浙江大學

本 科 毕 业 论 文 (设 计)

文献综述和开题报告

题目	谱方法某些结果的数值验证
姓名与学号	3080100913 李言迪
指导教师	叶兴德
年级与专业	2008 数学与应用数学
所在学院	理学院

一、题目：谱方法某些结果的数值验证

二、指导教师对开题报告、外文翻译和文献综述的具体要求：

文献综述：要求对谱方法进行综述，特别是Chebyshev方法，为毕业论文作一些必要的准备。

开题报告：要求对Chebyshev-谱tau方法作出详细的介绍，特别是如何利用Chebyshev多项式的递推关系，用不同方法实现tau方法。对如何设计拟三对角方程组的求解算法的思路作出说明。

指导教师（签名）：

年　　月　　日

目录

毕业设计—文献综述部分
谱方法

1 概述和分类

谱方法是一类求解微分方程的方法。他的基本想法是把微分方程的待求解用截断级数展开来近似，然后目的是确定这组展开式系数，使这个近似的函数在某种意义下尽可能满足微分方程和边界条件。谱方法的理论原型于 Fourier 级数方法，后来使用的基函数选取扩展到了 Chebyshev 正交多项式，Lengendre 正交多项式等。

谱方法具有比有限差分和有限元方法更高的精度并具有指数型的快速收敛性。其理论早在 1944 年就已被提出，但是问题在于，为了达到希望的精度，作级数展开需要较大的计算量；因此在计算的精度和速度提高之前，谱方法相比于有限差分法和有限元方法并没有大规模地应用。但是，到了 1970 年后，随着计算机水平的提高以及联系谱系数和函数值的高效算法—快速 Fourier 变换（FFT）—的出现，谱方法的又重新得到了重视，并从 90 年代开始成为主流方法。

正如前面所说，谱方法可以根据问题的性质选取不同的基函数，因而按照这个标准，可以分为 Fourier 方法，Chebyshev 方法等。Fourier 方法适用于周期性的问题；而对于非周期问题，则是需要用 Chebyshev 方法，可以避免像 Fourier 方法那样在边界出现 Gibbs 现象。而且，实际上这两种方法在本质上是一脉相承的，因此 Chebyshev 方法同时也享有着 Fourier 方法的优点，即它的指数收敛性以及可用 FFT 算法的通道。最后，Lengendre 方法也是可选的一种。此法在算子离散和数值积分上有一些不错的性质，然而由于没有找到一个快速的变换公式，因此应用还不普遍。

经典的谱方法有三个版本，即“Galerkin 型”的两个方法（经典 Galerkin 法，谱 tau 法）和配置法（collocation）。这三个版本的区别需要从谱方法求解微分方程的过程，及如何应用赋权剩余方法（Finlayson and Scriven (1966)）说起。对微分方程 $Lu - f = 0$ 的待求解 $u(x)$ 做一个截断级数展开的近似，

$$u_N(x) = \sum_{k=0}^{N} \hat{u}_k \varphi_k(x), \quad \alpha \le x \le \beta,$$

其中的基函数（记为 φ_k）可以是三角基函数 e^{ikx}，Chebyshev 多项式 $T_k(x)$ 或 Lengendre 多项式 $L_k(x)$（并在各自对应的权函数 w 下单位正交）。我们希望余项

$$R_N(x) = Lu_N - f,$$

即使不能期待恰为 0（那样近似就是准确的了），但至少在下面的“弱”的意义下为 0，即再选定一个内积的“检验权函数”w_* 和“检验基函数”（记为 ψ_i），对 $\forall i$ 成立

$$(R_N, \psi_i)_{w_*} = \int_{\alpha}^{\beta} R_N \psi_i w_* dx = 0, \quad i \in I_N, \tag{1.1}$$

这里面的检验权函数 w_* 和检验基函数 ψ_i 的选取，就决定了谱方法的另一种分类：

1. “Galerkin”型的方法，包括经典 Galerkin 法和谱 tau 法，是指选取基函数 φ_i 和其定义内积相应的权函数 w，为检验基函数 ψ_i 和检验权函数 w_*。区别在于，在经典 Galerkin 法中，要求基函数本身满足部分或者全部的边界条件；而谱 tau 方法是指基函数并不需满足边界条件，它会通过增加一组描述边界条件的方程来引入边界信息。

2. 配置法的检验权函数和检验基函数如下选择：

$$\psi_i = \delta(x - x_i), \quad w_* = 1, \tag{1.2}$$

其中 δ 为 Dirac delta 函数，点 x_i 在 $[\alpha, \beta]$ 中特定选取。将 (1.2) 代入 (1.1) 会得到配置法满足一个等价的插值条件

$$R_N(x_i) = 0,$$

小结一下，可以说“Galerkin 型“的方法是让余项在弱平均的意义下为 0，而配置法是让余项在特定点集上为 0。并以此为基础，要最终确定截断级数的展式系数。

2 研究历史和现状

配置法雏形的可能是Slater (1934) 和L.V. Kantorovic (1934) 提出的。由R.A. Frazer (1937) 发展为了求解常微分方程的一般方法。Lanczos 首先指出解的精度取决于如何选择基函数以及如何选择配置点，建立了正交配置方法的基础。20 年后，Clenshaw and Society (1957),Clenshaw and Norton (1963) 和Wright (1964) 重新拾起了配置法，并开始使用 Chebyshev 多项式作为基函数，应用到初值问题上；Villadsen and Stewart (1967) 应用到边值问题上。将之发展为求解周期性的偏微分方程的是Kreiss and Oliger (1972) 和Orszag (1972)。配置法的优势在于，它可以很容易地处理变系数的线性微分方程，甚至是非线性问题。

Galerkin 型的方法曾于Silberman (1954) 提出来解决偏微分方程，这也是最先应用到偏微分方程的谱方法。然而，直到Orszag (1969, 1970) 和Eliasen et al. (1970) 设计出计算二次非线性项的卷积变换公式之后，“Galerkin 型”方法才作为一个可以高精度计算的方法站上历史舞台。但是，如果涉及更复杂的非线性性，“Galerkin 型”方法仍然不如受影响更小的有限元方法。对于周期性问题，可以选用 Fourier 三角函数基，应用经典 Galerkin 方法；而对于非周期性问题，则可以选用 Chebyshev 多项式等，应用谱 tau 方法。Lanczos (1938) 设计了谱 tau 方法。然后，Orszag (1971) 应用 Chebyshev-tau 方法高精度地求解了一个流体力学的问题。这一成功使得谱 tau 方法开始在数值计算的诸多方面得到广泛应用，比如计算特征值，求解常系数微分方程等。到了 80 年代中期，谱方法开始和 Gauss 积分公式结合来，并选用 Gauss 点作为配置法的插值点(Gottlieb and Orszag (1993)，Mercier (1989))。在同一时期，对各种方法的稳定性和收敛性的理论结果得以给出。到 80 年代末，经典谱方法已经成熟，并成为流体物理学研究湍流的主流方法了。

90 年代后，谱方法的研究转向了复杂几何定义域上的问题。Karniadakis and Sherwin (1999) 给出了一个应用谱元方法到有结构域和无结构域的一个统一框架。Canuto et al. (2007) 则给出了更多关于复杂域上问题的研究进展。最后，Canuto et al. (2006); Peyret (2002) 可以作为谱方法的一个非常全面的参考。

参考文献

C. Canuto, MY Hussaini, A. Quarteroni, and TA Zang. *Spectral methods: fundamentals in single domain*. Springer, 2006. 2

C. Canuto, MY Hussaini, A. Quarteroni, and TA Zang. *Spectral methods: evolution to complex domains and applications to fluid dynamics*. Springer, NY, 2007. 2

CW Clenshaw and HJ Norton. The solution of nonlinear ordinary differential equations in chebyshev series. *The Computer Journal*, 6(1):88--92, 1963. 2

CW Clenshaw and Cambridge Philosophical Society. *The numerical solution of linear differential equations in Chebyshev series*. Cambridge Univ Press, 1957. 2

E. Eliasen, B. Machenhauer, and E. Rasmussen. *On a numerical method for integration of the hydrodynamical equations with a spectral representation of the horizontal fields*. 1970. 2

BA Finlayson and LE Scriven. The method of weighted residual--a review. *Appl. Mech. Rev*, 19(9):735--748, 1966. 1

D. Gottlieb and S.A. Orszag. *Numerical analysis of spectral methods: theory and applications*, volume 26. Society for industrial and applied mathematics, 1993. 2

G. Karniadakis and S.J. Sherwin. *Spectral/hp element methods for CFD*. Oxford University Press, USA, 1999. 2

H.O. Kreiss and J. Oliger. Comparison of accurate methods for the integration of hyperbolic equations. *Tellus*, 24(3): 199--215, 1972. 2

C. Lanczos. Trigonometric interpolation of empirical and analytical functions. *J. Math. Phys*, 17:123--199, 1938. 2

L.V. Kantorovic. On a new method of approximate solution of partial differential equations. *Dokl. Akad. Nauk SSSR*, 4:532--536, 1934. 2

B. Mercier. *An introduction to the numerical analysis of spectral methods*. Springer-Verlag New York, Inc., 1989. 2

S.A. Orszag. Numerical methods for the simulation of turbulence. *Physics of Fluids*, 12(12):II--250, 1969. 2

S.A. Orszag. Transform method for the calculation of vector-coupled sums: Application to the spectral form of the vorticity equation. *Journal of Atmospheric Sciences*, 27:890--895, 1970. 2

S.A. Orszag. Accurate solution of the orr-sommerfeld stability equation. *J. Fluid Mech*, 50(4):689--703, 1971. 2

S.A. Orszag. Comparison of pseudospectral and spectral approximation. *Stud. Appl. Math*, 51(3):253--259, 1972. 2

R. Peyret. *Spectral methods for incompressible viscous flow*, volume 148. Springer Verlag, 2002. 2

S.W. Skan R.A. Frazer, W.P. Jones. Approximation to Functions and to the Solution of Differential Equations. *Rep. and Mem*, 1937. 1799 (Aeronautical Research Council, London). 2

I. Silberman. Planetary waves in the atmosphere. *Journal of Atmospheric Sciences*, 11:27--34, 1954. 2

J.C. Slater. Electronic energy bands in metals. *Physical Review*, 45(11):794, 1934. 2

JV Villadsen and WE Stewart. Solution of boundary-value problems by orthogonal collocation. *Chemical Engineering Science*, 22(11):1483--1501, 1967. 2

K. Wright. Chebyshev collocation methods for ordinary differential equations. *The Computer Journal*, 6(4):358--365, 1964. 2

毕业设计—开题报告部分

1 背景和动机概述

本次毕业设计属于数值分析领域的谱方法这一块内容。谱方法的原型是用 Fourier 级数展开方法求解微分方程。它具有比有限差分和有限元方法更高的精度，并具有指数型的快速收敛性。其理论早在 1944 年就已被提出，但是问题在于，为了达到希望的精度，作级数展开需要较大的计算量；因此在计算的精度和速度提高之前，谱方法相比于有限差分法和有限元方法应用得更少。但是，到了 1970 年后，随着计算机水平的提高以及联系谱系数和函数值的高效算法—快速 Fourier 变换（FFT）—的出现，谱方法的又重新得到了重视，并从 90 年代开始成为主流方法。本文着眼于用谱方法来求解常微分方程的边值问题，便是其诸多应用当中的一种。

谱方法是用截断的级数展开来作近似，会根据问题的性质选取不同的基函数。因而按照这个标准，可以分为选用三角函数基的 Fourier 方法，选用 Chebyshev 正交多项式的 Chebyshev 方法等。Fourier 方法适用于周期性的问题；而对于非周期问题，则是需要用 Chebyshev 方法，可以避免像 Fourier 方法那样在边界出现 Gibbs 现象。而且，实际上这两种方法在本质上是一脉相承的，因此 Chebyshev 方法同时也享有着 Fourier 方法的优点，即它的指数收敛性以及可用 FFT 算法的通道。

使用谱方法求解微分方程时，微分方程是在一定的意义下被级数展式所满足；根据这个意义的不同选择，可以得到谱方法的三种主要版本，即“Galerkin 型”的两个方法（经典 Galerkin 法，谱 tau 法）和配置法（collocation）。本次毕业设计就是在 Chebyshev-谱 tau 方法的框架下进行的，具体 Chebyshev-谱 tau 方法会放在在第2.1节介绍。更多关于谱方法类型的介绍，请参看文献综述部分。

Canuto 等人的书 [1] 对谱方法的理论和应用做了非常详尽讨论。在这本书中，有一段关于 Chebyshev-tau 方法在求解常微分 Helmholtz 方程的边值问题的两种处理方法的记载。简单地说，是先求解未知函数，然后通过“谱微分”运算给出 1、2 阶导函数；还是先求其二阶导函数，再通过“谱积分”运算给出 1 阶导函数和未知函数。对于这两种算法，书中援引了 Greengard[2] 的一些相关结论，即“谱微分”运算过程会放大初始误差 $O(N^2)$ 倍，而“谱积分”运算只会放大 2.4 倍，从而猜测，在 1、2 阶导函数的计算上，使用“谱微分”的算法会在实际计算中不如“谱积分”的算法。这一猜测尚没有学者给出一个验证，本次毕设的主题就通过数值实验测试这两种算法的优劣。

2 问题详述

2.1 Chebyshev-tau 方法

2.1.1 Chebyshev 多项式

记 k 次的第一类 Chebyshev 多项式为 $T_k(x)$, 它们定义在 $x \in [-1, 1]$，满足

$$T_k(x) = \cos(k\cos^{-1}x), k = 0, 1, 2, ... \tag{2.1.1}$$

很明显，其值域为 $|T_k(x)| \leq 1$。对 Chebyshev 多项式来说，在理论和计算中更多用到下面的等价形式，即设 $x = \cos z$，有

$$T_k(z) = \cos kz \tag{2.1.2}$$

由该式不难写出 Chebyshev 多项式的前几项为

$$T_0 = 1,\ T_1 = \cos z = x,\ T_2 = \cos 2z = 2\cos^2 z - 1 = 2x^2 - 1, ...$$

2.1.2 两点边值问题

我们将应用 Chebyshev-tau 方法考虑如下的 Helmholtz 方程的 Dirichlet 问题，

$$\begin{cases} -\dfrac{d^2u}{dx^2} + \lambda u = f, & x \in (-1, 1) \\ u(-1) = a, \quad u(1) = b. \end{cases} \tag{2.1.3}$$

Chebyshev-tau 方法是这样进行的。即首先对微分方程的未知函数 $u(x)$ 和右端非齐次项 $f(x)$，作 N 阶截断的 Chebyshev 展开式，即

$$u_N(x) = \sum_{k=0}^{N} \hat{u}_k T_k(x), \quad f_N(x) = \sum_{k=0}^{N} \hat{f}_k T_k(x),$$

其中，展开式的系数我们称为（0 阶）谱系数。

然后，Chebyshev 多项式构成了一组单位正交基。我们让 $u(x)$ 的逼近 $u_N(x)$ 在弱的意义下满足微分方程，意即在不超过 $N-2$[1]次的 Chebyshev 多项式构成的子空间中成立，即

$$\int_{-1}^{1} \left[-\frac{d^2u_N}{dx^2} + \lambda u_N - f \right] \cdot T_k(x) w(x) dx = 0, \quad k = 0, 1, \ldots, N-2, \tag{2.1.4}$$

其中 $w(x) = (1-x^2)^{-1/2}$ 为内积的权函数。若记 $\hat{u}_k^{(2)}$ 表示 $u_N''(x)$ 的谱系数，即 2 阶谱系数。则上式由 $\{T_k\}_{k=0,\ldots,N}$ 的单位正交性可以写成

$$-\hat{u}_k^{(2)} + \lambda \hat{u}_k = \hat{f}_k, \quad k = 0, 1, \ldots, N-2, \tag{2.1.5}$$

最后，由 Chebyshev 多项式性质 $T_k(1) = 1$，$T_k(-1) = (-1)^k$，得到边界条件表达为

$$\sum_{k=0}^{N} \hat{u}_k = a, \quad \sum_{k=0}^{N} (-1)^k \hat{u}_k = b. \tag{2.1.6}$$

通过 (2.1.5) 和 (2.1.6) 两式求解弱解 $u_N(x)$ 是 Chebyshev-tau 方法下面需要求解的问题。

[1]考虑到我们仅需要 $N+1$ 个方程来确定 $N+1$ 个待定的谱系数 $\{\hat{u}_k\}_{k=0,\ldots,N}$，因此，(2.1.4)，(2.1.5) 中的 k 只要在 $0, \ldots, N-2$ 中成立，即构成一个恰好适定的线性方程组。

2.2 谱系数的递推关系

我们需要知道 $u_N(x)$ 的导函数在 Chebyshev 基下展开的谱系数(高阶谱系数),并将看到,高阶谱系数可以由原函数的谱系数以线性组合的方式给出。这是我们两种算法的基础。设

$$u'_N(x) = \sum_{k=0}^{N} \hat{u}_k T'_k(x) = \sum_{k=0}^{N} \hat{u}_k^{(1)} T_k(x) \tag{2.2.1}$$

利用三角函数公式得到的递推关系 $\frac{T'_{k+1}}{k+1} - \frac{T'_{k-1}}{k-1} = 2T_k$,可以得到

$$T'_k(x) = 2k \sum_{n=0}^{[(k-1)/2]} \frac{1}{c_{k-1-2n}} T_{k-1-2n}(x) \tag{2.2.2}$$

把 (2.2.2) 代入 (2.2.1) 可以导出一个 1 阶谱系数 $\hat{u}_k^{(1)}$ 的表达式:

$$\hat{u}_k^{(1)} = \frac{2}{c_k} \sum_{\substack{p=k+1 \\ p:=p+2}}^{N} p\hat{u}_p, k = 0, \ldots, N-1, \text{且有} \hat{u}_N^{(1)} = 0. \tag{2.2.3}$$

一般的 p 阶与 $p-1$ 阶谱系数也存在同样的递推关系。这便是由低阶谱系数获得高阶谱系数的“谱微分运算”。该关系也可以写成所谓微分矩阵的形式 $\hat{U}^{(1)} = \hat{D}\hat{U}$。递推用两次,也可以得到 2 阶谱系数计算如下

$$\hat{u}_k^{(2)} = \frac{1}{c_k} \sum_{\substack{p=k+2 \\ p:=p+2}}^{N} p(p^2 - k^2)\hat{u}_p, k = 0, \ldots, N-2, \text{且有} \hat{u}_N^{(2)} = \hat{u}_{N-1}^{(2)} = 0. \tag{2.2.4}$$

同样的公式 (2.2.3),隔项作差,又可以得到由高阶谱系数获得低阶谱系数的“谱积分运算”。

$$\hat{u}_k^{(p-1)} = \frac{1}{2k} \left(c_{k-1}\hat{u}_{k-1}^{(p)} - \hat{u}_{k+1}^{(p)} \right), \quad k \geq 1 \tag{2.2.5}$$

且有末端项为,1 阶谱系数时,

$$\hat{u}_N^{(1)} = 0, \ \hat{u}_{N-1}^{(1)} = 2N\hat{u}_N. \tag{2.2.6}$$

2 阶谱系数时,

$$\hat{u}_N^{(2)} = \hat{u}_{N-1}^{(2)} = 0, \ \hat{u}_{N-2}^{(2)} = 2(N-1)\hat{u}_{N-1}^{(1)} = 4N(N-1)\hat{u}_N. \tag{2.2.7}$$

2.3 两种求解过程概述

2.3.1 方法 I:化未知量为 0 阶谱系数

虽然我们可以利用 (2.2.4) 来直接消去 (2.1.5) 中的 2 阶谱系数,但是这种方式构成的线性系统一方面需要 $O(N^2)$ 阶的较多运算量,一方面也很可能是比较奇异的(Greengard[2])。同样是保留 0 阶谱,一种更有效率的方式如 Canuto 等人 [1] 所述。利用 $q = 2$ 时的三项递推关系 (2.2.5)

$$2k\hat{u}_k^{(1)} = c_{k-1}\hat{u}_{k-1}^{(2)} - \hat{u}_{k+1}^{(2)}, \quad k = 1, \ldots, N-1$$

将 (2.1.5)(此式对 $k=0,\dots,N-2$ 成立)代入得到

$$2k\hat{u}_k^{(1)} = c_{k-1}\left(-\hat{f}_{k-1}+\lambda\hat{u}_{k-1}\right)-\left(-\hat{f}_{k+1}+\lambda\hat{u}_{k+1}\right),\quad k=1,\dots,N-3.$$

再代入 $q=1$ 时的递推 (2.2.5),消去 1 阶谱系数得到

$$\begin{aligned}2k\hat{u}_k &= \frac{c_{k-1}}{2(k-1)}\left[c_{k-2}(-\hat{f}_{k-2}+\lambda\hat{u}_{k-2})-(-\hat{f}_k+\lambda\hat{u}_k)\right]\\ &\quad -\frac{1}{2(k+1)}\left[(-\hat{f}_k+\lambda\hat{u}_k)-(-\hat{f}_{k+2}+\lambda\hat{u}_{k+2})\right],\quad k=2,\dots,N-4.\end{aligned}$$

略去一些细节的讨论(包括 $k=N-3,\dots,N$ 的情形),最终得到

$$\begin{aligned}&\frac{c_{k-2}\lambda}{4k(k-1)}\hat{u}_{k-2}+\left(-1-\frac{\lambda\beta_k}{2(k^2-1)}\right)\hat{u}_k+\frac{\lambda\beta_{k+2}}{4k(k+1)}\hat{u}_{k+2}\\ =\;&\frac{c_{k-2}}{4k(k-1)}\hat{f}_{k-2}+\left(-\frac{\beta_k}{2(k^2-1)}\right)\hat{f}_k+\frac{\beta_{k+2}}{4k(k+1)}\hat{f}_{k+2},\quad k=2,\dots,N.\end{aligned}\tag{2.3.1}$$

其中

$$\beta_k=\begin{cases}1, & 0\le k\le N-2,\\ 0, & k>N-2.\end{cases}$$

然后加上奇、偶项分离的边界条件

$$\sum_{\substack{k=0\\ k\text{为偶数}}}^{N}\hat{u}_k=\frac{b+a}{2},\qquad \sum_{\substack{k=0\\ k\text{为奇数}}}^{N}\hat{u}_k=\frac{b-a}{2},\tag{2.3.2}$$

我们得到,方法 I 需要求解如下的线性系统:

$$\begin{pmatrix}1&1&1&\dots&&&1\\ *&*&*&&&&\\ &*&*&*&&&\\ &&&\dots&&&\\ &&&*&*&*&\\ &&&&*&*&\\ &&&&&*&*\end{pmatrix}\begin{pmatrix}\hat{u}_0\\ \hat{u}_2\\ \hat{u}_4\\ \vdots\\ \hat{u}_{N\text{偶}-4}\\ \hat{u}_{N\text{偶}-2}\\ \hat{u}_{N\text{偶}}\end{pmatrix}=\begin{pmatrix}\frac{b+a}{2}\\ \hat{g}_2\\ \hat{g}_4\\ \vdots\\ \hat{g}_{N\text{偶}-4}\\ \hat{g}_{N\text{偶}-2}\\ \hat{g}_{N\text{偶}}\end{pmatrix},\quad\begin{pmatrix}1&1&1&\dots&&&1\\ *&*&*&&&&\\ &*&*&*&&&\\ &&&\dots&&&\\ &&&*&*&*&\\ &&&&*&*&\\ &&&&&*&*\end{pmatrix}\begin{pmatrix}\hat{u}_1\\ \hat{u}_3\\ \hat{u}_5\\ \vdots\\ \hat{u}_{N\text{奇}-4}\\ \hat{u}_{N\text{奇}-2}\\ \hat{u}_{N\text{奇}}\end{pmatrix}=\begin{pmatrix}\frac{b-a}{2}\\ \hat{g}_3\\ \hat{g}_5\\ \vdots\\ \hat{g}_{N\text{奇}-4}\\ \hat{g}_{N\text{奇}-2}\\ \hat{g}_{N\text{奇}}\end{pmatrix},\tag{2.3.3}$$

其中,$N_{\text{偶}}$($N_{\text{奇}}$)表示最大偶(奇)数项,$*$ 表示系数矩阵中非零项,$\hat{g}_k$ 是 (2.3.1) 的相应 k 时的右端项。我们定义上面的系统为“拟三对角系统"。当求出解后,进行两次“谱微分”运算,便可以得到 1、2 阶谱系数,从而得到 1、2 阶导函数的截断级数近似。

2.3.2 方法 II:化未知量为 2 阶谱系数

针对谱空间中离散的方程 (2.1.5) 的另一种处理方式(见 Canuto 等 [1]),是利用低阶谱系数与高阶谱系数的递推关系在 (2.1.5) 中消去 0 阶谱系数,而保留 2 阶谱系数。

调用两次谱积分运算 (2.2.5)，可以得到这个递推关系：

$$\hat{u}_k = P_k\hat{u}_{k-2}^{(2)} + Q_k\hat{u}_k^{(2)} + R_k\hat{u}_{k+2}^{(2)}, \quad 2 \le k \le N, \tag{2.3.4}$$

其中$P_k = \dfrac{c_{k-2}}{4k(k-1)},\ Q_k = \dfrac{-e_{k+2}}{2(k^2-1)},\ R_k = \dfrac{e_{k+4}}{4k(k+1)}, \quad e_j = \begin{cases} 1, & j \le N, \\ 0, & j > N. \end{cases}$

将 (2.3.4) 代入方程 (2.1.5)，略去一些细节，化简可得：

$$\begin{aligned} \lambda\hat{u}_0 - \hat{u}_0^{(2)} &= \hat{f}_0, \\ \lambda\hat{u}_0^{(1)} + \left(-1 - \frac{\lambda}{8}\right)\hat{u}_1^{(2)} + \frac{\lambda}{8}\hat{u}_3^{(2)} &= \hat{f}_1, \\ \frac{\lambda}{4k(k-1)}\hat{u}_{k-2}^{(2)} + \left(-1 - \frac{\lambda}{2(k^2-1)}\right)\hat{u}_k^{(2)} + \frac{\lambda\beta_{k+2}}{4k(k+1)}\hat{u}_{k+2}^{(2)} &= \hat{f}_k, \quad k = 2, \dots, N-2. \end{aligned} \tag{2.3.5}$$

然后有边界条件为

$$\begin{aligned} \hat{u}_0 + \frac{1}{4}\hat{u}_0^{(2)} - \frac{7}{48}\hat{u}_2^{(2)} + \sum_{\substack{k=4 \\ k\text{为偶}}}^{N-2} \frac{3}{(k^2-1)(k^2-4)}\hat{u}_k^{(2)} &= \frac{b+a}{2}, \\ \hat{u}_0^{(1)} - \frac{1}{12}\hat{u}_1^{(2)} + \sum_{\substack{k=3 \\ k\text{为奇}}}^{N-2} \frac{3}{(k^2-1)(k^2-4)}\hat{u}_k^{(2)} &= \frac{b-a}{2}. \end{aligned} \tag{2.3.6}$$

我们得到，方法 II 需要求解如下的线性系统：

$$\begin{pmatrix} 1 & \frac{1}{4} & \frac{-7}{48} & \dots & * & * & * \\ * & * & & & & & \\ & * & * & * & & & \\ & & & \dots & & & \\ & & & * & * & * & \\ & & & & * & * & * \\ & & & & & * & * \end{pmatrix} \begin{pmatrix} \hat{u}_0 \\ \hat{u}_0^{(2)} \\ \hat{u}_2^{(2)} \\ \vdots \\ \hat{u}_{N\text{偶}-6}^{(2)} \\ \hat{u}_{N\text{偶}-4}^{(2)} \\ \hat{u}_{N\text{偶}-2}^{(2)} \end{pmatrix} = \begin{pmatrix} \frac{b+a}{2} \\ \hat{f}_0 \\ \hat{f}_2 \\ \vdots \\ \hat{f}_{N\text{偶}-6} \\ \hat{f}_{N\text{偶}-4} \\ \hat{f}_{N\text{偶}-2} \end{pmatrix}, \begin{pmatrix} 1 & \frac{-1}{12} & * & \dots & * & * & * \\ * & * & * & & & & \\ & * & * & * & & & \\ & & & \dots & & & \\ & & & * & * & * & \\ & & & & * & * & * \\ & & & & & * & * \end{pmatrix} \begin{pmatrix} \hat{u}_0^{(1)} \\ \hat{u}_1^{(2)} \\ \hat{u}_3^{(2)} \\ \vdots \\ \hat{u}_{N\text{奇}-6}^{(2)} \\ \hat{u}_{N\text{奇}-4}^{(2)} \\ \hat{u}_{N\text{奇}-2}^{(2)} \end{pmatrix} = \begin{pmatrix} \frac{b-a}{2} \\ \hat{f}_1 \\ \hat{f}_3 \\ \vdots \\ \hat{f}_{N\text{奇}-6} \\ \hat{f}_{N\text{奇}-4} \\ \hat{f}_{N\text{奇}-2} \end{pmatrix}, \tag{2.3.7}$$

其中，$N_{\text{偶}}$（$N_{\text{奇}}$）表示最大偶（奇）数项，$*$ 表示系数矩阵中非零项。我们看到，这也是一个拟三对角系统。当求出解后，进行两次“谱积分”运算，可以得到 0、1 阶谱系数，从而得到未知函数及其导函数的截断级数近似。

3 研究方案和可行性

3.1 研究方案

我们计划把研究分为如下几个阶段

1. 根据第2节的介绍，我们注意到两种过程首先都是面对一个拟三对角系统，形如 (3.1.1)。我们知道采用常规的 Gauss 消元法的运算量是 $O(N^3)$。由于我们的问题中 N 可能是比较大的，研究首先需要设计一种处

理拟三对角系统的高效、稳定的算法。

$$\begin{pmatrix} * & * & * & \dots & * & * \\ * & * & * & & & \\ & * & * & * & & \\ & & & \dots & & \\ & & & * & * & * \\ & & & & * & * \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \\ \vdots \\ \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \\ \vdots \\ \\ b_n \end{pmatrix}. \tag{3.1.1}$$

现在已有思路如下。我们其实本质上仍然是按照 Gauss 消元法的步骤进行，只是需要省去一些浪费在“零乘”和“零加”的浪费即可。类比于“追赶法”的一种处理 (追赶法参看 [4])，可以对矩阵作 LU 分解，然后通过待定系数的想法，按照特定的顺序，可依次得到 L 和 U 矩阵。最后，求解 $Ly = b$ 和 $Ux = y$ 即可。

2. 用 MATLAB 编程，完整实现我们的两种算法，并求解实际的例子。

3. 通过尝试不同的实例，作一个比较。
   由于方法 I 直接求解 0 阶谱，没有误差放大，直观上会比方法 II 放大两次后的结果精确；同理，方法 II 在 2 阶谱上会精确。因此，我们比较关心在 1 阶谱的计算上，哪种方法表现更好。验证 Greengard 的猜测是否正确。

3.2 难点分析

1. 在阅读文献 Canuto 的 [1] 和 Peyret 的 [3] 时，发现文本中存在着一些错误，因此并不能直接按照上面的过程动手编程。我们需要重新依其思路，推导出一个准确的算法，再编程。这一过程可能相当耗费时间，且由于细节部分较多，需要一定耐心。

2. 编程过程本身。

   (a) 首先，程序量会比较大，因此如何合理组织代码，如何封装是考验 MATLAB 编程技巧的部分。

   (b) 其次，涉及到大规模的数据运算，程序运行的时间可能就会长，这对于代码的效率就有一定要求：需要尽量减少冗余，并充分利用 BLAS 代数系统。

   (c) 最后，我们要进行很高精度的比较，因此程序中不能出现任何错误，结果才有意义；鉴于程序的规模可知，在调试和交叉检验上的是相当有挑战性的。

参考文献

[1] C. Canuto, MY Hussaini, A. Quarteroni, and TA Zang. *Spectral methods: fundamentals in single domain*, chapter 4, pages 173--177. Springer, 2006. 4, 6, 7, 9

[2] L. Greengard. Spectral integration and two-point boundary value problems. *SIAM Journal on Numerical Analysis*, pages 1071--1080, 1991. 4, 6

[3] R. Peyret. *Spectral methods for incompressible viscous flow*, volume 148, chapter 3, pages 39--53. Springer Verlag, 2002. 9

[4] 叶兴德，程晓良，陈明飞，薛莲. *数值分析基础*. 浙江大学出版社, 2008. 9

毕业设计—文献翻译部分

Chebyshev 方法

章 3

Chebyshev 方法

由于边界上 Gibbs 现象的存在，Fourier 方法一般只适用于周期性的问题。在不具有周期性的问题上，我们最好选择一些更加合适的基函数。像 Chebyshev 正交多项式，可以作为这种情况下 Fourier 方法的一种合适的替代方法。Chebyshev 级数的展开其实可以看作 Fourier 展开中的余弦项，因此它保留了 Fourier 方法的一些重要的好的性质，比如收敛性和 FFT 的应用。并在同时，它可以很好地避免边界上出现 Gibbs 现象。

另外也可以选择 Legendre 多项式作为正交基，它与 Chebyshev 多项式一样拥有一些相同的好性质，并且在离散算子和数值积分有关的问题上更有优势。但是，目前并没找到足够迅速的算法来实现 Legendre 多项式的变换。本书中我们只讨论 Chebyshev 多项式，但其实大部分算法和方法也适用于 Legendre 多项式，只有一些小的技术性细节不同。更多关于 Legendre 多项式的性质和应用的讨论，我们可以参考下面的书：Gottlieb 和 Orszag (1977)，Canuto et al.(1988)，Bernardi 和 Maday(1992)，或 Funaro(1992)。

这一章我们讲述 Chebyshev 多项式和它在边值问题求解上的应用，并讨论两种经典方法，即 Galerkin-type (tau 方法)和插值法。在后一种方法中我们会指出，级数展开的截断部分和可以看作是在插值点上的 Lagrange 插值多项式。我们还将给出求解 Chebyshev 逼近所推导出的代数方程，以及求解它的直接法和及迭代法。

3.1 Chebyshev 多项式的一般性质

第一类 Chebyshev 的 k 次多项式记作 $T_k(x)$，定义在 $x \in [-1, 1]$，

$$T_k(x) = \cos(k\cos^{-1}x), k = 0, 1, 2, ... \tag{3.1.1}$$

显然，$-1 \leq T_k \leq 1$。若设 $x = \cos z$ 则有

$$T_k = \cos kz \tag{3.1.2}$$

由该式不难推导得到 Chebyshev 多项式的前几项为

$$T_0 = 1,\ T_1 = \cos z = x,\ T_2 = \cos 2z = 2\cos^2 z - 1 = 2x^2 - 1, ...$$

更一般的可以利用 Moivre 公式，得到下面结果

$$\cos kz = Re\{(\cos z + i \sin z)^k\}$$

然后利用二项展开定理，我们可以获得 $T_k(x)$ 的表达如下

$$T_k = \frac{k}{2} \sum_{m=0}^{[k/2]} (-1)^m \frac{(k-m-1)!}{m!(k-2m)!} (2x)^{k-2m}, \tag{3.1.3}$$

其中 $[\phi]$ 表示 ϕ 的下取整函数。

利用三角函数恒等式

$$\cos(k+1)z + \cos(k-1)z = 2\cos z \cos kz$$

我们可以推出如下的递推关系

$$T_{k+1} - 2xT_k + T_{k-1} = 0, k \geq 1, \tag{3.1.4}$$

利用这个递推式，我们可以根据 T_0 和 T_1 推出多项式 $T_k, k \geq 1$ 的一般表达式。最初的几个多项式的图形如图 3.1。

(3.1.3) 式仅在一些特殊的场合用到，而式 (3.1.2) 则广泛地应用于理论和计算中。

现在我们给出一些 Chebyshev 多项式的性质，以便之后更好地理解 Chebyshev 多项式是如何应用在求解常微分方程和偏微分方程中的；其余的性质会在之后需要的时候再提到。更全的公式，读者可在附录 A 中找到。

多项式 T_k 以及它的一阶导函数 $T_k{}'$ 在 $x = \pm 1$ 处的取值分别可以算出

$$T_k(\pm 1) = (\pm 1)^k, \; T_k(\pm 1) = (\pm 1)^{k+1} k^2. \tag{3.1.5}$$

这个值在我们限定边界条件时将起到作用。还应该注意到

$$T_k(-x) = (-1)^k T_k(x), \tag{3.1.6}$$

即是说多项式的 parity 与它的次数 k 是一样的。

Chebyshev 多项式 T_k 的零点 x_i（也称为 Gauss 点），即它们是取值如下的点

$$x_i = \cos(i + \frac{1}{2})\frac{\pi}{k}, \; i = 0, ..., k-1 \tag{3.1.7}$$

T_k 取到其极值 ± 1 的点 x_i（称为 Gauss-Lobatto 点），即它们是取值如下的点

$$x_i = \cos \frac{\pi i}{k}, \; i = 0, ..., k \tag{3.1.8}$$

值得注意的是这些点可以看作是由多项式 $(1-x^2)T_k{}'(x)$ 的全部零点组成。

我们也可以如下很容易地得到关于导函数多项式的一个递推关系。首先，对 T_k 求导得到

$$T_k{}' = \frac{d}{dz}(\cos kz)\frac{dz}{dx} = k\frac{\sin kz}{\sin z},$$

其中用到了 (3.1.2) 式的 T_k 的表现型。接着，可以利用三角函数公式得到下面关系

$$\frac{T'_{k+1}}{k+1} - \frac{T'_{k-1}}{k-1} = 2T_k, \tag{3.1.9}$$

对于 k>1 时有意义。借由不断对 (3.1.9) 式进行微分，可以得到形式相同的关于 T_k 的任意第 p 阶导函数的递推式。

Chebyshev 多项式是 [-1,1] 区间上的正交多项式，此时权函数取为

$$w = (1-x^2)^{-1/2}. \tag{3.1.10}$$

相应定义的内积为

$$(u,v)_w = \int_{-1}^{1} uvwdx, \tag{3.1.11}$$

则有正交关系

$$(T_k, T_l)_w = \int_{-1}^{1} T_k T_l w dx = \frac{\pi}{2} c_k \delta_{k,l}, \tag{3.1.12}$$

其中，$\delta_{k,l}$ 是 Kronecker delta 而参数 c_k 定义如下

$$c_k = \begin{cases} 2 & \text{当} k = 0\text{时}, \\ 1 & \text{当} k \geq 1\text{时}. \end{cases} \tag{3.1.13}$$

Chebyshev 逼近方法将用到很多 Gauss 积分公式。利用 Gauss-Lobatto 点 $x_i = \cos \pi i/N, i = 0, ...N$（见公式 (3.1.8)）和插值法，对任意函数 $p(x)$ 的数值积分可表达为

$$\int_{-1}^{1} pwdx \cong \frac{\pi}{N}\sum_{i=0}^{N}\frac{p(x_i)}{\bar{c}_i}, \tag{3.1.14}$$

其中

$$\bar{c}_k = \begin{cases} 2 & \text{当} k = 0\text{时} \\ 1 & \text{当} 1 \leq k \leq N-1\text{时} \\ 2 & \text{当} k = N\text{ 时} \end{cases} \tag{3.1.15}$$

当 $p(x)$ 的次数不超过 $2N-1$ 时，式子 (3.1.14) 是严格成立的。[对于其他插值点下的积分公式和证明可参考 Mercier(1989)]

从式 (3.1.14) 中我们可以导出 Gauss-Lobatto 插值点 $x_i, i = 0, ..., N$ 下的离散的正交关系的表示。考虑

$k \neq N$或$l \neq N$ 时，把数值积分公式 (3.1.14) 用到正交关系 (3.1.12) 上，由于 $T_k T_l$ 的次数不超过 $2N-1$，故可以得到的是一个准确的逼近:

$$\frac{\pi}{2} c_k \delta_{k,l} = \int_{-1}^{1} T_k T_l w dx = \frac{\pi}{N} \sum_{i=0}^{N} \frac{1}{\bar{c}_i} T_k(x_i) T_l(x_i)$$

而对于 $k = l = N$ 的情况，只要将左手边的参数 c_k 换成 $\bar{c}_k$(=2) 就仍然成立。因此，离散的正交关系式就可以写成，对于所有 $0 \leq k, l \leq N$ 有如下的关系：

$$\sum_{i=0}^{N} \frac{1}{\bar{c}_i} T_k(x_i) T_l(x_i) = \frac{\bar{c}_k}{2} N \delta_{k,l}. \tag{3.1.16}$$

3.2 Chebyshev 级数的截断和

3.2.1 计算 Chebyshev 系数

我们考虑对一个定义在 $x \in [-1, 1]$ 上的函数 $u(x)$ 作 Chebyshev 逼近 $u_N(x)$，定义为：

$$u_N(x) = \sum_{k=0}^{N} \hat{u}_k T_k(x) \tag{3.2.1}$$

展开式的系数 $\hat{u}_k, k = 0, ..., N$ 将根据 1.2.1 节描述的 Galerkin-type 技术来加以确定。余项 $R_N = u - u_N$ 在弱平均的意义下是零，即

$$(R_N, T_l)_w = 0, l = 0, \ldots, N \tag{3.2.2}$$

也即，

$$\int_{-1}^{1} (u T_l \, w - \sum_{k=0}^{N} \hat{u}_k T_k T_l w) dx = 0, l = 0, \ldots, N,$$

然后考虑到得到的正交条件 (3.1.12)，我们就可以得到 Chebyshev 展开式系数的表达式了

$$\hat{u}_k = \frac{2}{\pi c_k} \int_{-1}^{1} u T_k w dx. \tag{3.2.3}$$

我们也有必要用 (3.1.2) 的表现型，即 $T_k = \cos kz, x = \cos z$，来重写式 (3.2.1) 为

$$u_N = \sum_{k=0}^{N} \hat{u}_k \cos kz, \tag{3.2.4}$$

便可以看出，关于变量 x 的 Chebyshev 展开等价于关于变量 z 的余弦 Fourier 展开。实际上，函数

$$v(z) = u(\cos z) = u(x)$$

是定义在 $0 \le z \le 2\pi$ 上是偶函数并且具有周期性 $v(z+2\pi) = v(z)$。而且在 $0 \le z \le \pi$ 上的 $v(z)$ 比 $-1 \le x \le 1$ 上的 $u(x)$ 具有更多的有界的导函数，从而余弦 Fourier 展开的收敛性可以由 2.1.2 节的结果保证。此外，由于 $v(z)$ 是周期的，它的表达式 (3.2.4) 是在极值点 $z = 0$ 和 $z = \pi$ 处是连续的，故在这些点处并不会出现 Gibbs 现象的问题。更多关于 Chebyshev 逼近法收敛性的细节将在 3.6 节给出。

3.2.2 微分

导函数在 Chebyshev 基下的表达式要比 Fourier 基下来得更复杂，事实上 $T_k(x)$ 的导函数表达式会涉及到所有互异 parity 的多项式，而 e^{ikx} 的导函数是很简单的 ike^{ikx} 而已。这使得这两种逼近在计算方面产生了较大不同：Chebyshev 的微分算子矩阵在谱空间和物理空间上都是满的（不稀疏），而类似的 Fourier 微分矩阵仅在物理空间是满的。

从 (3.1.9) 的递推关系可以得到

$$T'_k(x) = 2k \sum_{n=0}^{K} \frac{1}{c_{k-1-2n}} T_{k-1-2n}(x) \tag{3.2.5}$$

其中的 $K = [(k-1)/2]$。然后考虑一阶导函数

$$u'_N(x) = \sum_{k=0}^{N} \hat{u}_k T'_k(x) = \sum_{k=0}^{N} \hat{u}_k^{(1)} T_k(x) \tag{3.2.6}$$

并把 (3.2.5) 代入进去，我们可以导出一个关于系数 $\hat{u}_k^{(1)}$ 的表达式：

$$\hat{u}_k^{(1)} = \frac{2}{c_k} \sum_{\substack{p=k+1 \\ (p+k)\text{为奇数}}}^{N} p\hat{u}_p, k = 0, \ldots, N-1, \tag{3.2.7}$$

且有 $\hat{u}_N^{(1)} = 0$。把这个关系写成矩阵形式为

$$\hat{U}^{(1)} = \hat{D}\hat{U} \tag{3.2.8}$$

其中 $\hat{U} = (\hat{u}_0, ..., \hat{u}_N)^T, \hat{U}^{(1)} = (\hat{u}_0^{(1)}, ..., \hat{u}_N^{(1)})^T$，$\hat{D}$ 是严格上三角矩阵，其项由 (3.2.7) 式导出。

二阶导函数的展开式为

$$u''_N(x) = \sum_{k=0}^{N} \hat{u}_k^{(2)} T_k(x) \tag{3.2.9}$$

有

$$\hat{u}_k^{(2)} = \frac{1}{c_k} \sum_{\substack{p=k+2 \\ (p+k)\text{为偶数}}}^{N} p(p^2-k^2)\hat{u}_p, k = 0, \ldots, N-2 \tag{3.2.10}$$

且有 $\hat{u}_{N-1}^{(2)} = \hat{u}_N^{(2)} = 0$。写成矩阵形式为

$$\hat{U}^{(2)} = \hat{D}^2\hat{U} \tag{3.2.11}$$

其中 $\hat{U}^{(2)} = (\hat{u}_0^{(2)}, ..., \hat{u}_N^{(2)})^T$。

当导函数展开式的系数涉及到代数关系时，式 (3.2.7) 与 (3.2.9) 这两种解析的表达会有用。而另一方面，如果仅需要系数的数值的话，则可以通过 (3.2.8) 和 (3.2.11) 的矩阵与向量的乘积式来计算，或者通过 (3.1.9) 式的递推关系来求。具体来说就是，把 (3.1.9) 式 T_k 的表达式代入 (3.2.5) 中。然后通过比较 $T_k{}'$ 同项，我们可以得到一阶导函数的递推关系。一般 p 阶导数的系数 $\hat{u}_k^{(p)}$ 间的递推关系可以通过不断微分得到，令

$$c_{k-1}\hat{u}_{k-1}^{(p)} = \hat{u}_{k+1}^{(p)} + 2k\hat{u}_k^{(p-1)}, k \geq 1 \tag{3.2.12}$$

对一阶导数，有初始值

$$\hat{u}_N^{(1)} = 0,\ \hat{u}_{N-1}^{(1)} = 2N\hat{u}_N \tag{3.2.13}$$

对于二阶导数，有初始值[1]

$$\hat{u}_N^{(2)} = \hat{u}_{N-1}^{(2)} = 0,\ \hat{u}_{N-2}^{(2)} = 2(N-1)\hat{u}_{N-1}^{(1)} = 4N(N-1)\hat{u}_N \tag{3.2.14}$$

此 $p = 2$ 时的递推关系 (3.2.12) 也可以用一个直接连接 $\hat{u}_k^{(2)}$ 与 $\hat{u}_k$ 的式子代替。这关系可以通过考虑 $p = 2$ 的时候的 (3.2.12)，把 $k-1$ 和 $k+1$ 的式子写出来，例如 $\hat{u}_{k-1}^{(1)}$ 和 $\hat{u}_{k+1}^{(1)}$。然后这两个方程联立以便通过 $p = 1$ 时的 (3.2.12) 式消去 $\hat{u}_{k-1}^{(1)}$ 和 $\hat{u}_{k+1}^{(1)}$。最后得到的递推关系为

$$P_k\hat{u}_{k-2}^{(2)} + Q_k\hat{u}_k^{(2)} + R_k\hat{u}_{k+2}^{(2)} = \hat{u}_k,\ 2 \leq k \leq N \tag{3.2.15}$$

$$P_k = \frac{c_{k-2}}{4k(k-1)}, Q_k = \frac{-e_{k+2}}{2(k^2-1)}, R_k = \frac{e_{k+4}}{4k(k+1)} \tag{3.2.16}$$

$$e_j = \begin{cases} 1 & \text{若} j \leq N, \\ 0 & \text{若} j > N. \end{cases} \tag{3.2.17}$$

关于递推算法 (3.2.12)，Wengle 和 Seifeld(1978) 已经指出它可能不是数值稳定的 (ill-conditioned)，其最小系数 $\hat{u}_k^{(p-1)}$ 的误差会放大以致影响到所有的系数，甚至最大的系数 $\hat{u}_k^{(p)}$。不过，这个问题可以通过设一个取决于计算机精度的阈值，将足够小的系数都令为 0，而加以解决。

[1] 原文此公式误为 $\hat{u}_N^{(2)} = \hat{u}_{N-1}^{(2)} = 0,\ \hat{u}_{N-2}^{(2)} = 2(N-1)\hat{u}_{N-1}^{(1)} = 2N(N-1)\hat{u}_N$

3.3 离散 Chebyshev 级数与插值

本节讨论用来逼近给定函数的 Chebyshev 插值技术。首先会考虑离散的 Chebyshev 截断和的系数计算。然后会建立微分的矩阵。最后将等价地通过 Lagrange 插值多项式来讨论 Chebyshev 展式。

这里用到的插值点是 (3.8) 处定义的 Gauss-Lobatto 点。另外的插值点集在一些情况下也是有用的(参看 Gottlieb 等人,1984; Canuto 等人,1988; Mercier,1989)。例如,若希望边界点 $x = \pm 1$ 不作为插值点,则 (3.1.7) 式的点集可能不错。若希望排除边界点 $x = -1$(比如在圆柱坐标系下考虑时,$x = -1$ 对应的是轴),那么选用 Gauss-Radau 点可能很合适。而与之相对的,当考虑边值问题时,Gauss-Lobatto 点的包含边界的性质就是不可或缺的。

3.3.1 计算 Chebyshev 系数

考虑 Chebyshev 展式 (3.2.1),我们希望通过 1.2.1 节介绍的插值方法来计算 $\hat{u}_k$。方法是通过将插值点 $x_i = \cos \pi i/N, i = 0, \ldots, N$ 处的余项 $R_N = u - u_N$ 设为零得到的。令

$$u(x_i) = u_N(x_i) = \sum_{k=0}^{N} \hat{u}_k T_k(x_i), i = 0, ..., N \tag{3.3.1}$$

记 $u_i = u(x_i) = u_N(x_i)$,将定义 (3.1.1) 代入上式得到:

$$u_i = \sum_{k=0}^{N} \hat{u}_k \cos \frac{k\pi i}{N}, i = 0, ..., N \tag{3.3.2}$$

方程 (3.3.1)[或 (3.3.2)] 给出了一个 $2N + 1$ 的代数系统,可以求解这 $2N + 1$ 个系数 $\hat{u}_k$。该矩阵 $T = [\cos k\pi i/N], k, i = 0, ..., N$ 是可逆的;实际上,它的逆可由下面的 (3.3.4) 给出,即 $T^{-1} = [2(\cos \pi i/N)/(\bar{c}_k \bar{c}_i N)], k, i = 0, ..., N$。

可以通过使用 (3.1.16) 的正交关系来求解 (3.3.1)。先在 (3.3.1) 两侧同乘 $T_l(x_i)/\bar{c}_i$,然后对 i 从 0 到 N 进行加和,注意到 (3.1.16) 式,可以得到

$$\hat{u}_k = \frac{2}{\bar{c}_k N} \sum_{i=0}^{N} \frac{1}{\bar{c}_i} u_i T_k(x_i), k = 0, ..., N \tag{3.3.3}$$

或

$$\hat{u}_k = \frac{2}{\bar{c}_k N} \sum_{i=0}^{N} \frac{1}{\bar{c}_i} u_i \cos \frac{k\pi i}{N}, k = 0, ..., N \tag{3.3.4}$$

应当指出这个表达式只不过是 (3.2.3) 式基于 Gauss-Lobatto 点的数值积分近似的结果。

(3.3.2) 和 (3.3.3) 显示,格点处函数值 u_i 与系数 $\hat{u}_k$ 之间可以通过 Fourier 余弦级数的截断和的形式联系到一起。因此,利用 FFT 算法来联系物理空间(格点处函数值空间)与谱空间(系数空间)是一个可能的用法。

或者，也可以简单地利用矩阵与向量的乘积

$$U = T\hat{U}, \hat{U} = T^{-1}U \tag{3.3.5}$$

其中的 U 和 $\hat{U}$ 分别是格点处函数值的向量和展式系数的向量。当乘积的项数不是很大时（即 60 到 100 项），矩阵向量乘积方法在计算时间上比 FFT 更有效率，当然，这取决于所用的计算机和使用的例程。

关于插值逼近 (3.3.1) 和 (3.3.3) 的误差的结果在 3.6 节给出。

3.3.2 插值计算的系数与 Galerkin 系数间的关系

本节来精确给出由 (3.2.3) 积分式定义的展式系数和由 (3.3.3) 加和式计算出来的系数间的关系。第一种系数记为 $\hat{u}_k^e$，第二种记为 $\hat{u}_k^c$。在 2.3 节给出的用来分析 Fourier 级数的框架在这里也适用。由 (3.2.3) 和 (3.3.3) 可以得到

$$\hat{u}_k^e = \frac{2}{\pi c_k}\int_{-1}^{1} u(x)T_k(x)w(x)dx, k = 0, ..., N \tag{3.3.6}$$

$$\hat{u}_k^c = \frac{2}{\bar{c}_k N}\sum_{i=0}^{N}\frac{1}{\bar{c}_i}u_i T_k(x_i), k = 0, ..., N \tag{3.3.7}$$

现在我们将 (3.3.7) 中的 $u(x_i)$ 用它的级数展开

$$u(x) = \sum_{k=0}^{\infty}\hat{u}_k^e T_k(x)$$

代替并假设是绝对收敛的。然后我们把这个无限和分成如下两部分[2]

$$\begin{aligned}\hat{u}_k^c &= \frac{2}{\bar{c}_k N}\sum_{l=0}^{N}\hat{u}_l^e[\sum_{i=0}^{N}\frac{1}{\bar{c}_i}T_k(x_i)T_l(x_i)] \\ &\quad + \frac{2}{\bar{c}_i N}\sum_{l=0}^{\infty}\hat{u}_l^e[\sum_{i=0}^{N}\frac{1}{\bar{c}_k}T_k(x_i)T_l(x_i)]\end{aligned}$$

方括号中的部分无非是 (3.1.16) 那样的离散表达的正交关系。在第一个括号中，由于 k 和 l 都是在 0 到 N 之间，从而 (3.1.16) 成立。而在第二个中，指标 l 是在 $N+1$ 到无穷大之间取值，因而 (3.1.16) 并不适用。上面的式子可以写为

$$\hat{u}_k^c = \hat{u}_k^e + \frac{2}{\bar{c}_k N}\sum_{l=N+1}^{\infty}C_{kl}\hat{u}_l^e$$

[2]原文将 $\hat{u}_l^e$ 误为 $\hat{u}_k^e$

其中

$$\begin{aligned} C_{kl} &= \sum_{i=0}^{N} \frac{1}{\bar{c}_i} T_k(x_i) T_l(x_i) = \sum_{i=0}^{N} \frac{1}{\bar{c}_i} \cos\frac{ki\pi}{N} \cos\frac{li\pi}{N} \\ &= \frac{1}{2} \sum_{i=0}^{N} \frac{1}{\bar{c}_i} [\cos\frac{k-l}{N} i\pi + \cos\frac{k+l}{N} i\pi] \end{aligned}$$

其中 $k = 0, ..., N$ 且 $l = N+1, ...$。然后利用对 $p \in \mathbb{Z}$ 成立的等式

$$\sum_{i=0}^{N} \cos\frac{pi\pi}{N} = \begin{cases} N+1 & \text{如果} p = 2mN, m = 0, \pm 1, \pm 2, ..., \\ \frac{1}{2}[1 + (-1)^p] & \text{其他情况} \end{cases}$$

我们可以算出 C_{kl} 然后得到最终的两种系数之间的关系为

$$\hat{u}_k^c = \hat{u}_k^e + \frac{1}{\bar{c}_k} [\sum_{\substack{m=1 \\ 2mN > N-k}}^{\infty} \hat{u}_{k+2mN}^e + \sum_{\substack{m=1 \\ 2mN > N+k}}^{\infty} \hat{u}_{k-2mN}^e] \tag{3.3.8}$$

方括号中的两项为异名项(alias),它们出现的原因与 Fourier 级数时的原因相同(2.2 节和 2.3 节)。这是因为,可以将以 x 为变量的 Chebyshev 展式看成,以 $z = \cos^{-1} x$ 为变量的 Fourier 余弦展式。

3.3.3 Lagrange 插值多项式

我们重新回到逼近上来,考虑

$$u_N(x) = \sum_{k=0}^{N} \hat{u}_k T_k(x) \tag{3.3.9}$$

其中系数 $\hat{u}_k, k = 0, ..., N$ 的确定是通过,令 $u_N(x)$ 与 $u(x)$ 在插值点 $x_i = \cos \pi i/N, i = 0, ..., N$ 处重合而得到的。因此,(3.3.9) 的 N 次多项式的也就是在点集 $\{x_i\}$ 上的 Lagrange 插值多项式,从而可以写成

$$u_N = \sum_{j=0}^{N} h_j(x) u(x_j) \tag{3.3.10}$$

满足 $u_N(x_i) = u(x_i)$,$h_j(x)$ 是如下定义的 N 次多项式

$$h_j(x) = \frac{(-1)^{j+1}(1-x^2) T'_N(x)}{\bar{c}_j N^2 (x - x_j)} \tag{3.3.11}$$

此式的构造,是注意到插值点 x_j 是 $(1-x^2)T'_N(x)$ 的零点(见 section §3.1节),以及 $(1-x^2)T'{}_N(x)/(x-x_j) \to (-1)^{j+1}\bar{c}_j N^2$ 当 $x \to x_j, j = 0, ..., N$ 时。

因而 (3.3.10) 的表达等价于 (3.3.9),并在一些场合中会有用,因为它并没有涉及到谱系数。

3.3.4 物理空间上的微分

1.1.3 节曾提到过，2.7.2 节也在 Fourier 级数的情形下讨论过，将插值法应用到求解微分方程上时，可以理解为，既不知道展开式的系数同时也不知道格点处的值。第一种方法在 Chebyshev 型的情形下很少用到（参看 Marion 和 Gay,1986）。而第二种方法（或称为“正交插值法”，参看 Finlayson,1972），则是计算流体力学的主流方法。该法最初是由 Wengle（1979）在流体扩散方程上的工作，以及 Orszag 和 Patera（1983），或 Ouazzani 和 Peyret（1984）在 Navier-Stokes 方程上的工作所首先引入的。

因此，应用插值法，而格点处的值是未知量时，应该用格点处的函数值来表达插值点处的导数值，即对于 p 阶导数 $u_N^{(p)}$：

$$u_N^{(p)}(x_i) = \sum_{j=0}^{N} d_{i,j}^{(p)} u_N(x_j), i = 0, \ldots, N \tag{3.3.12}$$

系数 $d_{i,j}^{(p)}$ 可由下面两种方式计算：

1. 利用 (3.3.3)，从导数

$$u_N^{(p)}(x_i) = \sum_{k=0}^{N} \hat{u}_k T_k^{(p)}(x_i)$$

中消去 $\hat{u}_k$。然后根据 $T_k = \cos kz$ 用三角函数来表达 $T_k(x_i)$ 和 p 阶导数 $T_k^{(p)}(x_i)$。最后，利用经典三角恒等式来计算这个和。

2. 对插值多项式 (3.3.10) 直接微分 p 次：

$$u_N^{(p)}(x_i) = \sum_{j=0}^{N} h_j^{(p)}(x_i) u_N(x_j)$$

从而，$d_{i,j}^{(p)} = h_j^{(p)}(x_i)$ 可以从 (3.3.11) 计算出来。

前两阶导数的系数 $d_{i,j}^{(p)}$ 的表达如下:

一阶导数

$$\begin{aligned} d_{i,j}^{(1)} &= \frac{\bar{c}_i}{\bar{c}_j} \frac{(-1)^{i+j}}{(x_i - x_j)}, && 0 \le i, j \le N, i \ne j \\ d_{i,i}^{(1)} &= -\frac{x_i}{2(1-x_i^2)}, && 1 \le i \le N-1 \\ d_{0,0}^{(1)} &= -d_{N,N}^{(1)} = \frac{2N^2+1}{6}, \end{aligned} \tag{3.3.13}$$

其中，$x_i = \cos(\pi i/N), \bar{c}_0 = \bar{c}_N = 2, \bar{c}_j = 1, \forall 1 \le j \le N-1$

二阶导数

$$
\begin{aligned}
d_{i,j}^{(2)} &= \frac{(-1)^{i+j}}{\bar{c}_j}\frac{x_i^2+x_ix_j-2}{(1-x_i^2)(x_i-x_j)^2}, && 1\le i\le N-1\\
& && 0\le j\le N, i\neq j\\
d_{i,i}^{(2)} &= -\frac{(N^2-1)(1-x_i^2)+3}{3(1-x_i^2)^2}, && 1\le i\le N-1\\
d_{0,j}^{(2)} &= \frac{2}{3}\frac{(-1)^j}{\bar{c}_j}\frac{(2N^2+1)(1-x_j)-6}{(1-x_j)^2}, && 1\le j\le N\\
d_{N,j}^{(2)} &= \frac{2}{3}\frac{(-1)^{j+N}}{\bar{c}_j}\frac{(2N^2+1)(1+x_j)-6}{(1+x_j)^2}, && 0\le j\le N-1\\
d_{0,0}^{(2)} &= d_{N,N}^{(2)} = \frac{N^4-1}{15}.
\end{aligned}
\tag{3.3.14}
$$

回忆起下面的关系也会有用

$$d_{i,j}^{(2)} = \sum_{k=0}^{N} d_{i,k}^{(1)} d_{k,i}^{(1)} \tag{3.3.15}$$

导数写成向量形式为

$$U^{(1)} = DU,\ U^{(2)} = D^2U \tag{3.3.16}$$

其中

$$U = (u_N(x_0),...,u_N(x_N))^T, U^{(p)} = (u_N^{(p)}(x_0),...,u_N^{(p)}(x_N))^T \tag{3.3.17}$$

其中 $p=1,2$。微分矩阵 D 定义为

$$D = [d_{i,j}^{(1)}],\ i,j=0,\dots N. \tag{3.3.18}$$

3
Chebyshev method

The Fourier method is appropriate for periodic problems, but is not adapted to nonperiodic problems because of the existence of the Gibbs phenomenon at the boundaries. In the case of nonperiodic problems, it is advisable to have recourse to better-suited basis functions. Orthogonal polynomials, like Chebyshev polynomials, constitute a proper alternative to the Fourier basis. The Chebyshev series expansion may be seen as a cosine Fourier series, so that it possesses the valuable properties of the latter concerning, in particular, the convergence and the possible use of the FFT. On the other hand, the Chebyshev series expansion is exempt from the Gibbs phenomenon at the boundaries.

Another possible choice of an orthogonal polynomial basis is constituted by the Legendre polynomials. These polynomials share a number of properties with the Chebyshev polynomials. They present some advantages concerning the properties of the discrete operators and the numerical quadrature. On the other hand, no fast transform algorithm is known for Legendre polynomials. Only Chebyshev polynomials are discussed in this book, but the methods and algorithms described also apply to Legendre polynomials with only technical changes required by their specific properties. We refer to the books by Gottlieb and Orszag (1977), Canuto *et al.* (1988), Bernardi and Maday (1992) or Funaro (1992) for discussions on the properties and applications of the Legendre polynomials.

The present chapter is intended to give a general view of Chebyshev polynomials and their applications to the solution of boundary value problems. Two classical approaches, Galerkin-type (tau method) and collocation, will be addressed. In this latter case, it will be pointed out that the Chebyshev

truncated series expansion can be seen as the Lagrange interpolation polynomial based on the collocation points. Direct and iterative methods for solving the algebraic systems, resulting from the Chebyshev approximation, will be described.

3.1 Generalities on Chebyshev polynomials

The Chebyshev polynomial of the first kind $T_k(x)$ is the polynomial of degree k defined for $x \in [-1, 1]$ by

$$T_k(x) = \cos\left(k \, \cos^{-1} x\right) \, , k = 0, 1, 2, \dots \, , \tag{3.1}$$

therefore, $-1 \leq T_k \leq 1$. By setting $x = \cos z$, we have

$$T_k = \cos kz \, , \tag{3.2}$$

from which it is easy to deduce the expressions for the first Chebyshev polynomials

$$T_0 = 1 \, , \;\; T_1 = \cos z = x \, , \;\; T_2 = \cos 2z = 2 \, \cos^2 z - 1 = 2 \, x^2 - 1 \, , \dots \, .$$

More generally, from the Moivre formula, we get

$$\cos kz = \mathcal{R}e \left\{ \left(\cos z + \underline{i} \, \sin z \right)^k \right\}$$

and then, by application of the binomial formula, we may express the polynomial T_k as

$$T_k = \frac{k}{2} \sum_{m=0}^{[k/2]} (-1)^m \, \frac{(k-m-1)!}{m! \; (k-2\,m)!} \, (2\,x)^{k-2\,m} \, , \tag{3.3}$$

where $[\phi]$ denotes the integer part of ϕ.

From the trigonometrical identity

$$\cos\left(k+1\right) z + \cos\left(k-1\right) z = 2 \, \cos z \, \cos kz$$

we deduce the recurrence relationship

$$T_{k+1} - 2\,x\,T_k + T_{k-1} = 0 \, , \;\; k \geq 1 \, , \tag{3.4}$$

which allows us, in particular, to deduce the expression of the polynomials T_k, $k \geq 2$, from the knowledge of T_0 and T_1. The graph of the first polynomials is shown in Fig. 3.1.

Expression (3.3) may be useful in some special circumstances but the representation (3.2) is generally used in computational as well as theoretical studies.

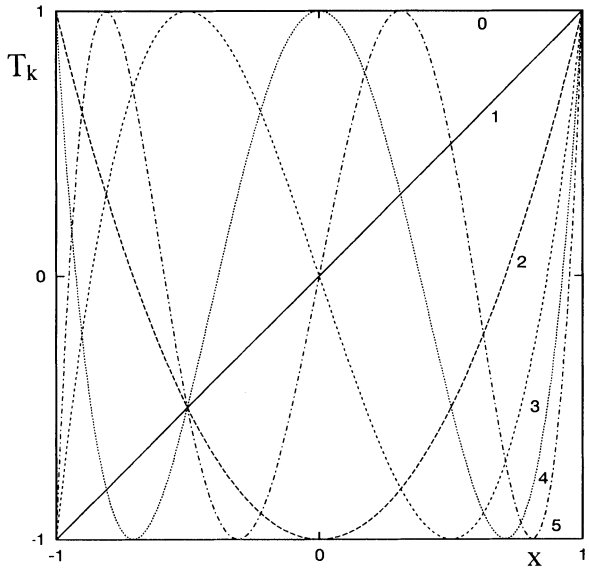

FIGURE 3.1. Graphs of the first Chebyshev polynomials, $T_k(x)$, for $k = 0, \cdots, 5$.

Now we list some properties useful for the understanding and application of Chebyshev polynomials to the solution of ordinary or partial differential equations ; other properties will be discussed in the sequel when necessary. And then, a richer set of formulas is given in Appendix A.

The values of T_k and its first-order derivative T_k' at $x = \pm 1$ are given by

$$T_k(\pm 1) = (\pm 1)^k \ , \quad T_k'(\pm 1) = (\pm 1)^{k+1} k^2 \, . \tag{3.5}$$

The knowledge of these values can be of interest when prescribing boundary conditions. It is important to note that

$$T_k(-x) = (-1)^k T_k(x) \, , \tag{3.6}$$

that is the parity of the polynomial is the same as its degree k.

The polynomial T_k vanishes at the points x_i (Gauss points) defined by

$$x_i = \cos\left(i + \frac{1}{2}\right)\frac{\pi}{k} \, , \quad i = 0, \ldots, k-1 \tag{3.7}$$

and it reaches its extremal values ± 1 at the points x_i (Gauss-Lobatto points) defined by

$$x_i = \cos\frac{\pi\, i}{k} \, , i = 0, \ldots, k \, . \tag{3.8}$$

Note that such points are the zeros of the polynomial $\left(1 - x^2\right) T_k'(x)$.

A recurrence relation on the derivative can easily be obtained. First, the differentiation of T_k gives

$$T'_k = \frac{d}{dz}\left(\cos kz\right)\frac{dz}{dx} = k\frac{\sin kz}{\sin z}\,,$$

where we have used the representation (3.2). Then, by the application of trigonometrical formulas, we get the relation

$$\frac{T'_{k+1}}{k+1} - \frac{T'_{k-1}}{k-1} = 2\,T_k \tag{3.9}$$

valid for $k > 1$. A similar formula for the pth derivative is obtained by successive differentiations of (3.9).

The Chebyshev polynomials are orthogonal on $[-1, 1]$ with the weight

$$w = \left(1 - x^2\right)^{-1/2}\,. \tag{3.10}$$

Let the scalar product be

$$\left(u, v\right)_w = \int_{-1}^{1} u\,v\,w\,dx\,, \tag{3.11}$$

so that the orthogonality property is

$$\left(T_k, T_l\right)_w = \int_{-1}^{1} T_k\,T_l\,w\,dx = \frac{\pi}{2}c_k\,\delta_{k,l}\,, \tag{3.12}$$

where $\delta_{k,l}$ is the Kronecker delta and c_k is defined by

$$c_k = \begin{cases} 2 & \text{if } k = 0\,, \\ 1 & \text{if } k \geq 1\,. \end{cases} \tag{3.13}$$

The Chebyshev approximation makes extensive use of the Gauss quadrature formulas. For the Gauss-Lobatto points $x_i = \cos \pi\, i/N$, $i = 0, \ldots, N$ [see Eq.(3.8)], generally used in collocation methods, the quadrature formula applied to any function $p(x)$ gives

$$\int_{-1}^{1} p\,w\,dx \cong \frac{\pi}{N}\sum_{i=0}^{N}\frac{p(x_i)}{\overline{c}_i}\,, \tag{3.14}$$

where

$$\overline{c}_k = \begin{cases} 2 & \text{if} \quad k = 0\,, \\ 1 & \text{if} \quad 1 \leq k \leq N-1\,, \\ 2 & \text{if} \quad k = N\,. \end{cases} \tag{3.15}$$

The relation (3.14) is exact if $p(x)$ is a polynomial of degree $2N - 1$ at most [see Mercier (1989) for the proof and formulas associated with other sets of points].

From Eq.(3.14) we may derive the discrete orthogonality relation based on the Gauss-Lobatto points x_i, $i = 0, \ldots, N$. For $k \neq N$ or $l \neq N$, the use of (3.14) gives an exact approximation to the integral in (3.12) since $T_k T_l$ is a polynomial of degree at most $2N - 1$:

$$\frac{\pi}{2} c_k \, \delta_{k,l} = \int_{-1}^{1} T_k \, T_l \, w \, dx = \frac{\pi}{N} \sum_{i=0}^{N} \frac{1}{\overline{c}_i} T_k(x_i) \, T_l(x_i) \, .$$

For $k = l = N$, this last formula remains exact provided c_k in the left-hand side is replaced by $\overline{c}_N$ $(= 2)$. Therefore, the discrete orthogonality relation is

$$\sum_{i=0}^{N} \frac{1}{\overline{c}_i} T_k(x_i) \, T_l(x_i) = \frac{\overline{c}_k}{2} \, N \, \delta_{k,l} \tag{3.16}$$

valid for $0 \leq k, l \leq N$.

3.2 Truncated Chebyshev series

3.2.1 Calculation of Chebyshev coefficients

Let us consider the Chebyshev approximation of the function $u(x)$ defined for $x \in [-1, 1]$:

$$u_N(x) = \sum_{k=0}^{N} \hat{u}_k \, T_k(x) \, . \tag{3.17}$$

The expansion coefficients $\hat{u}_k$, $k = 0, \ldots, N$, are determined by following the Galerkin-type technique described in Section 1.2.1. The residual $R_N = u - u_N$ is annuled in the weak average sense

$$(R_N, T_l)_w = 0 \, , \quad l = 0, \ldots, N \, , \tag{3.18}$$

namely,

$$\int_{-1}^{1} \left(u \, T_l \, w - \sum_{k=0}^{N} \hat{u}_k \, T_k \, T_l \, w \right) dx = 0 \, , \quad l = 0, \ldots, N \, .$$

Then, taking the orthogonality condition (3.12) into account, we obtain the expression for the Chebyshev expansion coefficients

$$\hat{u}_k = \frac{2}{\pi \, c_k} \int_{-1}^{1} u \, T_k \, w \, dx \, . \tag{3.19}$$

It seems worthwhile to express (3.17) by means of the representation (3.2), that is, $T_k = \cos kz$ with $x = \cos z$. The expansion (3.17) is then written as

$$u_N = \sum_{k=0}^{N} \hat{u}_k \, \cos kz \, , \tag{3.20}$$

showing that the Chebyshev expansion (3.17) with respect to x is equivalent to a cosine Fourier series in z. In fact, the function

$$v\,(z) = u\,(\cos z) = u\,(x)$$

defined in $0 \leq z \leq 2\pi$ is even and periodic since $v\,(z + 2\pi) = v\,(z)$. Moreover, $v\,(z)$ has as many bounded derivatives in $0 \leq z \leq \pi$ than $u\,(x)$ has in $-1 \leq x \leq 1$. Therefore, the convergence properties of the cosine Fourier series expansion (3.20) can be deduced from the results of Section 2.1.2. Moreover, since $v\,(z)$ is periodic, its representation (3.20) is continuous at the extremities $z = 0$ and $z = \pi$ and, consequently, is exempt from the Gibbs phenomenon at these points. More detailed results on the convergence of the Chebyshev approximation will be given in Section 3.6.

3.2.2 Differentiation

The expression of derivatives in the Chebyshev basis is more complicated than in the Fourier one. Indeed, the expression of the derivative of $T_k(x)$ involves all the polynomials of opposite parity and lower degree while the derivative of $e^{\underline{i}\,k\,x}$ is simply $\underline{i}\,k\,e^{\underline{i}\,k\,x}$. This makes the computational aspects of the two approximations very different : the Chebyshev differentiation matrices in the spectral and physical spaces are full while the analogous Fourier matrices are full only in the physical space.

From the recurrence relation (3.9), one obtains

$$T'_k(x) = 2\,k \sum_{n=0}^{K} \frac{1}{c_{k-1-2n}} T_{k-1-2n}(x)\,, \tag{3.21}$$

where $K = [(k-1)/2]$. Therefore, considering the first-order derivative

$$u'_N(x) = \sum_{k=0}^{N} \hat{u}_k\, T'_k(x) = \sum_{k=0}^{N} \hat{u}^{(1)}_k\, T_k(x) \tag{3.22}$$

and, taking Eq.(3.21) into account, we deduce the expression of the coefficient $\hat{u}^{(1)}_k$:

$$\hat{u}^{(1)}_k = \frac{2}{c_k} \sum_{\substack{p=k+1 \\ (p+k)\ \text{odd}}}^{N} p\,\hat{u}_p\,, \quad k = 0, \ldots, N-1\,, \tag{3.23}$$

and $\hat{u}^{(1)}_N = 0$. This can be written in matrix form as

$$\hat{U}^{(1)} = \hat{\mathcal{D}}\,\hat{U}\,, \tag{3.24}$$

where $\hat{U} = (\hat{u}_0, \ldots, \hat{u}_N)^T$, $\hat{U}^{(1)} = \left(\hat{u}^{(1)}_0, \ldots, \hat{u}^{(1)}_N\right)^T$ and $\hat{\mathcal{D}}$ is a strictly triangular upper matrix whose entries are deduced from (3.23).

The second-order derivative expansion is

$$u''_N(x) = \sum_{k=0}^{N} \hat{u}_k^{(2)}\, T_k(x) \tag{3.25}$$

with

$$\hat{u}_k^{(2)} = \frac{1}{c_k} \sum_{\substack{p=k+2 \\ (p+k)\ \text{even}}}^{N} p\, \left(p^2 - k^2\right)\, \hat{u}_p\,, \quad k = 0, \ldots, N-2 \tag{3.26}$$

and $\hat{u}_{N-1}^{(2)} = \hat{u}_N^{(2)} = 0$. This is written in matrix form as

$$\hat{U}^{(2)} = \hat{\mathcal{D}}^2\, \hat{U}\,, \tag{3.27}$$

where $\hat{U}^{(2)} = \left(\hat{u}_0^{(2)}, \ldots, \hat{u}_N^{(2)}\right)^T$.

The analytical expressions (3.23) and (3.25) are of interest each time the expansion coefficients of the derivatives are involved in algebraic calculations. On the other hand, if only the numerical values of the coefficients are needed, they can be calculated either from the matrix-vector products (3.24) and (3.27) or from recurrence formulas deduced from (3.9). More precisely, the expression for T_k given by (3.9) is brought into (3.21). Then, by identification of the derivative T'_k with the same index, we obtain the recurrence formula for the first-order derivative. The general recurrence formula for the coefficients $\hat{u}_k^{(p)}$ of the pth derivative is obtained by successive differentiations, let

$$c_{k-1}\, \hat{u}_{k-1}^{(p)} = \hat{u}_{k+1}^{(p)} + 2\, k\, \hat{u}_k^{(p-1)}\,, \quad k \geq 1\,, \tag{3.28}$$

be complemented with the starting values, for the first-order derivative

$$\hat{u}_N^{(1)} = 0\,, \quad \hat{u}_{N-1}^{(1)} = 2\, N\, \hat{u}_N\,, \tag{3.29}$$

and, for the second-order derivative,

$$\hat{u}_N^{(2)} = \hat{u}_{N-1}^{(2)} = 0\,, \quad \hat{u}_{N-2}^{(2)} = 2\,(N-1)\, \hat{u}_{N-1}^{(1)} = 2\, N\,(N-1)\, \hat{u}_N\,. \tag{3.30}$$

The recurrence relation (3.28) for $p = 2$ can be replaced by another connecting directly the coefficients $\hat{u}_k^{(2)}$ to $\hat{u}_k$. Such a relation is obtained by considering (3.28) with $p = 2$ and written for $k-1$ and $k+1$, such as the quantities $\hat{u}_{k-1}^{(1)}$ and $\hat{u}_{k+1}^{(1)}$ appear in the right-hand sides. Then these two equations are combined so that the quantities $\hat{u}_{k-1}^{(1)}$ and $\hat{u}_{k+1}^{(1)}$ can be eliminated thanks to (3.28) considered for $p = 1$. The resulting recurrence relation is

$$P_k\, \hat{u}_{k-2}^{(2)} + Q_k\, \hat{u}_k^{(2)} + R_k\, \hat{u}_{k+2}^{(2)} = \hat{u}_k\,, \quad 2 \leq k \leq N \tag{3.31}$$

with

$$P_k = \frac{c_{k-2}}{4\,k(k-1)}\,,\quad Q_k = \frac{-e_{k+2}}{2(k^2-1)}\,,\quad R_k = \frac{e_{k+4}}{4\,k(k+1)}\,, \tag{3.32}$$

where

$$e_j = \begin{cases} 1 & \text{if } j \le N\,, \\ 0 & \text{if } j > N\,. \end{cases} \tag{3.33}$$

Concerning the recurrent algorithm (3.28), Wengle and Seinfeld (1978) have remarked that it may be ill-conditioned in the sense that errors in the smallest coefficient $\hat{u}_k^{(p-1)}$ are amplified such that the accuracy of all the coefficients, even the largest ones $\hat{u}_k^{(p)}$, is destroyed. This can be avoided by simply equating to zero the coefficients smaller than a given threshold, depending on the accuracy of the computer.

3.3 Discrete Chebyshev series and collocation

This section is devoted to the Chebyshev collocation (i.e., interpolation) technique for the approximation of a given function. First, considering the discrete truncated Chebyshev series, the calculation of the expansion coefficients will be developed. Then the expression of the differentiation matrices will be established. Finally, we shall discuss an equivalent way to consider the Chebyshev expansion, namely by introducing the notion of Lagrange interpolation polynomial.

The collocation points considered here are the Gauss-Lobatto points defined by Eq.(3.8). Other sets of points, of similar nature (see Gottlieb *et al.*, 1984 ; Canuto *et al.*, 1988 ; Mercier, 1989), can be useful in some circumstances. For example, the choice of the set (3.7) may be of interest if it is not desired that the boundary points $x = \pm 1$ belong to the set of collocation points. Also, the Gauss-Radau points (see Appendix A) can be used if one wants to exclude the boundary point $x = -1$, for example in problems in cylindrical coordinates where $x = -1$ would correspond to the axis. On the other hand, for the solution of boundary value problems, the property of the set of collocation points held by the Gauss-Lobatto points to contain the boundaries is indispensable.

3.3.1 Calculation of Chebyshev coefficients

Considering the Chebyshev expansion (3.17), we want to calculate the coefficients $\hat{u}_k$ by means of the collocation (or interpolation) technique shown in Section 1.2.1. The technique consists of setting to zero the residual $R_N = u - u_N$ at the collocation points $x_i = \cos \pi\, i/N$, $i = 0, \ldots, N$,

let

$$u(x_i) = u_N(x_i) = \sum_{k=0}^{N} \hat{u}_k \, T_k(x_i) \,, \quad i = 0, \ldots, N \,. \tag{3.34}$$

By denoting $u_i = u(x_i) = u_N(x_i)$, and using the definition (3.1), the above equation gives :

$$u_i = \sum_{k=0}^{N} \hat{u}_k \, \cos \frac{k \, \pi \, i}{N} \,, \quad i = 0, \ldots, N \,. \tag{3.35}$$

Equation (3.34) [or (3.35)] gives an algebraic system of $2\,N+1$ equations for determining the $2\,N+1$ coefficients $\hat{u}_k$. The associated matrix $\mathcal{T} = [\cos k\,\pi\,i/N]$, $k, i = 0, \ldots, N$, is invertible ; as a matter of fact, it will be found below [Eq.(3.37)] that its inverse is $\mathcal{T}^{-1} = [2\,(\cos \pi\,i/N)\,/\,(\overline{c}_k\,\overline{c}_i\,N)]$, $k, i = 0, \ldots, N$.

The expression for the coefficients $\hat{u}_k$ (i.e., the solution of the system (3.34)) is directly obtained by means of the discrete orthogonality relation (3.16). By multiplying each side of (3.34) by $T_l(x_i)/\overline{c}_i$, then summing from $i = 0$ to $i = N$, and using the relation (3.16), we obtain

$$\hat{u}_k = \frac{2}{\overline{c}_k \, N} \sum_{i=0}^{N} \frac{1}{\overline{c}_i} u_i \, T_k(x_i) \,, \quad k = 0, \ldots, N \,, \tag{3.36}$$

or

$$\hat{u}_k = \frac{2}{\overline{c}_k \, N} \sum_{i=0}^{N} \frac{1}{\overline{c}_i} u_i \, \cos \frac{k \, \pi \, i}{N} \,, \quad k = 0, \ldots, N \,. \tag{3.37}$$

It must be noted that such expressions are nothing other than the numerical approximation (based on the Gauss-Lobatto points) of the integral appearing in Eq.(3.19).

The relations (3.35) and (3.36) show that the grid values u_i, as well as the coefficients $\hat{u}_k$, are related by truncated discrete Fourier series in cosine. Therefore, it is possible to use the FFT algorithm (in its cosine version) to connect the physical space (space of the grid values) to the spectral space (space of the coefficients). Note that it is also possible to simply make use of the matrix-vector products

$$U = \mathcal{T} \, \hat{U} \,, \quad \hat{U} = \mathcal{T}^{-1} \, U \,, \tag{3.38}$$

where U and $\hat{U}$ are, respectively, the vectors containing the grid values and the expansion coefficients. Note that the matrix-vector product for a moderate number of terms (namely 60-100) is less expensive in computing time than the FFT, depending on the computer and the routines used.

Results on the error of the collocation approximations (3.34) and (3.36) are given in Section 3.6.

3.3.2 Relation between collocation and Galerkin coefficients

The objective of this section is to make precise the relationship between the expansion coefficients defined by the integral (3.19) and those calculated from the sum (3.36). The first set of coefficients will be denoted here by $\hat{u}_k^e$ and the second set by $\hat{u}_k^c$. The general lines of the analysis made in Section 2.3 for the Fourier series apply in the present case. From Eqs.(3.19) and (3.36), we have

$$\hat{u}_k^e = \frac{2}{\pi\, c_k} \int_{-1}^{1} u(x)\, T_k(x)\, w(x)\, dx\,, \quad k = 0, \ldots, N\,, \tag{3.39}$$

$$\hat{u}_k^c = \frac{2}{\overline{c}_k\, N} \sum_{i=0}^{N} \frac{1}{\overline{c}_i} u(x_i)\, T_k(x_i)\,, \quad k = 0, \ldots, N\,. \tag{3.40}$$

Now we replace $u(x_i)$ in Eq.(3.40) by its expression in terms of the infinite series

$$u(x) = \sum_{k=0}^{\infty} \hat{u}_k^e\, T_k(x)$$

assumed to be absolutely convergent. Then, we decompose the resulting infinite sum into two partial sums according to

$$\begin{aligned} \hat{u}_k^c = \frac{2}{\overline{c}_k\, N} \sum_{l=0}^{N} \hat{u}_k^e \left[\sum_{i=0}^{N} \frac{1}{\overline{c}_i} T_k(x_i)\, T_l(x_i) \right] \\ + \frac{2}{\overline{c}_k\, N} \sum_{l=N+1}^{\infty} \hat{u}_k^e \left[\sum_{i=0}^{N} \frac{1}{\overline{c}_i} T_k(x_i)\, T_l(x_i) \right]\,. \end{aligned}$$

The expression appearing in square brackets is nothing other than the left-hand side of the discrete orthogonality relation (3.16). In the first bracket, the indices k and l vary between 0 and N, so that the relation (3.16) holds. On the other hand, in the second bracket, the index l varies between $N+1$ and infinity, so that the relation (3.16) is not applicable. The above expression for $\hat{u}_k^c$ can be written as

$$\hat{u}_k^c = \hat{u}_k^e + \frac{2}{\overline{c}_k\, N} \sum_{l=N+1}^{\infty} C_{k\,l}\, \hat{u}_l^e\,,$$

where

$$\begin{aligned} C_{k\,l} &= \sum_{i=0}^{N} \frac{1}{\overline{c}_i} T_k(x_i)\, T_l(x_i) = \sum_{i=0}^{N} \frac{1}{\overline{c}_i} \cos\frac{k\,i\,\pi}{N}\, \cos\frac{l\,i\,\pi}{N} \\ &= \frac{1}{2} \sum_{i=0}^{N} \frac{1}{\overline{c}_i} \left[\cos\frac{k-l}{N} i\,\pi + \cos\frac{k+l}{N} i\,\pi \right] \end{aligned}$$

with $k = 0, \ldots, N$ and $l = N+1, \ldots$. Then, by using the identity (valid for $p \in Z$),

$$\sum_{i=0}^{N} \cos \frac{p\, i\, \pi}{N} = \begin{cases} N+1 & \text{if } \; p = 2mN\,, \; m = 0, \pm 1, \pm 2, \ldots\,, \\ \dfrac{1}{2}\left[1 + (-1)^p\right] & \text{otherwise}\,, \end{cases}$$

we may calculate $C_{k\,l}$ and finally get the relation connecting the collocation coefficients to the Galerkin ones

$$\hat{u}_k^c = \hat{u}_k^e + \frac{1}{\overline{c}_k} \left[\sum_{\substack{m=1 \\ 2mN > N-k}}^{\infty} \hat{u}^e_{k+2mN} + \sum_{\substack{m=1 \\ 2mN > N+k}}^{\infty} \hat{u}^e_{-k+2mN} \right]. \quad (3.41)$$

The terms in square brackets are alias terms. The reason for their presence is the same as that for discrete Fourier series (Sections 2.2 and 2.3). This is a consequence of the fact that the Chebyshev expansion in x can also be considered as a cosine Fourier in the variable $z = \cos^{-1} x$.

3.3.3 Lagrange interpolation polynomial

Let us return to the approximation

$$u_N(x) = \sum_{k=0}^{N} \hat{u}_k\, T_k(x)\,, \quad (3.42)$$

where the coefficients $\hat{u}_k$, $k = 0, \ldots, N$, are determined by asking $u_N(x)$ to coincide with $u(x)$ at the collocation points $x_i = \cos \pi\, i/N$, $i = 0, \ldots, N$. Therefore, the polynomial of degree N defined by Eq.(3.42) is nothing other than the Lagrange interpolation polynomial based on the set $\{x_i\}$. Hence, it can also be written in the form

$$u_N(x) = \sum_{j=0}^{N} h_j(x)\, u(x_j) \quad (3.43)$$

with $u_N(x_j) = u(x_j)$, and $h_j(x)$ is the polynomial of degree N defined by

$$h_j\,(x) = \frac{(-1)^{j+1}\left(1 - x^2\right)\, T'_N(x)}{\overline{c}_j\, N^2\, (x - x_j)}\,. \quad (3.44)$$

This expression for h_j is easily constructed by recalling that the collocation points x_j are the zeros of the polynomial $\left(1 - x^2\right) T'_N(x)$ (see Section 3.1.1) and by observing that $\left(1 - x^2\right)\, T'_N\,(x)\,/\,(x - x_j) \;\to\; (-1)^{j+1}\, \overline{c}_j\, N^2$ when $x \to x_j$, $j = 0, \ldots, N$.

Therefore, the representation (3.43) is equivalent to (3.42) and is useful in several circumstances because it does not involve the spectral coefficients.

3.3.4 Differentiation in the physical space

As mentioned in Section 1.1.3, and discussed in Section 2.7.2 for the Fourier case, the application of collocation methods to the solution of differential equations may consider as unknowns the expansion coefficients as well as the grid values. The first approach is seldom employed in the Chebyshev case (see, e.g., Marion and Gay, 1986). On the other hand, the second approach (also known as "orthogonal collocation," see Finlayson, 1972) is of current use in computational fluid mechanics since the works by Wengle (1979) for the advection-diffusion equation and by Orszag and Patera (1983) or Ouazzani and Peyret (1984) for the Navier-Stokes equations.

Therefore, in the context of the collocation method where the unknowns are the grid values, it is necessary to express the derivatives at any collocation point in terms of the grid values of the function, that is, for the pth derivative $u_N^{(p)}$:

$$u_N^{(p)}(x_i) = \sum_{j=0}^{N} d_{i,j}^{(p)}\, u_N(x_j)\,, \quad i = 0, \ldots, N\,. \tag{3.45}$$

The coefficients $d_{i,j}^{(p)}$ can be calculated according to either of the following two ways :

(i) Eliminate $\hat{u}_k$ from the derivative

$$u_N^{(p)}(x_i) = \sum_{k=0}^{N} \hat{u}_k\, T_k^{(p)}(x_i)$$

by using expression (3.36). Then, express $T_k(x_i)$ and the pth derivative $T_k^{(p)}(x_i)$ in terms of trigonometrical functions according to $T_k = \cos kz$. Finally, apply the classical trigonometrical identities to evaluate the sums.

(ii) Differentiate p times directly the interpolation polynomial (3.43) :

$$u_N^{(p)}(x) = \sum_{j=0}^{N} h_j^{(p)}(x_i)\, u_N(x_j)\,.$$

Therefore, $d_{i,j}^{(p)} = h_j^{(p)}(x_i)$ which has to be evaluated from expression (3.44).

The expression of the coefficients $d_{i,j}^{(p)}$ for the first two derivatives are :

First-order derivative

$$\begin{aligned}
d_{i,j}^{(1)} &= \frac{\overline{c}_i}{\overline{c}_j}\,\frac{(-1)^{i+j}}{(x_i - x_j)}\,, && 0 \le i, j \le N\,,\ \ i \ne j \\
d_{i,i}^{(1)} &= -\frac{x_i}{2(1 - x_i^2)}\,, && 1 \le i \le N-1 \\
d_{0,0}^{(1)} &= -d_{N,N}^{(1)} = \frac{2\,N^2 + 1}{6}\,,
\end{aligned} \tag{3.46}$$

where $x_i = \cos(\pi i/N)$, $\overline{c}_0 = \overline{c}_N = 2$, $\overline{c}_j = 1$ for $1 \le j \le N-1$.

Second-order derivative

$$
\begin{aligned}
d^{(2)}_{i,j} &= \frac{(-1)^{i+j}}{\overline{c}_j}\,\frac{x_i^2 + x_i\,x_j - 2}{(1-x_i^2)(x_i - x_j)^2}\,, && 1 \le i \le N-1\,,\\
& && 0 \le j \le N\,,\ \ i \ne j\\
d^{(2)}_{i,i} &= -\frac{(N^2-1)(1-x_i^2)+3}{3(1-x_i^2)^2}\,, && 1 \le i \le N-1\\
d^{(2)}_{0,j} &= \frac{2}{3}\,\frac{(-1)^j}{\overline{c}_j}\,\frac{(2\,N^2+1)(1-x_j)-6}{(1-x_j)^2}\,, && 1 \le j \le N\\
d^{(2)}_{N,j} &= \frac{2}{3}\,\frac{(-1)^{j+N}}{\overline{c}_j}\,\frac{(2\,N^2+1)(1+x_j)-6}{(1+x_j)^2}\,, && 0 \le j \le N-1\\
d^{(2)}_{0,0} &= d^{(2)}_{N,N} = \frac{N^4-1}{15}\,.
\end{aligned}
\tag{3.47}
$$

It may be useful to recall that

$$
d^{(2)}_{i,j} = \sum_{k=0}^{N} d^{(1)}_{i,k}\,d^{(1)}_{k,i}\,. \tag{3.48}
$$

In vector form, the derivatives may be expressed as

$$
U^{(1)} = \mathcal{D}\,U\,, \quad U^{(2)} = \mathcal{D}^2 U \tag{3.49}
$$

where

$$
U = \left(u_N(x_0), \ldots, u_N(x_N)\right)^T\,, \quad U^{(p)} = \left(u_N^{(p)}(x_0), \ldots, u_N^{(p)}(x_N)\right)^T \tag{3.50}
$$

with $p = 1, 2$. The differentiation matrix $\mathcal{D}$ is defined by

$$
\mathcal{D} = \left[d^{(1)}_{i,j}\right]\,, \quad i,j = 0, \ldots, N\,. \tag{3.51}
$$

3.3.5 Round-off errors

An important question, when dealing with Chebyshev approximations of high degree N, concerns the effect of round-off errors. A possible cause of error associated with the use of the recurrence formulas for calculating the coefficients of the Chebyshev expansion of derivatives, has been mentioned in Section 3.2.2. Another way to calculate the derivatives is to make use of the differentiation matrices whose entries have been given in the previous section. This technique is very often employed because it is not necessary

毕业论文（设计）文献综述开题报告考核

答辩小组对开题报告、外文翻译和文献综述评语及成绩评定：

成绩比例	开题报告占 (20%)	文献综述占 (10%)	外文翻译占 (10%)
分值			

答辩小组负责人（签名）_____

年　月　日