REPRESENTATION DES NOMBRES

Exercice 1 – Représentation flottante double précision ✦✦✦

A) Convertir $a=(1,01010101)_{10}$ en représentation flottante IEEE$_{754}$ au format double précision.

L'excédent (qui permet de coder et fixer les bornes de l'exposant dans ce format) est :

$$E_0 = 2^{|E|-1}-1 = 2^{10}-1 = 1023$$

La formule de décodage IEEE$_{754}$ double précision est donc : $\mathrm{fl}(a)=(-1)^S(1+M)2^{E-1023}$

Comme $1<a<2$ il vient immédiatement que : $\begin{cases} S=0 \\ E=(1023)_{10}=\left(01\underbrace{\ldots 1}_{10}\right)_2 \end{cases}$

Reste à convertir $(0,01010101)_{10}$ en binaire pour déterminer les 52 chiffres de la mantisse.

Le problème pratique est de savoir comment calculer, en général, la décomposition comme somme de puissances entières relatives de deux, d'un rationnel quelconque.

$$(0,01010101)_{10}=\frac{1}{10^2}+\frac{1}{10^4}+\frac{1}{10^6}+\frac{1}{10^8}=\frac{1-\frac{1}{10^{10}}}{1-\frac{1}{10^2}}-1=\frac{10^8-1}{10^8\left(10^2-1\right)}$$

L'encadrement suivant $\frac{1}{2^6}>\frac{10^8-1}{10^8\left(10^2-1\right)}>\frac{1}{2^7}$

Permet la décomposition[1] $\frac{10^8-1}{10^8\left(10^2-1\right)}=\frac{1}{2^7}\left(1+\frac{29\times10^8-2^7}{10^8\left(10^2-1\right)}\right)$ avec $0<\frac{29\times10^8-2^7}{10^8\left(10^2-1\right)}<1$

$\frac{1}{2^1}>\frac{29\times10^8-128}{10^8\left(10^2-1\right)}\approx\frac{1}{3}>\frac{1}{2^2}$ et donc $\frac{29\times10^8-2^7}{10^8\left(10^2-1\right)}=\frac{1}{2^2}\left(1+\frac{17\times10^8-2^9}{10^8\left(10^2-1\right)}\right)$

[1] Voir en annexe de ce corrigé, quelques compléments théoriques et pratiques sur le développement de la mantisse d'un rationnel (toutes bases)

Pour gagner du temps (surtout sans calculatrice)

On remarque qu'à l'étape k, le reste est de la forme $\dfrac{\alpha_k 10^8 - 2^{\sum_{i=1}^{k} q_i}}{10^8\left(10^2-1\right)}$

Etant donné l'exposant q_{k+1} obtenu par encadrement de ce reste, il vient que :

$$\frac{\alpha_k 10^8 - 2^{\sum_{i=1}^{k} q_i}}{10^8\left(10^2-1\right)} = \frac{1}{2^{q_{k+1}}}\left(1+\frac{\left(2^{q_{k+1}}\alpha_k - 10^2+1\right)10^8 - 2^{\sum_{i=1}^{k+1} q_i}}{10^8\left(10^2-1\right)}\right)$$

C'est-à-dire : $\begin{cases} \alpha_{k+1} = 2^{q_{k+1}}\alpha_k - 10^2 + 1 \\ q_{k+1} = \min\left\{ q \mid q \in \mathbb{N}^* \wedge \left(2^q \alpha_k - 10^2+1\right)10^8 - 2^{q+\sum_{i=1}^{k} q_i} > 0 \right\} \end{cases}$

Notons par ailleurs que $2^{26} < 10^8 < 2^{27}$ et donc, pour $\sum_{i=1}^{k} q_i < 27$,

nous avons $\dfrac{\alpha_k 10^8 - 2^{\sum_{i=1}^{k} q_i}}{10^8\left(10^2-1\right)} \approx \dfrac{\alpha_k}{10^2}$, ce qui facilite les encadrements jusqu'à mi-parcours (26/52).

En itérant le procédé, on obtient :

$$(0,01010101)_{10} = \frac{1}{2^7}\left(1+\frac{1}{2^2}\left(1+\frac{1}{2^3}\left(1+\frac{1}{2^2}\left(1+\frac{1}{2^2}\left(1+\frac{1}{2^1}\left(1+\frac{1}{2^1}\left(1+\frac{1}{2^1}\left(1+\frac{1}{2^1}\left(1+\frac{1}{2^1}\left(1+\frac{1}{2^2}\left(1+\frac{1}{2^2}\left(1+\frac{1}{2^1}\left(1+\frac{1}{2^2}\left(1+\frac{1}{2^3}\left(1+\frac{1}{2^1}\left(1+\frac{1}{2^1}\left(1+\frac{1}{2^3}\left(1+\frac{1}{2^1}\left(1+\frac{1}{2^2}\left(1+\frac{1}{2^1}\left(1+\frac{1}{2^2}\left(1+\frac{1}{2^3}\left(1+\frac{1}{2^7}\left(1+\frac{65380096}{10^8}\right)$$

Et donc :

$$(0,01010101)_{10} = (0,M_{52})_2 + \left(\varepsilon < 2^{-52}\right)$$

$$M_{52} = \underbrace{0000001}_{7}\underbrace{01}_{2}\underbrace{001}_{3}\underbrace{01}_{2}\underbrace{01}_{2}\underbrace{11111}_{5\,fois 1}\underbrace{01}_{2}0110100111001101101001000000 1$$

B) Quelle est la plus petite valeur positive normalisée représentable au format IEEE$_{754}$ double précision ? Expliquez. On rappelle que l'exposant occupe 11 bits dans ce format.

La plus petite valeur **(strictement) positive <u>normalisée</u>** représentable au format IEEE$_{754}$ double précision est : $(-1)^0(1+0)2^{1-1023} = 2^{-1022}$

C) Y-a-t-il moyen de représenter une valeur encore plus petite ? Lequel ? Quel est alors la plus petite valeur représentable ?

Le format IEEE$_{754}$ double précision dénormalisé est $(-1)^0(0,M)2^{-1022}$ avec $M \neq 0$.

Dans ce format, le plus petit nombre strictement positif représentable est $2^{-52}2^{-1022} = 2^{-1074}$.

SYSTEMES D'EQUATIONS

Exercice 2 – Pivot de Gauss total ✦

A) Résoudre le système suivant par la méthode du pivot de Gauss total :

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 3 \\ 1 & 3 & 6 & 9 \\ 2 & 6 & 15 & 21 \end{array}\right)$$

$$\left(\begin{array}{ccc|c} 1 & 1 & 2 & 3 \\ 1 & 3 & 6 & 9 \\ 2 & 6 & 15 & 21 \end{array}\right) \begin{array}{l} L_1 \leftrightarrow L_3 \\ C_1 \leftrightarrow C_3 \\ \sigma = (1 \quad 3) \end{array}$$

$$\left(\begin{array}{ccc|c} 15 & 6 & 2 & 21 \\ 6 & 3 & 1 & 9 \\ 2 & 1 & 1 & 3 \end{array}\right) \begin{array}{l} L_2 \leftarrow L_2 - \frac{2}{5}L_1 \\ L_3 \leftarrow L_3 - \frac{2}{15}L_1 \end{array}$$

$$\left(\begin{array}{ccc|c} 15 & 2 & 6 & 21 \\ 0 & 3/5 & 1/5 & 3/5 \\ 0 & 1/5 & 11/15 & 1/5 \end{array}\right) \begin{array}{l} L_2 \leftrightarrow L_3 \\ C_2 \leftrightarrow C_3 \\ \sigma = (1 \quad 3)(2 \quad 3) \end{array}$$

$$\left(\begin{array}{ccc|c} 15 & 2 & 6 & 21 \\ 0 & 11/15 & 1/5 & 1/5 \\ 0 & 1/5 & 3/5 & 3/5 \end{array}\right) \quad L_3 \leftrightarrow L_3 - \frac{1}{11}L_2$$

$$\left(\begin{array}{ccc|c} 15 & 2 & 6 & 21 \\ 0 & 11/5 & 1/5 & 1/5 \\ 0 & 0 & 6/11 & 6/11 \end{array}\right) \Rightarrow \begin{cases} x_{\sigma_3} = x_2 = 1 \\ x_{\sigma_2} = x_1 = 0 \\ x_{\sigma_1} = x_3 = 1 \end{cases}$$

Notons que la validité de cette solution est assez facile à vérifier.

B) Est-il pertinent d'utiliser une méthode itérative pour résoudre ce système ? Justifiez votre réponse.

$\forall(i \quad j) \in [1 \quad 2] \times [1 \quad 3] \left(|A_{ii}| < \sum_{j \neq i} |A_{ij}| \right)$: la matrice n'est donc pas à diagonale strictement dominante (voir lemme d'Hadamard) il n'est donc pas certain que la méthode itérative converge. Par ailleurs, une méthode itérative, de complexité quadratique, devient plus intéressante que le pivot de Gauss, de complexité cubique, pour une matrice d'ordre supérieur à 100.

INTERPOLATIONS

Exercice 4 – Connaissances et réflexion théorique et pratique ✦✦✦

Soit n points (x_i, y_i) donnés sans ordre précis.

A) Les polynômes d'interpolation de Lagrange et de Newton de ces n points diffèrent-ils ? Justifiez votre réponse.

D'après le **théorème d'unisolvance**, deux polynômes de degré inférieur ou égal à *n – 1* interpolateurs des mêmes *n* points deux à deux distincts sont nécessairement égaux.

B) Combien existe-t-il de façons différentes de construire ces polynômes respectivement avec la méthode de Lagrange et avec la méthode de Newton ? Justifiez votre réponse.

La méthode de Lagrange consiste à donner la définition de n fonctions cardinales qui forment un ensemble invariant par permutation de l'ensemble d'indices des points. Dis autrement, le procédé de construction ne dépend pas de l'ordre dans lequel on considère les points.

Plus formellement : $\forall \sigma \in \mathfrak{S}_n \forall k \in [1 \quad n] : L_{n-1}(x) = \sum_{k=1}^{n} y_{\sigma_k} l_{\sigma_k}(x) = \sum_{k=1}^{n} y_k l_k(x)$

En revanche, pour la méthode de Newton, le procédé itératif des différences divisées dépend de l'ordre dans lequel on traite les points et l'on peut donc considérer qu'il existe $n!$ variantes du procédé pour obtenir l('uniqu)e polynôme interpolateur.

Plus formellement : les valeurs des coefficients $a_{i \in [1 \quad n]}$ du polynôme dépendent de la permutation, à commencer par $a_1 = \nabla^0_{y_{\sigma_1}} = y_{\sigma_1}$

C) Etant donné l'ensemble de points suivant :

$$\left(\begin{array}{c|cccccc} i & 0 & 1 & 2 & 3 & 4 & 5 \\ x_i & -a^2 & -a & -1 & 1 & a & a^2 \\ y_i & \left(a^2+1\right)^2 & 1 & 0 & 0 & 1 & \left(a^2+1\right)^2 \end{array}\right)$$

1) En donner, sous forme réduite et factorisée, le polynôme d'interpolation de Lagrange.

$$L_{<n}(x)=\sum_{k=1}^{n} y_k l_k(x)=\sum_{k=1}^{n} y_k \prod_{i=1,i\neq k}^{n} \frac{(x-x_i)}{(x_k-x_i)}$$

Conseil

Une première façon de gagner du temps est d'**éviter de calculer les fonctions cardinales pour les points dont l'ordonnée est nulle**. En effet : $\forall i\left(y_i=0\rightarrow y_i l_i(x)=0\right)$.

Ensuite, il s'agit de profiter des symétries remarquables et notamment celle des fonctions cardinales d'abscisses opposées et de même ordonnée : leur factorisation est immédiate.

$$l_{-a^2}(x)=\frac{\left(x^2-1\right)\left(x^2-a^2\right)\left(x-a^2\right)}{\left(a^4-1\right)\left(a^4-a^2\right)\left(-a^2-a^2\right)} \qquad l_{a^2}(x)=\frac{\left(x^2-1\right)\left(x^2-a^2\right)\left(x+a^2\right)}{\left(a^4-1\right)\left(a^4-a^2\right)\left(a^2+a^2\right)}$$

$$l_{-a}(x)=\frac{\left(x^2-1\right)(x-a)\left(x^2-a^4\right)}{\left(a^2-1\right)(-a-a)\left(a^2-a^4\right)} \qquad l_{a}(x)=\frac{\left(x^2-1\right)(x+a)\left(x^2-a^4\right)}{\left(a^2-1\right)(a+a)\left(a^2-a^4\right)}$$

$$L_{<6}(x)=\left(a^2+1\right)^2\left(l_{a^2}(x)+l_{-a^2}(x)\right)+\left(l_a(x)+l_{-a}(x)\right)=$$

$$\frac{\left(a^2+1\right)^2\left(x^2-1\right)\left(x^2-a^2\right)}{2a^4\left(a^4-1\right)\left(a^2-1\right)}\left(\left(x+a^2\right)-\left(x-a^2\right)\right)+\frac{\left(x^2-1\right)\left(x^2-a^4\right)}{2a^3\left(a^2-1\right)\left(1-a^2\right)}\left((x+a)-(x-a)\right)=$$

$$\left(x^2-1\right)\left(\frac{\left(a^2+1\right)\left(x^2-a^2\right)}{a^2\left(a^2-1\right)^2}-\frac{\left(x^2-a^4\right)}{a^2\left(a^2-1\right)^2}\right)=\frac{\left(x^2-1\right)}{a^2\left(a^2-1\right)^2}\left(a^2x^2-a^2\right)=\left(\frac{x^2-1}{a^2-1}\right)^2$$

2) En donner, sous forme réduite et factorisée, le polynôme d'interpolation de Newton.

Conseil

L'application du procédé de Lagrange pour déterminer le polynôme interpolateur nous apporte quelques indices sur la meilleure façon d'appliquer le procédé de Newton.

D'abord, nous savons que le polynôme est de degré quatre et que le dernier coefficient $P_0(x) = a_6 = \nabla^5 y_6$ est donc nécessairement nul :

- Cela est utile pour vérifier qu'il n'y a pas d'erreur au terme de l'application du procédé.

- En revanche cela n'aide pas à déterminer la mise en œuvre la plus simple.

Par tactique, considérons d'une part la possibilité vue plus haut de choisir une configuration initiale parmi $n!$, d'autre part la présence manifeste de facteurs en $(a+1)$ et en $(a-1)$.

Le regroupement par lignes successives de points de même ordonnée peut conduire à maximiser la production de termes nuls sur la diagonale de la table de calcul, c'est-à-dire des coefficients qui sont ensuite employés dans la récurrence qui permet l'obtention du polynôme.

Placer les lignes à abscisse négative avant les lignes à abscisse positive permet d'obtenir des dénominateurs positifs et d'éviter les changements de signes propices aux erreurs.

Placer en dernière position les lignes à ordonnée algébriquement complexe permet de les faire intervenir le plus tard possible dans les calculs (en effet, plus un terme calculé est vers le haut et vers la droite du tableau de calcul, plus les autres termes à calculer en dépendent, et donc, plus son impact sur l'ensemble du calcul est important).

Le changement de variable : $\begin{cases} \alpha = a+1 \\ \beta = a-1 \end{cases}$ permet de diminuer la complexité des calculs, en contenant le nombre de termes, et d'éviter les excès de distribution.

$$\begin{pmatrix} \sigma_i & x_{\sigma_i} & \nabla^0 y_{\sigma_i} & \nabla^1 y_{\sigma_i} & \nabla^2 y_{\sigma_i} & \nabla^3 y_{\sigma_i} & \nabla^4 y_{\sigma_i} & \nabla^5 y_{\sigma_i} \\ 1 & -1 & 0 & & & & & \\ 2 & 1 & 0 & 0 & & & & \\ 3 & -a & 1 & -\frac{1}{\beta} & \frac{1}{\alpha\beta} & & & \\ 4 & a & 1 & \frac{1}{\alpha} & \frac{1}{\alpha\beta} & 0 & & \\ 5 & -a^2 & (a^2+1)^2 & -\frac{(a^2+1)^2}{\alpha\beta} & \frac{a^2+1}{\alpha\beta} & -\frac{a}{\alpha\beta^2} & \frac{1}{(\alpha\beta)^2} & \\ 6 & a^2 & (a^2+1)^2 & a^2+1 & \frac{a^2+1}{\alpha\beta} & \frac{a}{\alpha^2\beta} & \frac{1}{(\alpha\beta)^2} & 0 \end{pmatrix}$$

Pour construire le polynôme, nous utilisons la récurrence : $\begin{aligned} P_0(x) &= a_n = \nabla^{n-1} y_n \\ P_k(x) &= a_{n-k} + (x - x_{n-k}) P_{k-1}(x) \end{aligned}$

$$P_0(x) = 0 \qquad P_1(x) = \frac{1}{(\alpha\beta)^2} \qquad P_2(x) = (x-a) P_1(x) = \frac{x-a}{(\alpha\beta)^2}$$

$$P_3(x) = \frac{1}{\alpha\beta} + (x+a) P_2(x) = \frac{1}{(\alpha\beta)^2}\left(\alpha\beta + \left(x^2 - a^2\right)\right) = \frac{x^2-1}{(\alpha\beta)^2}$$

$$P_4(x) = (x-1) P_3(x) = \frac{(x-1)\left(x^2-1\right)}{(\alpha\beta)^2}$$

$$P_5(x) = (x+1) P_4(x) = \left(\frac{x^2-1}{\alpha\beta}\right)^2 = \left(\frac{x^2-1}{a^2-1}\right)^2$$

ANNEXES

Décomposition euclidienne employée à l'exercice 1

On part du reste après une première étape de normalisation.

$$(1.1)\ \forall 0 < \frac{n_k}{d_k} < 1 \in \mathbb{Q} \forall b \in \mathbb{N} \setminus \{0 \quad 1\} \exists ! q \,/\, \frac{1}{b^{q-1}} > \frac{n_k}{d_k} \geq \frac{1}{b^q}$$

$$(1.2)\ \exists r_k \,/\, \frac{n_k}{d_k} = \frac{1}{b^q} + r_k \wedge \frac{1}{b^{q-1}} > r_k \geq 0$$

Démonstration :

$$\frac{1}{b^{q-1}} > \frac{n_k}{d_k} \geq \frac{1}{b^q} \Leftrightarrow \frac{1}{b^{q-1}} - \frac{1}{b^q} > \frac{n_k}{d_k} - \frac{1}{b^q} \geq 0 \Leftrightarrow \frac{b-1}{b^q} > \frac{n_k}{d_k} - \frac{1}{b^q} \geq 0 \Rightarrow \frac{1}{b^{q-1}} > \frac{n_k}{d_k} - \frac{1}{b^q} \geq 0$$

$$(1.3)\ \exists r_k \,/\, \frac{n_k}{d_k} = \frac{1}{2^q} + r_k \wedge \frac{1}{2^q} > r_k \geq 0$$

Démonstration spécifique base deux : $\frac{2-1}{2^q} > \frac{n_k}{d_k} - \frac{1}{2^q} \geq 0 \Leftrightarrow \frac{1}{2^q} > \frac{n_k}{d_k} - \frac{1}{2^q} \geq 0$

En base deux, de (1.3) on dérive :

$$\frac{1}{2^{q-1}} > \frac{n_k}{d_k} \geq \frac{1}{2^q} \Leftrightarrow \exists ! \varepsilon_k = \frac{n_{k+1}}{d_{k+1}} \,/\, \frac{n_k}{d_k} = \frac{1}{2^q}(1+\varepsilon_k) \wedge 0 \leq \varepsilon_k < 1$$

Partant de là, on dispose d'un procédé itératif de décomposition euclidienne qui permet de réaliser un développement limité $\frac{n_1}{d_1} = \frac{1}{2^{q_1}}\left(1 + \frac{1}{2^{q_2}}\left(1 + \ldots \frac{1}{2^{q_{n-1}}}\left(1 + \frac{1}{2^{q_n}}(1+\varepsilon_n)\right)\ldots\right)\right) \wedge 0 \leq \varepsilon_n < 1$

Qui une fois distribué donne : $\frac{n_1}{d_1} = \frac{1}{2^{q_1}} + \frac{1}{2^{q_1+q_2}} + \ldots + \frac{1}{2^{\sum_{i=1}^{n-1} q_i}} + \frac{1}{2^{\sum_{i=1}^{n} q_i}}(1+\varepsilon_n) \wedge 0 \leq \varepsilon_n < 1$

Pour des exercices du type l'exercice 1 : période de la mantisse

Selon le théorème d'Euler : $(1.4)\ \forall b\in\mathbb{N}^*\forall n\in\mathbb{N}\left(b^{\varphi(n)}\equiv 1[n]\right)$

Et donc, en particulier : $(1.5)\ \forall b\in\mathbb{N}^*\forall p\in\mathcal{P}\left(b^{p-1}\equiv 1[p]\right)$ (Petit Fermat)

$$(1.4)\Leftrightarrow \exists M_n\in\mathbb{N}/\, b^{\varphi(n)}=1+nM_n \Leftrightarrow \exists M_n\in\mathbb{N}/\, M_n=\frac{b^{\varphi(n)}-1}{n}$$

$$\Leftrightarrow \exists M_n\in\mathbb{N}/\, \frac{1}{n}=\frac{M_n}{b^{\varphi(n)}-1}=b^{-\varphi(n)}M_n\left(\frac{1}{1-b^{-\varphi(n)}}\right)=b^{-\varphi(n)}M_n\sum_{k=0}^{+\infty}b^{-k\varphi(n)}$$

Cela signifie que $\frac{1}{n}$ se décompose en un produit d'un décalage gauche (normalisation) de $b^{-\varphi(n)}$, avec un motif $M_n=\frac{b^{\varphi(n)}-1}{n}$ et le support $\sum_{k=0}^{+\infty}b^{-k\varphi(n)}$ de répétition périodique de M_n.

Il suffit de voir que la taille de M_n, $|M_n|\le\varphi(n)$ (car $M_n<b^{\varphi(n)}$) pour vérifier que deux répétitions du motif ne se chevauchent pas, et pour déterminer l'écriture de $\frac{1}{n}$ en base b à $\left(0,\underbrace{0\ldots0}_{\varphi(n)}\overbrace{0\ldots0}^{\varphi(n)-|M_n|}\underline{M_n}\right)_b$ (la notation soulignée signifie la répétition).

Application directe : $\begin{cases}b=10\\ n=3^3\end{cases}\Rightarrow\begin{cases}\varphi(27)=(3-1)3^{3-1}=18\\ M_n=\frac{10^{18}-1}{27}=37037037037037037\end{cases}$

Cet exemple pour montrer que si le raisonnement ci-dessus permet de prouver l'existence de, et de calculer effectivement le motif périodique d'une mantisse, celui-ci n'est pas unique, et ce procédé de calcul n'est pas optimal (sous-motifs, motifs décalés).

Pour faire mieux, à vous de jouer !

Vous pouvez, par exemple, vous exercer en tentant de déterminer la période et le motif minimaux de la mantisse de l'exercice 1.