数理解析研究所講究録
1040巻 1998年 92-99

偏微分方程式の任意精度数値シミュレーションについて

徳島大学 工学部 今井 仁司 (Hitoshi Imai)
徳島大学 工学部 竹内 敏己 (Toshiki Takeuchi)
徳島大学 工学部 坂口 秀雄 (Hideo Sakaguchi)
徳島大学 工学部 篠原 能材 (Yoshitane Shinohara)
徳島大学大学院 工学研究科 Tarmizi

1 はじめに

偏微分方程式を数値的に解く場合、単に数値計算が行えるだけではなく高精度の数値計算が要求されることがある。例えばある偏微分方程式は、一見定常に見えるが長い時間を経過した後に爆発現象がおきる、という性質を持っている。この例では定常と思われる部分の数値計算をかなりの高精度で行わないと爆発現象までシミュレートすることはできない。また、自由境界問題や逆問題においても、信頼できる解を得るためにかなり高い精度で数値計算を行わなければならないことがある。特に逆問題においては、数値計算の過程で誤差が指数的に増加する問題もあり [4]、通常の精度の計算では数値計算が不可能となることもある。

このような場合においても信頼できる解を得るために、偏微分方程式に対し任意の精度で数値計算を行う手法の開発が必要である。任意の精度で数値計算を行うためには、偏微分方程式の離散化で生じる打切誤差と、実数を計算機上で表現するときに生じる丸め誤差の二種類の誤差を要求されるだけ小さくする手法が必要である。

さて、偏微分方程式に対する数値解法としては、有限差分法、有限要素法、境界要素法等がよく知られている。それぞれの方法においては、離散化公式の次数を上げることにより打切誤差を小さくすることが可能であるが、これらの方法では次数が上がるにつれて離散化公式を導出する手続きが著しく煩雑になり、実用上打切誤差を任意に小さくすることが困難となる。そこで本研究では、離散化の精度を容易に上げることができるスペクトル法を用いる。その中でも特に、境界条件の組み込みが容易である Gauss-Lobatto 選点を用いたチェビシェフ多項式によるスペクトル選点法を用いる。また、丸め誤差を小さくする手法として、多倍長表現による浮動小数点数を用いた多倍長計算を行う。これにより任意の桁

数の浮動小数点演算が可能となり、丸め誤差を要求されるだけ小さくすることができる。本研究では、 Gauss-Lobatto 選点を用いたチェビシェフ多項式によるスペクトル選点法に多倍長計算を組み合わせることにより、打切誤差、丸め誤差を任意に小さくできる数値シミュレーション手法を提案するものである。ただしスペクトル法を用いるため、本研究で対象とする偏微分方程式の解は十分滑らかであるとの仮定が必要である。

2 スペクトル選点法

本研究で用いるスペクトル法では、チェビシェフ多項式、

$$T_k(x) = \cos(k \ \arccos x) \quad (-1 \leq x \leq 1) \tag{1}$$

を用いる。ここで、$k = 0, 1, 2, \cdots$である。これらの多項式を用いて関数を

$$\sum_{k=0}^{N} \tilde{u}_k \, T_k(x) \tag{2}$$

と近似する。Nがスペクトル法の次数である。また、本研究ではスペクトル法の中でも次式で定義される Gauss-Lobatto 選点、

$$x_j = \cos \frac{\pi j}{N} \qquad (j = 0, 1, \cdots, N) \tag{3}$$

を使用したスペクトル選点法を用いる。Gauss-Lobatto 選点では、境界上の点 $x = -1, 1$ に選点が存在するため、境界条件の処理が容易である。

さて、1 次元の関数 $u(x)$ に対し、(3) 式で定義された点 $x = x_j$ でのスペクトル法による近似値を u_jとおく。Gauss-Lobatto 選点によるスペクトル選点法では、u の一階導関数 u' と u_j の関係式、および二階導関数 u'' と u_j の関係式は次のように表される。

$$u'(x_l) \ = \ \sum_{j=0}^{N} (D_x)_{l,j} \ u_j, \tag{4}$$

$$u''(x_l) \ = \ \sum_{j=0}^{N} (D_{xx})_{l,j} \ u_j. \tag{5}$$

ここに現れる $N+1$ 次の正方行列 D_x, D_{xx}は次式で定義される。

$(D_x)_{l,j}$

$$= \begin{cases} \dfrac{\bar{c}_l}{\bar{c}_j} \dfrac{(-1)^{l+j+1}}{2\sin\frac{(l+j)\pi}{2N}\sin\frac{(l-j)\pi}{2N}} & l \neq j \\ -\dfrac{\cos\frac{j\pi}{N}}{2\sin^2\frac{j\pi}{N}} & 1 \leq l = j \leq N-1 \\ \dfrac{2N^2+1}{6} & l = j = 0 \\ -\dfrac{2N^2+1}{6} & l = j = N. \end{cases} \tag{6}$$

$(D_{xx})_{l,j}$

$$= \begin{cases} \dfrac{2}{3N\bar{c}_j}\displaystyle\sum_{k=0}^{N}\frac{k^2-1}{\bar{c}_k}k^2\cos\frac{kj\pi}{N} & l = 0 \\ \dfrac{2}{3N\bar{c}_j}\displaystyle\sum_{k=0}^{N}\frac{(-1)^k}{\bar{c}_k}k^2(k^2-1)\cos\frac{kj\pi}{N} & l = N \\ \dfrac{(-1)^{l+j+1}}{2\bar{c}_j\sin^2\frac{l\pi}{N}}\left\{\dfrac{\cos\frac{l\pi}{N}}{\sin\frac{(l+j)\pi}{2N}\sin\frac{(l-j)\pi}{2N}}\right. & \\ \qquad\left. +\dfrac{1}{\sin^2\frac{(l+j)\pi}{2N}}+\dfrac{1}{\sin^2\frac{(l-j)\pi}{2N}}\right\} & l \neq j,\ \ 1 \leq l \leq N-1 \\ \dfrac{-1}{2\bar{c}_j\sin^2\frac{l\pi}{N}}\left\{\dfrac{1+\cos^2\frac{l\pi}{N}}{\sin^2\frac{l\pi}{N}}+\dfrac{2N^2+1}{3}\right\} & 1 \leq l = j \leq N-1. \end{cases} \tag{7}$$

これら D_x, D_{xx}の要素は N, j, l が決まれば容易に計算できる。関係式(4)、(5) が Gauss-Lobatto 選点によるスペクトル選点法の離散化の公式であり、1 次元の問題であれば N次の離散化は $N+1$ 個の格子点により実現できる。離散化の次数を上げるには Nを大きくするだけでよく、容易に離散化の次数を上げることができる。なお、有限差分法と同じく、2 次元以上の離散化公式は 1 次元の離散化公式の重ね合わせによって得られる。

3 多倍長計算

多倍長実数の計算は多大な時間を必要とするため、その高速化が必要である。スペクトル選点法を用いた数値計算では、四則演算に加えて三角関数の計算が必要になる。しかし、三角関数の計算はTaylor展開の公式を用いれば四則演算に帰着できる。また除算は、

$$f(x) = \frac{1}{x} - a \tag{8}$$

に対するニュートン法、

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n(2 - ax_n) \tag{9}$$

を使って a の逆数を求め、それをかけることにより加減乗算に帰着できる。加減乗算の中では乗算に対する計算時間が、加減算に比べて大きい。加減算が多倍長表現での桁数に比例した計算時間ですむのに対して、通常の方法による乗算では、桁数の2乗に比例した計算時間が必要となる。しかし、これに対してはFFTを利用する方法が知られており [1]、計算時間を大幅に削減することができる。

本研究においては、FORTRANを用いてプログラムを作成した。多倍長実数を表現するためには2バイトの整数型の配列変数を使用した。例えば多倍長実数 x を表す場合、配列変数の中には次のように数を格納する。

$$\begin{aligned}
x(-1) &: \text{符号部 } (-1 \text{ or } 1) \\
x(0) &: \text{指数部 } (-16384 \sim 16383) \\
x(1) &: \text{実数部 } (0 \sim 9999) \\
x(2) &: \text{実数部 } (0 \sim 9999) \\
&\vdots \\
x(n) &: \text{実数部 } (0 \sim 9999)
\end{aligned}$$

これは、

$$x = x(-1)\cdot 10^{4x(0)}\cdot\{x(1)\cdot 10^{-4} + x(2)\cdot 10^{-8} + \cdots + x(n)\cdot 10^{-4n}\}$$

に対する多倍長表現となっている。なお、$x \neq 0$ に対しては、指数部を調節して $x(1) \neq 0$ となるように正規化する。この表現方法により $4n$ 桁の数が扱える。

さて、多倍長実数 x, y を前述の多倍長表現で表したときの実数部である $x(1), x(2), \cdots, x(n)$ と $y(1), y(2), \cdots, y(n)$ に対し、x と yの積の計算で必要となる

$$\begin{aligned} & x(n)y(n) \\ & x(n-1)y(n) + x(n)y(n-1) \\ & \vdots \\ & x(1)y(2) + x(2)y(1) \\ & x(1)y(1) \end{aligned}$$

を FFT を利用して求める方法を述べる。まず、n の 2 倍の長さの数列 $u_0, u_1, \cdots, u_{m-1}$ および $u_0, u_1, \cdots, u_{m-1}$ を用意する。ここで $m = 2n$ である。これらの値としては

$$u_k = \begin{cases} x(n-k) & k = 0, 1, \cdots, n-1 \\ 0 & k = n, n+1, \cdots, m-1 \end{cases} \tag{10}$$

$$v_k = \begin{cases} y(n-k) & k = 0, 1, \cdots, n-1 \\ 0 & k = n, n+1, \cdots, m-1 \end{cases} \tag{11}$$

を用いる。次にこれらの数に対して離散フーリエ変換

$$\hat{u}_j = \sum_{s=0}^{m-1} \zeta^{js} u_s, \quad \hat{v}_j = \sum_{t=0}^{m-1} \zeta^{jt} v_t \quad (j = 0, 1, \cdots, m-1) \tag{12}$$

を施す。ここで、

$$\zeta = e^{-\frac{2\pi \boldsymbol{i}}{m}} \tag{13}$$

である。さらに、

$$\hat{w}_j = \hat{u}_j \hat{v}_j \quad (j = 0, 1, \cdots, m-1) \tag{14}$$

によってそれぞれの積 $\hat{w}_j$ を計算し、最後に離散逆フーリエ変換

$$w_k = \frac{1}{m} \sum_{j=0}^{m-1} \zeta^{-kj} \hat{w}_j \quad (k = 0, 1, \cdots, m-1) \tag{15}$$

を施す。最後に得られた $w_0, w_1, \cdots, w_{m-1}$ は

$$\begin{aligned} w_0 &= x(n)y(n) \\ w_1 &= x(n-1)y(n) + x(n)y(n-1) \\ &\vdots \\ w_{m-2} &= x(1)y(2) + x(2)y(1) \\ w_{m-1} &= x(1)y(1) \end{aligned}$$

を満たし、これにより多倍長実数 x と yの積が計算できる。この手順中、2 回の離散フーリエ変換と 1 回の離散逆フーリエ変換は FFT を利用しての高速計算が可能であり、多倍長の積の計算を高速に行うことができる。実際の FFT の計算では、複素数を用いて本来の FFT の計算を行う方法と、ある素数を法とする有限体上で FFT を行う二つの方法がある。本研究では、整数演算を行うことで数値誤差を混入させずに直接計算することができる後者の方法を使用した。また、大きな数の計算を避けるために、複数の素数を法とするそれぞれの有限体上で FFT を実行し、最後にここで使用したすべての素数の積を法とする整数を合同式を解いて求める、という方法を用いた。なお、この合同式の解の存在および一意性は中国の剰余定理 (Chinese Remainder Theorem) により保証されている。また、FORTRAN プログラムでは、素数を法とする計算を 4 バイトの整数型の変数を用いて行い、すべての素数の積を法とする整数を求める計算のみ 8 バイトの実数型の変数を用いて行った。実際にこの方法により、3 個の素数を用いて 64、128、256、512、1024、2048、4096 桁の 7 種類の多倍長実数用のプログラムを作成した。

4 数値計算結果

本報告では、次の 1 次元境界値問題においてスペクトル法と多倍長計算の組み合わせにより高精度数値計算が可能であることを示す。

$$u_{xx} = -\frac{\pi^2}{16}\sin\frac{(x+1)\pi}{4} \qquad \text{in } (-1 < x < 1), \tag{16}$$

$$u(-1) = 0, \qquad u_x(1) = 0. \tag{17}$$

このモデル問題に対し、Gauss-Lobatto 選点によるスペクトル選点法を 1024 桁の多倍長実数を用いて実行した。ここで用いた境界値問題の厳

密解は、

$$u(x) = \sin \frac{(x+1)\pi}{4} \tag{18}$$

であり、スペクトル法で得られた数値解とこの厳密解の最大誤差により計算結果を評価した。

数値解の厳密解に対する最大誤差を表1に示す。ここでNはスペクトル法の次数を表す。N次のスペクトル選点法では、使用する格子点数は$N+1$個であり、最大誤差は(4)、(5)式から得られる$N+1$元の連立一次方程式を数値的に解いた結果である。なお、連立一次方程式の解法として、Gaussの消去法を用いた。

Table 1. Maximum errors for the model problem.

N	最大誤差	N	最大誤差
10	4.89×10^{-11}	120	6.38×10^{-248}
20	6.64×10^{-27}	130	5.76×10^{-273}
30	5.39×10^{-45}	140	2.42×10^{-298}
40	1.54×10^{-64}	150	4.97×10^{-324}
50	3.62×10^{-85}	160	5.25×10^{-350}
60	1.16×10^{-106}	170	2.98×10^{-376}
70	8.01×10^{-129}	180	9.38×10^{-403}
80	1.03×10^{-151}	190	1.70×10^{-429}
90	4.32×10^{-175}	200	1.82×10^{-456}
100	6.01×10^{-199}	250	2.24×10^{-594}
110	3.08×10^{-223}	300	1.20×10^{-736}

このモデル問題に対しては、スペクトル法の次数の約2倍程度の有効桁数の数値計算が可能であることがわかる。特に、$N = 300$の場合には有効桁数が700桁以上という数値計算結果が得られた。また多倍長実数を

使用しない場合、4倍精度の数値計算でもすでに $N = 30$ で丸め誤差の影響によりスペクトル法が本来持っている精度を保つことができなくなることもわかった。

なお N がいずれの値の時も、$x = 1$ での関数 u の数値解と厳密解の絶対誤差の値が最大誤差となった。

5 まとめ

チェビシェフ多項式を使ったスペクトル選点法に多倍長計算を組み合わせることにより、任意精度の数値シミュレーションを行う手法を提案した。また、1次元の境界値問題にその方法を適用し、数百桁の精度の数値計算が可能であることを示した。さらに、通常のプログラム言語が持っている精度の実数を使用している限りは、丸め誤差の影響により比較的低い次数のスペクトル法においてもスペクトル法が本来持っている精度を達成できないことがわかった。スペクトル法を有効活用するためにも、多倍長計算により十分な桁数で計算を行う本研究の手法は有効であるといえる。今後は2次元の問題に対して適用していく予定である。

なお、本研究の一部は、文部省科学研究費 (基盤研究 (B)、課題番号 09440080) の支援を受けて行ったものである。

参考文献

[1] Knuth, D. E., The Art of Computer Programming, Vol.2: Seminumerical Algorithms, Addison-Wesley, Reading, Massachusetts, 1981.

[2] Canuto, C. et al., Spectral Methods in Fluid Dynamics, Springer, 1988.

[3] A Practical Method for an Ill-Conditioned Optimal Shape Design of a Vessel in Which Plasma is Confined, A.Sasamoto, H.Imai and H.Kawarada, Inverse Problems in Engineering Sciences, Springer-Verlag, pp.120-125, 1991.

[4] An application of the Fuzzy Theory for an Ill-Posed Problem, H.Imai, A.Sasamoto, H.Kawarada and M.Natori, Inverse Problems in Engineering Mechanics, Springer-Verlag, pp.31-38, 1993.

[5] application of the Fuzzy Theory and Spectral Collocation Methods to an Ill-Posed Shape Design Problem with a Free Boundary, H.Imai, Inverse Problems in Mechanics, American Soc. Mech. Eng. AMD-Vol.186, pp.103-107, 1994.