

Analyse Numerique

David Wiedemann

Table des matières

1	Representation de nombres en arithmetique finie	2
1.1	Representation des nombres dans les ordinateurs	2
1.2	Approximation de \mathbb{R} dans $\mathcal{F}(2, 53, -1021, 1024)$	2
1.3	Operations dans \mathcal{F}	3
1.4	Parenthese sur le concept de stabilite	3
2	Integration Numerique	3
2.1	Formules d'integration de Newton-Cotes	4
2.2	Formules de quadrature d'ordre optimal	7

List of Theorems

2	Proposition	2
1	Definition	3
2	Definition (Formule de Quadrature)	4
3	Definition	5
4	Theorème	5
7	Theorème (Thm. fondamental de la theorie de l'integration) . . .	7

Lecture 1: Representation de nombres en arithmetique finie

Thu 03 Mar

1 Representation de nombres en arithmetique finie

Notons $\mathcal{F}(\beta, t, L, U)$ l'ensemble des nombres representables sous la forme $(-1)^s(0, \alpha_1 \dots \alpha_t)_\beta \beta^e$ ou e est l'exposant, $L \leq e \leq U, 0 \leq \alpha_i < \beta, \alpha_1, \dots, \alpha_t$ est la mantisse et s le signe.

Cette representation est la representation floating point.

1.1 Representation des nombres dans les ordinateurs

On appelle les nombres en double precision l'ensemble

$$\mathcal{F}(2, 53, -1021, 1024)$$

Bien que les valeurs maximales et minimales sont tres grandes ($2 \cdot 10^{-308}$ et $2 \cdot 10^{308}$), mais on en saute beaucoup.

Tous les nombres dans \mathcal{F} sont de la forme $\frac{p}{2^n}, p \in \mathbb{N}$.

On regarde la distance entre deux nombres consecutifs de \mathcal{F} .

Pour un exposant fixe, $[2^p, 2^{p+1}]$, le premier nombre apres 2^p est

$$(0.10 \dots 01)2^{p+1} = 2^p + 2^{p+1-t}$$

Donc dans ce cas, on a que le spacing est donne par 2^{p-52} .

Remarque

Si on a que des entiers dans un intervalle $[\beta^p, \beta^{p+1}]$, alors $\beta^{p+1-t} = 1$.

1.2 Approximation de \mathbb{R} dans $\mathcal{F}(2, 53, -1021, 1024)$

Soit $x \in \mathbb{R}$, on appelle $fl(x) \in \mathcal{F}(2, 53, -1021, 1024)$.

Notons $x = (-1)^s(0, \alpha_1 \dots \alpha_{t-1} \alpha_t \alpha_{t+1} \dots)_\beta \beta^e$, on definit alors

$$fl(x) = (-1)^s(0, \alpha_1 \dots \alpha_{t-1} \tilde{\alpha}_t)_\beta \beta^e$$

on fait l'hypothese ici que au moins un des α_i est non nul.

On veut borner $|x - fl(x)| \leq \frac{1}{2} \text{spacing} = \frac{1}{2} \beta^{e-t}$.

Bien que l'erreur absolue est, en principe, grande, l'erreur relative sera bornee, on a en effet

$$\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2} \beta^{e-t} \frac{1}{|x|} \leq \frac{1}{2} \beta^{1-t} (\simeq 10^{-16} \text{ dans notre systeme })$$

On appelle cette erreur la "machine precision" et on la note u

Proposition 2

On peut egalement ecrire que

$$x \in \mathbb{R} \quad fl(x) = x(1 + \epsilon), |\epsilon| \leq u$$

1.3 Operations dans \mathcal{F}

Soit $x, y \in \mathbb{R}$, $x + y \mapsto fl[fl(x) + fl(y)]$, qu'elle est l'erreur relative commise ?

$$\frac{|fl[fl(x) + fl(y)] - (x + y)|}{|x + y|}$$

En utilisant la proposition ci-dessus, notons $fl(x) = x(1 + \epsilon_1)$, $fl(y) = y(1 + \epsilon_2)$, on a alors

$$\begin{aligned} |(x(1+\epsilon_1)+y(1+\epsilon_2))(1+\epsilon_3)-(x+y)| \cdot \frac{1}{|x+y|} &\leq \frac{x\epsilon_1 + y\epsilon_2 + \epsilon_3(x+y) - (x+y)}{|x+y|} + \text{petit} \\ &\leq \left(\frac{|x|}{|x+y|} + \frac{|y|}{|x+y|} + 1 \right) u \end{aligned}$$

On remarque que si $x > 0, y < 0$, il est possible de commettre une erreur tres grande.

On dit que la soustraction est une operation instable.

1.4 Parenthese sur le concept de stabilite

On veut resoudre $y = G(x)$.

Definition 1

La resolution de $y = G(x)$ est stable si une petite perturbation de x correspond a une petite perturbation de y , ie.

$$y + \delta y = G(x + \delta x)$$

On appelle alors le conditionnement absolu du probleme

$$\kappa_{abs} = \sup_{\delta x} \frac{\|\delta y\|}{\|\delta x\|}$$

Et on appelle perturbation relative du probleme

$$\kappa_{rel} = \sup_{\delta x} \frac{\|\delta y\| / \|y\|}{\|\delta x\| / \|x\|}$$

Lecture 2: Integration Numerique

Thu 10 Mar

2 Integration Numerique

On veut construire des algorithmes pour calculer de maniere approchee $\int_a^b f(x)dx$

2.1 Formules d'intégration de Newton-Cotes

On écrit

$$\int_a^b f(x)dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)dx$$

Chacun des termes de la somme se réécrit comme

$$\int_{x_i}^{x_{i+1}} f(x)dx = \int_0^1 f(x_i + th_i)h_i dt$$

Et on trouve

$$\int_a^b f(x)dx = \sum_{i=1}^{N-1} h_i \int_0^1 f(x_i + th_i)dt$$

Ainsi, il suffit de trouver un algorithme pour calculer des intégrales de la forme $\int_0^1 g(t)dt$. La manière la plus naïve pour approximer cette intégrale serait de prendre $\int_0^1 g(t)dt \approx g(\frac{1}{2})$, et on note $Q_1^{nc}(g) = g(\frac{1}{2})$.

Une manière moins naïve de faire est d'approcher g par une fonction linéaire et de prendre l'approximation

$$\int_0^1 g(t)dt \approx \frac{1}{2} (g(0) + g(1)) = Q_2^{nc}(g) \text{ (formule de Newton-Cote a deux noeuds)}$$

ou encore

$$\int_0^1 \approx \frac{1}{6} (g(0) + 4g(\frac{1}{2}) + g(1)) = Q_3^{nc}(g) \text{ (formule de cote a trois noeuds ou formule de Simpson)}$$

De manière générale, on appelle formule de Newton-Cotes à S noeuds

$$\int_0^1 g(t)dt \approx \int_0^1 p(t)dt$$

où $p(t)$ est le polynôme de degré $s - 1$ passant par les points $(c_i, g(c_i))$, où $0 \leq c_1 \leq \dots \leq c_{s-1} < c_s \leq 1$.

Ainsi, de manière générale

$$Q_S^{nc}(g) = \sum_{i=1}^s b_i g(c_i)$$

où b_i sont les poids des formules de N.C.

On veut donc essayer de trouver des formules qui donnent les poids de l'intégration de Newton-Cotes.

Definition 2 (Formule de Quadrature)

Une formule de quadrature $Q_s(f)$ est donnée par n'importe quelle en-

semble de couples $(\{b_i\}_{i=1}^s, \{c_i\}_{i=1}^s)$:

$$Q_s(f) = \sum_{i=1}^N b_i f(c_i)$$

Definition 3

$Q_s(\cdot)$ est d'ordre s quand elle est exacte sur tout polynome de degre $\leq s-1$

Remarque

Par definition les formules Q_s^{nc} sont d'ordre s .

Theorème 4

Etant donne s noeuds distincts $\{c_i\}_{i=1}^N$, la formule donnee par $(\{b_i\}, \{c_i\})$ est d'ordre s si et seulement si les poids verifient

$$\sum_{i=1}^s c_i^{q-1} b_i = \frac{1}{q} \quad \forall q = 1, \dots, s$$

Preuve

Supposons que Q est d'ordre s , alors prenons

$$p(t) = t^q \quad q = 1 \dots s$$

On ecrit

$$\int_0^1 p(t) dt = \int_0^1 t^{q-1} dt = \frac{1}{q}$$

d'autre part

$$\sum_{i=1}^s b_i p(c_i) = \sum_{i=1}^s b_i p(c_i) = \sum_{i=1}^s b_i c_i^{q-1}$$

Dans l'autre sens, si $\sum_{i=1}^s c_i^{q-1} b_i = \frac{1}{q}$, alors la formule est exacte sur tout monome (par le raisonnement ci-dessus), par linearite, elle sera donc exacte sur n'importe quel polynome. \square

On montre maintenant qu'en fait les poids b_i sont uniques étant donné les c_i , en effet, étant donné le théorème ci-dessus, on a

$$\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ c_1 & c_2 & c_3 & \dots & c_s \\ c_1^2 & c_2^2 & c_3^2 & \dots & c_s^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_1^{s-1} & c_2^{s-1} & c_3^{s-1} & \dots & c_s^{s-1} \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ \vdots \\ b_s \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{3} \\ \vdots \\ \frac{1}{s} \end{pmatrix}$$

Ainsi, soit la matrice A ci-dessus est inversible, alors il y a un seul choix de poids pour la formule de N.C.

Par un théorème d'algèbre linéaire, la matrice est inversible. En appliquant donc ceci à une fonction f générale, on trouve

$$\begin{aligned} \int_a^b f(x) dx &= \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} f(x) dx = \sum_{j=0}^{N-1} h_j \int_0^1 f(x_j + th_j) dt \\ &= \sum_{j=0}^{N-1} h_j Q_s^{nc}(f(x_j + th_j)) = \sum_{j=0}^{N-1} h_j \sum_{i=1}^s b_i f(x_j + c_i h_j) \end{aligned}$$

Remarque

Pour les noeuds c_i fixes, il existe un seul choix de poids qui garantit que Q_s est d'ordre s .

Quel est le choix optimal des noeuds ?

- **Choix 1** Choisir des noeuds équidistants.

Ce choix rend le calcul instable en arithmétique finie.

En effet, supposons qu'on veut intégrer $f(x) > 0$, on aura $\sum_{i=1}^s f(ih)b_i$.

Alors les poids oscillent fortement.

- **Choix 2** On cherche à comprendre où placer les noeuds pour maximiser l'ordre de la formule.

Exemple

On considère à nouveau la formule de Simpson

$$Q_3^{nc}(g) = \frac{1}{6} \left[g(0) + 4g\left(\frac{1}{2}\right) + g(1) \right]$$

Ainsi, pour $c_i = 0, \frac{1}{2}, 1$ on a les poids $b_i = \frac{1}{6}, \frac{2}{3}, \frac{1}{6}$. Est-ce que cette formule est d'ordre 4 ?

$$\int_0^1 t^3 dt = \frac{1}{4} = \sum_i b_i c_i^3 = \frac{1}{4} \text{ (en substituant les valeurs)}$$

Est-elle aussi d'ordre 5 ?

$$\int_0^1 t^4 dt = \frac{1}{5} = \sum_i b_i c_i^4 = \frac{2}{3} \frac{1}{16} + \frac{1}{6} \neq \frac{1}{5}$$

2.2 Formules de quadrature d'ordre optimal

On veut donc choisir des noeuds c_1, \dots, c_s pour maximiser l'ordre de la formule de quadrature

Theorème 7 (Thm. fondamental de la theorie de l'integration)

Soit $(\{b_i\}, \{c_i\})$ une formule de quadrature d'ordre s , $Q_s(\cdot)$.

Soit $M(t) = (t - c_1)(t - c_2) \dots (t - c_s)$, alors la formule $Q_s(\cdot)$ est d'ordre $p \geq s + m$ si et seulement si

$$\int_0^1 M(t)g(t) dt = 0$$

Preuve

Soit $f(t)$ un polynome de degre $s + m - 1$, prenons $r(t)$ un polynome de degre $s - 1$ passant par les points $(c_i, f(c_i))$.

Alors $f(t) - r(t)$ est un polynome de degre $s + m - 1$ est un polynome s'annulant sur tous les noeuds.

Ainsi

$$f(t) - r(t) = M(t)g_f(t) \text{ avec } \deg g_f \leq m - 1$$

\Leftarrow

Supposons que $\int_0^1 M(t)g(t) dt = 0 \forall$ polynome $g(t) : \deg g \leq m - 1$.

On demontre que la formule est d'ordre $s + m - 1$.

Soit f un polynome $\deg f \leq s + m - 1$, on peut donc ecrire

$$f(t) = r(t) + \underbrace{\int_0^1 M(t)g_f(t) dt}_{=0}$$

De meme, on a que

$$Q_s(f) = \sum_{i=1}^s b_i f(c_i) = \sum_{i=1}^s b_i \left[r(c_i) + \underbrace{M(c_i)g_f(c_i)}_{=0} \right] = \int_0^1 r(t) dt$$

Et donc la formule est exacte

\Rightarrow

Supposons que la formule est d'ordre $s + m$, demontrons que $\int_0^1 M(t)g(t) dt = 0 \forall g, \deg g \leq m - 1$, ainsi

$$\int_0^1 M(t)g(t) dt = \sum_{i=1}^s b_i M(c_i)g(c_i) = 0$$

□