Ann. Sc. Math. Québec, Vol. III, No 2, pp. 227-235

UNE MÉTHODE DE QUADRATURE FAISANT APPEL À DES SUBDIVISIONS BINAIRES

Gilbert Labelle

0. INTRODUCTION

Soit $f : [0,1] \to \mathbb{R}$ une fonction de classe C^{k+1} où $k \geq 1$ est fixé et désignons par S_m sa somme de Riemann d'ordre $m \geq 1$:

$$S_m = \sum_{i=0}^{m-1} \frac{1}{m} f\left(\frac{i}{m}\right) \quad . \tag{0.1}$$

Un simple changement de variable dans la version classique [2] de la formule sommatoire d'Euler-Maclaurin permet d'écrire

$$S_m = a_o + \frac{a_1}{m} + \frac{a_2}{m^2} + \ldots + \frac{a_k}{m^k} + R_{m,k} \tag{0.2}$$

où

$$a_o = \int_o^1 f(x)dx \ , \quad a_\nu = B_\nu \cdot [f^{(\nu-1)}(1)-f^{(\nu-1)}(0)]/\nu! \ , \quad 1 \leq \nu \leq k \tag{0.3}$$

et

$$R_{m,k} = -\frac{1}{(k+1)!m^{k+1}} \int_o^1 \left[B_{k+1}(1-t)-B_{k+1}\right] \cdot \left[\frac{1}{m} \sum_{i=0}^{m-1} f^{(k+1)}\left(\frac{i+t}{m}\right)\right]dt \tag{0.4}$$

$$= O(m^{-k-1}) \ , \quad m \to \infty \ .$$

Les fonctions $B_j(x)$ et les nombres B_j sont respectivement les polynômes et les nombres de Bernoulli, $j = 0,1,2,\ldots$. La formule (0.2) ne nous fournit évidemment une méthode de quadrature vraiment efficace que si les premières dérivées de f sont calculables facilement. Dans le cas où seulement les valeurs de f sont

disponibles, une "extrapolation de la limite" nous permet d'éliminer $a_1, a_2, \ldots, a_k$ dans (0.2) par une combinaison linéaire adéquate des sommes $S_m, S_{m-1}, S_{m-2}, \ldots, S_{m-k}$ et d'obtenir ainsi

$$a_o = \int_o^1 f(x)dx \cong \sum_{i=0}^{k} \alpha_{m,k,i}\, S_{m-i} \tag{0.5}$$

où

$$\alpha_{m,k,i} = (-1)^i (m-i)^k [i!(k-i)!]^{-1} . \tag{0.6}$$

La quadrature (0.5) demande toutefois de calculer les valeurs de f en beaucoup de points (noeuds). Ce sont tous les points distincts de la forme $j(m-i)^{-1}$ où $i = 0,1,\ldots,k$, $0 \le j < m - i$. Le nombre de ces noeuds est beaucoup plus grand que m lorsque k est le moindrement grand. De plus, les coefficients (qui dépendent de m) ainsi que les noeuds ne sont pas très bien adaptés au calcul binaire direct.

Le but de cet article est de développer une variante de (0.5) permettant de réduire de beaucoup le nombre des noeuds (à moins de m en fait), d'éliminer m de la structure numérique des coefficients, de conserver une précision $0(m^{-k-1})$ et de rendre tous les nouveaux noeuds et coefficients d'un calcul pratique récursif en base 2 à multiple précision quel que soit leur nombre.

1. LA MÉTHODE

Voici d'abord quelques notations. Les nombres de Mersenne $2^i - 1$ (premiers ou non) seront représentés par μ_i , $i = 1,2,3,\ldots$. Les sommes particulières de Riemann S_m où $m = 2^n$ seront désignées par T_n , ainsi

$$T_n = 2^{-n} \sum_{i=0}^{2^n-1} f(2^{-n}i) . \tag{1.1}$$

On utilisera aussi le symbolisme

$$E_n = 2^{-(n-1)} \sum_{\substack{0<i<2^n \\ i \text{ impair}}} f(2^{-n}i) = \{f(2^{-n})+f(3\cdot 2^{-n})+\ldots+f((2^n-1)\cdot 2^{-n})\}/2^{n-1} . \tag{1.2}$$

Les sommes E_n sont, en quelque sorte, des sommes de Riemann "centrées" et "symétriques". En particulier, $E_0 = 0$, $E_1 = f(1/2)$, $E_2 = 1/2\ \{f(1/4)+f(3/4)\}$, $E_3 = 1/4\ \{f(1/8)+f(3/8)+f(5/8)+f(7/8)\}$, ...

Nous sommes maintenant en position d'énoncer le théorème suivant.

Théorème. Si $f : [0,1] \to \mathbb{R}$ est de classe C^{k+1} où $k \leq n$ est fixé, alors

$$\int_0^1 f(x)dx = \sum_{i=0}^{k-1} c_{k,i}\, E_{n-i} + \varepsilon_{n,k} \tag{1.3}$$

où $\varepsilon_{n,k} = 0(2^{-(k+1)n})$ quand $n \to \infty$ et où les coefficients $c_{k,i}$ sont indépendants de n et sont donnés par la récurrence

$$\begin{cases} c_{k,0} = \dfrac{\frac{k(k+1)}{2} - 1}{\mu_1\mu_2\cdots\mu_k} \\ \\ c_{k,i} = -\dfrac{\mu_{k-i}}{2^{k-i+1}\mu_i} \cdot c_{k,i-1} , \qquad i = 1,\ldots,k-1 . \end{cases} \tag{1.4}$$

Avant d'aborder la démonstration du théorème remarquons que les noeuds utilisés sont seulement au nombre de $2^{n-1} + 2^{n-2} + \ldots + 2^{n-k} = 2^n - 2^{n-k} < 2^n = m$, puisque E_j contient 2^{j-1} noeuds dans sa formation et que les noeuds sont "disjoints" pour des j distincts. Les sommes E_j s'engendrent par des récurrences binaires évidentes. Les coefficients $c_{k,i}$ se calculent par des multiplications selon

a) des puissances de 2 (i.e. translations du "point" binaire),

b) des nombres de Mersenne (i.e. par translations et soustractions binaires),

c) des inverses de nombres de Mersenne (i.e. par translations et additions binaires successives à cause du fait que $\mu_i^{-1} = 2^{-i} + 2^{-2i} + 2^{-3i} + \ldots$) .

On peut donc calculer très facilement les $c_{k,i}$ pour tous k et i par ordinateur et à multiple précision (une fois pour toutes ou encore, au fur et à mesure

qu'ils sont nécessaires) en utilisant, par exemple, un langage machine. Il y a $k+1$ termes dans la somme (0.5) tandis qu'il n'y en a que k dans (1.3).

Démonstration du théorème. La version (0.2) de la formule sommatoire d'Euler-Maclaurin avec successivement $m = 2^{n+i}$ $(i = 0,1,2,\ldots,k)$ permet d'écrire le système d'équations linéaires aux "inconnues" $a_0,a_1,\ldots,a_k$:

$$T_{n+i} = a_0 + 2^{-(n+i)}a_1 + 2^{-2(n+i)}a_2 + \ldots + 2^{-k(n+i)}a_k + \rho_{n,k,i} \qquad (1.5)$$

où $a_0 = \int_0^1 f(x)dx$ et $\rho_{n,k,i} = 0(2^{-(k+1)n})$ lorsque $n \to \infty$. Nous allons, dans une première étape, éliminer les coefficients $a_1,a_2,\ldots,a_k$ et résoudre pour a_0. Pour faire ceci, définissons un polynôme auxiliaire $P(t)$ en posant

$$P(t) = a_0 + a_1t + a_2t^2 + \ldots + a_kt^k \ . \qquad (1.6)$$

Le système (1.5) prend alors la forme du problème d'interpolation

$$P(2^{-n-i}) = U_{n,k,i} \ , \quad i = 0,1,\ldots,k \ , \quad U_{n,k,i} = T_{n+i} - \rho_{n,k,i} \ . \qquad (1.7)$$

Le polynôme (1.6) est donc identique au polynôme de Lagrange donné explicitement par

$$P(t) = \sum_{i=0}^{k} U_{n,k,i} \prod_{j\neq i} (t-2^{-n-j})(2^{-n-i}-2^{-n-j})^{-1} \ . \qquad (1.8)$$

Comme $\int_0^1 f(x)dx = a_0 = P(0)$, on déduit de (1.8) après substitution et réarrangement de termes, que

$$\int_0^1 f(x)dx = \sum_{i=0}^{k} \omega_{k,i}\, T_{n+i} + \varepsilon^*_{n,k} \qquad (1.9)$$

où

$$\omega_{k,i} = (-1)^{k-i}\, 2^{i(i+1)/2}\, \mu_1^{-1}\mu_2^{-1}\ldots\mu_i^{-1}\mu_1^{-1}\mu_2^{-1}\ldots\mu_{k-i}^{-1} \qquad (1.10)$$

et

$$\varepsilon^*_{n,k} = -\sum_{i=0}^{k} \omega_{k,i}\, \rho_{n,k,i} = 0(2^{-(k+1)n}) \qquad (1.11)$$

puisque les $\omega_{k,i}$ sont indépendants de n. La somme dans (1.9) fait cependant

intervenir plusieurs fois les mêmes noeuds. En effet, les $k + 1$ sommes de Riemann T_{n+i} sont affectées de coefficients distincts et leurs noeuds forment des ensembles emboîtés.

La deuxième étape (qui est l'étape vraiment centrale) sera donc de réécrire la somme (1.9) en utilisant les sommes E_j comme "classificateurs" des termes dans les sommes T_{n+i}. On vérifie immédiatement que

$$2^{n+i} T_{n+i} = f(0) + \sum_{j=1}^{n+i} 2^{j-1} E_j . \tag{1.12}$$

Une substitution de cette égalité dans la somme de (1.9) donne

$$\begin{aligned} \sum_{i=0}^{k} \omega_{k,i} T_{n+i} &= 2^{-n} \sum_{i=0}^{k} \left\{ 2^{-i} \omega_{k,i} \left(f(0) + \sum_{j=1}^{n+i} 2^{j-1} E_j \right) \right\} \\ &= 2^{-n} \left(\sum_{i=0}^{k} 2^{-i} \omega_{k,i} \right) \left(f(0) + \sum_{j=1}^{n} 2^{j-1} E_j \right) + \sum_{i=1}^{k} \left(2^{-i} \omega_{k,i} \sum_{j=1}^{i} 2^{j-1} E_{n+j} \right) . \end{aligned} \tag{1.13}$$

Le facteur $2^{-n} \left(\sum_{i=0}^{k} 2^{-i} \omega_{k,i} \right)$ entrant en jeu dans (1.13) est en fait nul. Pour le vérifier, il suffit de remarquer qu'il est égal à $Q(0)$ où $Q(t)$ est le polynôme de Lagrange associé au problème d'interpolation

$$Q(2^{-n-i}) = 2^{-n-i} , \qquad i = 0,1,\ldots,k \tag{1.14}$$

et que ce polynôme est égal, par unicité, au polynôme du premier degré $Q(t) = t$ qui s'annule donc en $t = 0$. Tenant compte de cette simplification importante et réarrangeant les termes de (1.13) on arrive à l'égalité

$$\sum_{i=0}^{k} \omega_{k,i} T_{n+i} = \sum_{i=1}^{k} 2^{i-1} \theta_{k,i} E_{n+i} \tag{1.15}$$

où

$$\theta_{k,i} = \sum_{j=i}^{k} 2^{-j} \omega_{k,j} , \qquad i = 1,2,\ldots,k . \tag{1.16}$$

Nous allons maintenant montrer que les $\theta_{k,i}$ ont une structure numérique beaucoup plus simple que la somme (1.16) le laisse à-priori supposer. En effet, nous allons vérifier, par induction sur i, que

$$\theta_{k,i} = (-1)^{k-i}\,\mu_k^{-1} \cdot 2^{i(i-1)/2} \cdot \mu_1^{-1}\mu_2^{-1}\ldots\mu_{i-1}^{-1}\mu_1^{-1}\mu_2^{-1}\ldots\mu_{k-i}^{-1} \,. \tag{1.17}$$

Tout d'abord,

$$\theta_{k,i} = \left(\sum_{j=0}^{k} - \sum_{j=0}^{i-1}\right)(2^{-j}\,\omega_{k,j}) = -\sum_{j=0}^{i-1} 2^{-j}\,\omega_{k,j} \tag{1.18}$$

puisque nous avons vu plus haut que la somme complète $\left(\sum_{j=0}^{k}\right)$ est nulle. A cause de (1.18) et (1.10), on a que $\theta_{k,1} = -\omega_{k,o} = (-1)^{k-1}\,\mu_1^{-1}\mu_2^{-1}\ldots\mu_k^{-1}$ ce qui est conforme à (1.17) pour $i = 1$. Supposons l'égalité (1.17) vraie pour i , alors le calcul suivant montre qu'elle est aussi vraie pour $i + 1$:

$$\begin{aligned}
\theta_{k,i+1} &= \theta_{k,i} - 2^{-i}\,\omega_{k,i} \quad \text{(par (1.18))} \\
&= (-1)^{k-i}\,\mu_k^{-1} \cdot 2^{i(i-1)/2} \cdot \mu_1^{-1}\ldots\mu_{i-1}^{-1}\mu_1^{-1}\ldots\mu_{k-i}^{-1} \\
&\qquad - (-1)^{k-i} \cdot 2^{i(i-1)/2}\,\mu_1^{-1}\ldots\mu_i^{-1}\mu_1^{-1}\ldots\mu_{k-i}^{-1} \\
&= (-1)^{k-i}\,\mu_k^{-1} \cdot 2^{i(i-1)/2}\,(\mu_i - \mu_k)\mu_1^{-1}\ldots\mu_i^{-1}\mu_1^{-1}\ldots\mu_{k-i}^{-1} \\
&= (-1)^{k-(i+1)}\,\mu_k^{-1} \cdot 2^{(i+1)i/2}\,\mu_1^{-1}\ldots\mu_i^{-1}\mu_1^{-1}\ldots\mu_{k-i-1}^{-1} \,,
\end{aligned}$$

car $\mu_i - \mu_k = -2^i\,\mu_{k-i}$.

Remplaçant maintenant n par $n - k$ dans (1.9) et (1.15) et réarrangeant les indices, on obtient alors la formule (1.3) avec

$$c_{k,i} = 2^{k-i-1}\,\theta_{k,k-i} \,, \qquad \varepsilon_{n,k} = \varepsilon^*_{n-k,k} \,, \qquad i = 0,1,\ldots,k-1 \,. \tag{1.19}$$

On vérifie finalement par calculs directs basés sur (1.17) que les $c_{k,i}$ ainsi définis satisfont bien à la récurrence (1.4). Ceci termine la démonstration du théorème.

Le corollaire qui suit fournit des formules asymptotiques pour les restes $\varepsilon_{n,k}$ en supposant une dérivabilité un peu plus forte sur f .

Corollaire. Soit $k \geq 1$ fixé avec $k \leq n$ et soit $\varepsilon_{n,k}$ le reste dans la quadrature (1.3). Alors

a) Si k est impair et f de classe C^{k+2}, on a

$$\varepsilon_{n,k} \sim \frac{1}{2^{(k+1)n}} \cdot \frac{2^{k(k+1)/2}\, B_{k+1}}{(k+1)!} \cdot [f^{(k)}(1)-f^{(k)}(0)] \tag{1.20}$$

lorsque $n \to \infty$.

b) Si k est pair et f de classe C^{k+3}, on a

$$\varepsilon_{n,k} \sim -\frac{1}{2^{(k+2)n}} \cdot \frac{2^{k(k+1)/2}\, \mu_{k+1}\, B_{k+2}}{(k+2)!} \cdot [f^{(k+1)}(1)-f^{(k+1)}(0)] \tag{1.21}$$

lorsque $n \to \infty$.

Démonstration. Définissons d'abord pour chaque $k \geq 0$ deux quantités u_k et v_k par

$$u_k = \sum_{j=0}^{k} \omega_{k,j} \quad \text{et} \quad v_k = \sum_{j=0}^{k} 2^j\, \omega_{k,j} . \tag{1.22}$$

Nous allons montrer que

$$u_k = 1 \quad \text{et} \quad v_k = \mu_{k+1} . \tag{1.23}$$

La première égalité de (1.23) est simple à établir. En effet, on vérifie par calcul direct que $u_k = V(0)$ où $V(t)$ est le polynôme de Lagrange associé au problème d'interpolation $V(2^{-i}) = 1$, $i = 0,1,\ldots,k$. Ce polynôme est donc identique au polynôme constant $V(t) = 1$. La deuxième égalité de (1.23) est plus délicate et nous allons la vérifier par induction sur k. Elle est trivialement vraie pour $k = 0$ puisque $v_0 = 1 = \mu_1$. Le calcul

$$\begin{aligned} v_{k+1} &= \sum_{i=0}^{k+1} 2^i\, \omega_{k+1,i} = \sum_{i=0}^{k+1} (1+\mu_i)\, \omega_{k+1,i} \\ &= u_{k+1} + \sum_{i=1}^{k+1} \mu_i\, \omega_{k+1,i} = u_{k+1} + 2 \cdot \sum_{j=0}^{k} 2^j\, \omega_{k,j} \\ &= 1 + 2v_k \end{aligned}$$

complète l'induction puisque $\mu_{k+2} = 1 + 2\mu_{k+1}$.

Montrons maintenant (1.20). Les formules (0.2) et (1.5) donnent (pour $k \geq 1$ impair et $f \in C^{k+2}$)

$$\rho_{n,k,i} = 2^{-(k+1)(n+i)}\, a_{k+1} + \rho_{n,k+1,i} = 2^{-(k+1)(n+i)}\, a_{k+1} + O(2^{-(k+2)n}),$$

on déduit alors de (1.11), (1.19) et (1.23) que

$$\varepsilon_{n,k} = \varepsilon^*_{n-k,k} \sim -2^{-(k+1)(n-k)}\, a_{k+1} \sum_{i=0}^{k} 2^{-(k+1)i}\, \omega_{k,i}$$

$$= 2^{-(k+1)n} \cdot 2^{k(k+1)/2}\, a_{k+1}\, u_k$$

car $2^{-(k+1)i}\, \omega_{k,i} = (-1)^k\, 2^{-k(k+1)/2}\, \omega_{k,k-i}$. Pour montrer (1.21), on procède comme suit. Les formules (0.2) et (1.5) donnent (pour $k > 1$ pair et $f \in C^{k+3}$)

$$\rho_{n,k,i} = 2^{-(k+1)(n+i)}\, a_{k+1} + 2^{-(k+2)(n+i)}\, a_{k+2} + \rho_{n,k+2,i}$$

$$= 2^{-(k+2)(n+i)}\, a_{k+2} + O(2^{-(k+3)n})$$

puisque (voir (0.3)) le nombre de Bernoulli B_{k+1} est nul dans le cas considéré. Les égalités (1.11), (1.19) et (1.23) donnent cette fois

$$\varepsilon_{n,k} = \varepsilon^*_{n-k,k} \sim -2^{-(k+2)(n-k)}\, a_{k+2} \sum_{i=0}^{k} 2^{-(k+2)i}\, \omega_{k,i}$$

$$= -2^{-(k+2)n} \cdot 2^{k(k+1)/2}\, a_{k+2}\, v_k$$

car $2^{-(k+2)i}\, \omega_{k,i} = (-1)^k\, 2^{-k(k+3)/2} \cdot 2^{k-i}\, \omega_{k,k-i}$. Ceci achève la démonstration du corollaire.

Il est possible de raffiner notre méthode de quadrature en exploitant plus systématiquement la propriété suivante des nombres de Bernoulli: $B_n = 0$ si $n > 1$ est impair. Cependant, la version parente de (1.3) qui en résulte contient toujours n termes (au lieu de k), les coefficients ne sont plus tous indépendants de n et les simplifications permettant de les exprimer sous une forme simple n'existent plus.

La quadrature classique de Gauss [3], bien que plus précise (à nombre de noeuds égal) que notre méthode, possède les inconvénients bien connus que ses noeuds et ses poids sont d'une structure numérique complexe reliée aux racines et valeurs des polynômes de Legendre; ils sont donnés, à précision limitée, par de grandes tables. Par exemple, Abramovitz et Stegun [1] en donnent une table à 20 décimales jusqu'à 96 noeuds (seulement). La quadrature contenue dans le présent texte ne possède pas, comme nous l'avons vu, ces inconvénients et mérite selon nous une attention spéciale.

RÉFÉRENCES

[1] ABRAMOVITZ, M., STEGUN, I.A., Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards, Applied Mathematics Series 55, U.S. Department of Commerce, June 1964.

[2] GUELFOND, A.O., Calcul des différences finies, Collection Universitaire de mathématiques, Dunod, Paris, 1963.

[3] WENDROFF, B., Theoretical Numerical Analysis, Academic Press, New York, 1967.

Département de mathématiques
Université du Québec à Montréal
C.P. 8888, Succ. A
Montréal, Qué.
Canada

Manuscrit reçu le 25 juillet 1978.
Revisé le 13 octobre 1978.