# MT 360 : Méthodes numériques et modèles déterministes

**Représentation des nombres en machine, erreur d'arrondi, stabilité et conditionnement numériques**
Séance de TD des 5 et 6 février 2025

**Exercice 1**
On considère une représentation en virgule flottante normalisée en base $\beta = 7$, avec $t = 5$ digits dans la mantisse et un exposant $e$ de 4 bits.

1. Quels sont les plus grand exposant, $e_M$, et plus petite exposant, $e_m$, dans cette représentation?

2. Calculer la valeur , $x_m$, du plus petit nombre machine en valeur absolue, et la valeur $x_M$, du plus grand nombre machine en valeur absolue

3. Calculer le nombre de nombres réels que l'on peut représenter exactement, c'est-à-dire le nombre de nombres machine, $\#(\mathbb{M})$

4. Calculer la précision machine, notée $h$

5. Déterminer les distances entre 2 nombres machines consécutifs d'exposant $e$ pour $e \in [e_m; e_M]$

6. Donner les représentations dans cette arithmétique en virgule flottante des nombres réels $\frac{1}{3}$ et $-\frac{4}{7}$ (exposants et mantisses), en utilisant l'approximation par troncature (technique dite de "*chopping*" ). Calculer les erreurs d'arrondis correspondantes, absolues et relatives

**Exercice 2** Cet exercice est un exercice d'application des notions d'erreurs engendrées et propagées, de conditionnement et de stabilité numérique. On se propose de calculer la plus grande racine réelle $x_M$ de l'équation du second ordre :

$$x^2 + 2px - q = 0 \tag{1}$$

avec $p, q \in \mathbb{R}$ et $q \geq 0$.

1. Calculer l'erreur propagée depuis les données (erreur inévitable) correspondant au problème (1) (i.e. le calcul de $x_M$). Y-a-t-il des valeurs des données $p$ et $q$ pour lesquelles ce problème est mal conditionné?

2. Proposer un algorithme de calcul direct de la solution $x_M$ du problème (1) et établir le graphe de calcul correspondant à cet algorithme. Le calcul direct se fonde sur l'évaluation de l'expression exacte

$$x_M := -p + \sqrt{p^2 + q}$$

3. En déduire l'expression de l'erreur relative totale sur le résultat. Préciser les parties de cette erreur correspondantes à l'erreur inévitable et à l'erreur propre à l'algorithme. Pour quels valeurs de $p$ et $q$ l'algorithme est-il instable?

4. Pour les valeurs des paramètres $p$ et $q$ pour lesquelles l'algorithme proposé est instable, proposer un autre algorithme (direct) et montrer que celui-ci est numériquement stable

**Exercice 3**

1. Soit $f : \mathbb{R} \to \mathbb{R} : x \mapsto f(x)$ une application de classe $\mathcal{C}^2$. Soit $x_0 \in \mathbb{R}^*$ tel que $f(x_0) \neq 0$. Soit $\bar{x}_0$ une valeur approchée de $x_0$ telle que $\bar{x}_0 = x_0(1 + \rho_0)$. Montrer, en utilisant le développement de Taylor[1] de $f$ au voisinage de $x_0$, que l'erreur relative sur $f(x_0)$ qui résulte de cette erreur sur $x_0$ est donnée par :

$$\rho(f(x_0)) = \frac{x_0 f'(x_0)}{f(x_0)} \rho_0 + \mathcal{O}(\rho_0^2) \tag{2}$$

2. Soient les applications numériques :

$$f_1(x_0) := (x_0 - 1)^6 \; ; \; f_2(x_0) := \left(\frac{1}{x_0 + 1}\right)^6 \; ; \; f_3(x_0) := 99 - 70x_0$$

Montrer que ces applications ont la même valeur exacte

$$f_1(x_0) = f_2(x_0) = f_3(x_0)$$

pour $x_0 = \sqrt{2}$.

[1] Le développement de Taylor de la fonction $f$ au voisinage du point $x_0 \in \mathbb{R}$ s'écrit :

$$f(x) = f(x_0) + (x - x_0) f'(x_0) + \mathcal{O}((x - x_0)^2)$$

3. Indiquer, pour chacune des fonctions $f_i(x_0 = \sqrt{2})$ du point précédent, la précision minimum nécessaire (nombre de chiffres significatifs ou erreur relative) sur $x_0$ pour représenter $f(x_0)$ à la précision machine (en double précision) . Utilisez pour montrer ces résultats les coefficients de propagation des fonctions $f_i$ tels que défini en (2). Quelle est de ce point de vue la formule la plus intéressante?

4. Calculer, à l'aide d'un graphe de calcul, pour chacune des fonctions $f_i(x_0)$ du point 2, l'erreur propre à l'algorithme associé. Comparer chacune de ces erreurs à l'erreur inévitable correspondante et discuter (conditionnement, stabilité numérique) en fonction des valeurs de $x_0$.

**Exercice 4**
On considère l'aplication :

$$f : \mathbb{R}^+ \to \mathbb{R} : x \mapsto f(x) := \sqrt{x+1} - \sqrt{x}$$

1. Evaluer le conditionnement de $f(x)$ pour $x \in \mathbb{R}^+$

2. Proposer un algorithme naïf $f^*(x)$ pour calculer $f(x)$ reposant sur l'évaluation littérale de l'expression (3) en appliquant les régles usuelles du calcul arithmétique dans $\mathbb{R}$

3. Evaluer $f^*(12345)$ en simple et en double précision. Conclure sur le conditionnement de $f(x)$ et la stabilité numérique de $f^*(x)$ pour $x = 12345$. Quelle est l'origine de l'instabilité numérique de l'algorithme $f^*(x)$ pour $x = 12345$

4. Proposer un algorithme alternatif $\tilde{f}(x)$ numériquement stable pour $x = 12345$. Evaluer l'erreur propre à l'algorithme $\tilde{f}$ et comparer la à l'erreur inévitable, pour $x = 12345$

**Exercice 5**
On considère l'aplication :

$$f : \mathbb{R}^+ \to \mathbb{R} : x \mapsto f(x) := (\cos x) \cdot e^{10x^2}$$

1. Calculer le coefficient de propagation de la fonction $f$ pour $x \in \left]-\frac{\pi}{2}; +\frac{\pi}{2}\right[$

2. Donner la précision nécessaire sur $x$ (i.e. une borne supérieure pour $\rho(x)$) afin de pouvoir obtenir une valeur $f(x)$ à la précision machine (en double précision)

3. Quelles sont les valeurs de $x$ pour lesquelles le problème $f(x)$ est mal conditionné?

**Exercice 6**
On travaille dans une arithmétique $FP(10, 4, cl)$ avec des exposants de 4 bits. On considère dans cette arithmétique la suite finie de nombres machines :

$x_1 := 0.1580$ ; $x_2 := 0.2653$ ; $x_3 := 0.2581 \cdot 10^1$ ; $x_4 := 0.2488 \cdot 10^1$ ;
$x_5 := 0.6266 \cdot 10^2$ ; $x_6 := 0.7555 \cdot 10^2$ ; $x_7 := 0.7889 \cdot 10^3$ ;
$x_8 := 0.7767 \cdot 10^3$ ; $x_9 := 0.8999 \cdot 10^4$

1. Calculer la somme de ces nombres par ordre croissant
2. Calculer la somme de ces nombres par ordre décroissant
3. Comparer les résultats et commenter

**Exercice 7**
On considère la suite définie par :

$$x_{n+1} := \frac{37}{6} x_n - x_{n-1} \ , \forall n \geq 1$$

avec $x_0 := 1$ et $x_1 := \frac{1}{6}$. Calculer les premières valeurs successives de $x_n$ calculées dans $FP(10, 4, cl)$ et dans $FP(10, 8, cl)$. Expliquer d'où provient la différence importante constatée (dès la cinquième étape).