

# 偏微分方程的数值方法

*NUMERICAL METHODS OF PARTIAL  
DIFFERENTIAL EQUATIONS*



2019 年 8 月 5 日-16 日

教师：张强 (qzh@nju.edu.cn)

办公室：南京大学鼓楼校区乙楼 210

---

# 目 录

---

<b>第一章 有限差分方法的基本概念</b>	<b>1</b>
1.1 差商离散 . . . . .	1
1.1.1 待定系数法 . . . . .	1
1.1.2 函数逼近理论 . . . . .	2
1.1.3 符号演算方法 . . . . .	3
1.2 基本设计思想 . . . . .	5
1.3 线性差分格式的基本理论 . . . . .	9
1.3.1 预备知识 . . . . .	9
1.3.2 相容性 . . . . .	11
1.3.3 稳定性 . . . . .	13
1.3.4 收敛性 . . . . .	18
1.3.5 Lax-Richtmyer 等价定理 . . . . .	19
<b>第二章 热传导方程</b>	<b>20</b>
2.1 Crank-Nicolson 格式 . . . . .	20
2.2 Du Fort-Frankel 格式 . . . . .	23
2.3 跳点格式 . . . . .	27
2.4 数值格式的健壮性 . . . . .	28
2.5 线性扩散方程 . . . . .	33
2.5.1 冻结系数方法 . . . . .	33
2.5.2 积分插值方法 . . . . .	35
2.5.3 稳定性分析方法 . . . . .	37
2.5.4 具有间断系数的线性扩散方程 . . . . .	40
2.6 非线性扩散方程 . . . . .	43
2.7 高维扩散方程 . . . . .	45

<b>第三章 线性双曲型方程</b>	<b>54</b>
3.1 迎风格式和 Lax-Wendroff 格式	54
3.1.1 迎风格式	54
3.1.2 Lax-Wendroff 格式	55
3.2 稳定性分析方法	57
3.2.1 数值黏性方法	57
3.2.2 CFL 方法	58
3.2.3 单调格式与数值震荡	59
3.2.4 数值色散分析	61
3.3 双曲型方程组	63
3.4 高维对流方程	66
<b>第四章 非线性双曲守恒律</b>	<b>70</b>
4.1 弱解和熵解	70
4.2 守恒型差分格式	74
4.3 有限体积方法	77
4.3.1 基本框架	77
4.3.2 线性问题的有限体积格式	79
4.3.3 非线性问题的有限体积格式	81
4.4 Godunov 方法	82
4.4.1 EA 过程	82
4.4.2 REA 过程	86
4.5 稳定性和收敛性	87
4.5.1 单调保持格式	88
4.5.2 单调格式	89
4.5.3 TVD 格式	91
4.6 TVD 修正技术	93
4.6.1 数值通量修正技术	93
4.6.2 数值斜率修正技术	95

<b>第五章</b>	<b>边界条件的数值离散方法</b>	<b>98</b>
5.1	一维扩散方程的含导数边界条件	98
5.1.1	单侧离散方式	98
5.1.2	双侧离散方式	100
5.2	二维扩散方程的边界条件离散	102
5.2.1	本质边界条件	104
5.2.2	自然边界条件的处理	106
5.3	对流方程的人工边界条件	110
<b>第六章</b>	<b>椭圆型方程</b>	<b>113</b>
6.1	五点差分格式	113
6.1.1	规则内点的五点差分方程	114
6.1.2	非规则内点的五点差分方程	115
6.1.3	离散方程组	115
6.1.4	线性方程组的数值解法 <sup>‡</sup>	117
6.2	最大模估计	123
6.2.1	强最大值原理	123
6.2.2	简单估计	125
6.2.3	精细估计	126
6.3	提高数值精度的方法	128
6.3.1	Richardson 外推技术	128
6.3.2	九点格式	130
6.3.3	Kreiss 差分格式	131
6.4	有限元方法 <sup>‡</sup>	132
6.4.1	变分方法的基本理论	133
6.4.2	古典变分法	136
6.4.3	标准有限元方法	138

<b>第七章 对流扩散方程的数值方法</b>	<b>143</b>
7.1 数值困难 . . . . .	143
7.2 常用的解决方法 . . . . .	146
7.2.1 数值黏性修正方法 . . . . .	146
7.2.2 隐式格式 . . . . .	149
7.2.3 算子分裂方法 . . . . .	149
7.2.4 特征差分方法 . . . . .	150
7.2.5 有限元方法 . . . . .	151
<b>第八章 间断有限元方法</b>	<b>152</b>
<b>参考资料</b>	<b>152</b>
<b>附录</b>	<b>154</b>
A. 修正方程方法 <sup>‡</sup> . . . . .	154
B. 能量方法 <sup>‡</sup> . . . . .	157



---

# 第 1 章

## 有限差分方法的基本概念

---

用有限差商直接离散导数的数值方法，称为有限差分方法。数值实现较为方便，应用极其广泛。

### 1.1 差商离散

Newton 差商技术是主流方法。同时，以下三种技术也是广泛使用的。

#### 1.1.1 待定系数法

假设  $p(x)$  足够光滑，相应的  $m$  阶导数记为  $\mathcal{D}^m p(x)$ ，其中  $\mathcal{D}$  是微分算子。为简单起见，设离散模版  $\{x_{j+s}\}_{s=-l:r}$  是等距的，即

$$x_{j+s} = x_j + s\Delta x, \quad s = -l:r,$$

其中  $l$  和  $r$  是非负整数。设  $x_\star = x_j + \theta\Delta x$  为离散焦点，其中  $\theta \in [0, 1)$ 。利用待定系数法，可以确定  $\mathcal{D}^m p(x_\star)$  的差商公式

$$\mathcal{D}^m p(x_\star) = \sum_{s=-l}^r \alpha_s p(x_{j+s}) + \mathcal{O}((\Delta x)^\sigma), \quad (1.1)$$

其中  $\alpha_s$  是差商系数， $\sigma > 0$  是相容阶。

以离散焦点  $x_\star$  为中心，写出每个函数值的 Taylor 级数。由 (1.1) 可得

$$\mathcal{D}^m p(x_\star) = \sum_{k=0}^{\infty} \beta_k (\Delta x)^k \mathcal{D}^k p(x_\star) + \mathcal{O}((\Delta x)^\sigma), \quad (1.2a)$$

其中

$$\beta_k = \sum_{s=-l}^r \alpha_s \frac{(s-\theta)^k}{k!}, \quad k = 0, 1, 2, \dots \quad (1.2b)$$

比较等式 (1.2a) 两端的各阶导数, 由系数相等, 可知

$$\beta_0 = \beta_1 = \cdots = \beta_{m-1} = 0, \quad \beta_m = 1/(\Delta x)^m. \quad (1.3)$$

若能从线性方程组 (1.3) 解出  $\{\alpha_s\}_{s=-l}^r$ , 即可建立差商公式 (1.1)。继续用 (1.2b) 计算后续的  $\beta_k$ , 可知相容阶是

$$\sigma = \min\{k : k > m, \beta_k = 0\} - m. \quad (1.4)$$

请注意, (1.3) 可能多解或者无解。若其无解, 则说明离散模版的宽度不足, 需增加  $l$  或者  $r$ 。

**论题 1.1.** 设三个网格点  $x_{j-1}, x_j$  和  $x_{j+1}$  是等距分布的, 相应的间距是  $\Delta x$ 。建立二阶导数  $p_{xx}(x_j)$  的差商离散, 使相容阶尽可能的高。

答: 显然,  $\theta = 0$  和  $l = r = 1$ 。简单计算, 可知

$$\begin{aligned} \beta_0 &= \alpha_{-1} + \alpha_0 + \alpha_1, & \beta_1 &= -\alpha_{-1} + \alpha_1, & \beta_2 &= (\alpha_{-1} + \alpha_1)/2, \\ \beta_3 &= (-\alpha_{-1} + \alpha_1)/6, & \beta_4 &= (\alpha_{-1} + \alpha_1)/24, & \cdots \end{aligned}$$

令  $m = 2$ , 由 (1.3) 可知答案就是二阶中心差商, 即

$$p_{xx}(x_j) \approx \frac{\delta_x^2[p]_j}{(\Delta x)^2}. \quad (1.5)$$

由  $\beta_3 = 0$  和  $\beta_4 = 0$  可知, 它具有  $\mathcal{O}((\Delta x)^2)$  的逼近程度。  $\square$

### 1.1.2 函数逼近理论

在局部区域 (通常是离散模版的覆盖区域) 上, 构造相应的近似函数  $p_{\Delta x}(x)$ , 例如 Lagrange 多项式、Hermite 多项式、样条多项式、最佳逼近多项式、或者三角多项式等等。经典的函数逼近理论表明,  $p(x)$  和  $p_{\Delta x}(x)$  的两个导函数也具有近似关系。因此, 只要准确计算出近似函数  $p_{\Delta x}(x)$  在离散焦点的导数, 即可建立相应的差商离散。



**┆ 论题 1.2.** 在等距分布的三个网格点  $x_{j-1}, x_j, x_{j+1}$  上, 利用局部的抛物线插值过程, 给出  $p_{xx}(x_j)$  的有限差商离散。

答:  $p(x)$  在三个网格点上插值而成的抛物函数是

$$p_{\Delta x}(x) = p(x_j) + p[x_j, x_{j+1}](x - x_j) + \frac{1}{2}p[x_{j-1}, x_j, x_{j+1}](x - x_j)(x - x_{j-1}),$$

其中  $p[x_j, x_{j+1}]$  和  $p[x_{j-1}, x_j, x_{j+1}]$  分别是一阶和二阶 Newton 差商。基于函数插值理论, 有

$$p_{xx}(x_j) \approx p''_{\Delta x}(x_j) = p[x_{j-1}, x_j, x_{j+1}] = \frac{\delta_x^2[p]_j}{(\Delta x)^2}. \quad (1.6)$$

显然, 右端就是二阶中心差商离散, 具有二阶相容性。□

利用函数插值理论, 还有

$$p_{xx}(x_{j+1}) \approx p''_{\Delta x}(x_{j+1}) = p[x_{j-1}, x_j, x_{j+1}]$$

它给出  $p_{xx}(x_{j+1})$  的二阶偏心差商。简单计算可知, 它仅仅具有一阶相容性。请读者自行验证。

### 1.1.3 符号演算方法

设  $h$  是给定的移位距离。定义移位算子半群  $\{\mathbb{E}^s\}_{s \in \mathbb{R}}$ , 其中

$$\mathbb{E}^s p(x) = p(x + sh), \quad \forall p(x). \quad (1.7)$$

特别地,  $\mathbb{E} = \mathbb{E}^1$  是 (正向) 移位算子,  $\mathbb{E}^0 = \mathbb{1}$  是恒等算子,  $\mathbb{E}^{-1}$  是反向移位算子。注意到

$$\mathbb{E}^{a+b} = \mathbb{E}^a \cdot \mathbb{E}^b, \quad \forall a, b \in \mathbb{R}, \quad (1.8)$$

可知  $\mathbb{E}^{-1}$  是  $\mathbb{E}$  的逆。

在差分方法中, 移位算子扮演着重要的作用。首先, 所有的差分离散算子均可以用移位算子表示, 例如

1. 一阶向前差分算子:  $\Delta_+ \equiv \Delta = \mathbb{E} - \mathbb{1}$ ;
2. 一阶向后差分算子:  $\Delta_- = \mathbb{1} - \mathbb{E}^{-1}$ ;
3. 一步中心差分算子:  $\Delta_0 = \mathbb{E} - \mathbb{E}^{-1}$ ;
4. 半步中心差分算子:  $\delta = \mathbb{E}^{1/2} - \mathbb{E}^{-1/2}$ ;
5. 二阶中心差分算子:  $\delta^2 = \mathbb{E} - 2\mathbb{1} + \mathbb{E}^{-1}$ .

若要强调相应的操作变量, 通常将其标注在算子符号的右下角. 其次, 利用符号演算技巧, 差商离散的设计及其相容阶的推演, 均可以得到相应的简化. 换言之, 直接将算子符号视为普通变量, 将所有的算子运算转化为相应的函数运算. 具体内容可参见 Hildebrand (1956) 的工作.

**例 1.**

利用符号演算技巧, 移位算子的 Taylor 级数是

$$\mathbb{E} = \sum_{k=0}^{\infty} \frac{1}{k!} (h\mathcal{D})^k = e^{h\mathcal{D}}, \quad (1.9)$$

其中  $\mathcal{D}$  是微分算子. 注意到  $\mathbb{E} = \mathbb{1} + \Delta$ , 有

$$\mathcal{D} = \frac{1}{h} \ln \mathbb{E} = \frac{1}{h} \ln(\mathbb{1} + \Delta). \quad (1.10)$$

显然  $\Delta^m = \mathcal{O}(h^m)$ , 其中  $m$  是任意的正整数. 利用函数  $\ln(1+z)$  的 Taylor 展开公式, 可得

$$\mathcal{D} = \frac{1}{h} \Delta + \mathcal{O}(h), \quad (1.11a)$$

$$\mathcal{D} = \frac{1}{h} \left( \Delta - \frac{1}{2} \Delta^2 \right) + \mathcal{O}(h^2). \quad (1.11b)$$

借用函数  $\ln(1+z)$  的有理逼近技术<sup>1</sup>, 有

$$\mathcal{D} = \frac{1}{h} \frac{\Delta}{\mathbb{1} + \Delta/2} + \mathcal{O}(h^2), \quad (1.12)$$

---

<sup>1</sup>有理逼近也称为 Páde 逼近, 即  $f(x) \approx Q_m(x)/Q_n(x)$ , 其中  $Q_m(x)$  和  $Q_n(x)$  分别是  $m$  次多项式和  $n$  次多项式.

其中除法运算应当理解为左逆运算。  $\square$

**例 2.**

注意到  $\delta = e^{h\mathcal{D}/2} - e^{-h\mathcal{D}/2}$ , 有

$$\mathcal{D} = \frac{2}{h} \sinh^{-1} \left( \frac{1}{2} \delta \right). \quad (1.13)$$

显然  $\delta^m = \mathcal{O}(h^m)$ , 其中  $m$  是任意的正整数。利用函数  $\sinh^{-1}(z/2)$  的 Taylor 展开公式, 可得

$$\mathcal{D} = \frac{1}{h} \left[ \delta - \frac{1}{24} \delta^3 + \frac{3}{640} \delta^5 \right] + \mathcal{O}(h^6), \quad (1.14a)$$

$$\mathcal{D}^2 = \frac{1}{h^2} \left[ \delta^2 - \frac{1}{12} \delta^4 + \frac{1}{90} \delta^6 \right] + \mathcal{O}(h^6). \quad (1.14b)$$

利用函数  $\sinh^{-1}(z/2)$  的有理逼近技术, 有

$$\mathcal{D} = \frac{1}{2h} \frac{\Delta_0}{\mathbb{I} + \delta^2/6} + \mathcal{O}(h^4), \quad \mathcal{D}^2 = \frac{1}{h^2} \frac{\delta^2}{\mathbb{I} + \delta^2/12} + \mathcal{O}(h^4), \quad (1.15)$$

其中除法运算也应当理解为左逆运算。  $\square$

**注释 1.1.** 记  $\mathbf{i} = \sqrt{-1}$ 。指数函数  $e^{\mathbf{i}kx}$  可以用于快速检验差商离散的相容阶, 其中  $k$  是任意的实数。例如, 简单计算可知

$$\begin{aligned} \Delta e^{\mathbf{i}kx} &= e^{\mathbf{i}kx} [\mathbf{i}kh + \mathcal{O}(k^2h^2)] = h\mathcal{D}e^{\mathbf{i}kx} + \mathcal{O}(k^2h^2), \\ \delta^2 e^{\mathbf{i}kx} &= e^{\mathbf{i}kx} [-(kh)^2 + \mathcal{O}(k^4h^4)] = h^2\mathcal{D}^2 e^{\mathbf{i}kx} + \mathcal{O}(k^4h^4). \end{aligned}$$

因此说, 向前差商  $\Delta/h$  是一阶导数  $\mathcal{D}$  的一阶相容, 中心差商  $\delta^2/h^2$  是二阶导数  $\mathcal{D}^2$  的二阶相容。

## 1.2 基本设计思想

格式构造通常包括计算区域的离散、微分方程的离散和定解条件的离散。在三个设计步骤中, 微分方程的离散最为关键。

设  $T > 0$  是给定的终止时刻, 考虑一维热传导方程的周期边值问题:

$$u_t = au_{xx} + f(x, t), \quad (x, t) \in (0, 1) \times (0, T], \quad (1.16a)$$

相应的初值是

$$u(x, 0) = u_0(x), \quad x \in [0, 1]. \quad (1.16b)$$

其中扩散系数  $a > 0$  是给定常数,  $f(x, t)$  和  $u_0(x)$  是已知函数。

数值操作均基于某种结构的**离散网格**。对于模型问题 (HD) 而言, 等距时空网格<sup>2</sup>

$$\mathcal{T}_{\Delta x, \Delta t} = \left\{ (x_j, t^n) : x_j = j\Delta x, t^n = n\Delta t \right\}_{j=0:J}^{n=0:N} \quad (1.17)$$

是最常用的, 其中  $\Delta x = 1/J$  称为**空间步长**,  $\Delta t = T/N$  称为**时间步长**,  $N$  和  $J$  是给定的正整数。它由分别平行于空间轴和时间轴的两个直线 (段) 族交叉而成, 具有笛卡尔乘积型结构。平行于坐标轴的直线 (段) 称为网格线, 网格线的交点称为网格点。

将真解  $u(x, t)$  限制在离散网格  $\mathcal{T}_{\Delta x, \Delta t}$  上, 相应的离散数据集合

$$\{[u]_j^n = u(x_j, t^n)\}_{j=0:J}^{n=0:N} \quad (1.18)$$

是差分方法的数值逼近目标。换言之, 在网格点  $(x_j, t^n)$  上, 建立  $[u]_j^n$  的近似值  $u_j^n$ 。通常, 称其为数值解。

为实现上述目标, 我们需要离散偏微分方程和初边值条件。设真解  $[u]$  足够光滑。利用 Newton 差商理论或者 Taylor 展开公式, 可得<sup>3</sup>

$$[u_t]_j^n = \frac{[u]_j^{n+1} - [u]_j^n}{\Delta t} + \mathcal{O}(\Delta t), \quad (1.19a)$$

$$[u_{xx}]_j^n = \frac{[u]_{j+1}^n - 2[u]_j^n + [u]_{j-1}^n}{(\Delta x)^2} + \mathcal{O}((\Delta x)^2). \quad (1.19b)$$

<sup>2</sup> 设  $A \leq B$  是两个整数, 符号  $A : B$  表示从  $A$  到  $B$  的所有整数。

<sup>3</sup> 符号  $\mathcal{O}(\eta)$  的含义是指: 存在某个  $\eta_0$ , 使得当  $0 < \eta < \eta_0$  时有  $|\mathcal{O}(\eta)| \leq C\eta$ , 其中界定常数  $C > 0$  同  $\eta$  无关。

换言之, 时间导数离散为一阶向前差商, 空间导数离散为二阶中心差商。由于热传导方程 (1.16a) 在网格点  $(x_j, t^n)$  上精确成立, 有

$$\frac{[u]_j^{n+1} - [u]_j^n}{\Delta t} - a \frac{[u]_{j+1}^n - 2[u]_j^n + [u]_{j-1}^n}{(\Delta x)^2} = f_j^n + O((\Delta x)^2 + \Delta t),$$

其中  $f_j^n = f(x_j, t^n)$  是已知信息,  $j = 1 : J - 1$  和  $n = 0 : N - 1$ 。略去无穷小量, 用数值解替换真解, 可得 (1.16a) 的差分方程

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - a \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2} = f_j^n. \quad (1.20a)$$

通常, 将差分方程 (1.20a) 等价变形为

$$\Delta_t u_j^n = \mu a \delta_x^2 u_j^n + \Delta t f_j^n, \quad (1.20b)$$

其中  $\Delta_t u_j^n = u_j^{n+1} - u_j^n$  是时间方向的一阶向前差分,  $\mu = \Delta t / (\Delta x)^2$  是网比,  $\delta_x^2 u_j^n = u_{j-1}^n - 2u_j^n + u_{j+1}^n$  是空间方向的二阶中心差分。

类似地, 时间导数离散为一阶向后差商, 空间导数依旧离散为二阶中心差商, 可得差分方程

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - a \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{(\Delta x)^2} = f_j^{n+1}, \quad (1.21a)$$

或者等价的

$$\Delta_t u_j^n = \mu a \delta_x^2 u_j^{n+1} + \Delta t f_j^{n+1}. \quad (1.21b)$$

图 1.1: 离散模版。左: 全显离散(1.20); 右: 全隐离散(1.21)



在差分方程中出现的网格点集, 称为相应的**离散模版**。参见图 1.1, 差分方程 (1.20) 和 (1.21) 的离散模版具有不同的结构。前者称为**显式离散**的, 因

为离散模版的顶端只含一个网格点值；后者称为**隐式离散**的，因为离散模版的顶端同时含有三个网格点值，要多个差分方程耦合起来才能解出。

对于模型问题 (HD) 而言，定解条件的离散是非常简单的，只需在网格点直接赋值即可。

将所有出现的差分方程汇总起来，即可建立模型问题 (HD) 的两个古典格式。基于等距时空网格 (1.17)，**全显格式**是

$$\Delta_t u_j^n = \mu a \delta_x^2 u_j^n + \Delta t f_j^n, \quad j = 1 : J - 1, n = 0 : N - 1, \quad (1.22)$$


**全隐格式**（或者 Lasonen 格式）是


$$\Delta_t u_j^n = \mu a \delta_x^2 u_j^{n+1} + \Delta t f_j^{n+1}, \quad j = 1 : J - 1, n = 0 : N - 1, \quad (1.23)$$

相应的数值初值均是

$$u_j^0 = u_0(x_j), \quad j = 0 : J. \quad (1.24)$$

它们的主要差异是热传导方程的离散方式，全显格式基于显式离散的差分方程 (1.20)，而全隐格式基于隐式离散的差分方程 (1.21)。

 **注释 1.2.** 由于微分方程的离散是差分格式的核心，相应的差分方程常常被称为差分格式。例如，显式离散的差分方程 (1.20) 也称为全显格式。

 **论题 1.3.** 可行性和效率分析。

**答：**全显格式可以直接计算，而全隐格式涉及到线性方程组的数值求解。若采用三对角矩阵的追赶法进行求解，全隐格式的单步计算量约为全显格式的 2 倍。但是，全隐格式可以采用大的时间步长，整体的计算效率更高。 □

## 1.3 线性差分格式的基本理论

### 1.3.1 预备知识

通常，**时空网格**是空间网格  $\mathcal{T}_{\Delta x} = \{x_j\}_{\forall j}$  和时间网格  $\mathcal{T}_{\Delta t} = \{t^n\}_{\forall n}$  的笛卡尔乘积，即

$$\mathcal{T}_{\Delta x, \Delta t} = \mathcal{T}_{\Delta x} \otimes \mathcal{T}_{\Delta t} = \{(x_j, t^n)\}_{\forall j}, \quad (1.25)$$

其中  $\Delta x$  称为空间步长， $\Delta t$  称为时间步长，符号  $\forall j$  和  $\forall n$  模糊指定了网格点的编号范围。为简单起见，默认时空网格  $\mathcal{T}_{\Delta x, \Delta t}$  是等距的，相应的空间网格点和时间网格点分别定义为

$$x_j = x_0 + j\Delta x, \quad t^n = t^0 + n\Delta t,$$

其中  $(x_0, t^0)$  是参考网格点。

位于时空网格  $\mathcal{T}_{\Delta x, \Delta t}$  或相应子集上的离散函数<sup>4</sup>，通常称为**网格函数**。基于逐层推进的策略，**空间网格函数**

$$u^n = \{u_j^n\}_{\forall j}, \quad \forall n, \quad (1.26)$$

是备受关注的研究对象。除简单直观的逐点描述之外，空间网格函数还可以借用**离散范数**进行整体度量。本讲义主要使用两种离散范数，即

$$\|u^n\|_{\infty} \equiv \|u^n\|_{\infty, \Delta x} = \max_{\forall j} |u_j^n|, \quad (1.27a)$$


$$\|u^n\|_2 \equiv \|u^n\|_{2, \Delta x} = \left[ \sum_{\forall j} |u_j^n|^2 \Delta x \right]^{\frac{1}{2}}. \quad (1.27b)$$

前者称为（离散）最大模，后者称为（离散） $L^2$  模。要注意：**离散范数的定义同空间网格密切相关**。

---

<sup>4</sup>离散和连续是相对的两个概念。离散（型）函数是指定义域集合的元素是有限的，或是可数的；连续（型）函数是指定义域集合的元素同实数域等势。

差分格式的描述方式有两种。其一是直观的局部描述，也就是位于不同网格点的差分方程。例如，差分方程 (1.20a) 是全显格式在内部网格点的局部描述，差分方程 (1.21a) 是全隐格式在内部网格点的局部描述；它们的离散对象非常清晰，都是热传导方程 (1.16a)。

 **注释 1.3.** 对于初边值条件，差分方程也具有相应的局部描述；略。

其二是繁琐的整体描述，即逼近模型问题的完整离散系统。换言之，经过适当的等价变形（例如数乘、消元或局部的可逆变换等），相关网格点的差分方程可以汇总为

$$\mathbb{B}_1 u^{n+1} = \mathbb{B}_0 u^n + \Delta t G^n, \quad \forall n, \quad (1.28)$$

其中  $\Delta t$  是时间步长， $u^n \equiv \{u_j^n\}_{\forall j}$  是  $t^n$  时刻的未知网格函数， $G^n$  是已知网格函数， $\mathbb{B}_1$  和  $\mathbb{B}_0$  是网格函数空间到自身的线性算子<sup>5</sup>。

为简单起见，不妨将 (1.28) 作为双层格式的一个抽象定义。基于可行性要求，默认  $\mathbb{B}_1$  是可逆的，定义相应的**规范形式**

$$u^{n+1} = \mathbb{B} u^n + \Delta t H^n, \quad \forall n, \quad (1.29)$$

其中  $\mathbb{B} = \mathbb{B}_1^{-1} \mathbb{B}_0$  和  $H^n = \mathbb{B}_1^{-1} G^n$ 。若差分格式具有相同的规范形式，则它们是相同的。

通常，某某格式是指依赖网格参数的一族差分格式。在网格加密的过程中，它呈现出来的数值现象是重要的研究内容。事实上，网格参数  $\Delta x$  和  $\Delta t$  可以独自趋于零，但是理论分析将略显繁琐和困难。通常，假定两个网格参数沿着**加密路径**

$$\Delta x = \mathcal{G}(\Delta t) \quad (1.30)$$

趋于零，其中  $\mathcal{G}(\cdot)$  是满足  $\mathcal{G}(0) = 0$  的连续函数。若某个数值现象或数值概念（例如即将介绍的相容性、稳定性和收敛性）同加密路径无关，则称其是**无条件的**；否则，称其是**有条件的**。

---

<sup>5</sup>事实上， $\mathbb{B}_0$  和  $\mathbb{B}_1$  还可以同时时间层数和网格函数有关。若同网格函数有关，则差分格式是非线性的。



### 1.3.2 相容性

相容性概念描述了差分格式同离散对象的逼近程度, 有两种定义方式, 其一是逐点相容性, 其二是整体相容性。前者基于局部描述, 简单直观, 应用广泛。后者基于整体描述, 理论严谨, 但是略显繁琐。事实上, 两种相容性概念具有密切的联系, 均通过**局部截断误差**来展现。

🕒 **定义 1.1.** 若差分方程在网格点  $(x_j, t^n)$  处具有局部描述

$$\mathcal{L}_{\Delta x, \Delta t} u_j^n = g_j^n, \quad (1.31)$$

则相应的**局部截断误差**是

$$\tau_j^n = \mathcal{L}_{\Delta x, \Delta t} [u]_j^n - g_j^n, \quad (1.32)$$

其中  $[u]$  是满足偏微分方程

$$\mathcal{L}[u] = g \quad (1.33)$$

的充分光滑函数。当  $\Delta x$  和  $\Delta t$  趋于零<sup>6</sup>时, 若

$$\tau_j^n \rightarrow 0,$$

则称差分方程 (1.31) **逐点相容**于偏微分方程 (1.33)。若存在不可改善的两个正数  $m_1$  和  $m_2$ , 使得

$$\tau_j^n = O((\Delta x)^{m_1} + (\Delta t)^{m_2}),$$

则称差分方程 (1.31) 的**局部截断误差阶**是  $(m_1, m_2)$ 。

换言之, 局部截断误差就是数值离散过程中我们所丢弃的那些无穷小量。要强调指出, 在逐点相容性概念中, **差分方程要同其离散对象具有清楚的逐项对应关系**。一个相对简便的原则是, 只要差分方程同离散对象具有相同的物理量纲<sup>7</sup>, 相应的局部截断误差阶就是正确的。

<sup>6</sup>  $\Delta x$  和  $\Delta t$  趋于零的默认含义是指, 它们沿着某条加密路径趋于零。

<sup>7</sup> 物理量纲是指米、秒和千克等物理单位。

**论题 1.4.** 计算全显格式 (1.20a) 的局部截断误差阶。

**答：**由于差分方程 (1.20a) 同热传导方程 (1.16a) 具有清晰的逐项对应关系，相应的局部截断误差为

$$\tau_j^n = \frac{[u]_j^{n+1} - [u]_j^n}{\Delta t} - a \frac{[u]_{j+1}^n - 2[u]_j^n + [u]_{j-1}^n}{(\Delta x)^2} - f_j^n, \quad (1.34)$$

其中  $[u]$  满足热传导方程 (1.16a)。假设  $[u]$  足够光滑，使得下面的 Taylor 展开都是合法操作。以  $(x_j, t^n)$  为展开中心，有

$$\begin{aligned} [u]_{j\pm 1}^n &= [u]_j^n \pm [\mathcal{D}_x u]_j^n \Delta x + \frac{1}{2} [\mathcal{D}_x^2 u]_j^n (\Delta x)^2 \\ &\quad \pm \frac{1}{6} [\mathcal{D}_x^3 u]_j^n (\Delta x)^3 + \frac{1}{4!} [\mathcal{D}_x^4 u]_j^n (\Delta x)^4 + \cdots, \\ [u]_j^{n+1} &= [u]_j^n + [\mathcal{D}_t u]_j^n \Delta t + \frac{1}{2} [\mathcal{D}_t^2 u]_j^n (\Delta t)^2 + \cdots, \end{aligned}$$

其中  $\mathcal{D}$  是微分算子符号，其下标指明操作变量的名称，上标表示求导运算的阶数。将所有展开式代入到 (1.34)，整理可得

$$\begin{aligned} \tau_j^n &= \left[ \mathcal{D}_t + \frac{1}{2} \Delta t \mathcal{D}_t^2 + \cdots \right] [u]_j^n \\ &\quad - a \left[ \frac{2}{2!} \mathcal{D}_x^2 + \frac{2}{4!} (\Delta x)^2 \mathcal{D}_x^4 + \cdots \right] [u]_j^n - f_j^n. \end{aligned}$$


注意到  $[\mathcal{D}_t - a\mathcal{D}_x^2][u]_j^n = f_j^n$ ，可知  $\tau_j^n$  的零阶项部分等于零。换言之，

$$\tau_j^n = \frac{1}{2} \mathcal{D}_t^2 [u]_j^n \Delta t - \frac{a}{12} \mathcal{D}_x^4 [u]_j^n (\Delta x)^2 + \cdots = \mathcal{O}(\Delta x^2 + \Delta t). \quad (1.35)$$

一般而言，这个结果已经是最优的；由于它同加密路径无关，故而全显格式无条件具有 (2, 1) 阶局部截断误差。□

类似地，全隐格式 (1.21a) 也无条件具有 (2, 1) 阶局部截断误差。

**注 1.4.** 局部截断误差阶同推导过程无关。换言之，在任意（一个或多个）位置执行 Taylor 展开，最终的理论结果都是一样的。留作练习。


 **注释 1.5.** 通常, 在一个差分格式中, 位于不同网格点的差分方程可能具有明显的差异, 比如离散对象不同, 或者离散方式不同。逐点相容性概念无法描述差分方程的交互影响, 只有整体相容性概念才能刻画差分格式的整体离散效果。整体相容性概念基于差分格式的规范形式, 整体相容阶同度量方式有关。略。

### 1.3.3 稳定性

稳定性概念同微分方程定解问题的真解无关, 是差分格式的固有性质, 刻画数值解关于定解数据的连续依赖性。考虑线性差分格式 (1.29) 或者

$$u^{n+1} = \mathbb{B}u^n + \Delta t H^n, \quad n = 0 : N-1, \quad (1.36)$$

其中  $N = \lfloor T/\Delta t \rfloor$  是不超过  $T/\Delta t$  的最大整数,  $T > 0$  是给定的终止时刻。基于线性叠加原理, 差分格式 (1.36) 的稳定性可以分解为两个基本概念。

 **定义 1.2.** 令  $H^n \equiv 0$ , 对应 (1.36) 的齐次线性差分格式是

$$u^{n+1} = \mathbb{B}u^n, \quad n = 0 : N-1. \quad (1.37)$$

给定离散范数  $\|\cdot\| = \|\cdot\|_{\Delta x}$ 。当  $\Delta x$  和  $\Delta t$  趋于零时, 若 (1.37) 的数值解满足

$$\|u^n\| \leq K\|u^0\|, \quad \forall n = 0 : N, \quad (1.38)$$

其中界定常数  $K > 0$  同  $\Delta x, \Delta t$  和  $u^0$  均无关<sup>8</sup>, 则称差分格式 (1.36) 按  $\|\cdot\|$  模具有初值稳定性。


事实上, (1.38) 仅仅指明数值解关于初值的有界性。注意到差分格式的线性结构, 若  $\{u^n\}_{n=0:N}$  和  $\{w^n\}_{n=0:N}$  均满足差分格式 (1.36), 则由初值稳定性的定义 1.2 可知

$$\|u^n - w^n\| \leq K\|u^0 - w^0\|, \quad n = 0 : N. \quad (1.39)$$

---


<sup>8</sup>若界定常数  $K$  同  $T$  有关, 则称差分格式具有短时间的初值稳定性。若界定常数同  $T$  无关, 则称差分格式具有长时间的初值稳定性。

换言之，初值扰动不会造成数值解的巨变，数值解连续依赖于初值。这才是初值稳定性的根本含义。


 **定义 1.3.** 设差分格式 (1.36) 的初值是零，即  $u^0 \equiv 0$ 。给定离散范数  $\|\cdot\| = \|\cdot\|_{\Delta x}$ 。若当  $\Delta t$  和  $\Delta x$  趋于零时，数值解满足

$$\|u^n\| \leq M \sum_{m=0}^{n-1} \|H^m\| \Delta t, \quad \forall n = 1 : N, \quad (1.40)$$

其中界定常数  $M > 0$  同  $\Delta x$  和  $\Delta t$  均无关，则称差分格式 (1.36) 具有**右端项稳定性**。

 **注释 1.6.** 平行于线性微分方程的 *Duhamel* 原理，线性差分格式也具有类似的结论：齐次线性差分格式 (1.37) 的初值稳定性，蕴含非齐次线性差分格式 (1.36) 的右端项稳定性。

### 最大模稳定性

 **论题 1.5.** 当且仅当  $\mu a \leq 1/2$  时，模型问题 (HP) 的全显格式具有最大模初值稳定性。

**答：**作为连续问题的数值离散，差分格式的稳定性表现最好能够继承或者接近定解问题的适定性表现。为简单起见，设  $f \equiv 0$ ，即连续问题和相应的差分格式都是齐次的。由偏微分方程的经典理论可知，此时的模型问题 (HP) 满足**最大模原理**，即**真解的最大模不增**。因此，希望全显格式也满足**离散最大模原理**，即**数值解的离散最大模也是不增的**。

当  $\mu a \leq 1/2$  时，位于任意网格点的差分方程

$$u_j^{n+1} = \mu a(u_{j-1}^n + u_{j+1}^n) + (1 - 2\mu a)u_j^n, \quad \forall j, \quad (1.41)$$

都具有显式的凸组合系数结构，即右侧的差分系数都是非负的，且系数总和不超过 1。此时，有

$$|u_j^{n+1}| \leq \max(|u_{j-1}^n|, |u_j^n|, |u_{j+1}^n|), \quad \forall j.$$

因此, 数值解满足  $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$ , 全显格式具有最大模初值稳定性。

事实上,  $\mu a \leq 1/2$  也是必要条件。否则, (1.41) 的等号右端出现负系数, 凸组合的系数结构不复存在, 离散的最大模原理遭到破坏。下面给出一个反例。设  $2J\Delta x = 1$ , 定义

$$u_j^0 = (-1)^j, \quad j = 0 : 2J.$$

利用数学归纳法, 可知模型问题 (HP) 的全显格式具有数值解

$$u_j^n = (1 - 4\mu a)^n (-1)^j. \quad (1.42)$$

由于  $|1 - 4\mu a| > 1$ , 数值解将趋于无穷, 故而全显格式是不稳定的。证毕。□

‡ **论题 1.6.** 模型问题 (HP) 的全隐格式无条件具有最大模初值稳定性。

答: 设  $f \equiv 0$ , 即连续问题和差分格式都是齐次的。此时, 对于任意的网比  $\mu$ , 全隐格式在所有网格点的差分方程

$$(1 + 2\mu a)u_j^{n+1} = u_j^n + \mu a(u_{j-1}^{n+1} + u_{j+1}^{n+1})$$

都具有隐式的凸组合系数结构, 即右端系数都是正的, 且左端 (离散焦点) 的系数不超过右端系数之和。因此, 模型问题 (HP) 的全隐格式满足离散最大模原理。

相对直白的论证过程如下: 设最大模在某个网格点取到, 不妨设  $|u_{j_0}^{n+1}| = \|u^{n+1}\|_\infty$ 。利用  $j_0$  点的差分方程, 有

$$\begin{aligned} (1 + 2\mu a)|u_{j_0}^{n+1}| &= |u_{j_0}^n + \mu a(u_{j_0-1}^{n+1} + u_{j_0+1}^{n+1})| \\ &\leq \|u^n\|_\infty + 2\mu a\|u^{n+1}\|_\infty. \end{aligned}$$

因此, 数值解满足  $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$ 。换言之, 全隐格式无条件具有最大模初值稳定性。□

对于模型问题 (HI) 和零边值的模型问题 (HD), 相应的古典格式具有完全相同的最大模初值稳定性结论: **当且仅当  $\mu a \leq 1/2$  时, 全显格式稳定; 全**

**隐格式是无条件稳定的。**相关证明是简单的，除了全显格式的必要性条件。具体而言，是

1. 对于模型问题 (HI)，必要条件的证明有两步。首先，证明

$$u_j^n = \sum_{k=0}^{\infty} e^{-2^k} \left[ 1 - 4\mu a \sin^2(2^{k-1}\pi\Delta x) \right]^n \cos(2^k\pi j\Delta x)$$

是全显格式的数值解。然后，令  $\Delta x = 2^{-m}$ ，证明：当  $\mu a > 1/2$  时，在  $x_0 = 0$  点的数值解满足  $\lim_{m \rightarrow \infty} u_0^m = \infty$ ，故而全显格式是不稳定的。

2. 对于零边值的模型问题 (HD)，必要条件可以利用直接矩阵方法给出。

事实上，上述三种模型问题的古典格式具有类似的  $L^2$  模稳定性结论：**当且仅当  $\mu a \leq 1/2$  时，全显格式稳定；全隐格式是无条件稳定的。**对于模型问题 (HI) 和 (HP)，Fourier 方法堪称是最简便的论证方式；对于零边值的模型问题 (HD)，直接矩阵方法是有效的论证方式。

## $L^2$ 模稳定性

对于均匀网格，且位于任意网格点的差分方程都具有相同形式（纯初值问题或者周期边值问题）的差分格式，Fourier 方法是  $L^2$  模稳定性的分析利器。

### 增长因子的计算

任取波数  $k \in \mathfrak{R}$ ，将模态解

$$u_j^n = \lambda^n e^{ikj\Delta x}, \quad \forall j \forall n \quad (1.43)$$

代入到差分方程。利用简单的代数演算，即可导出增长因子  $\lambda = \lambda(k)$ 。

### von Neumann 条件的判定

利用 Parseval 恒等式，可知双层格式的数值解满足

$$\begin{aligned} \|u^n\|_{2,\Delta x} &= \|\hat{u}^n\|_{L^2(\mathfrak{R})} \leq \sup_{k \in \mathfrak{R}} |\lambda(k)| \|\hat{u}^{n-1}\|_{L^2(\mathfrak{R})} \leq \cdots \\ &\leq \left[ \sup_{k \in \mathfrak{R}} |\lambda(k)| \right]^n \|\hat{u}^0\|_{L^2(\mathfrak{R})} = \left[ \sup_{k \in \mathfrak{R}} |\lambda(k)| \right]^n \|u^0\|_{2,\Delta x}. \end{aligned} \quad (1.44)$$

基于此, 可以导出著名的 **von Neumann 条件**: 当  $\Delta t$  适当小时, 有

$$|\lambda(k)| \leq 1 + C\Delta t, \quad \forall k \in \mathbb{R}, \quad (1.45)$$

其中界定常数  $C > 0$  同  $k$  和  $\Delta t$  均无关。反之亦然。

若增长因子没有显式出现时间步长  $\Delta t$ , 则 von Neumann 条件 (1.45) 的界定常数可取  $C = 0$ , 即

$$|\lambda(k)| \leq 1, \quad \forall k \in \mathbb{R}. \quad (1.46)$$

以示区别, 不妨称其为严格的 von Neumann 条件。

**¶ 论题 1.7.** 考虑模型问题 (HP) 和 (HI) 的两个古典格式, 其中  $f \equiv 0$ 。利用 Fourier 方法, 建立相应的  $L^2$  模初值稳定性结论。

答: 将模态解  $u_j^n = \lambda^n e^{ikj\Delta x}$  代入到全显格式, 简单计算可得

$$\begin{aligned} \lambda(k) &= 1 + \mu a(e^{ik\Delta x} - 2 + e^{-ik\Delta x}) \\ &= 1 - 4\mu a \sin^2\left(\frac{1}{2}k\Delta x\right). \end{aligned} \quad (1.47)$$

增长因子没有显式出现  $\Delta t$ , 相应的 von Neumann 条件是

$$-1 \leq 1 - 4\mu a \sin^2\left(\frac{1}{2}k\Delta x\right) \leq 1, \quad \forall k.$$

它等价于时空约束条件  $\mu a \leq 1/2$ 。由于增长因子是一个数, 它也是全显格式具有  $L^2$  模稳定性的充要条件。

类似地, 全隐格式的增长因子是

$$\lambda(k) = \left[1 + 4\mu a \sin^2\left(\frac{1}{2}k\Delta x\right)\right]^{-1}. \quad (1.48)$$

对于任意的网比, von Neumann 条件都是恒成立的。因此, 全隐格式无条件具有  $L^2$  模初值稳定性。  $\square$

### 1.3.4 收敛性

一个差分格式是否有用, 最终要看数值解在计算机上的近似结果, 是否有效地逼近真解。这个问题的回答, 涉及到两个方面。其一是前节给出的稳定性概念, 即差分格式的近似数值解是否接近差分格式的(准确)数值解。其二是本节关注的收敛性概念, 即差分格式的(准确)数值解是否逼近问题的真解。前者同舍入误差密切相关, 是差分格式的固有性质。后者同方法误差相关, 没有考虑舍入误差的影响。这种理想化的假设, 也是收敛性概念同稳定性概念的本质区别之一。

③ **定义 1.4.** 给定离散范数  $\|\cdot\| = \|\cdot\|_{\Delta x}$ 。当  $\Delta x$  和  $\Delta t$  趋于零时, 若数值误差(空间)网格函数  $e^n = \{e_j^n\}_{\forall j}$  满足

$$\|e^n\| \rightarrow 0, \quad \forall n,$$

则称差分格式按  $\|\cdot\|$  模 **收敛** 于定解问题。若存在不可改善的两个正数  $m_1$  和  $m_2$ , 使得

$$\|e^n\| = \mathcal{O}((\Delta x)^{m_1} + (\Delta t)^{m_2}), \quad \forall n,$$

则称差分格式按  $\|\cdot\|$  模具有  $(m_1, m_2)$  阶**精度**(或者误差)。

针对收敛性概念的理论分析, 收敛分析(convergence analysis)和误差估计(error estimate)是常常被混用的两个术语。事实上, 它们存在细微的差别, 特别是真解的光滑性假定不同。通常, 收敛分析要求偏低, 而误差估计偏高。

⚓ **论题 1.8.** 设模型问题 (HP)、(HD) 和 (HI) 的真解足够光滑, 建立相应古典格式的最大模误差估计。

**答:** 以模型问题 (HP) 的全显格式为例。注意到差分格式的线性结构, 利用逐点相容性概念, 可得误差方程

$$e_j^{n+1} = \mu a(e_{j-1}^n + e_{j+1}^n) + (1 - 2\mu a)e_j^n + \Delta t \tau_j^n, \quad \forall j \forall n, \quad (1.49)$$

其中  $\tau_j^n$  是局部截断误差, 参见 (1.35)。假设  $[u_{xxxx}]$  在  $[0, 1] \times [0, T]$  上连续



有界, 则

$$\max_{\forall j \forall n} |\tau_j^n| = \mathcal{O}((\Delta x)^2 + \Delta t).$$

注意到误差方程 (1.49) 同全显格式的表述极其相似, 故而仿照右端稳定性概念的论证过程, 建立相邻时刻的误差度量递推关系式。

当  $\mu a \leq 1/2$  时, 位于 (1.49) 右端关于误差函数的三个系数都是非负的, 且它们的总和不超过 1。注意到误差函数的周期性, 有

$$\|e^{n+1}\|_\infty \leq \|e^n\|_\infty + \Delta t \max_{\forall j} |\tau_j^n|.$$

利用数学归纳法, 可知<sup>9</sup>

$$\|e^n\|_\infty \leq \|e^0\|_\infty + \sum_{m=0}^{n-1} \max_{\forall j} |\tau_j^m| \Delta t, \quad n\Delta t \leq T.$$

注意到  $\|e^0\|_\infty = 0$  和  $n\Delta t \leq T$ , 模型问题 (HP) 的全显格式满足

$$\max_{n\Delta t \leq T} \|e^n\|_\infty = \mathcal{O}((\Delta x)^2 + \Delta t). \quad (1.50)$$

换言之, 全显格式的最大模误差阶是  $(2, 1)$ , 恰好等于它的相容阶。□

### 1.3.5 Lax-Richtmyer 等价定理

在 1956 年, 著名的 **Lax-Richtmyer 等价定理**<sup>10</sup> 就已经指出:

假设线性微分方程定解问题是适定的。若线性差分格式同它是相容的, 则稳定性和收敛性是等价的, 且误差阶不低于相容阶。

简而言之, 相容性和稳定性可以蕴含收敛性。

<sup>9</sup>若求和号的上标大于下标, 则它等于零。

<sup>10</sup>P. D. Lax and R. D. Richtmyer, *Survey of the stability of linear finite difference equations*, Comm. Pure Appl. Math., 9 (1956), 267-293

---

## 第 2 章

# 热传导方程

---

考虑线性常系数热传导方程

$$u_t = au_{xx}, \quad a > 0 \quad (2.1)$$

的纯初值问题或者周期边值问题

### 2.1 Crank-Nicolson 格式

设  $\theta \in [0, 1]$  是给定的权重。将热传导方程 (2.1) 的两个古典格式组合起来, 可得加权平均格式

$$\Delta_t u_j^n = \theta \mu a \delta_x^2 u_j^{n+1} + (1 - \theta) \mu a \delta_x^2 u_j^n. \quad (2.2)$$

它有时也称为六点格式或 Rose 格式。

**‡ 论题 2.1.** 计算加权平均格式 (2.2) 的局部截断误差。

**答:** 将网格点值按照某种对称方式进行分组, 相应的推导过程可以更加轻松。此时, 加权平均格式的局部截断误差是

$$\begin{aligned} \tau_j^n &= \frac{1}{\Delta t} \left( [u]_j^{n+1} - [u]_j^n \right) \\ &\quad - \frac{a}{(\Delta x)^2} \left[ \frac{1}{2} \delta_x^2 \left( [u]_j^{n+1} + [u]_j^n \right) + \left( \theta - \frac{1}{2} \right) \delta_x^2 \left( [u]_j^{n+1} - [u]_j^n \right) \right], \end{aligned}$$

其中  $[u]$  是热传导方程 (2.1) 的真解。

以  $(x_j, t^{n+1/2})$  为展开中心, 其中  $t^{n+1/2} = (t^n + t^{n+1})/2$ 。利用移位算子和符号演算技巧, 有

$$\begin{aligned} [u]_j^{n+1} - [u]_j^n &= \left[ e^{\frac{\Delta t}{2} \mathcal{D}_t} - e^{-\frac{\Delta t}{2} \mathcal{D}_t} \right] [u]_j^{n+\frac{1}{2}}, \\ \delta_x^2 ([u]_j^{n+1} \pm [u]_j^n) &= \left[ e^{\frac{\Delta t}{2} \mathcal{D}_t} \pm e^{-\frac{\Delta t}{2} \mathcal{D}_t} \right] \left[ e^{-\Delta x \mathcal{D}_x} - 2 + e^{\Delta x \mathcal{D}_x} \right] [u]_j^{n+\frac{1}{2}}, \end{aligned}$$

其中  $\mathcal{D}_t$  和  $\mathcal{D}_x$  是相应方向的微分算子。继续利用符号演算技巧, 写出所有指数运算的 Taylor 展开公式。代入到局部截断误差的定义, 利用热传导方程 (2.1) 进行化简和整理, 可得

$$\tau_j^n = -\Delta t(\theta - \frac{1}{2})[\mathcal{D}_t^2 u]_j^{n+\frac{1}{2}} - \frac{a(\Delta x)^2}{12}[\mathcal{D}_x^4 u]_j^{n+\frac{1}{2}} + \mathcal{O}((\Delta x)^4 + (\Delta t)^2).$$

换言之, 加权平均格式至少具有 (2, 1) 阶局部截断误差。  $\square$

在局部截断误差的上述表达式中, 等号右端的前两项就是主项部分。若前两项整体或者部分消失为零, 即可得到高阶相容的差分格式:

1. 令  $\theta = 1/2$ , 相应的加权平均格式

$$\Delta_t u_j^n = \frac{1}{2}\mu a(\delta_x^2 u_j^{n+1} + \delta_x^2 u_j^n) \quad (2.3)$$

就是著名的 Crank-Nicolson (CN) 格式<sup>1</sup>。它无条件具有 (2, 2) 阶局部截断误差。

事实上, CN 格式的构造可以解读如下: 在中心  $(x_j, t^{n+1/2})$  处, 利用**算术平均和中心离散技术**, 建立时间导数和空间导数的对称离散。换言之, 对称离散在精度阶方面具有优势。

2. 注意到  $[\mathcal{D}_t^2 u] = a^2[\mathcal{D}_x^4 u]$ , 将局部截断误差的表达式改写为

$$\tau_j^n = -\Delta t \left[ \frac{1}{12\mu a} + \theta - \frac{1}{2} \right] [\mathcal{D}_t^2 u]_j^{n+1/2} + \mathcal{O}((\Delta x)^4 + (\Delta t)^2).$$

若等距时空网格满足

$$\mu a = [6(1 - 2\theta)]^{-1}, \quad (2.4)$$

---

<sup>1</sup>J. Crank and P. Nicolson, *A practical method for numerical evaluation of solution of partial differential equations of the heat-conduction*, Proc. Camb. Philos. Soc., 43 (1947), 50–67

则相应的加权平均格式就是著名的 Douglas 格式<sup>23</sup>。其局部截断误差满足  $\tau_j^n = \mathcal{O}((\Delta x)^4)$ ，达到**整体四阶相容**。

利用 Lax-Richtmyer 等价定理可知：若 CN 格式和 Douglas 格式是稳定的，则它们将具有高阶精度。

同古典格式相比，加权平均格式的稳定性结论同度量方式有关，

**↓ 论题 2.2.** 讨论加权平均格式 (2.2) 的  $L^2$  模稳定性。

答：设  $k \in \Re$  是任意波数。将模态解  $u_j^n = \lambda^n e^{ikj\Delta x}$  代入到 (2.2)，简单计算可得增长因子

$$\lambda = \lambda(k) = \frac{1 - 4\mu a(1 - \theta) \sin^2(\frac{1}{2}k\Delta x)}{1 + 4\mu a\theta \sin^2(\frac{1}{2}k\Delta x)}. \quad (2.5)$$

注意到增长因子没有显式出现  $\Delta t$ ，加权平均格式具有  $L^2$  模稳定性的充要条件是严格的 von Neumann 条件，即  $|\lambda(k)| \leq 1$  或者等价的

$$-1 - 4\mu a\theta s \leq 1 - 4\mu a(1 - \theta)s \leq 1 + 4\mu a\theta s,$$

其中  $s = \sin^2(\frac{1}{2}k\Delta x) \in [0, 1]$ 。显然，右端不等式恒成立，而左端不等式成立的充要条件是

$$\mu a(1 - 2\theta) \leq 1/2. \quad (2.6)$$

因此，当且仅当 (2.6) 成立时，加权平均格式 (2.2) 具有  $L^2$  模稳定性。具体来说，就是：当  $\theta < 1/2$  时，它称作偏显格式，是有条件稳定；当  $\theta \geq 1/2$  时，它称作偏隐格式，是无条件稳定。□

CN 格式无条件具有  $L^2$  模稳定性和高阶相容性，是非常理想的数值格式，得到相当广泛的应用。

<sup>2</sup>S. H. Crandall, *An optimal implicit recurrence formula for the heat conduction equation*, Quarterly of Applied Mathematics, 13 (1955), 318-320

<sup>3</sup>J. Jr. Douglas, *The solution of the diffusion equation by a high order correct difference equation*, J. Mathematics and Physics, 35 (1956), 145-151

‡ 论题 2.3. 讨论加权平均格式 (2.2) 的最大模稳定性。

答：仿照古典格式的论证过程，继续采用离散最大模原理，探讨加权平均格式的最大模稳定性。为此，将 (2.2) 改写为

$$\begin{aligned} (1 + 2\theta\mu a)u_j^{n+1} &= \theta\mu a \left[ u_{j-1}^{n+1} + u_{j+1}^{n+1} \right] \\ &+ \left[ 1 - 2(1 - \theta)\mu a \right] u_j^n + (1 - \theta)\mu a \left[ u_{j-1}^n + u_{j+1}^n \right]. \end{aligned} \quad (2.7)$$

若网比满足

$$\mu a(1 - \theta) \leq 1/2, \quad (2.8)$$

则差分方程 (2.7) 的右端系数都是非负的。设  $|u_{j_*}^{n+1}| = \|u^{n+1}\|_\infty$ ，有

$$\begin{aligned} (1 + 2\theta\mu a)\|u^{n+1}\|_\infty &= (1 + 2\theta\mu a)|u_{j_*}^{n+1}| \\ &\leq 2\theta\mu a\|u^{n+1}\|_\infty + \left[ |1 - 2(1 - \theta)\mu a| + 2|(1 - \theta)\mu a| \right] \|u^n\|_\infty \\ &\leq 2\theta\mu a\|u^{n+1}\|_\infty + \|u^n\|_\infty, \end{aligned}$$

即离散最大模原理成立。因此，当 (2.8) 成立时，有

$$\|u^{n+1}\|_\infty \leq \|u^n\|_\infty \leq \cdots \leq \|u^0\|_\infty, \quad \forall n,$$

加权平均格式具有最大模稳定性<sup>4</sup>。具体来讲，就是：只有全隐格式 ( $\theta = 1$ ) 是无条件的，其它格式都是有条件的。□

## 2.2 Du Fort-Frankel 格式

最简单的多层格式是三层格式。对于热传导方程 (2.1)，非常自然的想法是利用一阶中心差商离散时间导数，利用二阶中心差商离散空间导数，可得

---

<sup>4</sup>此时，离散最大模原理仅仅是格式最大模稳定的充分条件。事实上，时空约束条件 (2.8) 可以放宽到 (2.6)。但是，相应的稳定性结论要放宽到  $\|u^n\|_\infty \leq K\|u^0\|_\infty$ ，其中的界定常数要大于 1。已知的最佳结果是  $K = 23$ 。

Richardson (1910) 格式

$$u_j^{n+1} = u_j^{n-1} + 2\mu a \delta_x^2 u_j^n. \quad (2.9)$$

显然, 它是显式格式, 无条件具有 (2, 2) 阶局部截断误差. 参见其离散模版, 它也称为实心十字架格式.

**↓ 论题 2.4.** Richardson 格式 (2.9) 是无条件线性  $L^2$  模不稳定的. 换言之, 对于任意的网比  $\mu > 0$ , 它都不具有  $L^2$  模稳定性.

**答:** 令  $v^n = u^{n-1}$ , 定义向量型函数  $\mathbf{w}^n = (u^n, v^n)^\top$ , 将标量型三层 Richardson 格式 (2.9) 改写为向量型双层格式

$$\mathbf{w}_j^{n+1} = \begin{bmatrix} 2\mu a & 0 \\ 0 & 0 \end{bmatrix} (\mathbf{w}_{j-1}^n + \mathbf{w}_{j+1}^n) + \begin{bmatrix} -4\mu a & 1 \\ 1 & 0 \end{bmatrix} \mathbf{w}_j^n. \quad (2.10)$$

代入模态解, 简单计算可得增长矩阵

$$\mathbb{G}(k; \Delta t) = \begin{bmatrix} -8\mu a \sin^2(\frac{1}{2}k\Delta x) & 1 \\ 1 & 0 \end{bmatrix},$$

相应的两个特征值为

$$\lambda_{1,2} = -4\mu a \sin^2\left(\frac{1}{2}k\Delta x\right) \pm \sqrt{1 + 16\mu^2 a^2 \sin^4\left(\frac{1}{2}k\Delta x\right)}.$$

显然, 存在波数  $k$ , 使得  $\sin^2(\frac{1}{2}k\Delta x) > \frac{1}{2}$ . 因此, 有

$$\max(|\lambda_1|, |\lambda_2|) > \mu a + \sqrt{1 + \mu^2 a^2} > 1 + \mu a.$$

换言之, von Neumann 条件不成立, Richardson 格式在  $L^2$  模度量下是不稳定的.  $\square$

虚化 Richardson 格式的中心点值  $u_j^n$ , 将其替换为相邻时刻网格点值的

算术平均值  $(u_j^{n+1} + u_j^{n-1})/2$ , 即得著名的 Du Fort-Frankel (DF) 格式<sup>5</sup>

$$u_j^{n+1} = u_j^{n-1} + 2\mu a(u_{j-1}^n - u_j^{n+1} - u_j^{n-1} + u_{j+1}^n). \quad (2.11)$$

注意到离散模版的形状, 它也称作空心十字架格式。

**¶ 论题 2.5.** 讨论 DF 格式 (2.11) 的局部截断误差。

答: DF 格式是 Richardson 格式的修正, 即

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} = a \frac{\delta_x^2 u_j^n}{(\Delta x)^2} - \frac{a(\Delta t)^2}{(\Delta x)^2} \frac{\delta_t^2 u_j^n}{(\Delta t)^2}. \quad (2.12)$$

利用 Richardson 格式的相容性结果可知, DF 格式的局部截断误差是

$$\tau_j^n = \mathcal{O}\left((\Delta x)^2 + (\Delta t)^2 + \frac{(\Delta t)^2}{(\Delta x)^2}\right).$$

当  $\Delta t/(\Delta x)^2$  固定时, 局部截断误差是  $\mathcal{O}((\Delta x)^2)$ ; 当  $\Delta t/\Delta x$  固定时, 局部截断误差是  $\mathcal{O}(1)$  的。换言之, 相容性结论依赖于具体的加密路径, 即 DF 格式是有条件相容的。□

观察 (2.12) 的局部截断误差推导过程, 可知: 同热传导方程 (2.1) 相比, DF 格式更加靠近一个含有网格参数的偏微分方程, 即

$$u_t = au_{xx} - \mu a \Delta t u_{tt}.$$

对于给定的网格参数  $\Delta x$  和  $\Delta t$ , 它属于电报方程, 具有更加健壮的适定性表现, 故而我们可以大胆猜测 DF 格式的稳定性表现强于 Richardson 格式。上述论证过程已经展现出修正方程方法的基本思想; 详细内容见 §A。

**¶ 论题 2.6.** DF 格式 (2.11) 无条件具有  $L^2$  模稳定性。

---

<sup>5</sup>E. C. Du Fort and S. P. Frankel, *Stability conditions in the numerical treatment of parabolic differential equations*, Math. Tables and other Aids to Computation, 7 (1953), 135-152

答: 令  $v_j^n = u_j^{n-1}$ , 定义向量型网格函数  $\mathbf{w}_j^n = (u_j^n, v_j^n)^\top$ , 将 DF 格式改写为向量型双层格式

$$\begin{bmatrix} 1+2\mu a & 0 \\ 0 & 1 \end{bmatrix} \mathbf{w}_j^{n+1} = \begin{bmatrix} 2\mu a & 0 \\ 0 & 0 \end{bmatrix} (\mathbf{w}_{j-1}^n + \mathbf{w}_{j+1}^n) + \begin{bmatrix} 0 & 1-2\mu a \\ 1 & 0 \end{bmatrix} \mathbf{w}_j^n.$$

简单计算可得增长矩阵

$$\mathbb{G}(k, \Delta t) = \frac{1}{1+2\mu a} \begin{bmatrix} 4\mu a \cos k \Delta x & 1-2\mu a \\ 1+2\mu a & 0 \end{bmatrix}$$

和特征方程

$$\lambda^2 - \frac{4\mu a \cos k \Delta x}{1+2\mu a} \lambda - \frac{1-2\mu a}{1+2\mu a} = 0. \quad (2.13)$$


此时, 利用特征值的具体表达式, 或者特征方程的系数结构<sup>6</sup>, 都可以建立两个特征值按模均不超过 1 的充要条件。由于特征方程 (2.13) 的系数满足

$$\left| \frac{4\mu a \cos k \Delta x}{1+2\mu a} \right| \leq 1 - \frac{1-2\mu a}{1+2\mu a} < 2,$$

故而严格 von Neumann 条件无条件成立。利用韦达定理, 由 (2.13) 可知

$$|\lambda_1| |\lambda_2| = \left| \frac{1-2\mu a}{1+2\mu a} \right| < 1,$$

换言之, Kreiss 定理<sup>7</sup> 成立, 严格 von Neumann 条件是  $L^2$  模稳定的充要条件。因此, DF 格式无条件具有  $L^2$  模稳定性。□

 **注释 2.1.** 同 Richardson 格式相比, DF 格式展示了算术平均的稳定化作用。这个过程也说明了数值方法研究的特色: 数值格式的微弱变化可能造成数值表现的明显差异。

<sup>6</sup>对于实系数二次方程  $x^2 + bx + c = 0$ , 两个根按模均不超过一的充要条件是  $|b| \leq 1 + c \leq 2$ 。对于复系数二次方程, 相应的充要条件较繁; 可参阅 [2] 的引理 5.2。

<sup>7</sup>??



利用 Lax-Richtmyer 等价定理可知, 当网比  $\mu$  固定时, DF 格式具有整体二阶误差。因此说, DF 格式是可用的三层格式。

三层格式却需要两个初值, 而定解问题仅提供一个初值。通常, 第零层的初值设置是容易的, 譬如  $u_j^0 = u_0(x_j)$ 。但是, 第一层的初值设置需要借用其它方法, 例如:

1. 假设真解充分光滑, 偏微分方程在初始时刻也成立。利用时间方向的 Taylor 公式, 将初始时刻的时间导数转化为相应的空间信息, 定义

$$u_j^1 = u_0(x_j) + a\Delta t u_0''(x_j). \quad (2.14)$$

若二阶导数采用中心差商进行离散, 则它恰好就是热传导方程 (2.1) 的全显格式  $u_j^1 = u_j^0 + \mu a \delta_x^2 u_j^0$ 。

2. 利用双层格式, 数值计算出第一层的初值。事实上, 若三层格式具有时间方向的二阶局部截断误差, 则启动格式在时间方向达到一阶相容性即可。因此, 对于 DF 格式, CN、全隐或者加权平均格式都是可行的选择。

由于使用次数有限, 启动格式无需满足时空约束条件, 例如全显格式的网比可以满足  $\mu a > 1/2$ 。上述启动策略具有普适性, 适用于任意层数的格式。

## 2.3 跳点格式

数值计算可以同时使用多个格式。本节以两个古典格式为例, 介绍它们的杂交算法。换言之, 将时空网格点按照奇偶属性分为两组, 其中一组使用全显格式, 另一组使用全隐格式。依据不同分类方式, 具体操作模式有三种。

若网格点的时空指标之和是奇(偶)数, 则称其是奇(偶)数点。若奇数点采用全显格式, 偶数点采用全隐格式, 可得跳点(hopscotch)格式<sup>8</sup>

$$u_j^{n+1} = u_j^n + \mu a \delta_x^2 u_j^n, \quad \text{若 } n+j = \text{奇数}, \quad (2.15a)$$

$$u_j^{n+1} = u_j^n + \mu a \delta_x^2 u_j^{n+1}, \quad \text{若 } n+j = \text{偶数}. \quad (2.15b)$$

---

<sup>8</sup>首先由 Gordon(1965) 提出, 而后被 Gourlay(1970) 命名为跳点格式。

显然, 它保持古典格式的局部截断误差阶。它的计算过程也是显式的: 先利用 (2.15a) 计算后续时刻的偶数点值, 再利用 (2.15b) 计算后续时刻的奇数点值。因此说, 跳点格式 (2.15) 也是半隐的。

**↓ 论题 2.7.** 跳点格式 (2.15) 等同于偶数点集上的 DF 格式。因此, 跳点格式无条件具有  $L^2$  模稳定性。

答: 当  $n+j$  为偶数时, 由 (2.15a) 可得差分方程

$$u_j^{n+2} = u_j^{n+1} + \mu a \delta_x^2 u_j^{n+1}. \quad (2.16)$$

将 (2.16) 同 (2.15b) 相加或者相减, 有

$$u_j^{n+2} - u_j^n = 2\mu a \delta_x^2 u_j^{n+1}, \quad (2.17a)$$

和

$$u_j^{n+2} = 2u_j^{n+1} - u_j^n. \quad (2.17b)$$

当  $n+j$  是偶数的时候, 两式联立, 消去 (2.17a) 中的奇数点值  $u_j^{n+1}$ , 可得

$$u_j^{n+2} = u_j^n + 2\mu a [u_{j-1}^{n+1} - u_j^n - u_j^{n+2} + u_{j+1}^n].$$

它构成偶数点集上的 DF 格式。当网比固定时, 跳点格式具有二阶相容性。□

在数值计算时, 我们可以不必保留奇数点值, 直接在偶数点集上执行 DF 格式即可。虽然跳点格式 (2.15) 同 DF 格式具有偶数集上的等价性, 但是它成功回避了多层格式的原值启动困难。

## 2.4 数值格式的健壮性

明确数值误差的真正来源和具体表现, 保证数值结果的可靠性, 是误差估计或收敛分析的主要工作。作为重要的理论研究, 它可以指明数值格式的高效性和健壮性。具体分析路线同真解的光滑程度相关。

当 Lax-Richtmyer 等价定理所需的强正则性假设无法满足的时候, 差分格式还能给出可靠的数值结果吗? 换言之, 数值格式的健壮性成为它能否成功应用的关键。

要回答这个问题, 不妨举例说明。考虑热传导方程 (2.1) 的周期边值问题, 相应的初值是间断函数

$$u_0(x) = u_0^{(1)}(x) = \begin{cases} 1, & x \in [-\frac{\pi}{2}, \frac{\pi}{2}], \\ 0, & x \in [-\pi, -\frac{\pi}{2}) \cup (\frac{\pi}{2}, \pi]; \end{cases} \quad (2.18a)$$

或者导数间断的连续函数

$$u_0(x) = u_0^{(2)}(x) = \pi - |x|, \quad x \in [-\pi, \pi]. \quad (2.18b)$$

由于初值达不到二阶连续可微的光滑度, 真解在初始时刻附近的导数不再有界。为行文简便, 统称为模型问题 (HP-WEAK)。

为简单起见, 设扩散系数为  $a = 1$ , 终止时刻为  $T = 1$ 。利用全显格式模拟模型问题 (HP-WEAK)。给定正整数  $J$ , 定义等距空间网格

$$\mathcal{T}_J = \{x_j = j\pi/J\}_{j=-J}^J, \quad J = 18, 36, 72, \dots,$$

相应的初值是  $u_j^0 = u_0(x_j)$ 。在网格加密的过程中, 网比  $\mu = 0.4$  保持不变。换言之, 若空间网格是  $\mathcal{T}_J$ , 则相应的时间步长是

$$(\Delta t)_J = \mu(\Delta x)_J^2 = \mu\left(\frac{\pi}{J}\right)^2.$$

当然, 最后的时间步长需要调整, 使得  $t^{N_J} = T$ , 让最后一个时间层恰好就是给定的终止时刻。在表 2.1 中, 我们给出了全显格式在终止时刻的  $L^2$  模误差

$$\mathcal{E}_J = \|e(T)\|_{2,\pi/J} = \left( \sum_{j=-J}^J \left[ u_j^{N_J} - [u]_j^{N_J} \right]^2 \Delta x \right)^{\frac{1}{2}},$$

和相应的  $L^2$  模误差阶

$$o_J = \frac{1}{\ln 2} \left[ \ln \mathcal{E}_{J/2} - \ln \mathcal{E}_J \right].$$

表格中的数据清楚说明数值解依旧收敛到真解，甚至还呈现出一阶或二阶的精度<sup>9</sup>。换言之，全显格式具有满意的健壮性。

	$u_0 = u_0^{(1)}$		$u_0 = u_0^{(2)}$	
$J$	误差	误差阶	误差	误差阶
18	6.970e-2		8.557e-4	
36	3.483e-2	1.00	2.110e-4	2.02
72	1.741e-2	1.00	5.273e-5	2.00
144	8.707e-3	1.00	1.317e-5	2.00
288	4.353e-3	1.00	3.293e-6	2.00

表 2.1: 全显格式的  $L^2$  模误差和误差阶

这个数值现象的理论证明较为困难，主要原因是真解的高阶导数（例如时间二阶导数或空间四阶导数）在  $[-\pi, \pi] \times [0, T]$  上无界，全显格式的局部截断误差缺乏清晰明瞭的整体控制。即使采用宽松的离散  $L^2$  范数作为度量，局部截断误差也没有明确阶数。

**论题 2.8.** 假设初值函数  $u_0(x)$  平方可积且有界。当  $\mu a \leq 1/2$  时，模型问题 (HP-WEAK) 的全显格式是收敛的。

答：借助于分离变量法，问题的真解可以准确地表示为

$$[u]_j^n = \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{+\infty} a_k e^{ikj\Delta x} [e^{-ak^2\Delta t}]^n, \quad (2.19a)$$

其中  $e^{-ak^2\Delta t}$  是真实增长因子，按模不超过一；类似地，借助于 Fourier 方法（或者分离变量方法），全显格式的数值解也可以精确地表示为

$$u_j^n = \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{+\infty} A_k e^{ikj\Delta x} [\lambda(k)]^n, \quad (2.19b)$$

<sup>9</sup>补充定义 (2.18a) 在间断点的取值为 0.5，全显格式依旧呈现出二阶精度。

其中  $\lambda(k) = 1 - 4\mu a \sin^2(\frac{1}{2}k\Delta x)$  是数值增长因子。当  $\mu a \leq 1/2$  时, 它按模也不超过一。在 (2.19) 中, 相应的展开系数是

$$a_k = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{-ikx} u_0(x) dx, \quad (2.20a)$$

$$A_k = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{-ikx} \tilde{u}_0(x) dx, \quad (2.20b)$$

其中  $\tilde{u}_0(x): [-\pi, \pi] \rightarrow \mathfrak{R}$  是初值  $\{u_j^0 = u_0(x_j)\}_{j=-J}^J$  逐点常值延拓而成的周期阶梯函数。

设  $\varepsilon$  是任意给定的正数。选取适当的正整数  $M \leq J$ , 将数值误差分裂为三部分, 即

$$e_j^n = [u]_j^n - u_j^n = \Pi_j^n + \Theta_j^n + \Upsilon_j^n, \quad (2.21)$$

其中

$$\Pi_j^n = \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{+\infty} (a_k - A_k) e^{ikj\Delta x} [\lambda(k)]^n, \quad (2.22a)$$

$$\Theta_j^n = \frac{1}{\sqrt{2\pi}} \sum_{|k| > M} a_k e^{ikj\Delta x} \left[ e^{-ak^2 n \Delta t} - (\lambda(k))^n \right], \quad (2.22b)$$

$$\Upsilon_j^n = \frac{1}{\sqrt{2\pi}} \sum_{|k| \leq M} a_k e^{ikj\Delta x} \left[ e^{-ak^2 n \Delta t} - (\lambda(k))^n \right]. \quad (2.22c)$$

每个部分对应一个网格函数。下面证明: 当网格充分密集 (等价于  $J$  充分大) 的时候, 上面三个网格函数的  $L^2$  模均小于  $\varepsilon$ 。

利用 Fourier 级数理论, 前两个网格函数的讨论是容易的。由初值设置方式<sup>10</sup>可知, 当网格充分密集时, 有

$$\|u_0 - \tilde{u}_0(x)\|_{L^2[-\pi, \pi]} < \varepsilon.$$

注意到数值增长因子的模不超过 1, 利用 Parseval 恒等式可得

$$\|\Pi^n\|_{2, \Delta x} \leq \left[ \sum_{k=-\infty}^{\infty} |a_k - A_k|^2 \right]^{\frac{1}{2}} = \|u_0 - \tilde{u}_0(x)\|_{L^2[-\pi, \pi]} < \varepsilon, \quad (2.23)$$

---

<sup>10</sup> 选取合适的空间网格, 可以使  $\tilde{u}_0^{(1)}(x) = u_0^{(1)}(x)$ ; 参见前注。

其中  $\Pi^n = \{\Pi_j^n\}_{\forall j}$ 。类似地, 注意到  $u_0(x)$  是平方可积的速降函数, 可知: 当  $M$  充分大 (蕴含网格充分密集) 时, 也有

$$\|\Theta^n\|_{2,\Delta x} \leq 2 \left[ \sum_{|k|>M} a_k^2 \right]^{1/2} < \varepsilon, \quad (2.24)$$

其中  $\Theta^n = \{\Theta_j^n\}_{\forall j}$ 。当然,  $J$  和  $M$  的取值都依赖  $u_0(x)$  的光滑程度。

网格函数  $\Upsilon^n = \{\Upsilon_j^n\}_{\forall j}$  反映数值格式局部误差<sup>11</sup>的累积。利用 Taylor 展开公式可知, 真实增长因子和数值增长因子的差距满足

$$|\lambda(k) - e^{-ak^2\Delta t}| \leq Ck^4(\Delta t)^2, \quad \forall |k| \leq M, \quad (2.25)$$

其中界定常数  $C = C(\mu a; k)$  同网格参数无关。因此说, 全显格式具有  $\mathcal{O}(\Delta t)^2$  的局部误差。利用  $n\Delta t \leq T$  和简单的不等式

$$|a^n - b^n| \leq n|a - b|, \quad \text{若 } |a| \leq 1, |b| \leq 1,$$

可知

$$\begin{aligned} \|\Upsilon^n\|_{2,\Delta x} &\leq \left[ \sum_{|k| \leq M} |a_k|^2 \left| [\lambda(k)]^n - e^{-ak^2n\Delta t} \right|^2 \right]^{1/2} \\ &\leq \left[ \sum_{|k| \leq M} |a_k|^2 k^8 \right]^{1/2} CT\Delta t. \end{aligned} \quad (2.26)$$

无论  $u_0(x)$  是否具有四阶导数, 亦或  $\|D_x^4 u_0\|_{L^2[-\pi,\pi]}$  是否有限, 均可断言  $\sum_{|k| \leq M} |a_k|^2 k^8 < +\infty$ 。因此, 只要时间步长  $\Delta t$  充分小, 就有

$$\|\Upsilon^n\|_{2,\Delta x} < \varepsilon. \quad (2.27)$$

网比  $\mu$  固定时,  $\Delta t$  充分小等价于  $\Delta x$  充分小, 或者说离散网格充分密集。

---

<sup>11</sup>假设当前时刻的数值计算是精确的, 即数值解就是真解。当数值格式推进一个时间步长之后, 相应的数值误差称为局部误差。

综上所述, 利用 (2.21) 和三角不等式, 即可证明命题结论。□

(2.21) 清楚地指明了数值误差的两个主要影响因素: 初值误差以及局部误差。前者同真解光滑度密切相关, 而后者同数值增长因子和真实增长因子的差距有关。在两者之中, 局部误差更为重要, 而初值误差常被忽略。

## 2.5 线性扩散方程

依据不同的物理近似过程, 非均匀介质内的热传导现象可以描述为两种形式的线性扩散方程。其一是非守恒型扩散方程

$$u_t = a(x, t)u_{xx}, \quad (2.28a)$$

其二是守恒型 (或散度型) 扩散方程

$$u_t = (a(x, t)u_x)_x, \quad (2.28b)$$

其中  $a(x, t)$  称为扩散系数, 在计算区域上具有正的下确界。当扩散系数变化缓慢的时候, 上述两种表达形式是非常接近的。

暂且假设  $a(x, t)$  和  $u(x, t)$  均足够光滑, 扩散方程 (2.28) 的差分格式的真解存在唯一且充分光滑。

### 2.5.1 冻结系数方法

在离散焦点直接冻结扩散系数, 可得 (2.28a) 的全显格式

$$\Delta_t u_j^n = \mu a_j^n \delta_x^2 u_j^n \quad (2.29a)$$

和全隐格式

$$\Delta_t u_j^n = \mu a_j^{n+1} \delta_x^2 u_j^{n+1}. \quad (2.29b)$$

利用 Taylor 展开技术可知, 它们均无条件具有 (2, 1) 阶局部截断误差, 保持了它们在定常情形 (线性常系数问题) 的相容阶。

加权平均格式是全显格式和全隐格式的线性组合，相应的扩散系数有两种冻结策略。其一是**多焦点策略**，定义

$$\Delta_t u_j^n = \mu \left[ \theta a_j^{n+1} \delta_x^2 u_j^{n+1} + (1 - \theta) a_j^n \delta_x^2 u_j^n \right], \quad (2.30a)$$

其中  $\theta \in [0, 1]$  是给定的权重。一般而言，它无条件具有  $(2, 1)$  阶局部截断误差；特别地，当  $\theta = 1/2$  时，它称为 Crank-Nicolson 格式，无条件具有  $(2, 2)$  阶局部截断误差。换言之，(2.30a) 也保持了其在定常情形的相容阶。其二是**单焦点策略**，定义

$$\Delta_t u_j^n = \mu a_j^* \left[ \theta \delta_x^2 u_j^{n+1} + (1 - \theta) \delta_x^2 u_j^n \right], \quad (2.30b)$$

其中  $a_j^*$  是扩散系数的局部冻结：

1. 当  $\theta = 1/2$  时，令  $a_j^* = a(x_j, t^*)$ ，其中  $t^* \in [t^n, t^{n+1}]$ 。此时，差分方程 (2.30b) 无条件具有  $(2, 1)$  阶局部截断误差，保持了它在定常情形的相容阶。
2. 当  $\theta = 1/2$  时，扩散系数的局部冻结要细心设置，才能保持定常情形的局部相容阶。借鉴 (2.30a) 的双焦点策略，可以定义

$$a_j^* = \frac{1}{2}(a_j^n + a_j^{n+1}). \quad (2.31a)$$

回忆线性常数 CN 格式的构造过程，扩散系数还可以直接冻结在最佳的离散焦点上，即定义

$$a_j^* = a_j^{\frac{n+1}{2}} \equiv a(x_j, (t^n + t^{n+1})/2). \quad (2.31b)$$

通常，基于 (2.31a) 和 (2.31b) 两种冻结方式的数值格式 (2.30b) 均称为 **Crank-Nicolson 格式**。事实上，前者基于时间积分的梯形公式，而后者基于时间积分的中点矩形公式。

由于  $a(x, t)$  足够光滑，上述两种冻结方式具有  $\mathcal{O}((\Delta t)^2)$  的差距，相应的 CN 格式都具有  $(2, 2)$  阶局部截断误差。但是，它们关于扩散系数的最低光滑性要求是不同的。请读者自行推导。



随着相容阶的增高，扩散系数的局部冻结技术也将变得繁琐。

**┆ 论题 2.9.** 为简单起见，设  $a(x, t) \equiv a(x)$ ，即扩散系数同时间无关。基于加权平均格式的离散模版，构造扩散方程 (2.28a) 的整体四阶格式。

答：回顾 (2.28a) 的 CN 格式，由构造过程可知

$$\begin{aligned} & \frac{[u]_j^{n+1} - [u]_j^n}{a_j \Delta t} - \frac{1}{2(\Delta x)^2} \left[ \delta_x^2 [u]_j^n + \delta_x^2 [u]_j^{n+1} \right] \\ &= -\frac{(\Delta x)^2}{12} [u_{xxxx}]_j^{n+\frac{1}{2}} + \mathcal{O}((\Delta x)^4 + (\Delta x \Delta t)^2 + (\Delta t)^2). \end{aligned} \quad (2.32)$$

要构造出整体四阶的差分格式，只需建立  $[u_{xxxx}]_j^{n+\frac{1}{2}}$  的二阶相容离散。首先，利用偏微分方程 (2.28a)，将空间导数转化为时间导数。然后，利用相应的中心差商离散，可以建立

$$\begin{aligned} [u_{xxxx}]_j^{n+\frac{1}{2}} &= [(a^{-1}u_t)_{xx}]_j^{n+\frac{1}{2}} = [(a^{-1}u)_{xxt}]_j^{n+\frac{1}{2}} \\ &= \frac{\delta_x^2 [a^{-1}u]_j^{n+1} - \delta_x^2 [a^{-1}u]_j^n}{(\Delta x)^2 \Delta t} + \mathcal{O}((\Delta x)^2 + (\Delta t)^2). \end{aligned} \quad (2.33)$$

两式联立，略去无穷小量，用数值解替换真解，可得整体四阶的差分方程

$$\frac{\Delta_t u_{j+1}^n}{12a_{j+1}} + \frac{5\Delta_t u_j^n}{6a_j} + \frac{\Delta_t u_{j-1}^n}{12a_{j-1}} = \frac{1}{2}\mu(\delta_x^2 u_j^{n+1} + \delta_x^2 u_j^n). \quad (2.34)$$

若  $a(x)$  是常值函数，它就是 **Douglas 格式**。 □

**┆ 注释 2.2.** 在 Douglas 格式 (2.34) 的设计过程中，有两个关键技术非常值得回味。首先，以偏微分方程为桥梁，将时空方向的导数进行恰当的转换，克服了时空信息分布不均的困难，使得离散模版的空间网格点分布更加紧凑。其次，填补局部截断误差主项的差商离散，将低阶格式修正到高阶格式。上述两种设计思想简单有效，应用范围极其广泛。

### 2.5.2 积分插值方法

积分插值方法是散度型导数的常用离散技术。其设计思想非常简单，就是离散对象在某个局部区域的积分近似：

1. 选取适当的局部区域, 积分偏微分方程或者散度型导数。基于散度定理, 高阶导数的高维积分可以转化为低阶导数的低维积分。要注意, 这个推导过程是精确的。

2. 离散低阶导数, 采用适当的数值积分公式, 近似低阶导数的低维积分。

由于积分维数和导数阶数均降低, 数值格式的设计过程变得相对简单。在近代文献中, 积分插值方法常常被收录到有限体积方法或广义差分方法。

**↓ 论题 2.10.** 利用积分插值方法, 建立守恒型扩散方程 (2.28b) 的全显格式。

答: 在局部区域  $(x_{j-1/2}, x_{j+1/2}) \times (t^n, t^{n+1})$  内, 考虑守恒型扩散方程 (2.28b) 的二维积分, 可得精确成立的积分恒等式

$$\begin{aligned} & \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t^{n+1}) dx - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t^n) dx \\ &= \int_{t^n}^{t^{n+1}} W(x_{j+\frac{1}{2}}, t) dt - \int_{t^n}^{t^{n+1}} W(x_{j-\frac{1}{2}}, t) dt, \end{aligned} \quad (2.35)$$

其中  $x_{j\pm 1/2} = x_j \pm \Delta x/2$  为控制区间  $I(x_j)$  的端点。左侧积分采用中点公式近似, 右侧积分采用左矩形公式近似, 有

$$([u]_j^{n+1} - [u]_j^n) \Delta x \approx ([W]_{j+\frac{1}{2}}^n - [W]_{j-\frac{1}{2}}^n) \Delta t. \quad (2.36)$$

利用一阶中心差商技术和冻结系数方法, 离散热流量, 有

$$[W]_{j+\frac{1}{2}}^n \approx a_{j+\frac{1}{2}}^n \frac{[u]_{j+1}^n - [u]_j^n}{\Delta x}, \quad (2.37)$$

其中  $a_{j+1/2}^n$  是扩散系数的局部冻结, 比如

$$a_{j+\frac{1}{2}}^n = a(x_{j+\frac{1}{2}}, t^n) \quad \text{或者} \quad a_{j+\frac{1}{2}}^n = \frac{1}{2}(a_j^n + a_{j+1}^n). \quad (2.38)$$

综上所述, 略去无穷小量, 用数值解替换精确解, 即得守恒型扩散方程 (2.28b) 的全显格式

$$\Delta_t u_j^n = \mu \delta_x (a_j^n \delta_x u_j^n) = \mu \Delta_{-,x} (a_{j+\frac{1}{2}}^n \Delta_{+,x} u_j^n). \quad (2.39)$$

利用 Taylor 展开技术可知, 它无条件具有  $(2, 1)$  阶局部截断误差。若  $a(x, t)$  恒等于常数  $a$ , 它就是线性常系数热传导方程 (2.1) 的全显格式 (1.22)。□

在 (2.38) 中, 第一种方式基于直接定义, 第二种方式基于算术平均。由于扩散系数  $a(x, t)$  足够光滑, 两种方式具有  $O((\Delta x)^2)$  的差距, 相应格式的数值结果差距甚微。为行文简便, 统称它们为算术平均方式。

### 2.5.3 稳定性分析方法

Lax-Richtmyer 等价定理依旧成立。换言之, 若线性变系数差分格式相容于某个适定的线性偏微分方程定解问题, 则它的稳定性和收敛性是彼此等价的。因此, 我们仍以相容性和稳定性概念为重点讨论对象。

对于线性变系数差分格式, 相容性分析依旧是容易的。基本工具仍是 Taylor 展开技术, 只不过推导过程变得繁琐而已。但是, 稳定性分析可能会遇到严重困难。比如, 简便快捷的 Fourier 方法仅仅适用于线性常系数差分格式, 不能直接应用于线性变系数差分格式。

#### 冻结系数稳定性分析方法

冻结系数稳定性分析方法堪称是应用最广的方法。它的想法非常朴素, 直接将线性变系数差分格式视为某个线性 (或者分片) 常系数差分格式的微小扰动。利用通常的数值观念——彼此“靠近”的差分格式具有“接近”的数值表现——诱导出“启发性”的稳定性结论: 若作为参考对象的线性常系数差分格式是 (不) 稳定的, 则线性变系数差分格式也是 (不) 稳定的。

分析过程如下: 首先, 将差分系数冻结为某个常数, 某个线性常系数差分格式; 然后, 利用其它的准确分析技术, 给出相应的稳定性结论; 最后, 综合所有合理的系数冻结方式, 确定稳定性结论的交集。这就是冻结系数方法给出的结果。


**论题 2.11.** 利用冻结系数方法, 给出全显格式 (2.29a) 的最大模稳定性条件和  $L^2$  模稳定性条件。

答: 将扩散系数  $a_j^n$  锁定为某个常数  $a$ , 令全显格式 (2.29a) 转化为线性常系数差分格式  $u_j^{n+1} = u_j^n + \mu a \delta_x^2 u_j^n$ 。利用已知的稳定性结果可知, 其最大模稳定性条件和  $L^2$  模稳定性条件都是  $\mu a \leq 1/2$ 。然而, 这个结论只能近似反映网格点  $(x_j, t^n)$  附近的情形。因此, 我们需要综合考虑所有的网格点。令  $a$  遍历  $\{a_j^n\}_{\forall j}^n$  的取值范围, 相应稳定性结论的交集

$$\max_{\forall x \forall t} a(x, t) \mu \leq \frac{1}{2} \quad (2.40)$$

就是全显格式 (2.29a) 稳定性条件。  $\square$

利用离散最大模原理可知, 时空约束条件 (2.40) 保证全显格式 (2.29a) 的最大模稳定性。但是, 在其临界状态下, 全显格式的  $L^2$  模稳定性无法严格论证; 参见论题 2.12。尽管如此, 稳定性结论 (2.40) 还是相对准确的。换言之, 对于线性变系数差分格式, 冻结系数方法给出的结论通常都具有足够的指导价值, 其时空约束条件可以被“模糊”地看作充要条件<sup>12</sup>。为减少数值模拟不稳定的风险, 时空约束条件的上界常常被缩至原来的 60% ~ 80%。

 **注释 2.3.** 冻结系数方法简单便捷, 能较好地反映稳定性结果, 却完全忽略了系数变化带来的数值影响。对于线性常系数问题稳定的某些格式, 有可能对于某些线性变系数问题出现“数值共振”现象, 使得部分简谐波呈现出无法控制的增长。这种现象称为**线性变系数不稳定现象**。它特别容易出现在双曲型方程的无耗散格式中。

## 能量方法

其推导过程类似于偏微分方程的能量方法, 通常包含如下三个步骤:

1. 选取适当的检验函数, 建立**能量范数的递推关系式**;
2. 指出**能量范数同离散  $L^2$  模**的等价关系;

<sup>12</sup>事实上, 它仅是格式稳定的必要条件。换言之, 若数值格式对于线性常系数问题都是不稳定的, 它必然没有应用价值。

3. 导出差分格式关于  $L^2$  模稳定性, 给出相应的充分条件。

Fourier 方法在频域空间进行操作, 而能量方法直接在时域空间进行操作, 应用范围更加广泛, 可以处理线性变系数问题、非周期边界条件以及非等距网格等等。

**¶ 论题 2.12.** 利用能量方法, 给出全显格式 (2.39) 具有  $L^2$  模稳定性的充分条件。

答: 为简单起见, 设  $a_{j+1/2}^n \equiv a_{j+1/2}$ 。当扩散系数同时间相关时, 相应的能量方法是类似的; 留作练习。

在差分方程 (2.39) 的两端同乘  $u_j^{n+1} + u_j^n$ , 其中  $j = 0 : J-1$ 。将  $J$  个恒等式相加, 可得

$$\begin{aligned} \text{LHS} &\equiv \sum_{j=0}^{J-1} (u_j^{n+1} - u_j^n)(u_j^{n+1} + u_j^n) \Delta x \\ &= \mu \sum_{j=0}^{J-1} \Delta_{-,x} (a_{j+\frac{1}{2}} \Delta_{+,x} u_j^n) (u_j^{n+1} + u_j^n) \Delta x \equiv \text{RHS}. \end{aligned} \quad (2.41)$$

下面估计 (2.41) 的两端。显然, 有

$$\text{LHS} = \sum_{j=0}^{J-1} (u_j^{n+1})^2 \Delta x - \sum_{j=0}^{J-1} (u_j^n)^2 \Delta x. \quad (2.42a)$$

注意到周期边界条件  $u_0^n = u_J^n$  和  $u_{J+1}^n = u_1^n$ , 以及扩散系数  $a(x)$  的空间周期性, 调整求和次序, 可得

$$\text{RHS} = -\mu \sum_{j=0}^{J-1} a_{j+\frac{1}{2}} \Delta_{+,x} u_j^n \Delta_{+,x} (u_j^{n+1} + u_j^n) \Delta x. \quad (2.42b)$$

注意到  $p(p+q) = \frac{1}{2}(p+q)^2 + \frac{1}{2}[p^2 - q^2]$ , 有

$$\mathcal{E}(u^{n+1}) - \mathcal{E}(u^n) = -\frac{1}{2}\mu \sum_{j=0}^{J-1} a_{j+\frac{1}{2}} \left[ \Delta_{+,x} (u_j^{n+1} + u_j^n) \right]^2 \Delta x \leq 0,$$

其中

$$\mathcal{E}(u^n) \equiv \sum_{j=1}^{J-1} (u_j^n)^2 \Delta x - \frac{1}{2} \mu \sum_{j=0}^{J-1} a_{j+\frac{1}{2}} (\Delta_{+,x} u_j^n)^2 \Delta x$$

是整个离散系统的能量范数<sup>13</sup>。换言之， $\mathcal{E}(u^n)$  是不增的。利用算术平均值不等式，可得

$$\left[1 - 2\mu A\right] \sum_{j=1}^{J-1} (u_j^n)^2 \Delta x \leq \mathcal{E}(u^n) \leq \cdots \leq \mathcal{E}(u^0) \leq \sum_{j=1}^{J-1} (u_j^0)^2 \Delta x,$$

其中  $A = \max_{x \in (0,1)} a(x) > 0$ 。若存在正常数  $\delta > 0$ ，使得

$$1 - 2\mu A \geq \delta, \quad (2.43)$$

则有  $L^2$  模稳定性，即

$$\sum_{j=1}^{J-1} (u_j^n)^2 \Delta x \leq \frac{1}{\delta} \sum_{j=1}^{J-1} (u_j^0)^2 \Delta x. \quad (2.44)$$

由 (2.43) 可知  $\delta < 1$ ，因此稳定性结论 (2.44) 弱于偏微分方程定解问题的适定性结论。□

在时空约束条件 (2.40) 的临界状态下，变系数全显格式 (2.39) 的  $L^2$  模稳定性结论是不明确的。但是，当扩散系数是常数的时候，利用直接矩阵方法可证：在临界状态下，全显格式也具有  $L^2$  模稳定性。在某种程度上，它也暗示冻结系数稳定性分析的风险。

## 2.5.4 具有间断系数的线性扩散方程

将两种材质焊接在一起，整个系统的导热现象仍可用守恒型扩散方程 (2.28b) 描述。此时，扩散系数在焊接点  $x_*$  出现第一类间断。为简单起见，设扩散系数是分片常数函数，即

$$a(x, t) = \begin{cases} a_L, & x < x_*, \\ a_R, & x > x_*, \end{cases} \quad (2.45)$$

<sup>13</sup>在适当的时空条件下，它才会真正成为离散范数。

其中  $a_L = a_R$ 。此时，真解不是（整体）古典解，而是满足联接条件

$$u(x_\star^+, t) = u(x_\star^-, t), \quad a_R u_x(x_\star^+, t) = a_L u_x(x_\star^-, t), \quad \forall t > 0 \quad (2.46)$$

的分片古典解<sup>14</sup>。它的直观含义是温度和热流量处处连续，无论扩散系数是否间断。

此时，全显格式 (2.39) 依旧可以给出收敛的数值结果，但误差表现并不理想。其根源是扩散系数的冻结方式。为此借用积分插值方法的设计思想，构造更加有效的冻结方式。考虑  $W(x, t^n)/a(x, t^n)$  在局部区间  $(x_j, x_{j+1})$  的积分近似。暂时假设  $a(x, t)$  连续。注意到  $W(x, t^n)$  的空间连续性，利用积分中值定理可得

$$[u]_{j+1}^n - [u]_j^n = \int_{x_j}^{x_{j+1}} \frac{W(x, t^n)}{a(x, t^n)} dx \approx [W]_{j+\frac{1}{2}}^n \int_{x_j}^{x_{j+1}} \frac{dx}{a(x, t^n)}.$$

换言之，扩散系数可以局部冻结为

$$a_{j+\frac{1}{2}}^n = \Delta x \left[ \int_{x_j}^{x_{j+1}} \frac{dx}{a(x, t^n)} \right]^{-1}. \quad (2.47)$$

事实上，它也适用于间断扩散系数。

(2.47) 可以用数值积分进行近似。例如，在间断点两侧分别采用左矩形公式和右矩形公式，将 (2.47) 近似为两侧扩散系数的（加权）调和平均，给出新的局部冻结方式

$$a_{j+\frac{1}{2}}^n = \left[ \frac{\theta_{j+\frac{1}{2}}^n}{a_j^n} + \frac{1 - \theta_{j+\frac{1}{2}}^n}{a_{j+1}^n} \right]^{-1}, \quad (2.48)$$

其中

$$\theta_{j+\frac{1}{2}}^n = \begin{cases} (x_\star - x_j)/\Delta x, & x_\star \in [x_j, x_{j+1}] \\ 1/2, & \text{其它.} \end{cases} \quad (2.49)$$

为行文简便，将 (2.47) 和 (2.48) 统称为调和平均方式。

---

<sup>14</sup>准确的定义应是“弱解”。

当扩散系数具有连续有界的二阶导数时, 调和平均方式和算术平均方式是非常接近的, 即

$$2 \left[ \frac{1}{a_j^n} + \frac{1}{a_{j+1}^n} \right]^{-1} - \frac{1}{2} [a_j^n + a_{j+1}^n] = \mathcal{O}((\Delta x)^2),$$

相应的全显格式 (2.39) 具有相同的收敛阶, 数值差别是微乎其微的。但是, 当存在第一类间断点时, 调和平均方式的数值优势将会得以显现。

**↓ 论题 2.13.** 利用直观的物理观点进行解释, 调和平均方式 (2.48) 更加准确地保持了热流通量在间断点两侧的连续性。

**答:** 设间断点  $x_*$  落在网格点  $x_j$  和  $x_{j+1}$  之间。若间断点两侧的扩散系数分别是  $a_j^n$  和  $a_{j+1}^n$ , 则左右两侧的热流通量可以分别近似为

$$[W]_L^n \approx a_j^n \frac{[u]_*^n - [u]_j^n}{\theta_{j+1/2}^n \Delta x}, \quad [W]_R^n \approx a_{j+1}^n \frac{[u]_{j+1}^n - [u]_*^n}{(1 - \theta_{j+1/2}^n) \Delta x},$$

其中  $[u]_*^n = u(x_*, t^n)$  是位于间断点的未知温度。上述过程相当于物理学中的均匀化技术。若间断点两侧的不同材质也被视为某种 (虚拟的) 均匀材质, 相应的扩散系数是待定的常数  $a_*^n$ , 则位于间断点  $x_*$  的热流通量可以近似为

$$[W]_*^n \approx a_*^n \frac{[u]_{j+1}^n - [u]_j^n}{\Delta x}.$$

基于物理观点, 上述三种刻画方式应当近似相等。因此, 有

$$\frac{[u]_{j+1}^n - [u]_j^n}{\Delta x / a_*^n} \approx \frac{[u]_*^n - [u]_j^n}{\theta_{j+1/2}^n \Delta x / a_j^n} \approx \frac{[u]_{j+1}^n - [u]_*^n}{(1 - \theta_{j+1/2}^n) \Delta x / a_{j+1}^n}.$$

将其看作等式关系, 解出的  $a_*^n$  就是调和平均扩散系数 (2.48)。□

**☞ 注释 2.4.** 在扩散系数的第一类间断点附近, 扩散系数的不同冻结方式可使差分方程 (2.39) 具有不同的相容阶。若基于算术平均方式 (2.38), 其局部截断误差是  $\mathcal{O}((\Delta x)^{-1})$ ; 若基于调和平均方式 (2.48), 其局部截断误差是  $\mathcal{O}(1)$ 。相应的推导过程较为繁琐, 因为间断点两侧的 Taylor 展开公式不同, 局部截断误差的化简要充分利用联接条件 (2.46)。具体内容可参阅 [4]。



## 2.6 非线性扩散方程

为简单起见, 本节以非线性热传导方程<sup>15</sup>

$$u_t = b(u)u_{xx} \quad (2.50)$$

的纯初值问题或周期边值问题为例, 其中扩散系数  $b(\cdot): \mathbb{R} \rightarrow \mathbb{R}^+$  具有正的下确界。因篇幅有限, 我们跳过适定性和正则性的深入讨论, 直接假设 (2.50) 具有足够光滑的唯一真解。

一般而言, 对于非线性扩散方程, 前面的各种数值离散技术依旧有效, 相应的格式设计是相对简单的。但是, 计算效率和理论分析将面临严峻的挑战。将系数冻结方法和差商离散技术相结合, 即可建立相应的全显格式、全隐格式和 Crank-Nicolson 格式

$$u_j^{n+1} = u_j^n + \mu b(u_j^n) \delta_x^2 u_j^n, \quad (2.51a)$$

$$u_j^{n+1} = u_j^n + \mu b(u_j^{n+1}) \delta_x^2 u_j^{n+1}, \quad (2.51b)$$

$$u_j^{n+1} = u_j^n + \frac{1}{2} \mu [b(u_j^n) \delta_x^2 u_j^n + b(u_j^{n+1}) \delta_x^2 u_j^{n+1}]. \quad (2.51c)$$

利用 Taylor 展开技术可知, 前两个格式具有  $(2, 1)$  阶局部截断误差, 最后一个具有  $(2, 2)$  阶局部截断误差。

平行于线性差分格式的 Lax-Richtmyer 等价定理, 非线性差分格式具有 **Strang 定理: 若非线性差分格式相容于某个适定的非线性问题, 则稳定性也是收敛性的充要条件**。相容性分析依旧是最容易的, 相应的 Taylor 展开将会变得更加繁琐。对于非线性差分格式, 稳定性分析将变得非常困难, 甚至在某种程度上会超过误差估计。尽管如此, 冻结系数方法依旧简单有效, 可以“启发式地”建立非线性差分格式的稳定性结论。

**↓ 论题 2.14.** 利用冻结系数方法, 建立差分格式 (2.51a) 和 (2.51c) 的  $L^2$  模稳定性结论。

---

<sup>15</sup>严格来说, 它是相对简单的半线性扩散问题。至于完全的非线性扩散问题, 相应的数值方法是非常困难的前沿课题, 本书不与讨论。

答：将系数  $b(u_j^n)$  看作其取值范围内的某个常数  $b$ ，非线性差分格式 (2.51a) 可以转化为线性常系数差分格式

$$u_j^{n+1} = u_j^n + \mu b \delta_x^2 u_j^n.$$

利用熟知的结果可知，其  $L^2$  模稳定的充要条件是  $\mu b \leq 1/2$ 。因此，差分格式 (2.51a) 的  $L^2$  模稳定性条件是

$$\mu \max_{j \forall n} b(u_j^n) \leq \frac{1}{2}.$$

类似地，可以断定差分格式 (2.51c) 无条件具有  $L^2$  模稳定性。 □

由冻结系数方法给出的时空约束条件，通常只是非线性差分格式数值稳定的必要条件而已。因此说，非线性问题的数值计算存在风险，数值格式的可靠性需要长期的实践验证和理论支持。

在数值实现方面，非线性差分格式可能遇到效率的困扰。例如，对于全隐格式 (2.51b) 和 CN 格式 (2.51c)，单步时间推进都将导致大规模的非线性方程组。即使采用高效的 Newton 方法，相应的求解过程也会耗费大量的 CPU 时间，整体的计算效率将大打折扣。事实上，即使每个非线性方程组都得到准确解，差分格式的数值结果依旧是定解问题的近似解而已。因此说，精确求解是没有必要的，相对合理的迭代近似即可。

**局部线性化**技术是常用的方法，即将非线性差分格式转换为一组线性差分格式<sup>16</sup>。以 CN 格式 (2.51c) 为例，常用的处理方式如下。

1. **时间延迟技术**用  $b(u_j^n)$  替换  $b(u_j^{n+1})$ ，相应的差分方程是

$$u_j^{n+1} = u_j^n + \frac{1}{2} \mu b(u_j^n) \delta_x^2 [u_j^n + u_j^{n+1}]. \quad (2.52)$$

可以证明：它具有 (2, 1) 阶局部截断误差。

---

<sup>16</sup>是指差分格式关于待解的网格函数是线性的

2. 预测校正方法连续执行两次线性化过程, 可得差分方程

$$\tilde{u}_j^{n+1} = u_j^n + \frac{1}{2}\mu \left[ b(u_j^n) \delta_x^2 u_j^n + b(u_j^n) \delta_x^2 \tilde{u}_j^{n+1} \right], \quad (2.53a)$$

$$u_j^{n+1} = u_j^n + \frac{1}{2}\mu \left[ b(u_j^n) \delta_x^2 u_j^n + b(\tilde{u}_j^{n+1}) \delta_x^2 u_j^{n+1} \right]. \quad (2.53b)$$

可以证明: 它具有 (2, 2) 阶局部截断误差。

此时, 某些多层格式具有计算效率的优势。例如, 利用已知时间层信息进行多项式外推, 给出扩散系数的高阶近似, 构造非线性热传导方程 (2.50) 的外推 CN 格式

$$u_j^{n+1} = u_j^n + \frac{1}{2}\mu b(u_j^{n+\frac{1}{2}}) \left[ \delta_x^2 u_j^{n+1} + \delta_x^2 u_j^n \right], \quad (2.54a)$$

其中

$$u_j^{n+\frac{1}{2}} = \frac{3}{2}u_j^n - \frac{1}{2}u_j^{n-1}. \quad (2.54b)$$

显然, 它关于网格函数  $u^{n+1}$  是线性的。利用 Taylor 展开技术可知, 它具有 (2, 2) 阶局部截断误差。利用冻结系数方法, 可证它无条件  $L^2$  模稳定。

## 2.7 高维扩散方程

前面介绍的数值离散技术和理论分析方法, 都可以顺利地推广到高维问题。为简单起见, 以二维线性常系数扩散方程

$$u_t = au_{xx} + bu_{yy} \quad (2.55)$$

为模型方程, 其中  $a$  和  $b$  是给定的正常数。随着空间维数的增高, 我们会遇到两个棘手的问题, 其一是计算效率的严重下降, 其二是边界条件的离散困难。

定义时空网格

$$\mathcal{T}_{\Delta x, \Delta y, \Delta t} = \mathcal{T}_{\Delta x, \Delta y} \otimes \mathcal{T}_{\Delta t} \equiv \{(x_j, y_k, t^n)\}_{\forall j, \forall k}^{\forall n}, \quad (2.56)$$

其中  $\Delta x$  和  $\Delta y$  分别是  $x$  方向和  $y$  方向的空间步长,  $\Delta t$  是时间步长。显然, 它是二维空间网格

$$\mathcal{T}_{\Delta x, \Delta y} = \mathcal{T}_{\Delta x} \otimes \mathcal{T}_{\Delta y} \equiv \{(x_j, y_k)\}_{j \forall k} \quad (2.57)$$

和时间网格  $\mathcal{T}_{\Delta t} = \{t^n\}_{n \forall}$  的笛卡尔乘积。通常,  $\mathcal{T}_{\Delta x, \Delta y}$  也是笛卡尔乘积型网格, 两族网格线分别平行于两个空间坐标轴。默认  $\mathcal{T}_{\Delta x, \Delta y, \Delta t}$  是等距的时空网格, 两个方向的网比记作

$$\mu_x = \Delta t / (\Delta x)^2, \quad \mu_y = \Delta t / (\Delta y)^2. \quad (2.58)$$

在网格点  $(x_j, y_k, t^n)$ , 真解用  $[u]_{jk}^n$  来表示, 数值解用  $u_{jk}^n$  来表示。换言之, 空间信息采用双下标标注方法。

**论题 2.15.** 构造二维扩散方程 (2.55) 的加权平均格式和 Du Fort-Frankel 格式。

**答: 逐维离散技术**也适用于高维扩散方程。换言之, 沿着各自的方向离散偏导数, 可得二维加权平均格式

$$\begin{aligned} u_{jk}^{n+1} = & u_{jk}^n + \theta \left[ \mu_x a \delta_x^2 u_{jk}^{n+1} + \mu_y b \delta_y^2 u_{jk}^{n+1} \right] \\ & + (1 - \theta) \left[ \mu_x a \delta_x^2 u_{jk}^n + \mu_y b \delta_y^2 u_{jk}^n \right], \end{aligned} \quad (2.59)$$

其中  $\theta \in [0, 1]$  是给定的权重。当  $\theta$  是 0, 1/2 和 1 时, 它依次称为二维全显格式、二维 Crank-Nicolson 格式和二维全隐格式。

类似地, 二维 Du Fort-Frankel 格式定义为

$$\begin{aligned} u_{j,k}^{n+1} = & u_{j,k}^{n-1} + \mu_x a \left[ u_{j-1,k}^n - u_{j,k}^{n-1} - u_{j,k}^{n+1} - u_{j+1,k}^n \right] \\ & + \mu_y b \left[ u_{j,k-1}^n - u_{j,k}^{n-1} - u_{j,k}^{n+1} - u_{j,k+1}^n \right]. \end{aligned} \quad (2.60)$$

换言之, 二维 Richardson 格式的中心点值被上下两个时间层的算术平均值所替代。□

相容性、稳定性和收敛性概念的基本含义同空间维数无关, 相应的 Lax-Richtmyer 等价定理依旧成立: **设线性偏微分方程的定解问题是适定的。若线性差分格式是相容的, 则稳定性和收敛性是彼此等价的, 且收敛阶不会低于相容阶。**对于高维差分格式, 我们依旧重点讨论相容性和稳定性概念, 跳过收敛性概念的严格论证。

事实上, 空间维数仅仅影响离散范数的具体定义而已。例如, 基于时空网格 (2.56), 二维 (空间) 网格函数  $u^n = \{u_{jk}^n\}_{\forall j \forall k}$  的 (二维) 最大模和  $L^2$  模分别定义为

$$\|u^n\|_\infty = \max_{\forall j \forall k} |u_{jk}^n|, \quad \|u^n\|_2 = \left( \sum_{\forall j} \sum_{\forall k} (u_{jk}^n)^2 \Delta x \Delta y \right)^{1/2}. \quad (2.61)$$

**↓ 论题 2.16.** 给出二维加权平均格式 (2.59) 的局部截断误差阶。

答: 将二维扩散方程 (2.55) 的真解  $[u]$  代入到差分方程 (2.59), 等号两侧的差距就是局部截断误差, 即

$$\begin{aligned} \tau_{jk}^n \equiv & \frac{[u]_{jk}^{n+1} - [u]_{jk}^n}{\Delta t} - \theta \left[ \frac{a}{(\Delta x)^2} \delta_x^2 [u]_{jk}^{n+1} + \frac{b}{(\Delta y)^2} \delta_y^2 [u]_{jk}^{n+1} \right] \\ & - (1 - \theta) \left[ \frac{a}{(\Delta x)^2} \delta_x^2 [u]_{jk}^n + \frac{b}{(\Delta y)^2} \delta_y^2 [u]_{jk}^n \right]. \end{aligned} \quad (2.62)$$

由于二维加权平均格式具有典型的逐维离散思想, 它的局部截断误差就是偏导数离散的局部截断误差叠加在一起。利用 Taylor 展开技术, 可知

$$\tau_{jk}^n = \mathcal{O}((\Delta x)^2 + (\Delta y)^2 + (2\theta - 1)\Delta t + (\Delta t)^2).$$

因此, 当  $\theta = 1/2$  时, 二维加权平均格式 (2.59) 具有  $(2, 2, 1)$  阶局部截断误差。当  $\theta = 1/2$  时, 它对应 CN 格式, 具有  $(2, 2, 2)$  阶局部截断误差。□

**↓ 论题 2.17.** 给出 (2.59) 最大模稳定的充分条件。

答: 既然离散对象具有最大模原理, 数值格式也应当满足离散最大模原理。仿照前面的讨论, 将 (2.59) 改写为

$$\left[ 1 + 2\theta(\mu_x a + \mu_y b) \right] u_{jk}^{n+1} = \cdots + \left[ 1 - 2(1 - \theta)(\mu_x a + \mu_y b) \right] u_{jk}^n,$$

其中省略部分的差分系数都是非负的。若时间步长满足

$$\mu_x a + \mu_y b \leq \frac{1}{2(1-\theta)}, \quad (2.63)$$

则等号右端的差分系数都是非负的，满足离散最大模原理。因此，数值解满足  $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$ ，相应的二维加权平均格式 (2.59) 具有最大模稳定性。□

**论题 2.18.** 给出二维加权平均格式 (2.59) 的  $L^2$  模稳定性。

答：设  $\ell_1$  和  $\ell_2$  是任意实数，对应两个空间的波数。将二维模态解

$$u_{jk}^n = \lambda^n e^{i(j\ell_1 \Delta x + k\ell_2 \Delta y)} \quad (2.64)$$

代入到二维加权平均格式 (2.59)，简单计算可得增长因子

$$\lambda = \lambda(\ell_1, \ell_2) = \frac{1 - 4(1-\theta)s}{1 + 4\theta s}, \quad (2.65)$$

其中  $s = \mu_x a \sin^2(\ell_1 \Delta x / 2) + \mu_y b \sin^2(\ell_2 \Delta y / 2)$ 。

标量双层格式  $L^2$  模稳定的充要条件<sup>17</sup>，依旧是著名的 von Neumann 条件，即

$$|\lambda(\ell_1, \ell_2)| \leq 1 + C\Delta t, \quad \forall \ell_1, \forall \ell_2, \quad (2.66)$$

其中界定常数  $C \geq 0$  同  $\Delta x, \Delta t, \ell_1, \ell_2$  和  $n$  均无关。若增长因子没有显式含有时间步长  $\Delta t$ ，则只需考虑严格的 von Neumann 条件

$$|\lambda(\ell_1, \ell_2)| \leq 1, \quad \forall \ell_1, \forall \ell_2. \quad (2.67)$$

注意到 (2.65)，二维加权平均格式  $L^2$  模稳定的充要条件是严格的 von Neumann 条件。由于  $s \in [0, \mu_x a + \mu_y b]$ ，相应的等价条件是

$$(1 - 2\theta)(\mu_x a + \mu_y b) \leq \frac{1}{2}. \quad (2.68)$$

<sup>17</sup>当然，Fourier 方法也可以推广到向量型格式或者多层格式。此时，增长因子变为增长矩阵。相应的 von Neumann 条件仅仅是格式按  $L^2$  模稳定的必要条件。必要条件能否成为充分条件，需要合适的 Kreiss 矩阵定理来验证。

换言之, 当  $\theta \geq 1/2$  时, 偏隐格式 (包括 CN 格式和全隐格式) 无条件  $L^2$  模稳定。当  $\theta < 1/2$  时, 偏显格式 (包括全显格式) 有条件  $L^2$  模稳定。  $\square$

对于高维扩散方程而言, 显式格式的时间推进受到严重限制, 在计算效率上缺乏竞争力。但是, 隐式格式的数值求解需要付出额外的代价, 相应的快速求解研究成为一个重要方向。

交替方向隐式 (Alternative Direction Implicit, 简称 ADI) 方法和局部一维化 (Local One Dimensional, 简称 LOD) 方法都属于分数步长方法, 可以实现快速求解的目标。它们的共同点是二维差分格式近似转化为一组“具有一维求解属性”的差分格式, 提升单步时间推进的效率。时至今日, 它们已经划归到算子分裂 (Operator Splitting) 方法的框架。

### ADI 格式

Peaceman 和 Rachford (PR) 格式是 Peaceman 和 Rachford 在模拟石油储藏模型时, 为解决二维 CN 格式的计算效率而最早提出的。在二维 CN 格式的等号右端添加修正项  $\frac{1}{4}\mu_x\mu_y a b \delta_x^2 \delta_y^2 (u^n - u^{n+1})$ , 差分方程不仅保持 (2, 2, 2) 阶局部截断误差, 而且具有漂亮的 (因式分解) 形式<sup>18</sup>

$$\begin{aligned} \left[ \mathbb{1} - \frac{1}{2}\mu_x a \delta_x^2 \right] \left[ \mathbb{1} - \frac{1}{2}\mu_y b \delta_y^2 \right] u^{n+1} \\ = \left[ \mathbb{1} + \frac{1}{2}\mu_x a \delta_x^2 \right] \left[ \mathbb{1} + \frac{1}{2}\mu_y b \delta_y^2 \right] u^n, \end{aligned} \quad (2.69)$$

其中  $\mathbb{1}$  是恒等算子。引进辅助网格函数  $u^{n+1/2}$ , 将 (2.69) 分裂为两步, 即得著名的 PR 格式<sup>19</sup>:

$$\left[ \mathbb{1} - \frac{1}{2}\mu_x a \delta_x^2 \right] u^{n+\frac{1}{2}} = \left[ \mathbb{1} + \frac{1}{2}\mu_y b \delta_y^2 \right] u^n, \quad (2.70a)$$

$$\left[ \mathbb{1} - \frac{1}{2}\mu_y b \delta_y^2 \right] u^{n+1} = \left[ \mathbb{1} + \frac{1}{2}\mu_x a \delta_x^2 \right] u^{n+\frac{1}{2}}. \quad (2.70b)$$

<sup>18</sup>略去网格函数的空间下标。

<sup>19</sup>D. W. Peaceman and H. H. Jr Rachford, *The numerical solution of parabolic and elliptic differential equations*, J. Soc. Indust. Appl. Math., 3 (1955), 28–41

由于时间上标出现分数，它也被称为分数步长方法。

$u^{n+1/2}$  仅仅是辅助函数而已，同中间时刻的真解  $[u]^{n+1/2}$  无直接关系。但是，若将  $u^{n+1/2}$  视为  $[u]^{n+1/2}$  的某种逼近，则 PR 格式的两步可以直观地解释为两个“半步时间”的推进过程：

1. (2.70a) 可视为  $t^n$  到  $t^{n+1/2} = t^n + \Delta t/2$  的计算过程，其中  $x$  方向采用全隐方式， $y$  方向采用全显方式。相应的时间推进距离是  $\Delta t/2$ 。
2. (2.70b) 可视为  $t^{n+1/2}$  到  $t^{n+1}$  的计算过程，其中  $x$  方向采用全显方式， $y$  方向采用全隐方式。相应的时间推进距离是  $\Delta t/2$ 。

事实上，ADI 方法的名程就源于此。

**┆ 论题 2.19.** 讨论 PR 格式 (2.70) 的  $L^2$  模稳定性。

答：设  $\ell_1$  和  $\ell_2$  是任意实数，分别对应两个方向的波数。将模态解

$$u_{j,k}^m = \hat{u}^m e^{i(\ell_1 j \Delta x + \ell_2 k \Delta y)}, \quad m = n, n+1/2,$$

代入到 PR 格式 (2.70) 或与其等价的 (2.69)，可得相应的增长因子

$$\lambda(\ell_1, \ell_2) = \frac{1 - 2\mu_y b \sin^2(\frac{1}{2}\ell_2 \Delta y)}{1 + 2\mu_x a \sin^2(\frac{1}{2}\ell_1 \Delta x)} \cdot \frac{1 - 2\mu_x a \sin^2(\frac{1}{2}\ell_1 \Delta x)}{1 + 2\mu_y b \sin^2(\frac{1}{2}\ell_2 \Delta y)}.$$

对于任意的网比  $\mu > 0$ ，严格的 von Neumann 条件都成立。因此，PR 格式无条件  $L^2$  模稳定。 □

在相容性和  $L^2$  模稳定性方面，二维 PR 格式 (2.70) 和二维 CN 格式难分伯仲。但是，在计算效率方面，二维 PR 格式 (2.70) 具有绝对优势，计算复杂度的下降程度达到开根号量级。

由于显著的高效性，PR 格式的设计思想引起数值工作者的浓厚兴趣。各式各样的 ADI 格式被相继提出，重要成果包括二维扩散方程 (2.55) 的 Douglas



格式<sup>20</sup>

$$u^{n+\frac{1}{2}} - u^n = \mu_x a \delta_x^2 \frac{u^{n+\frac{1}{2}} + u^n}{2} + \mu_y b \delta_y^2 u^n, \quad (2.71a)$$

$$u^{n+1} - u^n = \mu_x a \delta_x^2 \frac{u^{n+\frac{1}{2}} + u^n}{2} + \mu_y b \delta_y^2 \frac{u^n + u^{n+1}}{2}. \quad (2.71b)$$

显然, 前后两步分别对应逐行扫描和逐列扫描过程, 具有同二维 PR 格式一样的计算效率。

**论题 2.20.** 二维 Douglas 格式 (2.71) 同二维 PR 格式 (2.70) 等价。

答: 消去 Douglas 格式的辅助网格函数即可。由 (2.71a) 可知

$$\left[ \mathbb{1} - \frac{1}{2} \mu_x a \delta_x^2 \right] \left[ u^{n+\frac{1}{2}} - u^n \right] = \left[ \mu_x a \delta_x^2 + \mu_y b \delta_y^2 \right] u^n. \quad (2.72)$$

将 Douglas 格式的两个差分方程相减, 有

$$u^{n+1} - u^{n+\frac{1}{2}} = \frac{1}{2} \mu_y b \delta_y^2 \left[ u^{n+1} - u^n \right], \quad (2.73)$$

从而

$$u^{n+\frac{1}{2}} - u^n = \left[ \mathbb{1} - \frac{1}{2} \mu_y b \delta_y^2 \right] \left[ u^{n+1} - u^n \right].$$

将其代入到 (2.72) 的左端, 可得

$$\left[ \mathbb{1} - \frac{1}{2} \mu_x a \delta_x^2 \right] \left[ \mathbb{1} - \frac{1}{2} \mu_y b \delta_y^2 \right] \left[ u^{n+1} - u^n \right] = \left[ \mu_x a \delta_x^2 + \mu_y b \delta_y^2 \right] u^n.$$

简单整理可知, 它就是二维 PR 格式在分裂之前的表达形式 (2.69)。命题得证。□

因此说, 二维 Douglas 格式同二维 PR 格式的理论性质和数值表现是完全相同的。但是, 它们的数值实现过程还是略有区别的。二维 Douglas 格式需要额外的存储空间, 来记录辅助网格函数的信息。换言之, 同二维 PR 格式相比, 二维 Douglas 格式没有任何数值优势。但是, 当它们的思想被推广到三维扩散问题的时候, 三维 Douglas 格式不再等价三维 PR 格式, 具有相应的数值优势。

<sup>20</sup>J. Jr. Douglas, *Alternating direction methods for three space variables*, Numer. Math., 4 (1962), 41–63

### 局部一维化方法

最初思想源于前苏联学者 Bagrinovskii 和 Godunov (1957), 主要研究结果由 Yanenko (1959) 完成。在 Yanenko (1965) 的论著中, LOD 方法被称为“分数步长方法”。

LOD 方法具有扎实的理论基础和物理背景。设相应的时间区间是  $[0, T]$ 。给定正整数  $N$ , 构造  $[0, T]$  的等距离散网格

$$\mathcal{T}_{\Delta t} = \{t^n = n\Delta t\}_{n=0}^N,$$

其中  $\Delta t = T/N$  是时间步长。记  $t^{n+1/2} = (t^n + t^{n+1})/2$  是中间时刻。保持空间变量的连续性, 考虑 (时间) 半离散问题:

1. 定义  $u^{\Delta t}(x, y, 0) = u(x, y, 0)$ , 保持初值的一致性;
2. 对  $n = 0 : N - 1$ , 依次求解两个“具有一维结构”的二维偏微分方程定解问题

$$\frac{1}{2}u_t^{\Delta t} = au_{xx}^{\Delta t}, \quad t \in (t^n, t^{n+\frac{1}{2}}]; \quad (2.74a)$$

$$\frac{1}{2}u_t^{\Delta t} = bu_{yy}^{\Delta t}, \quad t \in (t^{n+\frac{1}{2}}, t^{n+1}]. \quad (2.74b)$$

简单计算可知  $u^{\Delta t}(x, y, T) = u(x, y, T)$ 。事实上, 半离散问题可以解读为热量传导的逐维分解, 符合物理学的基本定律。

半离散问题 (2.74) 的各种数值格式, 都称为二维热传导方程 (2.55) 的 LOD 格式。例如, 利用 CN 格式, 可得经典 LOD 格式

$$\left(1 - \frac{1}{2}\mu_x a \delta_x^2\right) u^{n+\frac{1}{2}} = \left(1 + \frac{1}{2}\mu_x a \delta_x^2\right) u^n, \quad (2.75a)$$

$$\left(1 - \frac{1}{2}\mu_y b \delta_y^2\right) u^{n+1} = \left(1 + \frac{1}{2}\mu_y b \delta_y^2\right) u^{n+\frac{1}{2}}. \quad (2.75b)$$

**论题 2.21.** 对于纯初值问题或周期边值问题, 有

$$\delta_x^2 \delta_y^2 = \delta_y^2 \delta_x^2, \quad (2.76)$$

即二阶差分算子可交换。此时,  $LOD$  格式 (2.75) 同  $PR$  格式 (2.70) 是等价的, 无条件具有  $L^2$  模稳定性和  $(2, 2, 2)$  阶局部截断误差。

答: 将算子  $\mathbb{1} - \frac{1}{2}\mu_x a \delta_x^2$  作用到 (2.75b), 利用 (2.75a) 进行整理, 即可得到两个格式的等价性。□

$LOD$  方法常常用于 **预测校正格式** 的预测值计算。例如, 二维扩散方程 (2.55) 的 Yanenko 格式是

$$u^{n+\frac{1}{4}} = u^n + \frac{1}{2}\mu_x a \delta_x^2 u^{n+\frac{1}{4}}, \quad (2.77a)$$

$$u^{n+\frac{1}{2}} = u^{n+\frac{1}{4}} + \frac{1}{2}\mu_y b \delta_y^2 u^{n+\frac{1}{2}}, \quad (2.77b)$$

$$u^{n+1} = u^n + \mu_x a \delta_x^2 u^{n+\frac{1}{2}} + \mu_y b \delta_y^2 u^{n+\frac{1}{2}}. \quad (2.77c)$$

**¶ 论题 2.22.** 若  $\delta_x^2 \delta_y^2 = \delta_y^2 \delta_x^2$ , 则 Yanenko 格式 (2.77) 同二维 Douglas (或  $PR$ ) 格式是等价的。

答: 利用 Yanenko 格式的前两式, 消去  $u^{n+1/4}$ , 得到

$$\left(\mathbb{1} - \frac{1}{2}\mu_x a \delta_x^2\right) \left(\mathbb{1} - \frac{1}{2}\mu_y b \delta_y^2\right) u^{n+\frac{1}{2}} = u^n.$$

将其代入到 Yanenko 格式的最后一式, 可得

$$u^{n+1} - u^n = \left(\mu_x \delta_x^2 + \mu_y b \delta_y^2\right) \left(\mathbb{1} - \frac{1}{2}\mu_y b \delta_y^2\right)^{-1} \left(\mathbb{1} - \frac{1}{2}\mu_x a \delta_x^2\right)^{-1} u^n.$$

将算子  $(\mathbb{1} - \frac{1}{2}\mu_x a \delta_x^2)(\mathbb{1} - \frac{1}{2}\mu_y b \delta_y^2)$  作用到上式, 利用  $\delta_x^2 \delta_y^2 = \delta_y^2 \delta_x^2$  可得

$$\left(\mathbb{1} - \frac{1}{2}\mu_x a \delta_x^2\right) \left(\mathbb{1} - \frac{1}{2}\mu_y b \delta_y^2\right) (u^{n+1} - u^n) = \left(\mu_x a \delta_x^2 + \mu_y b \delta_y^2\right) u^n.$$

因此, Yanenko 格式等价于二维 Douglas (或  $PR$ ) 格式, 具有  $(2, 2, 2)$  阶局部截断误差, 并且按  $L^2$  模是无条件稳定的。□

---

## 第 3 章

# 线性双曲型方程

---

粗略地讲，抛物型方程的数值离散技巧和理论分析方法，也同样适用于双曲型方程。双曲型方程本身缺乏耗散机制，可以同时存在光滑解和间断界面，相应的数值困难更为突出。

### 3.1 迎风格式和 Lax-Wendroff 格式

本节介绍线性常系数对流方程

$$u_t + au_x = 0 \quad (3.1)$$

的迎风 (upwind) 格式和 Lax-Wendroff (LW) 格式。它们的数值表现截然不同，可以反映出双曲型方程的典型困难。

#### 3.1.1 迎风格式

最易想到的方法是用向前差商离散时间导数  $[u_t]_j^n$ ，同时用中心差商离散空间导数  $[u_x]_j^n$ ，建立对流方程 (3.1) 的中心差商显格式

$$u_j^{n+1} = u_j^n - \frac{1}{2}\nu a(u_{j+1}^n - u_{j-1}^n). \quad (3.2)$$

虽然无条件具有 (2, 1) 阶局部截断误差，它是无条件线性  $L^2$  模不稳定的。换言之，对于任意的网比  $\nu$ ，它都是  $L^2$  模不稳定的。

主要有两种解决方案。放弃高阶相容的中心差商离散，采用低阶相容的单侧差商离散，可得 (3.1) 的偏风格式

$$u_j^{n+1} = u_j^n - \nu a \Delta_{\pm x} u_j^n,$$

分别称为左偏心格式或右偏心格式。基于单侧差商离散，它仅仅具有  $(1, 1)$  阶局部截断误差。在对流方程 (3.1) 中， $a$  的符号指明了流动的方向。若  $a > 0$ ，则流动从左到右，左侧是上游方向；若  $a < 0$ ，则流动从右到左，右侧是上游方向。因此，下面两种状态的偏心格式

$$u_j^{n+1} = u_j^n - \nu a \Delta_- u_j^n, \quad a > 0, \quad (3.3a)$$

$$u_j^{n+1} = u_j^n - \nu a \Delta_+ u_j^n, \quad a < 0, \quad (3.3b)$$

称为迎风格式，因为其空间方向的离散模版均处于上游方向。

**┆ 论题 3.1.** 迎风格式 (3.3) 有条件具有  $L^2$  模稳定性。

答：以 (3.3a) 为例。代入模态解  $u_j^n = \lambda^n e^{ikj\Delta x}$ ，可得增长因子

$$\lambda = \lambda(k) = 1 - \nu a(1 - e^{-ik\Delta x}).$$

分离  $\lambda$  的实部和虚部，简单计算可得

$$|\lambda_{\text{upw}}|^2 = 1 - 4\nu a(1 - \nu a) \sin^2\left(\frac{1}{2}k\Delta x\right).$$

当且仅当  $0 < \nu a \leq 1$  时，严格的 von Neumann 条件成立，迎风格式 (3.3a) 具有  $L^2$  模稳定性。  $\square$

若偏心方向指向下游，则偏心格式是无条件线性不稳定的。因此说，在迎风格式 (3.3) 中，流动方向的准确判定是非常必要的。

### 3.1.2 Lax-Wendroff 格式

基于中心差商显格式的离散模版，可以构造二阶的 LW 格式<sup>1</sup>

$$u_j^{n+1} = u_j^n - \frac{1}{2}\nu a(u_{j+1}^n - u_{j-1}^n) + \frac{1}{2}\nu^2 a^2 \delta_x^2 u_j^n. \quad (3.4)$$

它蕴含相当丰富的数值设计思想，例如时间 Taylor 方法、待定系数方法、特征线方法和数值黏性修正方法等等。

<sup>1</sup>P. D Lax and B. Wendroff, *System of conservation laws*, Commun. Pure Appl. Math., 13 (1960), 217-237.

**‡ 论题 3.2.** 利用时间 Taylor 方法, 构建 LW 格式 (3.4)。

答: 利用时间方向的 Taylor 展开公式, 有

$$[u]_j^{n+1} = [u]_j^n + \Delta t [u_t]_j^n + \frac{1}{2}(\Delta t)^2 [u_{tt}]_j^n + \mathcal{O}((\Delta t)^3).$$

利用微分方程 (3.1), 将时间导数转化为空间导数, 有

$$\begin{aligned} [u_t]_j^n &= -[au_x]_j^n = -\frac{a}{2\Delta x} \Delta_{0x} [u]_j^n + \mathcal{O}((\Delta x)^2), \\ [u_{tt}]_j^n &= a^2 [u_{xx}]_j^n = \frac{a^2}{(\Delta x)^2} \delta_x^2 [u]_j^n + \mathcal{O}((\Delta x)^2), \end{aligned}$$

其中空间导数利用中心差商进行离散。综上所述, 有

$$[u]_j^{n+1} = [u]_j^n - \frac{\nu a}{2} \Delta_{0x} [u]_j^n + \frac{(\nu a)^2}{2} \delta_x^2 [u]_j^n + \mathcal{O}((\Delta x)^2 \Delta t + (\Delta t)^3).$$

略去无穷小量, 用数值解替换真解, 即得 LW 格式 (3.4)。构造过程清楚地表明: LW 格式无条件具有 (2, 2) 阶局部截断误差。□

**‡ 论题 3.3.** 当且仅当  $|\nu a| \leq 1$  时, LW 格式 (3.4) 具有  $L^2$  模稳定性。换言之, 其稳定性结论同  $a$  的符号无关。

答: 简单计算可得增长因子


$$\lambda(k) = 1 - \mathrm{i}\nu a \sin \xi - 2\nu^2 a^2 \sin^2 \left( \frac{1}{2} \xi \right),$$

其中  $\xi = k\Delta x$ 。分离相应的实部和虚部, 可得

$$\begin{aligned} |\lambda(k)|^2 &= \left[ 1 - 2\nu^2 a^2 \sin^2 \left( \frac{1}{2} \xi \right) \right]^2 + \left[ 2\nu a \sin \left( \frac{1}{2} \xi \right) \cos \left( \frac{1}{2} \xi \right) \right]^2 \\ &= 1 - 4\nu^2 a^2 (1 - \nu^2 a^2) \sin^4 \left( \frac{1}{2} \xi \right). \end{aligned} \quad (3.5)$$

当且仅当  $|\nu a| \leq 1$  时, 严格的 von Neumann 条件成立, LW 格式 (3.4) 具有  $L^2$  模稳定性。□

当  $|\nu a| \leq 1$  时, 迎风格式 (3.3) 满足离散最大模原理, 故而具有最大模稳定性。但是, 当  $0 < |\nu a| < 1$  时, LW 格式 (3.4) 的右端出现负系数, 离散最大模原理不再成立。数值实验和理论分析均表明: 除非  $|\nu a| = 1$ , LW 格式不具有最大模稳定性。

 **注释 3.1.** 当  $0 < |\nu a| < 1$  时, LW 格式 (3.4) 满足<sup>2</sup>

$$\|u^n\|_\infty \leq C n^{\frac{1}{12}} \|u^0\|_\infty,$$

其中界定常数  $C$  同  $\Delta x, \Delta t$  和  $n$  均无关。数值解趋于无穷的速度较慢, 常常被误读为最大模有界。

## 3.2 稳定性分析方法

### 3.2.1 数值黏性方法

迎风格式 (3.3) 可以视为中心差商显格式的黏性修正。利用绝对值的运算性质

$$\max(a, 0) = \frac{1}{2}(a + |a|), \quad \min(a, 0) = \frac{1}{2}(a - |a|),$$

迎风格式 (3.3) 可以统一写作

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = \frac{|a|\Delta x}{2} \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2}. \quad (3.6)$$

同中心差商显格式 (3.2) 相比较, (3.6) 的等号右端就是新增的数值黏性 (修正) 项, 其中  $\frac{1}{2}|a|\Delta x$  称为数值黏性系数, 可以随着网格加密而消失。

类似地, LW 格式 (3.4) 也可视为中心差商显格式的修正, 相应的数值黏性系数是  $\frac{1}{2}|a|^2\Delta t$ 。

---

<sup>2</sup>S. Larsson and V. Thomée, *Partial Differential Equations with Numerical Methods*, Springer-Verlag Berlin Heidelberg, 2003

上述论证表明：迎风格式和 LW 格式的构造过程，相当于利用二阶相容的中心差商离散带有数值黏性的对流扩散方程


$$u_t + au_x = \frac{1}{2}|a|\Delta x u_{xx}, \quad u_t + au_x = \frac{1}{2}|a|^2 \Delta t u_{xx}.$$

事实上，它们分别称为迎风格式和 LW 格式的修正方程。相比于对流方程 (3.1)，迎风格式和 LW 格式更加靠近相应的修正方程，数值格式的性质可以用修正方程的性质来“描述”。由微分方程理论或者 Fourier 理论可知，对流扩散方程

$$u_t + au_x = bu_{xx}, \quad b > 0$$

的扩散系数  $b$  越大，简谐波的衰减速度越快，微分方程的适定性表现越好。综上所述，可以断言：若数值黏性系数越大，则数值稳定性表现越好。例如，当  $|\nu a| \leq 1$  时，LW 格式的数值黏性系数弱于迎风格式，其数值稳定性表现也弱于迎风格式。后面的数值实验将会表明，LW 格式产生虚假的数值震荡，而迎风格式却没有。

强调指出，数值黏性的增加可以增加稳定性，却有可能导致相容性的下降，例如 LW 格式是二阶相容，而迎风格式是一阶相容。

 **注释 3.2.** 上述论证过程过于粗糙，没有给出差分格式稳定的时空约束条件。更多的细节参见 §A. 的修正方程方法。

### 3.2.2 CFL 方法

直接利用物理的流动观点或者数学的特征线理论进行诠释，迎风格式和 LW 格式的稳定性结论是显而易见的。

早在 1928 年，Courant、Fridrichs 和 Lewy 三位学者就已经提出著名的结论<sup>3</sup>：设差分方程同微分方程是相容的。若其稳定，则必然满足 **CFL 条件**

微分方程的依赖区域必须包含于差分方程的依赖区域，至少当  $\Delta x$  和  $\Delta t$  趋于零的时候。

<sup>3</sup>R. Courant, K. Friedrichs and H. Lewy, *Über die partiellen differenzengleichungen der mathematischen physik*, Math. Ann., 100 (1928), 32-74.



换言之, CFL 条件是相容格式具有稳定性的必要条件。它具有简单直接的特性, 被广泛地应用于双曲型方程数值方法的研究。

**‡ 论题 3.4.** 设  $a > 0$ , 给出两个偏心格式的 CFL 条件。

答: 考虑单步推进的 CFL 条件即可。不妨以网格点  $(x_j, t^{n+1})$  为参考点。在前一时刻  $t^n$ , 左偏心格式 (3.3a) 的数值依赖区域是  $[x_j - \Delta x, x_j]$ , 而 PDE 的依赖区域是点集  $\{x_j - a\Delta t\}$ 。相应的 CFL 条件是  $\{x_j - a\Delta t\} \subset [x_j - \Delta x, x_j]$ , 即  $|a|\Delta t \leq \Delta x$ , 或者说, 相应的 Courant 数或 CFL 数  $\gamma_{\text{cfl}}$  必须满足

$$\gamma_{\text{cfl}} \equiv \frac{|a|\Delta t}{\Delta x} \leq 1. \quad (3.7)$$

它恰好就是 Fourier 方法给出的稳定性条件。

类似地, 右偏心格式 (3.3b) 在右侧方向依次展开一个空间网格, 不包含对流方程的依赖区域。换言之, CFL 条件不成立, 右偏心格式是不稳定的。□

**‡ 论题 3.5.** 给出 LW 格式的 CFL 条件。

答: 在单步推进中, LW 格式在两侧方向均展开一个空间网格。因此, 相应的 CFL 条件也是 (3.7), 就是 Fourier 方法给出的稳定性条件。□

对于偏心格式和 LW 格式而言, 由 CFL 方法和 Fourier 方法给出的稳定性结论是相同的。但是, 要强调指出: **CFL 条件只是格式稳定的必要条件, 且稳定性概念没有明确指出具体的离散范数**。例如, 满足 CFL 条件 (3.7) 的中心差商显格式就是不稳定的, 即使在较弱的  $L^2$  模度量之下; 此外, 满足 CFL 条件的 LW 格式具有  $L^2$  模稳定性, 却不具有最大模稳定性。

### 3.2.3 单调格式与数值震荡

基于行波解结构可知, 对流方程 (3.1) 具有单调保持性质: **若初值是单调的, 则真解将一直保持相同的单调性**。理想的差分格式

$$u_j^{n+1} = \sum_{s=-l}^r \alpha_s u_{j+s}^n, \quad \forall j \forall n, \quad (3.8)$$

应当具有相同的性质：若数值初值是单调的，则数值解将一直保持相同的单调性。否则，单调性刻画出现错误，数值震荡现象随之产生。因此说，**单调保持性质**是非常重要的，它是数值格式避免数值震荡的前提条件。

单调保持性质同差分系数的符号有关。假设数值格式 (3.8) 存在负系数，不妨设  $\alpha_\ell < 0$ 。取**单增的初值函数**

$$u_j^0 = \begin{cases} 0, & j \leq \ell, \\ 1, & j > \ell. \end{cases}$$

将其代入到差分方程 (3.8)，简单计算可知

$$u_1^1 - u_0^1 = \sum_{s=-l}^r \alpha_s [u_{1+s}^0 - u_s^0] = \alpha_\ell < 0.$$

换言之，数值解不再保持单增性质，出现虚假的数值震荡。基于上述事实，数值工作者提出单调格式的概念。

**定义 3.1.** 若差分系数  $\{\alpha_s\}_{s=-l}^r$  都是**非负**的，则称差分方程 (3.8) 是**单调格式**。有时，也称作**正格式**。

显然，当 CFL 条件成立时，迎风格式 (3.3) 是单调格式；但是，LW 格式 (3.4) 不是单调格式，除非  $\nu a = \pm 1$ 。

**定理 3.1.** 单调格式具有**单调保持**性质；反之亦然。

**证明：**利用简单的数学归纳，即证。 □

相容的单调格式具有**凸组合的系数结构**，当前时刻数值解必是前一时刻数值解的凸组合。因此，最小值不减，最大值不增。换言之，相容的单调格式满足离散最大模原理，具有最大模稳定性。

**单调格式存在严重的缺点，就是它的相容阶不高。**相关结论收录在下面的 Godunov 定理<sup>4</sup>。

---

<sup>4</sup>A. Harten, J. M. Hyman and P. D. Lax, *On the finite difference approximation and entropy conditions for shocks*, Comm. Pure Appl. Math., 29 (1976), 297-321

**定理 3.2.** 单调格式 (3.8) 至多具有一阶局部截断误差。

**证明:** 设它是相容的。利用 Taylor 展开技术可知, 其局部截断误差是

$$\tau_j^n = \frac{\Delta x}{2\nu} \left[ a^2 \nu^2 - \sum_{s=-l}^r s^2 \alpha_s \right] [u_{xx}]_j^n + \mathcal{O}((\Delta t)^2 + (\Delta x)^3 / \Delta t).$$

对于单调格式而言, 差分系数  $\{\alpha_s\}_{s=-l}^r$  都是非负的。利用 Cauchy-Schwartz 不等式可知

$$a^2 \nu^2 = \left[ \sum_{s=-l}^r s \alpha_s \right]^2 \leq \sum_{s=-l}^r s^2 \alpha_s \sum_{s=-l}^r \alpha_s = \sum_{s=-l}^r s^2 \alpha_s.$$

等号成立的条件是  $\{\alpha_s\}_{s=-l}^r$  只有一个非零系数; 此时, 相应的差分方程没有任何实际价值, 无需考虑。换言之, 在通常情况下,

$$a^2 \nu^2 - \sum_{s=-l}^r s^2 \alpha_s \neq 0,$$

因此单调格式的局部截断误差至多一阶。定理得证。  $\square$

Godunov 定理表明: 高阶相容的线性格式必定存在负系数, 数值震荡现象是不可避免的。要建立高阶无震荡格式, 我们必须跳出单调格式的框架。

### 3.2.4 数值色散分析

产生数值震荡的根本原因, 是各种波数 (或频率) 的数值简谐波具有不同的传播速度。Fourier 方法的 (数值) 增长因子蕴含许多信息, 例如振幅变化率、相位速度和传播速度等。这个过程统称为数值色散分析<sup>5</sup>。

在微分系统和差分系统中, 初始时刻的单位简谐波  $u(x, 0) = e^{ikx}$  具有不同的表现。显然, 对流方程 (3.1) 的真解是  $u(x, t) = e^{i(kx + \omega t)}$ , 其相位速度

<sup>5</sup>L. N. Trefethen, *Group velocity in finite difference scheme*, SIAM Rev., 24 (1982), 113-136.

$\omega = \omega(k) = -ak$  是线性的，没有色散现象。利用分离变量方法或 Fourier 方法，差分格式 (3.8) 的数值解是

$$u_j^n = e^{i(kj\Delta x + \omega^* n\Delta t)} = [\lambda(k)]^n e^{ikj\Delta x}, \quad (3.9)$$

其中  $\omega^* = \omega^*(k)$  称为广义数值色散关系， $\lambda(k)$  是增长因子。注意到

$$\lambda(k) = e^{i\omega^*(k)\Delta t} = e^{-\omega_{\text{Im}}^*(k)\Delta t} \cdot e^{i\omega_{\text{Re}}^*(k)\Delta t} = |\lambda(k)| e^{i \arg \lambda(k)},$$

可知

$$|\lambda(k)| = e^{-\omega_{\text{Im}}^*(k)\Delta t}, \quad \arg \lambda(k) = \omega_{\text{Re}}^*(k)\Delta t \quad (3.10)$$

分别表示推进  $\Delta t$  的振幅变化率，和相位改变量。换言之，增长因子蕴含丰富的数值波动信息。

#### 1. 增长因子的模决定数值耗散性质。

- 若  $\omega^*(k)$  的虚部为正，则相应的简谐波振幅（或者能量）将会衰减，形成数值耗散现象。此时，差分格式称为有耗散的。  
通常，在双曲方程的差分格式中，高波数简谐波承受更强的数值耗散。格式的稳定性结论主要取决于低波数简谐波的数值表现。
- 若  $\omega^*(k)$  的虚部为负，则相应的简谐波振幅将会膨胀，形成反数值耗散现象。若反数值耗散现象极其严重，破坏了 von Neumann 条件，则  $L^2$  模稳定性也将丧失。
- 若  $\omega^*(k)$  的虚部是零，则相应的简谐波振幅保持不变。若  $\omega^*(k)$  恒为实数，则差分格式称为无耗散的。

#### 2. 增长因子的辐角决定数值色散性质。数值相位速度

$$\arg \lambda(k) / \Delta t = \omega_{\text{Re}}^*(k), \quad (3.11)$$

或者利用相位速度相对误差

$$\frac{\omega_{\text{Re}}^*(k)}{\omega(k)} - 1 = -\frac{\arg \lambda(k)}{ka\Delta t} - 1, \quad (3.12)$$

均可以说明数值简谐波同真实简谐波的传播快慢。换言之，若 (3.12) 为正（负），则数值简谐波超前（滞后）于真实简谐波。

通常，数值相位速度 (3.11) 是非线性的，不同波数的数值简谐波具有不同波速，产生数值色散现象。在无耗散格式中，数值色散现象尤其突出。

由于数值色散现象，数值解的整体波形可能出现明显变化。若出现虚假的数值震荡现象<sup>6</sup>，其位置可以采用如前判断：若 (3.12) 恒正（负），则数值震荡出现在波前（后）。

**┆ 论题 3.6.** 在  $LW$  格式中，数值震荡必然出现在波后。

**答：**假设  $\xi$  足够小，或者说数值简谐波具有较低的波数，在空间网格上具有较高的辨识度。利用 Taylor 展开技术，有

$$\begin{aligned}\arg \lambda(k) &= -\arctan \left[ \frac{\nu a \sin \xi}{1 - 2\nu^2 a^2 \sin^2 \frac{1}{2}\xi} \right] \\ &= -\nu a \xi \left[ 1 - \frac{1}{6}(1 - \nu^2 a^2)\xi^2 + \dots \right].\end{aligned}\quad (3.13)$$

在方括号内，数字 1 后面的部分就是相位速度相对误差。当  $\nu|a| < 1$  时，它是非正的。数值传播速度低于真实传播速度，数值震荡出现在波后。  $\square$

### 3.3 双曲型方程组

设  $\mathbf{u}(x, t) = (u_1, u_2, \dots, u_m)^\top$  是向量值函数， $\mathbb{A}$  是给定的  $m$  阶实矩阵。考虑线性常系数偏微分方程组

$$\mathbf{u}_t + \mathbb{A}\mathbf{u}_x = 0. \quad (3.14)$$

若  $\mathbb{A}$  的特征值  $\{d_\ell\}_{\ell=1}^m$  都是实数，且（右）特征向量  $\{\mathbf{r}_\ell\}_{\ell=1}^m$  线性无关，则称 (3.14) 是双曲型的。

<sup>6</sup>数值色散不一定产生数值震荡，例如单调的迎风格式。

换言之, 矩阵  $\mathbf{A}$  具有特征分解  $\mathbf{A} = \mathbb{R}\mathbb{D}\mathbb{R}^{-1}$ , 其中  $\mathbb{D} = \text{diag}\{d_\ell\}_{\ell=1}^m$  是特征值构成的对角阵,  $\mathbb{R} = (\mathbf{r}_1, \dots, \mathbf{r}_m)$  是特征向量构成的相似变换阵. 令

$$\mathbb{D}^\oplus = \text{diag}\{\max(d_\ell, 0)\}_{\ell=1}^m, \quad \mathbb{D}^\ominus = \text{diag}\{\min(d_\ell, 0)\}_{\ell=1}^m,$$

定义矩阵  $\mathbf{A}$  的正 (负) 部和绝对值

$$\mathbf{A}^\oplus = \mathbb{R}\mathbb{D}^\oplus\mathbb{R}^{-1}, \quad \mathbf{A}^\ominus = \mathbb{R}\mathbb{D}^\ominus\mathbb{R}^{-1}, \quad |\mathbf{A}| = \mathbf{A}^\oplus - \mathbf{A}^\ominus = \mathbb{R}|\mathbb{D}|\mathbb{R}^{-1}. \quad (3.15)$$

**论题 3.7.** 构造双曲型方程组 (3.14) 的迎风格式。

答: 利用特征分解, 由 (3.14) 可以得到完全解耦的双曲型方程组

$$\mathbf{v}_t + \mathbb{D}\mathbf{v}_x = 0, \quad (3.16)$$

其中  $\mathbf{v}$  是  $\mathbf{u}$  在特征 (向量) 空间  $\text{span}\{\mathbf{r}_1, \dots, \mathbf{r}_m\}$  的投影坐标, 即

$$\mathbf{v} = \mathbb{R}^{-1}\mathbf{u} = (v_1, v_2, \dots, v_m)^\top.$$

换言之, 有  $(v_\ell)_t + d_\ell(v_\ell)_x = 0$ . 利用标量方程的迎风离散技术, 写出相应的差分方程

$$\begin{aligned} (v_\ell)_j^{n+1} = & (v_\ell)_j^n - \nu \max(d_\ell, 0) \left[ (v_\ell)_j^n - (v_\ell)_{j-1}^n \right] \\ & - \nu \min(d_\ell, 0) \left[ (v_\ell)_{j+1}^n - (v_\ell)_j^n \right], \end{aligned} \quad (3.17)$$

其中  $\ell = 1 : m$ . 将它们整合起来, (3.16) 的迎风格式可以简述为

$$\mathbf{v}_j^{n+1} = \mathbf{v}_j^n - \nu \mathbb{D}^\oplus \left[ \mathbf{v}_j^n - \mathbf{v}_{j-1}^n \right] - \nu \mathbb{D}^\ominus \left[ \mathbf{v}_{j+1}^n - \mathbf{v}_j^n \right]. \quad (3.18)$$

将  $\mathbf{v}$  变换到  $\mathbf{u}$ , 可得双曲型方程组 (3.14) 的迎风格式

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \nu \mathbf{A}^\oplus \left[ \mathbf{u}_j^n - \mathbf{u}_{j-1}^n \right] - \nu \mathbf{A}^\ominus \left[ \mathbf{u}_{j+1}^n - \mathbf{u}_j^n \right]. \quad (3.19)$$

由于表达形式同标量方程的 CIR 格式基本相同, 故而也称为 CIR 格式。 □

事实上, 矩阵的特征分解隐含着“通量分裂技术”的基本思想, 即

$$\mathbb{A}\mathbf{u} \equiv \mathbf{f}(\mathbf{u}) = \mathbf{f}^{\oplus}(\mathbf{u}) + \mathbf{f}^{\ominus}(\mathbf{u}) \equiv \mathbb{A}^{\oplus}\mathbf{u} + \mathbb{A}^{\ominus}\mathbf{u}.$$

由于通量函数  $\mathbf{f}^{\oplus}(\mathbf{u})$  和  $\mathbf{f}^{\ominus}(\mathbf{u})$  均具有明确的上游方向, 相应的迎风离散是显而易见的, 让迎风格式 (3.19) 的构造变得水到渠成。

**¶ 论题 3.8.** 建立迎风格式 (3.19) 的  $L^2$  模稳定性结果。

答: 利用 Fourier 方法或熟知的  $L^2$  模稳定性结论, 可知: 当且仅当  $|d_{\ell}|\Delta t \leq \Delta x$  时, 差分格式 (3.17) 具有  $L^2$  模稳定性, 即

$$\|v_{\ell}^n\|_2 \leq \|v_{\ell}^0\|_2, \quad \forall n, \quad \ell = 1 : m.$$

注意到 (3.18) 是完全解耦的, 可以断言: 当且仅当

$$\rho(\mathbb{A})\Delta t = \max_{\ell=1:m} |d_{\ell}|\Delta t \leq \Delta x, \quad (3.20)$$

差分格式 (3.18) 具有  $L^2$  模稳定性结论

$$\|\mathbf{v}^n\|_2 \leq \|\mathbf{v}^0\|_2 \equiv \left( \sum_{\ell=1}^m \|v_{\ell}^0\|_2^2 \right)^{\frac{1}{2}}, \quad \forall n. \quad (3.21)$$

注意到  $\mathbf{v} = \mathbb{R}^{-1}\mathbf{u}$ , 两个网格函数的  $L^2$  模要么同时稳定, 或者同时不稳定。事实上, 由 (3.21) 可得

$$\begin{aligned} \|\mathbf{u}^n\|_2 &\leq \|\mathbb{R}\|_2 \|\mathbb{R}^{-1}\mathbf{u}^n\|_2 = \|\mathbb{R}\|_2 \|\mathbf{v}^n\|_2 \leq \|\mathbb{R}\|_2 \|\mathbf{v}^0\|_2 \\ &= \|\mathbb{R}\|_2 \|\mathbb{R}^{-1}\mathbf{u}^0\|_2 \leq \|\mathbb{R}\|_2 \|\mathbb{R}^{-1}\|_2 \|\mathbf{u}^0\|_2, \end{aligned} \quad (3.22)$$

即存在界定常数  $C = \|\mathbb{R}\|_2 \|\mathbb{R}^{-1}\|_2$ , 使得

$$\|\mathbf{u}^n\|_2 \leq C \|\mathbf{u}^0\|_2, \quad \forall n. \quad (3.23)$$

反之亦然。因此, 迎风格式 (3.19) 具有  $L^2$  模稳定性的充要条件也是 (3.20)。

□

在迎风格式中, 特征分解是必须的。对于线性变系数 (或者半线性) 问题, 特征分解需要在每个网格点上执行, 消耗大量的 CPU 时间。因此, 无需特征分解的数值格式更受欢迎, 例如 Lax 格式

$$\mathbf{u}_j^{n+1} = \frac{1}{2}(\mathbf{u}_{j-1}^n + \mathbf{u}_{j+1}^n) - \frac{1}{2}\nu\mathbb{A}\Delta_{0,x}\mathbf{u}_j^n, \quad (3.24)$$

和 LW 格式

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \frac{1}{2}\nu\mathbb{A}\Delta_{0,x}\mathbf{u}_j^n + \frac{1}{2}\nu^2\mathbb{A}^2\delta_x^2\mathbf{u}_j^n. \quad (3.25)$$

与标量方程的同名格式相比, 它们的表达形式基本相同, 仅仅是数  $a$  变成了矩阵  $\mathbb{A}$  而已, 相应的  $L^2$  模稳定性条件都是 (3.20)。

### 3.4 高维对流方程

设  $a$  和  $b$  是给定的两个正常数, 考虑二维线性常系数对流方程

$$u_t + au_x + bu_y = 0. \quad (3.26)$$

设  $\mathcal{T}_{\Delta x, \Delta y, \Delta t}$  是等距的时空网格, 其中  $\Delta x$  和  $\Delta y$  是两个空间方向的空间步长,  $\Delta t$  是时间步长。对应两个空间方向, 相应的网比分别记作  $\nu_x = \Delta t / \Delta x$  和  $\nu_y = \Delta t / \Delta y$ 。

基于一阶空间导数的迎风离散策略, 利用逐维离散化技术, 即可轻松地建立 (3.26) 的二维迎风格式

$$u_{jk}^{n+1} = u_{jk}^n - \nu_x a \left[ u_{j,k}^n - u_{j-1,k}^n \right] - \nu_y b \left[ u_{j,k}^n - u_{j,k-1}^n \right]. \quad (3.27)$$

显然, 它无条件具有  $(1, 1, 1)$  阶局部截断误差。

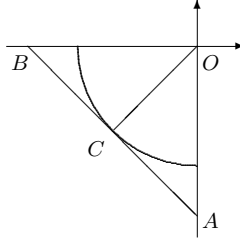
**论题 3.9.** 为简单起见, 设  $a = b = 1$  且  $\Delta x = \Delta y = h$ , 记  $r = \Delta t / h$ , 则二维迎风格式 (3.27) 可以简写为

$$u_{jk}^{n+1} = (1 - 2r)u_{jk}^n + r \left[ u_{j,k-1}^n + u_{j-1,k}^n \right]. \quad (3.28)$$

建立相应的 CFL 条件、 $L^2$  模稳定性结论和最大模稳定性结论。



图 3.1: 数值依赖区域和真实依赖区域



答: 任取一个时空网格点, 考虑其在  $\Delta t$  回溯之后的依赖区域。真实依赖区域是为其为顶点的四分之一圆锥, 锥底是以  $|OC| = \Delta t$  为半径的扇形区域; 数值依赖区域是为其为顶点的三角锥, 锥底是以  $|OA| = |OB| = h$  为直边的三角形; 参见图 3.1。因此, 迎风格式 (3.28) 的 CFL 条件是扇形区域不超出直角三角形, 即

$$r \leq 1/\sqrt{2}. \quad (3.29)$$

设  $\ell_1$  和  $\ell_2$  是任意实数, 对应两个方向的波数。将二维模态解

$$u_{jk}^n = [\lambda(\ell_1, \ell_2)]^n e^{i(\ell_1 j h + \ell_2 k h)}$$

代入到差分方程 (3.28), 简单计算, 有

$$\begin{aligned} |\lambda(\ell_1, \ell_2)|^2 &= (1 - 2r)^2 + 2r^2 + 2r(1 - 2r)[\cos(\ell_1 h) + \cos(\ell_2 h)] \\ &\quad + 2r^2 \cos[(\ell_1 - \ell_2)h] \\ &\leq \max(1, (1 - 4r)^2), \quad \forall \ell_1, \forall \ell_2. \end{aligned}$$

这是一个简单的二元函数极值问题, 具体推导过程略。由于右端上界可以取到, 故而结论是无法改进的。此时, 迎风格式 (3.28) 具有  $L^2$  模稳定性的充要条件是严格的 von Neumann 条件, 即

$$\max(1, (1 - 4r)^2) \leq 1,$$

相应的等价条件是  $r \leq 1/2$ 。

当  $r \leq 1/2$  时, 迎风格式 (3.28) 的右端系数都是非负的。由离散最大模原理可知, 它具有最大模稳定性。□

**‡ 论题 3.10.** 构造二维对流方程 (3.26) 的 **LW 格式**。

答: 利用时间 Taylor 方法, 有

$$[u]_{jk}^{n+1} = [u]_{jk}^n + \Delta t [u_t]_{jk}^n + \frac{1}{2} \Delta t^2 [u_{tt}]_{jk}^n + \mathcal{O}((\Delta t)^3),$$

其中对流方程 (3.26) 蕴含

$$\begin{aligned} [u_t]_{jk}^n &= -a[u_x]_{jk}^n - b[u_y]_{jk}^n, \\ [u_{tt}]_{jk}^n &= a^2[u_{xx}]_{jk}^n + 2ab[u_{xy}]_{jk}^n + b^2[u_{yy}]_{jk}^n. \end{aligned}$$

利用中心差商离散空间导数, 可得二维 LW 格式

$$\begin{aligned} u_{jk}^{n+1} &= u_{jk}^n - \frac{1}{2}(\nu_x a \Delta_{0x} u_{jk}^n + \nu_y b \Delta_{0y} u_{jk}^n) \\ &\quad + \frac{1}{2}(\nu_x^2 a^2 \delta_x^2 u_{jk}^n + \nu_y^2 b^2 \delta_y^2 u_{jk}^n) + \frac{1}{4} \nu_x \nu_y ab \Delta_{0x} \Delta_{0y} u_{jk}^n. \end{aligned} \quad (3.30)$$

它的离散模版共含有九个空间网格点。利用 Fourier 方法可知,


$$|\nu_x a| \leq \frac{1}{2\sqrt{2}}, \quad |\nu_y b| \leq \frac{1}{2\sqrt{2}} \quad (3.31)$$

是二维 LW 格式 (3.30) 的  $L^2$  模稳定性条件。同一维 LW 格式 (3.4) 相比, 它的时空约束条件更加苛刻。□

二维 LW 格式 (3.30) 不是一维 LW 格式 (3.4) 的直接推广。若采用逐维离散的思路, 可得离散模版更为简洁的差分方程

$$\begin{aligned} u_{jk}^{n+1} &= u_{jk}^n - \frac{1}{2}(\nu_x a \Delta_{0x} u_{jk}^n + \nu_y b \Delta_{0y} u_{jk}^n) \\ &\quad + \frac{1}{2}(\nu_x^2 a^2 \delta_x^2 u_{jk}^n + \nu_y^2 b^2 \delta_y^2 u_{jk}^n). \end{aligned} \quad (3.32)$$

但是, 它丧失了时间方向的二阶相容, 而且是线性无条件  $L^2$  模不稳定的, 没有应用价值。

 **注释 3.3.** 算子分裂方法可以有效改善高维双曲型方程差分格式的計算效率，例如 *LOD* 格式

$$u^{n+\frac{1}{2}} = u^n - \frac{1}{2}\nu_x a \Delta_{0x} u^n + \frac{1}{2}\nu_x^2 a^2 \delta_x^2 u^n, \quad (3.33a)$$

$$u^{n+1} = u^{n+\frac{1}{2}} - \frac{1}{2}\nu_y b \Delta_{0y} u^{n+\frac{1}{2}} + \frac{1}{2}\nu_y^2 b^2 \delta_y^2 u^{n+\frac{1}{2}} \quad (3.33b)$$

具有更为宽松的时空约束条件。当然，采用 *Strang* 镜像策略，数值效果更为理想。

---

## 第 4 章

# 非线性双曲守恒律

---

**非线性**双曲守恒律方程具有广泛的应用背景，例如 Euler 方程组和浅水波方程组等。为简单起见，考虑一维标量**非线性双曲守恒律**

$$u_t + f(u)_x = 0, \quad (x, t) \in \mathbb{R} \times \mathbb{R}^+, \quad (4.1)$$

其中  $u: \mathbb{R} \times \mathbb{R}^+ \rightarrow \mathbb{R}$  是未知函数， $f: \mathbb{R} \rightarrow \mathbb{R}$  是**连续可微**的已知通量函数。即使初值充分光滑，它也可能演化出激波、稀疏波和接触间断等各种复杂多变的结构。由于**局部光滑程度可能发生突变**，用于线性问题的数值格式可能给出完全错误的计算结果。

### 4.1 弱解和熵解

即使初值充分光滑，非线性双曲守恒律依旧有可能出现特征线相交的情形。因此，古典解概念需要进行相应的拓展。

🕒 **定义 4.1.** 定义在**上半时空平面**  $\mathbb{R} \times \mathbb{R}^+$  且具有**紧支集**<sup>1</sup>的**无穷光滑**函数全体，构成**检验函数空间**  $\mathcal{H}$ 。称**可测**的**有界函数**

$$u(x, t): \mathbb{R} \times \mathbb{R}^+ \rightarrow \mathbb{R}$$

是双曲守恒律 (4.1) 的**弱解**，若它对于任意的检验函数  $\phi(x, t) \in \mathcal{H}$  均成立

$$\iint_{t \geq 0} [u\phi_t + f(u)\phi_x] dx dt + \int_{t=0} u_0(x)\phi(x, 0) dx = 0, \quad (4.2)$$

其中  $u_0(x)$  是给定的初值。

---

<sup>1</sup>取值非零的函数点集闭包是有界的。

为简单起见, 本书的弱解概念仅仅局限于有限片段的古典解。假设某条连续可微的时空界面曲线

$$\Gamma: x = x(t), \quad t \geq 0 \quad (4.3)$$

将  $\mathbb{R} \times \mathbb{R}^+$  划分为左右两块区域, 相应的古典解分别记为  $u_1(x, t)$  和  $u_2(x, t)$ 。双曲守恒律 (4.1) 的弱解

$$u(x, t) = \begin{cases} u_1(x, t), & x < x(t), \\ u_2(x, t), & x > x(t), \end{cases} \quad (4.4)$$

必须满足著名的 Rankie Hugoniot (RH) 跳跃条件:

$$s(u_+ - u_-) = f(u_+) - f(u_-), \quad (4.5)$$

其中  $s \equiv x'(t)$  是界面曲线  $\Gamma$  的移动速度,  $u_{\pm} = \lim_{x \rightarrow x(t) \pm 0} u(x, t)$  是两侧函数的左右 (空间) 极限。

**注释 4.1.** 事实上, RH 跳跃条件可以由弱解的定义直接导出。它阐述了双曲守恒律的局部守恒性质, 无论时空界面曲线  $\Gamma$  两侧的函数是否连续。

弱解是古典解的拓展, 但是唯一性常常遭到破坏。通常引进适当的限制条件, 从弱解集合中筛选出唯一的物理解。

**定义 4.2.** 设  $u(x, t)$  是双曲守恒律 (4.1) 的弱解, 在时空界面曲线上满足 RH 跳跃条件 (4.5)。若还成立 Oleinik 熵条件: 对于  $u^-$  和  $u^+$  之间的任意值  $v$ , 均有

$$\frac{f(u_-) - f(v)}{u_- - v} \geq s \geq \frac{f(u_+) - f(v)}{u_+ - v}, \quad (4.6)$$

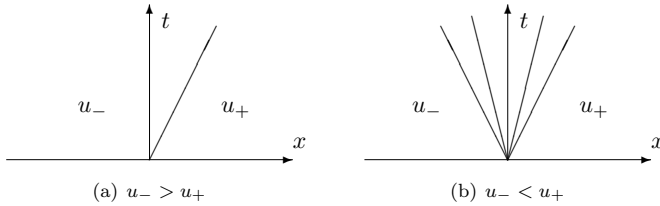
则称  $u(x, t)$  是双曲守恒律 (4.1) 的熵解。

Oleinik (1957) 证明: 标量双曲守恒律 (4.1) 的熵解存在且唯一。若  $f(\cdot)$  是凸可微的, 则 Oleinik 熵条件可以简化为著名的 Osher 熵条件:

$$f'(u_-) \geq s \geq f'(u_+). \quad (4.7)$$

换言之, 对应熵解的特征线不会远离时空界面曲线。

图 4.1: 激波和稀疏解



跳出连续可微的古典解范畴, 假设熵解在  $(x_*, t^*)$  点出现间断。依据间断点两侧的状态, 后续时刻的熵解可以局部演化出下面的结构:

1. 若后续时刻的时空区域处处有特征线穿过, 则间断点  $(x_*, t^*)$  将演化成间断界面  $x = x(t)$ , 相应的熵解具有局部的间断结构。

(a) 若熵条件 (4.6) 或 (4.7) 的不等式严格成立, 则两侧的特征线均交汇到间断界面。相应的局部间断结构称为**激波**, 其移动速度  $s = x'(t)$  称为**激波速度**。

(b) 若熵条件 (4.6) 或 (4.7) 局部退化为恒等式, 则两侧的特征线同间断界面是局部平行的。此时, 双曲守恒律局部退化为线性双曲型方程, 相应的局部间断结构称为**接触间断**。事实上, 线性常系数对流方程的间断界面就是接触间断。

2. 若在后续时刻的某个扇形(时空)区域没有特征线穿过, 则间断点  $(x_*, t^*)$  将会消失, 相应的熵解局部具有**稀疏波**结构。在扇形区域内部, 熵解具有自相似结构

$$u(x, t) = u\left(\frac{x - x_*}{t - t^*}\right),$$

将扇形区域外侧的两个状态连续地联接起来。换言之, 稀疏波是局部连续的, 但是位于两条射线的空间导数是间断的。

考虑 Burgers 方程 ( $f(u) = u^2/2$ ) 的 Riemann 问题, 相应的初值是简单的分段常值函数, 即

$$u(x, 0) = \begin{cases} u_-, & x < 0; \\ u_+, & x > 0. \end{cases} \quad (4.8)$$

当  $u_- = u_+$  时, 熵解就是一个常值函数, 属于古典解。当  $u_- > u_+$  时, 熵解是激波, 即 (图 4.1 的左侧)

$$u(x, t) = \begin{cases} u_-, & x - st < 0, \\ u_+, & x - st > 0, \end{cases} \quad (4.9a)$$

其中  $s = \frac{1}{2}(u_- + u_+)$  是激波速度。当  $u_- < u_+$  时, 熵解是稀疏波, 即 (图 4.1 的右侧)

$$u(x, t) = \begin{cases} u_-, & x - u_-t < 0; \\ u_+, & x - u_+t > 0; \\ x/t, & \text{其它}. \end{cases} \quad (4.9b)$$

事实上, (4.9a) 永远都是弱解。但是, 当  $u_- < u_+$  时, 它违背熵条件 (4.7), 不是熵解。

由于不断演化的各种细微结构, 非线性双曲守恒律的数值方法面临非常严峻的挑战。我们需要同时解决以下问题:

1. 激波速度的刻画要准确, 间断界面的捕捉要准确;
2. 在真解相对光滑区域, 相容阶要高, 计算效率要高;
3. 间断界面附近的数值震荡现象要得到控制;
4. 数值解要收敛到唯一的熵解。

能够实现上述目标的格式称为高精度高分辨率格式。常用的构造方法有两种:

- 其一是激波装配技术, 基本思路是先用特殊的算法确定间断界面的位置, 再用高效高精度格式计算界面之间的光滑解。

- 其二是激波捕捉技术，基本思路是不追踪间断界面的位置，而是直接建立统一的数值操作过程，可以自动地适用于光滑解和间断解的数值模拟。

本讲义仅讨论激波捕捉技术。

## 4.2 守恒型差分格式

记  $f_j^n = f(u_j^n)$ 。采用上游信息进行流量导数的单侧逼近，即可得到 (4.1) 的迎风格式

$$u_j^{n+1} = \begin{cases} u_j^n - \nu [f_j^n - f_{j-1}^n], & \text{若 } a(u_j^n) > 0; \\ u_j^n - \nu [f_{j+1}^n - f_j^n], & \text{若 } a(u_j^n) \leq 0. \end{cases} \quad (4.10)$$

利用算术平均技术改善中心差商全显格式的稳定性，可得 (4.1) 的 Lax 格式

$$u_j^{n+1} = \frac{1}{2}(u_{j-1}^n + u_{j+1}^n) - \frac{1}{2}\nu [f_{j+1}^n - f_{j-1}^n]. \quad (4.11)$$

显然，迎风格式 (4.10) 是无条件相容，而 Lax 格式 (4.11) 是有条件相容。当网比  $\nu$  固定时，它们均具有整体一阶的局部截断误差。

基于时间 Taylor 方法和积分插值方法，可得相应的 LW 格式

$$\begin{aligned} u_j^{n+1} = & u_j^n - \frac{\nu}{2} [f_{j+1}^n - f_{j-1}^n] \\ & + \frac{\nu^2}{2} \left\{ A_{j+\frac{1}{2}}^n [f_{j+1}^n - f_j^n] - A_{j-\frac{1}{2}}^n [f_j^n - f_{j-1}^n] \right\}, \end{aligned} \quad (4.12a)$$

其中

$$A_{j+\frac{1}{2}}^n = a(u_{j+\frac{1}{2}}^n) = f' \left( \frac{1}{2}(u_j^n + u_{j+1}^n) \right) \quad (4.12b)$$

是位于网格点中间位置的流场速度。

利用系数冻结分析方法或者 CFL 方法，可知它们稳定的模糊条件都是

$$\max |f'(u)|\nu \leq 1. \quad (4.13)$$



**论题 4.1.** 考虑 Burgers 方程的 Riemann 问题。若初始状态是  $u_- = 1$  和  $u_+ = 0$ , 则相应的真解是右行速度为  $1/2$  的激波。定义数值初值

$$u_j^0 = 1, j \leq 0; \quad u_j^0 = 0, j > 0,$$

观察迎风格式 (4.10) 的数值结果。

**答:** 简单计算可知, 两个格式的数值解在任意时刻都等于初值, 数值间断界面没有移动起来。因此说, 数值解彻底违背 RH 联接条件, 其极限不可能是 Riemann 问题的弱解, 更谈不上熵解。□

两个格式失败的主要原因是局部守恒性质遭到严重破坏。基于此, Lax 和 Wendroff (1960) 提出了守恒型格式的概念。

**定义 4.3.** 称差分格式是**守恒型格式**, 若它可以统一表述为

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left[ \hat{f}_{j+\frac{1}{2}}^n - \hat{f}_{j-\frac{1}{2}}^n \right], \quad \forall j, \quad (4.14)$$

其中  $\hat{f}_{j+\frac{1}{2}}^n$  称为数值通量, 具有统一的表达方式。通常, 定义

$$\hat{f}_{j+\frac{1}{2}}^n = \hat{f}(u_{j-l+1}^n, u_{j-l+2}^n, \dots, u_j^n, u_{j+1}^n, \dots, u_{j+r}^n), \quad (4.15)$$

其中  $l$  和  $r$  是给定的正整数,  $\hat{f}$  是给定的数值通量函数<sup>2</sup>。

数值通量函数和守恒型格式是一一对应的, 常常被冠以相同的名称。数值通量函数通常具有以下两条性质:

- 关于每个变元都是局部 Lipschitz 连续的;
- 相容性条件, 即  $\hat{f}(p, \dots, p) = f(p)$ ;

前者控制舍入误差的影响, 后者保证格式的相容性。最简单的定义方式是仅仅依赖左右网格点值的两点型数值通量

$$\hat{f}_{j+\frac{1}{2}}^n \equiv \hat{f}(u_j^n, u_{j+1}^n). \quad (4.16)$$

下面, 我们给出两个具体的实例。

<sup>2</sup>在隐式格式中, 数值通量函数还会同  $t^{n+1}$  时刻的网格函数有关。

‡ 论题 4.2. Lax 格式 (4.11) 和 LW 格式 (4.12) 都是守恒型的。

答：在 Lax 格式，相应的数值通量是

$$(\hat{f}^{\text{Lax}})_{j+1/2}^n = \frac{1}{2}[f_j^n + f_{j+1}^n] - \frac{1}{2\nu}(u_{j+1}^n - u_j^n); \quad (4.17)$$

在 LW 格式中，相应的数值通量是

$$(\hat{f}^{\text{LW}})_{j+1/2}^n = \frac{1}{2}[f_j^n + f_{j+1}^n] - \frac{\nu}{2}A_{j+\frac{1}{2}}[f_{j+1}^n - f_j^n]. \quad (4.18)$$

因此，Lax 格式和 LW 格式都是守恒型的。简单验证可知，两个数值通量函数均满足（局部 Lipschitz）连续性和相容性条件。□

事实上，两个数值通量具有明显不同的性质。当 CFL 条件

$$\max_{\forall u} |f'(u)| \frac{\Delta t}{\Delta x} \leq 1$$

成立时，Lax 数值通量 (4.17) 关于第一个变元不减，关于第二个变元不增，是熵数值通量 (entropy numerical flux) 或者单调数值通量 (monotone numerical flux)。恰恰相反，LW 数值通量 (4.18) 不是单调数值通量。

针对双曲守恒律的数值模拟，守恒型格式占据非常重要的地位。它获得成功的主要根源是以下三个优点：

1. 数值误差  $e_j^n = u_j^n - [u]_j^n$  在空间方向的整体求和，同时间层数无关。这个性质符合整体质量守恒的计算目标。
2. 数值解的局部守恒性质内蕴 RH 跳跃条件的近似满足，间断界面（或激波）的位置可以获得相对可靠的数值追踪。
3. 数值收敛性结论相对完美。著名的 Lax-Wendroff (LW) 定理指出：设守恒型差分格式同双曲守恒律相容。当网格尺度趋于零时，若数值解几乎处处有界收敛到某个函数，则它必定是问题的弱解。尽管 LW 定理还不能完全保障数值解收敛到弱解，但是它至少表明守恒型格式的数值解是相对合理的。

## 4.3 有限体积方法

有限体积方法自动实现数值解的局部守恒性质，可以用于守恒型差分格式的设计。此时，数值通量函数的含义可以获得准确的直观理解。事实上，有限体积方法可视为特殊的差分方法。同差分方法相比，有限体积方法可以灵活地应用于复杂的空间网格。

### 4.3.1 基本框架

下面以双曲守恒律 (4.1) 为例，阐述有限体积格式的设计思路。有别于差分方法，空间变量和时间变量采用不同的离散方式。操作过程如下：

1. 将空间区域分割为互不重叠的工作单元组。例如，对于一维空间而言，相应的单元剖分是

$$\mathcal{T}_{\Delta x} = \{I_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})\}_{\forall j}, \quad (4.19)$$

其中  $I_j$  称为工作单元。记  $\Delta x_j = x_{j+1/2} - x_{j-1/2}$  是  $I_j$  的单元长度，称  $\Delta x = \max_{\forall j} \Delta x_j$  是单元剖分的参数。

2. 将时间区间离散为有限个点，构造时间网格  $\mathcal{T}_{\Delta t} = \{t^n\}_{n \geq 0}$ ，其中  $\Delta t^n = t^{n+1} - t^n$  是局部时间步长， $\Delta t = \max_{\forall n} \Delta t^n$  是时间步长。

在有限体积方法中，数值逼近的目标是

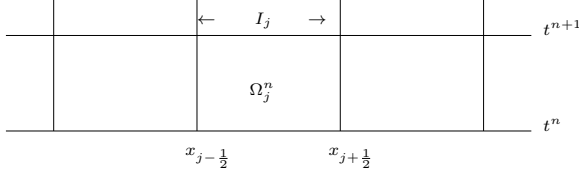
$$[\bar{u}]_j^n \equiv [\bar{u}]_j(t^n) \equiv \frac{1}{\Delta x_j} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t^n) dx, \quad \forall j \forall n, \quad (4.20)$$

即真解在不同时刻不同单元的均值。

参见图 4.2，在  $\Omega_j^n = I_j \times (t^n, t^{n+1})$  内积分偏微分方程。利用 Green 公式，将二维积分转化为一维线积分，可得

$$0 = \iint_{\Omega_j^n} \{u_t + f(u)_x\} dx dt = \oint_{\partial \Omega_j^n} \{f(u) dt - u dx\}.$$

图 4.2: 有限体积方法的基本框架



它蕴含的精确等式

$$[\bar{u}]_j^{n+1} = [\bar{u}]_j^n - \frac{\Delta t^n}{\Delta x_j} (F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n), \quad (4.21a)$$

是有限体积格式的设计起点，其中

$$F_{j+\frac{1}{2}}^n \equiv \frac{1}{\Delta t^n} \int_{t^n}^{t^{n+1}} f(u(x_{j+\frac{1}{2}}, t)) dt \quad (4.21b)$$

是位于单元界面  $x_{j+1/2}$  的**真实平均通量**。假设存在某个**数值通量函数**  $\mathcal{H}$ ，使得

$$F_{j+\frac{1}{2}}^n \approx \mathcal{H}([\bar{u}]_{j-l+1}^n, \dots, [\bar{u}]_{j+r}^n), \quad (4.22)$$

其中  $l$  和  $r$  是给定的正整数。两式联立，将近似关系视为相等，用数值均值  $\bar{u}_j^n$  代替真实均值  $[\bar{u}]_j^n$ ，即可得到 (4.1) 的有限体积格式

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\Delta t^n}{\Delta x_j} [\hat{f}_{j+\frac{1}{2}}^n - \hat{f}_{j-\frac{1}{2}}^n], \quad (4.23)$$

其中  $\hat{f}_{j+\frac{1}{2}}^n = \mathcal{H}(\bar{u}_{j-l+1}^n, \dots, \bar{u}_{j+r-1}^n, \bar{u}_{j+r}^n)$  是**数值通量**。通常， $\mathcal{H}$  定义为两点型数值通量，即

$$\hat{f}_{j+\frac{1}{2}}^n \equiv \mathcal{H}(\bar{u}_j^n, \bar{u}_{j+1}^n). \quad (4.24)$$

换言之，数值通量仅仅依赖单元界面两侧的单元均值。

有限体积方法也具有相容性、稳定性和收敛性概念。定义方式同有限差分方程基本相同，此处不再赘述。

### 4.3.2 线性问题的有限体积格式

设通量函数是  $f(u) = au$ ，其中  $a$  为给定的常数。换言之，双曲守恒律 (4.1) 就是线性常数对流方程 (3.1)。

**论题 4.3.** 为简单起见，设离散网格是一致的，即工作单元的长度都是  $\Delta x$ ，局部时间步长都是  $\Delta t$ 。构造 (3.1) 的迎风有限体积格式和 Lax 有限体积格式。

答：在单元界面  $x_{j+1/2}$  处，定义数值通量

$$\begin{aligned} (\hat{f}^{\text{upw}})^n_{j+\frac{1}{2}} &= \begin{cases} a\bar{u}_j^n, & \text{当 } a \geq 0, \\ a\bar{u}_{j+1}^n, & \text{当 } a < 0, \end{cases} \\ &= \frac{a + |a|}{2}\bar{u}_j^n + \frac{a - |a|}{2}\bar{u}_{j+1}^n. \end{aligned} \quad (4.25)$$

由于它只依赖上游单元的数据，故而称做（完全）迎风数值通量。将其代入到 (4.23)，可得迎风有限体积格式

$$\frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} + \frac{a + |a|}{2\Delta x} [\bar{u}_j^n - \bar{u}_{j-1}^n] + \frac{a - |a|}{2\Delta x} [\bar{u}_{j+1}^n - \bar{u}_j^n] = 0. \quad (4.26)$$

事实上，迎风数值通量 (4.25) 可以视为中心型数值通量

$$(\hat{f}^{\text{cen}})^n_{j+\frac{1}{2}} = \frac{a\bar{u}_j^n + a\bar{u}_{j+1}^n}{2} \quad (4.27)$$

的某种数值黏性修正，因为它具有等价形式

$$(\hat{f}^{\text{upw}})^n_{j+\frac{1}{2}} = \frac{a\bar{u}_j^n + a\bar{u}_{j+1}^n}{2} - \frac{|a|}{2} [\bar{u}_{j+1}^n - \bar{u}_j^n].$$

右端的第二项就是数值黏性修正项，其中  $\bar{u}_{j+1}^n - \bar{u}_j^n$  是界面位置的数值跳跃， $|a|/2$  称为修正强度。

仿照上述观点，其它形式的数值通量函数可以类似定义。例如，著名的 Lax 数值通量是


$$(\hat{f}^{\text{Lx}})^n_{j+\frac{1}{2}} = \frac{a\bar{u}_j^n + a\bar{u}_{j+1}^n}{2} - \frac{\Delta x}{2\Delta t} [\bar{u}_{j+1}^n - \bar{u}_j^n], \quad (4.28)$$

由它导出的有限体积格式


$$\bar{u}_j^{n+1} = \frac{1}{2}(\bar{u}_{j-1}^n + \bar{u}_{j+1}^n) - \frac{a\Delta t}{2\Delta x}(\bar{u}_{j+1}^n - \bar{u}_{j-1}^n), \quad (4.29)$$

称为 Lax 有限体积格式。  $\square$

有限体积方法和有限差分方法密切相关。若将单元均值  $\bar{u}_j^n$  理解为单元中心的点值  $u_j^n$ , 则上述两个有限体积格式都可诠释为同名的有限差分格式, 相应的空间离散网格由工作单元的中心点构成。反之, 若将网格点值  $u_j^n$  视为控制单元的均值  $\bar{u}_j^n$ , 则有限差分格式也可诠释为同名的有限体积格式, 相应的剖分是网格点的控制区间。对于足够光滑的函数, 单元均值和中心点值的差距是  $\mathcal{O}((\Delta x)^2)$ 。因此, 二阶以内的有限差分格式和有限体积格式具有相近的数值结果和理论概念。

 **注释 4.2.** 虽然有限体积方法可以视为特殊的有限差分方法, 它的收敛性证明需要借用有限元方法的分析技术, 特别是当单元剖分不是均匀的时候。

积分插值方法和有限体积方法的设计思想非常接近, 它们均基于偏微分方程在局部区域的积分近似过程。积分插值方法属于差分方法的范畴, 数值积分的信息来源是周边的点值信息, 而有限体积方法的信息来源是周边的均值信息。

 **论题 4.4.** 为简单起见, 设  $\mathcal{T}_{\Delta x, \Delta t} = \{(x_j, t^n)\}_{\forall j}^n$  是等距的时空网格, 其中  $\Delta x$  和  $\Delta t$  分别是空间步长和时间步长。利用积分插值方法, 构造 (3.1) 的迎风有限差分格式。

答: 以  $a > 0$  为例。选取局部区域  $(x_{j-1}, x_j) \times (t^n, t^{n+1})$ , 积分 (3.1), 有

$$\int_{x_{j-1}}^{x_j} [u(x, t^{n+1}) - u(x, t^n)] dx = a \int_{t^n}^{t^{n+1}} [u(x_j, t) - u(x_{j-1}, t)] dt.$$

空间积分用右矩形公式近似, 而时间积分用左矩形公式近似, 有

$$([u]_j^{n+1} - [u]_j^n) \Delta x \approx a([u]_j^n - [u]_{j-1}^n) \Delta t.$$

用数值解替换真解，即可得到迎风格式(3.3)。□

**■ 注释 4.3.** 基于上述观点，本书常常略去有限体积格式的均值符号。换言之，离散信息可以理解为中心点值或单元均值，数值格式可以视为有限体积格式或有限差分格式。

### 4.3.3 非线性问题的有限体积格式

对于非线性双曲守恒律 (4.1)，有限体积格式的构造是类似的。

**↯ 论题 4.5.** 设离散网格是一致的，构造 (4.1) 的迎风有限体积格式。

答：为行文简便，不妨直接用数值解进行描述。在  $t^n$  时刻，位于单元界面  $x_{j+1/2}$  的流速可以定义为

$$A_{j+1/2}^n = f'(\{\bar{u}\}_{j+1/2}^n), \quad (4.30)$$

其中  $\{\bar{u}\}_{j+1/2}^n = \frac{1}{2}(\bar{u}_j^n + \bar{u}_{j+1}^n)$  是两侧单元均值的算术平均。

利用  $A_{j+1/2}^n$  的符号，断定当时当地的上游方向，定义（完全依赖于上游信息的）迎风数值通量

$$\begin{aligned} \hat{f}_{j+\frac{1}{2}}^n &= \begin{cases} f_j^n, & \text{当 } A_{j+\frac{1}{2}} \geq 0, \\ f_{j+1}^n, & \text{当 } A_{j+\frac{1}{2}} < 0 \end{cases} \\ &= \frac{1}{2} [f_j^n + f_{j+1}^n] - \frac{1}{2} \text{sgn} A_{j+\frac{1}{2}}^n [f_{j+1}^n - f_j^n], \end{aligned} \quad (4.31)$$

其中  $f_j^n = f(\bar{u}_j^n)$ 。由基本框架 (4.23) 可知，相应的守恒型迎风格式为

$$\begin{aligned} \bar{u}_j^{n+1} &= \bar{u}_j^n - \frac{\nu}{2} \left\{ \left[ 1 - \text{sgn} A_{j+\frac{1}{2}}^n \right] \Delta_{+x} f_j^n \right. \\ &\quad \left. + \left[ 1 + \text{sgn} A_{j-\frac{1}{2}}^n \right] \Delta_{-x} f_j^n \right\}, \end{aligned} \quad (4.32)$$

其中  $\Delta_{\pm x}$  是向前/向后差分算子， $\nu = \Delta t / \Delta x$  为网比。□

视 (4.32) 为有限差分格式，可得守恒型迎风格式。它同简单迎风格式(4.10) 具有一个细微的区别，即上游方向的判断位置是不同的。守恒型迎风格式基

于控制区域的界面位置，而简单迎风格式直接基于网格点位置。如果流场方向保持恒定，它们是完全相同的。但是，当流场方向发生变化时，它们将有所区别。

类似地，可得 (4.1) 的 Lax 有限体积格式

$$\bar{u}_j^{n+1} = \frac{1}{2}(\bar{u}_{j-1}^n + \bar{u}_{j+1}^n) - \frac{\nu}{2} \left( f(\bar{u}_{j+1}^n) - f(\bar{u}_{j-1}^n) \right). \quad (4.33)$$

和 LW 有限体积格式

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\nu}{2} \left\{ \left[ 1 - \nu A_{j+\frac{1}{2}}^n \right] \Delta_{+x} f_j^n + \left[ 1 + \nu A_{j-\frac{1}{2}}^n \right] \Delta_{-x} f_j^n \right\}, \quad (4.34)$$

其中  $f_j^n = f(\bar{u}_j^n)$  且  $A_{j+1/2}^n = f'(\{\bar{u}\}_{j+1/2}^n)$ 。

## 4.4 Godunov 方法

Godunov 方法是由前苏联的著名数学家 Godunov 提出的<sup>3</sup>，用于证明气动力学方程组存在唯一的熵解。在某种程度上，它被公认为首个成功模拟非线性双曲守恒律的有限体积格式。Godunov 方法具有清晰的物理背景，已经发展为求解非线性双曲守恒律（组）的主流数值方法之一。

下面以双曲守恒律 (4.1) 为模型，简要描述 Godunov 方法的两种实现过程。为简单起见，设离散网格是一致的。

### 4.4.1 EA 过程

参见图 4.3；假设  $t^n$  时刻的单元均值  $\{\bar{u}_j^n\}_{\forall j}$  是已知的， $t^{n+1}$  时刻的单元均值  $\{\bar{u}_j^{n+1}\}_{\forall j}$  可以按照下面的过程给出：

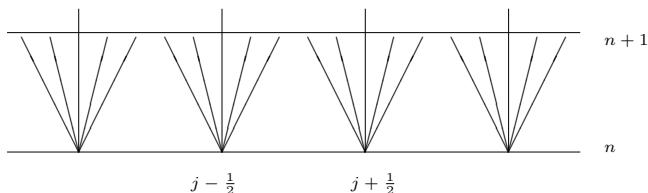
1. **局部推进**：在单元边界  $x_{j+1/2}$  处，构造 (4.1) 的局部 Riemann 问题，相

---

<sup>3</sup>S. K. Godunov, *A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics*, Mat. Sb., 47 (1959), 271–306



图 4.3: Godunov 方法的主要思想



应的初值是

$$u(x, t^n) = \begin{cases} u_{j+\frac{1}{2}}^-, & \text{当 } x < x_{j+\frac{1}{2}}, \\ u_{j+\frac{1}{2}}^+, & \text{当 } x > x_{j+\frac{1}{2}}, \end{cases} \quad (4.35)$$

其中  $u_{j+\frac{1}{2}}^\pm$  是已知函数在单元边界的左右极限, 即

$$u_{j+\frac{1}{2}}^- = \bar{u}_j^n, \quad u_{j+\frac{1}{2}}^+ = \bar{u}_{j+1}^n. \quad (4.36)$$

基于初值 (4.35) 在单元界面两侧的不同状态, 局部 Riemann 问题的真解 (简称为局部 Riemann 解) 可能是古典解、激波、接触间断或者稀疏波等。当  $\bar{u}_{j+1}^n = \bar{u}_j^n$  时, 用

$$s_{j+\frac{1}{2}}^n = \frac{f(\bar{u}_{j+1}^n) - f(\bar{u}_j^n)}{\bar{u}_{j+1}^n - \bar{u}_j^n},$$

表示 (可能存在的) 激波速度。若时间步长足够小, 使得

$$\max_{\forall j} \left( |s_{j+\frac{1}{2}}^n|, |f'(\bar{u}_j^n)| \right) \frac{\Delta t}{\Delta x} \leq \frac{1}{2}, \quad (4.37)$$

则相邻的局部 Riemann 解不会产生双值冲突, 可以拼接出  $t^{n+1}$  时刻的函数  $\tilde{u}(x, t^{n+1})$ 。

2. 单元平均: 计算  $\tilde{u}(x, t^{n+1})$  的单元均值, 即

$$\bar{u}_j^{n+1} \equiv \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} \tilde{u}(x, t^{n+1}) dx, \quad \forall j. \quad (4.38)$$

显然, EA 过程的数值误差主要源于单元平均。受限于常值函数的逼近效果, 它的局部截断误差只能达到整体一阶。

**论题 4.6.** 设双曲守恒律 (4.1) 是线性的, 即  $f(u) = au$ , 其中  $a$  是给定的常数。利用 EA 过程, 构造相应的 Godunov 格式。

答: 局部 Riemann 问题 (4.35) 具有精确解

$$u_{j+\frac{1}{2}}(x, t^{n+1}) = \begin{cases} \bar{u}_j^n, & \text{当 } x < x_{j+\frac{1}{2}} + a\Delta t, \\ \bar{u}_{j+1}^n, & \text{其他.} \end{cases} \quad (4.39)$$

当 (4.37) 或者等价的  $\nu a \leq 1/2$  成立时<sup>4</sup>, 相邻的局部 Riemann 解 (4.39) 没有出现双值冲突。不妨设  $a > 0$ , 计算单元均值

$$\begin{aligned} \bar{u}_j^{n+1} &= \frac{1}{\Delta x} \left[ \int_{x_{j-\frac{1}{2}}}^z u_{j-\frac{1}{2}}(x, t^{n+1}) dx + \int_z^{x_{j+\frac{1}{2}}} u_{j+\frac{1}{2}}(x, t^{n+1}) dx \right] \\ &= \frac{1}{\Delta x} [\bar{u}_{j-1}^n a\Delta t + (\Delta x - a\Delta t) \bar{u}_j^n], \end{aligned}$$

其中  $z = x_{j-1/2} + a\Delta t$ 。事实上, 它就是熟知的迎风有限体积格式, 相应的数值通量  $\hat{f}_{j+1/2}^n = a\bar{u}_j^n$  恰好就是局部 Riemann 解 (4.39) 在单元边界  $x_{j+1/2}$  的平均通量。□

事实上, 关于数值通量的上述论断是普遍成立的, 因为 EA 实现过程等同于双曲守恒律 (4.1) 在局部区域  $\Omega_j^n = I_j \times (t^n, t^{n+1})$  内的积分过程。换言之, EA 过程的数值解满足有限体积格式

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{\Delta t}{\Delta x} (\hat{f}_{j+\frac{1}{2}}^n - \hat{f}_{j-\frac{1}{2}}^n), \quad (4.40)$$

相应的数值通量是 Riemann 问题 (4.35) 的平均通量, 即

$$\hat{f}_{j+\frac{1}{2}}^n = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(u_{j+\frac{1}{2}}(x_{j+\frac{1}{2}}, t)) dt.$$

---

<sup>4</sup>可以放松到  $a\Delta t \leq \Delta x$ 。

注意到双曲守恒律的自相似性质, 局部 Riemann 解可以表示为

$$u_{j+\frac{1}{2}}(x, t) = \mathcal{R}\left(\frac{x - x_{j+\frac{1}{2}}}{t - t^n}; \bar{u}_j^n, \bar{u}_{j+1}^n\right), \quad (4.41)$$

其中  $\mathcal{R}$  是局部 Riemann 解算子。既然 (4.41) 在单元界面上保持不变, 相应的数值通量满足

$$\hat{f}_{j+\frac{1}{2}}^n = f(\mathcal{R}(0; \bar{u}_j^n, \bar{u}_{j+1}^n)). \quad (4.42)$$

换言之, 在有限体积方法中, 数值通量可以按照 (4.42) 进行构造。它等价于局部 Riemann 解 (4.41) 的计算。

对于非线性双曲守恒律 (4.1), 局部 Riemann 解 (4.41) 通常是无法准确给出的。此时, 可以将 (4.1) 局部近似为线性双曲型方程

$$u_t + A_{j+\frac{1}{2}}^n u_x = 0, \quad (4.43)$$

利用它的局部 Riemann 解作为 (4.1) 的局部 Riemann 近似解。

**‡ 论题 4.7.** 利用 Godunov 方法, 构造非线性双曲守恒律 (4.1) 的 Roe 型迎风格式。

答: 在线性双曲型方程 (4.43) 中, 定义

$$A_{j+1/2}^n = f'(\{\bar{u}\}_{j+1/2}^n), \quad (4.44)$$

计算相应的局部 Riemann 解, 利用 EA 过程即可给出迎风格式 (4.32)。推导过程类似, 略。

更加有效的局部线性化方式是将 (4.43) 中的  $A_{j+1/2}$  定义为单元界面的 Roe 平均值, 即

$$A_{j+1/2}^n = \begin{cases} \frac{f(\bar{u}_{j+1}^n) - f(\bar{u}_j^n)}{\bar{u}_{j+1}^n - \bar{u}_j^n}, & \bar{u}_{j+1}^n = \bar{u}_j^n, \\ f'(\bar{u}_j^n), & \text{否则}. \end{cases} \quad (4.45)$$

换言之, Roe 平均值满足离散版本的 RH 跳跃条件

$$A_{j+\frac{1}{2}}^n [\bar{u}_{j+1}^n - \bar{u}_j^n] = f(\bar{u}_{j+1}^n) - f(\bar{u}_j^n), \quad (4.46)$$

可以较为准确地刻画激波速度。(4.46) 也称为 Roe 平均条件<sup>5</sup>。此时, 利用 EA 过程即可导出著名的 Roe 型迎风格式, 即迎风格式 (4.32) 中的  $A_{j+1/2}^n$  被替换为 Roe 平均值 (4.45)。相应的格式称为 Roe 型迎风格式。□

当真解充分光滑时, Roe 平均值同算术平均值的差距非常小。它们具有形式上的二阶误差, 即

$$A_{j+\frac{1}{2}}^n - f'(\{\bar{u}\}_{j+\frac{1}{2}}^n) = \mathcal{O}(|\bar{u}_{j+1}^n - \bar{u}_j^n|^2).$$

但是, 当遇到间断解时, 两者将产生明显的区别。

#### 4.4.2 REA 过程

在进行局部推进和单元平均操作之前, 增加一个高阶重构的操作, 则 EA 过程升级到 REA 过程。换言之, 利用有限个工作单元的均值信息, 将目标单元的常值函数提升到多项式函数, 改善局部 Riemann 问题的逼近效果, 从而提升 Godunov 方法的相容阶。

下面以分片线性多项式为例。重构函数的单元均值应保持不变, 每个工作单元的线性多项式可以定义为

$$u(x, t^n) = \bar{u}_j^n + \sigma_j^n \frac{x - x_j}{\Delta x}, \quad x \in I_j, \quad (4.47)$$

其中  $\sigma_j^n$  是需要重构的广义斜率, 可以利用各种方式构造, 例如

$$\sigma_j^n = \bar{u}_{j+1}^n - \bar{u}_j^n; \quad (4.48a)$$

$$\sigma_j^n = \bar{u}_j^n - \bar{u}_{j-1}^n; \quad (4.48b)$$

$$\sigma_j^n = (\bar{u}_{j+1}^n - \bar{u}_{j-1}^n)/2. \quad (4.48c)$$

若  $\sigma_j^n \equiv 0$ , 则 REA 过程退化为 EA 过程。

---

<sup>5</sup>P. L. Roe, *Approximate Riemann solvers, parameter vectors, and difference schemes*, J. Comput. Phys., 43 (1981), 357-372

**¶ 论题 4.8.** 设双曲守恒律 (4.1) 是线性的, 即  $f(u) = au$ , 其中  $a$  是给定的常数. 按照 (4.48a) 定义线性多项式的广义斜率, 构造相应的 REA 过程.

答: 在重构分片线性多项式之后, 局部 Riemann 问题的初值条件 (4.35) 重新定义为

$$\begin{aligned} u_{j+\frac{1}{2}}^- &= \bar{u}_j^n + \frac{1}{2}\sigma_j^n = \frac{1}{2}\bar{u}_j^n + \frac{1}{2}\bar{u}_{j+1}^n, \\ u_{j+\frac{1}{2}}^+ &= \bar{u}_{j+1}^n - \frac{1}{2}\sigma_{j+1}^n = \frac{3}{2}\bar{u}_{j+1}^n - \frac{1}{2}\bar{u}_{j+2}^n. \end{aligned} \quad (4.49)$$

相应的局部 Riemann 解是

$$u_{j+\frac{1}{2}}(x, t^{n+1}) = \begin{cases} \bar{u}_j^n + \frac{1}{2}(\bar{u}_{j+1}^n - \bar{u}_j^n), & \text{当 } x < x_{j+\frac{1}{2}} + a\Delta t, \\ \bar{u}_{j+1}^n - \frac{1}{2}(\bar{u}_{j+2}^n - \bar{u}_{j+1}^n), & \text{其他.} \end{cases}$$

若  $|a|\Delta t \leq \Delta x$ , 则相邻的两个局部 Riemann 解没有发生双值冲突. 计算单元均值, 可得

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \frac{1}{2}\nu \left[ \bar{u}_{j+1}^n - \bar{u}_{j-1}^n \right] + \frac{1}{2}\nu^2 \left[ \bar{u}_{j+1}^n - 2\bar{u}_j^n + \bar{u}_{j-1}^n \right].$$

若将均值视为点值, 它就是 LW 格式(3.4). □

类似地, 基于同样的分片线性多项式重构技术, REA 方法可以导出非线性双曲守恒律 (4.1) 的 LW 格式 (4.34), 其中近似 Riemann 解由 (4.43) 给出,  $A_{j+1/2}^n$  由 (4.44) 给出. 若  $A_{j+1/2}^n$  由 (4.45) 给出, 相应的 LW 格式称为 Roe 型 LW 格式.

## 4.5 稳定性和收敛性

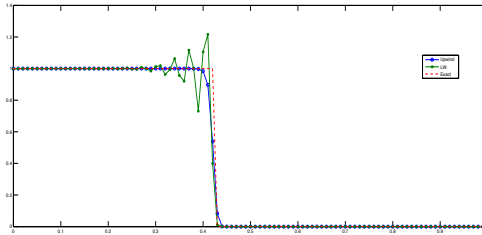
对于非线性双曲守恒律, 守恒型格式在某种程度上可以确保数值解收敛到弱解, 给出相对理想的数值结果, 但是高阶相容和数值震荡的矛盾依旧没有得到解决.

以 Burgers 方程为代表, 考察迎风格式 (4.32) 和 LW 格式 (4.12) 的数值表现. 由图 4.4 可知:

1. LW 格式 (4.12) 具有二阶局部截断误差<sup>6</sup>，在间断界面附近出现严重的数值震荡。即使网格不断加密，数值解的（上下）溢出现象也不会减弱。
2. 迎风格式 (4.32) 具有一阶局部截断误差，在间断界面附近没有出现数值震荡。但是，数值过渡区间包含较多的计算单元，相应的数值间断界面略显平坦。

因此，要提高守恒型格式的计算效果，高阶格式需改善间断间断界面的光滑度和陡峭度。

图 4.4: Burgers 方程的迎风格式和 LW 格式



### 4.5.1 单调保持格式

双曲守恒律具有单调保持性质：若初值是单增或单减函数，则熵解在任意时刻均保持相同的单调性。理想的数值格式应数值保持这个性质。

🌀 **定义 4.4.** 只要数值初值是单调的，则任意时刻的数值解均具有相同的单调性。具有上述性质的格式称为单调保持格式。

由定义可知，当数值初值和真实初值具有相同的单增（或单减）属性时，单调保持格式的数值解和双曲守恒律的真解将保持相同的单调性，非单调保

---

<sup>6</sup>指问题真解充分光滑的时候。

持格式必将出现错误的单调性刻画, 形成虚假的数值震荡。因此说, 避免数值震荡现象的前提是单调保持格式。

数值格式是否属于单调保持格式, 通常是难以验证的。

### 4.5.2 单调格式

相对于单调保持性质, 双曲守恒律 (4.1) 还具有更强的比较性质:

$$v(x, 0) \gtrless u(x, 0), \forall x \Rightarrow v(x, t) \gtrless u(x, t), \forall x, \forall t > 0,$$

其中  $u(x, t)$  和  $v(x, t)$  都是熵解。因此, 数值格式也应继承这个性质, 即数值解  $u$  和  $v$  满足

$$v_j^n \gtrless u_j^n, \forall j \Rightarrow v_j^{n+1} \gtrless u_j^{n+1}, \forall j. \quad (4.50)$$

对于线性格式, 前面介绍过的单调格式概念已经实现了这个目标。下面将其推广到非线性格式。

⊙ **定义 4.5.** 设  $l$  和  $r$  是给定的左右臂长。称数值格式

$$u_j^{n+1} = H(u_{j-l}^n, u_{j-l+1}^n, \dots, u_j^n, \dots, u_{j+r-1}^n, u_{j+r}^n) \quad (4.51)$$

是**单调格式**, 如果函数  $H$  关于每个变元都是非减的。

若函数  $H$  是可微的, 则定义 (4.5) 可以获得简化。换言之, 若  $H$  的一阶偏导数都是非负的, 则数值格式 (4.51) 是单调格式。

⌞ **论题 4.9.** 在相应的 CFL 条件下, Lax 格式 (4.11) 是单调格式。

答: 将 Lax 格式写成 (4.51) 的形式, 相应的函数是

$$H(a, b, c) = \frac{1}{2} [a + \nu f(a)] + \frac{1}{2} [c - \nu f(c)].$$

当 CFL 条件  $\nu \max_{\forall u} |f'(u)| \leq 1$  成立时,  $H$  的一阶偏导数

$$H'(a) = \frac{1}{2} [1 + \nu f'(a)], \quad H'(b) = 0, \quad H'(c) = \frac{1}{2} [1 - \nu f'(c)]$$

都是非负的。换言之，Lax 格式是单调格式。  $\square$

单调保持格式可以不是单调格式，但是**单调格式必然是单调保持格式**。局限于线性格式的范畴，单调保持格式和单调格式是完全等价的两个概念。

**↯ 论题 4.10.** LW 格式 (4.12) 不是单调格式。

**答：**对于线性问题，结论是显然的。对于非线性问题，论证过程是类似的。构造具体实例，或者参见图 4.4，可以说明 LW 格式不具备单调保持性质。因此，它不是单调格式。  $\square$

Harten、Hyman 和 Lax 已经证明<sup>7</sup>：**单调格式的数值解一致有界，必然收敛到双曲守恒律的熵解**。但是，Godunov 定理依旧成立：**单调格式至多具有一阶局部截断误差**。换言之，在相容阶方面，非线性单调格式同线性单调格式没有差别。类似地，非线性单调格式通常也会抹平数值间断界面，产生较宽的数值过渡区间。

**↯ 论题 4.11.** 在相应的 CFL 条件下，Roe 型迎风格式

$$u_j^{n+1} = u_j^n - \frac{\nu}{2} \left\{ \left[ 1 - \operatorname{sgn} A_{j+\frac{1}{2}}^n \right] \Delta_{+x} f_j^n + \left[ 1 + \operatorname{sgn} A_{j-\frac{1}{2}}^n \right] \Delta_{-x} f_j^n \right\}$$

不是单调格式，其中  $A_{j+1/2}$  是 Roe 平均值 (4.45)。

**答：**以 Burgers 方程的 Riemann 问题为例。若初值是  $u_- = -1$  和  $u_+ = 1$ ，则相应的真解是稀疏波。定义空间网格  $\{x_j = j\Delta x\}_{j \in \mathbb{Z}}$  和相应的数值初值

$$u_j^0 = 1, \quad j \geq 0; \quad u_j = -1, \quad j < 0.$$

简单计算可知，Roe 型迎风格式的数值解保持不变。换言之，数值解仅仅收敛到某个弱解而已，没有收敛到问题的熵解。因此，前面的理论结果说明，守恒型 Roe 迎风格式不是单调格式。  $\square$

---

<sup>7</sup> A. Harten, J. M. Hyman and P. D. Lax, *On finite-difference approximations and entropy conditions for shocks*, Comm. Pure Appl. Math., 29 (1976), 297-322



### 4.5.3 TVD 格式

设  $v: \mathfrak{R} \rightarrow \mathfrak{R}$  是 Lebesgue 可测函数, 其震荡强度可以用连续 (型) 函数的全变差<sup>8</sup>

$$\mathrm{TV}(v) = \limsup_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \int_{-\infty}^{\infty} |v(x) - v(x - \varepsilon)| dx = \int_{-\infty}^{\infty} |v'(x)| dx$$

来描述。可以理论证明, 双曲守恒律 (4.1) 的熵解满足全变差不增, 即

$$\mathrm{TV}(u(x, t_2)) \leq \mathrm{TV}(u(x, t_1)), \quad t_2 > t_1$$

基于这个性质的数值保持, 下面的重要概念被提出。

⊙ 定义 4.6. 称数值格式是全变差不增 (TVD) 的, 若它的数值解恒满足

$$\mathrm{TV}(u^{n+1}) \leq \mathrm{TV}(u^n), \quad \forall n$$

其中  $\mathrm{TV}(u^n) = \sum_j |u_{j+1}^n - u_j^n|$ 。

引理 4.1. (Harten 引理) 设数值格式具有增量形式<sup>9</sup>

$$u_j^{n+1} = u_j^n - C_{j-\frac{1}{2}} \left[ u_j^n - u_{j-1}^n \right] + D_{j+\frac{1}{2}} \left[ u_{j+1}^n - u_j^n \right], \quad (4.52)$$

且处处成立

$$C_{j+1/2} \geq 0, \quad D_{j+1/2} \geq 0, \quad C_{j+1/2} + D_{j+1/2} \leq 1, \quad (4.53)$$

则它是 TVD 格式。

证明: 证明是简单的。利用三角不等式和 (4.53) 可知

$$\begin{aligned} \mathrm{TV}(u^{n+1}) &\leq \sum_j \left[ 1 - C_{j+\frac{1}{2}} - D_{j+\frac{1}{2}} \right] |u_{j+1}^n - u_j^n| \\ &\quad + \sum_j C_{j-\frac{1}{2}} |u_j^n - u_{j-1}^n| + \sum_j D_{j+\frac{3}{2}} |u_{j+2}^n - u_{j+1}^n|. \end{aligned}$$

<sup>8</sup>这里的导数应当理解为分布导数, 例如间断函数的导数是  $\delta$  函数。

<sup>9</sup>增量系数  $C$  和  $D$  也可以依赖数值解。

平移最后两项的求和指标, 即证  $\text{TV}(u^{n+1}) \leq \text{TV}(u^n)$ .  $\square$

**论题 4.12.** 在相应的 CFL 条件下, Roe 型迎风格式 (参见论题 4.11) 是 TVD 格式。

答: 将 Roe 型迎风格式改写为等价的增量形式 (4.52), 其中

$$C_{j+\frac{1}{2}} = \frac{\nu}{2} \left[ 1 + \text{sgn} A_{j+\frac{1}{2}}^n \right] \frac{\Delta_{+x} f_j^n}{\Delta_{+x} u_j^n},$$

$$D_{j+\frac{1}{2}} = -\frac{\nu}{2} \left[ 1 - \text{sgn} A_{j+\frac{1}{2}}^n \right] \frac{\Delta_{+x} f_j^n}{\Delta_{+x} u_j^n}.$$

注意到 Roe 平均值的定义 (4.46), 可知  $C_{j+1/2}$  和  $D_{j+1/2}$  都是非负的。当 CFL 条件  $\nu \max_u |f'(u)| \leq 1$  成立时, 有

$$C_{j+\frac{1}{2}} + D_{j+\frac{1}{2}} = \nu \text{sgn} A_{j+\frac{1}{2}}^n \frac{\Delta_{+x} f_j^n}{\Delta_{+x} u_j^n} = \nu |A_{j+\frac{1}{2}}^n| \leq 1.$$

利用 Harten 引理可知, Roe 型迎风格式是 TVD 的。  $\square$

**单调格式是 TVD 格式, 且 TVD 格式是单调保持格式。**但是, **逆命题是不成立的。**例如, Roe 型迎风格式只是 TVD 格式, 不是单调格式。**局限于线性差分格式的范畴, 单调格式、TVD 格式和单调保持格式是彼此等价的。**利用 Godunov 定理可知, 线性的 TVD 格式至多具有一阶局部截断误差。换言之, 高阶的 TVD 格式必须是非线性的, 即使离散对象是线性的双曲守恒律。

**论题 4.13.** Roe 型 LW 格式既不是 TVD 格式, 也不是单调格式。

答: 既然它不是单调保持格式, 结论是显然的。  $\square$

由论题 4.11 和 4.12 可知, Roe 型迎风格式是守恒型 TVD 格式, 不是单调格式。虽然它避免了剧烈的数值震荡, 数值解却可能收敛到非熵解的某个弱解。究其原因是, 计算过程中遇到的所有间断结构被均武断地归结为激波, 完全忽略了稀疏波的存在性。要想走出上述困境, 数值计算需要区别激波和稀

疏波两种结构。换言之，在局部 Riemann 解的近似过程，或者数值通量的定义中，引进适当的“熵修正”技术是非常必要的。典型工作是带有熵修正的 Roe 型数值通量或 Engquist-Osher 数值通量<sup>10</sup>

$$\begin{aligned}\hat{f}_{j+\frac{1}{2}}^n = & \frac{1}{2} \left[ 1 + \operatorname{sgn} f'(u_j^n) \right] f(u_j^n) + \frac{1}{2} \left[ 1 - \operatorname{sgn} f'(u_{j+1}^n) \right] f(u_{j+1}^n) \\ & + \frac{1}{2} \left[ \operatorname{sgn} f'(u_{j+1}^n) - \operatorname{sgn} f'(u_j^n) \right] f(u_s),\end{aligned}\quad (4.54)$$

其中  $u_s$  满足  $f'(u_s) = 0$ ，称为声波点。若  $f(u)$  是可微的凸函数，可证：在相应的 CFL 条件下，带有熵修正的 Roe 型迎风格式是 TVD 格式。事实上，它是单调格式，可以保证数值解收敛到熵解。

## 4.6 TVD 修正技术

所谓的 TVD 修正技术，就是利用恰当的非线性限制器技术，**将高阶（但震荡）格式和低阶（但单调）格式动态地结合起来**，进而建立双曲守恒律的高精度高分辨率格式。

### 4.6.1 数值通量修正技术

原始思想可以追溯到**人工黏性法或者人工跳转方法**：**若数值解被判断为“局部间断”的，则局部增加数值黏性，压制附近的数值震荡。**通常，低阶（但单调）格式具有较强的数值黏性，高阶（非单调）格式具有较弱的数值黏性。对于守恒型格式，数值黏性的差异通过数值通量来体现。因此，上述操作等价于高阶格式和低阶格式的自动跳转，或者高阶数值通量和低阶数值通量的自动跳转。

---

<sup>10</sup>B. Engquist and O. Osher, *One-sided difference approximations for nonlinear conservation laws*, Math. Comp., 36(1981), 321-352

已知双曲守恒律 (4.1) 的高阶格式和低阶格式，分别记为

$$u_j^{n+1} = u_j^n - \nu \left[ (\hat{f}_{\mathcal{H}})^n_{j+\frac{1}{2}} - (\hat{f}_{\mathcal{H}})^n_{j-\frac{1}{2}} \right], \quad (4.55a)$$

$$u_j^{n+1} = u_j^n - \nu \left[ (\hat{f}_{\mathcal{L}})^n_{j+\frac{1}{2}} - (\hat{f}_{\mathcal{L}})^n_{j-\frac{1}{2}} \right], \quad (4.55b)$$

其中  $\hat{f}_{\mathcal{H}}$  称为高阶数值通量， $\hat{f}_{\mathcal{L}}$  称为低阶数值通量。它们可以是守恒型差分格式，也可以是守恒型体积格式。基于数值通量修正技术，定义新的守恒型格式

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} \left[ (\hat{f}_{\mathcal{M}})^n_{j+\frac{1}{2}} - (\hat{f}_{\mathcal{M}})^n_{j-\frac{1}{2}} \right], \quad (4.56a)$$

相应的数值通量是

$$(\hat{f}_{\mathcal{M}})^n_{j+\frac{1}{2}} = \theta_{j+\frac{1}{2}}^n (\hat{f}_{\mathcal{H}})^n_{j+\frac{1}{2}} + (1 - \theta_{j+\frac{1}{2}}^n) (\hat{f}_{\mathcal{L}})^n_{j+\frac{1}{2}}. \quad (4.56b)$$

这里， $\theta_{j+1/2}^n$  是适当构造的**开关函数或者通量限制器**，可以实现高阶数值通量和低阶数值通量之间的自动跳转。具体来说，数值格式要实现下述功能：

- 若数值解被判断为局部“光滑”的，则开关函数趋向一，相应格式趋于高阶格式；
- 若数值解被判断为局部“间断”的，则开关函数趋向零，相应格式趋于低阶格式。

由 Godunov 定理可知，若要 (4.56) 是高阶的 TVD 格式，相应的通量限制器一定是非线性的。典型的例子是<sup>11</sup>

$$\theta_{j+\frac{1}{2}}^n \equiv \max(0, \min(1, s_{j+\frac{1}{2}}^n)), \quad (4.57a)$$

其中  $A_{j+1/2}^n$  是界面位置的 Roe 平均值，

$$s_{j+\frac{1}{2}}^n = \begin{cases} (u_j^n - u_{j-1}^n)/(u_{j+1}^n - u_j^n), & \text{当 } A_{j+\frac{1}{2}}^n \geq 0; \\ (u_{j+2}^n - u_{j+1}^n)/(u_{j+1}^n - u_j^n), & \text{当 } A_{j+\frac{1}{2}}^n < 0. \end{cases} \quad (4.57b)$$

<sup>11</sup> A. Harten and G. Zwas, *Self-adjusting hybrid schemes for shock computations*, J. Comput. Phys., 9 (1972), 568-583

若高阶格式是 Roe 型 LW 格式, 低阶格式是 Roe 型迎风格式, 则守恒型差分格式 (4.56) 是 (形式上) 二阶的 TVD 格式。证明过程是 Harten 引理的直接应用; 因篇幅有限, 详略。

### 4.6.2 数值斜率修正技术

数值斜率修正技术主要用于有限体积格式, 相应的开创性研究工作是 MUSCL 格式<sup>12</sup>。它基于 LW 格式的 REA 过程, 其主要贡献是分片线性函数

$$u_j^n = \bar{u}_j^n + \sigma_j^n \frac{x - x_j}{\Delta x}, \quad x \in I_j \quad (4.58)$$

的广义斜率重构过程。

为简单起见, 以线性双曲守恒律 (4.1) 为例, 即  $f(u) = au$ , 其中  $a$  是给定的正常数。在 LW 格式中, 广义斜率定义为  $\sigma_j^n = \bar{u}_{j+1}^n - \bar{u}_j^n$ 。它主要产生两个作用: 其一是提高光滑解区域的相容阶, 其二是导致间断界面附近出现数值震荡。若数值解已经被判定为“局部间断”, 则削减目标单元的广义斜率, 降低局部区域的数值震荡强度。比如, 借用 minmod 限制器

$$\begin{aligned} & \text{minmod}(a_1, a_2, a_3) \\ &= \begin{cases} s \min\{|a_1|, |a_2|, |a_3|\}, & s = \text{sgn}(a_1) = \text{sgn}(a_2) = \text{sgn}(a_3), \\ 0, & \text{其它,} \end{cases} \end{aligned}$$

将 (4.58) 的广义斜率定义为

$$\sigma_j^n = \text{minmod}\left(\bar{u}_{j+1}^n - \bar{u}_j^n, \bar{u}_j^n - \bar{u}_{j-1}^n, \frac{1}{2}(\bar{u}_{j+1}^n - \bar{u}_{j-1}^n)\right). \quad (4.59)$$

利用 Harten 引理, 可以证明: 在相应的 CFL 条件下, MUSCL 格式是 TVD 的; 详略。

下面阐述斜率限制器 (4.59) 的工作机制。若它的返回值等于第一个参数, 则 MUSCL 格式退化为 LW 格式, 具有二阶局部截断误差。否则, 将有以下两种情况发生:

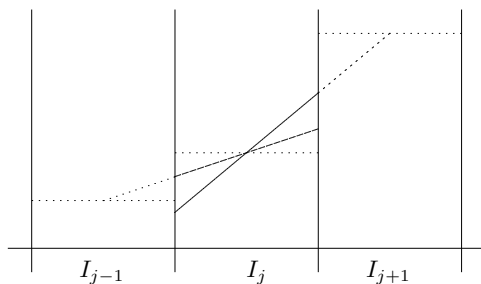
---

<sup>12</sup>van Leer, *Towards the ultimate conservative difference scheme V. a second order sequel to Godunov's method*, J. Comput. Phys., 32 (1976), 101–136

1. 当三个参数的符号不同时,相邻三个单元的均值有增有减。此时,minmod 函数认为真解含有间断界面,导致局部区域出现数值震荡现象。为削弱数值震荡的强度,minmod 函数将目标单元的广义斜率直接清零,相应的 MUSCL 格式局部退化为单调的迎风格式。
2. 当三个参数符号相同时,相邻三个单元的均值是局部单调的。此时,minmod 函数判定局部区域的真解是光滑和单调的。但是,若目标单元的数值解在重构之后,相应的高次多项式(实斜线)超出两侧单元的均值,则 minmod 函数会认为局部区域产生了较弱的数值震荡;参见图 4.5。此时,它自动调整目标单元的广义斜率,使得校正解(虚斜线)整体落在两侧单元均值之间,从而改善了局部区域的数值震荡现象。

换言之,斜率限制器(4.59)自动压制目标单元的广义斜率,使得 MUSCL 格式呈现出理想的数值效果。

图 4.5: 斜率限制器: 实线是修正前,虚线是修正后。



数值经验表明: 对于一维非线性双曲守恒律而言,基于 TVD 修正技术的上述高分辨率格式是成功和有效的。但是,不足之处依旧然存在,例如:

1. 非线性限制器技术还需深入和完善。例如,基于 minmod 函数的斜率限制器,常常将具有极值点的光滑解误判为间断解。由于三个参数异号,

`minmod` 函数的返回值是零，MUSCL 格式退化为单调格式，局部截断误差变为一阶，TVD 格式的高阶相容性遭到了破坏。目前，许多新型的非线性限制器被相继提出，正在努力减少或者避免“误判和漏判”现象的出现。

2. 可以理论证明，一维 TVD 格式至多具有二阶局部截断误差；但是，二维 TVD 格式至多具有一阶局部截断误差，无法实现高精度和高分辨率。为解决上述困难，全变差有界（TVB）格式、本质不震荡（ENO）格式和加权本质不震荡（WENO）格式等新型算法被相继提出，并取得了良好的数值效果。

简而言之，非线性双曲守恒律的数值方法极具挑战性，现在还有很多的关键问题没有解决和完善。

---

## 第 5 章

# 边界条件的数值离散方法

---

### 5.1 一维扩散方程的含导数边界条件

考虑热传导方程 (2.1) 的混合边值问题 (HX), 相应的边界条件是

$$-au_x(0, t) + \sigma u(0, t) = \phi_0(t), \quad t \in (0, T], \quad (5.1a)$$

$$u(1, t) = \phi_1(t), \quad t \in (0, T], \quad (5.1b)$$

其中  $\phi_0(t)$  和  $\phi_1(t)$  是已知函数,  $\sigma \geq 0$  是给定常数,  $T > 0$  是终止时刻。为简单起见, 设时空网格是等距的, 热传导方程 (2.1) 采用简单的全显格式或全隐格式进行离散。对于本质边界条件 (5.1b), 数值离散是简单的, 只需将其设置为空间网格点, 进行简单赋值即可。但是, 对于自然边界条件 (5.1a), 我们需要解决两个关键问题。其一是边界导数的差商离散方式, 其二是差分格式的相容性、稳定性<sup>1</sup>和收敛性概念是否受到影响。

#### 5.1.1 单侧离散方式

对于自然边界条件, 单侧离散方式是最直接的选择。此时, 自然边界点也要设置为空间网格点。因此, 适用于混合边界条件 (5.1) 的等距空间网格是

$$\mathcal{T}_{\Delta x}^{(1)} = \{x_j = j\Delta x\}_{j=0}^J, \quad (5.2)$$

其中  $J$  是给定的正整数,  $\Delta x = 1/J$  是空间步长。

---

<sup>1</sup>即使差分格式面对纯初值问题或周期边值问题是稳定的, 但是它可能受到数值边界条件的影响而变得不再稳定。在抛物型方程的差分方法中, 边界条件的数值离散通常具有较弱的破坏程度。若无特别需要, 我们略过相关内容的讨论。



在  $t^n = n\Delta t$  时刻, 本质边界条件 (5.1b) 的差分方程是

$$u_J^n = \phi_1(t^n). \quad (5.3)$$

既然在端点  $x = 0$  的左侧没有其它网格点, 我们自然选用单侧差商离散边界导数, 得到自然边界条件 (5.1a) 的差分方程

$$-a \frac{u_1^n - u_0^n}{\Delta x} + \sigma u_0^n = \phi_0(t^n). \quad (5.4)$$

若以全显格式离散热传导方程 (2.1), 则位于内部 (空间) 网格点的差分方程是

$$u_j^{n+1} = u_j^n + \mu a \delta_x^2 u_j^n, \quad j = 1 : J - 1. \quad (5.5)$$

将其同数值边界条件 (5.3) 和 (5.4) 联立, 即可得到一个封闭的离散系统。通常, 称其为模型问题 (HX) 的全显格式。

模型问题 (HX) 的全隐格式可类似定义。换言之, 以全隐格式离散热传导方程 (2.1), 则位于内部 (空间) 网格点的差分方程是

$$u_j^{n+1} = u_j^n + \mu a \delta_x^2 u_j^{n+1}, \quad j = 1 : J - 1. \quad (5.6)$$

将其同数值边界条件 (5.3) 和 (5.4) 联立, 所得的封闭离散系统称为模型问题 (HX) 的全隐格式。

在上述两个格式中, 不同网格点的差分方程具有不同的属性。当网格点远离边界时, 差分方程同边界条件无关, 其局部截断误差的推导和结论都是明确的, 逐点相容性的概念和结论没有变化; 前面章节已经讨论过, 此处无需赘述。我们只需关注那些位于边界附近的有限个差分方程, 指出它们受到数值边界条件影响而产生的相容性变化。

**↓ 论题 5.1.** 若采用单侧离散方式处理自然边界条件, 古典格式在  $x_1$  点的局部截断误差是  $\mathcal{O}(1)$  的。

答: 以全显格式为例。差分方程在  $x_1$  点的局部描述是

$$\frac{u_1^{n+1} - u_1^n}{\Delta t} = a \frac{u_2^n - 2u_1^n}{(\Delta x)^2} + \frac{\tilde{\sigma}}{\Delta x} \left[ \frac{au_1^n}{\Delta x} + \phi_0(t^n) \right], \quad (5.7)$$

相应的局部截断误差是

$$\tau_1^n \equiv \frac{[u]_1^{n+1} - [u]_1^n}{\Delta t} - a \frac{[u]_2^n - 2[u]_1^n}{(\Delta x)^2} - \frac{\tilde{\sigma}}{\Delta x} \left[ \frac{a[u]_1^n}{\Delta x} + \phi_0(t^n) \right], \quad (5.8)$$

其中  $[u]$  满足模型问题 (HX)。利用 Taylor 展开技术, 即可得到相应的局部截断误差阶。等价的简化推导过程如下。回顾差分方程 (5.7) 的生成过程可知, 它源于数值边界条件 (5.4) 和当  $j = 1$  时差分方程 (5.5) 的线性组合。前者的离散对象是自然边界条件, 后者的离散对象是偏微分方程, 相应的局部截断误差分别是

$$\tau_{\text{pde}} = \frac{[u]_1^{n+1} - [u]_1^n}{\Delta t} - a \frac{\delta_x^2 [u]_1^n}{(\Delta x)^2} = \mathcal{O}((\Delta x)^2 + \Delta t), \quad (5.9a)$$

$$\tau_{\text{bry}} = -a \frac{[u]_1^n - [u]_0^n}{\Delta x} + \sigma[u]_0^n - \phi_0(t^n) = \mathcal{O}(\Delta x). \quad (5.9b)$$

沿用同样的组合方式, 在 (5.9b) 的两侧乘以  $\tilde{\sigma}/\Delta x$ , 同 (5.9a) 相加, 即可消去边界点信息  $[u]_0^n$ , 得到局部截断误差 (5.8) 和相应的估计

$$\tau_1^n = \tau_{\text{pde}} + \frac{\tilde{\sigma}}{\Delta x} \tau_{\text{bry}} = \mathcal{O}(1).$$

因此说, 全显格式在  $x_1$  点的差分方程是不相容的, 同其它位置的整体二阶相容是不匹配的。□

上述分析过程表明: 要改善差分方程在边界附近网格点的相容性, 数值边界条件 (特别是边界导数) 的相容阶必须得到相应的提高。常用的边界导数离散策略有两种: 其一是扩大离散模版的宽度, 建立高阶相容的单侧离散, 例如

$$[u_x]_0^n = \frac{1}{2\Delta x} \left[ 3[u]_2^n - 4[u]_1^n + [u]_0^n \right] + \mathcal{O}(\Delta x)^2;$$

其二是利用对称模版的数值优势, 基于特殊结构的空间网格, 建立边界导数的双侧离散。

### 5.1.2 双侧离散方式

虚拟点 (ghost point) 方法和半网格 (offset mesh) 方法的主要区别是空间网格的设置方式。下面以全显格式为例, 说明两种方法的实现过程。

### 虚拟点方法

在虚拟点方法中,自然边界点要设置为空间网格点。整个空间网格是 (5.2) 的拓展,即在计算区域的外部增加少量的辅助网格点。具体来说,适用于混合边界条件 (5.1) 的等距空间网格是

$$\mathcal{T}_{\Delta x}^{(2)} = \{x_j = j\Delta x\}_{j=-1}^J, \quad (5.10)$$

其中  $J$  是给定的正整数,  $\Delta x = 1/J$  是空间步长。辅助网格点  $x_{-1}$  位于计算区域之外,故被称为虚拟网格点。实际上,它就是网格点  $x_1$  关于空间边界  $x = x_0$  的镜像对称点。

关于本质边界条件 (5.1b),相应的差分方程依旧定义为 (5.3)。至于自然边界条件 (5.1a),以边界  $x = x_0$  为离散焦点,利用一步中心差商离散边界导数,可得

$$-a \frac{u_1^n - u_{-1}^n}{2\Delta x} + \sigma u_0^n = \phi_0(t^n). \quad (5.11)$$

显然,它具有二阶(空间)相容性。假设热传导方程 (2.1) 可以拓展到区域外侧,将  $x_0$  视为空间内点,得到显式离散的差分方程

$$u_j^{n+1} = u_j^n + \mu a \delta_x^2 u_j^n, \quad j = 0 : J - 1. \quad (5.12)$$

综上所述,模型问题 (HX) 的全显格式定义完毕。

**↓ 论题 5.2.** 若采用虚拟点方法处理自然边界条件,全显格式在网格点  $x_0$  处的局部截断误差是  $\mathcal{O}(\Delta x + \Delta t)$ 。

### 半网格方法

在半网格方法中,自然边界点要设置在空间网格点的正中间。因此,要兼顾本质边界点  $x = 1$  也是一个网格点,适用于混合边界条件 (5.1) 的等距空间网格定义为

$$\mathcal{T}_{\Delta x}^{(3)} = \left\{ x_{j-\frac{1}{2}} = \left( j - \frac{1}{2} \right) \Delta x \right\}_{j=0}^J, \quad (5.13)$$

其中  $J$  是给定的正常数,  $\Delta x = 2/(2J+1)$  是空间步长。由于空间网格点采用半点方式进行编号, 基于此网格的边界离散方法常常称为半网格方法。当然, 空间网格点采用整点进行编号, 也是可以的。

关于本质边界条件 (5.1b), 相应的差分方程是

$$u_{J-\frac{1}{2}}^n = \phi_1(t^n). \quad (5.14)$$

至于自然边界条件 (5.1a), 以真实边界  $x = x_0$  为离散焦点, 利用半步中心差商离散边界导数, 利用算术平均技术离散边界点值, 可得二阶 (空间) 相容的差分方程

$$-\frac{a}{\Delta x} \left( u_{\frac{1}{2}}^n - u_{-\frac{1}{2}}^n \right) + \frac{\sigma}{2} \left( u_{-\frac{1}{2}}^n + u_{\frac{1}{2}}^n \right) = \phi_0(t^n), \quad (5.15)$$

其中  $u_{-1/2}^n$  也是虚拟点值。同 (5.11) 相比, 它的离散模版更为紧凑。假设热传导方程 (2.1) 可以拓展到区域外侧, 将  $x_{1/2}$  视为空间内点, 得到显式离散的差分方程

$$u_{j+\frac{1}{2}}^{n+1} = u_{j+\frac{1}{2}}^n + \mu a \delta_x^2 u_{j+\frac{1}{2}}^n, \quad j = 0 : J-1. \quad (5.16)$$

综上所述, 模型问题 (HX) 的全显格式定义完毕。

**论题 5.3.** 若采用半网格方法处理自然边界条件, 全显格式在网格点  $x_{1/2}$  的局部截断误差是  $\mathcal{O}(\Delta x + \Delta t)$ 。

## 5.2 二维扩散方程的边界条件离散

设  $\Omega$  是二维有界区域, 考虑热传导方程

$$u_t = \Delta u = u_{xx} + u_{yy}, \quad (x, y) \in \Omega \quad (5.17)$$

初边值问题。本节以全隐格式为例, 介绍 Dirichlet 边界条件和 Riemann 边界条件的数值离散方法。

当空间区域形状复杂时, 相对完美的空间网格<sup>2</sup>通常是难以构造的。事实上, 网格设计是一项精细的前期准备工作, 需要花费巨大的精力。因篇幅限制, 本节跳过这个主题的讨论, 直接利用二维平面的正方形网格

$$\mathcal{M}_h = \{(x_j, y_k) : x_j = x_0 + jh, y_k = y_0 + kh\}_{j=-\infty:+\infty, k=-\infty:+\infty}$$

在空间区域  $\Omega$  的适当剪裁, 给出相应的离散网格

$$\bar{\Omega}_h \equiv \Omega_h \cup \Gamma_h. \quad (5.18)$$

具体含义如下:

1. 若  $\mathcal{M}_h$  的网格点落在  $\Omega$  内, 则它称为网格内点。其全体构成网格内点集  $\Omega_h \equiv \mathcal{M}_h \cap \Omega$ 。

网格内点分为两类。若相应位置的差分方程同边界信息无关, 则称其为**规则内点**; 否则, 称其为**非规则内点**。

2.  $\mathcal{M}_h$  的网格线同区域边界  $\Gamma = \partial\Omega$  的交点, 称为网格边界点。其全体构成网格边界点集  $\Gamma_h$ 。请注意, 网格边界点不一定是网格点。

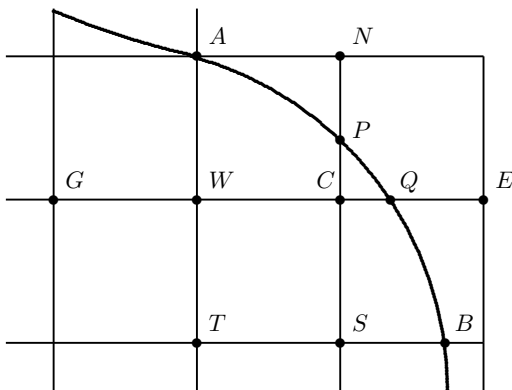
图 5.1 给出了二维空间网格的局部结构示意图, 其中  $\widehat{APQB}$  是位于东北角的边界曲线。显然,  $A, B, P$  和  $Q$  是网格边界点,  $G, T, W, C$  和  $S$  是网格内点。当采用全隐格式离散偏微分方程时,  $G$  和  $T$  是规则内点,  $W, C$  和  $S$  是非规则内点。

要构造初边值问题的差分格式, 只需在所有网格内点写出相应的差分方程。在规则内点, 差分方程仅同偏微分方程有关; 相应的离散方法已经介绍过, 无需赘述。在非规则内点, 差分方程要受到边界条件的影响, 其构造过程同边界条件的具体离散相关。

下面以图 5.1 的两个非规则内点  $W$  和  $C$  为例, 介绍两类边界条件的数值离散过程。为行文简便, 我们直接用空间网格点的符号替代标号。

<sup>2</sup>离散网格的结构和质量, 例如网格点分布情况同真解的匹配程度, 或者网格形状同计算区域的匹配程度, 都会影响数值格式的计算效果。事实上, 离散网格的最佳设计是较为独立的一个研究方向, 同计算机辅助设计 (CAD) 和计算机辅助工程 (CAE) 具有紧密的联系。

图 5.1: 二维空间网格的局部示意性描述



### 5.2.1 本质边界条件

考虑 Dirichlet 边界条件或者本质边界条件

$$u|_{\widehat{APQB}} = g(x, y), \quad (5.19)$$

其中  $g$  是已知的边界函数。假定  $g$  同时间无关，仅仅是为了陈述方便。若其同时间有关，相应的处理方式是类似的。

在非规则内点  $W$  处，相邻的网格边界点  $A$  恰好也是网格点，边界条件离散是非常简单的。只需在原有的全隐格式中，直接代入已知的边界条件，即可得到  $W$  点的差分方程

$$-\mu \left[ u_G^{n+1} + u_C^{n+1} + u_T^{n+1} \right] + (1 + 4\mu) u_W^{n+1} = u_W^n + \mu g(A).$$

但是，在非规则内点  $C$  处，相邻的两个网格边界点  $P$  和  $Q$  不是网格点，边界条件离散将略显复杂。常见的实现策略有两种。

**论题 5.4.** 第一种实现策略是利用插值逼近技术，将最靠近的边界点信息直接迁移到非规则内点上。

答：在图 5.1 中，最靠近非规则内点  $C$  的网格边界点是  $Q$ 。利用常值延拓技术，定义

$$u_C^{n+1} = g(Q), \quad (5.20)$$

或者利用  $Q$  和  $W$  两点的线性插值技术，定义

$$u_C^{n+1} = \frac{s_1}{s_1 + s_3} u_W^{n+1} + \frac{s_3}{s_1 + s_3} g(Q). \quad (5.21)$$

前者的局部截断误差<sup>3</sup>是  $\mathcal{O}(h)$ 。后者的局部截断误差是  $\mathcal{O}(h^2)$ 。具体的操作过程均同偏微分方程无关。  $\square$

**论题 5.5.** 第二种实现策略是将网格边界点收录到非规则内点的离散模版中，构造热传导方程 (5.17) 的非等臂长差分方程。

答：以非规则内点  $C$  为离散焦点。参见图 5.1，四个方向的空间臂长分别记为

$$|CQ| = s_1 h, \quad |CP| = s_2 h, \quad |CW| = s_3 h, \quad |CS| = s_4 h,$$

其中  $s_1 < 1$  和  $s_2 < 1$  对应网格边界点，而  $s_3 = s_4 = 1$  对应网格内点。此时，两个二阶空间导数满足

$$\begin{aligned} [u_{xx}]_C &\approx \frac{1}{\frac{1}{2}(s_1 + s_3)h} \left[ \frac{[u]_Q - [u]_C}{s_1 h} - \frac{[u]_C - [u]_W}{s_3 h} \right], \\ [u_{yy}]_C &\approx \frac{1}{\frac{1}{2}(s_2 + s_4)h} \left[ \frac{[u]_P - [u]_C}{s_2 h} - \frac{[u]_C - [u]_S}{s_4 h} \right]. \end{aligned}$$

采用向后 Euler 差商离散时间导数，有

$$[u_t]_C \approx \frac{1}{\Delta t} \left[ [u]_C^{n+1} - [u]_C^n \right].$$

因此，二维热传导方程 (5.17) 的非等臂长全隐格式可以定义为

$$u_C^{n+1} - u_C^n = \mu \sum_{\sharp \in \{Q, P, W, S\}} \beta_{\sharp} \left[ u_{\sharp}^{n+1} - u_C^{n+1} \right], \quad (5.22a)$$

<sup>3</sup>局部截断误差是指插值逼近带来的误差，即真解代入差分方程后等式两端的差距。

其中

$$\begin{aligned}\beta_Q &= \frac{2}{s_1(s_1 + s_3)}, & \beta_P &= \frac{2}{s_2(s_2 + s_4)}, \\ \beta_W &= \frac{2}{s_3(s_1 + s_3)}, & \beta_S &= \frac{2}{s_4(s_2 + s_4)}.\end{aligned}\quad (5.22b)$$

利用 Taylor 展开技术可知, 它相容于二维热传导方程 (5.17); 但是, 由于  $s_1 = 1$  和  $s_2 = 1$ , 其局部截断误差仅仅达到  $(1, 1, 1)$  阶。□

若在非规则内点采用非等臂长全隐格式 (5.22), 在规则内点采用等臂长全隐格式, 则汇总而成的线性方程组很难保持对称性。为此, 可以强行修正差分系数, 重新定义非等臂长全隐格式

$$u_C^{n+1} - u_C^n = \mu \sum_{\sharp \in \{Q, P, W, S\}} \bar{\beta}_{\sharp} [u_{\sharp}^{n+1} - u_C^{n+1}], \quad (5.23a)$$

其中

$$\bar{\beta}_Q = \frac{1}{s_1 s_3}, \quad \bar{\beta}_P = \frac{1}{s_2 s_4}, \quad \bar{\beta}_W = \frac{1}{s_3^2}, \quad \bar{\beta}_S = \frac{1}{s_4^2}. \quad (5.23b)$$

由于  $s_1 = 1$  和  $s_2 = 1$ , 差分方程 (5.23) 不相容于二维热传导方程 (5.17), 在空间方向的局部截断误差是  $\mathcal{O}(1)$  的。

### 5.2.2 自然边界条件的处理

设 Neumann 边界条件或者自然边界条件是

$$\nabla u \cdot \gamma \Big|_{\widehat{APQB}} = g(x, y), \quad (5.24)$$

其中  $\gamma = (\gamma_1, \gamma_2)^\top$  是单位外法向量, 已知边界函数  $g$  同时间无关。若  $g$  同时时间有关, 相应的处理方式是类似的。

在非规则内点  $W$  处, 相邻的网格边界点  $A$  也是网格点, 边界条件离散相对简单。利用标准的全隐格式, 二维热传导方程 (5.17) 可以在  $W$  点离散为

$$-\mu \left[ u_G^{n+1} + u_C^{n+1} + u_T^{n+1} + u_A^{n+1} \right] + (1 + 4\mu) u_W^{n+1} = u_W^n.$$



利用单侧差商离散技术, 可以建立 Neumann 边界条件的差分离散

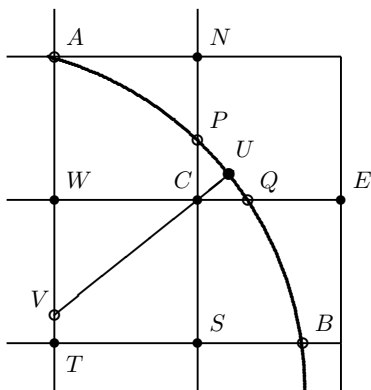
$$\gamma_1 [u_A^{n+1} - u_D^{n+1}] + \gamma_2 [u_A^{n+1} - u_W^{n+1}] = hg(A). \quad (5.25)$$

两式联立, 消去网格边界点信息  $u_A^{n+1}$ , 可得  $W$  点的差分方程

$$\begin{aligned} -\mu [u_G^{n+1} + u_C^{n+1} + u_T^{n+1} + \gamma_1 u_D^{n+1}] \\ + \left[ (1 + 4\mu) - \frac{\mu\gamma_2}{\gamma_1 + \gamma_2} \right] u_W^{n+1} = u_W^n + \frac{\mu hg(A)}{\gamma_1 + \gamma_2}. \end{aligned} \quad (5.26)$$

但是, 在非规则内点  $C$  处, 相邻的两个边界网格点  $P$  和  $Q$  不是网格点, 边界条件离散变得极其繁琐。离散策略<sup>4</sup>主要有两种。

图 5.2: 含导数边界处理的网格描述: 最近迁移方法



**论题 5.6.** 第一种策略也是边界条件的迁移技术, 具体实现过程同偏微分方程没有关系。

答: 图 5.2 是  $C$  点附近的局部放大图。数值离散过程如下:

<sup>4</sup>在网格边界点  $P$  和  $Q$  处, 利用单侧离散技术处理法向导数。联立  $C$  点的非等臂长差分格式, 也可以建立相应的差分方程。详略。

1. 确定  $C$  点到边界  $\Gamma$  (或者近似线段  $PQ$ ) 的垂线, 找到相应的垂足  $U$ ;
2. 确定垂线  $CU$  同内部网格线的交点, 记为  $V$ ;
3. 利用周围的网格点信息, 给出  $[u]_V$  的插值逼近, 例如

$$[u]_V \approx \frac{|VT|}{h}[u]_W + \frac{|WV|}{h}[u]_T,$$

其中  $|VT|$  和  $|WV|$  是两个直线段的长度;

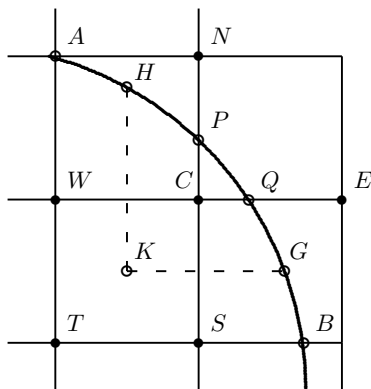
4. 利用  $C$  和  $V$  的函数信息, 建立法向导数的单侧逼近, 即

$$g(U) = \left[ \frac{\partial u}{\partial \gamma} \right]_U \approx \frac{[u]_C - [u]_V}{|VC|}.$$

它等同于将  $U$  点的自然边界条件迁移到  $C$  点。

综上所述, 略去无穷小量, 用数值解替换真解, 即得非规则内点  $C$  处的差分方程。具体表达, 略。  $\square$

图 5.3: 含导数边界处理的网格描述: 积分插值方法



**┆ 论题 5.7.** 第二种策略是利用积分插值方法, 将自然边界条件融合到偏微分方程的离散过程中。

答: 积分插值方法适用于微分方程的离散, 也适用于边界条件的离散。图 5.3 是  $C$  点附近的局部放大图。数值离散过程如下:

1. 确定  $C$  点的控制区域  $\Omega_C$ 。它是一个封闭区域, 通常由边界曲线、单元中心点连线、单元中心点到边界曲线的垂直线段联接而成。在图 5.3 中, 控制区域被简化为曲边三角形  $HKG$ , 其中  $K$  是正方形  $WCST$  的中心。
2. 考虑热传导方程 (5.17) 在控制区域  $\Omega_C$  的积分。利用散度定理, 将二维的面积分转化为一维的曲线积分, 可得恒等式

$$\text{LHS} = \int_{\triangle HKG} [u_t]^{n+1} dx dy = \oint_{\partial HKG} \left[ \frac{\partial u}{\partial \gamma} \right]^{n+1} ds = \text{RHS},$$

其中  $\gamma$  是  $\triangle HKG$  的单位外法向量。

3. 差商离散一阶导数, 积分近似恒等式的两端, 有

$$\begin{aligned} \text{LHS} &\approx |\triangle HKG| [u_t]_C^{n+1} \approx |\triangle HKG| \frac{[u]_C^{n+1} - [u]_C^n}{\Delta t}, \\ \text{RHS} &\approx \frac{[u]_C^{n+1} - [u]_W^{n+1}}{\Delta x} |HK| \\ &\quad + \frac{[u]_C^{n+1} - [u]_S^{n+1}}{\Delta y} |KG| + \int_{\widehat{HPQG}} g(x, y) ds, \end{aligned}$$

其中  $|\triangle HKG|$  是曲边三角形的面积,  $|HK|$  和  $|KG|$  是直线段的长度, 最后的曲线积分可以利用数值积分公式计算。

综上所述, 略去无穷小量, 用数值解替换真解, 即得非规则内点  $C$  处的差分方程。详略。  $\square$

将 Dirichlet 和 Neumann 两类边界条件的数值离散方法结合起来, 即可建立 Robin 边界条件的数值离散方法。详略。

### 5.3 对流方程的人工边界条件

位于有界区间的对流方程

$$u_t = u_x, \quad x \in (0, 1), \quad t > 0 \quad (5.27a)$$

只能在  $x = 1$  处提供入流边界条件

$$u(1, t) = 0, \quad t > 0, \quad (5.27b)$$

不能在  $x = 0$  处提供出流边界条件。利用简单的能量方法, 可知模型问题 (5.27) 是适定的, 即其  $L^2$  模不减, 满足

$$\int_0^1 u^2(x, t) dx \leq \int_0^1 u_0^2(x, t) dx, \quad \forall t, \quad (5.28)$$

其中  $u_0(x)$  是给定的初值函数。

下面考虑模型问题 (5.27) 的有限差分方法。为简单起见, 设  $\mathcal{T}_{\Delta x, \Delta t}$  是等距的时空网格, 相应的空间网格是

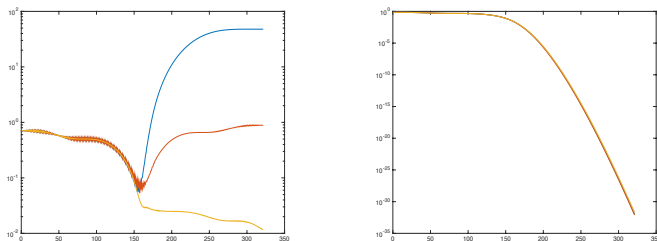
$$\mathcal{T}_{\Delta x} = \{x_j = j\Delta x\}_{j=0}^J,$$

其中  $\Delta x = 1/J$  是空间步长,  $J$  是给定的正整数。设时间步长是  $\Delta t$ , 相应的网比记作  $\nu = \Delta t / \Delta x$ 。

在内部网格点  $\{x_j\}_{j=1}^{J-1}$  处, 本章给出的差分格式都是适用的。在入流边界点  $x_J = 1$  处, 数值边界条件是相对简单的, 可以直接赋值

$$u_J^n = 0, \quad \forall n.$$

但是, 在出流边界点  $x_0 = 0$  处, 数值计算可能遇到困难。对于单边离散模版的迎风格式, 出流边界点值可以直接利用格式给出; 对于双边离散模版的 Lax 格式、LW 格式或者蛙跳格式, 出流边界点值无法直接利用格式给出, 需要人工定义。

图 5.4: 蛙跳格式 (左) 和 LW 格式 (右) 的离散  $L^2$  模演变过程

人工边界条件的定义要非常小心。若其设置不当, 数值结果可能变得极其糟糕, 造成稳定性和精度阶的丢失、局部守恒性的破坏, 以及出现虚假的波反弹现象等等。因篇幅限制, 本节仅仅给出两种常见的设置方式:

1. 利用内部数值解构造多项式, 进行相应的外插逼近。例如,

- 常值外插:  $u_0^n = u_1^n$ , 它具有一阶局部截断误差。
- 线性外插:  $u_0^n = 2u_1^n - u_2^n$ , 它具有二阶局部截断误差。

换言之, 具体操作过程不依赖微分方程和数值格式。

2. 由于出流边界点的特征线是指向区域外部的, 于是出流边界条件可以利用特征线回溯理论进行设置, 即


$$u_0^{n+1} = u_0^n + \nu(u_1^n - u_0^n). \quad (5.29)$$


它就是局部的迎风格式。

为直观理解上述三种人工边界条件的数值差异, 下面观察 LW 格式和蛙跳格式的  $L^2$  模稳定性变化。设模型问题 (5.27a) 的初值为

$$u(x, 0) = \sin(2\pi x), \quad x \in [0, 1]. \quad (5.30)$$

由入流边界条件 (5.27b) 可知, 问题的真解最终演化为零。图 5.4 绘制了两个格式的数值解  $L^2$  模演变过程。左侧子图对应蛙跳格式, 从上到下的三条曲线分别对应线性插值、常值插值和 (5.29) 给出的人工边界条件。实验结果表明, 前两种人工边界条件对于蛙跳格式而言是不稳定的。右侧子图对应 LW 格式, 基于不同人工边界条件的三条曲线几乎重叠在一起。换言之, 上述三种人工边界条件对于 LW 格式而言都是可行的。

 **注释 5.1.** 关于数值边界条件的稳定性影响, 严格的理论分析通常是较难实现的。目前, 较为成功的方法有分离变量法、能量方法和 GKS 理论。

 **注释 5.2.** 当然, 虚拟网格方法也可用于出流边界条件的设置; 略。

---

## 第 6 章

# 椭圆型方程

---

为简单起见，考虑二维 Poisson 方程

$$-\Delta u \equiv -(u_{xx} + u_{yy}) = f(x, y), \quad (x, y) \in \Omega, \quad (6.1)$$

其中  $f$  是已知函数， $\Omega$  是边界逐段光滑的二维有界区域。显然，它是一个稳态问题，同二维扩散方程  $u_t = \Delta u + f(x, y)$  密切相关。换言之，二维 Poisson 方程的解就是二维扩散方程的稳态解，前面的数值离散技术都可以借鉴过来。

### 6.1 五点差分格式

稳态问题的离散网格就是发展型问题的空间网格。为简单起见，不妨设离散网格  $\bar{\Omega}_h$  具有正方形结构，由二维平面的正方形网格

$$\mathcal{T}_h = \{(x_j, y_k): x_j = x_\star + jh, y_k = y_\star + kh\}_{\forall j \forall k}$$

同  $\Omega$  相交而成，其中  $(x_\star, y_\star)$  是参考点， $h$  是网格参数或空间步长。如前，离散网格  $\bar{\Omega}_h$  包含两个部分，即

$$\bar{\Omega}_h = \Omega_h \cup \Gamma_h, \quad (6.2)$$

其中  $\Omega_h$  是网格内点集合， $\Gamma_h$  是边界点集。

对于二维 Poisson 方程 (6.1)，五点差分格式是最简单的。它需要在所有网格内点给出相应的差分方程。在规则内点和非规则内点，差分方程的形式略有不同。

### 6.1.1 规则内点的五点差分方程

为行文简便，我们不采用双下标注方法，而是采用罗盘标注方法。换言之，离散焦点记为中心点  $c$ ，相应位置的真解和数值解分别记为  $[u]_c$  和  $u_c$ ；其它网格点利用相对方位来表示，例如东面的网格点记为  $e$ ，相应位置的真解和数值解分别记为  $[u]_e$  和  $u_e$ 。

若离散焦点  $c$  是规则内点，则格式构造同边界条件无关。利用标准的二阶中心差商，逐维离散两个空间偏导数，有

$$[u_{xx}]_c = \frac{1}{h^2} ([u]_e - 2[u]_c + [u]_w) + \mathcal{O}(h^2), \quad (6.3)$$

$$[u_{yy}]_c = \frac{1}{h^2} ([u]_n - 2[u]_c + [u]_s) + \mathcal{O}(h^2). \quad (6.4)$$

代入到二维 Poisson 方程 (6.1)，略去无穷小量，用数值解替换真解，即得标准的**五点差分方程**：

$$\mathcal{L}_h u_c \equiv \frac{1}{h^2} [4u_c - u_e - u_n - u_s - u_w] = f_c. \quad (6.5)$$

以它为主体的差分格式，本书均简称为五点差分格式。

同发展型差分格式一样，稳态的差分格式也具有相容性、稳定性和收敛性等数值概念。它们的基本含义和描述目标是相同的，仅仅是相关陈述略有不同。例如，相容性概念依旧同局部截断误差密切相关。将 Poisson 方程 (6.1) 的真解代入到五点差分方程 (6.5)，两端的差距就是局部截断误差<sup>1</sup>

$$\tau_c = \frac{1}{h^2} \{4[u]_c - [u]_e - [u]_n - [u]_s - [u]_w\} = \mathcal{O}(h^2),$$

因此说，五点差分格式 (6.5) 是**二阶相容于二维 Poisson 方程**的。稳定性概念依旧描述数值解关于定解条件的连续依赖关系，而收敛性概念直接描述数值解同真解的逼近程度。具体陈述，略。类似地，上述三个数值概念也满足 Lax–Richtmyer 等价定理。

<sup>1</sup>当然，差分格式的量纲要同椭圆型方程保持一致。



### 6.1.2 非规则内点的五点差分方程

对于椭圆型方程的定解问题, Dirichlet、Neumann 和 Robin 边界条件都是常见的边界条件。相应的边界条件离散方法同二维扩散问题类似, 具体内容可参见 §5.2。为行文需要, 下面再次明确指出 Dirichlet 边界条件的离散方法。

↓ **论题 6.1.** 参见图 5.1, 设离散焦点  $c$  是紧邻 Dirichlet 边界的非规则内点。构造二维 Poisson 方程 (6.1) 的非等臂长五点差分方程。

答: 仿照论题 5.5 的空间导数离散过程, 非等臂长的五点差分方程可以定义为

$$\frac{1}{h^2} [\beta_c u_c - \beta_e u_e - \beta_n u_n - \beta_s u_s - \beta_w u_w] = f_c, \quad (6.6)$$

其中  $\beta_c = \beta_e + \beta_s + \beta_w + \beta_n$ 。余下四个系数有两种定义方式, 其一是

$$\begin{aligned} \beta_e &= \frac{2}{s_1(s_1 + s_3)}, & \beta_s &= \frac{2}{s_2(s_2 + s_4)}, \\ \beta_w &= \frac{2}{s_3(s_1 + s_3)}, & \beta_n &= \frac{2}{s_4(s_2 + s_4)}, \end{aligned} \quad (6.7)$$

其二是

$$\beta_e = \frac{1}{s_1 s_3}, \quad \beta_s = \frac{1}{s_2 s_4}, \quad \beta_w = \frac{1}{s_3^2}, \quad \beta_n = \frac{1}{s_4^2}. \quad (6.8)$$

当四个方向的臂长相等时, (6.6) 的局部截断误差是  $\mathcal{O}(h^2)$ 。当臂长不相等的时候, 基于 (6.7) 的五点格式具有  $\mathcal{O}(h)$  的局部截断误差, 基于 (6.8) 的五点格式具有  $\mathcal{O}(1)$  的局部截断误差。□

基于 (6.8) 的非等臂长五点差分方程, 可以保证五点差分格式的离散系统也具有相应的对称性。

### 6.1.3 离散方程组

将网格内点的差分方程汇总起来, 稳态差分格式可以转化为一个规模庞大的线性方程组。下面给出一个简单实例。

**论题 6.2.** 设  $\Omega = (0, 1) \times (0, 1)$ , 考虑二维 Poisson 方程 (6.1) 的 Dirichlet 零边值问题. 给定正整数  $J$ , 定义正方形网格

$$\bar{\Omega}_h = \{(x_j, y_k) = (jh, kh)\}_{j=0:J}^{k=0:J},$$

其中  $h = 1/J$ . 写出五点差分格式对应的线性方程组.

**答:** 按照先后列 (从左到右, 从下到上) 的编号次序, 将二维 (空间) 网格函数  $u_h = \{u_{jk}\}_{j=1:J-1}^{k=1:J-1}$  改写为列向量

$$\mathbf{u}_h = [\mathbf{u}_1^\top, \mathbf{u}_2^\top, \dots, \mathbf{u}_{J-2}^\top, \mathbf{u}_{J-1}^\top]^\top, \quad (6.9)$$

其中  $\mathbf{u}_k = [u_{1,k}, u_{2,k}, \dots, u_{J-2,k}, u_{J-1,k}]^\top$  是其在水平线  $y = y_k$  上的限制. 若采用相同的编号次序排列差分方程, 则线性方程组

$$\mathbf{A}_h \mathbf{u}_h = \mathbf{f}_h \quad (6.10)$$

的刚度矩阵是

$$\begin{aligned} \mathbf{A}_h &= \frac{1}{h^2} [\mathbf{C}_h \otimes \mathbf{1}_h + \mathbf{1}_h \otimes \mathbf{C}_h] \\ &= \frac{1}{h^2} \text{tridiag}\{-\mathbf{1}_h, \mathbf{C}_h + 2\mathbf{1}_h, -\mathbf{1}_h\} \\ &= \frac{1}{h^2} \begin{bmatrix} \mathbf{C}_h + 2\mathbf{1}_h & & & & \\ -\mathbf{1}_h & \mathbf{C}_h + 2\mathbf{1}_h & & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -\mathbf{1}_h \\ & & & -\mathbf{1}_h & \mathbf{C}_h + 2\mathbf{1}_h \end{bmatrix}, \end{aligned} \quad (6.11a)$$

其中  $\mathbf{1}_h$  是  $J-1$  阶单位矩阵,  $\mathbf{C}_h = \text{tridiag}\{-1, 2, -1\}$  是同阶的三对角矩阵. 在 (6.10) 中, 右端向量称为荷载向量, 具体定义是

$$\mathbf{f}_h \equiv [\mathbf{f}_1^\top, \mathbf{f}_2^\top, \dots, \mathbf{f}_{J-2}^\top, \mathbf{f}_{J-1}^\top]^\top, \quad (6.11b)$$

其中  $\mathbf{f}_k = [f_{1,k}, f_{2,k}, \dots, f_{J-2,k}, f_{J-1,k}]^\top$  是已知源项在水平线  $y = y_k$  上的网格限制。□

由于刚度矩阵  $\mathbb{A}_h$  具有对角占优（或者对称正定）性质，故而线性方程组 (6.10) 唯一可解。利用附录内容可知， $\mathbb{A}_h$  具有  $(J-1)^2$  个特征值

$$\lambda^{(p,q)} = \lambda^{(p)} + \lambda^{(q)}, \quad p = 1 : J-1, \quad q = 1 : J-1, \quad (6.12)$$

其中

$$\lambda^{(s)} = \frac{4}{h^2} \sin^2\left(\frac{s\pi}{2J}\right) = \frac{4}{h^2} \sin^2\left(\frac{sh\pi}{2}\right), \quad s = 1 : J-1$$

是三对角矩阵  $\mathbb{C}_h$  的特征值。因此，刚度矩阵的谱条件数是

$$\text{cond}(\mathbb{A}_h) \equiv \frac{\max_{p,q} \lambda^{(p,q)}}{\min_{p,q} \lambda^{(p,q)}} = \cot^2\left(\frac{\pi h}{2}\right) = \mathcal{O}(h^{-2}). \quad (6.13)$$

换言之，当离散网格变密时，线性方程组 (6.10) 不仅计算规模膨胀，而且病态程度加剧。此时，线性方程组的数值求解效率，将成为五点差分格式能否成功的关键。

#### 6.1.4 线性方程组的数值解法<sup>‡</sup>

对于线性方程组 (6.10)，常用的直接法和迭代法都是高效的数值求解器，例如 Gauss 消元方法、Gauss-Seidel 方法、超松弛方法、共轭斜量方法和预处理方法等。具体内容可查阅数值代数的相关资料，略。上述数值方法都是纯代数的，没用充分考虑问题的产生背景。事实上，直接利用微分方程的数值离散技术，也可以构造线性方程组 (6.10) 的高效数值求解器，例如交替方向 (alternating direction) 方法、多重网格 (multi-grid) 方法和区域分解 (domain decomposition) 方法。因篇幅有限，下面简要描述它们的数值实现过程。

##### 交替方向方法

由于线性方程组 (6.10) 来自二维 Poisson 方程的差分格式

$$-\mathcal{L}_h u_c = f_c,$$

故而二维扩散方程（基于相同空间离散处理）半离散格式

$$\frac{du_c}{dt} = \mathcal{L}_h u_c + f_c$$

的稳态数值解计算过程，可以看作 (6.10) 的迭代方法。为提高数值解趋于稳态的计算效率，不妨采用二维扩散方程的 ADI 方法或者 LOD 方法。

下面给出一个具体实例。基于  $u_t = u_{xx} + u_{yy} + f$  的 PR 格式，线性方程组 (6.10) 的交替方向方法可以定义为

$$\mathbf{u}_h^{k+\frac{1}{2}} = \mathbf{u}_h^n - \tau_k \left[ \mathbb{L}_1 \mathbf{u}_h^{k+\frac{1}{2}} + \mathbb{L}_2 \mathbf{u}_h^k - \mathbf{f}_h \right], \quad (6.14a)$$

$$\mathbf{u}_h^{k+1} = \mathbf{u}_h^n - \tau_k \left[ \mathbb{L}_1 \mathbf{u}_h^{k+\frac{1}{2}} + \mathbb{L}_2 \mathbf{u}_h^{k+1} - \mathbf{f}_h \right], \quad (6.14b)$$

其中  $\tau_k$  称为虚拟时间步长，

$$\mathbb{L}_1 = \frac{1}{h^2} \mathbb{C}_h \otimes \mathbb{I}_h, \quad \mathbb{L}_2 = \frac{1}{h^2} \mathbb{I}_h \otimes \mathbb{C}_h, \quad (6.15)$$

分别是  $x$  方向和  $y$  方向的二阶中心差商（整体）算子。由 (6.14) 可知，交替方向方法的迭代矩阵是

$$\mathbb{T}_k = \tau_k^2 (\mathbb{I} + \tau_k \mathbb{L}_2)^{-1} \mathbb{L}_1 (\mathbb{I} + \tau_k \mathbb{L}_1)^{-1} \mathbb{L}_2.$$

考虑相应的一阶定常迭代，即  $\tau_k \equiv \tau$  和  $\mathbb{T}_k \equiv \mathbb{T}$ 。若

$$\tau = \tau_{\text{opt}} = \frac{1}{2} \left[ \sin \pi h \right]^{-1}, \quad (6.16)$$

则相应的迭代矩阵（记为  $\mathbb{T}_{\text{opt}}$ ）具有最小的谱半径，即

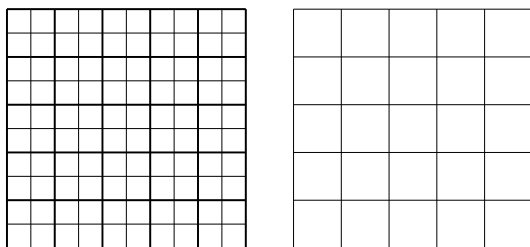
$$\rho(\mathbb{T}_{\text{opt}}) = \left[ \frac{1 - \tan \frac{\pi h}{2}}{1 + \tan \frac{\pi h}{2}} \right]^2. \quad (6.17)$$

此时，交替方向方法的收敛速度可以媲美带有最佳因子的超松弛方法。具体内容可参见 [3]，详略。

### 多重网格法<sup>‡</sup>

对于线性方程组 (6.10) 及其迭代算法, Fourier 理论也是一种极为有效的分析工具。对于 Jacobi 方法或者 Gauss-Seidel 方法, 不同波数的简谐波具有不同的迭代收敛速度。具体来讲, 高波数 (或高频) 的简谐波误差收敛很快, 而低波数 (或低频) 的简谐波误差收敛较慢。换言之, 上述简单迭代方法可以快速滤去那些高频 (简谐波) 误差, 它们的迭代收敛速度完全受限于低频 (简谐波) 误差趋于零的速度。

图 6.1: 嵌套网格: 左侧的粗线对应右侧的网格



事实上, 某个波数的简谐波到底归属于高频还是低频, 要视它所在的空间网格尺寸。通常, 细网格上的低频简谐波可能是粗网格上的高频波。因此, 如果细网格上的低频误差能够转化为粗网格上的高频误差, 则粗网格上的简单迭代方法可以令上述误差快速地衰减到零。不断重复这个过程, 迭代算法的效率有望获得明显的提升。这就是多重网格方法的设计初衷, 其核心技术是低频误差与高频误差在粗细网格之间的相互转化。

最简单的实现过程是基于嵌套网格的二重网格方法。为简单起见, 设细网格是  $\bar{\Omega}_h$ , 由粗网格  $\bar{\Omega}_{2h}$  加密而成, 即粗网格的网格点集包含于细网格的网格点集。参见图 6.1。设线性方程组 (6.10) 是二维 Poisson 方程 (6.1) 在细网格  $\bar{\Omega}_h$  上离散得到的, 相应的迭代解可以按照以下步骤进行计算:

1. 基于细网格的光滑过程: 以猜测值或者迭代解  $\mathbf{u}_h^k$  为初值, 执行  $m$  次简

单迭代（例如 Jacobi 或者 Gauss-Seidel 方法），得到相对光滑的数值解  $\bar{\mathbf{u}}_h^k$ 。通常， $m$  同空间步长  $h$  无关。

2. 基于粗网格的校正过程：这是二重网格方法的核心部分。

$$\text{a) 细网格的残量计算：} \quad \mathbf{r}_h^k = \mathbf{f}_h - \mathbb{A}_h \bar{\mathbf{u}}_h^k, \quad (6.18\text{a})$$

$$\text{b) 粗网格的残量限制：} \quad \mathbf{r}_{2h}^k = \mathbb{I}_h^{2h} \mathbf{r}_h^k, \quad (6.18\text{b})$$

$$\text{c) 粗网格的方程组求解：} \quad \mathbb{A}_{2h} \mathbf{e}_{2h}^k = -\mathbf{r}_{2h}^k, \quad (6.18\text{c})$$

$$\text{d) 粗网格到细网格的延拓：} \quad \bar{\mathbf{e}}_h^k = \mathbb{I}_{2h}^h \mathbf{e}_{2h}^k, \quad (6.18\text{d})$$


$$\text{e) 细网格的数值解校正：} \quad \mathbf{u}_h^{k+1} = \bar{\mathbf{u}}_h^k - \bar{\mathbf{e}}_h^k. \quad (6.18\text{e})$$

其中  $\mathbb{I}_h^{2h}$  称为细网格到粗网格的限制算子， $\mathbb{I}_{2h}^h$  称为粗网格到细网格的延拓算子。通常，两者互为逆算子。

综上所述，二重网格方法的迭代过程可以描述为

$$\mathbf{u}_h^{k+1} = \bar{\mathbf{u}}_h^k - \mathbb{I}_{2h}^h \mathbb{A}_{2h}^{-1} \mathbb{I}_h^{2h} (\mathbb{A}_h \bar{\mathbf{u}}_h^k - \mathbf{f}_h). \quad (6.19)$$

理论分析表明：要使迭代误差达到用户要求，二重网格方法的迭代步数同计算规模或网格步长  $h$  无关。整个算法的计算复杂度是  $\mathcal{O}(n \log n)$ ，堪称最优的线性方程组计算效率，其中  $n = \mathcal{O}(h^{-2})$  是未知量的个数。

 **注释 6.1.** 事实上，粗网格的代数方程组 (6.18c) 可以继续采用二重网格方法求解。如此嵌套下去，即可形成各种形式的多重网格方法。因此说，二重网格方法是多重网格方法的基础。

下面给出限制算子  $\mathbb{I}_h^{2h}$  和延拓算子  $\mathbb{I}_{2h}^h$  的具体定义。为行文简便，不妨采用二维笛卡尔网格的星型图表示法。

1. 限制算子  $\mathbb{I}_h^{2h}$  是细网格函数到粗网格函数的加权平均算子，对应的星型图是

$$\mathbb{I}_h^{2h} \equiv \frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}_h^{2h}, \quad (6.20)$$

其含义是：将矩阵中心置于关注的粗网格点上，矩阵的每个元素表示周围九个细网格点的对应权重。在粗网格点上的限制值，就是九个细网格点值按照给定权重进行平均。

2. 延拓算子  $\mathbb{I}_{2h}^h$  是粗网格函数到细网格函数的双线性插值算子，对应的星型图是

$$\mathbb{I}_{2h}^h \equiv \frac{1}{4} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}_{2h}^h, \quad (6.21)$$

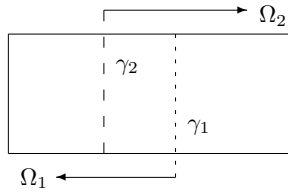
其含义是：将矩阵中心置于关注的粗网格点上，矩阵的每个元素就是周围九个细网格点的对应权重。此时，粗网格点值将按照权重，逐一分配到周围的九个细网格点。在所有粗网格点上执行上述操作，最终的叠加结果就是网格函数从粗网格到细网格的插值延拓。

当然，限制算子和延拓算子是不唯一的，可以差分格式进行相应的设置。

### 区域分解方法<sup>‡</sup>

基本思想就是将一个整体区域分割为有限个子域，把一个大规模的计算问题转化为多个小规模的计算问题。如果离散系统的基本算法具有高次多项式的计算复杂度，则上述操作通常都可以明显地改善基本算法的计算效率。此外，每个子域的数值计算可以相对独立地进行，区域分解方法具有本质并行机制。

图 6.2: 区域分解方法的重叠区域



区域分解方法具有多种实现途径。因篇幅限制, 本节仅仅介绍重叠型 Schwarz 方法。参见图 6.2, 设整体区域  $\Omega$  是两个重叠区域的并集, 即  $\Omega = \Omega_1 \cup \Omega_2$ , 且  $\Omega_1 \cap \Omega_2$  的测度严格大于零。为简单起见, 假设  $\Omega_\kappa$  的内部边界

$$\gamma_\kappa = \partial\Omega_\kappa \cap \Omega, \quad \kappa = 1, 2$$

是由整体网格  $\bar{\Omega}_h$  的部分网格线联接而成。设线性方程组 (6.10) 是二维 Poisson 方程 (6.1) 在细网格  $\bar{\Omega}_h$  上利用五点差分格式 (6.5) 离散得到的, 相应的迭代解可以按照以下步骤进行计算:

1. 基于  $\bar{\Omega}_h$  在  $\bar{\Omega}_1$  限制而成的子网格, 利用五点差分格式, 计算  $\Omega_1$  内的二维 Poisson 方程, 相应的边界条件<sup>2</sup>是

$$u|_{\gamma_1} = \mathbf{u}_h^k, \quad u|_{\partial\Omega \cap \partial\Omega_1} = 0,$$

其中  $\mathbf{u}_h^k$  是猜测初值或者迭代解。相应的数值解记为  $\tilde{\mathbf{u}}_h^{k+1}$ 。

2. 基于  $\bar{\Omega}_h$  在  $\bar{\Omega}_2$  限制而成的子网格, 利用五点差分格式, 计算  $\Omega_2$  内的二维 Poisson 方程, 相应的边界条件是

$$u|_{\gamma_2} = \tilde{\mathbf{u}}_h^{k+1}, \quad u|_{\partial\Omega \cap \partial\Omega_2} = 0.$$

相应的数值解记为  $\mathbf{u}_h^{k+1}$ 。

若重叠区域  $\Omega_1 \cap \Omega_2$  的相邻数值解  $\mathbf{u}_h^{k+1}$  和  $\tilde{\mathbf{u}}_h^{k+1}$  充分接近, 则迭代可以停止。理论分析表明: 重叠型 Schwarz 方法是收敛的, 其收敛速度同重叠区域的面积成正比。换言之, 若重叠区域的面积越大, 则迭代收敛的速度越快。

在上述算法中, 计算流程称为异步并行方式, 边界信息交换方式称为 Dirichlet-Dirichlet 方式。事实上, 计算流程和边界信息交换方式还有其他策略, 例如同步并行方式和 Dirichlet-Neumann 方式。

---

<sup>2</sup>位于内部边界的函数可以理解为相应网格点信息形成的线性插值函数。



## 6.2 最大模估计

本节以二维 Poisson 方程 Dirichlet 边值问题的五点差分格式为例, 建立椭圆型差分格式的最大模稳定性和误差估计<sup>3</sup>。理论分析的核心内容是强最大值原理。

### 6.2.1 强最大值原理

为行文简便, 本节采用单下标标注方法。换言之, 离散点直接利用整数  $\ell$  来表示, 相应位置的数值解用  $u_\ell$  来表示。

回忆 (6.2), 设离散网格为  $\bar{\Omega}_h = \Omega_h \cup \Gamma_h$ , 其中  $\Omega_h$  是网格内点集,  $\Gamma_h$  是网格边界点集。设  $\{u_j\}_{\forall j \in \bar{\Omega}_h}$  是未知网格函数, 满足差分格式

$$\mathcal{D}_h u_j \equiv d_{jj} u_j - \sum_{k \in \mathcal{O}(j)} d_{jk} u_k = f_j, \quad j \in \Omega_h, \quad (6.22a)$$

$$u_j = g_j, \quad j \in \Gamma_h, \quad (6.22b)$$

其中  $f_j$  和  $g_j$  是已知的网格函数,  $\mathcal{O}(j)$  表示网格内点  $j$  的空心邻域, 包含同  $j$  关联的有限个网格点。若  $\{d_{jj}\}_{\forall j}$  和  $\{d_{jk}\}_{\forall j \forall k}$  均是给定的正数, 且满足

$$d_{jj} \geq \sum_{k \in \mathcal{O}(j)} d_{jk}, \quad \forall j, \quad (6.23)$$

则称  $\mathcal{D}_h$  是椭圆型差分算子, (6.22) 是椭圆型差分格式。

默认离散网格  $\bar{\Omega}_h$  关于差分格式 (6.22) 是**连通**的。换言之, 对于任意网格内点  $\ell_\star$  和  $\ell^\star$ , 均存在网格内点形成的路径

$$\ell_\star = \ell_0, \ell_1, \ell_2, \dots, \ell_{m-1}, \ell_m = \ell^\star, \quad (6.24a)$$

其中

$$\ell_r \in \mathcal{O}(\ell_{r-1}), \quad r = 1 : m. \quad (6.24b)$$

---

<sup>3</sup>至于  $L^2$  模度稳定性和误差估计, 可以利用能量方法或矩阵直接方法给出; 详略。

由于网格边界点至少落在某个网格内点的空心邻域内, 两个网格内点可以推广到一个网格内点和一个网格边界点。

**引理 6.1.** (强最大值原理) 若网格函数  $u = \{u_j\}_{j \in \bar{\Omega}_h}$  不恒等于常值, 且处处满足

$$\mathcal{D}_h u_j \leq 0, \quad j \in \Omega_h, \quad (6.25)$$

则  $u$  不可能在网格内点集  $\Omega_h$  上取到正的最大值。

**证明:** 反证。假设  $u$  在网格内点  $j_0$  处取到正的最大值。注意到

$$\mathcal{D}_h u_j = \left[ d_{jj} - \sum_{k \in \mathcal{O}(j)} d_{jk} \right] u_j + \sum_{k \in \mathcal{O}(j)} d_{jk} [u_j - u_k], \quad (6.26)$$

利用椭圆型差分算子的性质, 可知  $\mathcal{D}_h u_{j_0} \geq 0$ 。因此, 假设条件 (6.25) 蕴含  $\mathcal{D}_h u_{j_0} = 0$ , 进而得到

$$d_{j_0 j_0} = \sum_{k \in \mathcal{O}(j_0)} d_{j_0 k}, \quad \text{且 } u_k = u_{j_0}, \quad \forall k \in \mathcal{O}(j_0).$$

换言之, 同  $j_0$  关联的网格点也取到正的最大值。由于离散网格  $\bar{\Omega}_h$  是连通的, 利用离散模版的漂移, 可知所有网格点都取到正的最大值, 同引理条件 (非定常假设) 矛盾! 因此, 命题得证。  $\square$

事实上, 引理 6.1 是椭圆型方程强最大值原理的数值刻画。作为一个直接应用, 我们可以建立椭圆型差分格式的优函数理论。

**引理 6.2.** 若椭圆型差分格式

$$\begin{cases} \mathcal{D}_h U_j = F_j, & j \in \Omega_h, \\ U_j = G_j, & j \in \Gamma_h. \end{cases} \quad (6.27)$$

和椭圆型差分格式 (6.22) 具有相同的算子结构, 且

$$|f_j| \leq F_j, \quad \forall j \in \Omega_h; \quad |g_j| \leq G_j, \quad \forall j \in \Gamma_h,$$

则称网格函数  $U_j$  是椭圆型差分格式 (6.22) 的优函数, 即

$$|u_j| \leq U_j, \quad \forall j \in \bar{\Omega}_h. \quad (6.28)$$

**证明：**记  $z_j = u_j - U_j$ 。利用差分格式的线性结构，有

$$\begin{cases} \mathcal{D}_h z_j = -F_j + g_j \leq 0, & j \in \Omega_h, \\ z_j = -G_j + g_j \leq 0, & j \in \Gamma_h. \end{cases} \quad (6.29)$$

利用引理 6.1，可知  $z_j \leq 0$  处处成立。类似地，可证  $u_j + U_j \geq 0$  处处成立。证毕。  $\square$

### 6.2.2 简单估计

下面建立椭圆型差分格式 (6.22) 的最大模稳定性结论和最大模误差估计。为简单起见，设差分方程都具有 (6.6) 的形式，即规则内点的五点差分方程是等臂长的，非规则内点的五点差分方程可能是非等臂长的。

**定理 6.1.** 椭圆型差分格式 (6.22) 满足最大模稳定性，即

$$\|u_h\|_{\bar{\Omega}_{h,\infty}} \leq C_1 [\|f\|_{\Omega_{h,\infty}} + \|g\|_{\Gamma_{h,\infty}}], \quad (6.30)$$

其中界定常数  $C_1 > 0$  同空间步长  $h$  无关。

**证明：**设  $(x_\star, y_\star)$  是计算区域  $\Omega$  的中心， $\rho$  是区域半径，定义

$$U(x, y) = K [\rho^2 - (x - x_\star)^2 - (y - y_\star)^2], \quad (6.31)$$

其中  $K$  是需要适当选取的正常数。简单计算，可知  $\mathcal{D}_h U_j = \theta K$ ；对于五点差分方程 (6.6) 而言，在规则内点有  $\theta = 4$ ，在非规则内点也有  $\theta > 2$ 。因此，若取

$$K = \frac{1}{2} \max_{\Omega_h} |f_j|,$$

则 (6.31) 定义的网格函数  $\{U_j = U(x_j, y_j)\}_{\forall j \in \bar{\Omega}_h}$  是椭圆型差分格式

$$\begin{cases} \mathcal{D}_h u_j^{(1)} = f_j, & j \in \Omega_h, \\ u_j = 0, & j \in \Gamma_h, \end{cases} \quad (6.32)$$

的优函数。由引理 6.2，可知  $\|u^{(1)}\|_{\bar{\Omega}_{h,\infty}} \leq \frac{1}{2} \rho^2 \|f\|_{\Omega_{h,\infty}}$ 。

显然, 常值网格函数  $\{V_j = \max_{\Gamma_h} |g_j|\}_{j \in \bar{\Omega}_h}$  是椭圆型差分格式

$$\begin{cases} \mathcal{D}_h u_j^{(2)} = 0, & j \in \Omega_h, \\ u_j = g_j, & j \in \Gamma_h, \end{cases} \quad (6.33)$$

的优函数。因此,  $\|u^{(2)}\|_{\bar{\Omega}_h, \infty} \leq \|g\|_{\Gamma_h, \infty}$ 。

注意到  $u_h = u^{(1)} + u^{(2)}$  和三角不等式, 可知

$$\|u_h\|_{\bar{\Omega}_h, \infty} = \|u^{(1)} + u^{(2)}\|_{\bar{\Omega}_h, \infty} \leq \|u^{(1)}\|_{\bar{\Omega}_h, \infty} + \|u^{(2)}\|_{\bar{\Omega}_h, \infty}.$$

综上所述, 定理即证。  $\square$

下面建立五点差分格式的最大模误差估计。设  $\Omega = (0, 1) \times (0, 1)$  是正方形区域, 相应的离散网格

$$\bar{\Omega}_h = \{(x_j, y_k) = (jh, kh)\}_{j=0:J}^{k=0:J}, \quad h = 1/J$$

是完美匹配的。换言之, 非规则内点的五点差分方程 (6.6) 也是等臂长的。利用线性结构, 可得误差方程

$$\begin{cases} \mathcal{D}_h e_j = \tau_j, & j \in \Omega_h, \\ e_j = 0, & j \in \Gamma_h, \end{cases} \quad (6.34)$$

其中  $e_j = [u]_j - u_j$  是数值误差,  $\tau_j = \mathcal{O}((\Delta x)^2)$  是局部截断误差。因此, 由定理 6.1 可知,

$$\|e\|_{\bar{\Omega}_h, \infty} \leq C_1 \|\tau\|_{\Omega_h, \infty} \leq C_2 (\Delta x)^2.$$

换言之, 五点差分格式具有二阶的最大模误差。

### 6.2.3 精细估计

回顾 §6.1.2 可知: 对于任意的二维有界区域, 非规则内点的五点差分方程可能是非等臂长的, 相应的局部截断误差至多一阶。此时, 直接利用定理 6.1, 相应的最大模误差估计也至多一阶。但是, 数值经验表明最大模误差依旧可以达到二阶。换言之, 理论结果同真实表现存在差距。

为弥补上述差距, 我们需要完善定理 6.1 的结论, 建立更加精细的最大模稳定性结论. 基于线性叠加原理, 将椭圆型差分格式 (6.32) 的数值解再次分裂为两个部分, 即

$$u^{(1)} = u^* + u^{**}, \quad (6.35)$$

其中  $u^*$  和  $u^{**}$  分别满足椭圆型差分格式

$$\begin{cases} \mathcal{D}_h u_j^* = f_j, & \mathcal{D}_h u_j^{**} = 0, & j \in \Omega_h^*, \\ \mathcal{D}_h u_j^* = 0, & \mathcal{D}_h u_j^{**} = f_j, & j \in \Omega_h^{**}, \\ u_j^* = 0, & u_j^{**} = 0, & j \in \Gamma_h, \end{cases} \quad (6.36)$$

在 (6.36) 中,  $\Omega_h^*$  是规则内点集,  $\Omega_h^{**}$  是非规则内点集. 类似于定理 6.1 的证明, 借助同样的优函数可知

$$\|u^*\|_{\bar{\Omega}_h, \infty} \leq C_3 \|f\|_{\Omega_h^*, \infty}, \quad (6.37)$$

其中定解常数  $C_3 > 0$  同空间步长无关. 要估计  $u^{**}$ , 我们需要完成下面两步:

- 首先, 仿照引理 6.1 的证明过程, 建立增强版的强最大值原理: 若  $u^{**}$  不是常值函数, 则正的最大值和负的最小值均不会出现在规则内点集  $\Omega_h^*$  上. 换言之,  $u^{**}$  的最大绝对值必然出现在非规则内点集  $\Omega_h^{**}$  上.
- 然后, 深入观察  $\Omega_h^{**}$  内的五点差分方程, 有

$$\tilde{d}_{jj} u_j + \sum_{k \in \mathcal{O}(j), k \in \Gamma_h} d_{jk} [u_j - u_k] = f_j, \quad (6.38a)$$

且

$$\tilde{d}_{jj} \equiv d_{jj} - \sum_{k \in \mathcal{O}(j), k \in \Gamma_h} d_{jk} \geq \frac{C_4}{h^2}, \quad \forall j \in \Omega_h^{**}, \quad (6.38b)$$

其中界定常数  $C_4 > 0$  同空间步长无关. 设正的最大值 (或负的最小值) 在某个非规则内点取到, 考虑相应位置的差分方程 (6.38a). 利用 (6.38b), 可知

$$\|u^{**}\|_{\Omega_h^{**}, \infty} \leq C_5 h^2 \|f\|_{\Omega_h^{**}, \infty}, \quad (6.39)$$

其中界定常数  $C_5 > 0$  同空间步长无关.

综上所述, 椭圆型差分格式 (6.22) 具有稳定性结论

$$\|u_h\|_{\bar{\Omega}_h, \infty} \leq C_6 \left[ \|f\|_{\Omega_h^*, \infty} + h^2 \|f\|_{\Omega_h^{**}, \infty} + \|g\|_{\Gamma_h, \infty} \right], \quad (6.40)$$

其中界定常数  $C_6 > 0$  同空间步长无关。

**┆ 论题 6.3.** 证明: 无论 *Dirichlet* 边界条件的数值处理是否完美, 五点差分格式 (6.6) 都具有二阶的最大模误差估计。

答: 此时, 误差方程 (6.34) 依旧成立。由 (6.40), 可知

$$\|e\|_{\bar{\Omega}_h, \infty} \leq C_6 \left[ \|\tau\|_{\Omega_h^*, \infty} + h^2 \|\tau\|_{\Omega_h^{**}, \infty} \right].$$

于是, 论题得证。 □

**┆ 注释 6.2.** 类似地, 在适当的时空约束条件下, 二维扩散方程 *Dirichlet* 问题的偏隐格式也具有最优的最大模误差估计, 无论边界条件的数值离散是否完美。例如, 对于任意的网比, 全隐格式也具有强最大值原理, 相应的最大模误差都是  $\mathcal{O}(h^2 + \Delta t)$ , 其中  $h$  是空间步长,  $\Delta t$  是时间步长。

## 6.3 提高数值精度的方法

五点差分格式只有二阶精度, 相应的计算效率较低。比如, 数值误差要降至原有的 25%, 空间步长需缩小 50%, 计算规模需膨胀 4 倍。若线性方程组的计算复杂度是非线性的, 则相应的 CPU 时间将急剧地增长。本节给出三种解决途径。

### 6.3.1 Richardson 外推技术

Richardson 外推技术是简单易行的事后处理方法, 可以基于较粗的网格和较少的运算量, 获得更为准确的数值解。相应的理论基础是误差渐近展开公式。

**‡ 论题 6.4.** 考虑论题 6.2 中的五点差分格式, 建立相应的误差渐近展开公式。

答: 考虑辅助的椭圆型方程

$$-\Delta w = [u_{xxxx}] + [u_{yyyy}], \quad (x, y) \in \Omega = (0, 1) \times (0, 1), \quad (6.41a)$$

$$w = 0, \quad (x, y) \in \Gamma, \quad (6.41b)$$

其中  $w(x, y)$  是未知函数,  $u(x, y)$  是二维 Poisson 方程定解问题的真解。采用双下标标注方法, 令

$$\eta_{jk} = u_{jk} - [u]_{jk} - \frac{1}{12}h^2[w]_{jk},$$

其中  $u_{jk}$  是五点差分格式的数值解。注意到五点差分格式的局部截断误差, 简单计算可知

$$\mathcal{L}_h \eta_{jk} = \mathcal{O}(h^4), \quad (x_j, y_k) \in \Omega_h, \quad (6.42a)$$

$$\eta_{jk} = 0, \quad (x_j, y_k) \in \Gamma_h, \quad (6.42b)$$

其中  $\mathcal{L}_h$  的具体定义见 (6.5),  $\Omega_h$  是网格内点集,  $\Gamma_h$  是网格边界点集。利用椭圆型差分格式的优函数理论, 可知  $\|\eta\|_{\bar{\Omega}_h, \infty} = \mathcal{O}(h^4)$ , 即五点差分格式具有误差渐近展开公式


$$u_{jk} = [u]_{jk} + \frac{1}{12}h^2[w]_{jk} + \mathcal{O}(h^4). \quad (6.43)$$

证毕。 □

设  $\bar{\Omega}_h$  是由  $\bar{\Omega}_{2h}$  加密而成的嵌套网格, 相应的五点差分格式给出两个数值解  $u^h$  和  $u^{2h}$ 。它们都是真解  $[u]$  的二阶逼近。基于误差渐近展开公式 (6.43), 定义粗网格  $\bar{\Omega}_{2h}$  上的 Richardson 外推组合

$$\tilde{u}^{2h} = \frac{4}{3}u^h - \frac{1}{3}u^{2h}. \quad (6.44)$$

显然, 它是真解  $[u]$  的四阶逼近, 即  $\|\tilde{u}^{2h} - [u]^{2h}\|_{\Omega_{2h}, \infty} = \mathcal{O}(h^4)$ 。

 **注释 6.3.** 当区域形状变得复杂或者边界条件类型发生变化时, 边界条件离散方法可能影响误差渐近展开公式的完美结构。一般而言, 对于本质边界条件, Richardson 外推公式 (6.44) 依旧有效; 但是, 对于自然边界条件, 相应的数值外推表现不够理想。

### 6.3.2 九点格式

直接提高数值格式的相容阶, 也是有效的解决途径。通过离散模版的扩张, 即可实现上述目标。例如, 基于正方形网格, 二维 Poisson 方程 (6.1) 的 **四阶九点格式**可以定义为


$$\mathcal{N}_h u_c := \left[ \frac{2}{3} \mathcal{L}_h + \frac{1}{3} \mathcal{S}_h \right] u_c = f_c - \frac{1}{12} \mathcal{L}_h f_c, \quad (6.45)$$

其中  $\mathcal{L}_h$  是正五点格式 (6.5) 的差分算子,  $\mathcal{S}_h$  是**斜五点格式**<sup>4</sup>

$$\mathcal{S}_h u_c \equiv \frac{1}{2h^2} [4u_c - u_{ne} - u_{nw} - u_{se} - u_{sw}] = f_c \quad (6.46)$$

的差分算子。基于正方形网格, 在 (6.45) 的右侧增加适当的高阶修正, 可得二维 Poisson 方程 (6.1) 的**六阶九点格式**

$$\left[ \frac{2}{3} \mathcal{L}_h + \frac{1}{3} \mathcal{S}_h \right] u_c = f_c - \frac{1}{12} \mathcal{L}_h f_c - \frac{h^4}{240} (\delta_x^4 + \delta_y^4) f_c + \frac{h^4}{90} \delta_x^2 \delta_y^2 f_c.$$

 **注释 6.4.** 基于非正方形的空间 (矩形) 网格, 二维 Poisson 方程 (6.1) 的四阶九点格式将略显复杂。事实上, 它可以定义为

$$-\left[ \frac{\delta_x^2}{(\Delta x)^2} + \frac{\delta_y^2}{(\Delta y)^2} + \frac{(\Delta x)^2 + (\Delta y)^2}{12(\Delta x)^2(\Delta y)^2} \delta_x^2 \delta_y^2 \right] u_c = \left[ 1 + \frac{\delta_x^2 + \delta_y^2}{12} \right] f_c,$$

其中  $\Delta x$  和  $\Delta y$  是两个方向的空间步长。要确保它是椭圆型差分格式,  $\Delta x$  和  $\Delta y$  需要满足限制条件  $\frac{1}{\sqrt{5}} \leq \frac{\Delta x}{\Delta y} \leq \sqrt{5}$ 。

随着离散模版的扩张, 数值格式的相容阶得以提高。但是, 系数矩阵变得越来越稠密, 线性方程组的数值求解变得越来越困难。

<sup>4</sup>在旋转变换下, Laplace 算子的形式保持不变。将直角坐标系旋转  $\pi/4$ , 正五点格式 (6.5) 可以导出斜五点格式 (6.46)。相应的局部截断误差也是  $\mathcal{O}(h^2)$ 。



### 6.3.3 Kreiss 差分格式

回忆 (1.15)，二阶导数具有紧凑的四阶逼近方式

$$D^2 = \frac{1}{h^2} \frac{\delta^2}{I + \delta^2/12} + \mathcal{O}(h^4), \quad (6.47)$$

其中  $h$  为空间步长。基于此，我们可以构造二维 Poisson 方程 (6.1) 的 Kreiss 格式。它是四阶相容的**紧凑型差分格式**，将真解和导数同时作为逼近目标。

空间网格设置如前。Kreiss 格式的设计过程如下：同时引进三个网格函数  $u, p$  和  $q$ ，分别逼近问题真解和二阶导数，即

$$u \approx [u], \quad p \approx [u_{xx}], \quad q \approx [u_{yy}].$$

为行文简便，下面采用双下标标注方法。以网格点  $(x_j, y_k)$  为离散焦点。对应二维 Poisson 方程 (6.1)，定义差分方程

$$-(p_{jk} + q_{jk}) = f_{jk}. \quad (6.48a)$$

利用 (6.47) 离散两个空间导数，可以建立差分方程

$$\frac{1}{12}p_{j+1,k} + \frac{5}{6}p_{jk} + \frac{1}{12}p_{j-1,k} = \frac{1}{h^2}\delta_x^2 u_{jk}, \quad (6.48b)$$

$$\frac{1}{12}q_{j,k+1} + \frac{5}{6}q_{jk} + \frac{1}{12}q_{j,k-1} = \frac{1}{h^2}\delta_y^2 u_{jk}. \quad (6.48c)$$

联立上述三种差分方程，即可得到网格函数  $u, p, q$  的线性方程组，其未知量的总数是网格点数的三倍。

这个大规模的线性方程组可以利用迭代方法求解。以 Dirichlet 零边值问题为例。若猜测初值或迭代解  $u^n$  是已知的，则迭代解  $u^{n+1}$  可以按照以下步骤进行计算：

- 利用 (6.48b) 和 (6.48c)，计算相应的  $p^n$  和  $q^n$ 。此时，位于边界位置的二阶导数可以利用  $u^n$  的二阶单侧差商来表示。在每条直线段上，相应的线性方程组具有对角占优的三对角系数矩阵，可以利用 Thomas 算法快速地求解。因此，相应的乘除法运算次数同直线上的未知量个数成正比。

- 利用 (6.48a) 的简单迭代格式

$$\frac{u_{jk}^{n+1} - u_{jk}^n}{\tau^n} - [p_{jk}^n + q_{jk}^n] = f_{jk}, \quad (6.49)$$

可以给出下一步的迭代解  $u^{n+1}$ , 其中  $\tau^n$  是迭代参数, 称为虚拟时间步长。

若相邻误差  $\|u^{n+1} - u^n\|_2$  达到用户需求, 则迭代可以停止。

## 6.4 有限元方法<sup>‡</sup>

当区域形状复杂或者边界条件含有导数的时候, 基于正交网格的有限差分方法常常变得捉襟见肘, 相应的数值实现过程变得繁琐。此时, 有限元方法将展现其灵活性和数值优势, 其成功主要源于下面三个原因:

- 基于变分表述, 自然边界条件的处理变得简单。有限元方法继承了古典变分方法的基本框架, 相应的数值计算具有扎实的理论基础。
- 基于分片多项式理论, 有限元方法成功地克服了古典变分方法的缺点, 可以用于任意形状的区域。
- 在有限元方法中, 刚度矩阵和荷载向量的组装具有标准的操作流程, 相应的程序编程和实际应用具有统一框架。

时至今日, 有限元方法已经广泛应用于各种类型的偏微分方程, 成为一种主流数值方法。

在 R.W.Clough (1960) 关于平面弹性力学问题的学术论文中, 有限元方法的名称首次出现。同时, 以冯康院士为首的中国数值计算专家也独立提出了相同的方法, 当时的名称是基于变分原理的差分方法。事实上, 著名的应用数学家 R. Courant 在 1943 年就提出过类似的思想。

### 6.4.1 变分方法的基本理论

对于弹性力学问题，相应的数学描述通常有两种方式，其一是基于牛顿定律导出的微分方程定解问题，其二是基于能量原理导出变分问题。两者相比，后者更能准确反映数学物理问题的本质，因为它具有下面的优点：

1. 关于解函数的可导性要求，微分方程定解问题要强于变分问题。在很多实际问题中，解函数可能无法逐点满足微分方程定解问题，却可以（在积分意义下）整体满足变分问题。
2. 微分方程需要明确给出边界条件，而变分问题可以将边界条件融入到试探函数空间的定义中。因此，变分问题可以较为轻松处理各种边界条件，特别是含有导数的自然边界条件。

为简单起见，本节以二维 Poisson 方程的 Dirichlet 零边值问题为例，介绍椭圆型方程的变分方法，其中  $\Omega$  是二维凸多边形区域。至于自然边界条件，我们将会给予简要说明。

Dirichlet 边界条件也称为本质边界条件，必须出现在相应的函数空间内。引进赋范线性空间<sup>5</sup>

$$\mathcal{H} = \{v : v, v_x, v_y \in L^2(\Omega) \text{ 且 } v|_{\Gamma} = 0\}, \quad (6.50)$$

它是函数空间  $C_0^1(\Omega)$  关于范数

$$\|v\|_{\mathcal{H}} = \left( \|v\|^2 + \|v_x\|^2 + \|v_y\|^2 \right)^{\frac{1}{2}} \quad (6.51)$$

的完备化。由于  $C_0^1(\Omega)$  在  $\mathcal{H}$  中稠密，不妨将  $\mathcal{H}$  粗略地理解为  $C_0^1(\Omega)$ 。

对于模型问题，常用的变分描述主要有两个。因篇幅有限，我们跳过详细的推导过程，直接给出它们的具体形式。

1. Rietz 变分问题：求  $u \in \mathcal{H}$ ，使

$$E(u) = \min_{v \in \mathcal{H}} E(v), \quad (6.52)$$

---

<sup>5</sup>导数应当理解为“广义导数”，边界取值应当理解为“迹”。

其中

$$E(v) = \int_{\Omega} \left[ \frac{1}{2}(v_x^2 + v_y^2) - fv \right] dx. \quad (6.53)$$

它对应物理学的最小势能原理。

2. Galerkin 变分问题: 求  $u \in \mathcal{H}$ , 使得

$$A(u, v) = F(v), \quad \forall v \in \mathcal{H}, \quad (6.54a)$$

其中

$$A(u, v) \equiv \int_{\Omega} \nabla u \cdot \nabla v dx, \quad F(v) \equiv \int_{\Omega} f v dx. \quad (6.54b)$$

它对应物理学的虚功原理。

在 (6.54) 中,  $u$  称为试探函数,  $v$  称为检验函数, 其所属空间分别称为试探函数空间和检验函数空间。由于两个函数空间相同, 故而 (6.54) 称为 Galerkin 变分问题。

同 Rietz 变分问题相比, Galerkin 变分问题具有更加广泛的适用范围。本节将以其为讨论重点。

**定理 6.2. (Lax-Milgram 定理)** 设  $\mathcal{H}$  是以范数  $\|\cdot\|$  为度量的 Banach 空间, 双线性泛函数  $A(w, v): \mathcal{H} \times \mathcal{H} \rightarrow \mathfrak{R}$  满足

1. 对称:  $A(w, v) = A(v, w)$ ;
2. 正定: 存在固定常数  $M_1 > 0$ , 使得  $A(v, v) \geq M_1 \|v\|^2$ ;
3. 有界: 存在固定常数  $M_2 > 0$ , 使得  $|A(w, v)| \leq M_2 \|w\| \|v\|$ .

若  $F(v): \mathcal{H} \rightarrow \mathfrak{R}$  是有界线性泛函, 则 Galerkin 变分问题 (6.54) 的解  $u \in \mathcal{H}$  存在且唯一, 满足先验估计

$$\|u\|_{\mathcal{H}} \leq \frac{1}{M_1} \|F\|_{\mathcal{H}^*} = \frac{1}{M_1} \sup_{v=0} \frac{\|F(v)\|}{\|v\|}.$$

因篇幅限制, 我们跳过 Lax-Milgram 定理的证明, 直接将其用于 Galerkin 变分问题 (6.54)。显然,  $A(\cdot, \cdot)$  是对称的双线性泛函。利用 Cauchy-Schwartz 不等式, 可知

$$A(u, v) \leq \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}, \quad \forall u, v \in \mathcal{H},$$


即  $A(\cdot, \cdot)$  满足有界性。注意到  $u|_{\Gamma} = 0$ , 利用著名的 Poincaré 不等式, 简单计算可知

$$A(u, u) = \|u_x\|^2 + \|u_y\|^2 \geq M_1 \|u\|_{\mathcal{H}}^2.$$

显然,  $F(\cdot)$  是有界线性泛函。因此, Lax-Milgram 定理的条件均满足, Galerkin 变分问题 (6.54) 的解  $u \in \mathcal{H}$  存在且唯一。

**定理 6.3.** 模型问题的古典解必定满足满足 Galerkin 变分问题 (6.54)。反之, 若 Galerkin 变分问题 (6.54) 的解满足  $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$ , 则它也是模型问题的古典解。

因此, Galerkin 变分问题 (6.54) 的解常常定义为模型问题的弱解。在后续的讨论中, 两个概念是完全等同的。

 **注释 6.5.** Neumann 和 Robin 边界条件称为自然边界条件, 因为它们可以直接体现在变分方程中, 不用出现在试探函数空间或者检验函数空间中。若二维 Poisson 方程 (6.1) 具有边界条件

$$\frac{\partial u}{\partial \gamma} + \sigma u = g(x, y),$$

其中  $\sigma \geq 0$  是给定的常数,  $\gamma = (\gamma_1, \gamma_2)$  是单位外法向量, 相应的 Galerkin 变分问题是: 求  $u \in \tilde{\mathcal{H}} \equiv \{v : v, v_x, v_y \in L^2(\Omega)\}$ , 使得

$$\int_{\Omega} \nabla u \nabla v dx dy + \int_{\Gamma} \sigma u v ds = \int_{\Omega} f v dx + \int_{\Gamma} g v ds, \quad \forall v \in \tilde{\mathcal{H}}.$$

因篇幅限制, 具体的推导过程略。

### 6.4.2 古典变分法

由于  $\mathcal{H}$  是无穷维空间, 变分问题 (6.54) 通常是无法准确求解的。古典变分法是较为有效的近似求解方法, 其实现过程如下。设

$$\mathcal{H}_n = \text{span}\{\psi_1(x), \psi_2(x), \dots, \psi_n(x)\} \quad (6.55)$$

是  $\mathcal{H}$  的一个有限维子空间, 其中  $\psi_1(x), \psi_2(x), \dots, \psi_n(x)$  是线性无关的基函数。将变分问题 (6.54) 局限在  $\mathcal{H}_n$  内, 可以得到有限维变分问题: 求  $u_n(x) \in \mathcal{H}_n$ , 使得

$$A(u_n, v_n) = F(v_n), \quad \forall v \in \mathcal{H}_n. \quad (6.56)$$

既然  $u_n(x) \in \mathcal{H}_n$ , 试探函数可以表示为

$$u_n = \sum_{j=1}^n \alpha_j \psi_j(x), \quad (6.57)$$

其中  $\{\alpha_j\}_{j=1}^n$  是待定系数。令检验函数为  $v_n(x) = \psi_j(x), j = 1, 2, \dots, n$ , 则有限维变分问题 (6.56) 等价转化为  $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_n]^\top$  的线性方程组

$$\mathbb{A}\boldsymbol{\alpha} = \mathbf{b}, \quad (6.58)$$

其中

$$\mathbb{A} = \left( A(\psi_j, \psi_i) \right)_{i=1:n}^{j=1:n}, \quad \mathbf{b} = [F(\psi_1), F(\psi_2), \dots, F(\psi_n)]^\top, \quad (6.59)$$

分别称作刚度矩阵和荷载向量。

由于双线性泛函  $A(u, v)$  是对称正定的, 于是刚度矩阵  $\mathbb{A}$  是对称正定的。因此, 线性方程组 (6.58) 存在唯一解  $\boldsymbol{\alpha}$ , 相应的  $u_n \in \mathcal{H}_n$  也是变分问题 (6.56) 的唯一解。

**引理 6.3. (Ceá 引理)** 设  $\mathcal{H}$  为 Banach 空间, 双线性泛函  $A(\cdot, \cdot)$  满足定理 6.2 的有界性和正定性。若  $\mathcal{H}_n$  是  $\mathcal{H}$  的有限维子空间, 则存在仅仅依赖  $M_1$  和  $M_2$  的正常数  $\beta$ , 使得

$$\|u - u_n\|_{\mathcal{H}} \leq \beta \inf_{v \in \mathcal{H}_n} \|u - v\|_{\mathcal{H}}, \quad (6.60)$$

其中  $u$  满足变分问题 (6.54),  $u_n$  满足变分问题 (6.56)。

**证明:** 注意到  $\mathcal{H}_n \subset \mathcal{H}$ , 将 (6.54) 和 (6.56) 相减, 可知误差  $u - u_n$  满足正交性质

$$A(u - u_n, v) = 0, \quad \forall v \in \mathcal{H}_n. \quad (6.61)$$

利用  $A(\cdot, \cdot)$  的正定性和有界性, 有

$$\begin{aligned} M_1 \|u - u_n\|_{\mathcal{H}}^2 &\leq A(u - u_n, u - u_n) = A(u - u_n, u - v) \\ &\leq M_2 \|u - u_n\|_{\mathcal{H}} \|u - v\|_{\mathcal{H}}, \end{aligned} \quad (6.62)$$

其中  $v \in \mathcal{H}$  是任意的函数。取正常数  $\beta = M_2/M_1$ , 即证。  $\square$

**定理 6.4. (投影定理)** 假设引理 6.3 的条件成立, 且双线性泛函  $A(\cdot, \cdot)$  还具有对称性, 则

$$\|u - u_n\| = \inf_{v \in \mathcal{H}_n} \|u - v\|, \quad (6.63)$$

其中  $\|v\| = \sqrt{a(v, v)}$  称做能量模。换言之,  $u_n$  是真解  $u$  在  $\mathcal{H}_n$  的最佳能量模逼近。

**证明:** 定义二元运算

$$[w, v] \equiv A(w, v), \quad w, v \in \mathcal{H}.$$

不难验证, 它是空间  $\mathcal{H}$  的内积, 相应的诱导范数  $\|v\| = [v, v]^{1/2}$  就是能量模。由误差正交性质 (6.61) 可知

$$\begin{aligned} \|u - u_n\|^2 &= A(u - u_n, u - u_n) \\ &= A(u - u_n, u - v), \quad \forall v \in \mathcal{H}_n, \\ &\leq \|u - u_n\| \cdot \|u - v\|, \quad \forall v \in \mathcal{H}_n. \end{aligned} \quad (6.64)$$

注意到  $u_n \in \mathcal{H}_n$ , 定理结论得证。  $\square$

在投影定理 6.4 的条件下, 利用  $A(\cdot, \cdot)$  的正定性和有界性, 由 (6.64) 可以导出比 Céa 引理更加精确的估计

$$\|u - u_n\|_{\mathcal{H}} \leq \sqrt{\frac{M_2}{M_1}} \inf_{v \in \mathcal{H}_n} \|u - v\|_{\mathcal{H}}. \quad (6.65)$$

利用误差的正交性质, 有  $\|u\|^2 = \|u - u_n\|^2 + \|u_n\|^2$ . 因此, 数值解具有“能量单侧逼近”性质, 即

$$\|u_n\| \leq \|u\|. \quad (6.66)$$

这个性质非常适宜弹性静力学的数值计算。

利用 Ceá 引理或者投影定理可知, 古典变分法的误差估计可以转化为相应的函数逼近问题。只要有限维子空间  $\mathcal{H}_n$  具有良好的函数逼近性质, 则古典变分法给出的近似解  $u_n$  就是可靠的。

**定理 6.5.** 假设引理 6.3 的条件成立。若  $\mathcal{H}$  是 Hilbert 空间, 具有完全的正交系  $\{\psi_i\}_{i=1}^\infty$ , 则有限维子空间  $\mathcal{H}_n = \text{span}\{\psi_1, \psi_2, \dots, \psi_n\}$  可以保证  $\lim_{n \rightarrow \infty} \|u - u_n\|_{\mathcal{H}} = 0$ 。

**证明:** 设  $\varepsilon$  是任意的正数。由于  $\{\psi_i\}_{i=1}^\infty$  是  $\mathcal{H}$  的完全正交系, 存在  $N_\varepsilon \in \mathbb{Z}^+$  和  $\{\alpha_i\}_{i=1}^{N_\varepsilon}$ , 使得

$$\left\| u - \sum_{i=1}^{N_\varepsilon} \alpha_i \psi_i \right\|_{\mathcal{H}} \leq \varepsilon / \beta.$$

当  $n \geq N_\varepsilon$ , 有  $\mathcal{H}_{N_\varepsilon} \subset \mathcal{H}_n$ , 由引理 6.3 可得

$$\|u - u_n\|_{\mathcal{H}} \leq \beta \inf_{v \in \mathcal{H}_n} \|u - v\|_{\mathcal{H}} \leq \beta \left\| u - \sum_{i=1}^{N_\varepsilon} \alpha_i \psi_i \right\|_{\mathcal{H}} \leq \varepsilon,$$

即证。 □

由定理 6.5 可知, 若  $\mathcal{H}_n$  是正交的三角多项式空间, 即

$$\mathcal{H}_n = \text{span}\{\sin(\pi x), \sin(2\pi x), \dots, \sin(n\pi x)\},$$

则相应的古典变分法称为谱方法, 保证数值解收敛到真解。

### 6.4.3 标准有限元方法

对于高维椭圆型方程, 古典变分法仍会遇到一些困难, 特别是有限维子空间的构造以及数值格式的计算效率。具体而言, 是



1. 作为函数空间的强制约束, 有限维子空间的基函数需要满足 Dirichlet 边界条件。当计算区域形状复杂的时候, 这个目标是难以实现的。
2. 刚度矩阵和荷载向量涉及大量的积分运算。在古典变分法中, 基函数通常是 (几乎) 处处非零的。从数值计算的角度出发, 它导致两个非常严重的数值困扰:
  - 由于数值积分要遍历整个计算区域, 线性方程组的组装过程需要消耗大量的 CPU 时间。
  - 刚度矩阵通常是稠密的, 含有大量的非零元素。这将导致数据存储的困难, 以及线性方程组的求解代价过高。

基于上述原因, 数值工作者提出不同的基函数构造技术, 在古典变分法的基础上, 发展出很多极富特色的数值方法, 例如有限元方法、配置法和谱 (元) 方法等等。

在有限元方法中, 有限维子空间的基函数是具有紧支集的分片多项式。构造过程如下:

1. 构造区域  $\Omega$  的网格剖分  $\mathcal{T}_h = \{K_\ell\}_{\ell=1}^J$ , 其中  $K_\ell$  是工作单元, 使得

$$(i) \quad \bar{\Omega} = \bigcup_{\ell=1}^J \bar{K}_\ell, \quad (ii) \quad K_{\ell_1} \cap K_{\ell_2} = \emptyset, \ell_1 \neq \ell_2. \quad (6.67)$$

通常, 工作单元具有简单的几何结构, 例如一维的区间、二维的三角形、矩形或者四边形等等。这里, 网格参数  $h$  是单元半径的最大值。

2. 在每个工作单元  $K_\ell$  上, 构造有限次数的多项式空间  $\mathcal{P}(K_\ell)$ ; 按照某种原则, 它们可以整体拼接出有限维空间

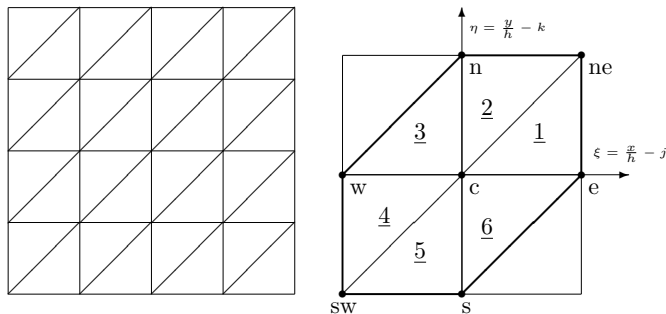
$$\mathcal{H}_n = \prod_{\ell=1}^J \mathcal{P}(K_\ell). \quad (6.68)$$

它就是基于网格剖分  $\mathcal{T}_h$  的有限元空间, 通常记为  $V_h$ ,

若  $V_h \subset \mathcal{H}$ , 则相应的古典变分法称为 (标准) 有限元方法。

对于模型问题而言, 无穷维空间  $\mathcal{H}$  的函数都是连续的<sup>6</sup>。标准有限元方法要求  $V_h \subset \mathcal{H}$ , 相应的网格剖分  $\mathcal{T}_h$  需要具有适当的协调性, 单元顶点不能出现在其它工作单元的边内。换言之, 网格剖分不能出现悬挂点。否则, 在跨越单元边界时, 有限元空间的函数无法保证连续性。

图 6.3: 网格剖分及其基本结构



**论题 6.5.** 设  $\Omega = (0, 1) \times (0, 1)$  是单位正方形, 相应的网格剖分由边长为  $h$  的正方形沿某方向一分为二而成; 参见图 6.3 的左侧。定义线性有限元空间

$$V_h = \{v \in C_0(\Omega): v|_K \text{ 是线性多项式}, \forall K \in \mathcal{T}_h\}, \quad (6.69)$$

建立模型问题的标准有限元方法。

**答:** 任取正方形内部的某个三角形顶点  $c(x_j, y_k)$ , 其中  $x_j = jh$  和  $y_k = kh$ 。相应的基函数  $\psi^c(x, y) \in V_h$  是分片线性多项式, 仅仅在  $c$  点取值为 1, 在其它节点均取值为零。参见图 6.3 的右侧, 粗线围成的六边形区域是基函数  $\psi^c(x, y)$  的支集。

<sup>6</sup>请参阅 Sobolev 空间的嵌入定理。

对应支集内的 6 个三角形单元, 有限元解  $u_h$  和基函数  $\psi^c$  的具体定义如下, 即

单元 $K$ 编号	有限元解 $u_h$	基函数 $\psi^c$
1	$(1 - \xi)u_c + (\xi - \eta)u_e$	$1 - \xi$
2	$(1 - \eta)u_c + (\eta - \xi)u_n + \xi u_{ne}$	$1 - \eta$
3	$(1 + \xi - \eta)u_c + \eta u_n - \xi u_w$	$1 + \xi - \eta$
4	$(1 + \xi)u_c + (\eta - \xi)u_w - \eta u_{sw}$	$1 + \xi$
5	$(1 + \xi)u_c + (\xi - \eta)u_s + \xi u_{sw}$	$1 + \eta$
6	$(1 - \xi + \eta)u_c + \xi u_e - \eta u_s$	$1 - \xi + \eta$

其中  $u_c$  和  $u_e$  等是  $u_h$  在相应节点的函数值,

$$\xi = \frac{x - x_j}{h}, \quad \eta = \frac{y - y_k}{h} \quad (6.70)$$

是局部区域的直角坐标系。

模型问题的标准有限元方法是: 求  $u_h \in V_h$ , 使得

$$A(u_h, v_h) = F(v_h), \quad \forall v_h \in V_h, \quad (6.71)$$

其中  $A(\cdot, \cdot)$  和  $F(\cdot)$  的定义已经在 (6.54) 给出。令  $v_h = \psi^c(x, y)$ , 精确计算 (6.71) 的左端  $A(u_h, \psi^c)$ , 用数值积分公式

$$\int_{\triangle ABC} g(x, y) dx dy \approx \frac{|\triangle ABC|}{3} [g(A) + g(B) + g(C)]$$

近似 (6.71) 的右端  $F(\psi^c)$ 。对应支集内的 6 个三角形单元, 相应的计算结果列表如下, 即


单元 $K$ 编号	$\int_K (u_{h,x} \psi_x^c + u_{h,y} \psi_y^c) dx dy$	$\int_K f \psi^c dx dy$ 的近似
1	$\frac{1}{2}(u_c - u_e)$	$\frac{1}{6} f_c h^2$
2	$\frac{1}{2}(u_c - u_n)$	$\frac{1}{6} f_c h^2$
3	$\frac{1}{2}(2u_c - u_n - u_w)$	$\frac{1}{6} f_c h^2$
4	$\frac{1}{2}(u_c - u_w)$	$\frac{1}{6} f_c h^2$
5	$\frac{1}{2}(u_c - u_s)$	$\frac{1}{6} f_c h^2$
6	$\frac{1}{2}(2u_c - u_e - u_s)$	$\frac{1}{6} f_c h^2$

将 6 个单元的计算结果相加, 即可得到  $c$  点的差分方程

$$4u_c - [u_e + u_n + u_w + u_s] = f_c h^2.$$

它恰好就是五点差分格式 (6.5)。因此说, 有限元方法和有限差分方法具有紧密的联系。□

对于非一致的网格结构或者自然边界条件, 有限元方法的数值操作依旧简单, 数值效果依旧理想。

 **注释 6.6.** 利用有限元方法的误差估计技巧, 可证: 若二维 Poisson 方程定解问题的真解充分光滑, 在适当的网格要求下, 有限元格式均可以达到二阶 ( $L^2$  模或最大模) 误差。详见 [3]。

---

## 第 7 章

# 对流扩散方程的数值方法

---

一个系统可以同时存在对流现象和扩散现象，典型的数学模型是对流扩散方程。以一维问题为例，它通常具有如下形式

$$u_t + cu_x = au_{xx}, \quad (7.1)$$

其中  $c$  是流动速度， $a \geq 0$  是扩散速度。当  $a$  和  $c$  同时非零的时候，简单的数值离散技术可能遇到困难。为简单起见，设  $c$  和  $a$  是给定的常数。

### 7.1 数值困难

设  $\mathcal{T}_{\Delta x, \Delta t}$  是等距时空网格<sup>1</sup>。最自然的设计思路是借鉴前面的数值经验，将已知的导数离散技术结合起来。比如，用中心差商离散两个空间导数，用向前差商离散时间导数，可得 (7.1) 的中心差商显格式

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c(u_{j+1}^n - u_{j-1}^n)}{2\Delta x} = \frac{a\delta_x^2 u_j^n}{(\Delta x)^2}. \quad (7.2)$$

显然，对于任意的  $c$  和  $a$ ，它都无条件具有 (2,1) 阶局部截断误差。

✚ **论题 7.1.** 讨论中心差商显格式 (7.2) 的  $L^2$  模稳定性。

答：设  $k$  是任意的波数。将模态解  $u_j^n = \lambda^n e^{ikj\Delta x}$  代入到 (7.2)，其中  $\lambda = \lambda(k)$  是增长因子。简单计算，可得

$$\lambda(k) = 1 - i\nu c \sin(k\Delta x) - 4\mu a \sin^2\left(\frac{1}{2}k\Delta x\right), \quad (7.3)$$

---

<sup>1</sup>事实上，采用同方程系数相匹配的非均匀网格，数值结果将会更加完美。

其中  $\nu = \Delta t / \Delta x$  称为对流网比,  $\mu = \Delta t / (\Delta x)^2$  称为扩散网比。分离实部和虚部, 有

$$|\lambda(k)|^2 = (1 - 4\mu a s)^2 + 4\nu^2 c^2 s(1 - s), \quad (7.4)$$

其中  $s = \sin^2(k\Delta x/2) \in [0, 1]$ 。取  $s = 1$ , 由  $|\lambda(k)| \leq 1$  可知

$$\mu a \leq 1/2. \quad (7.5)$$

注意到  $4s(1-s) \leq 1$  和  $\nu^2 = \mu\Delta t$ , 由 (7.4) 和 (7.5) 可知 von Neumann 条件成立, 即

$$|\lambda(k)| \leq \left[1 + \frac{1}{2} \frac{c^2}{a} \Delta t\right]^{1/2} \leq 1 + \frac{1}{4} \frac{c^2}{a} \Delta t, \quad \forall k. \quad (7.6)$$

因此, 中心差商显格式 (7.2) 具有  $L^2$  模稳定性, 即

$$\|u^n\|_2 \leq \left[1 + \frac{1}{4} \frac{c^2}{a} \Delta t\right]^n \|u^0\|_2 \leq e^{\frac{1}{4} \frac{c^2 T}{a}} \|u^0\|_2, \quad \forall n : n\Delta t \leq T, \quad (7.7)$$

其中  $T$  是给定的终止时刻。□

注意到 Fourier 方法的理论基础, 由于估计 (7.6) 是可以等号成立的, 于是稳定性表现 (7.7) 是不可改进的。当对流占优 ( $a \ll |c|$ ) 的时候, 右端的界定常数变得非常巨大。换言之, 微小的扰动可能放大到无法忍受的程度, 最终的计算结果完全失去参考价值。因此说, (7.7) 给出的理论结果缺乏实际意义, 因为差分格式的稳定性表现已经严重偏离微分方程的适定性表现。我们需要重新审视稳定性概念, 给出更加合理的符合实际需求的描述。

**定义 7.1.** 若数值解的稳定性表现同真解的适定性表现保持一致, 则称相应的数值格式具有**强稳定性**。

事实上, 对于线性常系数纯扩散 (或纯对流) 问题的差分格式, 前面章节采用的稳定性概念就是强稳定性概念。

**论题 7.2.** 确定中心差商显格式 (7.2) 的  $L^2$  模强稳定性条件。

答：由于对流扩散方程 (7.1) 具有  $L^2$  模不增的性质，格式具有  $L^2$  模强稳定的充要条件是  $|\lambda(k)| \leq 1$  恒成立，即

$$0 \leq (1 - 4\mu as)^2 + 4\nu^2 c^2 s(1 - s) \leq 1, \quad \forall s \in [0, 1].$$

简单推导可知，它等价于

$$(\nu c)^2 \leq 2\mu a \leq 1. \quad (7.8)$$

同 (7.5) 相比，它有一个额外的时空约束条件  $\Delta t \leq 2a/c^2$ 。换言之，当  $a \ll |c|$  时，时间推进的速度将慢得无法接受。  $\square$

利用离散最大模原理可知，中心差商显格式 (7.2) 具有最大模强稳定性的充要条件是

$$\nu|c| \leq 2\mu a \leq 1. \quad (7.9)$$

它也蕴含一个苛刻的时空约束条件  $\Delta x \leq 2a/|c|$ 。当  $a \ll |c|$  时，空间网格必须足够密集。否则，数值解可能产生剧烈的数值震荡，出现明显的（上下）溢出现象；参见图 7.1。

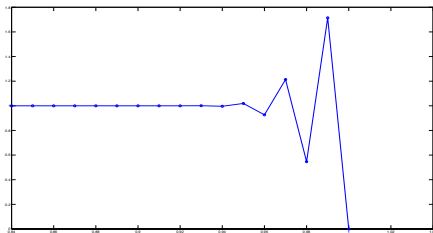


图 7.1: 中心差商显格式的（粗网格上）数值效果

**注释 7.1.** 对流占优扩散问题的数值困难不仅源于强稳定性概念带来的苛刻时空约束条件，还源于真解通常具有的恶劣光滑性表现。例如，它常常具有“固定边界层”或者“移动内层”等突变结构，局部截断误差的表现通常都是

相当糟糕的。相关问题的数值研究主要包括两个方向，其一是设计更理想的格式，其二是构造更合适的网格。

## 7.2 常用的解决方法

为行文简便，假设  $c > 0$ 。当  $a \ll c$  时，对流扩散方程 (7.1) 的数值模拟变得极富挑战性。理想的数值格式应当具有宽松的强稳定性条件，可以基于粗糙的时空网格给出相对准确的数值结果。

### 7.2.1 数值黏性修正方法

最直接的处理方法是借用双曲型方程的一阶导数离散技术，在数值离散过程中引入适当的数值黏性。此时，强稳定性和高精度之间的平衡，将成为数值研究的关键。

1. 直接采用一阶导数的迎风离散机制，定义

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c(u_j^n - u_{j-1}^n)}{\Delta x} = a \frac{\delta_x^2 u_j^n}{(\Delta x)^2}. \quad (7.10)$$

显然，它无条件具有 (1, 1) 阶局部截断误差。

**┆ 论题 7.3.** (7.10) 具有  $L^2$  模强稳定性的充要条件是

$$2\mu a + \nu c \leq 1. \quad (7.11)$$

答：数值格式 (7.10) 的等价形式是

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c(u_{j+1}^n - u_{j-1}^n)}{2\Delta x} = \left[ a + \frac{c\Delta x}{2} \right] \frac{\delta_x^2 u_j^n}{(\Delta x)^2}.$$

利用中心差商显格式 (7.2) 的  $L^2$  模强稳定性条件 (7.8)，可知

$$(\nu c)^2 \leq 2\mu \left[ a + \frac{c}{2} \Delta x \right] \leq 1.$$



左侧是自动成立的，右侧的不等式就是 (7.11)。  $\square$

同中心差商显格式 (7.2) 相比，(7.10) 的时空约束条件变得相当宽松，不用担心  $a/c$  是否过小。

2. 双曲型方程 LW 格式的离散方式，也是可以直接借用的。换言之，首先忽略微分方程的扩散部分，建立双曲部分的 LW 格式；然后填补刚刚忽略的扩散部分，建立相应的二阶中心差商离散，即可建立 (7.1) 的 **修正中心差商显格式**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c(u_{j+1}^n - u_{j-1}^n)}{2\Delta x} = \left[ a + \frac{c^2 \Delta t}{2} \right] \frac{\delta_x^2 u_j^n}{(\Delta x)^2}. \quad (7.12)$$

显而易见，它具有 (2,1) 阶局部截断误差。

**论题 7.4.** (7.12) 具有  $L^2$  模强稳定性的充要条件是

$$2\mu a + (\nu c)^2 \leq 1. \quad (7.13)$$

**答：**借用中心差商显格式 (7.2) 的强稳定性结论 (7.8)，可以直接得到修正中心差商显格式 (7.12) 的  $L^2$  模强稳定性条件

$$(\nu c)^2 \leq 2\mu \left( a + \frac{c^2 \Delta t}{2} \right) \leq 1.$$

注意到  $\mu \Delta t = \nu^2$ ，本论题的结论是显然的。  $\square$

3. 利用精细的设计策略，可以构造出 (7.1) 的指数型格式

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c(u_{j+1}^n - u_{j-1}^n)}{2\Delta x} = a\sigma \frac{\delta_x^2 u_j^n}{(\Delta x)^2}, \quad (7.14a)$$

其中**拟合因子**

$$\sigma = \frac{c\Delta x}{2a} \coth \left( \frac{c\Delta x}{2a} \right) = R \coth R, \quad (7.14b)$$


可以提供额外的数值黏性，

$$R = \frac{c\Delta x}{2a} \quad (7.14c)$$

称为网格 Péclet 数。可以证明, 指数型格式具有  $(2, 1)$  阶局部截断误差, 相应的  $L^2$  模强稳定条件是宽松的

$$\sigma\mu a \leq \frac{1}{2}. \quad (7.15)$$

指数型差分格式还具有很多优点, 例如它的收敛性结论关于  $a$  是一致的, 即误差估计的界定常数同  $a^{-1}$  无关。

 **注释 7.2.** 指数型格式的设计思想具有极高的理论价值。事实上, 它源于稳态方程

$$d + cu_x = au_{xx}$$

的指数型格式, 其中  $d$  是任意给定的常数。我们希望差分方程

$$\alpha u_{j-1} + \beta u_j + \gamma u_{j+1} = d, \quad \forall j,$$

可以在所有网格点上的数值误差都等于零, 其中  $\alpha, \beta$  和  $\gamma$  是同  $d$  无关的待定系数。精确写出微分方程和差分方程通解结构, 即可计算出待定的三个差分系数。略去具体的推导过程, 相应的指数型格式是

$$d + \frac{c(u_{j+1} - u_{j-1})}{2\Delta x} = a\sigma \frac{\delta_x^2 u_j}{(\Delta x)^2},$$

其中  $\sigma$  就是 (7.14b) 定义的拟合因子。最后, 将其推广到发展型方程。令  $d = u_t$ , 利用向前 Euler 差商进行离散, 即可得到对流扩散方程 (7.1) 的指数型格式 (7.14)。

4. 等价变形指数型格式 (7.14), 将一阶导数的中心差商离散改写为向前 (迎风) 差商离散, 有

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c(u_j^n - u_{j-1}^n)}{\Delta x} = a \frac{2R}{e^{2R} - 1} \frac{\delta_x^2 u_j^n}{(\Delta x)^2}.$$


截取指数函数的 Taylor 展开前三项, 即

$$e^{2R} \approx 1 + 2R + 2R^2,$$

则指数型格式 (7.14) 可以简化为 Samapckii 格式

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c(u_j^n - u_{j-1}^n)}{\Delta x} = \frac{a}{1 + R(\Delta x)^2} \delta_x^2 u_j^n. \quad (7.16)$$

可以证明, 它保持指数型格式 (7.14) 的相容阶。同指数型格式相比, Samapckii 格式成功回避了指数函数的计算困难。首先, 当  $a \ll c$  时,  $e^{2R}$  的计算无需特殊的程序代码处理; 其次, 对于变系数或者非线性问题, 大量的指数函数运算时间可以得到明显的节省。

 **注释 7.3.** 若截取指数的 Taylor 级数前两项, 则指数型格式 (7.14) 可以简化为迎风格式 (7.10)。

### 7.2.2 隐式格式

隐式格式也是有效的离散策略, 例如中心全隐格式

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c(u_{j+1}^{n+1} - u_{j-1}^{n+1})}{2\Delta x} = \frac{a\delta_x^2 u_j^{n+1}}{(\Delta x)^2}. \quad (7.17)$$

利用 Fourier 方法可证, 它无条件具有  $L^2$  模强稳定性。当然, 结合前面的数值黏性修正方法, 数值格式可以获得更好的数值效果。

在此强调: 具有  $L^2$  模强稳定性的数值格式依旧可能产生数值震荡, 除非它满足离散最大模原理。基于粗糙的时空网格, 实现离散最大模原理, 也是对流占优扩散问题数值方法的重要课题。因篇幅有限, 详略。

### 7.2.3 算子分裂方法

算子分裂方法可以解决对流占优带来的数值困难, 其实现过程类似于二维热传导方程的 LOD 方法。

例如, 先用 LW 格式离散时间区间  $[t^n, t^{n+1/2}]$  的纯对流问题

$$\frac{1}{2}u_t + cu_x = 0,$$

再用全显格式离散时间区间  $[t^{n+1/2}, t^{n+1}]$  的纯扩散问题

$$\frac{1}{2}u_t = u_{xx},$$

其中  $t^{n+1/2} = (t^n + t^{n+1})/2$ 。因此, 对流扩散方程 (7.1) 的算子分裂格式可以定义为

$$u_j^{n+\frac{1}{2}} = u_j^n - \frac{1}{2}\nu c \left[ u_{j+1}^n - u_{j-1}^n \right] + \frac{1}{2}(\nu c)^2 \delta_x^2 u_j^n, \quad (7.18a)$$

$$u_j^{n+1} = u_j^{n+\frac{1}{2}} + \mu a \delta_x^2 u_j^{n+\frac{1}{2}}. \quad (7.18b)$$

由于两个计算步都具有完美的稳定性结论, 整个格式也自然地具有宽松的  $L^2$  模强稳定性条件

$$\max(\nu|c|, 2\mu a) \leq 1. \quad (7.19)$$

当  $a \approx |c|\Delta x$  时, 时间步长  $\Delta t$  和空间步长  $\Delta x$  是同阶的。

### 7.2.4 特征差分方法

当  $a = 0$  时, 对流扩散方程 (7.1) 退化到双曲型方程  $u_t + cu_x = 0$ , 相应的特征线是  $x - ct$  恒定的直线段。沿着特征线, 引进方向导数或**时间全导数**

$$\bar{D}_t u = \frac{1}{\sqrt{1+c^2}}u_t + \frac{c}{\sqrt{1+c^2}}u_x, \quad (7.20)$$

直角坐标系下的对流扩散方程 (7.1) 可以转化为斜坐标系下的纯扩散问题

$$\sqrt{1+c^2}\bar{D}_t u = au_{xx}. \quad (7.21)$$

这是特征差分方法<sup>2</sup>的构造起点。

---

<sup>2</sup>J. Jr. Douglas and F. R. Thomas, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19:5 (1982), 871-885

下面以网格点  $(x_j, t^{n+1})$  为离散焦点, 离散 (7.21) 的两个导数。利用向后差商离散时间全导数, 可得

$$[\bar{D}_t u]_j^{n+1} = \frac{[u]_j^{n+1} - [\tilde{u}]_j^n}{\sqrt{1+c^2}\Delta t} + \mathcal{O}(\Delta t), \quad (7.22a)$$

其中  $[\tilde{u}]_j^n \equiv u(\tilde{x}_j, t^n)$  且  $\tilde{x}_j = x_j - c\Delta t$  是离散焦点的回溯点。设  $\tilde{x}_j$  落在两个相邻网格点  $x_{j,L} = x_j - \kappa\Delta x$  和  $x_{j,R} = x_{j,L} + \Delta x$  之间, 其中  $\kappa$  是事先给定的正整数。利用线性插值理论, 可知

$$[\tilde{u}]_j^n = \left[1 - \frac{\tilde{x}_j - x_{j,L}}{\Delta x}\right][u]_{j,L}^n + \frac{\tilde{x}_j - x_{j,L}}{\Delta x}[u]_{j,R}^n + \mathcal{O}((\Delta x)^2). \quad (7.22b)$$

采用二阶中心差商离散空间导数, 有

$$[u_{xx}]_j^{n+1} = \frac{\delta_x^2[u]_j^{n+1}}{(\Delta x)^2} + \mathcal{O}((\Delta x)^2). \quad (7.23)$$

综上所述, 略去无穷小量, 用数值解替换真解, 可得对流扩散方程 (7.1) 的特征差分格式

$$u_j^{n+1} = \tilde{u}_j^n + \mu a \delta_x^2 u_j^{n+1}, \quad (7.24a)$$

其中

$$\begin{aligned} \tilde{u}_j^n &= \left[1 - \frac{x_j - c\Delta t - x_{j,L}}{\Delta x}\right]u_{j,L}^n + \frac{x_j - c\Delta t - x_{j,L}}{\Delta x}u_{j,R}^n \\ &= u_{j-\kappa}^n + (\kappa - \nu c)\Delta_{+,x}u_{j-\kappa}^n. \end{aligned} \quad (7.24b)$$

可以证明: 特征差分格式 (7.24) 无条件具有  $L^2$  模稳定性, 数值误差达到整体一阶, 界定常数同真解沿特征线的变化率相关。详略。

当微分方程 (7.1) 是对流占优的, 真解沿特征线的变化极其缓慢。因此, 特征差分格式 (7.24) 可以使用较大的时间步长, 即可获得同样大小的数值误差。

### 7.2.5 有限元方法

常用的数值方法有特征有限元方法、流线扩散方法、最小二乘方法、子网格方法、气泡函数方法等等。

---

## 第 8 章

# 间断有限元方法

---

见综述论文 [5]。

---

## 参 考 资 料

---

- [1] R. J. Leveque, *Finite volume methods for hyperbolic problems*, Cambridge University Press, 2002
- [2] K. W. Morton and D. F. Mayers, *Numerical solutions for partial differential equations*, Cambridge University Press, 2005
- [3] 胡健伟、汤怀民, 微分方程数值方法, 南开大学出版社, 2004
- [4] 张强, 偏微分方程的有限差分方法, 科学出版社, 2018
- [5] B. Cockburn and C.- W. Shu, *Runge-Kutta Discontinuous Galerkin methods for convection-dominated problem*, J. Sci. Comput, 16:3 (2001), pp. 173-261

---

# 附录

---

## A. 修正方程方法<sup>‡</sup>

修正方程方法可以借用已有的偏微分方程理论结果，启发式地判断数值格式的稳定性结论。其原始思想源于 Yanenko 和 Shokin (1969) 的微分逼近方法，现在的名称是由 Warming 和 Hyett<sup>1</sup>给出的。由于简单易行，它的适用范围较为广泛。

操作过程主要包括两步。**第一步至关重要，由差分方程出发，导出含有网格参数的修正方程。**以原有的离散对象为参照对象，差分方程更加相容于修正方程。粗略地将，**修正方程包含了局部截断误差的主项**，相应的推导过程可以理解为局部截断误差推导的相反过程。具体陈述如下：

1. 假设网格函数  $\{u_j^n\}_{j,j}^n$  由光滑函数  $w(x, t)$  限制而成的。逐点进行 Taylor 级数展开，**由差分方程导出一个处处成立的微分恒等式**。通常，它是一个无穷级数，同时含有  $w$  及其导数。
2. 假设微分恒等式可以逐项求导。**依次将不同阶数的时间导数转化为空间导数的无穷级数，进行适当的截断和化简**，最终得到的偏微分方程

$$w_t = \mathcal{L}w \equiv \sum_{\ell=0}^m \alpha_{\ell} D_x^{\ell} w \quad (\text{A.1})$$

就是**修正方程**，其中  $m$  是适当选取的正整数，系数  $\{\alpha_{\ell}\}_{\ell=0}^m$  可能同网格参数相关。

**第二步是启发性的，利用修正方程的性质解释差分方程的数值表现。**具体而言，是


---

<sup>1</sup>R. F. Warming and B. J. Hyett, *The modified equation approach to the stability and accuracy analysis of finite difference methods*, J. Comput. Phys., 14 (1974), 159-179.




1. 若修正方程是不适定的, 则数值格式必然是不稳定的。若修正方程是适定的, 则数值格式可以断言是稳定的。
2. 若修正方程具有耗散或色散效应, 则数值格式具有相近的数值耗散和数值色散效应。

请注意, 类似的思想在 §3.2.1 中曾经出现过。

 **注释 8.1.** 对于线性常系数偏微分方程 (A.1), *Fourier* 理论是非常高效的分析工具。当  $\mathcal{L}$  仅仅包含一个空间导数时, 有:

1. 奇数阶空间导数描述波动现象, 其中一阶导数刻画对流现象, 而其它奇数阶导数刻画色散现象。
2. 偶数阶空间导数描述扩散或反扩散现象。正系数的二阶导数和负系数的四阶导数均刻画扩散现象, 相应的真解不断衰减, 微分系统是适定的。负系数的二阶导数和正系数的四阶导数刻画反扩散现象, 相应的真解不断膨胀, 微分系统是不适定的。

当  $\mathcal{L}$  包含不同阶数的空间导数时, 上述结论可以叠加起来, 最终的结论将略显复杂。详略。

 **论题 8.1.** 考虑对流方程  $u_t + u_x = 0$  的中心差商显格式

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0.$$

利用修正方程方法, 说明它是数值不稳定的。

答: 利用 Taylor 展开技术, 有

$$w_t + w_x + \frac{\Delta t}{2} w_{tt} + \mathcal{O}(\Delta x^2 + \Delta t^2) = 0. \quad (\text{A.2})$$

关于时间和空间求导, 有

$$\begin{aligned} w_{tt} + w_{xt} + \frac{\Delta t}{2} w_{ttt} + \mathcal{O}(\Delta x^2 + \Delta t^2) &= 0, \\ w_{tx} + w_{xx} + \frac{\Delta t}{2} w_{xtt} + \mathcal{O}(\Delta x^2 + \Delta t^2) &= 0. \end{aligned}$$

代入到 (A.2), 有

$$w_t + w_x = -\frac{\Delta t}{2} w_{xx} + \mathcal{O}(\Delta x^2 + \Delta t^2).$$

略去高阶小量, 即得中心差商显格式的修正方程

$$w_t + w_x = -\frac{\Delta t}{2} w_{xx}. \quad (\text{A.3})$$

由于扩散系数为负, 相应的修正方程是不适定的。因此, 中心差商显格式是不稳定的。□

↯ **论题 8.2.** 考虑对流方程  $u_t + u_x = 0$  的迎风格式

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{u_j^n - u_{j-1}^n}{\Delta x} = 0.$$

利用修正方程方法, 给出它的稳定性结果。

答: 利用 Taylor 展开技术, 由迎风格式可知

$$w_t + w_x + \frac{\Delta t}{2} w_{tt} - \frac{\Delta x}{2} w_{xx} + \mathcal{O}(\Delta x^2 + \Delta t^2) = 0. \quad (\text{A.4})$$

关于时间和空间变量, 依次进行求导。代入上式, 可得

$$w_t + w_x + \frac{\Delta t - \Delta x}{2} w_{xx} + \mathcal{O}(\Delta x^2 + \Delta x \Delta t + \Delta t^2) = 0.$$

略去高阶小量, 即得迎风格的修正方程

$$w_t + w_x = \frac{\Delta x - \Delta t}{2} w_{xx}. \quad (\text{A.5})$$

当  $\Delta t > \Delta x$  时, 扩散系数是负的, 修正方程是不适定的, 故而迎风格式是不稳定的。而当  $\Delta t < \Delta x$  时, 修正方程是适定的, 故而迎风格式是稳定的。□

在修正方程方法中, Taylor 级数展开的合理性是至关重要的。否则, 相应的分析结果可能是无意义的。

1. 在双曲型方程的差分格式中, 低频简谐波的数值简谐波更为重要。由于低频简谐波具有较好的光滑度表现, 相应的 Taylor 级数展开是合理的。

2. 在抛物型方程的差分格式中, 数值不稳定现象主要源于高频简谐波。由于高频简谐波可视为光滑度极差的函数, 相应的 Taylor 级数展开缺乏合理性。

因此说, 修正方程的推导需要适当的技巧性和针对性。

**‡ 论题 8.3.** 考虑热传导方程  $u_t = u_{xx}$  的全显格式

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2}.$$

利用修正方程方法, 说明: 当  $\mu > 1/2$  时, 全显格式是不稳定的。

答: 数值解可以分裂为低频成份和高频成份之和, 即

$$u_j^n = v_j^n + (-1)^{j+n} \phi_j^n,$$

其中  $v$  是低频成份,  $(-1)^{j+n} \phi_j^n$  是高频成份, 并且均满足全显格式。在高频成份中, 高频震荡效应主要体现在因子  $(-1)^{j+n}$  中, 而  $\{\phi_j^n\}_{v_j}^n$  对应某个光滑函数。显然, 成立差分方程

$$-\phi_j^{n+1} = \phi_j^n - \mu [\phi_{j-1}^n + 2\phi_j^n + \phi_{j+1}^n].$$

逐点实施 Taylor 级数展开技术, 可得修正方程

$$\phi_t = \frac{2(2\mu - 1)}{\Delta t} \phi.$$

为行文简便, 这里继续沿用旧的符号。简单计算可知, 当  $\mu > 1/2$  时, 修正方程的解  $\phi$  将以指数方式增长到无穷。因此说, 相应的全显格式是不稳定的。□

## B. 能量方法<sup>‡</sup>

能量方法曾经在 § 2.5.3 介绍过。它是一种普适的分析方法, 可以建立  $L^2$  模稳定的充分条件。本节系统介绍能量方法的分析技巧, 给出更多的具体实例。

不同于偏微分方程的能量方法, 此时的研究对象不是函数的微积分运算, 而是离散数据的差分与求和运算。为简单起见, 设

$$\mathcal{T}_{\Delta x} = \{x_j = j\Delta x\}_{j=0}^J$$

是区间  $[0, 1]$  的等距空间网格, 其中  $\Delta x = 1/J$  是空间步长。对于任意的网格函数

$$u = \{u_j\}_{j=0}^J, \quad v = \{v_j\}_{j=0}^J,$$

定义各种类型的离散内积和诱导范数, 例如<sup>2</sup>

$$\begin{aligned} \langle u, v \rangle &= \sum_{j=1}^{J-1} u_j v_j \Delta x, \quad \|u\|_2 = \langle u, u \rangle^{\frac{1}{2}}, \\ \langle u, v \rangle &= \sum_{j=1}^J u_j v_j \Delta x, \quad \|u\|_2 = \langle u, u \rangle^{\frac{1}{2}}, \\ [u, v] &= \sum_{j=0}^{J-1} u_j v_j \Delta x, \quad \|u\|_2 = [u, u]^{\frac{1}{2}}, \\ [u, v] &= \sum_{j=0}^J u_j v_j \Delta x, \quad \|u\|_2 = [u, u]^{\frac{1}{2}}. \end{aligned} \tag{B.6}$$

它们均可看作  $L^2(0, 1)$  内积及其  $L^2$  范数的离散表示。平行于分部积分公式, 可以建立各种版本的分部求和公式, 例如

$$\langle \Delta_+ u, v \rangle = -\langle u, \Delta_- v \rangle + u_J v_J - u_1 v_0, \tag{B.7}$$

其中  $\Delta_{\pm}$  是向前（后）差分算子, 即

$$\Delta_+ u_j = u_{j+1} - u_j, \quad \Delta_- u_j = u_j - u_{j-1}.$$

在此基础上, 依次建立 Green 公式的离散版本

$$\begin{aligned} &\langle \Delta_+(a\Delta_- u), v \rangle \\ &= -\langle a\Delta_- u, \Delta_- v \rangle + a_J \Delta_- u_J v_J - a_1 \Delta_- u_1 v_0, \end{aligned} \tag{B.8a}$$

<sup>2</sup>有时候, 网格函数的边界点值默认为零。

和

$$\begin{aligned} & \langle \Delta_+(a\Delta_-u), v \rangle - \langle \Delta_+(a\Delta_-v), u \rangle \\ &= a_J(v\Delta_-u_J - u\Delta_-v_J) - a_1(v\Delta_+u_0 - u\Delta_+v_0), \end{aligned} \quad (\text{B.8b})$$

其中  $a$  是已知的网格函数。事实上，上述三个公式的本质都是求和运算的次序重排。证明是简单的，略。

能量方法的技术路线是基本相同的，用到的分析工具较多。例如，下面的不等式是经常使用的。

1.  $\varepsilon$ - $ab$  不等式：

$$|ab| \leq \varepsilon a^2 + \frac{1}{4\varepsilon} b^2, \quad \varepsilon > 0. \quad (\text{B.9})$$

2. Cauchy-Schwartz 不等式：

$$|\langle u, v \rangle| \leq \|u\|_2^{\frac{1}{2}} \|v\|_2^{\frac{1}{2}}, \quad |\langle u, v \rangle| \leq \|u\|_2^{\frac{1}{2}} \|v\|_2^{\frac{1}{2}}, \quad \dots \quad (\text{B.10})$$

3. 各种离散范数的相互控制关系：

比如，当网格函数满足  $u_0 = u_J = 0$  时，有

$$\frac{1}{2} \|\Delta_+ u\|_2 \leq \|u\|_2 \leq \frac{L}{\sqrt{8}\Delta x} \|\Delta_+ u\|_2. \quad (\text{B.11})$$

左端是显然的，右端是 Poincaré 不等式的离散描述。证明是简单的，留作练习。

更多内容不再赘述，可参见 [3] 和相关文献。

**‡ 论题 8.4.** 设  $a(x, t)$  恒大于零且  $a_x(x, t)$  有界。讨论纯初值问题的迎风格式

$$u_j^{n+1} = \nu a_j^n u_{j-1}^n + (1 - \nu a_j^n) u_j^n \quad (\text{B.12})$$

的  $L^2$  模稳定性，其中  $\nu = \Delta t / \Delta x$  为网比。

答: 记  $A = \max a(x, t)$  和  $B = \max |a_x(x, t)|$ 。当 CFL 条件

$$A\nu \leq 1$$

满足时, 差分方程的右端系数  $\nu a_j^n$  和  $1 - \nu a_j^n$  都是非负的。平方 (B.12) 的两端。因为平方函数是凸的, 利用 Jensen 不等式可得

$$(u_j^{n+1})^2 \leq \nu a_j^n (u_{j-1}^n)^2 + (1 - \nu a_j^n) (u_j^n)^2.$$

注意到  $|a_j^n - a_{j-1}^n| \leq B\Delta t$  恒成立, 有

$$(u_j^{n+1})^2 \leq \nu a_{j-1}^n (u_{j-1}^n)^2 + (1 - \nu a_j^n) (u_j^n)^2 + B\nu (u_{j-1}^n)^2 \Delta t.$$

在所有空间网格点上求和。适当平移空间指标, 有

$$\|u^{n+1}\|_2^2 \leq (1 + B\nu\Delta t) \|u^n\|_2^2, \quad \forall n.$$

因此, 迎风格式具有  $L^2$  模稳定性结论

$$\|u^n\|_2^2 \leq (1 + B\nu\Delta t)^n \|u^0\|_2^2 \leq e^{B\nu T} \|u^0\|_2^2, \quad (\text{B.13})$$

其中  $T > 0$  是终止时刻。  $\square$

能量方法也可用于数值边界条件的设置, 确保数值格式具有  $L^2$  模稳定性。下面给出一个简单实例。

考虑对流方程  $u_t + u_x = 0$  的初边值问题, 其中空间区间是  $(0, 1)$ , 入流边界条件是  $u(0, t) = 0$ 。显然, 真解的  $L^2$  模不增, 即

$$\int_0^1 u^2(x, t) dx \leq \int_0^1 u^2(x, 0) dx.$$

设时空网格  $\mathcal{T}_{\Delta x, \Delta t} = \mathcal{T}_{\Delta x} \times \mathcal{T}_{\Delta t}$  是等距的, 其中  $\Delta x = 1/J$  是空间步长,  $\Delta t$  是时间步长, 相应的离散网格是

$$\mathcal{T}_{\Delta x} = \{x_j = j\Delta x\}_{j=0}^J, \quad \mathcal{T}_{\Delta t} = \{t^n = n\Delta t\}_{n \geq 0}.$$

在内部空间网格点, 蛙跳格式定义为

$$u_j^{n+1} = u_j^{n-1} - \nu(u_{j+1}^n - u_{j-1}^n) = 0, \quad j = 1 : J-1. \quad (\text{B.14})$$

其中  $u_0^n = 0$  是入流边界条件。

**‡ 论题 8.5.** 给出蛙跳格式 (B.14) 的人工出流边界条件, 使其在  $\nu \leq \nu_0 < 1$  的条件下依旧具有  $L^2$  模稳定性。

**答:** 在 (B.14) 的两端同乘  $u_j^{n+1} + u_j^{n-1}$ , 将位于内部空间网格点的恒等式叠加起来。利用分部求和公式进行整理, 可得

$$\begin{aligned} \|u^{n+1}\|_2^2 - \|u^{n-1}\|_2^2 &= -\nu \left\langle u_j^{n+1} + u_j^{n-1}, \Delta_{0x} u_j^n \right\rangle \\ &= -\nu \left\langle u_j^{n+1}, \Delta_{0x} u_j^n \right\rangle' + \nu \left\langle u_j^n, \Delta_{0x} u_j^{n-1} \right\rangle' + \Pi, \end{aligned}$$

其中  $\|u^n\|$  的定义参见 (B.6), 表达式  $\langle \cdot, \cdot \rangle'$  剔除了  $\langle \cdot, \cdot \rangle$  的出流边界点信息, 而

$$\begin{aligned} \Pi &= \nu \left( u_{J-1}^n u_J^{n-1} - u_{J-1}^{n+1} u_J^n \right) \Delta x - \Delta t \left( u_{J-1}^{n-1} u_J^n + u_J^{n-1} u_{J-1}^n \right) \\ &= -\Delta t \left( u_{J-1}^{n-1} + u_{J-1}^{n+1} \right) u_J^n \end{aligned} \quad (\text{B.15})$$

是同出流边界点信息相关的其余部分。若定义人工出流边界条件

$$u_J^n = \frac{1}{2}(u_{J-1}^{n-1} + u_{J-1}^{n+1}), \quad (\text{B.16})$$

显然成立  $\Pi \leq 0$ 。因此, 有

$$S^n \equiv \|u^{n+1}\|_2^2 + \|u^n\|_2^2 + \nu \left\langle u_j^{n+1}, \Delta_{0x} u_j^n \right\rangle' \leq S^{n-1} \leq \dots \leq S^0.$$

注意到  $|\langle u_j^{n+1}, \Delta_{0x} u_j^n \rangle'| \leq \|u^{n+1}\|_2^2 + \|u^n\|_2^2$ , 有


$$S^n \geq (1 - \nu) \left( \|u^{n+1}\|_2^2 + \|u^n\|_2^2 \right).$$

于是, 当  $|\nu| \leq \nu_0 < 1$  时, 蛙跳格式具有  $L^2$  模稳定性。  $\square$

联立蛙跳格式, 可知人工出流边界条件 (B.16) 等价于

$$u_J^n = \frac{2}{2 + \nu a} u_{J-1}^{n-1} + \frac{\nu a}{2 + \nu a} u_{J-2}^n. \quad (\text{B.17})$$

这才是实际可行的人工边界条件。

 **注释 8.2.** 能量方法也适用于线性变系数问题和非线性问题。除稳定性分析之外, 它也可用于格式的  $L^2$  模误差估计。因篇幅有限, 详略。