Institut de Formation et de Recherche en Informatique

Licence 2 : Statistique-Informatique

Année académique : 2020 – 2021
L2 : Analyse Numérique
Periode : Semestre 3 **Enseignant :** Prof. Guy DEGLA.

I. Utilité et limitation des méthodes numériques

La résolution des problèmes scientifiques passe par une modélisation (représentation mathématique) des phénomènes mis en jeu. Pour parvenir à les représenter, il faut souvent négliger certains phénomènes et simplifier d'autres en ne prenant en compte que les grandeurs et les variables essentielles.

Malgré ces simplifications, les équations obtenues sont souvent insolubles par les méthodes algébriques ou analytiques classiques. D'où la nécessité de recourrir à des méthodes numériques.

L'essor des méthodes numériques résulte principalement de la conjoncture de trois éléments à savoir

- La plupart des problèmes "simples" ayant déjà été résolus, on est depuis une cinquantaine d'années confronté à des problèmes de plus en plus compliqués et insolubles par les méthodes mathématiques traditionnelles.

- On a développé depuis la fin de la deuxième guerre mondiale (1945) des ordinateurs et calculateurs electroniques de puissance et de rapidité extraordinaires, sans cesse croissantes, à des prix de plus en plus bas, accessibles à une très grande masse d'utilisateurs et sans cesse croissante.

- Dans le même temps, les Mathématiciens ont développé des techniques de résolution de plus en plus efficaces et applicables à une variété de problèmes mathématiques.

On note deux limitations à l'utilisation des méthodes numériques dues au fait que:

- Certains programmes sont si colossaux (importants) qu'ils dépassent les capacités des ordinateurs actuels. Soit le nombre de données dépasse la capacité mémoire, soit la résolution dure trop longtemps.
  Dans ce cas, la possibilité d'utiliser des méthodes numériques, dépend de la disponibilité du prix à payer pour résoudre le problème.

- Il n'existe pas encore de modèles mathématiques complets et précis pour certains problèmes.

Pour plus de détails, se référer à [J.P. Nougier].

II. Sources d'erreurs dans un modèle numérique

En général nous distinguons dans un modèle numérique les sources d'erreurs suivantes:

1. **les erreurs de modélisation**, qui peuvent être contrôlées par un choix convenable du modèle mathématique. Comme leur nom l'indique, ces erreurs proviennent de l'étape de mathématisation du phénomène physique auquel on s'intéresse. C'est l'étape qui consiste à faire ressortir les causes les plus déterminantes du phénomène observé et à les mettre sous forme d'équations (algébriques ou différentielles le plus souvent). Lorsque le phénomène est très complexe, il faut simplifier et négliger ses composantes qui paraissent moins importantes ou qui rendent la résolution numérique trop difficile;

2. **les erreurs sur les données**, qui peuvent être réduites en améliorant la précision des mesures. ;

3. **les erreurs de troncature**, qui proviennent du fait qu'on a remplacé dans le modèle numérique des passages à la limite par des opérations mettant en jeu un nombre fini d'étapes.
   Ces erreurs proviennent principalement de l'utilisation du développement de Taylor, permettant par exemple de remplacer une équation différentielle par une équation algébrique.
   Le développement de Taylor est le principal outil mathématique du numéricien. C'est donc primordial d'en maîtriser l'énoncé et ses conséquences;

4. **les erreurs d'arrondi**, qui proviennent principalement des représentations des nombres (sur l'ordinateur). En effet, la représentation des nombres sur ordinateur, généralement binaire et finie, introduit souvent des erreurs. Même initialement infirmes, ces erreurs peuvent s'accumuler quand on effectue un très grand nombre d'opérations. C'est des erreurs qui se propagent au fil des calculs et qui peuvent même compromettre la précision des résultats.

Les erreurs des points 3 et 4 constituent **l'erreur numérique**. *Une méthode numérique est dite convergente* si cette erreur peut être rendue arbitrairement petite quand on augmante l'effort de calcul.
Naturellement la convergence est le but principal (mais pas le seul) d'une méthode numérique; les autres buts étant la *précision*, la *fiabilité* et l'*efficacité*.

La **précision** d'une méthode numérique signifie que les erreurs sont petites par rapport à une tolérance fixée. Elle est généralement mesurée par l'ordre infinitésimal de l'erreur e_n par rapport au paramètre de discrétisation. Noter que la *précision de la machine* ne limite par théoriquement la précision de la méthode.

La **fiabilité** signifie qu'il est possible de garantir que l'erreur globale se situe en dessous d'une certaine tolérance. Naturellement, un modèle numérique peut être considéré comme fiable seulement s'il a été *testé*, c'est-à-dire validé par plusieurs cas tests.

L'**efficacité** signifie que la complexité du calcul (i.e., la quantité d'opérations et la taille de mémoire requise) nécessaire pour maîtriser l'erreur est aussi petite que possible.

Rappelons que par **algorithme**, nous entendons une démarche qui décrit, à l'aide d'opérations élémentaires finies, toutes les étapes nécessaires à la résolution d'un problème spécifique.
Un algorithme peut à son tour contenir des sous-algorithmes. Il doit avoir la propriété de s'achever après un nombre fini d'opérations élémentaires. Celui qui exécute l'algorithme (une machine ou un être humain) doit y trouver toutes les instructions pour résoudre complètement le problème considéré, pourvu que les ressources nécessaires à son éxécution soient disponibles.

Enfin, la *complexité d'un algorithme* est une mesure de son temps d'éxécution. Calculer la complexité d'un algorithme fait alors partie de l'analyse de l'éfficacité d'une méthode numérique.
Plusieurs algorithmes, de complexité différentes, peuvent être utilisés pour résoudre un même problème P. On introduit alors la notion de *complexité d'un problème*. Cette dernière se définit comme étant la complexité de l'Algorithme qui a la complexité la plus petite parmi ceux qui resolvent le problème P. La complexité d'un problème est typiquement mesurée par un paramètre directement associé à P. Par exemple, dans le cas du produit de deux matrices carrées d'ordre n, la complexité du calcul peut être exprimée en fonction d'une puissance de la taille n.

III. Les formes les plus courantes de représentation des nombres sur ordinateur

La structure interne de la plupart des ordinateurs s'appuie sur le système binaire. Dans ce cas, l'**unité d'information** ou **bit** prend la valeur 0 ou 1. Il est évident que peu d'information peut être stockée au moyen d'un seul bit. On regroupe donc les bits en *mots* (codes) de longueur variable dont les plus courantes sont les longueurs de 8, de 16, de 32 ou de 64. Les nombres, entiers ou réels, sont représentés de cette façon, bien que leur **mode précis de représentation** dépende du fabriquant.

3.1. Représentation des entiers signés

Nous considérons la Représentation binaire habituelle d'un entier naturel, la Représentation signe et grandeur d'un entier relatif, la Représentation en complément à 2 d'un entier relatif, et puis la Représentation par excès d'un entier relatif.

3.1.1 Représentation binaire habituelle d'un entier naturel

On rappelle que pour tout entier positif non nul N, il existe un unique entier naturel non nul n tel que

$$2^{n-1} \;\leq\; N \;<\; 2^n\,,$$

et puis n entiers naturels (*chiffres*)

$$a_0\,, \quad \ldots\,, \quad a_{n-1}$$

satisfaisant

$$1 \leq a_{n-1} < 2 \qquad \text{et} \quad 0 \leq a_{n-i} < 2 \quad \text{si} \;\; 1 < i \leq n\,;$$

c'est-à-dire

$$a_{n-1} = 1 \qquad \text{et} \quad a_{n-i} \in \{0,\,1\} \quad \text{si} \;\; i = 2,\,\ldots,\,n\,;$$

et tels que :

$$\begin{aligned} N &= a_{n-1} \times 2^{n-1} + a_{n-2} \times 2^{n-2} + a_{n-3} \times 2^{n-3} + \ldots + a_1 \times 2^1 + a_0 \times 2^0 \\ &= \sum_{i=1}^{n} a_{n-i}\, 2^{n-i}\,. \end{aligned}$$

Le nombre N (considéré dans le système décimal) est alors représenté par

$$\big(a_{n-1}\, a_{n-2}\, a_{n-3}\, \cdots\, a_1\, a_0\big)_2$$

ou parfois

$$\overline{a_{n-1}\, a_{n-2}\, a_{n-3}\, \cdots\, a_1\, a_0}^{\,2}\,.$$

Et donc

$$(N)_{10} \;=\; \big(a_{n-1}\, a_{n-2}\, a_{n-3}\, \cdots\, a_1\, a_0\big)_2\,.$$

Dans ce qui précède, le sous-indice de chacune des parenthèses indique la base correspondante.
En pratique, on obtient successivement les valeurs $a_0\,,\ \ldots,\ a_{n-1}$ *en procédant de la façon suivante :*

- le chiffre a_0 est le reste de la division euclidienne de N par 2 ;
- on refait le même raisonnement avec la partie entière de $N/2$ pour obtenir a_1 ;
- on continue jusqu'à obtenir une partie entière nulle.

Par exemples,

⋆ si $N = (35)_{10}$, alors on a:

$$\begin{array}{rcccccc} 35/2 & \longrightarrow & 17 & \text{reste } 1 & \text{ainsi} & a_0 & = & 1 \\ 17/2 & \longrightarrow & 8 & \text{reste } 1 & \text{ainsi} & a_1 & = & 1 \\ 8/2 & \longrightarrow & 4 & \text{reste } 0 & \text{ainsi} & a_2 & = & 0 \\ 4/2 & \longrightarrow & 2 & \text{reste } 0 & \text{ainsi} & a_3 & = & 0 \\ 2/2 & \longrightarrow & 1 & \text{reste } 0 & \text{ainsi} & a_4 & = & 0 \\ 1/2 & \longrightarrow & 0 & \text{reste } 1 & \text{ainsi} & a_5 & = & 1. \end{array}$$

Donc l'entier naturel 35 s'écrit 100011 en base 2. En effet on a bien

$$35 \;=\; 2^5 + 2^1 + 2^0 \,.$$

⋆ si $N = (100)_{10}$, alors on a:

$$\begin{array}{rcccccc} 100/2 & \longrightarrow & 50 & \text{reste } 0 & \text{ainsi} & a_0 & = & 0 \\ 50/2 & \longrightarrow & 25 & \text{reste } 0 & \text{ainsi} & a_1 & = & 0 \\ 25/2 & \longrightarrow & 12 & \text{reste } 1 & \text{ainsi} & a_2 & = & 1 \\ 12/2 & \longrightarrow & 6 & \text{reste } 0 & \text{ainsi} & a_3 & = & 0 \\ 6/2 & \longrightarrow & 3 & \text{reste } 0 & \text{ainsi} & a_4 & = & 0 \\ 3/2 & \longrightarrow & 1 & \text{reste } 1 & \text{ainsi} & a_5 & = & 1 \\ 1/2 & \longrightarrow & 0 & \text{reste } 1 & \text{ainsi} & a_6 & = & 1. \end{array}$$

Donc l'entier naturel 100 s'écrit 1100100 en base 2. En effet on a bien

$$100 \;=\; 2^6 + 2^5 + 2^2 \,.$$

Questions : Donner en base 2 les représentations respectives des entiers naturels suivants: 0, 2, 8, 9, 10, 12, 21 et 1000.

3.1.2 Représentation signe et grandeur d'un entier relatif

Dans cette représentation, un bit (le premier) est consacré au signe :

0 pour un entier positif
1 pour un entier negatif;

$$\big(\textit{ceci peut se justifier par le fait que } (-1)^{\mathbf{0}} = +1 \textit{ et } \; (-1)^{\mathbf{1}} = -1 \;\big),$$

et les autres bits servent à la représentation de la valeur absolue de l'entier.

Par exemple, en considérant une représentation signe et grandeur avec $n = 16$ bits, on a:

$$0100\,0000\,0000\,0000 \quad = \quad +2^{14}\,,$$

$$1100\,0000\,0000\,0000 \quad = \quad -2^{14}\,.$$

Le plus grand entier représentable dans ce cas est:

$$\begin{aligned} 0111\,1111\,1111\,1111 \quad &= \quad (-1)^0 \left(\textstyle\sum_{i=0}^{14} 2^{n-i}\right) \\ &= \quad +\left(2^{15} - 1\right) \\ &= \quad +32\,767 \quad \text{(en base 10).} \end{aligned}$$

C'est-à-dire, plus précisement, qu'en représentation signe et grandeur

$$(0111\,1111\,1111\,1111)_2 \;=\; (+32\,767)_{10}\,.$$

Noter que le nombre 0 peut être représenté de deux manières à savoir:

$$\begin{aligned} +0 &= 0000\,0000\,0000\,0000\,, \\ -0 &= 1000\,0000\,0000\,0000\,. \end{aligned}$$

Aussi dans le cas de la représentation signe et grandeur avec 16 bits, si un calcul sur des nombres entiers donne un entier supérieur au nombre 32 767, alors le compilateur enverra un message d'erreur indiquant un **débordement (overflow)**.

Par ailleurs dans la représentation signe et grandeur, et également dans les représentations utilisées dans la suite, nous optons pour la convention selon laquelle le premier bit est celui situé le plus à gauche. Cependant, soulignons qu'en Informatique, il arrive souvent de considérer une numérotation des bits allant de 0 à $n-1$ en commençant par le bit le plus à droite dit **le moins significatif**.

Questions :

- ⋆ Quel est le nombre le plus petit que l'on peut écrire dans la représentation signe et grandeur avec 16 bits?
- ⋆ Quelle est la représentation signe et grandeur avec 16 bits du nombre 23 716 ?

3.1.3 Représentation en complément à 2 d'un entier relatif

La représentation en complément à 2 s'utilise fréquemment.
Lorsqu'on dispose de n bits ($n \geq 2$) pour exprimer un entier relatif N, on procède à la décomposition:

$$\begin{aligned} N &= -a_{n-1} \times 2^{n-1} + a_{n-2} \times 2^{n-2} + a_{n-3} \times 2^{n-3} + \cdots + a_1 \times 2^1 + a_0 \times 2^0 \\ &= -a_{n-1} \times 2^{n-1} + \sum_{i=2}^{n} a_{n-i}\, 2^{n-i}. \end{aligned}$$

Il faut remarquer le signe négatif présent devant le terme a_{n-1} et constater facilement que tous les entiers positifs (entiers naturels) vérifient:

$$a_{n-1} \;=\; 0\,.$$

Les entiers positifs sont donc représentés par 0 suivi de leur expression binaire habituelle en $(n-1)$ bits.
Quant à celle d'un nombre négatif; $-2^{n-1} < N \leq -2^{n-2}$, il suffit de lui ajouter 2^{n-1} et de transformer le résultat en forme binaire.

Par exemples,

- ⋆ La représentation en complément à 2 sur 4 bits 0101 vaut:

$$-0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0\,,$$

soit 5 en forme décimale.

- ⋆ La représentation en complément à 2 sur 4 bits 1101 vaut:

$$-1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 \,,$$

soit -3 en forme décimale.

- ⋆ Inversement, la représentation en complément à 2 du nombre décimal -6 sera 1 suivi de la représentation en complément à 2 sur 3 bits de

$$-6 + 2^3 = 2$$

qui est 010. C'est-à-dire que

$$(-6)_{10} = (1010)_2 \qquad \text{dans la représentation en complément à 2}\,.$$

Questions: :

- ⋆ Quel est le nombre décimal qui vaut $(01011)_2$ dans la représentation en complément à 2 ?
- ⋆ Quel est le nombre décimal qui vaut $(11010)_2$ dans la représentation en complément à 2 ?
- ⋆ Quelles sont respectivement les représentations en complément à 2 des nombres décimaux 15 et -13 ?

3.1.4 Représentation par excès d'un entier relatif

Pour représenter un entier décimal N par excès, il suffit de lui ajouter un **excès** d et de donner le résultat sous forme binaire.
Inversement, si on a la représentation binaire par excès d'un entier, il suffit de calculer sa valeur en base 10 et de soustraire d pour trouver l'entier recherché.

En général, avec un mot de n bits, la valeur de d est 2^{n-1} et on peut alors représenter au plus 2^n entiers différents, y compris les entiers négatifs. Ainsi avec $n = 4$ bits et $d = 2^3$, la représentation par excès a l'avantage d'ordonner la représentation binaire en assignant à 0000 le plus petit entier décimal représentable; à savoir $-d$. Donc on a:

Forme binaire	Forme décimale
0000	-8
0001	-7
$\vdots$	$\vdots$
1110	$+6$
1111	$+7$

Inversement, par exemple avec un mot de 8 bits et un excès $d = 2^{8-1} = 2^7 = 128$, pour représenter $(-100)_{10}$, il suffit d'ajouter 128 à -100, ce qui donne 28, et d'exprimer ce résultat sur 8 bits, soit 0001 1100.

Questions : Avec 8 bits et un excès 2^7, donner les représentations respectives des nombres décimaux -128, -28, 0, 30 et 127.

3.2 Représentation des nombres réels

La tâche de représentation des nombres réels (non entiers en particulier) est plus complexe. Dans le système décimal, on représente souvent un nombre décimal x par

$$x \;=\; m \times 10^k$$

où m est la **mantisse**, k l'**exposant** et 10 la **base**.

De façon générale, selon une base $b \in \mathbb{N} \setminus \{0, 1\}$ quelconque, on peut écrire un nombre décimal x comme suit:

$$x \;=\; m \times b^k\,;$$

avec pour forme générale de la mantisse:

$$m \;=\; 0, d_1 d_2 d_3 \cdots d_n$$

signifiant par définition que

$$\begin{aligned} m \quad &= \quad d_1 \times b^{-1} + d_2 \times b^{-2} + d_3 \times b^{-3} + \cdots + d_n \times b^{-n} \\ &= \quad \textstyle\sum_{i=1}^{n} d_i\, b^{-i}\,, \end{aligned}$$

où n est le nombre de chiffres (significatifs) de la mantisse et les chiffres d_i satisfont:

$$\begin{array}{rcccl} 1 & \leq & d_1 & \leq & b-1 \\ 0 & \leq & d_i & \leq & b-1\,, \quad \text{pour } i = 2,\, 3,\, \ldots,\, n. \end{array}$$

Le fait que d_1 soit supérieur ou égal à 1 signifie que la mantisse est **normalisée**; c'est-à-dire que son premier chiffre est toujours différent de 0. Cette normalisation assure l'unicité de la représentation et permet d'eviter les ambiguités entre

$$0,2016 \times 10^2 \quad \text{et} \quad 0,02016 \times 10^3$$

pour représenter le nombre $20,16$. La dernière expression ci-dessus n'est jamais retenue en machine. Dans cet exemple, on a considéré la base $b = 10$ et $n = 4$ dans la mantisse.
Ainsi la mantisse satisfait toujours

$$\frac{1}{b} \leq m < 1\,.$$

Ce sont ces considérations qui servent de lignes directrices pour la représentation d'un nombre réel sur ordinateur. Nous nous intéresserons généralement au cas où la base est 2 ($b = 2$). *Il faut alors trouver un moyen (une convention) de représenter la mantisse (fraction), l'exposant (un entier signé) et le signe de ce nombre.*

Remarque : Les calculatrices de poche se distinguent des ordinateurs principalement par le fait qu'elles utilisent la base 10 ($b = 10$) et une mantisse d'une longueur d'environ 10 ($n = 10$) et un exposant variant généralement entre -100 et 100.

Par exemple, considérons un mot de 8 bits comme

0	1	0	1	1	0	1	1

Alors, dans le cas d'une représentation signe et grandeur de l'exposant,

(i) le premier bit (0) donne le signe du nombre, soit

$$(-1)^0 \quad \longrightarrow \quad +\,,$$

(ii) on retient trois bits suivants (101) des sept restants pour représenter l'exposant, soit

$$101 \quad \longrightarrow \quad (-1)^1 \times \left(0 \times 2^1 \,+\, 1 \times 2^0\right) \;=\; -1\,,$$

(iii) et les quatre derniers bits (1011) représentent la mantisse, soit

$$0,1011 \quad \longrightarrow \quad 1 \times 2^{-1} \,+\, 0 \times 2^{-2} \,+\, 1 \times 2^{-3} \,+\, 1 \times 2^{-4} \;=\; 0,6875\,.$$

Donc 0101 1011 représente le nombres

$$0,6875 \times 2^{-1} \;=\; 0,34375\,.$$

Questions :

- ⋆ Justifier que dans la base b, la mantisse vérifie toujours $\frac{1}{b} \leq m < 1$?
- ⋆ Montrer que dans le cas d'une représentation en complément à 2 de l'exposant, le nombre réel signé de 8 bits $(0101\,1011)_2$ vaut $0,085\,9375$ en base 10.

3.3 Conversion d'une fraction décimale en valeur binaire

La méthode de conversion d'une fraction décimale en en valeur binaire est similaire à celle que l'on utilise dans le cas des entiers.
Soit f une fraction décimale comprise entre 0 et 1. Il s'agit de trouver les chiffres d_i tels que

$$(f)_{10} \;=\; (0, d_1 d_2 d_3 \cdots)_2$$

ou encore

$$f \;=\; d_1 \times 2^{-1} \,+\, d_2 \times 2^{-2} \,+\, d_3 \times 2^{-3} \,+\, \cdots \quad .$$

- Si on multiplie f par 2 (ce qui revient à diviser f par 2^{-1}), on obtient d_1 plus une fraction.
- En appliquant le même raisonnement à $(2f - d_1)$, on obtient d_2.
- On réitère ainsi le raisonnement jusqu'à ce que la partie fractionnaire soit nulle ou que l'on ait atteint le nombre maximal de chiffres de la mantisse.

Par exemples,

- ⋆ on a avec $f = 0,0625$:

$$\begin{array}{lcllcl} 0,0625 \times 2 & = & 0,1250 & \text{ainsi} & d_1 & = & 0 \\ 0,1250 \times 2 & = & 0,2500 & \text{ainsi} & d_2 & = & 0 \\ 0,2500 \times 2 & = & 0,5000 & \text{ainsi} & d_3 & = & 0 \\ 0,5000 \times 2 & = & 1,0000 & \text{ainsi} & d_4 & = & 1\,. \end{array}$$

Donc (en pratique)

$$(0,0625)_{10} \;=\; (0,0001)_2\,.$$

$\star$ on a avec $f = \frac{1}{3}$:

$$\begin{array}{rcll} \frac{1}{3} \times 2 & = & 0 + \frac{2}{3} & \text{ainsi} \quad d_1 = 0 \\ \frac{2}{3} \times 2 & = & 1 + \frac{1}{3} & \text{ainsi} \quad d_2 = 1 \\ \frac{1}{3} \times 2 & = & 0 + \frac{2}{3} & \text{ainsi} \quad d_3 = 0 \\ \frac{2}{3} \times 2 & = & 1 + \frac{1}{3} & \text{ainsi} \quad d_4 = 1 \\ \vdots & \vdots & \vdots & \vdots \end{array}$$

On peut poursuivre la conversion à l'infini et montrer que

$$\frac{1}{3} = (0,010101\cdots)_2 .$$

Et puisqu'en pratique, on n'utilise qu'un nombre fini de chiffres dans la mantisse, il faudra s'arrêter après un certain nombre n de bits.

Questions : Completer les tableaux suivants:

Forme décimale	Forme binaire
2	...
1,5	...
1,25	...
1,125	...
1,0625	...
0,5	...
0,2	...

Forme binaire	Forme décimale
0,1	...
0,01	...
0,001	...
0,1101	...
0,01011	...
10,01	...
1011	...

3.4 Norme IEEE

L '*Institute for Electrical and Electronic Engineering* (IEEE) s'éfforce de rendre aussi uniformes que possibles les représentations sur ordinateur. Il propose une représentation des nombres réels en **simple précision** sur 32 bits et en **double précision** sur 64 bits (*Convention IEEE*-754).
La représentation en simple (respectivement, double) précision est construite comme suit:

- le premier bit indique le signe du nombre,
- les 8 bits suivants (11 bits en double précision) déterminent l'exposant avec un excès de $127 = 2^{8-1} - 1$ $\big(1023 = 2^{11-1} - 1$ en double précision$\big)$, et
- les 23 derniers bits 23 (52 en double précision) sont pour la mantisse normalisée.

Puisque l'on normalise la mantisse, son premier chiffre est toujours 1 et il n'est pas nécessaire de le garder en mémoire. La mantisse normalisée peut donc commencer par un 0 tout en conservant la même précision

avec 24 bits (53 en double précision).

Ainsi donc, selon la représentation de Cheney & Kincaid, les 32 bits de la représentation en simple précision, le mot

$$(d_1 d_2 d_3 \cdots d_{31} d_{32})_2$$

désigne le nombre décimal

$$(-1)^{d_1} \times 2^{(d_2 d_3 \cdots d_9)_2} \times 2^{-127} \times (1, d_{10} d_{11} \cdots d_{32})_2 \, .$$

On observe immédiatement ls trois différentes composantes: le bit de signe, l'exposant avec un excès de 127 et la mantisse normalisée par l'ajout du 1 manquant.

Par exemple, en simple précision IEEE, les 32 bits

$$1100\,0001\,1110\,0000\,0000\,0000\,0000\,0000$$

se décomposent en

$$(-1)^1 \times 2^{(10000011)_2} \times 2^{-127} \times (1,11)_2 \;=\; -28\,.$$

Question : Vérifier que

$$(30,0625)_{10} \;=\; (11110,0001)_2 \;=\; 1,11100001 \times 2^4$$

et montrer que sa représentation en simple précision IEEE est:

$$0\ 10000011\ 11100001000000000000000\,.$$

Remarque : Noter que la norme IEEE traite le nombre 0 de façon particulière.

3.5. Erreurs dues aux repésentations

La repésentation en point flottant, par **troncature (chopping)** ou bien, par **arrondi** (rounding), induit une erreur relative qui dépend du nombre de bits de la mantisse, de l'utilisation de la troncature ou de l'arrondi, ains que du nombre à représenter. Notons que l'intervalle entre les nombres représentables varie en longueur selon l'exposant devant la mantisse.

Tout nombre réel s'écrit (se développe) dans le système décimal sous la forme

$$\mathbf{x} \;=\; \mathbf{0, d_1 d_2 d_3 \cdots d_n d_{n+1} \cdots \times 10^k}$$

où les d_i sont des chiffres avec d_1 non nul; c'est la représentation flottante de x à l'aide de la mantisse normalisée.

3.5.1 Arithmétique flottante

- La représentation de x par troncature en **notation flottante** à n chiffres, se définit par

$$\mathrm{fl}(x) \;=\; 0, d_1 d_2 d_3 \cdots d_n \times 10^k \, .$$

- La représentation de x par arrondi en **notation flottante** à n chiffres, se définit par

$$\mathrm{fl}(x) \;=\; \begin{cases} 0, d_1 d_2 d_3 \cdots d_n \times 10^k & \text{si} \quad 0 \le d_{n+1} \le 4 \\ (0, d_1 d_2 d_3 \cdots d_n \,+\, 10^{-n}) \times 10^k & \text{si} \quad 5 \le d_{n+1} \le 9\,. \end{cases}$$

3.5.2. Remarques : Troncature et Arrondi.

- La troncature à n chiffres d'un nombre décimal positif $x = 0, d_1 d_2 d_3 \cdots d_n d_{n+1} \cdots \times 10^k$ en notation flottante consiste à retrancher les chiffres à partir de la position $n+1$. Donc on a toujours $\mathrm{fl}(x) \leq x$ et on dit que *la troncature est biaisée.*
  Dans ce cas, on a une troncature à 10^{-n+k} près.

  Noter que pour une précision donnée, on tronque un nombre décimal positif aux sous-multiples de l'unité.

- Quant à l'arrondi, il s'agit d'ajouter 5 au $(n+1)$-ième chiffre de la mantisse avant d'effectuer la troncature. En d'autres termes, *l'arrondi à n chiffres d'un nombre décimal x en notation flottante, s'obtient par la troncature du nombre obtenu en ajoutant à x le nombre $0,5 \times 10^{-n}$ ou encore $5 \times 10^{-n-1}$.* L'arrondi vérifie $\mathrm{fl}(x) \leq x$ ou $\mathrm{fl}(x) \geq x$, et on dit qu'il est *non biaisé.* Dans ce cas, on a un arrondi à 10^{-n+k} près.

  On peut aussi arrondir un nombre réel positif pour obtenir une valeur approchée décimale ou entière!

- En pratique, étant donné un entier relatif n ($n \in \mathbb{Z}$),

  – la **troncature** d'un nombre décimal positif x à 10^n près, est le nombre décimal de la forme $m \times 10^n$ avec $m \in \mathbb{N}$ et tel que:

  $$m \times 10^n \leq x < (m+1) \times 10^n.$$

  – l'**arrondi** d'un nombre réel positif x à 10^n près, est le nombre décimal de la forme $m \times 10^n$ avec $m \in \mathbb{N}$ et tel que:

  $$x - 0,5 \times 10^n < m \times 10^n \leq x + 0,5 \times 10^n .$$

  Il peut être plus grand que x contrairement à la valeur obtenue par troncature!

  Noter que l'arrondi d'un nombre réel positif x à 10^n près, est le plus grand nombre décimal de la forme $m \times 10^n$; avec $m \in \mathbb{N}$, qui est à distance minimale de x.

  Remarque. En informatique, l'arrondi d'un nombre négatif diffère cependant selon les logiciels utillisés. Généralement, pour arrondir un nombre négatif, on prend l'opposé de l'arrondi de sa valeur absolue.

3.5.3. Exemples

- La troncature de $3,14159$ à l'unité près; c'est-à-dire à 10^0 près, est : 3.
- La troncature de $3,14159$ à 10^{-2} près est : $3,14$.
- La troncature de $3,14159$ à 10^{-4} près est : $3,1415$.
- L'arrondi de $3,14159$ à 10^{-2} près est : $3,14$.
- L'arrondi de $3,1415$ à 10^{-3} près est : $3,142$. C'est le plus grand des seuls nombres 3141×10^{-3} et 3142×10^{-3} qui sont à distance minimale de $3,1415$.
- L'arrondi de $3,14159$ à l'unité près; c'est-à-dire à 10^0 près, est : 3.
- L'arrondi de 85 à la dizaine près; c'est-à-dire à 10^1 près, est : 90.

- L'arrondi de 185 à la centaine près; c'est-à-dire à 10^1 près, est : 200.
- L'arrondi de 2021 à l'unité de mille près; c'est-à-dire à 10^3 près, est : 2000.

L'on peut définir l'arrondi d'un nombre réel quelconque dans n'importe quelle base appropriée b (base dans laquelle tout nombre réel possède un développement éventuellement illimité).
En effet l'arrondi d'un nombre réel positif x dans une base $b \geq 2$ (e.g; $b = 10$ ou $b = 2$) avec une certaine précision b^{-n} ($n \in \mathbb{N}$) est le nombre

$$e_m \times b^m + e_{m-1} \times b^{m-1} + \ldots + e_0 + d_1 \times b^{-1} + \ldots + d_n \times b^{-n}$$

le plus proche de x pour lequel tous les chiffres correspondant aux puissances b^{-k} allant en dessous de cette précision; c'est-à-dire $-k < -n$ ou encore $k < n$, sont nuls.
Par convention, lorsqu'il existe deux de ces nombres plus proches possible (à distance minimale), l'arrondi est alors le plus grand.

La norme IEEE recommande l'utilisation de l'arrondi dans la représentation binaire des nombres réels.

Questions : On choisit $n = 4$ (i.e., la représentation de la mantisse normalisée avec 4 chiffres) dans le système décimal.
Trouver alors

$$\begin{aligned} \mathrm{fl}(1/3) &= \ldots \\ \mathrm{fl}(2,0166) &= \ldots \\ \mathrm{fl}(12,4551) &= \ldots \\ \mathrm{fl}(\pi) &= \ldots . \end{aligned}$$

3.5.3. Chiffres significatifs

Si l'erreur absolue commise dans la représentation (ou l'approximation) d'un nombre réel x vérifie

$$\Delta x \ \leq \ 0,5 \times 10^k \qquad \text{où } k \text{ est un entier relatif;}$$

alors le chiffre correspondant à la k-ième puissance de 10 est dit **significatif** *et tous ceux qui sont à sa gauche (correspondant aux puissances de 10 supérieures à k) le sont aussi à l'exception des zéros qui ne sont précédés d'aucun chiffre non nul.*

Questions :

- ⋆ En faisant l'approximation de π au moyen de la quantité $^{22}/_7$, quelle erreur absolue commet-on? Quels sont les chiffres significatifs?
- ⋆ En retenant comme approximation de π, le nombre $3,1416$, quels sont les chiffres significatifs?

Nous reviendront dans la suite sur la notion des chiffres significatifs et la méthode courante de détermination de l'arrondi.

3.5.4. Précision machine

La **précision machine** ε est la plus grande erreur relative que l'on puisse commettre en représentant un

nombre réel sur ordinateur en utilisant la troncature.
Si on utilise l'arrondi alors la précision machine vaut $\varepsilon/2$.

Question : Justifier que la précisionn machine vérifie

$$\varepsilon \ \leq \ b^{1-n}$$

où b est la base utilisée et n le nombre de bits de la mantisse.

3.6. Opérations élémentaires en Arithmétique Flottante

On rappelle que les opérations élémentaires sont l'addition, la soustraction, la multiplication et la division.
Pour effectuer une opération élémentaire en Arithmétique Flottante, on doit d'abord représenter les deux opérandes en notation flottante, effectuer l'opération de la façon habituelle et puis exprimer le résultat en notation flottante.
En d'autres termes (symboliques), on a:

$$\begin{aligned} x+y \quad &\rightarrow \quad \mathrm{fl}\Big(\mathrm{fl}(x)+\mathrm{fl}(y)\Big) \\ x-y \quad &\rightarrow \quad \mathrm{fl}\Big(\mathrm{fl}(x)-\mathrm{fl}(y)\Big) \\ x\div y \quad &\rightarrow \quad \mathrm{fl}\Big(\mathrm{fl}(x)\div\mathrm{fl}(y)\Big) \\ x\times y \quad &\rightarrow \quad \mathrm{fl}\Big(\mathrm{fl}(x)\times\mathrm{fl}(y)\Big)\,. \end{aligned}$$

Questions : Avec $n=4$, effectuer les opérations suivantes en notation flottante.

a) $(1/3)\times 3\,.$

b) $\big(0,4035\times 10^{6}\big)\times\big(0,1978\times 10^{-1}\big)\ .$

c) $\big(0,56789\times 10^{4}\big)\div\big(0,1234321\times 10^{-3}\big)\ .$

d) $\big(0,4035\times 10^{6}\big)+\big(0,1978\times 10^{4}\big)\ .$

e) $\big(0,56789\times 10^{4}\big)-\big(0,1234321\times 10^{6}\big)\ .$

Avertissements : Des opérations mathématiquement équivalentes ne le sont pas forcément en Arithmétique Flottante.

1) La propriété de distributivité de la multiplication sur l'addition n'est pas toujours respectée en Arithmétique Flottante.
   Par exemple, avec $n=3$,

$$\begin{aligned} 122\times(333+695) \quad &\rightarrow \quad 0,126\times 10^{6}\,. \\ (122\times 333)\ +\ (122\times 695) \quad &\rightarrow \quad 0,125\times 10^{6}\,. \end{aligned}$$

2) La propriété d'associativité de l'addition n'est pas toujours respectée en Arithmétique Flottante.
   Par exemple, avec $n=3$,

$$\begin{aligned} 2+(0,004+0,003) \quad &\rightarrow \quad 0,2\times 10^{1} \quad = \quad 2\,. \\ (2+0,004)+0,003 \quad &\rightarrow \quad 0,201\times 10^{1} \quad = \quad 2,01\,. \end{aligned}$$

3) Opération risquée: l'addition de deux nombres dont les ordres de grandeur sont très différent peut entraîner la disparution complète du plus petit nombre devant le plus grand.

Par exemple, avec $n = 4$,

$$\begin{aligned} \left(0,2016 \times 10^5\right) \, + \, \left(0,1000 \times 10^{-2}\right) \quad &\rightarrow \quad \text{fl}\left(0,2016 \times 10^5 \, + \, 0,00000001 \times 10^5\right) \\ &= \quad \text{fl}\left(0,20160001 \times 10^5\right) \\ &= \quad 0,2016 \times 10^5 \, . \end{aligned}$$

4) Opération risquée: la soustraction de deux nombres presque identiques peut faire apparaître des chiffres non significatifs!

Par exemple, avec $n = 4$, le calcul en notation flottante de

$$\begin{aligned} \left(0,2016 \times 10^6\right) \, - \, \left(0,2015 \times 10^6\right) \quad &\rightarrow \quad \text{fl}\left(0,0001 \times 10^6\right) \\ &= \quad 0,1000 \times 10^3 \end{aligned}$$

fait apparaître dans $0,0001 \times 10^6$ trois 0 non significatifs.

IV. Rappels sur les chiffres significatifs

On rappelle que la notion de chiffres significatifs est une convention universellement adoptée pour représenter une valeur approchée décimale d'un nombre ou de la mesure d'une grandeur (physique) avec une erreur implicitement associée. Cette convention simplifie la notion d'incertitude de mesure qui exprime explicitement l'erreur ou une majoration de l'erreur.

Le principe fondamental de la notion de chiffres significatifs est que dans l'écriture scientifique d'une valeur approchée, sans mention explicite de précision, seul le dernier chiffre peut être incertain. Par exemple dans une table de données, une valeur comme $12,94$ traduit que la valeur exacte de la grandeur est comprise entre $12,94 - 0,01 = 12,93$ et $12,94 + 0,01 = 12,95$; c'est-à-dire que l'incertitude est de $0,01$. Par contre la valeur approchée décimale $12,940$ traduit que la valeur exacte est comprise entre $12,940 - 0,001 = 12,939$ et $12,940 + 0,001 = 12,941$ (ici l'incertitude est de $0,001$). Donc en matière d'approximation, la valeur $12,940$ est plus précise que $12,94$ (bien que les nombres décimaux $12,940$ et $12,94$ soient parfaitement égaux). Par ailleurs, même en matière d'approximation les valeurs $012,94$ et $12,94$ ont la même précision puisqu'elles sont parfaitement égales.

De même, la valeur $12,00$ signifie que la valeur exacte est comprise entre $11,99$ et $12,01$ et est plus précise que la valeur $12,0$. Par ailleurs, notons que les valeurs $12,0$ et $012,0$ ont la même précision étant donné qu'elles sont égales.

- Dans un nombre décimal non nul :
  - Aucun zéro (0) précédant le premier chiffre non nul n'est significatif.
  - Les chiffres significatifs sont le premier chiffre non nul et tous les autres chiffres situés à sa droite (y compris le dernier zéro).
  - Le nombre de chiffres situés après le dernier zéro (0) non significatif représente le nombre de chiffres significatifs.
- Dans l'expression décimale d'une valeur approchée, si un certain chiffre est significatif, alors tous ceux qui sont à sa gauche sont aussi significatifs à l'exception des zéros qui ne sont précédés d'aucun chiffre non nul.
  Si l'erreur absolue commise dans la représentation (ou l'approximation) d'un nombre réel x vérifie
  $$\Delta x \ \leq \ 0,5 \times 10^k \qquad \text{où } k \text{ est un entier relatif;}$$
  alors le chiffre correspondant à la k-ième puissance de 10 est significatif et tous ceux qui sont à sa gauche (correspondant aux puissances de 10 supérieures à k) le sont aussi à l'exception des zéros qui ne sont précédés d'aucun chiffre non nul.
- En **notation scientifique**, on écrit une valeur approchée décimale sous la forme $m \cdot 10^k$ oú $m \in [0,\, 10[$ est la mantisse (ou la significande) dont tous les chiffres sont significatifs et k est un entier relatif.
- En **notation ingénieur**, on écrit une valeur approchée décimale sous la forme $m \cdot 10^{k\times 3}$ oú $m \in [0,\, 1000[$ est la mantisse (ou la significande) dont tous les chiffres sont significatifs et k est un entier relatif.

  Notons qu'en ingéniérie, on a les noms suivants pour les puissances de 1.000 :

Puissance de 10	10^{-24}	10^{-21}	10^{-18}	10^{-15}	10^{-12}	10^{-9}	10^{-6}	10^{-3}
Nom	yocto	zepto	atto	femto	pico	nano	micro	milli
Symbole	y	z	a	f	p	n	μ	m

Puissance de 10	10^3	10^6	10^9	10^{12}	10^{15}	10^{18}	10^{21}	10^{24}
Nom	Kilo	Mega	Giga	Tera	Peta	Exa	Zeta	Yolta
Symbole	k	M	G	T	P	E	Z	Y

Par exemple:

- 9 a un chiffre significatif tout comme les nombres $0,9$ et $0,09$.
- $0,90$ a 2 chiffres significatifs.
- $0,0103$ a 3 chiffres significatifs.
- $5,3 \times 10^9$ a 2 chiffres significatifs.

De plus, lorsque suite à la mesure d'une grandeur on trouve 510,

- si un seul chiffre est significatif, il ne peut qu'être 5 et on écrira alors le résultat final sous la forne 5×10^2 ou encore $0,5 \times 10^3$ (dans ce cas aucun des chiffres n'est exact),
- si seulement deux chiffres sont significatifs, ils ne peuvent qu'être 5 et 1, et on écrira alors le résultat final sous la forne $5,1 \times 10^2$ ou encore $0,51 \times 10^3$ (dans ce cas seul le chiffre 5 est surêment exact),
- si trois chiffres sont significatifs, alors on écrira le résultat final sous la forne $5,10 \times 10^2$ ou encore $0,510 \times 10^3$, ou encore 510 (dans ce cas les deux chiffres 5 et 1 sont surêment exacts) et l'incertitude est d'une unité.

Remarque
Compte tenu de certaines conventions, la notion (du nombre) de chiffres peut comporter des subtilités. C'est le cas des tableaux de logarithmes dans lesquels la règle consiste à avoir autant que possible de chiffres significatifs après la virgule dans le logarithme (obtenu par approximation) que dans la valeur (dont on calcule le logarithme). Ceci est en réalité une conséquence du nombre de chiffres significatifs défini dans la représentation scientifique d'un nombre décimal, puisque dans le logarithme, le nombre avant la virgule n'est rien d'autre que la valeur de l'exposant.
Par exemple $4,1 \times 10^3$ a deux chiffres et on convient de dire que son logarithme décimal avec 2 chiffres significatifs est $3,61$ et non $3,6$.

Questions

1. Donner le nombre de chiffres significatifs de chacun des nombres décimaux suivants:

$$12,0. \quad 0,12. \quad 0,120. \quad 0,012. \quad 12,00. \quad 0,0012. \quad 0,120. \quad 12,001. \quad 0,02019. \quad 011,02019.$$

2. Donner le nombre de chiffres significatifs de chacun des nombres décimaux suivants:

$$3,5 \times 10^2. \quad 0,350 \times 10^2. \quad 3,50 \times 10^3. \quad 0,014 \times 10^{-3}.$$

3. La mesure d'une grandeur donne $19,47$ comme résultat brut avec une incertitude égale à $0,50$.
   i) Dans cette valeur approchée décimale, combien y a-t-il de chiffre(s) exact(s)?
   (On pourra d'abord faire un encadrement).
   ii) Dans l'expression décimale du résultat final, combien de chiffres significatifs a-t-on (au maximum)?
   iii) Donner le résultat final (sans avoir à préciser l'incertitude).

4. La mesure d'une grandeur donne $19,47$ comme résultat brut avec une incertitude égale à $0,05$.

   i) Dans cette valeur approchée décimale, combien y a-t-il de chiffres exacts?

   ii) Dans l'expression décimale du résultat final, combien de chiffres significatifs a-t-on (au maximum)?

   iii) Donner le résultat final (sans avoir à préciser l'incertitude).

V. Erreurs absolue et relative

5.1 Erreur absolue

Soit x un nombre réel (ou une grandeur exacte). Pour une valeur approchée x^* de ce nombre, l'erreur absolue est définie par $\Delta x = |x - x^*|$.

5.2 Erreur relative

Soit x un nombre réel non nul (ou une grandeur exacte non nulle). Pour une valeur approchée x^* de ce nombre, l'erreur absolue est définie par

$$E_r(x) \;=\; \frac{|x - x^*|}{|x|} \;=\; \frac{\Delta x}{|x|}\,.$$

L' erreur relative en pourcentage s'obtient en multipliant l'erreur relative par 100%.

Remarque: En pratique on a autant de difficulté à trouver la valeur exacte d'un nombre qu'à trouver l'erreur absolue (ou relative) de son approximation par un nombre x^*. C'est l'exemple des grandeurs mésurées dont on ne dispose que de valeurs approximatives rendant impossible la détermination des erreurs absolues. Dans ce cas, on essaie de trouver une borne supérieure (raisonnable) Δx de $|x - x^*|$, si bien qu'on a

$$|x - x^*| \;\leq\; \Delta x$$

ce qui équivaut à

$$x^* \,-\, \Delta x \;\leq\; x \;\leq\; x^* \,+\, \Delta x\,.$$

On interprète cela en disant que x^* est une estimation de la valeur exacte x avec une incertitude de Δx de part et d'autre (i.e., par excès ou par défaut).
Noter que l'erreur absolue indique la mesure quantitative commise tandis que l'erreur relative en mesure l'importance (à travers le pourcentage).

Activité II.

1. Donner une approximation du nombre π^e par troncature avec deux chiffres après la virgule.
2. Donner aussi une approximation de ce nombre par arrondi avec deux chiffres après la virgule.
3. Quelle est la valeur approchée décimale de π^e avec 4 chiffres significatifs.

Activité III.
Les mesures d'une feuille A4 donnent respectivement $L = 29,7\,\text{cm}$ pour la longueur et $l = 21,0\,\text{cm}$ avec une précision de l'ordre du millimètre ($\Delta L = \Delta l = 0,1\,\text{cm}$).

1. Calculer alors le périmètre de cette feuille en précisant l'erreur absolue.
2. Calculer aussi l'aire de cette feuille en précisant l'erreur absolue.

Activité IV.
La mesure de la longueur d'un côté d'une boîte cubique $\mathcal{B}_1$ donne $l^* = 10,2\,\text{cm}$ avec une précision de l'ordre du millimètre ($\Delta l = 0,1\,\text{cm}$).

1. Calculer la valeur approximative du volume de cette boîte et l'erreur liée à cette valeur.
2. Quels sont les chiffres significatifs de cette valeur approximative du volume de cette boîte.

Activité V.
Une fonction numérique f définie d'un intervalle non vide I vers $\mathbb{R}$ possède un zéro a de multiplicité $k \in \mathbb{N}^*$, s'il existe une fonction $g : I \longrightarrow \mathbb{R}$ telle que

$$\begin{cases} f(x) = (x-a)^k g(x) \quad \forall x \in I, \\ g(a) \neq 0. \end{cases}$$

Dans le cas où f est k fois dérivable en a, cela signifie que

$$\begin{cases} f^{(i)}(a) = 0 \quad \text{pour tout} \quad 0 \leq i \leq k-1, \\ f^{(k)}(a) \neq 0. \end{cases}$$

1. On considère le polynôme $P(x) = x(x-1)(x^2-1)(3x+1)^4$.
   i) Quels sont les zéros de P ? Préciser l'ordre de multiplicité de chacun d'un.
   ii) Parmi ces zéros, lesquels sont simples ?
2. Montrer que π est un zéro simple de la fonction sinus.
3. 0 est-il un zéro simple de la fonction numérique $f : x \mapsto 1 - \cos x$?
   Sinon quel est son multiplicité ?

Activité VI.
On pose $f(x) = x^2 - 2$ et on s'interesse à l'approximation de la racine positive α de f.

1. Montrer, sans chercher à déterminer α, que $\alpha \in]1, 2[$.
2. En utilisant la méthode de la bissection (encore dite de dichotomie), trouver une valeur approchée de α à 10^{-2} près.
   Vérifier le résultat à l'aide d'une calculatrice (par résolution graphique ou autrement).

Activité VII.

1. Montrer que l'équation $xe^x - 1 = 0$ admet une seule solution $\bar{x}$ dans $\mathbb{R}$.
2. Montrer que $\bar{x} \in]0, 1[$.
3. En utilisant la méthode de la dichotomie, trouver une valeur approchée de $\bar{x}$ à 10^{-2} près.

Activité VIII.

1. Montrer que l'équation $x + \ln x = 0$ admet une seule solution β dans $]0, +\infty[$.
2. Montrer que $\beta \in]0, 1[$.
3. En utilisant la méthode de la dichotomie, trouver une valeur approchée de β à 10^{-2} près.
4. Vérifier le résultat à l'aide de votre ordinateur (par résolution graphique ou autrement).

Activité IX.
Soit f une fonction réelle continûment dérivable (i.e., dérivable et de fonction dérivée continue) sur un intervalle I d'intérieur non vide.
On suppose que f' n'est jamais nulle dans I et on choisit un certain élément $x_0 \in I$.

1. Montrer que l'abcisse x_1 du point d'intersection de l'axe des abcisses et de la tangente à la courbe représentative de f au point d'abcisse x_0 est:
$$x_1 \;=\; x_0 - \frac{f(x_0)}{f'(x_0)}\,.$$
2. Etant donné (ou assuré) que $x_n \in I$, pour $n \in \mathbb{N}$, on pose
$$x_{n+1} \;=\; x_n - \frac{f(x_n)}{f'(x_n)}\,.$$
C'est l'algorithme de Newton !
3. On suppose de plus que tous les termes x_n (avec $n \in \mathbb{N}$) sont définis et que la suite $(x_n)_{n\in\mathbb{N}}$ converge vers un réel x^* dans I.
Que représente x^* pour f ?
4. Quelle est l'utilité de l'algorithme de Newton ?

Activité X.

1. Ecrire l'algorithme de Newton dans le cas où $f(x) = x^2 - 2$ pour $x \in [1,\, 2]$ et $x_0 = 1$.
2. Construire le graphe de f puis représenter (sans calcul) les termes x_1 , x_2.
3. Calculer x_2.
4. Pour $n \in \mathbb{N}^*$, que peut-on dire de x_n pour f ?
5. Que représente x_2 pour $\sqrt{2}$?

Activité XI.

1. Soit I un intervalle non vide et $f : I \to \mathbb{R}$ une fonction numérique.
Montrer que f est une fonction de Lipschitz si et seulement si ses taux de variation moyens sont bornés; c'est-à-dire, qu'il existe une constante réelle $K \geq 0$ tel que
$$\left|\frac{f(x) - f(y)}{x - y}\right| \leq K \quad \text{pour tous éléments distincts } x \text{ et } y \text{ de } I\,.$$
2. Montrer que les fonctions suivantes sont de Lipschitz:
   i) Toute fonction affine f de $\mathbb{R}$; c'est-à-dire que $f(x) = ax + b$ où a et b sont des constantes réelles.
   Précisez le rapport de Lipschitz?
   ii) La fonction numérique $g : [0,\, 1] \to \mathbb{R}$ définie par $g(x) = x^2$.
   iii) La fonction numérique $h : [-1,\, 1] \to \mathbb{R}$ définie par $h(x) = \frac{1}{2+x}$.

iv) La fonction numérique $u : [1, 4] \to [0, 4]$ définie par $u(x) = \sqrt{x}$.

3. Montrer que les fonctions suivantes ne sont pas de Lipschitz:

   i) La fonction numérique $\varphi : [0, 1] \to [0, 1]$ définie par $\varphi(x) = \sqrt{x}$.
   On pourra raisonner par l'absurde tout en considérant $\lim\limits_{x\to 0^+} \frac{\varphi(x)-\varphi(0)}{x-0}$.

   ii) La fonction numérique $\psi : \mathbb{R} \to \mathbb{R}$ définie par $\psi(x) = x^2$.
   On pourra raisonner par l'absurde tout en considérant $\lim\limits_{x\to+\infty} \frac{\psi(x)-\psi(0)}{x-0}$.

4. Quelles leçons tirez-vous des réponses aux questions 2.ii), 2.iv), 3.i), 3.ii).

<u>Activité</u> XII.
Soient a et b deux nombres réels tels que $a < b$.

1. i) Montrer que si $f : [a, b] \longrightarrow \mathbb{R}$ est continûment dérivable, alors f est une application de Lipschitz dont on précisera la constante de Lipschitz.
   On pourra appliquer le Théorème (l'inégalité) des accroissements finis.

   ii) En déduire que les fonctions suivantes sont de Lipschitz.
   - La fonction $f : [0, 2\pi] \to \mathbb{R}$ définie par $f(x) = \sin x$.
   - La fonction $g : [0, 1] \to \mathbb{R}$ définie par $g(x) = \ln(1 + x)$.

2. i) Soit $g : [a, b] \longrightarrow [a, b]$ une application continûment dérivable. On pose

$$||g'||_\infty = \max_{a\le x\le b} |g'(x)| \qquad \text{ou simplement} \quad K = \max_{a\le x\le b} |g'(x)| .$$

   Montrer que g est une contraction stricte si et seulement si $||g'||_\infty < 1$.

   ii) Soit I un intervalle ouvert non vide et $g : I \longrightarrow I$ une application continûment dérivable. On pose

$$K = \sup_{x\in I} |g'(x)| .$$

   Montrer que g est une contraction stricte si et seulement si $K < 1$.
   En déduire que la fonction $f : [0, +\infty[\to [0, +\infty[,\ x \mapsto \frac{\ln(1+x)}{2}$ est une contraction stricte.

3. On considère les applications suivantes:

$$\begin{array}{cccc} g_1 : & [0,1] & \to & [0,1] \\ & x & \mapsto & 1-x \end{array} \qquad \begin{array}{cccc} g_2 : & [0,1] & \to & [0,1] \\ & x & \mapsto & \sqrt{x} \end{array} \qquad \begin{array}{cccc} g_3 : & [-1,1] & \to & [-1,1] \\ & x & \mapsto & x^3 \end{array}$$

$$\begin{array}{cccc} g_4 : & [0,1] & \to & [0,1] \\ & x & \mapsto & \frac{1+x}{2} \end{array} \qquad \begin{array}{cccc} g_5 : & \mathbb{R} & \to & \mathbb{R} \\ & x & \mapsto & x+1 \end{array} \qquad \begin{array}{cccc} g_6 : & [-1,1] & \to & [-1,1] \\ & x & \mapsto & \sqrt{1-x^2} \end{array}$$

$$\begin{array}{cccc} g_7 : &]0,1[& \to &]0,1[\\ & x & \mapsto & \frac{x}{2} \end{array} \qquad \begin{array}{cccc} g_8 : & \mathbb{R} & \to & \mathbb{R} \\ & x & \mapsto & 2x \end{array} \qquad \begin{array}{cccc} g_9 : & [0,1] & \to & [0,1] \\ & x & \mapsto & x^2 \end{array}$$

   i) Lesquelles des applications ci-dessus possèdent de point(s) fixe(s) ?

   ii) Lesquelles des (neuf) applications ci-dessus sont des contractions strictes ?

   iii) Quelles leçons tirez-vous des deux réponses précédentes ?

Activité XIII.
Principe de l'application contractante (ou Théorème du point fixe de Banach)

Soient $a < b$ deux nombres réels et g une contraction stricte de l'intervalle férmé (et borné) $[a, b]$. Alors g admet un point fixe unique x^* dans $[a, b]$.
*De plus on a l'***algorithme du point fixe suivant***:*
Tout point $p_0 \in [a, b]$ appartient au bassin d'attraction de x^* en ce sens que la suite récurrente définie par

$$\begin{aligned} x_0 &= p_0 \\ x_{n+1} &= g(x_n)\,, \quad \forall n = 0,\, 1,\, 2,\, 3,\, \ldots\,. \end{aligned}$$

converge vers x^*.
En outre

$$|x_n - x^*| \;\leq\; \frac{K^n}{1-K}|x_1 - x_0| \qquad \forall\, n \geq 1$$

où K est une constante de Lipschitz de g appartenant à $[0, 1[$.
NB. Un tel K existe bien car g est par hypothèse une contractipon stricte.

Application
On pose

$$g(x) \;=\; \frac{x}{2} + \frac{1}{x}\,, \quad \forall\, x \in [1,\, 2]\,.$$

1. i) Vérifier que g est dérivable et que $g(x) \in [0,\, 1]$ pour tout $x \in [1,\, 2]$.
   ii) Montrer que g est une contraction stricte de $[1,\, 2]$.
2. Définir la suite $(x_n)_n$ de termes générés par l'algorithme du point fixe de g en prenant $x_0 = 1$.
3. Représenter le graphe de g et puis trouver graphiquement (sans calcul) les termes x_1 et x_2.
4. Montrer que
$$\left|x_n - \sqrt{2}\right| \;\leq\; \frac{1}{2^{n-1}}\,, \quad \forall n \geq 0\,.$$
5. Quelle est la plus petite valeur q de n, á partir de laquelle x_n est une approximation de $\sqrt{2}$ à 10^{-3} près?
6. Comparer cet algorithme à l'algorithme de Newton pour la détermination du zéro de $x^2 - 2$ dans $[1,\, 2]$.

Activité XIV.
On considère la fonction numérique f définie sur $\mathbb{R}$ par

$$f(x) \;=\; x^3 + 2x - 1\,.$$

1. Montrer que l'équation algébrique
$$f(x) \;=\; 0 \tag{E_1}$$
admet une unique solution réelle γ et que de plus $0 < \gamma < 1$.

   Maintenant on cherche à résoudre l'équation algébrique $f(x) = 0$ par un algorithme (itératif) du point fixe.

2. On pose
$$g(x) \;=\; \frac{1-x^3}{2}\,, \qquad x \in \mathbb{R}\,.$$

   i) Montrer que la racine réelle de l'équation $f(x)=0$ est le seul point fixe de g et que g réalise une contraction stricte de $\left[0,\,\frac{1}{2}\right]$.
   En déduire que $\left[0,\,\frac{1}{2}\right]$ est contenu dans le bassin d'attraction du point fixe de g.
   Pour tout $x_0 \in \left[0,\,\frac{1}{2}\right]$*, on analysera le comportement de la suite* $(x_n)_{n\in\mathbb{N}}$ *telle que* $x_{n+1} = g(x_n)$.

   ii) Montrer que $g([0,1]) \subset \left[0,\,\frac{1}{2}\right]$.
   En déduire que $[0,1]$ est contenu dans le bassin d'attraction du point fixe de g.

3. On s'intéresse aussi aux fonctions ci-dessous dont les points fixes coïncident avec la solution γ de (E_1) :
$$g_1(x) \;=\; -x^3 - x + 1\,, \quad g_2(x) \;=\; \frac{-x^3+1}{2} \quad \text{et} \quad g_3(x) \;=\; \frac{1}{x^2+2}\;; \qquad 0 \le x \le 1\,.$$

   Pour quelle(s) fonction(s) ci-dessus, suggérez-vous l'algorithme du point fixe pour la détermination de γ ?

   Justifier brièvement votre réponse.

<u>Activité</u> XV.
On considère les applications définies respectivement de [0, 1] vers [0, 1] par:
$$f(x) \;=\; \frac{x}{2} \quad \text{et} \qquad g(x) \;=\; \frac{1-x}{3}\,.$$

1. Déterminer respectivement l'ensemble des points fixes $\mathrm{Inv}(f)$ et $\mathrm{Inv}(g)$ respectifs des fonctions f et g.

2. On considère la suite recurrente définie par:
$$\left\{\begin{array}{rcl} x_0 & = & 1 \\ x_{n+1} & = & f(x_n) \quad \text{pour tout } n \in \mathbb{N}. \end{array}\right.$$

   i) Calculer x_1, x_2 et plus généralement x_n pour tout entier naturel n.

   ii) Quelle est la limite de la suite $(x_n)_{n\in\mathbb{N}}$?

   iii) Le résultat était-il prévisible ?

3. On considère la suite recurrente définie par:
$$\left\{\begin{array}{rcl} y_0 & = & 0 \\ y_{n+1} & = & g(y_n) \quad \text{pour tout } n \in \mathbb{N}\,, \end{array}\right.$$
   et on pose $z_n = y_n - \frac{1}{4}$ pour tout entier naturel n.

   i) Calculer z_0 et puis z_{n+1} en fonction de z_n pour tout entier naturel n.
   En déduire z_n et puis y_n pour tout entier naturel n.

   ii) Quelle est la limite de la suite $(y_n)_{n\in\mathbb{N}}$?
   Le résultat était-il prévisible ?