MT09-Analyse numérique élémentaire

Chapitre 1 : Introduction au calcul flottant

Équipe de Mathématiques Appliquées

UTC

Septembre 2021

Chapitre 1
Introduction au calcul flottant

Ce chapitre vise à donner quelques connaissances de base sur le codage et le calcul avec des nombres flottants. En effet, les calculs numériques faits par un ordinateur contiennent toujours des erreurs, qui peuvent avoir plusieurs origines. Certaines erreurs proviennent du fait que les nombres réels sont codés de façon *approchée* et les calculs avec ces approximations de réels donnent des résultats également approchés. Les erreurs commises peuvent parfois être catastrophiques.

L'ordinateur donne donc rarement des résultats exacts, parfois les résultats sont complètement faux ! L'ingénieur doit en être conscient et doit avoir un œil critique sur les sorties d'un ordinateur.

Sommaire
Concepts

Exemples
Exercices
Documents

1.1 Nombres entiers sur ordinateur

Sommaire
Concepts

Exemples
Exercices
Documents

1.1.1 Nombres entiers (*integer*)

L'ensemble des nombres entiers mathématiques $\mathbb{Z}$ constitue un ensemble infini. Les ordinateurs disposent d'une *mémoire finie* : ils ne peuvent pas stocker tous les entiers. Sur un ordinateur, l'ensemble des entiers codés, que l'on note ici $\mathcal{Z}$, s'écrit sous la forme :

$$\mathcal{Z} = \{-m, -m+1, \ldots, -1, 0, 1, \ldots, M-1, M\} \subset \mathbb{Z}, \quad \text{card}(\mathcal{Z}) = M + m + 1 < +\infty,$$

où m et M sont des entiers positifs. On rappelle que le cardinal d'un ensemble fini est le nombre de ses éléments.

L'ensemble $\mathcal{Z}$ possède un plus petit entier (négatif) noté $-m$ et un plus grand entier (positif) noté M. Les valeurs de m et M dépendent de l'espace mémoire servant à coder l'entier. Pour les entiers codés sur 32 bits (les `int` standard en C ou C++ par exemple), m et M sont de l'ordre de $2^{31} = 2.(2^{10})^3 \approx 2.(10^3)^3 \approx 2.10^9$ (rappel : $2^{10} = 1024$.)

Si un résultat de calcul est trop petit ou trop grand, il ne peut pas être représenté dans $\mathcal{Z}$ et une erreur se produit. L'erreur est généralement appelée dépassement d'entier ou "integer overflow".

Important : chaque entier de $\mathbb{Z}$ compris entre $-m$ et M est représenté de manière *exacte*. Comme on va le voir, ce n'est pas le cas pour les réels.

Sommaire
Concepts

Exemples
Exercices
Documents

1.2 Nombres flottants

Les réels ne peuvent pas être programmés de façon exacte sur un ordinateur, car d'une part il existe une infinité de nombres réels, et d'autre part la plupart des réels s'écrivent avec une infinité de chiffres (par exemple : $1/3 = 0.333\ldots$ ou π). Un ordinateur possède une mémoire finie, il est impossible de stocker tous les réels, et pour un réel donné, il peut être impossible de stocker toutes ses décimales.

C'est pourquoi les réels sont *approchés* dans la plupart des ordinateurs par des *nombres à virgule flottante*, encore appelés *flottants*. Les calculs numériques sur ordinateur manipulent ces nombres flottants.

Sommaire
Concepts

Exemples
Exercices
Documents

1.2.1 Flottants en base 10

Exemples :
Exemple A.1.1

Les ordinateurs travaillent en base 2, mais nous présentons les concepts en base 10 pour simplifier. En base 10, l'ensemble $\mathscr{F}_{10}$ des flottants s'écrit :

$$\mathscr{F}_{10} = \left\{ \pm 0.d_1 d_2 \dots d_t \ 10^e \mid d_i \in \{0,1,2,\dots,9\} \ \forall i = 1,\dots,t, \quad d_1 \neq 0, \quad L \leq e \leq U \right\} \cup \{0\},$$

où $t > 0$ est le nombre de chiffres significatifs, L et U ($L \leq U$) constituent les bornes inférieure et supérieure de l'exposant e. Par convention, l'exposant e est choisi de façon que le premier chiffre d_1 soit toujours non-nul. Le nombre 0 est explicitement inséré dans $\mathscr{F}_{10}$ car 0 ne s'écrit pas comme un nombre flottant normal.

Voir l'exemple A.1.1.

Les entiers t, U et L définissent entièrement $\mathscr{F}_{10}$, qui est un ensemble fini, de cardinal

$$\text{card}(\mathscr{F}_{10}) = 2 \times 9 \times 10^{t-1}(U - L + 1) + 1 \quad < +\infty.$$

L'ensemble $\mathscr{F}_{10}$ (fini) est donc un ensemble discret, par opposition à $\mathbb{R}$ (infini) qui est dit "continu".

Sommaire
Concepts

Exemples
Exercices
Documents

Flottants en base 10

Le nombre $f = 0.d_1 d_2 \dots d_t$ s'appelle la mantisse et f vérifie :

$$f_{10}^{\min} = 0.10\dots0 = 10^{-1} \leq f \leq f_{10}^{\max} = 0.99\dots9 = 1 - 10^{-t} < 1.$$

La mantisse peut s'écrire comme la somme $f = \sum_{i=1}^{t} d_i 10^{-i}$.

Il faut remarquer que, contrairement à $\mathbb{R}$, $\mathscr{F}_{10}$ possède un plus grand flottant noté $F_{10}^{\max}$ et un plus petit flottant strictement positif noté $F_{10}^{\min}$:

$$\forall g \in \mathscr{F}_{10} \qquad 0 < F_{10}^{\min} = 0.10\dots0\ 10^{L} = 10^{L-1} \ \leq\ g\ \leq\ F_{10}^{\max} = 0.99\dots9\ 10^{U} \approx 10^{U}.$$

À part 0, il n'y a aucun flottant dans l'intervalle $]-10^{L-1}, 10^{L-1}[$: il y a un trou autour de 0.

1.2.2 Représentation des réels, écarts, erreur relative

Exemples :
Exemple A.1.2
Exemple A.1.3

Sur l'intervalle $[f_{10}^{\min}, 1]$, la différence entre deux flottants consécutifs est constant et vaut $\delta_0 = 10^{-t}$. On appelle ce nombre l'*écart absolu* par opposition à l'écart relatif qui va être déterminé plus bas. De même, pour un exposant e tel que $L \le e \le U$, l'écart absolu entre deux flottants consécutifs dans $[10^{e-1}, 10^e] = [f_{10}^{\min}, 1] \times 10^e$ est constant et vaut

$$\delta_e = 10^{-t+e}.$$

Voir l'exemple A.1.2.

Ceci implique que sur l'intervalle $[F_{10}^{\min}, F_{10}^{\max}]$, l'écart absolu entre deux flottants voisins est variable : les flottants sont plus denses pour des petites valeurs (positives) que pour les grandes valeurs (voir la figure de l'exemple A.1.4).

Cependant l'écart *relatif* noté δ_r reste majoré par une constante. En effet, soit un flottant $g \in [F_{10}^{\min}, F_{10}^{\max}]$. Il existe un unique exposant e dans $[L, U]$ et une unique mantisse f dans $[f_{10}^{\min}, f_{10}^{\max}]$ tels que $g = f\ 10^e$. Le flottant qui suit g est $g + \delta_e$. L'écart relatif entre

Sommaire
Concepts

Exemples
Exercices
Documents

deux flottants consécutifs s'écrit donc

$$\delta_r = \frac{|g+\delta_e - g|}{|g|} \leq \frac{\delta_e}{f_{10}^{\min}\ 10^e} = 10^{-t+1}.$$

Ce calcul permet de donner une majoration de l'*erreur relative* commise quand on approche un réel x par le flottant le plus proche de x, que l'on note $\mathrm{fl}(x)$. Soit donc un réel x dans l'intervalle $[F_{10}^{\min}, F_{10}^{\max}]$. Il existe un flottant $g = f\ 10^e$ de mantisse f et d'exposant e, tel que $g \leq x < g+\delta_e$. Le flottant $\mathrm{fl}(x)$ le plus proche de x sera soit g, soit $g+\delta_e$. Dans le cas où x est au milieu de l'intervalle $[g, g+\delta_e]$, une convention d'arrondi permet de choisir $\mathrm{fl}(x)$. Le réel x est éloigné de $\mathrm{fl}(x)$ au plus de $\frac{1}{2}\delta_e$ et $x \geq g \geq f_{10}^{\min}\ 10^e$. L'erreur relative entre x et le flottant $\mathrm{fl}(x)$ vérifie donc :

$$\frac{|x-\mathrm{fl}(x)|}{|x|} \leq \frac{\frac{1}{2}\delta_e}{f_{10}^{\min}\ 10^e} \leq \frac{1}{2}10^{-t+1} = \frac{\delta_1}{2} = \varepsilon_{\mathrm{mach},10}.$$

Cette précision relative constante ne dépend que du nombre de chiffres significatifs t. Le nombre $\varepsilon_{\mathrm{mach},10}$ appelé précision du codage de machine est caractéristique de $\mathscr{F}_{10}$: c'est la précision (relative) maximale que l'on peut attendre des calculs en flottants.

Voir l'exemple A.1.3.

1.2.3 Flottants en base 2

Exemples : Exemple A.1.4 **Documents** : Document B.1.1 **Exercices** : Exercice C.1.1

La base 2 est la base le plus souvent utilisée par un ordinateur. Le document B.1.1 précise quelques informations sur le codage des flottants respectant une norme internationale. Les flottants en base 2 se définissent de la même manière qu'en base 10,

$$\mathscr{F}_2 = \left\{ \pm 0.1 d_2 \dots d_t \; 2^e \mid d_i \in \{0,1\} \; \forall i = 2, \dots, t, \quad L \leq e \leq U \right\} \cup \{0\}.$$

Le premier chiffre d_1, qui est par convention non-nul, vaut donc toujours 1. L'ensemble fini $\mathscr{F}_2$ a pour cardinal :

$$\text{card}(\mathscr{F}_2) = 2^t (U - L + 1) + 1 \quad < +\infty.$$

La mantisse $f = 0.d_1 d_2 \dots d_t = \sum_{i=1}^{t} d_i 2^{-i}$ vérifie :

$$f_2^{\min} = 0.10 \dots 0 = 2^{-1} \leq f \leq f_2^{\max} = 0.11 \dots 1 = 1 - 2^{-t} < 1.$$

Voir l'exercice C.1.1 et l'exemple A.1.4.

L'ensemble $\mathscr{F}_2$ possède un plus grand flottant noté $F_2^{\max}$ et un plus petit flottant strictement positif noté $F_2^{\min}$:

$$\forall g \in \mathscr{F}_2 \qquad 0 < F_2^{\min} = 0.10 \dots 0 \; 2^L = 2^{L-1} \; \leq \; g \; \leq \; F_2^{\max} = 0.11 \dots 1 \; 2^U \approx 2^U.$$

Sommaire
Concepts

Exemples
Exercices
Documents

Les mêmes raisonnements que pour $\mathscr{F}_{10}$ permettent de déterminer les écarts absolus et relatifs, et l'erreur relative. Soit un flottant $g = f\ 2^e \in [F_2^{\min}, F_2^{\max}]$ (mantisse f dans $[f_2^{\min}, f_2^{\max}]$ et exposant e dans $[L,U]$). Les écarts absolus et relatifs entre g et son successeur vérifient

$$\delta_e = 2^{-t+e}, \qquad \delta_r = \frac{|g+\delta_e - g|}{|g|} \leq \frac{\delta_e}{f_2^{\min}\ 2^e} = 2^{-t+1}.$$

L'erreur relative entre un réel x appartenant à $[F_2^{\min}, F_2^{\max}]$ et le flottant de $\mathscr{F}_2$ le plus proche noté $\mathrm{fl}(x)$ vérifie :

$$\frac{|x-\mathrm{fl}(x)|}{|x|} \leq \frac{\frac{1}{2}2^{-t+e}}{f_2^{\min}\ 2^e} = 2^{-t} = \frac{\delta_1}{2} = \delta_0 = \varepsilon_{\mathrm{mach},2}.$$

1.3 Calcul avec les flottants

Sommaire
Concepts

Exemples
Exercices
Documents

1.3.1 Arithmétique flottante

Exemples :
Exemple A.1.5

Soit $\mathscr{F}$ un ensemble de flottants (la base peut être indifféremment 2 ou 10). Le résultat d'une opération sur deux flottants de $\mathscr{F}$ n'est en général *pas* un flottant de $\mathscr{F}$ (voir l'exemple A.1.5).

Les calculs en flottants respectant la norme`IEEE 754` doivent vérifier la règle suivante. On prend une opération arithmétique usuelle notée "$*$" (qui vaut "+", "−", "×" ou "/"). On note l'opération arithmétique flottante correspondante : "$\circledast$", qui n'est pas exacte.

Pour deux flottants $(x, y) \in \mathscr{F}$, le résultat de l'opération en arithmétique approchée $(x \circledast y)$ est l'arrondi de l'opération exacte $x * y$, ce qu'on écrit

$$(x \circledast y) = \mathrm{fl}(x * y).$$

Par exemple, pour l'addition, on a : $(x \oplus y) = \mathrm{fl}(x + y)$. Le résultat de l'addition approchée de deux flottants est le flottant le plus proche de l'addition exacte. La racine carrée vérifie également ce type de propriété.

Cette propriété permet de raisonner sur les calculs en flottants. En particulier, grâce aux résultats sur l'erreur relative, si $x * y \neq 0$, on a une majoration de l'erreur relative sur les opérations flottantes :

$$\frac{|(x \circledast y) - (x * y)|}{|x * y|} \leq \varepsilon_{\text{mach}}.$$

1.3.2 Addition flottante

Exemples :
Exemple A.1.6
Exemple A.1.7
Exemple A.1.8

Le principe (un peu simplifié) de l'addition flottante est le suivant. Soit x et y dans $\mathscr{F}$. On suppose que $x > y$ pour fixer les idées.

1. on décale la virgule dans y de façon que x et y aient le même exposant (on fait apparaître des 0 dans la mantisse de y),
2. on ajoute les mantisses (calcul exact).
3. on effectue un arrondi sur le résultat de façon à ne garder que t chiffres.

Voir l'exemple A.1.6.

Propriété de l'addition flottante :

L'arithmétique flottante est commutative, mais n'est *pas associative*. En général,

$$(x \oplus y) \oplus z \neq x \oplus (y \oplus z) \quad x, y, z \in \mathscr{F}.$$

Voir l'exemple A.1.7.

Sommaire
Concepts

Exemples
Exercices
Documents

Annulation destructrice :

Quand deux flottants ayant des ordres de grandeurs très différents sont ajoutés, une perte d'information est toujours réalisée : certains chiffres significatifs du plus petit nombre sont perdus. L'erreur relative qui est faite est de l'ordre de $\varepsilon_{\text{mach}}$. Il arrive que cette perte d'information s'avère catastrophique, quand dans la suite des calculs, on a justement besoin de cette information disparue : on parle alors d'annulation destructrice.

Voir l'exemple A.1.8.

Annexe A
Exemples

A.1 Exemples du chapitre 1

Sommaire
Concepts

Exemples
Exercices
Documents

Exemple A.1.1 Un ensemble flottant

Prenons $t = 3$ et $L = -1$ et $U = 2$. Les flottants strictement positifs sont donc :

$$\{0.100, 0.101, 0.102, \ldots, 0.998, 0.999\} \times 10^{-1}, \times 10^{0}, \times 10^{1}, \times 10^{2}.$$

retour au cours

Sommaire
Concepts

Exemples
Exercices
Documents

Exemple A.1.2 Exemple d'écart

On prend $t = 3$ et $e = 0$. Les flottants dans $[f_{10}^{\min}, 1]$ sont :

$$f_{10}^{\min}(= 0.100), 0.101, 0.102, 0.103, \ldots, 0.109, 0.110, 0.111, 0.112, \ldots, 0.998, f_{10}^{\max}(= 0.999), 1.$$

L'écart absolu entre deux flottants successifs vaut $0.001 = 10^{-t}$.

retour au cours

Sommaire
Concepts

Exemples
Exercices
Documents

Exemple A.1.3 Exemple d'erreur relative

On prend $t = 7$. L'approximation flottante du réel $x = 1789.0408$ s'écrit $\mathrm{fl}(x) = 0.1789041\ 10^4$, en arrondissant au plus proche. L'erreur absolue commise vaut $|x - \mathrm{fl}(x)| = 2\ 10^{-4}$, l'erreur relative vaut $\frac{|x-\mathrm{fl}(x)|}{|x|} \approx 1.12\ 10^{-7} \leq \varepsilon_{\mathrm{mach},10} = \frac{1}{2}10^{-6}$.

retour au cours

Sommaire
Concepts

Exemples
Exercices
Documents

Exemple A.1.4 Exemple d'ensemble flottants en base 2.

On prend $t = 3$, $L = -1$ et $U = 2$. Dans ce cas, $\text{card}(\mathscr{F}_2) = 33$.

On se donne un flottant $f \in \mathscr{F}_2$. L'exposant e de f est déterminé de façon à ce que le premier chiffre de la mantisse soit 1 : ainsi $\frac{3}{2} = 1 + \frac{1}{2} = 2(\frac{1}{2} + \frac{1}{4}) = (0.110)_2\ 2^1$, et $\frac{5}{2} = 2 + \frac{1}{2} = 4(\frac{1}{2} + \frac{1}{8}) = (0.101)_2\ 2^2$.

Pour $e = 0$, les flottants positifs de $\mathscr{F}_2$ sont compris entre 1/2 et 1 (exclus), et valent

$$\begin{array}{lll} 1/2 & = 1/2 & = (0.100)_2 \\ 5/8 & = 1/2 + 1/8 & = (0.101)_2 \\ 3/4 & = 1/2 + 1/4 & = (0.110)_2 \\ 7/8 & = 1/2 + 1/4 + 1/8 & = (0.111)_2. \end{array}$$

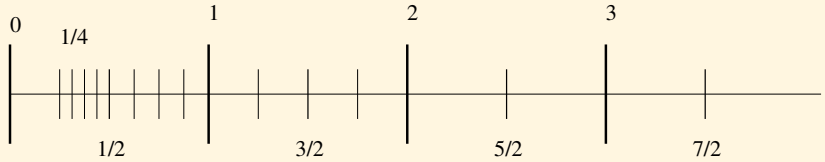


FIGURE A.1.1 – Répartition des flottants ≥ 0 en base 2, avec $t = 3$ chiffres significatifs, des bornes d'exposant $L = -1$, $U = 2$.

retour au cours

Sommaire
Concepts

Exemples
Exercices
Documents

Exemple A.1.5 Calcul flottant en base 2.

On se place en base 2, avec 3 chiffres significatifs ($t = 3$). Les nombres $\frac{5}{8} = (0.101)_2$ et $\frac{3}{4} = (0.110)_2$ appartiennent tous les deux à $\mathscr{F}_2$, mais $\frac{5}{8} + \frac{3}{4} = \frac{11}{8} = 1 + \frac{3}{8} \notin \mathscr{F}_2$.

En effet :

$$\begin{array}{rl} & 0.101 \\ + & 0.110 \\ \hline & 1.011 \\ = & 0.1011\ 2^1 \end{array}$$

Le dernier chiffre (le quatrième) ne peut pas être pris en compte. Ici, le résultat du calcul sera l'un des deux flottants les plus proches : $(0.101)_2\ 2^1 = 1 + \frac{1}{4} = \frac{5}{4}$ ou $(0.110)_2\ 2^1 = 1 + \frac{1}{2} = \frac{3}{2}$.

retour au cours

Sommaire
Concepts

Exemples
Exercices
Documents

Exemple A.1.6 Addition flottante

On prend $\mathscr{F}_{10}$ avec $t = 7$ chiffres.

Soit $a = 0.1234567$, $b = 0.4711325\ 10^4 \in \mathscr{F}_{10}$. Comme $a < b$, on effectue le décalage de mantisse de a de façon à avoir le même exposant 10^4 pour b et a, puis on les ajoute.

$$\begin{array}{rl} 0.47113250000 & 10^4 \\ +\ 0.00001234567 & 10^4 \\ \hline 0.47114484567 & 10^4 \end{array}$$

On arrondit au flottant le plus proche et on obtient $a \oplus b = 0.4711448\ 10^4$.

Il faut noter que les 4 derniers chiffres significatifs de a ont été éliminés lors de l'opération d'arrondi. L'erreur relative commise lors de cette addition vaut

$$e_r = \frac{|0.47114484567\ 10^4 - 0.4711448\ 10^4|}{|0.47114484567\ 10^4|} \approx \frac{4.567\ 10^{-4}}{4.7\ 10^3} \approx 10^{-7},$$

qui est inférieure à la précision machine $\varepsilon_{\text{mach},10} = \frac{1}{2}10^{-t+1} = 5\ 10^{-7}$. Il y a une perte d'information (les derniers chiffres de a), mais le résultat du calcul est précis.

retour au cours

Sommaire
Concepts

Exemples
Exercices
Documents

Exemple A.1.7 Addition flottante non associative.

On reprend l'exemple précédent avec $c = -b$.

Si on calcule $a \oplus (b \oplus c)$, on obtient $a = 0.1234567$ (le calcul est exact dans ce cas, car $b \oplus c = 0$ exactement).

Si on effectue $(a \oplus b) \oplus c$, on obtient $y = a \oplus b = 0.4711448\ 10^4$. Il reste à ajouter c :

$$\begin{array}{rl} 0.4711448 & 10^4 \\ \text{- } 0.4711325 & 10^4 \\ \hline 0.0000123 & 10^4 \end{array}$$

Dans cette addition, il n'y a pas besoin d'arrondir pour obtenir y (le calcul est exact dans ce cas : $y \oplus c = y + c$).

On obtient donc $(a \oplus b) \oplus c = 0.1230000$ qui est différent de $a \oplus (b \oplus c)$.

L'erreur relative commise lors du calcul vaut

$$e_r = \frac{|0.1230000 - 0.1234567|}{|0.1234567|} \approx 3.7\ 10^{-3},$$

et cette fois-ci l'erreur est beaucoup plus importante.

retour au cours

Sommaire
Concepts

Exemples
Exercices
Documents

Exemple A.1.8 Addition flottante de $(1+\varepsilon)$.

Soit un flottant $x \in \mathscr{F}$, tel que $0 < x < \varepsilon_{\text{mach}}$. Alors $(x \oplus 1) \ominus 1 = 0$, tandis que $x \oplus (1 \ominus 1) = x$.

retour au cours

Sommaire
Concepts

Exemples
Exercices
Documents

Annexe B
Documents

B.1 Documents du chapitre 1

Sommaire
Concepts

Exemples
Exercices
Documents

Document B.1.1 Codage informatique des flottants : la norme `IEEE 754`

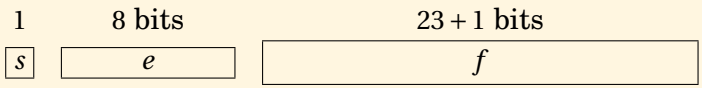
La norme `IEEE 754` établit des règles de codage, d'arrondi, de calcul... sur les flottants. La plupart des ordinateurs (utilisé par java, les processeurs intel...) la respectent.

Il existe en particulier deux flottants standard, le flottant simple précision (appelé `float` en langage C) codé sur 32 bits, et le flottant double précision (appelé `double` en langage C) codé sur 64 bits. Il existe également des doubles étendus dont nous ne parlerons pas.

Les `float`, très employés à l'origine en raison du faible espace mémoire disponible dans les anciens ordinateurs, sont à présents moins courants. On ne les présente ici que dans un but pédagogique. Les `double` sont utilisés presque partout (à l'exception de la plupart des calculatrices scientifiques). Le logiciel Scilab en particulier fonctionne avec des `double`.

Simple précision (`float`) : codés sur 32 bits, ils ont pour caractéristique :

$$t = 24,\ L = -126,\ U = 127,\ F^{\max} \approx 10^{38},\ F^{\min} \approx 10^{-38},\quad \varepsilon_{\text{mach}} = 2^{-24} \approx 5.96\ 10^{-8}.$$



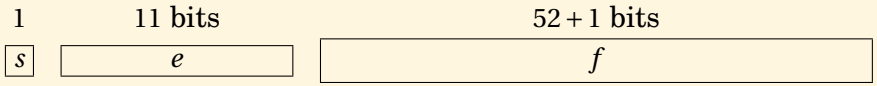
Un `simple` contient 1 bit pour le signe, 8 bits pour l'exposant et 23 bits pour la mantisse.Il faut noter qu'en pratique, le premier bit de la mantisse (qui vaut toujours 1 en base 2) n'est pas stocké, ce qui permet de gagner 1 bit de précision ($t = 23$ bits stockés + 1).

Sommaire
Concepts

Exemples
Exercices
Documents

Double précision (`double`) **:** codés sur 64 bits, ils ont pour caractéristique :

$$t = 53,\ L = -1022,\ U = 1023,\ F^{\max} \approx 10^{308},\ F^{\min} \approx 10^{-308} \quad \varepsilon_{\text{mach}} = 2^{-53} \approx 1.11\ 10^{-16}.$$



Scilab dispose de la variable prédéfinie $\%eps = 2^{-52} = 2\varepsilon_{\text{mach},2} \approx 2.22\ 10^{-16}$, qui correspond à l'écart entre 1 et le plus petit flottant supérieur à 1.

En plus des nombres flottants décrits ci-dessus (appelés flottants normalisés), les flottants `IEEE 754` comprennent :

— ± 0 (il y a deux 0, avec le signe + et le signe −),
— $\pm\infty$ (tout nombre plus grand en valeur absolue que $F^{\max}$),
— des nombres dénormalisés (autour de 0, leur premier bit de mantisse vaut 0 au lieu de 1),
— `NaN`, *"not a number"*, qui ne sont pas des nombres. Ils correspondent à des erreurs ou des calculs impossibles (ex. : l'opération ∞/∞ en Scilab retourne `NaN`).

Des règles précises sont définies pour les calculs d'arrondi (au plus proche, vers 0, vers l'infini).

retour au cours

Document B.1.1
Codage informatique des flottants : la norme `IEEE 754`

Sommaire
Concepts

Exemples
Exercices
Documents

Annexe C
Exercices

C.1 Exercices du chapitre 1

Exercice C.1.1 Série géométrique

On reprend les notations de la Section Flottants / Base 2.

1. Montrer que $f_2^{\max} = 0.11\ldots1 = 1 - 2^{-t}$.
   (Indication : utiliser la série géométrique $\sum_{i=1}^{t} 2^{-i}$).

retour au cours

Sommaire
Concepts

Exemples
Exercices
Documents

C.2 Exercices de TD du chapitre 1

Sommaire
Concepts

Exemples
Exercices
Documents

Exercice C.2.1 TD1-Exercice1 : $(1+x)-1=$?

1. Déterminer en fonction de $\varepsilon_{\text{mach},10}$ la valeur du flottant $u \in \mathscr{F}_{10}$, qui est le plus petit flottant strictement plus grand que 1.
2. Soit $x_1 \in [0,\ \varepsilon_{\text{mach},10}[$. Calculer $y_1 = \text{fl}(1+x_1)$, puis en déduire $z_1 = \text{fl}(1+x_1)-1$. Que vaut l'erreur relative pour z_1 ? (On supposera que toutes les opérations arithmétiques sont exactes).
3. Mêmes questions pour $x_2 \in]\varepsilon_{\text{mach},10},\ 2\varepsilon_{\text{mach},10}]$.
4. Mêmes questions pour $x_3 \in]-\varepsilon_{\text{mach},10}/10,\ 0[$, $(x_3 < 0)$.

Question 1 Aide 1 Aide 2
Question 2 Aide 1 Aide 2 Aide 3
Question 3 Aide 1 Aide 2
Question 4 Aide 1 Aide 2

Sommaire
Concepts

Exemples
Exercices
Documents

Exercice C.2.2 TD1-Exercice2 : Suite récurrente linéaire d'ordre 1

Soit α, β et u_0 trois réels. Soit la suite

$$\begin{cases} u_n = \alpha u_{n-1} + \beta & n \geq 1, \\ u_0 \quad \text{donné.} \end{cases}$$

1. Montrer que la suite admet pour solution $u_n = \alpha^n u_0 + \frac{\alpha^n - 1}{\alpha - 1}\beta$.
2. En fonction de α et β, discuter de l'existence de la limite de $(u_n)_{n\in\mathbb{N}}$ et donner la valeur de cette limite quand elle existe.
3. On prend $\alpha = 4$, $\beta = -1$ et $u_0 = \frac{1}{3}$. Que vaut u_n ? Quelle est sa limite éventuelle ?
4. On travaille à présent en arithmétique exacte. On tient compte du fait qu'une erreur d'arrondi est faite sur la condition initiale : $\tilde{u}_0 = 1/3\,(1-\delta)$ (avec δ petit). Calculer la solution perturbée $\tilde{u}_n$. Quelle est sa limite quand n tend vers l'infini ?

Sommaire
Concepts

Exemples
Exercices
Documents

Exercice C.2.3 TD1-Exercice3 : Suite récurrente linéaire d'ordre 2

On se donne la suite suivante, définie par récurrence :

$$\begin{cases} u_{n+2} &= \frac{9}{4}u_{n+1} - \frac{1}{2}u_n, \qquad n = 0, 1, \ldots, \\ u_0 &= \alpha, \\ u_1 &= \beta. \end{cases}$$

1. Donner et résoudre l'équation caractéristique de la suite.
2. Que vaut la solution exacte u_n en fonction de n, de α et β ? Quelle est la limite de u_n ?
3. Quand $\alpha = \frac{1}{3}$ et $\beta = \frac{1}{12}$, donner cette solution et sa limite quand n tend vers l'infini.
4. On travaille à présent avec une arithmétique exacte, mais en tenant compte des erreurs d'arrondi qui sont faites sur la condition initiale. On suppose donc que $\tilde{u}_0 = \tilde{\alpha} = 1/3\,(1+\tau_1)$ et $\tilde{u}_1 = \tilde{\beta} = 1/12\,(1+\tau_2)$ (avec τ_1 et τ_2 petits et inconnus) et que la suite $(\tilde{u}_n)_{n\in\mathbb{N}}$ satisfait la relation de récurrence ci-dessus. Calculer la solution pertubée $\tilde{u}_n$. Quelle est sa limite quand n tend vers l'infini ?
5. Étudier la fonction $\varphi(x) = \rho_1 e^{\mu_1 x} + \rho_2 e^{-\mu_2 x}$ où ρ_1, ρ_2, μ_1 et μ_2 sont des réels > 0.
6. On suppose pour simplifier que $\tau_2 = 8\tau_1$ et $\tau_1 = \varepsilon_{\text{mach},2}$ en double précision. En utilisant la question précédente, déterminer la valeur minimum de $\tilde{u}_n$. Pour quelle valeur de n est-elle atteinte ?

Sommaire
Concepts

Exemples
Exercices
Documents

Exercice C.2.3
TD1-Exercice4 : Suite récurrente linéaire d'ordre 2

Question 1 Aide 1 Aide 2
Question 2 Aide 1 Aide 2
Question 3 Aide 1
Question 4 Aide 1 Aide 2 Aide 3
Question 5 Aide 1
Question 6 Aide 1 Aide 2 Aide 3

Sommaire
Concepts

Exemples
Exercices
Documents

Exercice C.2.4 TD1-Exercice4 : Approximation de la dérivée par le taux d'accroissement.

Soit $a < b$ deux réels. Soit une fonction $f : [a,b] \to \mathbb{R}$ de classe C^2 sur $[a,b]$. On cherche à approcher la dérivée $f'(x)$ en un point $x \in]a,b[$ par le taux d'accroissement

$$\tau_h = \frac{f(x+h) - f(x)}{h},$$

avec $h \neq 0$, supposé assez petit pour que $x + h \in [a,b]$. On pose

$$M_0 = \max_{x\in[a,b]} |f(x)|, \qquad M_1 = \max_{x\in[a,b]} |f'(x)|, \qquad M_2 = \max_{x\in[a,b]} |f''(x)|.$$

1. En écrivant une formule de Taylor à l'ordre 2, relier $f'(x)$ à τ_h.
2. En déduire une majoration de l'écart $E(h) = |\tau_h - f'(x)|$ en fonction de M_2 et de h. Quelle est la limite de $E(h)$ quand h tend vers 0 ?
3. On suppose maintenant que l'évaluation de la fonction f en x n'est pas exacte, mais approchée par une valeur $\tilde{f}(x)$ qui vérifie

   $$\tilde{f}(x) = f(x)\,(1 + \delta(x)), \qquad \text{et} \quad |\delta(x)| \leq \varepsilon_{\text{mach}}.$$

   En utilisant une formule de Taylor, calculer une approximation de

   $$\tilde{\tau}_h = \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h}.$$

4. En déduire une majoration de l'écart $\tilde{E}(h) = |\tilde{\tau}_h - f'(x)|$. Que se passe-t-il quand h tend vers 0 ?

Sommaire
Concepts

Exemples
Exercices
Documents

5. Étudier la fonction $\varphi(x) = \alpha x + \frac{\beta}{x} + \gamma$ sur $]0, +\infty[$, où α, β et γ sont > 0. En déduire pour quelle valeur de $h = h_{\min} > 0$ l'écart $\tilde{E}(h)$ est minimal. Que vaut alors $\tilde{E}_{\min} = \tilde{E}(h_{\min})$?
6. On prend $f(x) = x$. Calculer exactement les erreurs dans ce cas.

Question 1 Aide 1
Question 2 Aide 1
Question 3 Aide 1 Aide 2
Question 4 Aide 1
Question 5 Aide 1 Aide 2

Exercice C.2.4
TD1-Exercice4 : Approximation de la dérivée par le taux d'accroissement.

Sommaire
Concepts

Exemples
Exercices
Documents

Index des concepts

Le gras indique un grain où le concept est défini ; l'italique indique un renvoi à un exercice ou un exemple, le gras italique à un document, et le romain à un grain où le concept est mentionné.

Sommaire
Concepts

Exemples
Exercices
Documents

Revoyez votre cours.
Écrivez 1 en écriture flottante. Que vaut l'exposant e ?
Écrivez u fonction de t.
Déterminer l'écart entre les 2 flottants en fonction de t.

Retour à l'exercice ▲

On a : $1 = 0.10\ldots00 \times 10^1 = f_{10}^{\min}\ 10^1$ et son successeur $u = 0.100\ldots01 \times 10^1$. L'exposant vaut $e = 1$.
L'écart entre les 2 est donc $10^{-t} \times 10 = \delta_1$. On a

$$u = 1 + \delta_1 = 1 + 10^{-t+1} = 1 + 2\varepsilon_{\text{mach},10}.$$

Retour à l'exercice ▲

De quel flottant $1 + x_1$ est-il le plus proche ?
Quel est le milieu de $[1,\ u]$? Faites un dessin pour vous aider.

Retour à l'exercice ▲

Le milieu de l'intervalle $[1,\ u]$ est $1+\delta_1/2 = 1+\varepsilon_{\text{mach},10}$.

Comme $1+x_1 \in [1,\ u]$ est plus proche de 1 que de son successeur $u = 1+2\varepsilon_{\text{mach},10}$, le flottant y_1 qui va approcher $1+x_1$ sera $y_1 = 1 \oplus x_1 = \text{fl}(1+x_1) = 1$.

Calculer l'erreur relative entre la solution exacte de $(1+x_1)-1$ et z.

Retour à l'exercice ▲

Donc $z_1 = 1 \ominus 1 = 1 - 1 = 0$. Le résultat du calcul exact est $(1 + x_1) - 1 = x_1$. L'erreur relative s'écrit

$$e_{r,1} = \frac{|z_1 - x_1|}{|x_1|} = \frac{|x_1|}{|x_1|} = 100\%.$$

Retour à l'exercice ▲

À présent, $1 + x_2 \in [1, u]$ est plus proche de u que de 1. Donc le flottant y_2 qui va approcher $1 + x_2$ sera $y_2 = 1 \oplus x_2 = \mathrm{fl}(1 + x_2) = u$.

Calculer l'erreur relative entre la solution exacte de $(1 + x_2) - 1$ et z_2.

Retour à l'exercice ▲

Donc $z_2 = (1 + 2\varepsilon_{\text{mach},10}) \ominus 1 = 2\varepsilon_{\text{mach},10}$. Le résultat du calcul exact est $(1 + x_2) - 1 = x_2$. L'erreur relative s'écrit

$$e_{r,2} = \frac{|z_2 - x_2|}{|x_2|} = \frac{|2\varepsilon_{\text{mach},10} - x_1|}{|x_1|}$$

qui varie entre 0% quand $x_2 = 2\varepsilon_{\text{mach},10}$ et 100% quand $x_2 = \varepsilon_{\text{mach},10}$.

Retour à l'exercice ▲

Attention, on se trouve sur l'intervalle $[\frac{1}{10}, 1[$, donc $e = -1$.
Trouver le flottant v immédiatement inférieur à 1. Quel est le milieu de $[v, 1]$? Faites un dessin pour vous aider.
De quel flottant $1 + x_3$ est-il le plus proche ?

Retour à l'exercice ▲

On trouve $y_3 = 1 \oplus x_3 = 1$. Le reste de la question est donc identique à la question 2.

Notez que la longueur de l'intervalle $]-\varepsilon_{\text{mach},10}/10,\ 0]$ est dix fois plus petite que celle de $[0,\ \varepsilon_{\text{mach},10}[$ (cf. question 2.). C'est dû au fait que l'exposant vaut ici $e = 0$, alors que dans la question 2., il valait $e = 1$.

Retour à l'exercice ▲

Revoyez votre cours sur les suites. Pour retrouver l'équation caractéristique (qui permet de déterminer les solutions de la suite), on cherche u_n sous la forme $u_n = A\lambda^n$. Injecter dans l'équation de récurrence et obtenez une équation polynomiale de degré deux à résoudre.

Retour à l'exercice ▲

Équation caractéristique

$$\lambda^2 - \frac{9}{4}\lambda + \frac{1}{2} = 0.$$

Retour à l'exercice ▲

Aide 1, Question 2, Exercice C.2.3

Comme l'équation caractéristique admet 2 racines distinctes λ_1 et λ_2, les solutions sont du type $u_n = A\lambda_1^n + B\lambda_2^n$.

Déterminer $\lambda_1 > 1$ et $0 < \lambda_2 < 1$.

Utiliser les conditions initiales u_0 et u_1 pour trouver A et B.

Retour à l'exercice ▲

On trouve

$$u_n = \frac{1}{7}(-\alpha + 4\beta)\,\lambda_1^n + \frac{1}{7}(8\alpha - 4\beta)\,\lambda_2^n.$$

La limite de u_n dépend du fait que $4\beta = \alpha$ ou non.

Retour à l'exercice ▲

Avec ces conditions initiales, on obtient $A = 0$.

Retour à l'exercice ▲

Avec ces conditions initiales particulières, la solution perturbée devient $\tilde{u}_n = \tilde{A}\lambda_1^n + \tilde{B}\lambda_2^n$. Pour calculer $\tilde{A}$ et $\tilde{B}$ en fonction de τ_1 et τ_2, il suffit de remplacer α et β dans l'expression de u_n par $\tilde{\alpha}$ et $\tilde{\beta}$.

Retour à l'exercice ▲

$\tilde{A}$ n'est plus nul et dépend de $\tau_1 - \tau_2$.

Retour à l'exercice ▲

On trouve $\tilde{A} = \frac{1}{21}(\tau_2 - \tau_1)$ (en général non nul) et $\tilde{B} = \frac{1}{3}(1 + \frac{8\tau_1 - \tau_2}{7}) \approx \frac{1}{3}$.

Retour à l'exercice ▲

Aide 1, Question 5, Exercice C.2.3

φ atteint un minimum en $x_{\min} = \frac{\ln(K)}{\mu_1+\mu_2}$, où K dépend de ρ_1, ρ_2, μ_1 et μ_2

Retour à l'exercice ▲

Étudier la fonction $\varphi(x) = \tilde{A}e^{x\ln(2)} + \tilde{B}e^{-2x\ln(2)}$.

Calculer le minimum de la fonction en utilisant le fait que $\varepsilon_{\text{mach},2} = 2^{-53}$.

En déduire n_0 tel que $\tilde{u}_{n_0}$ est minimal. Calculer $\tilde{u}_{n_0}$ (faire les applications numériques jusqu'au bout).

Retour à l'exercice ▲

On trouve $K = 2^{54}$ et $x_{\min} = 18$. On trouve donc un minimum pour $n_0 = 18$ et $\tilde{u}_{18} = \frac{1}{3}(2^{-53}2^{18} + 2^{-2\times 18}) = 2^{-36} \approx 1.45 \times 10^{-11}$.

Regardez ce que donne `scilab`.

Retour à l'exercice ▲

Le résultat obtenu avec ce modèle simplifié est très précis, car on obtient avec `scilab` un minimum pour $n = 19$ qui vaut $\mathtt{U}_{19} \approx 2.77 \times 10^{-12}$.

Retour à l'exercice ▲

D'après Taylor, il existe $\theta \in [0,1]$ tel que

$$\tau_h = f'(x) + \frac{h}{2} f''(x + \theta h).$$

Retour à l'exercice ▲

Aide 1, Question 2, Exercice C.2.4

Utiliser la question précédente pour obtenir une majoration de $E(h)$ comme $C|h|$, où C ne dépend que de M_2.

Retour à l'exercice ▲

Écrire

$$\tilde{\tau}_h \quad = \quad \frac{1}{h}\big(f(x+h)(1+\delta(x+h)) - f(x)(1+\delta(x))\big)$$

et utiliser le développement de Taylor de $f(x+h)$.

Retour à l'exercice ▲

Aide 2, Question 3, Exercice C.2.4

Dans $\tilde{\tau}_h$, trois termes apparaissent : un terme en $\frac{f(x)}{h}$, un terme en $f'(x)$ et un terme en $\frac{h}{2}f''(x+\theta h)$. Devant chaque terme, il y a des facteurs à déterminer, qui peuvent dépendre de $\delta(x)$ ou de $\delta(x+h)$.

Retour à l'exercice ▲

Obtenir une majoration du type :

$$\tilde{E}(h) \quad \leq \quad \frac{K_0}{|h|} + K_1 + |h| K_2,$$

où K_0, K_1 et K_2 sont à déterminer en fonction de $\varepsilon_{\text{mach}}$, de M_0, M_1 et M_2.

Retour à l'exercice ▲

Trouver une majoration du type :

$$\tilde{E}_{\min} \leq 2\sqrt{K_0 K_2} + K_1.$$

Retour à l'exercice ▲

Remarquer que, en pratique **le taux d'accroissement *ne tend pas* vers la dérivée !** Le minimum est de l'ordre de $\sqrt{\varepsilon_{\text{mach}}}$. En double précision, l'approximation de la dérivée par le taux d'accroissement sera au mieux de l'ordre de 10^{-8} (en précision relative).

Retour à l'exercice ▲