

On définit également la *précision machine* $u := \frac{\epsilon}{2}$.

Proposition 2.18. Soit $x \in \mathbb{R}$ et $\text{fl}(x) \in \mathbb{F}$. On a $\epsilon_x \leq \frac{\epsilon}{2}$.

Preuve. $|\text{fl}(x) - x|$ représente la distance entre x et son approximation. x est arrondi au nombre dans \mathbb{F} le plus proche, et donc $|\text{fl}(x) - x| \leq \frac{1}{2}|x_{i+1} - x_i|$ avec $x_i \leq x \leq x_{i+1}$. On trouve alors :

$$\epsilon_x = |\rho_x| \leq \frac{\epsilon}{2} = u.$$

□

Remarque. La précision machine u donne l'ordre de grandeur de la meilleure approximation possible pour un arrondi sur une machine donnée.

Remarque. On trouve alors :

$$\text{fl}(x) = x(1 + \rho_x).$$

2.4 Erreurs de propagation et de génération