

博士资格考试复习资料

# 计算与应用数学博士资格考试复习资料

误差·逼近·非线性·线性代数·ODE ·PDE ·渐近·优化

本资料面向博士资格考试复习。在读者具备基础微积分、线性代数和概率知识的前提下，可独立用于系统复习、证明训练和历年真题准备。本书覆盖八个考纲模块和六期正式试卷。数值 PDE 严格以 24 张板书、23 个独立黑板状态、30 个 training\_core 与 176 个公式级单元为不可回退基线；历史超纲题保留题面与唯一解答并标明身份。

2026 年 7 月

# 使用说明

本书先建立误差、条件数、稳定性与收敛语言，再进入七个算法模块。所有稳定性结论都必须指明方程、离散格式、范数和网格族；所有收敛率都必须说明光滑性、步长和起始误差条件。2026 春 Q2(c) 按打印题面保留，并明确标记原命题不成立及给出反例。

## 考试范围提醒

数值 PDE 当前培训主线与历史真题范围分开标注。谱方法、有限元、三阶色散和完整 Poisson 等历史题不冒充最新培训核心，但其正式题仍保留完整解答。

# 目录

|                                         |    |
|-----------------------------------------|----|
| 使用说明                                    | ii |
| 第一部分 理论、算法与考试主干                         | 1  |
| 第一章 数值分析的共同语言                           | 2  |
| 1.1 从精确问题到实际输出                          | 2  |
| 1.2 条件数：问题对数据的敏感性                       | 3  |
| 1.3 浮点模型与舍入传播                           | 4  |
| 1.4 稳定性与收敛：同名词的三个语境                     | 5  |
| 1.5 非线性问题的绝对与相对条件数                      | 6  |
| 1.6 后向稳定如何与条件数合成前向误差                    | 7  |
| 1.7 消去误差与等价公式重写                         | 7  |
| 1.8 三种稳定性的可执行对照                         | 8  |
| 第二章 插值、逼近与数值积分                          | 9  |
| 2.1 Lagrange 与 Newton 插值                | 9  |
| 2.2 节点选择与 Runge 现象                      | 10 |
| 2.3 Hermite 插值与三次样条                     | 10 |
| 2.4 正交投影与最佳平方逼近                         | 11 |
| 2.5 数值积分：精度、误差与复化                       | 12 |
| 2.6 Gauss 求积：为什么能达到 $2n - 1$ 次          | 13 |
| 2.7 正式真题一：周期复化梯形                        | 13 |
| 2.8 正式真题二：Chebyshev 极小与 Gauss–Chebyshev | 14 |
| 2.9 正式真题三：首一 Legendre、插值与 Gauss         | 16 |
| 2.10 插值余项的完整来源                          | 17 |
| 2.11 Lebesgue 常数与节点选择                   | 18 |
| 2.12 Hermite 插值与重节点条件                   | 18 |
| 2.13 三次样条的方程从哪里来                        | 19 |
| 2.14 Gauss 求积的节点、权重与正性                  | 19 |
| 2.15 复化公式与 Euler–Maclaurin 视角           | 20 |
| 2.16 等距三角插值、DFT 与 FFT                   | 21 |
| 2.17 有理逼近与 Padé 构造                      | 22 |
| 第三章 非线性方程                               | 24 |
| 3.1 二分法：慢但可证的全局基准                       | 24 |
| 3.2 不动点迭代与压缩映射                          | 24 |
| 3.3 收敛阶与误差常数                            | 25 |
| 3.4 Newton 法：局部二次收敛的完整链                 | 25 |
| 3.5 割线法与黄金阶                             | 26 |
| 3.6 无导数 Newton 型方法                      | 27 |
| 3.7 方程组 Newton 与线性化误差                   | 27 |
| 3.8 全局化、停止准则与条件性                        | 28 |
| 3.9 正式真题一：Newton 与割线法                   | 28 |
| 3.10 正式真题二：重根 Newton                    | 29 |
| 3.11 正式真题三：无导数 Newton 型迭代               | 29 |
| 3.12 Newton–Kantorovich 的可验证局部条件        | 31 |
| 3.13 阻尼 Newton 与充分下降                    | 31 |
| 3.14 方程组 Newton 的误差与线性求解                | 31 |
| 3.15 Steffensen 的 $C^2$ 边界              | 32 |
| 第四章 数值线性代数                              | 33 |
| 4.1 范数、条件数与稳定性                          | 33 |
| 4.2 直接法与分解                              | 33 |

|            |                                               |           |
|------------|-----------------------------------------------|-----------|
| 4.3        | Jacobi、Gauss–Seidel 与 SOR                     | 33        |
| 4.4        | Krylov 与共轭梯度                                  | 33        |
| 4.5        | 特征值、最小二乘与 SVD                                 | 34        |
| 4.6        | 向量/矩阵范数与谱半径                                   | 34        |
| 4.7        | 扰动、条件数、残量和后向稳定                                | 35        |
| 4.8        | Gaussian 消元、LU 与主元                            | 35        |
| 4.9        | Cholesky、QR 与最小二乘                             | 36        |
| 4.10       | 定常迭代的充要条件与充分条件                                | 36        |
| 4.11       | CG 的推导、正交性、最小化与有限终止                           | 37        |
| 4.12       | 幂法、反幂、Rayleigh 商与 QR                          | 38        |
| 4.13       | 特征值扰动、Gershgorin 与 SVD                        | 39        |
| 4.14       | 结构化更新、矩阵逆迭代与广义特征值                             | 40        |
| 4.15       | 真题迁移、方法比较与检查                                  | 41        |
| 4.16       | 非正规矩阵：特征值不足以描述敏感性                             | 42        |
| 4.17       | LU、主元与后向误差                                    | 42        |
| 4.18       | QR、最小二乘与秩亏                                    | 43        |
| 4.19       | Jacobi、Gauss–Seidel 与 SOR                     | 43        |
| 4.20       | 一致排序与 SOR 最优松弛因子                              | 45        |
| 4.21       | CG 的 Galerkin 条件与多项式误差                        | 46        |
| 4.22       | 前处理与 GMRES                                    | 47        |
| 4.23       | 特征值算法的位移与商                                    | 47        |
| <b>第五章</b> | <b>数值常微分方程</b>                                | <b>49</b> |
| 5.1        | IVP 适定性与连续依赖                                  | 49        |
| 5.2        | 一步法：一致性、稳定性与整体误差                              | 49        |
| 5.3        | Runge–Kutta 阶条件与稳定函数                          | 50        |
| 5.4        | 为什么需要多步法                                      | 51        |
| 5.5        | 从插值与积分构造 Adams 方法                             | 51        |
| 5.6        | BDF：对解插值后再求导                                  | 52        |
| 5.7        | 局部截断误差与阶条件                                    | 54        |
| 5.8        | 零稳定与根条件                                       | 55        |
| 5.9        | Dahlquist 等价定理：相容加稳定等于收敛                      | 56        |
| 5.10       | 绝对稳定：测试方程上的另一组根                               | 58        |
| 5.11       | 方法比较、刚性与考试母题                                  | 60        |
| 5.12       | 一套可执行的答题流程                                    | 61        |
| 5.13       | 自测与简答                                         | 61        |
| 5.14       | 边值问题、方法比较与历年迁移                                | 62        |
| 5.15       | Runge–Kutta 方法的树前阶条件                          | 64        |
| 5.16       | 稳定函数与 A/L 稳定                                  | 64        |
| 5.17       | 局部误差到全局误差的一步法证明                               | 65        |
| 5.18       | 刚性：精度允许的步长与稳定允许的步长                            | 65        |
| 5.19       | 边值问题：打靶与有限差分                                  | 65        |
| <b>第六章</b> | <b>数值偏微分方程：30 个培训核心与 176 个公式单元</b>            | <b>67</b> |
| 6.1        | 网格、模板、层数与统一参数                                 | 67        |
| 6.2        | Taylor 展开、相容、稳定与 Lax 等价                       | 68        |
| 6.3        | 离散 Fourier–Parseval 与矩阵幂判据                    | 69        |
| 6.4        | 常系数热方程：FTCS、后向 Euler 与 Crank–Nicolson         | 71        |
| 6.5        | Richardson 与 Du Fort–Frankel 的完整根分析           | 71        |
| 6.6        | Schrödinger 实虚部系统与严格 Jordan 临界                | 72        |
| 6.7        | 输运格式逐项 Fourier 比较与 CFL 必要条件                   | 74        |
| 6.8        | 变系数非散度型扩散：显式与 CN 型格式                          | 76        |
| 6.9        | 散度型变系数扩散：半网格通量与完整误差链                          | 77        |
| 6.10       | 离散最大值原理、唯一性与比较定理                              | 78        |
| 6.11       | 非线性扩散的中值定理凸组合                                 | 79        |
| 6.12       | 离散内积、范数、分部求和与 Green 公式                        | 79        |
| 6.13       | 热方程 FTCS 的完整离散能量估计                            | 80        |
| 6.14       | 有限区间 leapfrog：显式能量、边界通量与闭合                    | 81        |
| 6.15       | 板书主线、真题映射与复习检查                                | 83        |
| 6.16       | 板书范围的状态化阅读                                    | 83        |
| 6.17       | FTCS、后向 Euler 与 Crank–Nicolson 的统一 Fourier 推导 | 84        |
| 6.18       | Richardson 与 Du Fort–Frankel 的三层根             | 84        |
| 6.19       | 输运格式的耗散与色散                                    | 85        |
| 6.20       | 二次根、半单性与一致幂有界                                 | 85        |
| 6.21       | 离散能量与边界通量                                     | 86        |
| 6.22       | 变系数通量形式与最大值原理                                 | 86        |

|                               |            |
|-------------------------------|------------|
| <b>第七章 渐近分析</b>               | <b>87</b>  |
| 7.1 渐近序列、展开与误差                | 87         |
| 7.2 量纲分析、无量纲化与定量模型边界          | 87         |
| 7.3 Laplace 方法与 Watson 引理     | 88         |
| 7.4 驻相、最速下降与 WKB              | 88         |
| 7.5 边界层与多尺度                   | 88         |
| 7.6 渐近符号、渐近序列与唯一性             | 88         |
| 7.7 主导平衡、正则与奇异摄动              | 89         |
| 7.8 Watson 引理与端点贡献            | 89         |
| 7.9 Laplace 方法：端点与内点          | 90         |
| 7.10 非驻相衰减与驻相法                | 90         |
| 7.11 最速下降法的操作链                | 90         |
| 7.12 正则摄动、可解性与特征值修正           | 91         |
| 7.13 Poincaré–Lindstedt 与多重尺度 | 91         |
| 7.14 WKB、转向点与量子化              | 91         |
| 7.15 边界层、匹配与复合展开              | 92         |
| 7.16 方法选择、真题迁移与检查             | 92         |
| 7.17 渐近展开的运算规则与失效边界           | 93         |
| 7.18 Watson 引理的缩放证明           | 93         |
| 7.19 Laplace 方法的内点最大值         | 93         |
| 7.20 驻相法与非驻相端点展开              | 94         |
| 7.21 边界层匹配的完整例子               | 95         |
| 7.22 多尺度消除久期项                 | 95         |
| <b>第八章 优化：凸性、对偶、极集与算法</b>     | <b>97</b>  |
| 8.1 凸性、最优性与梯度下降               | 97         |
| 8.2 约束、Lagrange、对偶与 KKT       | 97         |
| 8.3 投影、近端、罚函数与算法边界            | 97         |
| 8.4 凸集、凸函数、强凸与光滑              | 97         |
| 8.5 一阶/二阶最优性与次梯度              | 98         |
| 8.6 梯度下降的两个基本收敛率              | 99         |
| 8.7 线搜索、Newton、拟 Newton 与信赖域  | 99         |
| 8.8 投影、近端与复合优化                | 101        |
| 8.9 约束资格、KKT、对偶与 Slater       | 102        |
| 8.10 线性规划、极集与掌握层级             | 102        |
| 8.11 同伦方法与预测—校正路径跟踪           | 103        |
| 8.12 优化问题的固定点形式               | 104        |
| 8.13 动态规划与 Bellman 最优性原理      | 104        |
| 8.14 方法比较、真题迁移与终章清单           | 105        |
| 8.15 一阶条件、二阶条件与约束几何           | 106        |
| 8.16 梯度下降的步长与精确二次分析           | 106        |
| 8.17 线搜索、Newton 与拟 Newton     | 107        |
| 8.18 对偶函数与弱/强对偶               | 107        |
| 8.19 投影、近端与罚函数的区别             | 108        |
| 8.20 极集与双极定理的证书               | 108        |
| <b>第二部分 六期正式试卷与完整解答</b>       | <b>110</b> |
| <b>第九章 2023 年秋季正式试卷</b>       | <b>111</b> |
| <b>第十章 2024 年春季正式试卷</b>       | <b>115</b> |
| <b>第十一章 2024 年秋季正式试卷</b>      | <b>119</b> |
| <b>第十二章 2025 年春季正式试卷</b>      | <b>123</b> |
| <b>第十三章 2025 年秋季正式试卷</b>      | <b>126</b> |
| <b>第十四章 2026 年春季正式试卷</b>      | <b>129</b> |
| <b>第十五章 跨年度题型索引与答题策略</b>      | <b>132</b> |

## 第一部分

### 理论、算法与考试主干

# 第一章 数值分析的共同语言

## 学习目标

本章统一后续各模块反复使用的语言。读者应能区分数据误差、舍入误差与截断误差；计算前向/后向条件数；解释“问题病态”和“算法不稳定”的差别；并针对线性代数、ODE、PDE 三类语境分别写清 consistency、stability 和 convergence。

## 1.1 从精确问题到实际输出

把数学问题写成映射

$$F: \mathcal{X} \rightarrow \mathcal{Y}, \quad y = F(x).$$

计算机实际拿到的是扰动数据  $\tilde{x}$ ，算法在有限精度下输出  $\hat{y}$ 。总误差至少有三层：

1. 数据误差：  $\tilde{x} - x$ ；
2. 离散/截断误差：连续问题被有限维近似时产生；
3. 舍入与算法误差：浮点执行没有精确实现所写算式。

定义 1.1 (前向与后向误差). 前向误差比较输出与精确答案：

$$\text{forward error} = \|\hat{y} - F(x)\|.$$

后向误差寻找最小输入扰动，使计算结果成为扰动问题的精确解：

$$\eta_{\text{back}} = \inf\{\|\Delta x\| : \hat{y} = F(x + \Delta x)\}.$$

相对误差把分母换成  $\|F(x)\|$  或  $\|x\|$ ，但分母接近零时应改用绝对误差或混合尺度。

后向稳定算法的含义是：它精确解了一个与原问题很接近的问题。它不保证前向误差小；若问题本身病态，小的输入后向误差也会被放大。

例题 1.1 (求根的前向/后向视角). 计算  $f(x) = 0$  得近似根  $\hat{x}$ 。前向误差是  $|\hat{x} - x^*|$ ，但需要知道真根。残量  $|f(\hat{x})|$  可解释为把方程右端从 0 扰动成  $f(\hat{x})$  的后向误差。若  $f'(x^*)$  很小，

$$\hat{x} - x^* \approx \frac{f(\hat{x})}{f'(x^*)},$$

残量很小仍可能对应很大的前向误差。

## 1.2 条件数：问题对数据的敏感性

**定义 1.2** (局部绝对与相对条件数). 若  $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$  在  $x$  可微, 则在给定范数下

$$\kappa_{\text{abs}}(x) = \|DF(x)\|, \quad \kappa_{\text{rel}}(x) = \frac{\|DF(x)\| \|x\|}{\|F(x)\|},$$

后式要求分母非零。更精确的方向条件数是  $\|DF(x)\Delta x\| / \|\Delta x\|$ , 它依赖扰动方向。

**证明路线** 由一阶展开

$$F(x + \Delta x) - F(x) = DF(x)\Delta x + o(\|\Delta x\|),$$

取上确界便得到局部绝对放大因子；再除以输入、输出尺度得到相对条件数。

**例题 1.2** (标量函数的缩放). 对  $F(x) = x^p$ 、 $x \neq 0$ ,

$$\kappa_{\text{rel}}(x) = \frac{|px^{p-1}||x|}{|x|^p} = |p|.$$

而  $F(x) = x - c$  的相对条件数为  $|x|/|x - c|$ , 当  $x \approx c$  时巨大。这解释了相近数相减的结果为何对输入相对扰动敏感；并非所有“消去”都由算法造成, 问题的相对尺度本身也可能病态。

### 1.2.1 线性方程的条件数

对非奇异  $Ax = b$ , 固定  $A$  只扰动  $b$ :

$$A\Delta x = \Delta b, \quad \frac{\|\Delta x\|}{\|x\|} \leq \|A^{-1}\| \frac{\|\Delta b\| \|b\|}{\|b\| \|x\|} \leq \kappa(A) \frac{\|\Delta b\|}{\|b\|},$$

其中

$$\kappa(A) = \|A\| \|A^{-1}\|.$$

在 2-范数下  $\kappa_2(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$ 。若  $\sigma_{\min}$  很小, 某些右端方向会被大幅放大。

若  $A$  也扰动, 设  $(A + \Delta A)(x + \Delta x) = b + \Delta b$ 。忽略二阶项,

$$A\Delta x \approx \Delta b - \Delta A x,$$

故相对误差同时受  $\|\Delta b\|/\|b\|$  与  $\|\Delta A\|/\|A\|$  控制；严格界还需  $\|A^{-1}\Delta A\| < 1$  以保证扰动矩阵仍可逆。

**定理 1.1** (残量到前向误差). 令  $A$  非奇异,  $\hat{x}$  的残量  $r = b - A\hat{x}$ 。则

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}.$$

**证明.**  $A(x - \hat{x}) = r$ , 故  $\|x - \hat{x}\| \leq \|A^{-1}\| \|r\|$ ; 同时  $\|b\| = \|Ax\| \leq \|A\| \|x\|$ 。两式相乘即得。□

### 1.3 浮点模型与舍入传播

设基数为  $\beta$ 、有效位数为  $t$ 。忽略上溢/下溢时，正确舍入满足

$$\text{fl}(x) = x(1 + \delta), \quad |\delta| \leq u,$$

其中  $u$  为 unit roundoff。对基本运算  $\circ \in \{+, -, \times, /\}$ ，常用模型为

$$\text{fl}(x \circ y) = (x \circ y)(1 + \delta), \quad |\delta| \leq u,$$

前提是精确结果在正常浮点范围内。

连续  $n$  次乘法中的误差可收拢为

$$\prod_{i=1}^n (1 + \delta_i) = 1 + \theta_n, \quad |\theta_n| \leq \gamma_n := \frac{nu}{1 - nu},$$

条件是  $nu < 1$ 。证明用  $\prod(1 + |\delta_i|) \leq (1 + u)^n$  再比较几何级数； $\gamma_n$  记号比粗写  $O(nu)$  更便于跟踪常数。

**例题 1.3** (灾难性消去与稳定改写)。当  $x$  很小时，直接计算

$$\sqrt{1+x} - 1$$

会相减两个接近数。代数有理化得

$$\sqrt{1+x} - 1 = \frac{x}{\sqrt{1+x} + 1},$$

右式避免丢失有效数字。两式在实数算术中等价，但浮点误差传播不同。

**例题 1.4** (二次方程的小根)。若  $b > 0$ 、 $b^2 \gg 4ac$ ，公式

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

含严重消去。先稳定计算大根

$$x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a},$$

再用  $x_1 = c/(ax_2)$ 。若  $b < 0$  则选择相反符号。稳定实现必须根据符号避免相减近数，而不是机械使用同一个公式。

#### 易错点

“出现相减”不等于算法必然不稳定。关键是相减结果的尺度、输入误差与后续放大。例如计算本来就很小的物理差量时，相对条件数可能天然很大；算法分析必须把问题条件性和浮点实现分开。

## 1.4 稳定性与收敛：同名词的三个语境

### 1.4.1 线性代数：后向稳定

直接法常以

$$(A + \Delta A)\hat{x} = b + \Delta b$$

为稳定性证书。若  $\|\Delta A\|/\|A\|$ 、 $\|\Delta b\|/\|b\| = O(u)$ ，称算法后向稳定。结合条件数才得到前向误差

$$\frac{\|\hat{x} - x\|}{\|x\|} \lesssim \kappa(A)u.$$

### 1.4.2 ODE：零稳定与误差传播

线性多步法的 consistency 是局部截断误差趋零；zero-stability 是齐次误差递推对启动扰动的幂有界性；两者共同推出固定时间区间上的 convergence。绝对稳定则是测试方程  $y' = \lambda y$  上对特定  $z = h\lambda$  的衰减性质，与 zero-stability 不是同一概念。

### 1.4.3 PDE：一致性、稳定性与 Lax 等价

对线性适定初值问题，离散算子  $L_h$  的 consistency 表示把光滑真解代入后残量趋零；stability 表示离散解算子在所选范数下对数据/强迫一致有界；convergence 表示数值解趋于真解。在线性适定框架和合适一致性条件下，Lax 等价定理给

$$\text{consistency} + \text{stability} \iff \text{convergence}.$$

非线性、非适定或范数不匹配时不能无条件套用。

#### 答题方法

遇到“证明算法收敛”先问四件事：

1. 精确对象是什么：方程解、积分值、特征子空间还是最优点？
2. 局部残量采用什么归一化？
3. 扰动传播在哪个范数、哪个时间区间控制？
4. 常数是否独立于网格、步长和迭代次数？

只有把四项写清， $O(h^p)$  或“稳定”才是可核验命题。

#### 课程母题：条件数与稳定算法辨析

设  $F(x) = 1/(1-x)$ ，在  $x = 0.99$  附近计算相对条件数。再比较直接计算  $(1 - \cos x)/x^2$  与改写

$$\frac{2\sin^2(x/2)}{x^2}$$

在  $x \rightarrow 0$  时的浮点表现。

## 完整解答（独立核验稿） 第一问

$$\kappa_{\text{rel}}(x) = \frac{|F'(x)||x|}{|F(x)|} = \frac{|x|}{|1-x|} = 99,$$

说明输入相对误差可被放大约两位数量级。第二问的数学函数在零附近趋于  $1/2$ ，问题本身良态；直接式中的  $1 - \cos x$  会消去，而正弦平方改写避免相近数相减，因此差异来自算法表达而非问题条件性。

## 自测

1. 推导  $F(x) = \log x$  的绝对和相对条件数，并讨论  $x \rightarrow 1$ 。
2. 若残量很小但  $\kappa(A)$  很大，能否保证线性方程前向误差小？
3. 证明  $|\theta_n| \leq \gamma_n$  的乘积误差引理。
4. 分别给出 zero-stability 与 absolute stability 的一句严格定义。
5. 解释为什么 PDE 稳定性必须写明范数和边界处理。

## 本节结论

数值分析的基本链条是

$$\text{后向误差} \xrightarrow{\text{问题条件数}} \text{前向误差}, \quad \text{局部残量} \xrightarrow{\text{稳定传播}} \text{全局误差}.$$

条件数描述问题，稳定性描述算法；把两者混为一谈会使后续每个模块的证明失去逻辑支点。

## 1.5 非线性问题的绝对与相对条件数

设标量问题为  $y = f(x)$ 。小扰动下

$$\Delta y = f'(x)\Delta x + o(|\Delta x|).$$

绝对条件数为

$$\kappa_{\text{abs}}(x) = |f'(x)|,$$

在  $x \neq 0, f(x) \neq 0$  时，相对条件数为

$$\kappa_{\text{rel}}(x) = \left| \frac{x f'(x)}{f(x)} \right|.$$

它们是问题本身的局部敏感性，不取决于用何种算法计算  $f(x)$ 。

**例题 1.5.**  $f(x) = \sqrt{x}$  在  $x > 0$  上

$$\kappa_{\text{abs}}(x) = \frac{1}{2\sqrt{x}}, \quad \kappa_{\text{rel}}(x) = \frac{1}{2}.$$

靠近零时绝对敏感性变大, 但相对敏感性保持常数。若题目不说明误差尺度, 只说“条件数很大”可能没有意义。

**例题 1.6** (根对系数的敏感性). 令  $p(r; \alpha) = r^2 - 2r + \alpha$ , 根

$$r_{\pm} = 1 \pm \sqrt{1 - \alpha}.$$

当  $\alpha \uparrow 1$  时,

$$\left| \frac{\partial r_{\pm}}{\partial \alpha} \right| = \frac{1}{2\sqrt{1 - \alpha}} \rightarrow \infty.$$

重根附近根对系数扰动高度敏感; 即使求根算法后向稳定, 前向根误差仍可能很大。

## 1.6 后向稳定如何与条件数合成前向误差

若算法输出  $\hat{y} = f(x + \Delta x)$ , 且

$$\frac{\|\Delta x\|}{\|x\|} \leq cu + O(u^2),$$

则算法后向稳定。若问题在  $x$  附近可微, 前向相对误差近似满足

$$\frac{\|\hat{y} - y\|}{\|y\|} \lesssim \kappa_{\text{rel}}(x) cu.$$

因此“后向稳定”并不承诺前向误差接近机器精度; 它承诺算法像是在稍有扰动的数据上精确求解。前向误差还要乘问题条件数。

## 1.7 消去误差与等价公式重写

当  $x$  很小时,

$$\sqrt{1+x} - 1$$

由两个接近的数相减, 直接计算会丢失有效位。乘共轭得到

$$\sqrt{1+x} - 1 = \frac{x}{\sqrt{1+x} + 1},$$

右侧避免相消。类似地, 对大正  $x$ , 二次方程

$$x^2 + bx + c = 0$$

不能总用同一符号直接计算两个根; 先稳定求一个根, 再用根积  $r_1 r_2 = c$  得另一个常更可靠。

### 易错点

代数等价不等于浮点等价。分析稳定公式时, 逐个标出舍入发生在哪个运算, 再判断是否存在近数相减、上溢、下溢或误差放大。

## 1.8 三种稳定性的可执行对照

| 场景        | 一致性/局部误差   | 稳定性          | 收敛结论           |
|-----------|------------|--------------|----------------|
| 线性代数      | 通常不使用“一致性” | 后向误差小或扰动增长受控 | 前向误差由条件数合成     |
| ODE 多步法   | 局部截断误差趋零   | 根条件/零稳定      | 相容加零稳定推出全局收敛   |
| 线性 PDE 差分 | 离散算子逼近连续算子 | 网格范数下一致有界    | 适定线性问题中 Lax 等价 |

### 答题方法

任何误差题先写四行：精确输入与精确输出；扰动输入与算法输出；采用绝对还是相对尺度；结论属于条件性、稳定性还是收敛性。把名词归类后再估计。

### 自测

1. 计算  $f(x) = e^x$  的绝对与相对条件数。
2. 构造一个后向稳定但前向误差可能很大的病态问题。
3. 为什么  $\sqrt{1+x} - 1$  的有理化形式数值上更稳定？
4. 对比 ODE 零稳定与 PDE Fourier 稳定的对象。

## 第二章 插值、逼近与数值积分

### 学习目标

本章从“用有限信息恢复函数”出发，建立插值存在唯一性、余项、节点选择、Hermite 与样条、正交投影和最佳逼近，再把相同多项式结构用于 Newton-Cotes 与 Gauss 求积。读者应能完整证明插值余项、Chebyshev 极小性质和 Gauss 精度上限，并完成三道历年综合题。

### 2.1 Lagrange 与 Newton 插值

给定互异节点  $x_0, \dots, x_n$  和函数值  $f_i$ ，插值问题寻找  $p \in \mathbb{P}_n$  使  $p(x_i) = f_i$ 。

**定理 2.1** (插值存在唯一性). 互异  $n+1$  节点上的次数不超过  $n$  的插值多项式存在且唯一，并可写为

$$p_n(x) = \sum_{i=0}^n f_i \ell_i(x), \quad \ell_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

*证明.*  $\ell_i(x_j) = \delta_{ij}$ ，故上式存在。若  $p, q \in \mathbb{P}_n$  都插值，则  $p - q$  有  $n+1$  个互异零点，只能恒为零。□

Lagrange 形式适合证明，Newton 形式适合增量计算：

$$p_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \dots, x_n] \prod_{j=0}^{n-1} (x - x_j).$$

均差递推为

$$f[x_i, \dots, x_{i+k}] = \frac{f[x_{i+1}, \dots, x_{i+k}] - f[x_i, \dots, x_{i+k-1}]}{x_{i+k} - x_i}.$$

增加一个节点时只需增加最后一项；不要重新求全部 Lagrange 基函数。

**定理 2.2** (插值余项). 若  $f \in C^{n+1}[a, b]$ ，节点和  $x$  位于区间内，则存在依赖于  $x$  的  $\xi_x \in (a, b)$  使

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \omega_{n+1}(x), \quad \omega_{n+1}(x) = \prod_{i=0}^n (x - x_i).$$

*证明.* 若  $x$  是节点，结论平凡。否则选常数

$$c = \frac{f(x) - p_n(x)}{\omega_{n+1}(x)}$$

并令

$$g(t) = f(t) - p_n(t) - c\omega_{n+1}(t).$$

$g$  在  $n+2$  个互异点  $x_0, \dots, x_n, x$  上为零。连续应用 Rolle 定理  $n+1$  次, 存在  $\xi_x$  使  $g^{(n+1)}(\xi_x) = 0$ 。因为  $p_n^{(n+1)} = 0$ ,  $\omega_{n+1}^{(n+1)} = (n+1)!$ , 解出  $c$  即得。□

### 易错点

$\xi_x$  随  $x$  改变。点态余项可直接给  $L^\infty$  界, 但不能把  $f^{(n+1)}(\xi_x)$  当作固定常数提出积分, 从而无条件得到  $L^2$  界。 $L^2$  估计应使用 Peano 核、正交投影或 Sobolev 型不等式。

## 2.2 节点选择与 Runge 现象

等距高次插值可能在端点剧烈振荡。问题不在插值唯一性, 而在  $\|\omega_{n+1}\|_\infty$  和 Lebesgue 常数

$$\Lambda_n = \max_{x \in [a,b]} \sum_{i=0}^n |\ell_i(x)|.$$

若数据有扰动  $\tilde{f}_i = f_i + \delta_i$ , 则

$$|\tilde{p}_n(x) - p_n(x)| \leq \Lambda_n \max_i |\delta_i|.$$

Chebyshev 节点使  $\Lambda_n$  只按  $O(\log n)$  增长, 而等距节点可指数增长。

**定理 2.3** (Chebyshev 首一多项式的极小性). 在所有首项系数为 1 的  $n$  次多项式中,

$$\min_{\substack{p \text{ monic} \\ \deg p = n}} \|p\|_{\infty, [-1,1]} = 2^{1-n},$$

唯一最小者为  $2^{1-n}T_n(x)$ , 其中  $T_n(\cos \theta) = \cos(n\theta)$ 。

*证明.*  $T_n$  首项系数为  $2^{n-1}$ , 且在  $x_j = \cos(j\pi/n)$  处交替取  $(-1)^j$ , 所以上界成立。若另一个 monic  $p$  满足  $\|p\|_\infty < 2^{1-n}$ , 令

$$q = p - 2^{1-n}T_n.$$

$q$  次数至多  $n-1$ 。在相邻极值点  $x_j$  上, 严格范数界迫使  $q(x_j)$  符号交替, 所以每个相邻区间至少有一个零, 共  $n$  个零, 与  $\deg q \leq n-1$  矛盾。等号情形用相同交错论证给唯一性。□

## 2.3 Hermite 插值与三次样条

若在节点  $x_i$  同时给定函数值和导数, 相当于节点在节点多项式中重复。最简单的两点三次 Hermite 插值写成

$$H(x) = f_0 h_{00}(t) + h f'_0 h_{10}(t) + f_1 h_{01}(t) + h f'_1 h_{11}(t), \quad t = \frac{x - x_0}{h},$$

其中

$$h_{00} = 2t^3 - 3t^2 + 1, \quad h_{10} = t^3 - 2t^2 + t, \quad h_{01} = -2t^3 + 3t^2, \quad h_{11} = t^3 - t^2.$$

若  $f \in C^4$ , 余项为

$$f(x) - H(x) = \frac{f^{(4)}(\xi_x)}{4!}(x - x_0)^2(x - x_1)^2.$$

全局高次插值容易振荡; 样条改用分段低次多项式。三次样条在每个  $[x_i, x_{i+1}]$  上为三次, 整体  $C^2$ 。设  $M_i = s''(x_i)$ , 步长  $h_i = x_{i+1} - x_i$ , 连续一阶导数给三对角方程

$$h_{i-1}M_{i-1} + 2(h_{i-1} + h_i)M_i + h_iM_{i+1} = 6 \left( \frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_{i-1}} \right).$$

还需两个边界条件: 自然样条取  $M_0 = M_n = 0$ ; 夹持样条给端点一阶导数; not-a-knot 则让前两段和后两段在内节点处共享三次式。

**定理 2.4** (自然三次样条的变分性质). 在所有插值给定节点值且二阶弱导数平方可积的函数中, 自然三次样条最小化

$$J(g) = \int_a^b |g''(x)|^2 dx.$$

*证明.* 任取另一插值函数  $g = s + v$ , 则  $v(x_i) = 0$ 。分段积分两次,

$$\int s''v'' = \sum_i ([s''v']_{x_i}^{x_{i+1}} - [s'''v]_{x_i}^{x_{i+1}}).$$

$v$  在全部节点为零;  $s''$  在内节点连续使相邻  $s''v'$  边界项抵消, 自然边界又消去两端项。因此交叉项为零,

$$J(g) = J(s) + \int |v''|^2 \geq J(s).$$

□

## 2.4 正交投影与最佳平方逼近

在带权内积

$$\langle f, g \rangle_w = \int_a^b f(x)g(x)w(x) dx$$

下, 若  $\phi_0, \dots, \phi_n$  正交, 则最佳平方逼近

$$p_n = \arg \min_{p \in \text{span}\{\phi_0, \dots, \phi_n\}} \|f - p\|_w$$

的系数为

$$c_j = \frac{\langle f, \phi_j \rangle_w}{\|\phi_j\|_w^2}.$$

**定理 2.5** (投影定理与勾股式).  $p_n$  最优当且仅当

$$\langle f - p_n, q \rangle_w = 0$$

对所有  $q$  属于逼近空间成立。此时任意  $p$  满足

$$\|f - p\|_w^2 = \|f - p_n\|_w^2 + \|p - p_n\|_w^2.$$

证明. 把  $f - p = (f - p_n) + (p_n - p)$  展开平方. 交叉项为零, 故  $p_n$  唯一最优. 反过来, 若存在方向  $q$  与残量不正交, 沿  $p_n + tq$  的二次函数在  $t = 0$  导数非零, 不可能最优.  $\square$

Legendre 多项式对应  $w = 1$ , Chebyshev 多项式对应  $w = (1 - x^2)^{-1/2}$ . 首一正交多项式满足三项递推; 这是乘法算子  $p \mapsto xp$  在正交基中为对称三对角矩阵的结果.

## 2.5 数值积分: 精度、误差与复化

插值型求积写为

$$Q_n(f) = \sum_{i=0}^n w_i f(x_i), \quad w_i = \int_a^b \ell_i(x) dx.$$

若节点固定, 它至少对  $\mathbb{P}_n$  精确. 代数精度  $r$  指对所有  $\deg p \leq r$  精确, 而对某个  $r + 1$  次多项式不精确.

闭 Newton-Cotes 使用等距且含端点的节点. 梯形和 Simpson 规则分别为

$$T_h(f) = \frac{h}{2}(f(a) + f(b)), \quad S_h(f) = \frac{h}{6}[f(a) + 4f((a+b)/2) + f(b)].$$

局部误差分别为

$$I - T_h = -\frac{h^3}{12}f''(\xi), \quad I - S_h = -\frac{h^5}{2880}f^{(4)}(\xi).$$

复化后区间数约  $1/h$ , 所以全局阶分别为  $O(h^2)$  与  $O(h^4)$ .

**命题 2.6** (闭 Newton-Cotes 的奇偶精度). 在  $n + 1$  个等距闭节点上构造的插值型 Newton-Cotes 公式至少具有  $n$  次代数精度; 若  $n$  为偶数, 则由于关于区间中点的对称性, 精度至少提升到  $n + 1$ . 当  $n$  为奇数时一般不能提升. 高阶闭公式的权重可出现负数, 因此提高节点数不等于得到更稳定的求积法.

证明. 插值型构造对  $\mathbb{P}_n$  精确. 把区间平移缩放为  $[-1, 1]$ , 节点与权重关于零点成对对称, 所以求积误差泛函  $E(f) = \int_{-1}^1 f - Q_n(f)$  对所有奇函数为零. 当  $n$  为偶数时,  $x^{n+1}$  为奇函数, 故新增一阶精确性; 当  $n$  为奇数时  $x^{n+1}$  为偶函数, 对称性不给额外消去. 更高阶权重的符号由  $w_j = \int \ell_j$  决定, 插值基函数振荡会使某些积分为负; 负权会放大函数值误差和舍入扰动.  $\square$

### 2.5.1 Richardson 与 Romberg

若

$$T(h) = I + c_2 h^2 + c_4 h^4 + \cdots,$$

则

$$R_1(h) = \frac{4T(h/2) - T(h)}{3} = I + O(h^4).$$

递推

$$R_{k,j} = R_{k,j-1} + \frac{R_{k,j-1} - R_{k-1,j-1}}{4^j - 1}$$

形成 Romberg 表. 外推成立依赖误差具有偶次渐近展开; 有端点奇性或不光滑时不能只按公式升阶.

## 2.6 Gauss 求积：为什么能达到 $2n - 1$ 次

取正权  $w(x) > 0$ ，令  $\pi_n$  为首一  $n$  次正交多项式，其  $n$  个零点  $x_1, \dots, x_n$  都位于区间内部且互异。

**定理 2.7** (Gauss 求积的精度与正权). 以  $x_i$  为节点、插值权

$$\lambda_i = \int_a^b \ell_i(x) w(x) dx$$

构造的求积公式对所有  $\deg p \leq 2n - 1$  精确，且  $\lambda_i > 0$ 。任何  $n$  点公式的代数精度不可能超过  $2n - 1$ 。

**证明.** 对  $\deg p \leq 2n - 1$ ，作多项式除法

$$p = q\pi_n + r, \quad \deg q \leq n - 1, \quad \deg r \leq n - 1.$$

正交性给  $\int q\pi_n w = 0$ ；在节点上  $\pi_n(x_i) = 0$ ，所以公式对  $p$  的离散和等于对  $r$  的离散和，而插值公式对  $r$  精确。故精度至少  $2n - 1$ 。

取节点基函数  $\ell_i$ 。因公式对  $\ell_i^2$  精确，

$$\lambda_i = Q_n(\ell_i^2) = \int_a^b \ell_i(x)^2 w(x) dx > 0.$$

最后，若某个  $n$  点公式对  $2n$  次也精确，取

$$p(x) = \prod_{i=1}^n (x - x_i)^2.$$

离散和为零，而正权积分严格为正，矛盾。 □

若  $f \in C^{2n}$ ，Gauss 误差可写为

$$I(f) - Q_n(f) = \frac{f^{(2n)}(\xi)}{(2n)!} \int_a^b \pi_n(x)^2 w(x) dx.$$

常数取决于首一正交多项式的平方范数。

## 2.7 正式真题一：周期复化梯形

正式真题：2025 春 Q1（原卷第 1 页，15 分）

令  $f$  为  $2\pi$ -周期函数，

$$I(f) = \int_0^{2\pi} f(x) dx, \quad I_h^{\text{Tr}}(f) = h \sum_{i=1}^n f(ih), \quad h = 2\pi/n.$$

1. 若  $f \in C^m[0, 2\pi]$ , 证明存在与  $h$  无关的  $C > 0$  使

$$|I(f) - I_h^{\text{Tr}}(f)| \leq Ch^m;$$

2. 若  $f$  解析, 证明指数收敛。

完整解答 (独立核验稿) 写 Fourier 级数

$$f(x) = \sum_{k \in \mathbb{Z}} \hat{f}_k e^{ikx}.$$

离散正交恒等式

$$\frac{1}{n} \sum_{j=1}^n e^{2\pi i k j / n} = \begin{cases} 1, & n \mid k, \\ 0, & n \nmid k \end{cases}$$

给

$$I_h^{\text{Tr}} = 2\pi \sum_{\ell \in \mathbb{Z}} \hat{f}_{\ell n}, \quad I = 2\pi \hat{f}_0.$$

误差因此是 aliasing 尾

$$2\pi \sum_{\ell \neq 0} \hat{f}_{\ell n}.$$

若  $f$  及直到  $m-1$  阶导数满足周期端点匹配, 分部积分  $m$  次给  $|\hat{f}_k| \leq C|k|^{-m}$ , 从而

$$|I - I_h| \leq Cn^{-m} \sum_{\ell \neq 0} |\ell|^{-m} = O(h^m)$$

(这一绝对求和写法直接覆盖  $m > 1$ ;  $m = 1$  时改用单元 Peano 估计或复化矩形误差,  $C^1$  已保证有界变差, 仍得  $O(h)$ )。若  $f$  在宽度  $a$  的复条带内解析且有界, 把积分路径上移/下移得到  $|\hat{f}_k| \leq Ce^{-a|k|}$ , 故

$$|I - I_h| = O(e^{-an}) = O(e^{-2\pi a/h}).$$

## 2.8 正式真题二: Chebyshev 极小与 Gauss–Chebyshev

正式真题: 2025 秋 Q2 (原卷第 1 页, 20 分)

设  $T_0 = 1, T_1 = x, T_{n+1} = 2xT_n - T_{n-1}$ 。

1. 证明  $T_n(x) = \cos(n \arccos x)$  ( $|x| \leq 1$ ) 以及

$$T_n(x) = \cosh(n \operatorname{arccosh} x) \quad (|x| > 1)$$

(负半轴按奇偶性解释);

2. 证明原卷打印的  $\min_{\deg p \leq n, p(1)=1} \|p\|_\infty = 1$ ;

3. 若  $L_n$  在  $T_{n+1}$  的  $n+1$  个零点插值  $f \in C^{n+1}[-1, 1]$ , 证明

$$\|f - L_n\|_{\infty} \leq \frac{\|f^{(n+1)}\|_{\infty}}{(n+1)!2^n};$$

4. 对  $f \in C^{2n+2}[-1, 1]$ , 推导

$$I(f) = \int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx$$

的  $(n+1)$  点 Gauss-Chebyshev 公式, 并证明

$$|I(f) - I_{n+1}(f)| \leq \frac{\pi \|f^{(2n+2)}\|_{\infty}}{(2n+2)!2^{2n+1}}.$$

**完整解答 (独立核验稿)** 余弦倍角递推与题给递推相同, 初值也相同, 故

$$T_n(\cos \theta) = \cos n\theta.$$

令  $x = \cosh t$  得  $|x| > 1$  的双曲表示 (负半轴结合奇偶性)。

第二问按原题约束其实是平凡极小:  $p(1) = 1$  强迫  $\|p\|_{\infty} \geq 1$ , 而常数多项式  $p \equiv 1$  已达到 1;  $T_n$  也达到 1, 但不是唯一最小者。这与 定理 2.3 的“固定首项系数”经典定理不同, 不能静默替换。

$T_{n+1}$  首项系数为  $2^n$ , 故节点多项式为  $\omega_{n+1} = 2^{-n}T_{n+1}$ ,  $\|\omega_{n+1}\|_{\infty} = 2^{-n}$ 。由 定理 2.2,

$$\|f - L_n\|_{\infty} \leq \frac{\|f^{(n+1)}\|_{\infty}}{(n+1)!2^n}.$$

令  $x = \cos \theta$ , 则

$$\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx = \int_0^{\pi} f(\cos \theta) d\theta.$$

节点与权为

$$x_j = \cos \frac{(2j-1)\pi}{2(n+1)}, \quad \lambda_j = \frac{\pi}{n+1}.$$

Gauss 误差公式和首一  $\pi_{n+1} = 2^{-n}T_{n+1}$  的平方范数给

$$|E(f)| \leq \frac{\pi \|f^{(2n+2)}\|_{\infty}}{(2n+2)!2^{2n+1}}.$$

## 2.9 正式真题三：首一 Legendre、插值与 Gauss

正式真题：2026 春 Q2（原卷第 1 页，20 分）

在  $[-1, 1]$  上取

$$\langle f, g \rangle = \int_{-1}^1 f(x)g(x) dx, \quad \|f\|_2 = \sqrt{\langle f, f \rangle}.$$

令

$$a_n = \frac{(2n)!}{2^n(n!)^2}, \quad P_0 = 1, \quad P_1 = x, \quad P_{n+1} = xP_n - \frac{n^2}{4n^2 - 1}P_{n-1}.$$

1. 计算  $\langle P_n, P_m \rangle$ ;
2. 证明  $P_n$  解下列问题并计算最小值：

$$\min\{\|p\|_2 : \deg p \leq n, p^{(n)}(1) = n!\};$$

3. 若  $L_n$  在  $P_{n+1}$  的零点插值  $f \in C^{n+1}[-1, 1]$ ，证明

$$\|f - L_n\|_2 \leq \frac{2 \|f^{(n+1)}\|_2}{(2n+3)a_{n+1}(n+1)!};$$

4. 推导  $(n+1)$  点 Gauss-Legendre 公式并估计其误差。

**完整解答（独立核验稿）** 若标准 Legendre 多项式记为  $L_n$ ，其首项系数为  $a_n$ ，故  $P_n = L_n/a_n$ 。于是

$$\langle P_n, P_m \rangle = 0 \quad (n \neq m), \quad \|P_n\|_2^2 = \frac{2}{(2n+1)a_n^2}.$$

约束  $p^{(n)}(1) = n!$  等价于  $p$  为首一  $n$  次。写  $p = P_n + q$ 、 $\deg q < n$ ，正交勾股给

$$\|p\|_2^2 = \|P_n\|_2^2 + \|q\|_2^2,$$

所以唯一最小者为  $P_n$ ，最小范数为  $\sqrt{2/(2n+1)}/a_n$ 。

第三问的打印常数事实上不能成立，不能用“Peano 核计算”跳过核范数。取

$$f = P_{n+1}, \quad L_n \equiv 0.$$

由第一问及  $P_{n+1}$  首一，

$$\|f - L_n\|_2 = \frac{\sqrt{2}}{a_{n+1}\sqrt{2n+3}}, \quad \|f^{(n+1)}\|_2 = (n+1)!\sqrt{2}.$$

题面右端因此等于

$$\frac{2\sqrt{2}}{(2n+3)a_{n+1}},$$

而左端与右端之比为  $\sqrt{2n+3}/2 > 1$  ( $n \geq 1$ )。例如  $n = 1$  时  $P_2 = x^2 - \frac{1}{3}$ ，左端平方为  $8/45$ ，题面右端平方为  $32/225$ 。因此该小问按打印形式应回答“命题不成立”，并给出上述反例；任何正

确常数至少不能小于

$$\frac{1}{a_{n+1}(n+1)!\sqrt{2n+3}}.$$

这是一项进入最终审计记录的原题/冻结解答冲突，而不是用候选答案覆盖正式答案。

Gauss-Legendre 节点是  $P_{n+1}$  的零点，权可写为

$$w_i = \int_{-1}^1 \ell_i(x) dx = \frac{2}{(1-x_i^2)[L'_{n+1}(x_i)]^2} > 0.$$

由 定理 2.7，精度为  $2n+1$ 。若  $f \in C^{2n+2}$ ,

$$E(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_{-1}^1 P_{n+1}(x)^2 dx = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \frac{2}{(2n+3)a_{n+1}^2}.$$

### 答题方法

插值/求积综合题按同一条链处理：

节点多项式  $\rightarrow$  正交或极小性质  $\rightarrow$  余项/Peano 核  $\rightarrow$  范数常数。

遇到原题条件与经典定理不一致时，先按原题求解，再明确指出经典版本的条件；不能把更熟悉的定理悄悄替换进题面。

### 自测

1. 推导 Newton 均差形式并证明最高阶均差关于节点对称。
2. 为什么等距插值的 Runge 现象不能靠提高次数自动消失？
3. 写出自然样条与夹持样条的两个边界条件。
4. 证明任意  $n$  点求积精度不超过  $2n-1$ 。
5. 推导两点 Gauss-Legendre 节点和权，并验证三次精度。
6. 解释周期梯形指数收敛为何依赖复平面解析条带。

### 本节结论

插值、最佳逼近和求积共享同一个多项式骨架：插值由零点多项式控制，平方逼近由正交投影控制，Gauss 求积把正交多项式零点变成最优节点。真正决定误差的不是公式名称，而是函数正则性、节点几何和所用范数。

## 2.10 插值余项的完整来源

设  $x_0, \dots, x_n$  互异， $p_n$  插值  $f$ 。对固定  $x \notin \{x_i\}$ ，令

$$\omega_{n+1}(t) = \prod_{i=0}^n (t - x_i), \quad K = \frac{f(x) - p_n(x)}{\omega_{n+1}(x)}.$$

构造

$$g(t) = f(t) - p_n(t) - K\omega_{n+1}(t).$$

$g$  在  $x_0, \dots, x_n, x$  共  $n+2$  个点为零。连续使用 Rolle 定理, 存在  $\xi$  位于这些点的凸包内, 使

$$g^{(n+1)}(\xi) = 0.$$

由于  $p_n^{(n+1)} = 0$ ,  $\omega_{n+1}^{(n+1)} = (n+1)!$ , 得到

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_{n+1}(x).$$

这里  $\xi$  依赖  $x$ , 不能在积分或求导时当作固定常数。

## 2.11 Lebesgue 常数与节点选择

Lagrange 插值算子

$$I_n f(x) = \sum_{i=0}^n f(x_i) \ell_i(x)$$

的算子范数为

$$\Lambda_n = \max_{x \in [a, b]} \sum_{i=0}^n |\ell_i(x)|.$$

若数据扰动  $|\delta_i| \leq \varepsilon$ , 则

$$\|I_n(f + \delta) - I_n f\|_\infty \leq \Lambda_n \varepsilon.$$

并且对任意  $p \in \mathbb{P}_n$ ,

$$f - I_n f = (f - p) - I_n(f - p),$$

故

$$\|f - I_n f\|_\infty \leq (1 + \Lambda_n) \inf_{p \in \mathbb{P}_n} \|f - p\|_\infty.$$

因此误差不仅取决于  $f^{(n+1)}$ , 还取决于插值算子的稳定性。等距节点的  $\Lambda_n$  快速增长, Chebyshev 型节点则只对数增长, 解释了 Runge 现象与节点聚集的必要性。

## 2.12 Hermite 插值与重节点条件

给定节点  $x_i$  上函数值和若干阶导数, 共  $N+1$  个线性条件, 存在唯一次数不超过  $N$  的 Hermite 多项式。若每个节点同时给  $f(x_i), f'(x_i)$ , 则误差含重节点积

$$f(x) - H(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \prod_{i=0}^n (x - x_i)^2.$$

唯一性证明: 两个解之差在每个  $x_i$  至少有二重零点, 故若次数不超过  $2n+1$ , 只能恒为零。

## 2.13 三次样条的方程从哪里来

在节点  $x_0 < \cdots < x_n$  上, 三次样条要求每段为三次多项式, 整体  $C^2$ , 并插值数据。记

$$h_i = x_{i+1} - x_i, \quad M_i = S''(x_i).$$

在区间  $[x_i, x_{i+1}]$ ,  $S''$  线性插值  $M_i, M_{i+1}$ 。积分两次并用端点函数值, 匹配一阶导数得到内点方程

$$h_{i-1}M_{i-1} + 2(h_{i-1} + h_i)M_i + h_iM_{i+1} = 6 \left( \frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}} \right).$$

还需两条边界条件: 自然样条取  $M_0 = M_n = 0$ ; 夹持样条指定两端一阶导; not-a-knot 要求头两段与尾两段的三阶导连续。不同边界条件定义不同样条, 不可只写“解三对角方程”。

## 2.14 Gauss 求积的节点、权重与正性

设权函数  $w(x) > 0$ ,  $\pi_n$  是首一  $n$  次正交多项式, 其  $n$  个零点为  $x_i$ 。对任意  $p \in \mathbb{P}_{2n-1}$ , 作除法

$$p = q\pi_n + r, \quad q, r \in \mathbb{P}_{n-1}.$$

正交性给

$$\int p w = \int r w.$$

又因  $\pi_n(x_i) = 0$ ,  $p(x_i) = r(x_i)$ 。选择对  $\mathbb{P}_{n-1}$  精确的插值型权重, 便对  $\mathbb{P}_{2n-1}$  精确。

权重正性可取 Lagrange 基函数  $\ell_i$ , 利用  $\ell_i^2 \in \mathbb{P}_{2n-2}$ :

$$w_i = \sum_j w_j \ell_i(x_j)^2 = \int \ell_i(x)^2 w(x) dx > 0.$$

精度不可能超过  $2n - 1$ , 因为

$$q(x) = \prod_{i=1}^n (x - x_i)^2$$

是  $2n$  次非负非零多项式, 节点求积值为零, 而其积分严格为正。

**定理 2.8** (Gauss 求积的 Hermite 余项). 设  $w \geq 0$  在有限区间  $[a, b]$  上可积且不恒为零, 首一正交多项式  $\pi_n$  的  $n$  个零点  $x_1, \dots, x_n$  位于区间内部。若  $f \in C^{2n}[a, b]$ , 则相应  $n$  点 Gauss 求积的误差满足

$$E_n(f) := \int_a^b f(x)w(x) dx - \sum_{j=1}^n w_j f(x_j) = \frac{f^{(2n)}(\xi)}{(2n)!} \int_a^b \pi_n(x)^2 w(x) dx$$

对某个  $\xi \in (a, b)$  成立。因而最高精确次数为  $2n - 1$ , 而第一项可能出现的导数恰是  $2n$  阶导数。

完整证明. 取次数不超过  $2n - 1$  的 Hermite 插值多项式  $H$ , 使

$$H(x_j) = f(x_j), \quad H'(x_j) = f'(x_j), \quad 1 \leq j \leq n.$$

重复节点插值余项给出: 对每个  $x \notin \{x_1, \dots, x_n\}$ , 存在位于  $[a, b]$  内的  $\xi_x$ , 使

$$f(x) - H(x) = \frac{f^{(2n)}(\xi_x)}{(2n)!} \prod_{j=1}^n (x - x_j)^2 = \frac{f^{(2n)}(\xi_x)}{(2n)!} \pi_n(x)^2.$$

Gauss 公式对  $H$  精确, 且节点处  $H = f$ , 所以

$$E_n(f) = \int_a^b [f(x) - H(x)] w(x) dx.$$

令  $m$  与  $M$  分别为  $f^{(2n)}$  在紧区间上的最小值和最大值。因为  $w\pi_n^2 \geq 0$ , 上式夹在

$$\frac{m}{(2n)!} \int_a^b \pi_n^2 w \quad \text{与} \quad \frac{M}{(2n)!} \int_a^b \pi_n^2 w$$

之间。积分因子严格为正; 由连续函数的介值性质, 存在  $\xi \in [a, b]$  使误差具有定理中的形式。若  $f^{(2n)}$  不在端点取得所需值, 可取  $\xi \in (a, b)$ ; 端点情形由极限解释。最后, 前述非负多项式  $\pi_n^2$  证明  $2n$  次精确不可能, 因此次数  $2n - 1$  确为最高。□

**例题 2.1** (二点 Gauss-Legendre 余项核对). 在  $[-1, 1]$  上取  $w \equiv 1$ 。二点公式的节点为  $\pm 1/\sqrt{3}$ , 权重均为 1。对  $f(x) = x^4$ ,

$$\int_{-1}^1 x^4 dx = \frac{2}{5}, \quad Q_2(f) = 2 \left( \frac{1}{\sqrt{3}} \right)^4 = \frac{2}{9},$$

故误差为  $8/45$ 。首一正交多项式  $\pi_2(x) = x^2 - \frac{1}{3}$ , 且

$$\int_{-1}^1 \pi_2(x)^2 dx = \frac{8}{45}, \quad \frac{f^{(4)}}{4!} = 1.$$

这与定理给出的误差常数完全一致, 也直接显示“精确到三次”与“四阶导数余项”是同一结构的两面。

## 2.15 复化公式与 Euler-Maclaurin 视角

复化梯形在一般光滑函数上全局二阶; 但周期函数若端点各阶导数匹配, Euler-Maclaurin 展开中的边界差项连续消失, 可获得高阶乃至谱式收敛。这个结论的条件不是“函数周期”四个字, 而是积分区间跨越完整周期且所需阶导端点匹配。

Richardson 外推若

$$T(h) = I + c_p h^p + c_{p+1} h^{p+1} + \dots,$$

则

$$R(h) = \frac{2^p T(h/2) - T(h)}{2^p - 1}$$

消去  $h^p$  主误差。若真实展开含对数项、奇次项或网格不成固定比例, 照搬公式可能失效。

**例题 2.2.** Simpson 公式可由梯形  $T(h)$  与中点  $M(h)$  组合得到

$$S(h) = \frac{1}{3}T(h) + \frac{2}{3}M(h).$$

对三次多项式逐项核验精确性；对一般  $C^4$  函数，误差常数来自四阶导，而不是因为“使用三个点所以三阶”。

## 2.16 等距三角插值、DFT 与 FFT

周期函数在等距节点上的插值天然落到离散 Fourier 变换。令

$$x_j = \frac{2\pi j}{N}, \quad f_j = f(x_j), \quad 0 \leq j < N,$$

并采用归一化约定

$$\hat{f}_k = \frac{1}{N} \sum_{j=0}^{N-1} f_j e^{-2\pi i j k / N}, \quad f_j = \sum_{k=0}^{N-1} \hat{f}_k e^{2\pi i j k / N}.$$

第二式由离散正交关系

$$\sum_{j=0}^{N-1} e^{2\pi i (k-\ell) j / N} = N \mathbf{1}_{\{k \equiv \ell \pmod{N}\}}$$

得到；它同时证明 DFT 矩阵除归一化外是酉矩阵，因此变换唯一且二范数稳定。当  $N = 2m + 1$  时，把频率指标改写为  $-m, \dots, m$ ，即得唯一的三角多项式插值

$$p_m(x) = \sum_{k=-m}^m \hat{f}_k e^{ikx}, \quad p_m(x_j) = f_j.$$

偶数  $N$  时 Nyquist 频率需要固定一个共轭对称约定；实数据满足  $\hat{f}_{N-k} = \overline{\hat{f}_k}$ 。

**定理 2.9** (Cooley–Tukey 偶奇分裂). 若  $N = 2M$ ，记  $\omega_N = e^{-2\pi i / N}$ ，则

$$\sum_{j=0}^{N-1} f_j \omega_N^{jk} = \sum_{r=0}^{M-1} f_{2r} \omega_M^{rk} \omega_N^k \sum_{s=0}^{M-1} f_{2r+1} \omega_M^{rk}.$$

因此长度  $N$  的 DFT 可递归化为两个长度  $M$  的 DFT，再做  $O(N)$  次蝶形合并；当  $N = 2^q$  时总成本为  $O(N \log N)$ ，而直接矩阵乘法为  $O(N^2)$ 。

**例题 2.3** (四点余弦数据). 取  $f_j = \cos(2\pi j/4) = (1, 0, -1, 0)$ 。按上述归一化，

$$\hat{f}_1 = \hat{f}_3 = \frac{1}{2}, \quad \hat{f}_0 = \hat{f}_2 = 0.$$

逆变换给  $p(x) = \frac{1}{2}e^{ix} + \frac{1}{2}e^{-ix} = \cos x$ 。若把正、逆变换的  $1/N$  位置同时照搬另一约定，会整体系数错  $N$  倍。

## 易错点

采样只能区分模  $N$  同余的频率：在节点上  $e^{i(k+N)x_j} = e^{ikx_j}$ ，这就是 aliasing。FFT 只降低计算复杂度，不会自动消除混叠，也不会把不光滑函数变成谱收敛。

## 2.17 有理逼近与 Padé 构造

有理函数

$$r_{m,n}(x) = \frac{p_m(x)}{q_n(x)}, \quad \deg p_m \leq m, \quad \deg q_n \leq n,$$

只有在目标区间上  $q_n$  不为零时才给出稳定逼近。与高次多项式相比，有理函数可以用分母表达极点或快速变化，但代价是可能产生伪极点。

若  $f(x) = \sum_{k \geq 0} a_k x^k$ ，Padé  $[m/n]$  逼近通过归一化  $q_n(0) = 1$  和

$$q_n(x)f(x) - p_m(x) = O(x^{m+n+1})$$

确定。比较  $x^{m+1}, \dots, x^{m+n}$  的系数先得到分母系数的  $n \times n$  线性系统，再由低阶系数恢复  $p_m$ 。若该系统奇异，指定次数的 Padé 逼近可能不唯一或实际降阶。

**例题 2.4** ( $e^x$  的一阶对一阶 Padé 逼近). 设  $p = 1 + ax$ 、 $q = 1 + bx$ 。要求

$$(1 + bx)e^x - (1 + ax) = O(x^3)$$

给出  $a - b = 1$ 、 $b + \frac{1}{2} = 0$ ，故

$$r_{1,1}(x) = \frac{1 + x/2}{1 - x/2} = 1 + x + \frac{x^2}{2} + O(x^3).$$

它比一次 Taylor 多匹配一个系数，但在  $x = 2$  有极点；讨论误差前必须先限定远离分母零点的区间。

插值型有理式也常写为重心形式

$$r(x) = \frac{\sum_{j=0}^n \frac{w_j f_j}{x - x_j}}{\sum_{j=0}^n \frac{w_j}{x - x_j}}.$$

在  $x = x_j$  处取连续延拓得到  $r(x_j) = f_j$ 。该形式计算稳定且避免显式展开高次多项式，但权重选择若使分母在区间内接近零，仍会造成巨大条件数。误差估计必须同时控制分子逼近误差和分母下界；只报告级数匹配阶并不能保证整个区间上的一致误差。

## 易错点

原打印题 2026 春 Q2(c) 的常数冲突已在唯一真题解答处冻结处理：题面不改， $f = P_{n+1}$  给出一般反例，左右端比为  $\sqrt{2n+3}/2 > 1$ ，并保留  $n = 1$  数值核验。它属于原题命题不

成立，不属于讲义解答缺失。

### 自测

1. 用辅助函数与 Rolle 定理证明 Lagrange 余项。
2. Lebesgue 常数同时控制哪两类误差？
3. 推导自然三次样条的三对角内点方程。
4. 证明 Gauss 权重严格为正，并说明最高精度为何不能超过  $2n - 1$ 。
5. 从离散正交关系推导 DFT 逆变换，并写出一次 FFT 偶奇分裂。
6. 构造  $e^x$  的  $[1/1]$  Padé 逼近并指出其极点。

## 第三章 非线性方程

### 学习目标

本章建立“良定义—不变邻域—误差递推—收敛阶”的完整证明链。读者应能比较二分、不动点、Newton、割线和无导数 Newton 型方法；处理重根；为方程组写出线性化与停止准则；并识别局部高速方法在初值、导数和条件性上的失败模式。

### 3.1 二分法：慢但可证的全局基准

若  $f \in C[a, b]$  且  $f(a)f(b) < 0$ ，二分法每次取中点并保留符号变化子区间。第  $k$  次后区间长度为

$$\frac{b-a}{2^k},$$

若以中点为近似根，则

$$|x_k - x^*| \leq \frac{b-a}{2^{k+1}}.$$

为使误差不超过  $\varepsilon$ ，足够取

$$k \geq \left\lceil \log_2 \frac{b-a}{2\varepsilon} \right\rceil.$$

二分只保证找到某个根，不保证区间内根唯一；若端点同号也不能推出无根。

### 3.2 不动点迭代与压缩映射

把  $f(x) = 0$  改写成

$$x = g(x), \quad x_{k+1} = g(x_k).$$

同一个方程有许多  $g$ ，收敛性取决于改写方式。

**定理 3.1** (压缩映射迭代). 设闭区间  $I$  满足  $g(I) \subseteq I$ ，且

$$|g(x) - g(y)| \leq q|x - y|, \quad 0 \leq q < 1.$$

则  $g$  在  $I$  有唯一不动点  $x^*$ ，任意  $x_0 \in I$  的迭代收敛，并有

$$|x_k - x^*| \leq q^k |x_0 - x^*| \leq \frac{q^k}{1-q} |x_1 - x_0|.$$

*证明.* 迭代差满足  $|x_{k+1} - x_k| \leq q^k |x_1 - x_0|$ ，故为 Cauchy 列；闭性给极限仍在  $I$ ，连续性给  $x^* = g(x^*)$ 。两个不动点间距满足  $|x^* - y^*| \leq q|x^* - y^*|$ ，所以唯一。误差估计由几何级数尾和得

到。 □

局部判据  $|g'(x^*)| < 1$  只给充分近初值的局部收敛；要证明全局区间收敛，还需  $g(I) \subseteq I$  和整个区间上的导数界。

### 3.3 收敛阶与误差常数

**定义 3.1** (收敛阶). 若  $e_k = x_k - x^* \rightarrow 0$ , 且存在  $p \geq 1$ ,  $C \in (0, \infty)$  使

$$\lim_{k \rightarrow \infty} \frac{|e_{k+1}|}{|e_k|^p} = C,$$

则称迭代以  $p$  阶收敛。 $p = 1$  时还要求  $C < 1$  才是线性收敛。

若  $x_{k+1} = g(x_k)$ ,  $g \in C^p$ , 且

$$g(x^*) = x^*, \quad g'(x^*) = \cdots = g^{(p-1)}(x^*) = 0, \quad g^{(p)}(x^*) \neq 0,$$

Taylor 展开直接给

$$e_{k+1} = \frac{g^{(p)}(x^*)}{p!} e_k^p + o(e_k^p).$$

但在使用该判阶前必须先证明迭代在邻域内良定义并确实收敛到  $x^*$ 。

### 3.4 Newton 法：局部二次收敛的完整链

Newton 迭代为

$$x_{k+1} = N(x_k) = x_k - \frac{f(x_k)}{f'(x_k)}.$$

**定理 3.2** (单根 Newton 局部二次收敛). 设  $f \in C^2$  于  $x^*$  的邻域,  $f(x^*) = 0$  且  $f'(x^*) \neq 0$ . 则存在  $r > 0$ , 使任意  $|x_0 - x^*| \leq r$  的 Newton 迭代良定义、留在该邻域并收敛到  $x^*$ . 并且

$$\lim_{k \rightarrow \infty} \frac{x_{k+1} - x^*}{(x_k - x^*)^2} = \frac{f''(x^*)}{2f'(x^*)}.$$

*证明.* 由  $f'$  连续且  $f'(x^*) \neq 0$ , 可取闭邻域  $I$  使

$$|f'(x)| \geq \mu > 0, \quad |f''(x)| \leq M.$$

对  $x_k \in I$ , Taylor 公式在  $x_k$  展开  $f(x^*)$ :

$$0 = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2}f''(\xi_k)(x^* - x_k)^2.$$

整理得

$$e_{k+1} = \frac{f''(\xi_k)}{2f'(x_k)} e_k^2, \quad |e_{k+1}| \leq \frac{M}{2\mu} |e_k|^2.$$

再把  $r$  取得满足  $Mr/(2\mu) \leq 1/2$ , 便有  $|e_{k+1}| \leq |e_k|/2$ , 所以迭代留在  $I$  并收敛。最后  $\xi_k, x_k \rightarrow x^*$ , 对误差式取极限。 □

### 3.4.1 重根与修正 Newton

若

$$f(x) = (x - \alpha)^m g(x), \quad g(\alpha) \neq 0, \quad m > 1,$$

则

$$\frac{f(x)}{f'(x)} = \frac{e}{m + e g'(x)/g(x)} = \frac{e}{m} + O(e^2).$$

标准 Newton 因而满足

$$e_{k+1} = \left(1 - \frac{1}{m}\right) e_k + O(e_k^2),$$

降为线性。若  $m$  已知, 使用

$$x_{k+1} = x_k - m \frac{f(x_k)}{f'(x_k)}$$

可消去一次项, 恢复二次收敛。

若  $m$  未知, 可对  $u = f/f'$  使用 Newton:

$$x_{k+1} = x_k - \frac{f(x_k)f'(x_k)}{[f'(x_k)]^2 - f(x_k)f''(x_k)}.$$

$u$  在  $\alpha$  处是单根, 因此在常规条件下局部二次。

## 3.5 割线法与黄金阶

割线法用两点斜率替代导数:

$$x_{k+1} = x_k - f(x_k) \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}.$$

它每步只需一次新函数值, 但误差是二步递推。

**定理 3.3** (割线法局部误差). 在 [定理 3.2](#) 的单根条件下, 充分近且不同的初值使割线法良定义并局部收敛, 且

$$\lim_{k \rightarrow \infty} \frac{e_{k+1}}{e_k e_{k-1}} = \frac{f''(x^*)}{2f'(x^*)}.$$

若渐近阶存在, 则

$$p = \frac{1 + \sqrt{5}}{2}.$$

**证明.** 割线更新是通过  $(x_{k-1}, f(x_{k-1}))$ 、 $(x_k, f(x_k))$  的一次插值多项式之零点。插值余项在  $x^*$  给

$$0 - f[x_{k-1}, x_k](x^* - x_{k+1}) = \frac{f''(\xi_k)}{2}(x^* - x_{k-1})(x^* - x_k),$$

从而

$$e_{k+1} = \frac{f''(\xi_k)}{2f[x_{k-1}, x_k]} e_k e_{k-1}.$$

在足够小邻域内, 均差远离零, 误差乘积使邻域不变并给收敛; 取极限即得常数。若  $|e_k| \asymp |e_{k-1}|^p$  且  $|e_{k+1}| \asymp |e_k|^p$ , 代入乘积递推得到  $p^2 = p + 1$ 。□

### 3.6 无导数 Newton 型方法

Steffensen 法把导数替换为前向差商:

$$x_{k+1} = x_k - \frac{f(x_k)^2}{f(x_k + f(x_k)) - f(x_k)}.$$

它只需函数值, 却在单根处通常保持二次收敛。令  $a = f'(x^*) \neq 0$ 、 $b = f''(x^*)/2$ 、 $e = x - x^*$ 。则

$$f(x) = ae + be^2 + o(e^2).$$

记  $h = f(x) = O(e)$ 。由  $f \in C^2$  的 Peano 型 Taylor 展开,

$$f(x+h) - f(x) = f'(x)h + \frac{1}{2}f''(x)h^2 + o(h^2).$$

又有  $f'(x) = a + 2be + o(e)$ 、 $f''(x) = 2b + o(1)$ , 所以分母

$$D(e) := f(x + f(x)) - f(x) = a^2e + ab(3+a)e^2 + o(e^2).$$

当  $e \neq 0$  足够小时,  $D(e)/e \rightarrow a^2 \neq 0$ , 故迭代分母非零。同时

$$f(x)^2 = a^2e^2 + 2abe^3 + o(e^3).$$

对分子、分母同时除以  $a^2e$ , 再用  $(1+u)^{-1} = 1 - u + o(u)$ , 得到

$$e_+ = \frac{b(1+a)}{a}e^2 + o(e^2) = \frac{f''(x^*)[1+f'(x^*)]}{2f'(x^*)}e^2 + o(e^2).$$

因此存在  $C, \delta > 0$ , 使  $0 < |e| \leq \delta$  时迭代良定义且  $|e_+| \leq C|e|^2$ 。再缩小  $\delta$  使  $C\delta \leq 1/2$ , 便有  $|e_+| \leq |e|/2$ : 该邻域不变, 且迭代收敛到  $x^*$ 。

若主项系数非零, 方法为二次收敛, 误差常数为

$$\left| \frac{f''(x^*)[1+f'(x^*)]}{2f'(x^*)} \right|.$$

若该系数为零, 在  $C^2$  假设下只能断言  $e_+ = o(e^2)$ ; 除非再增加更高阶光滑性和相应非退化条件, 不能据此声称存在确定的更高代数收敛阶。

### 3.7 方程组 Newton 与线性化误差

对  $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ , Newton 步解

$$J_F(x_k)s_k = -F(x_k), \quad x_{k+1} = x_k + s_k.$$

若  $J_F(x^*)$  非奇异且 Jacobian 局部 Lipschitz, 则与标量情形相同,

$$\|e_{k+1}\| \leq C \|e_k\|^2.$$

证明使用积分余项

$$F(x_k) - F(x^*) = \int_0^1 J_F(x^* + te_k) e_k dt$$

并控制  $J_F(x_k)^{-1}$ 。实际计算中不显式求逆，而是解线性方程。

若每步线性系统只近似求解，

$$\|J_F(x_k)s_k + F(x_k)\| \leq \eta_k \|F(x_k)\|,$$

得到 inexact Newton。若  $\eta_k \leq \bar{\eta} < 1$ ，通常保持局部线性； $\eta_k \rightarrow 0$  可得超线性； $\eta_k = O(\|F(x_k)\|)$  可恢复二次。

### 3.8 全局化、停止准则与条件性

Newton 是局部方法。阻尼 Newton 取

$$x_{k+1} = x_k + \alpha_k s_k, \quad 0 < \alpha_k \leq 1,$$

用 Armijo 条件选择  $\alpha_k$ ；信赖域则限制  $\|s\| \leq \Delta_k$ 。全局化的目标是从远处保证某个 merit function 下降，进入局部区后再接受全 Newton 步。

常见停止准则应组合使用：

$$\|F(x_k)\| \leq \tau_F, \quad \|s_k\| \leq \tau_x(1 + \|x_k\|).$$

小步长可能由线搜索过度缩短或 Jacobian 病态造成，并不自动说明残量小；小残量也需结合根的条件数判断前向误差。

### 3.9 正式真题一：Newton 与割线法

正式真题：2023 秋 Q1（原卷第 1 页）

设  $f \in C^2(a, b)$ ， $x^* \in (a, b)$  是单根。

1. 证明充分近初值下 Newton 法收敛，并求

$$\lim \frac{x_{k+1} - x^*}{(x_k - x^*)^2}.$$

2. 用两点割线斜率替换导数，证明局部收敛并求

$$\lim \frac{x_{k+1} - x^*}{(x_k - x^*)(x_{k-1} - x^*)}.$$

完整解答（独立核验稿） 第一问由 定理 3.2，极限为

$$\frac{f''(x^*)}{2f'(x^*)}.$$

第二问由 定理 3.3 的插值余项,

$$e_{k+1} = \frac{f''(\xi_k)}{2f[x_{k-1}, x_k]} e_k e_{k-1}.$$

在足够小不变邻域中分母远离零, 故迭代良定义并收敛; 取极限仍得

$$\frac{f''(x^*)}{2f'(x^*)}.$$

由  $e_{k+1} \sim C e_k e_{k-1}$  可推出黄金阶, 而不是二次。

### 3.10 正式真题二: 重根 Newton

正式真题: 2025 春 Q2 (原卷第 1 页, 15 分)

证明单根 Newton 局部二次收敛; 求  $m > 1$  重根处标准 Newton 的收敛率; 证明已知  $m$  时修正 Newton 恢复二次; 未知  $m$  时提出一种超线性方法。

完整解答 (独立核验稿) 单根部分见 定理 3.2。写

$$f(x) = (x - \alpha)^m g(x), \quad g(\alpha) \neq 0.$$

则

$$\frac{f}{f'} = \frac{e}{m + eg'/g} = \frac{e}{m} + O(e^2),$$

故标准 Newton

$$e_{k+1} = \left(1 - \frac{1}{m}\right) e_k + O(e_k^2)$$

仅线性收敛。乘  $m$  的修正 Newton 消去一次项, 得到  $e_{k+1} = O(e_k^2)$ 。未知  $m$  时对  $u = f/f'$  作 Newton:

$$x_{k+1} = x_k - \frac{ff'}{(f')^2 - ff''},$$

因为  $u$  在重根处为单根, 故局部二次。

### 3.11 正式真题三: 无导数 Newton 型迭代

正式真题: 2026 春 Q1 (原卷第 1 页, 10 分)

设  $f \in C^2$  在局部有唯一单根  $x^*$ , 且  $f'(x^*) \neq 0$ 。对

$$x_{k+1} = x_k - \frac{f(x_k)^2}{f(x_k + f(x_k)) - f(x_k)},$$

证明局部收敛并估计收敛阶。

完整解答 (独立核验稿) 令  $a = f'(x^*) \neq 0$ ,  $b = f''(x^*)/2$ ,  $e = x - x^*$ , 并记  $h = f(x)$ 。因为

题目只假设  $f \in C^2$ , 应使用 Peano 余项:

$$f(x) = ae + be^2 + o(e^2), \quad f(x+h) - f(x) = f'(x)h + \frac{1}{2}f''(x)h^2 + o(h^2).$$

由  $h = O(e)$ 、 $f'(x) = a + 2be + o(e)$  和  $f''(x) = 2b + o(1)$ , 分母与分子分别为

$$f(x + f(x)) - f(x) = a^2e + ab(3+a)e^2 + o(e^2),$$

$$f(x)^2 = a^2e^2 + 2abe^3 + o(e^3).$$

相除得到

$$e_+ = \frac{b(1+a)}{a}e^2 + o(e^2) = \frac{f''(x^*)[1+f'(x^*)]}{2f'(x^*)}e^2 + o(e^2).$$

由于分母除以  $e$  趋于  $a^2 \neq 0$ , 当  $e \neq 0$  足够小时分母不为零; 若已经到达  $x^*$  则算法停止。再缩小邻域使  $|e_+| \leq C|e|^2 \leq |e|/2$ , 得到迭代良定义、邻域不变和局部收敛。

若系数非零, 收敛阶为二, 误差常数是

$$\left| \frac{f''(x^*)[1+f'(x^*)]}{2f'(x^*)} \right|.$$

若系数为零, 只能由现有  $C^2$  条件推出  $e_+ = o(e^2)$ ; 没有更强光滑性时, 不能断言某个确定的更高代数阶。

### 易错点

只计算  $g'(x^*) = 0$  不能替代局部收敛证明。完整答案必须说明分母在邻域内不为零、迭代映射把小邻域映回自身, 并给出误差递推。

### 答题方法

局部迭代证明固定采用五步:

1. 检查根的单重性或明确重数;
2. 选邻域使分母/Jacobian 远离奇异;
3. 写 Taylor 或插值余项;
4. 得到  $e_{k+1}$  的主项并证明邻域不变;
5. 最后才取极限判阶和误差常数。

### 自测

1. 构造一个  $|g'(x^*)| < 1$  但从远处初值发散的不动点迭代。
2. 对  $f(x) = x^m$  直接计算标准与修正 Newton 误差。
3. 证明割线法每步只需一次新函数值, 并比较单位函数评估成本下的效率指数。
4. 为什么小残量在多重根附近不一定意味着小前向误差?

5. 写出方程组 Newton 的 inexact forcing 条件及其对局部阶的影响。

### 本节结论

二分法用区间保证换取线性速度; Newton 用导数和好初值换取二次速度; 割线与 Steffensen 用更多历史或函数值换取无导数超线性。所有高速结论都必须先经过良定义和不变邻域检查。

## 3.12 Newton–Kantorovich 的可验证局部条件

标量 Newton 的“初值足够近”可用具体邻域表达。设区间  $I = [x^* - r, x^* + r]$  上

$$|f'(x)| \geq m > 0, \quad |f''(x)| \leq M.$$

由 Taylor 定理,

$$e_{k+1} = e_k - \frac{f(x_k)}{f'(x_k)} = \frac{f''(\xi_k)}{2f'(x_k)} e_k^2,$$

故

$$|e_{k+1}| \leq \frac{M}{2m} |e_k|^2.$$

若  $r \leq 2m/M$ , 且  $x_0 \in I$ , 则  $|e_1| \leq |e_0| \leq r$ , 归纳得到迭代留在邻域中, 分母不为零, 并二次收敛。证明中“不变邻域—分母非零—误差递推”三步不可颠倒。

## 3.13 阻尼 Newton 与充分下降

纯 Newton 步  $p_k = -f(x_k)/f'(x_k)$  远离根时可能越出吸引域。可对 merit function  $\Phi(x) = \frac{1}{2}f(x)^2$  使用阻尼步

$$x_{k+1} = x_k + \alpha_k p_k, \quad 0 < \alpha_k \leq 1.$$

因

$$\Phi'(x_k)p_k = f(x_k)f'(x_k) \left( -\frac{f(x_k)}{f'(x_k)} \right) = -f(x_k)^2 < 0$$

(当  $f(x_k) \neq 0, f'(x_k) \neq 0$ ), Newton 方向是  $\Phi$  的下降方向。回溯可选到满足 Armijo 条件的  $\alpha_k$ 。进入局部二次邻域后通常接受  $\alpha_k = 1$ , 恢复快速率。

## 3.14 方程组 Newton 的误差与线性求解

对  $F: \mathbb{R}^d \rightarrow \mathbb{R}^d$ , Newton 步解

$$J_F(x_k)s_k = -F(x_k), \quad x_{k+1} = x_k + s_k.$$

若  $J_F(x^*)$  可逆且 Jacobian 局部 Lipschitz, 则

$$\|e_{k+1}\| \leq C\|e_k\|^2.$$

实际中通常不显式求逆，而解线性方程。若线性子问题只近似求解，

$$J_F(x_k)s_k = -F(x_k) + r_k,$$

需要 forcing term 控制  $\|r_k\| \leq \eta_k \|F(x_k)\|$ 。固定  $\eta_k < 1$  常给线性型局部率； $\eta_k \rightarrow 0$  可超线性；与残量同阶的更强控制可保二次率。

### 3.15 Steffensen 的 $C^2$ 边界

本章唯一解答位置已统一使用 Peano 余项：

$$e_{k+1} = \frac{f''(x^*)[1 + f'(x^*)]}{2f'(x^*)} e_k^2 + o(e_k^2).$$

当系数非零时，收敛阶为二，误差常数为其绝对值；当系数为零时，在只有  $f \in C^2$  的条件下只能断言  $e_{k+1} = o(e_k^2)$ ，不能虚构确定的更高代数阶。任何写成  $O(e_k^3)$  的版本都需要更强光滑性。

#### 自测

1. 给出保证 Newton 邻域不变的一组  $m, M, r$  条件。
2. 为什么 Newton 方向是平方残量的下降方向？
3. 近似 Newton 线性子问题的残量如何影响局部率？
4. Steffensen 二次项系数为零时，在  $C^2$  下能断言什么、不能断言什么？

## 第四章 数值线性代数

### 4.1 范数、条件数与稳定性

相容矩阵范数满足  $\|Ax\| \leq \|A\| \|x\|$ 。谱半径  $\rho(A) \leq \|A\|$ ，但一般不等号。线性系统相对扰动一阶界含  $\kappa(A) = \|A\| \|A^{-1}\|$ ；后向稳定算法的前向误差仍可能被大条件数放大。

### 4.2 直接法与分解

Gaussian 消元约  $2n^3/3$  flops；部分主元抑制增长因子，是通用实践但非无条件前向稳定保证。若所有顺序主子式非零，则存在单位下三角  $L$  与上三角  $U$ ；固定单位对角约定时唯一。Hermitian 正定矩阵唯一分解  $A = LL^*$  且  $L$  对角正；QR 对满列秩最小二乘避免正规方程把条件数平方。

**定理 4.1** (Cholesky 存在唯一性)。若  $A = A^* > 0$ ，则存在唯一正对角下三角  $L$  使  $A = LL^*$ 。

完整证明. 对维数归纳。写  $A = \begin{bmatrix} a & r^* \\ r & B \end{bmatrix}$ 。  $a > 0$ ，令  $\ell = \sqrt{a}$ 、 $v = r/\ell$ 。Schur 补  $S = B - vv^*$  正定，因为对任意  $y$  取  $x = -r^*y/a$  代入二次型。归纳得  $S = MM^*$ ，于是  $L = \begin{bmatrix} \ell & 0 \\ v & M \end{bmatrix}$ 。正对角逐列比较给唯一性。  $\square$

### 4.3 Jacobi、Gauss–Seidel 与 SOR

写  $A = D - L - U$ ，则

$$B_J = D^{-1}(L + U), \quad B_{GS} = (D - L)^{-1}U,$$

SOR 为  $x^{k+1} = (D - \omega L)^{-1}([(1 - \omega)D + \omega U]x^k + \omega b)$ 。定常迭代对任意初值收敛当且仅当  $\rho(B) < 1$ 。严格对角占优给 Jacobi/GS 常用充分条件；SPD 给 GS 收敛。SPD 下 SOR 的标准收敛区间是  $0 < \omega < 2$ ； $0 < \omega < 1$  只是其中的充分子区间，不是充要区间。

### 4.4 Krylov 与共轭梯度

$\mathcal{K}_k(A, r_0) = \text{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\}$ 。SPD 情形 CG 递推

$$\alpha_k = \frac{r_k^T r_k}{p_k^T A p_k}, \quad x_{k+1} = x_k + \alpha_k p_k, \quad r_{k+1} = r_k - \alpha_k A p_k, \quad \beta_k = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}, \quad p_{k+1} = r_{k+1} + \beta_k p_k.$$

残量彼此正交、方向  $A$ -共轭， $x_k$  在  $x_0 + \mathcal{K}_k$  最小化能量范数误差；精确算术至多  $n$  步终止。

**定理 4.2** (CG 条件数界). 若  $A$  SPD,  $\kappa = \lambda_{\max}/\lambda_{\min}$ , 则

$$\|x_k - x^*\|_A \leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x^*\|_A.$$

#### 证明思路

用最小化性质把误差写为  $\min_{p(0)=1, \deg p \leq k} \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |p(\lambda)|$ , 再用缩放 Chebyshev 多项式。标准通用上界保留前因子 2; 特殊归一化可产生更锐常数, 但不能将其当作通用结论。

预条件 CG 对  $M^{-1}A$  工作但应在  $M$ -内积/对称变换中保持自伴; GMRES 最小化残量, 适合一般非对称矩阵; MINRES 用于 Hermitian 不定系统, 考试主要掌握其定位与适用条件, 不求高级实现细节。

### 4.5 特征值、最小二乘与 SVD

幂法需唯一模最大特征值且初向量在其特征向量方向分量非零, 线性率约  $|\lambda_2/\lambda_1|$ 。反幂加移位收敛到离移位最近的特征值; Hermitian 情形 Rayleigh 商迭代局部可达三次。QR 迭代以相似变换保持特征值, 先 Hessenberg 化把每步从  $O(n^3)$  降至  $O(n^2)$ 。

SVD  $A = U\Sigma V^*$  给伪逆、最小二乘和低秩逼近。Eckart-Young: 谱范数最优误差  $\sigma_{k+1}$ , Frobenius 最优误差  $(\sum_{j>k} \sigma_j^2)^{1/2}$ 。正规方程  $A^*Ax = A^*b$  使条件数近似平方, QR/SVD 更稳定。

### 4.6 向量/矩阵范数与谱半径

常用向量范数为

$$\|x\|_1 = \sum_i |x_i|, \quad \|x\|_2 = (x^*x)^{1/2}, \quad \|x\|_\infty = \max_i |x_i|.$$

有限维空间中任意两范数等价, 但等价常数依赖维数, 例如  $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$ 。诱导矩阵范数  $\|A\| = \max_{x \neq 0} \|Ax\|/\|x\|$  自动满足相容性和次乘性,

$$\|A\|_1 = \max_j \sum_i |a_{ij}|, \quad \|A\|_\infty = \max_i \sum_j |a_{ij}|, \quad \|A\|_2 = \sqrt{\lambda_{\max}(A^*A)}.$$

Frobenius 范数不是由常用向量范数诱导, 但满足次乘性且酉不变。

**定理 4.3** (谱半径与矩阵幂). 对任意矩阵范数  $\rho(A) \leq \|A\|$ 。此外

$$\rho(A) = \lim_{k \rightarrow \infty} \|A^k\|^{1/k}, \quad A^k \rightarrow 0 \iff \rho(A) < 1.$$

完整证明. 若  $Av = \lambda v$  且  $v \neq 0$ , 则相容性给

$$|\lambda| \|v\| = \|Av\| \leq \|A\| \|v\|,$$

所以  $\rho(A) \leq \|A\|$ 。对任意  $k$ , 同理有  $\rho(A)^k = \rho(A^k) \leq \|A^k\|$ , 从而  $\rho(A) \leq \liminf_k \|A^k\|^{1/k}$ 。

为证反向不等式, 在复数域写  $A = SJS^{-1}$ 。大小为  $m$  的 Jordan 块  $J_\lambda = \lambda I + N$  满足

$$J_\lambda^k = \sum_{r=0}^{m-1} \binom{k}{r} \lambda^{k-r} N^r.$$

若  $\lambda \neq 0$ , 其范数至多为  $C_m k^{m-1} |\lambda|^{k-m+1}$ ; 若  $\lambda = 0$ , 当  $k \geq m$  时该块为零。因此对任意  $\eta > 0$ , 吸收有限个小  $k$  后存在  $C_\eta$  使

$$\|A^k\| \leq C_\eta(\rho(A) + \eta)^k.$$

取  $k$  次根、令  $k \rightarrow \infty$  再令  $\eta \downarrow 0$ , 得到  $\limsup_k \|A^k\|^{1/k} \leq \rho(A)$ , 故 Gelfand 公式成立。

若  $\rho(A) < 1$ , 上式取  $\eta > 0$  使  $\rho(A) + \eta < 1$ , 便有  $A^k \rightarrow 0$ 。反之, 若  $A^k \rightarrow 0$ , 对任一特征对  $A^k v = \lambda^k v \rightarrow 0$ , 故每个  $|\lambda| < 1$ , 即  $\rho(A) < 1$ 。这里不能把 Jordan 块静默当成可对角化矩阵; 多项式因子正是亏损情形仍需控制的部分。□

## 4.7 扰动、条件数、残量和后向稳定

若  $(A + \Delta A)(x + \Delta x) = b + \Delta b$  且  $\|A^{-1}\| \|\Delta A\| < 1$ , 则

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\|\Delta A\|/\|A\|} \left( \frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|} \right).$$

推导从  $(I + A^{-1}\Delta A)\Delta x = A^{-1}(\Delta b - \Delta Ax)$  开始, 并用 Neumann 级数控制逆。它明确说明相对误差界的分母条件; 省略“小扰动”假设会把局部界错误写成无条件结论。

前向误差衡量答案离真解多远; 后向误差问近似解恰好解了哪个邻近问题。带部分主元 Gaussian 消元通常具有小分量后向误差, 但增长因子可能在构造例上很大; 实践可靠不等于无条件小前向误差。真实误差还会被  $\kappa(A)$  放大, 见定理 4.13。

**例题 4.1** (条件数与扰动方向). 取  $A = \text{diag}(1, 10^{-6})$ ,  $b = (1, 10^{-6})^T$ , 则  $x = (1, 1)^T$ ,  $\kappa_2(A) = 10^6$ 。若  $\Delta b = (0, 10^{-8})^T$ , 输入相对扰动约  $10^{-8}$ , 第二个解分量改变  $10^{-2}$ 。界给最坏放大  $10^6$ , 实际放大取决于扰动是否对准最敏感奇异向量。

## 4.8 Gaussian 消元、LU 与主元

第  $k$  步消元以乘子  $m_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$  消去主元下方元素。把乘子存入  $L$ , 剩余上三角为  $U$ 。无主元稠密消元约  $\frac{2}{3}n^3 + O(n^2)$  flops, 随后每个右端前/回代各约  $n^2$ 。

**定理 4.4** (无主元 LU 的存在唯一性). 矩阵  $A \in \mathbb{C}^{n \times n}$  的所有顺序主子式非零, 当且仅当 Gaussian 消元无需换行即可完成, 并有  $A = LU$ , 其中  $L$  单位下三角、 $U$  上三角。此归一化下分解唯一。

**完整证明.** 非零顺序主子式保证每一步 Schur 补的主元非零, 故消元可进行并构造  $L, U$ 。反之若  $A = LU$ , 前  $k$  阶顺序主子式等于  $\det L_{1:k, 1:k} \det U_{1:k, 1:k} = \prod_{j=1}^k u_{jj}$ , 消元主元非零时它非零。若  $LU = \tilde{L}\tilde{U}$ , 则  $\tilde{L}^{-1}L = \tilde{U}U^{-1}$  同时为单位下三角和上三角, 只能是  $I$ , 故唯一。□

部分主元在当前列选择最大绝对值, 得到  $PA = LU$ , 保证  $|m_{ik}| \leq 1$ ; 完全主元在剩余子矩阵选最大元, 需行列置换。主元策略减少中间元素增长并避免除以小数, 但部分主元的增长因子最坏仍可指数大。带状矩阵若消元顺序能控制填充, 可把存储和成本降到与带宽有关; 具体复杂度必须同时说明上、下带宽与填充假设。

**例题 4.2** (两步 LU 手算).

$$A = \begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}.$$

若首元改为  $10^{-16}$  而第二行首元为 1, 无主元乘子为  $10^{16}$ ; 部分主元先换行可避免该放大。这是主元作用的最小示例。

## 4.9 Cholesky、QR 与最小二乘

Hermitian 正定性不仅保证 Cholesky 存在, 还保证所有 Schur 补正定, 因此无需选主元。Cholesky 约  $n^3/3$  flops, 是一般 LU 的一半。半正定但奇异时对角元可能为零, 唯一正对角 Cholesky 结论不再适用。

满列秩  $A \in \mathbb{C}^{m \times n}$  的约化 QR 为  $A = QR$ ,  $Q^*Q = I_n$ ,  $R$  上三角且若规定对角正则唯一。Householder 反射  $H = I - 2vv^*/(v^*v)$  一次消去一列下方分量, 数值稳定且适合块化; Givens 旋转一次消一个元素, 适合稀疏和增量更新。最小二乘转化为

$$\min_x \|Ax - b\|_2^2 = \min_x \left\| \begin{bmatrix} Rx - Q^*b \\ (I - QQ^*)b \end{bmatrix} \right\|_2^2,$$

故解三角系统  $Rx = Q^*b$ 。正规方程的条件数  $\kappa_2(A^*A) = \kappa_2(A)^2$ , 这是 QR/SVD 更稳的根本原因。

**例题 4.3** (QR 最小二乘). 令  $A = (1, 1)^T$ ,  $b = (1, 2)^T$ .  $Q = 2^{-1/2}(1, 1)^T$ ,  $R = \sqrt{2}$ , 故  $x = Q^*b/R = 3/2$ ; 残量  $(-1/2, 1/2)^T$  与  $A$  正交。若只写正规方程而不检查残量正交, 就漏掉最重要的几何解释。

## 4.10 定常迭代的充要条件与充分条件

一般分裂  $A = M - N$  给  $x^{k+1} = Bx^k + c$ ,  $B = M^{-1}N$ 。误差满足  $e^k = B^k e^0$ 。

**定理 4.5** (定常迭代收敛判据). 对有限维矩阵, 迭代对任意初值收敛到唯一解, 当且仅当  $\rho(B) < 1$ 。

**完整证明.** 若  $\rho(B) < 1$ , 由定理 4.3 得  $B^k \rightarrow 0$ , 故误差趋零。反之若对任意初值收敛, 则  $B^k e \rightarrow 0$  对任意  $e$  成立, 从而  $B^k \rightarrow 0$ , 再由谱半径判据得  $\rho(B) < 1$ 。□

严格行对角占优时, 若  $Bx = \lambda x$ , 选  $|x_i| = \|x\|_\infty$  的分量, 可对 Jacobi/GS 逐行估计得到  $|\lambda| < 1$ 。SPD 情形 Gauss-Seidel 可解释为对二次能量  $\frac{1}{2}x^T Ax - b^T x$  的逐坐标精确最小化, 能量严格下降并收敛。

**定理 4.6** (SPD 下 SOR 的准确区间). 若  $A$  Hermitian 正定, 则 SOR 对任意初值收敛当且仅当  $0 < \omega < 2$ 。

### 证明思路

把一次 SOR 看作带松弛的坐标校正, 可得能量差含正因子  $\omega(2 - \omega)$ ; 当  $0 < \omega < 2$  时能量单调下降并结合正定性得到误差趋零。若  $\omega \leq 0$  或  $\omega \geq 2$ , 可从迭代矩阵行列式/简单一维 SPD 例  $A = [1]$  (误差因子  $1 - \omega$ ) 看出不能对全部 SPD 矩阵收敛。因此  $0 < \omega < 2$  只

是充分子区间，不是充要区间。

**例题 4.4** (Jacobi 与 GS 谱半径). 对  $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ ,

$$B_J = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix}, \quad \rho(B_J) = 1/2, \quad B_{GS} = \begin{bmatrix} 0 & 1/2 \\ 0 & 1/4 \end{bmatrix}, \quad \rho(B_{GS}) = 1/4.$$

二者都收敛，GS 渐近因子更小。不能仅比较一次迭代误差就断言渐近速度。

## 4.11 CG 的推导、正交性、最小化与有限终止

对 SPD 系统，解  $Ax = b$  等价于最小化

$$\phi(x) = \frac{1}{2}x^T Ax - b^T x, \quad \nabla \phi(x) = -r.$$

在方向  $p_k$  上精确线搜索给  $\alpha_k = (r_k^T p_k)/(p_k^T A p_k)$ ；由递推正交性又等于  $r_k^T r_k/(p_k^T A p_k)$ 。选择  $p_{k+1} = r_{k+1} + \beta_k p_k$  并强制  $p_{k+1}^T A p_k = 0$ ，得到  $\beta_k = r_{k+1}^T r_{k+1}/(r_k^T r_k)$ 。

**定理 4.7** (CG 三个核心不变量). 精确算术下，对 SPD  $A$ ：

$$r_i^T r_j = 0 \quad (i \neq j), \quad p_i^T A p_j = 0 \quad (i \neq j),$$

且

$$x_k = \arg \min_{x \in x_0 + \mathcal{K}_k(A, r_0)} \|x - x^*\|_A.$$

因此至多  $n$  步终止；若  $A$  只有  $s$  个不同特征值，则至多  $s$  步终止。

**完整证明.** 用  $r_k = b - Ax_k$ ，可归纳证明  $\text{span}\{p_0, \dots, p_{k-1}\} = \mathcal{K}_k$ 。精确线搜索使  $r_{k+1} \perp p_k$ ，而  $r_{k+1} = r_k - \alpha_k A p_k$  与既有共轭性进一步给它正交于所有旧方向；由此递推得到残量两两正交和方向两两  $A$ -共轭。对任意  $v \in \mathcal{K}_k$ ，

$$(x^* - x_k)^T A v = r_k^T v = 0,$$

即误差与子空间在  $A$ -内积下正交，所以  $x_k$  是正交投影并具有最小化性质。非零残量彼此正交，至多有  $n$  个；最小多项式次数给不同特征值数细化。□

**完整证明. CG 条件数界的 Chebyshev 链.** 由最小化性质和  $e_k = p_k(A)e_0$ ，

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \min_{\substack{p(0)=1 \\ \deg p \leq k}} \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |p(\lambda)|.$$

把谱区间仿射映到  $[-1, 1]$ ，选缩放 Chebyshev 多项式

$$p(\lambda) = \frac{T_k\left(\frac{\lambda_{\max} + \lambda_{\min} - 2\lambda}{\lambda_{\max} - \lambda_{\min}}\right)}{T_k\left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right)}.$$

分子在谱区间绝对值不超过 1。利用  $T_k(t) = \frac{1}{2}[(t + \sqrt{t^2 - 1})^k + (t - \sqrt{t^2 - 1})^k]$  并代入  $\kappa = \lambda_{\max}/\lambda_{\min}$ , 得到

$$\|e_k\|_A \leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|e_0\|_A.$$

通用上界的前因子 2 不能因某个特殊归一化的简写而静默删除。□

预条件的目标不是让  $\|M^{-1}A\|$  小, 而是使对称预条件算子特征值聚集且应用  $M^{-1}$  便宜。PCG 要求  $M$  SPD, 并在  $M$ -内积或  $M^{-1/2}AM^{-1/2}$  上保持自伴。GMRES 用 Arnoldi 生成正交基并解小型最小二乘, 残量范数单调不增但存储随步数增长; MINRES 适合 Hermitian 不定系统, 正文仅要求理解三对角化与最小残量定位。

## 4.12 幂法、反幂、Rayleigh 商与 QR

若  $A$  可对角化且  $|\lambda_1| > |\lambda_2| \geq \dots$ , 初向量在  $v_1$  上系数非零, 则归一化幂法方向误差为  $O(|\lambda_2/\lambda_1|^k)$ 。若最大模特征值不唯一或初向量正交于主特征子空间, 结论失效。

**定理 4.8** (幂法的投影条件与模比收敛). 设  $A = V\Lambda V^{-1}$  可对角化,  $|\lambda_1| > |\lambda_2| \geq \dots$ , 且  $x_0 = \sum_{j=1}^n c_j v_j$  中  $c_1 \neq 0$ 。若每步以任一不使主分量消失的齐次归一化进行, 则归一化迭代方向收敛到  $v_1$  的方向, 误差为  $O(|\lambda_2/\lambda_1|^k)$ 。

完整证明. 直接展开

$$A^k x_0 = \lambda_1^k c_1 \left[ v_1 + \sum_{j=2}^n \frac{c_j}{c_1} \left( \frac{\lambda_j}{\lambda_1} \right)^k v_j \right].$$

括号中余项在任意有限维范数下至多为  $C(|\lambda_2/\lambda_1|^k)$ 。齐次归一化只除去公共因子  $\lambda_1^k c_1$ , 并在  $v_1$  的一个邻域内连续, 故方向误差具有同一量级。若  $c_1 = 0$ , 主特征向量方向从未出现; 若最大模特征值不唯一, 括号中相应项不衰减。这两种情形正说明假设不能删除。反幂法把同一证明用于  $(A - \mu I)^{-1}$ , 其主模特征值对应离  $\mu$  最近的原特征值。□

反幂法对  $(A - \mu I)^{-1}$  做幂法, 收敛到离移位  $\mu$  最近且投影非零的特征向量; 分解  $A - \mu I$  可在迭代中复用。Rayleigh 商迭代每步取  $\mu_k = x_k^* A x_k / (x_k^* x_k)$  并解移位系统; Hermitian 简单特征对附近具有局部三次收敛, 但一般非正规矩阵不能照搬。

QR 迭代  $A_k = Q_k R_k$ ,  $A_{k+1} = R_k Q_k = Q_k^* A_k Q_k$  是酉相似变换。先用 Householder 把一般矩阵化为上 Hessenberg (对称矩阵化为三对角), 再做隐式移位 QR; Hessenberg 结构把每步成本降为  $O(n^2)$ 。收敛表现为次对角元趋零并 deflation, 而不是“直接把  $R$  当特征值”。

**命题 4.9** (QR 步的相似性与结构入口). 若  $A_k - \mu_k I = Q_k R_k$  且  $Q_k$  酉, 则移位 QR 更新

$$A_{k+1} = R_k Q_k + \mu_k I = Q_k^* A_k Q_k$$

保持特征值。若  $A_k$  为上 Hessenberg, 使用相邻 Givens 旋转实现该步时,  $A_{k+1}$  仍为上 Hessenberg。

证明. 由  $R_k = Q_k^* (A_k - \mu_k I)$  立即得到相似式。对 Hessenberg 矩阵, 消去第  $j$  列唯一的次对角元只需作用于第  $j, j+1$  行; 把这些旋转从左侧依次“追赶”到右侧时, 任何新产生的非零元至多位于下一条次对角线, 并在下一步被消去。最终相乘仍只保留第一条次对角线。该证明只给结构保持; 次对角元是否趋零还依赖移位和谱分离条件。□

**例题 4.5** (幂法速度).  $A = \text{diag}(4, 2, 1)$ ,  $x_0 = (1, 1, 1)^T$ . 归一化前  $A^k x_0 = 4^k(1, 2^{-k}, 4^{-k})^T$ , 方向误差由  $2^{-k}$  主导, 与  $|\lambda_2/\lambda_1|^k$  一致。

### 4.13 特征值扰动、Gershgorin 与 SVD

Gershgorin 圆盘

$$\lambda(A) \subset \bigcup_i \{z : |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\}.$$

若某组  $k$  个圆盘与其余圆盘不相交, 则该连通组恰含  $k$  个特征值 (计代数重数)。Hermitian 扰动满足 Weyl 界  $|\lambda_i(A + E) - \lambda_i(A)| \leq \|E\|_2$ 。一般可对角化  $A = V\Lambda V^{-1}$  时, Bauer–Fike 给每个扰动特征值  $\mu$

$$\min_i |\mu - \lambda_i| \leq \kappa_2(V) \|E\|_2.$$

因此非正规矩阵的特征值可远比  $\|E\|$  敏感。

**定理 4.10** (Gershgorin 包含与 Bauer–Fike 扰动界). 设  $R_i = \sum_{j \neq i} |a_{ij}|$ 。每个  $\lambda \in \text{spec}(A)$  落在某个圆盘  $\{z : |z - a_{ii}| \leq R_i\}$  中。若  $A = V\Lambda V^{-1}$  可对角化,  $\mu \in \text{spec}(A + E)$ , 则

$$\min_i |\mu - \lambda_i| \leq \kappa_2(V) \|E\|_2.$$

**完整证明.** 取  $Ax = \lambda x$  的非零特征向量, 并选  $|x_i| = \|x\|_\infty$ 。第  $i$  行给

$$|\lambda - a_{ii}| |x_i| = \left| \sum_{j \neq i} a_{ij} x_j \right| \leq R_i |x_i|,$$

除以  $|x_i| > 0$  即得 Gershgorin 包含。

再取  $(A + E)y = \mu y$ ,  $y \neq 0$ 。若  $\mu \notin \text{spec}(A)$ , 则

$$y = (\mu I - A)^{-1} E y = V(\mu I - \Lambda)^{-1} V^{-1} E y.$$

取 2-范数并约去  $\|y\|_2$ :

$$1 \leq \|V\|_2 \|V^{-1}\|_2 \max_i |\mu - \lambda_i|^{-1} \|E\|_2.$$

整理即得所述界; 若  $\mu \in \text{spec}(A)$ , 左端最小距离为零, 结论自动成立。这也说明一般非正规情形不能删除  $\kappa_2(V)$ 。□

SVD  $A = U\Sigma V^*$  对任意矩阵存在。满列秩最小二乘解  $x = A^\dagger b$ ; 秩亏时  $A^\dagger b$  是所有最小二乘解中的最小 2-范数解。

**定理 4.11** (Eckart–Young–Mirsky). 设奇异值  $\sigma_1 \geq \dots \geq \sigma_r > 0$ , 截断  $A_k = \sum_{i=1}^k \sigma_i u_i v_i^*$ 。则

$$\min_{\text{rank}(B) \leq k} \|A - B\|_2 = \sigma_{k+1}, \quad \min_{\text{rank}(B) \leq k} \|A - B\|_F = \left( \sum_{i>k} \sigma_i^2 \right)^{1/2}.$$

**完整证明.** 先证谱范数下界。令  $V_{k+1} = \text{span}\{v_1, \dots, v_{k+1}\}$ 。若  $\text{rank}(B) \leq k$ , 则  $B$  在  $k+1$  维

空间  $V_{k+1}$  上的限制有非零核; 可取  $z \in V_{k+1} \cap \ker B$ ,  $\|z\|_2 = 1$ 。写  $z = \sum_{j=1}^{k+1} c_j v_j$ , 得到

$$\|A - B\|_2 \geq \|(A - B)z\|_2 = \|Az\|_2 = \left( \sum_{j=1}^{k+1} \sigma_j^2 |c_j|^2 \right)^{1/2} \geq \sigma_{k+1}.$$

而  $A - A_k = \sum_{j>k} \sigma_j u_j v_j^*$  的谱范数正是  $\sigma_{k+1}$ , 故下界可达。

再证 Frobenius 下界。设  $P$  为到  $\mathcal{R}(B)$  的正交投影, 则  $(I - P)B = 0$ , 并且

$$\|A - B\|_F^2 = \|(I - P)A\|_F^2 + \|P(A - B)\|_F^2 \geq \|(I - P)A\|_F^2.$$

由  $\text{rank}(P) \leq k$  与迹的循环性,

$$\|(I - P)A\|_F^2 = \text{tr}(A^*A) - \text{tr}(PAA^*) \geq \sum_j \sigma_j^2 - \sum_{j=1}^k \sigma_j^2 = \sum_{j>k} \sigma_j^2.$$

中间不等式是 Rayleigh–Ritz 原理: 秩不超过  $k$  的正交投影在  $AA^*$  上所能截取的迹至多为最大的  $k$  个特征值之和。截断 SVD  $A_k$  使两个正交分量恰好分离, 达到该下界。□

**例题 4.6** (重奇异值与最优解不唯一). 取  $A = \text{diag}(5, 3, 3, 1)$ , 令  $k = 2$ 。截断 SVD 给

$$\min_{\text{rank}(B) \leq 2} \|A - B\|_2 = 3, \quad \min_{\text{rank}(B) \leq 2} \|A - B\|_F = \sqrt{3^2 + 1^2} = \sqrt{10}.$$

由于截断位置满足  $\sigma_2 = \sigma_3 = 3$ , 在对应二维奇异子空间中保留任意一维方向都给出同一最优值, 最优低秩矩阵不唯一; 若  $\sigma_k > \sigma_{k+1}$ , 则在通常的非退化情形下截断子空间被唯一确定。

## 4.14 结构化更新、矩阵逆迭代与广义特征值

**命题 4.12** (Sherman–Morrison 秩一更新). 若  $A$  可逆且  $1 + v^* A^{-1} u \neq 0$ , 则

$$(A + uv^*)^{-1} = A^{-1} - \frac{A^{-1}uv^*A^{-1}}{1 + v^*A^{-1}u}.$$

分母为零时,  $A + uv^*$  奇异, 不能静默使用公式。

*证明.* 把右端记作  $B$ , 直接展开  $(A + uv^*)B$ 。含  $uv^*A^{-1}$  的两项合并系数为

$$1 - \frac{1 + v^*A^{-1}u}{1 + v^*A^{-1}u} = 0,$$

余下为  $I$ 。右乘验证相同。□

Newton–Schulz 迭代

$$X_{k+1} = 2X_k - X_k A X_k$$

可理解为矩阵求逆的 Newton 法。令右残量  $E_k = I - X_k A$ , 则

$$E_{k+1} = I - (2X_k - X_k A X_k)A = E_k^2.$$

因此若某个相容范数下  $\|E_0\| < 1$ , 则  $\|E_k\| \leq \|E_0\|^{2^k}$ , 并且  $X_k = (I - E_k)A^{-1} \rightarrow A^{-1}$ 。条件只控制局部收敛; 有限精度中还要防止矩阵乘法误差和不当缩放。

广义特征值问题写成正则矩阵铅笔

$$(A - \lambda B)x = 0, \quad \det(A - \lambda B) \neq 0.$$

广义 Schur (QZ) 分解以酉矩阵  $Q, Z$  给

$$Q^*AZ = S, \quad Q^*BZ = T,$$

其中  $S, T$  上三角; 射影特征值由对角对  $(s_{ii}, t_{ii})$  给出,  $t_{ii} = 0$  表示无穷特征值。酉等价保持铅笔的广义特征结构, 而单独对  $B^{-1}A$  求特征值在  $B$  病态或奇异时并不安全。

**例题 4.7** (矩阵最小二乘的向量化). 对  $X \in \mathbb{R}^{p \times q}$ ,

$$\min_X \|AXB - C\|_F \iff \min_x \|(B^\top \otimes A)x - \text{vec } C\|_2, \quad x = \text{vec } X.$$

故可解性、唯一性和最小范数解都归结为 Kronecker 设计矩阵。若要求任意  $C$  下  $X$  唯一, 需  $B^\top \otimes A$  满列秩; 等价地,  $A$  满列秩且  $B$  满行秩。实际计算优先利用 QR/SVD 和矩阵结构, 不显式形成可能巨大的 Kronecker 矩阵。

## 4.15 真题迁移、方法比较与检查

六期真题把本章多个“教材边缘”变成实际能力: 2023 秋 Q2 移位 QR; 2024 春 Q2 CG 与非 Hermitian Krylov; 2024 秋 Q1–Q3 分别考 SVD、GS 和谱半径; 2025 春 Q3 考 Sherman–Morrison 结构; 2025 秋 Q1/Q3 考 Newton–Schulz 与 QZ; 2026 春 Q3 考矩阵最小二乘。正文主干给定理条件; 结构化秩一更新、Newton–Schulz 和 QZ 属于真题驱动的低频扩展, 在真题章掌握题目所需推导即可。

### 易错点

不要把  $\rho(A)$  与任意矩阵范数写成等号; 不要把小残量直接当小误差; 不要无条件声称部分主元“稳定”; 不要对非 SPD 系统使用 CG; 不要把  $M^{-1}A$  的非对称实现直接称 PCG; 不要把正规方程当作最稳定的最小二乘算法。

### 复习检查

能否证明 LU 唯一性、定常迭代谱半径判据和 CG 三个不变量? 能否写出 SOR 的准确区间并区分充要区间与充分子区间? 能否从 Chebyshev 多项式得到带前因子 2 的 CG 界? 能否比较 LU/Cholesky/QR/SVD 的条件与成本? 能否说明幂法、反幂、RQI、QR 的不同收敛条件, 并使用 Gershgorin/Weyl/Bauer–Fike 做扰动判断?

## 4.16 非正规矩阵：特征值不足以描述敏感性

**定理 4.13** (残量与前向误差的双边界). 设  $A$  可逆,  $Ax = b$ , 近似解  $\hat{x}$  的残量为  $r = b - A\hat{x}$ 。对任意诱导矩阵范数,

$$\frac{1}{\kappa(A)} \frac{\|r\|}{\|b\|} \leq \frac{\|x - \hat{x}\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}, \quad \kappa(A) = \|A\| \|A^{-1}\|.$$

*证明.* 令  $e = x - \hat{x}$ , 则  $Ae = r$ 。由

$$\|e\| \leq \|A^{-1}\| \|r\|, \quad \|b\| = \|Ax\| \leq \|A\| \|x\|$$

相乘整理得上界。另一方面,

$$\|r\| = \|Ae\| \leq \|A\| \|e\|, \quad \|x\| = \|A^{-1}b\| \leq \|A^{-1}\| \|b\|$$

整理得下界。可逆性与诱导范数的相容性都是条件。  $\square$

正规矩阵可酉对角化, 函数与幂的范数直接由特征值控制。非正规矩阵即使所有特征值都在单位圆内, 也可能有显著暂态增长。例

$$A = \begin{pmatrix} \lambda & M \\ 0 & \lambda \end{pmatrix} = \lambda I + N, \quad N^2 = 0.$$

则

$$A^k = \lambda^k I + k\lambda^{k-1}N.$$

当  $|\lambda| < 1$  时最终衰减, 但  $M$  大时中间的  $k|\lambda|^{k-1}|M|$  可很大; 当  $|\lambda| = 1, M \neq 0$  时幂不有界。单位圆上的特征值必须半单, 正是 PDE 放大矩阵临界分析反复出现的条件。

Bauer–Fike 定理: 若  $A = V\Lambda V^{-1}$  可对角化, 则

$$\min_i |\tilde{\lambda} - \lambda_i| \leq \kappa(V) \|E\|$$

对  $A + E$  的任意特征值  $\tilde{\lambda}$  成立。非正规性通过  $\kappa(V)$  放大特征值扰动。

## 4.17 LU、主元与后向误差

Gaussian 消元把  $A$  分为  $LU$ 。没有选主元时, 若主元很小, 消元乘子会很大, 舍入误差被增长因子放大。部分选主元在每列选择绝对值最大的可用元素, 使乘子  $|\ell_{ij}| \leq 1$ 。经典分析给计算因子满足

$$A + \Delta A = \hat{L}\hat{U}, \quad |\Delta A| \lesssim u |\hat{L}| |\hat{U}|,$$

因此误差还取决于增长因子, 而不是无条件等于  $O(u)\|A\|$ 。

Cholesky 对称正定矩阵写成

$$A = LL^T.$$

正定性保证每一步平方根内部为正；若矩阵只对称不正定，分解可能中断，需要带主元的  $LDL^T$  等方法。

## 4.18 QR、最小二乘与秩亏

满列秩  $A \in \mathbb{R}^{m \times n}$  的最小二乘满足正规方程

$$A^T A x = A^T b.$$

但

$$\kappa_2(A^T A) = \kappa_2(A)^2,$$

显式形成正规方程会平方条件数。若  $A = QR$ ,  $Q^T Q = I$ , 则

$$\|Ax - b\|^2 = \|Rx - Q^T b\|^2 + \|(I - QQ^T)b\|^2,$$

解三角方程  $Rx = Q^T b$  更稳定。

秩亏时 SVD

$$A = U \Sigma V^T$$

给 Moore–Penrose 伪逆

$$A^+ = V \Sigma^+ U^T.$$

$x^+ = A^+ b$  是所有最小二乘解中二范数最小者。截断小奇异值或 Tikhonov 正则

$$\min_x \|Ax - b\|^2 + \lambda \|x\|^2$$

以偏差换稳定性。

## 4.19 Jacobi、Gauss–Seidel 与 SOR

采用课程中的符号约定

$$A = D - L - U,$$

其中  $D = \text{diag}(a_{11}, \dots, a_{nn})$ ,  $-L$  与  $-U$  分别是  $A$  的严格下三角和严格上三角部分。若另一本资料写  $A = D + L_+ + U_+$ , 则  $L_+ = -L$ ,  $U_+ = -U$ ; 代公式前必须先统一符号。三种迭代的矩阵形式与逐分量更新分别为

$$B_J = D^{-1}(L + U), \quad x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right),$$

$$B_{GS} = (D - L)^{-1}U, \quad x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j < i} a_{ij} x_j^{(k+1)} - \sum_{j > i} a_{ij} x_j^{(k)} \right),$$

$$B_\omega = (D - \omega L)^{-1}[(1 - \omega)D + \omega U],$$

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left( b_i - \sum_{j < i} a_{ij}x_j^{(k+1)} - \sum_{j > i} a_{ij}x_j^{(k)} \right).$$

显然  $\omega = 1$  时 SOR 退化为 Gauss-Seidel。

**定理 4.14** (定常迭代的充要谱半径判据). 分裂  $A = M - N$  给出  $x^{(k+1)} = Bx^{(k)} + c$ , 其中  $B = M^{-1}N$ . 对每个初值都收敛到  $Ax = b$  的解, 当且仅当  $\rho(B) < 1$ 。

完整证明. 若  $x^*$  是不动点, 则误差满足  $e^{(k)} = B^k e^{(0)}$ . 若  $\rho(B) < 1$ , 把  $B$  化为复 Jordan 形. 每个大小为  $s$  的 Jordan 块的  $k$  次幂由  $\lambda^k$  乘以次数至多  $s-1$  的  $k$  的多项式, 因  $|\lambda| < 1$  而趋于零; 故  $B^k \rightarrow 0$ , 任意初值均收敛. 反之, 若对任意初值收敛, 则  $B^k v \rightarrow 0$  对所有  $v$  成立. 取任一特征对  $Bv = \lambda v$ , 得到  $\lambda^k v \rightarrow 0$ , 所以  $|\lambda| < 1$ . 所有特征值均如此, 即  $\rho(B) < 1$ .  $\square$

**定理 4.15** (严格行对角占优下的 Jacobi 与 GS 收敛). 若

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \quad 1 \leq i \leq n,$$

则 Jacobi 与 Gauss-Seidel 迭代对任意初值都收敛。

完整证明. 对 Jacobi, 取  $B_J v = \lambda v$  及满足  $|v_i| = \|v\|_\infty$  的指标  $i$ . 由

$$\lambda a_{ii} v_i = - \sum_{j \neq i} a_{ij} v_j$$

得

$$|\lambda| |a_{ii}| |v_i| \leq \sum_{j \neq i} |a_{ij}| |v_j| \leq \sum_{j \neq i} |a_{ij}| |v_i|,$$

故  $|\lambda| < 1$ 。

对 GS, 设  $B_{GS} v = \lambda v$ . 若  $\lambda = 0$  无须处理; 否则取  $i$  为使  $|v_i| = \|v\|_\infty$  的最小指标. 特征方程的第  $i$  行为

$$\lambda a_{ii} v_i = -\lambda \sum_{j < i} a_{ij} v_j - \sum_{j > i} a_{ij} v_j.$$

假设  $|\lambda| \geq 1$ , 则

$$|a_{ii}| |v_i| \leq \sum_{j < i} |a_{ij}| |v_j| + |\lambda|^{-1} \sum_{j > i} |a_{ij}| |v_j| \leq \sum_{j \neq i} |a_{ij}| |v_i|,$$

与严格对角占优矛盾. 因此所有特征值模均小于一, 再用 [定理 4.14](#) 完成证明.  $\square$

**定理 4.16** (Hermitian 正定情形的 GS 与 SOR). 若  $A = A^* \succ 0$ , 则 Gauss-Seidel 在  $A$ -范数下收敛; SOR 对任意初值收敛当且仅当

$$0 < \omega < 2.$$

完整证明. 先看 GS. 令  $F(x) = \frac{1}{2} x^* A x - \Re(b^* x)$ . 一次 GS 扫描依次在每个坐标方向上精确最小化  $F$ , 故除非已在解处, 每个非零坐标改变量都使  $F$  严格下降. 又

$$F(x) - F(x^*) = \frac{1}{2} \|x - x^*\|_A^2,$$

所以  $A$ -范数误差单调下降。其极限若非零, 则某个坐标残量不为零, 下一次扫描仍会产生固定量的下降, 与能量收敛矛盾; 故误差趋零。

再看 SOR 的必要性。由三角矩阵行列式,

$$\det B_\omega = \frac{\det((1-\omega)D + \omega U)}{\det(D - \omega L)} = (1-\omega)^n.$$

若收敛, 则  $\rho(B_\omega) < 1$ , 从而  $|1-\omega|^n = |\det B_\omega| < 1$ , 即  $0 < \omega < 2$ 。

证充分性。设  $B_\omega z = \lambda z$ , 并记

$$d = z^* D z > 0, \quad a = z^* A z > 0, \quad z^*(U - L)z = i\tau \quad (\tau \in \mathbb{R}).$$

由  $2(D - \omega L) = (2 - \omega)D + \omega A + \omega(U - L)$  及  $2[(1 - \omega)D + \omega U] = (2 - \omega)D - \omega A + \omega(U - L)$ , 左乘  $z^*$ 、右乘  $z$  得

$$\lambda = \frac{(2 - \omega)d - \omega a + i\omega\tau}{(2 - \omega)d + \omega a + i\omega\tau}.$$

当  $0 < \omega < 2$  时, 分母模平方减去分子模平方为

$$4\omega(2 - \omega)ad > 0,$$

所以  $|\lambda| < 1$ 。由谱半径判据, SOR 收敛。取  $\omega = 1$  同时再次得到 GS 收敛。  $\square$

**例题 4.8** (二维 SPD 系统的三种迭代). 对

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix},$$

有

$$B_J = \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}, \quad B_{GS} = \begin{pmatrix} 0 & 1/2 \\ 0 & 1/4 \end{pmatrix}.$$

故  $\rho(B_J) = 1/2$ 、 $\rho(B_{GS}) = 1/4$ 。SOR 可按上式逐分量执行; 只要  $0 < \omega < 2$  即收敛, 但“收敛”并不意味着任意  $\omega$  都比 GS 快。

## 4.20 一致排序与 SOR 最优松弛因子

设  $A = D - L - U$ 。若对每个非零复数  $\alpha$ , 矩阵

$$C(\alpha) = \alpha D^{-1}L + \alpha^{-1}D^{-1}U$$

的特征值与  $\alpha$  无关, 则称  $A$  为一致排序矩阵。具有非零对角元的三对角矩阵是典型情形: 取  $S(\alpha) = \text{diag}(1, \alpha, \dots, \alpha^{n-1})$ , 可验证  $C(\alpha) = S(\alpha)^{-1}C(1)S(\alpha)$ 。这一假设不能由“ $A$  是 SPD”自动推出。

**定理 4.17** (一致排序矩阵的最优 SOR 参数). 设  $A$  一致排序, Jacobi 迭代矩阵  $B_J = D^{-1}(L + U)$

的谱为实数,  $\Lambda = \rho(B_J) < 1$ 。则 SOR 特征值  $\mu$  与某个  $\lambda \in \sigma(B_J)$  满足

$$(\mu + \omega - 1)^2 = \mu\omega^2\lambda^2.$$

在  $0 < \omega < 2$  中, 最小化  $\rho(B_\omega)$  的参数为

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \Lambda^2}}, \quad \rho(B_{\omega_{\text{opt}}}) = \omega_{\text{opt}} - 1 = \frac{1 - \sqrt{1 - \Lambda^2}}{1 + \sqrt{1 - \Lambda^2}}.$$

完整证明. 设  $B_\omega v = \mu v$ , 取平方根  $\eta^2 = \mu$ 。分量方程可重排为

$$[(\mu + \omega - 1)D - \omega\mu L - \omega U]v = 0.$$

除以  $\omega\eta$ , 并令  $\alpha = \eta$ , 得到

$$\left[ \frac{\mu + \omega - 1}{\omega\eta} I - (\eta D^{-1}L + \eta^{-1}D^{-1}U) \right] v = 0.$$

一致排序使括号内第二项与  $B_J = C(1)$  同谱, 故存在  $\lambda \in \sigma(B_J)$  使

$$\frac{\mu + \omega - 1}{\omega\sqrt{\mu}} = \lambda,$$

平方即得特征值关系;  $\mu = 0$  的情形由连续性包含。

对固定  $|\lambda| \leq \Lambda$ , 令  $r = \sqrt{|\mu|}$ 。当判别式  $\omega^2\Lambda^2 - 4(\omega - 1) \geq 0$  时, 最坏根的模随  $|\lambda|$  增大, 且

$$\rho(B_\omega) = \left( \frac{\omega\Lambda + \sqrt{\omega^2\Lambda^2 - 4(\omega - 1)}}{2} \right)^2.$$

当判别式为负时, 共轭根乘积为  $(\omega - 1)^2$ , 故  $\rho(B_\omega) = \omega - 1$ 。前一支随  $\omega$  下降, 后一支随  $\omega$  上升, 最小值在两支交界  $\omega^2\Lambda^2 = 4(\omega - 1)$  处取得。解出位于  $(1, 2)$  的根即为定理中的  $\omega_{\text{opt}}$ , 代回得到最优谱半径。□

**例题 4.9** (手算最优松弛因子). 对 [例题 4.8](#) 中的三对角 SPD 矩阵,  $\Lambda = \rho(B_J) = 1/2$ , 所以

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{3}/2} = 8 - 4\sqrt{3} \approx 1.0718, \quad \rho(B_{\omega_{\text{opt}}}) = 7 - 4\sqrt{3} \approx 0.0718.$$

若矩阵不满足一致排序, 或  $B_J$  具有非实谱, 就不能直接套用该公式; 此时应从实际  $B_\omega$  的谱或可验证界选择参数。

## 4.21 CG 的 Galerkin 条件与多项式误差

对 SPD 矩阵  $A$ , 二次函数

$$\phi(x) = \frac{1}{2}x^\top Ax - b^\top x$$

的最优点解  $Ax = b$ 。CG 在仿射 Krylov 空间

$$x_0 + \mathcal{K}_k(A, r_0), \quad \mathcal{K}_k = \text{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\}$$

中最小化  $A$ -范数误差。Galerkin 条件为

$$r_k \perp \mathcal{K}_k.$$

因此

$$e_k = p_k(A)e_0, \quad p_k(0) = 1,$$

且

$$\|e_k\|_A \leq \min_{\substack{p \in \mathbb{P}_k \\ p(0)=1}} \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |p(\lambda)| \|e_0\|_A.$$

选取缩放 Chebyshev 多项式得到

$$\|e_k\|_A \leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|e_0\|_A.$$

这个界依赖 SPD 与  $A$ -范数；非对称矩阵不能直接使用 CG。

## 4.22 前处理与 GMRES

前处理选择易解的  $M \approx A$ ，把系统变为

$$M^{-1}Ax = M^{-1}b$$

或对称前处理形式。目标不是使  $\|M - A\|$  单独最小，而是让预处理算子的谱/非正规结构更利于 Krylov 收敛，同时保持每次应用  $M^{-1}$  便宜。

GMRES 适用于一般非奇异矩阵，在  $x_0 + \mathcal{K}_k(A, r_0)$  中最小化二范数残量：

$$x_k = \arg \min_{x \in x_0 + \mathcal{K}_k} \|b - Ax\|_2.$$

Arnoldi 正交化构造

$$AQ_k = Q_{k+1}\overline{H}_k,$$

把大规模问题化为小型最小二乘。存储与正交化成本随  $k$  增长，restart 可控成本却可能破坏原本的收敛行为。

## 4.23 特征值算法的位移与商

幂法收敛因子约为  $|\lambda_2/\lambda_1|$ 。反幂法对  $(A - \sigma I)^{-1}$  用幂法，收敛到离  $\sigma$  最近的特征值。Rayleigh 商

$$\rho(x) = \frac{x^\top Ax}{x^\top x}$$

对称情形下是单位球约束优化的一阶驻点条件。Rayleigh quotient iteration 对称矩阵在局部常具三次收敛，但每步需解位移线性系统。

QR 迭代

$$A_k = Q_k R_k, \quad A_{k+1} = R_k Q_k = Q_k^\top A_k Q_k$$

保持相似性。实际算法先 Hessenberg 化、使用位移与 deflation；“每步直接做完整 QR”只是概念版本。

### 答题方法

线性代数算法题先判断矩阵结构：SPD、对称不定、一般方阵、长方或秩亏。结构决定可用分解、范数和收敛理论；然后再比较条件性、后向稳定性、每步成本与存储。

### 自测

1. 构造谱半径小于一但短期幂范数很大的非正规矩阵。
2. 为什么正规方程会平方二范数条件数？
3. 证明定常迭代对任意初值收敛当且仅当  $\rho(G) < 1$ 。
4. CG 的误差多项式为什么满足  $p(0) = 1$ ？
5. GMRES 与 CG 各自最小化什么量，适用什么矩阵？

## 第五章 数值常微分方程

### 学习目标

本章从初值问题适定性出发，建立一步法与多步法的局部误差、稳定、收敛和绝对稳定语言。线性多步部分完整保留 AB2 稳定区间、根条件和隐式误差吸收证明。

### 5.1 IVP 适定性与连续依赖

考虑

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0.$$

若  $f$  在矩形邻域连续并对  $y$  一致 Lipschitz，则 Picard 迭代在足够短区间给局部存在唯一。若两个解满足初值差  $\delta_0$ ，则

$$\|y(t) - z(t)\| \leq e^{L(t-t_0)} \|\delta_0\|.$$

证明由积分形式和 Gronwall 不等式得到。这个指数常数是问题条件性的一部分；数值方法的稳定常数应与它相容，而不是要求误差不增长。

**定理 5.1** (连续 Gronwall). 若非负连续函数  $u$  满足  $u(t) \leq a + \int_{t_0}^t b(s)u(s) ds$ ，其中  $b \geq 0$  可积，则

$$u(t) \leq a \exp\left(\int_{t_0}^t b(s) ds\right).$$

完整证明. 令  $v(t) = a + \int_{t_0}^t b(s)u(s) ds$ ，则  $u \leq v$  且  $v' \leq bv$ 。当  $a > 0$  时积分  $(\log v)' \leq b$  得结论； $a = 0$  可先用  $a + \varepsilon$  再令  $\varepsilon \downarrow 0$ 。□

### 5.2 一步法：一致性、稳定性与整体误差

一般一步法写成

$$y_{n+1} = y_n + h\Phi(t_n, y_n, h).$$

本书把未除以  $h$  的局部 defect 定义为

$$d_{n+1} = y(t_{n+1}) - y(t_n) - h\Phi(t_n, y(t_n), h).$$

若  $d_{n+1} = O(h^{p+1})$ ，称局部阶  $p+1$ ；若固定区间上  $\max_n \|y(t_n) - y_n\| = O(h^p)$ ，称整体  $p$  阶。另一种常用约定把  $d_{n+1}/h$  称局部截断误差，阶会少 1；比较结论前必须统一约定。

**定理 5.2** (一步法的离散 Gronwall 收敛). 若增量函数对  $y$  一致 Lipschitz:  $\|\Phi(t, y, h) - \Phi(t, z, h)\| \leq L\|y - z\|$ ，且  $\|d_{n+1}\| \leq Ch^{p+1}$ ，初值误差为  $O(h^p)$ ，则在  $nh \leq T$  上整体误差为  $O(h^p)$ 。

完整证明. 误差满足

$$\|e_{n+1}\| \leq (1 + hL)\|e_n\| + Ch^{p+1}.$$

迭代并求几何和:

$$\|e_n\| \leq e^{LT}\|e_0\| + \frac{C}{L}(e^{LT} - 1)h^p$$

( $L = 0$  时直接求和)。常数与  $h$  无关, 得到固定时间区间上的阶。□

$\theta$  法统一为

$$y_{n+1} = y_n + h[(1 - \theta)f(t_n, y_n) + \theta f(t_{n+1}, y_{n+1})].$$

$\theta = 0, 1/2, 1$  分别是显式 Euler、梯形、后向 Euler; 除  $\theta = 1/2$  外通常一阶。预测-校正可先用显式 Euler 预测, 再代入梯形右端校正; 若只校正有限次, 它仍是显式实现, 不等于精确求解隐式方程。

**例题 5.1** (局部与整体阶的约定). 对  $y' = y$ , Euler 一步精确值为  $e^h y_n = (1 + h + h^2/2 + O(h^3))y_n$ , 故未除  $h$  的 defect 为  $h^2 y_n/2 + O(h^3)$ , 局部二阶而整体一阶。若定义截断误差为 defect/ $h$ , 则称它一阶; 两说并不矛盾。

### 5.3 Runge-Kutta 阶条件与稳定函数

把精确解 Taylor 展开

$$y(t+h) = y + hf + \frac{h^2}{2}(f_t + f_y f) + O(h^3).$$

二阶段显式 RK

$$K_1 = f(t_n, y_n), \quad K_2 = f(t_n + c_2 h, y_n + a_{21} h K_1), \quad y_{n+1} = y_n + h(b_1 K_1 + b_2 K_2)$$

达到二阶的条件为  $b_1 + b_2 = 1$ ,  $b_2 c_2 = 1/2$ ,  $b_2 a_{21} = 1/2$ , 且显式一致性通常取  $c_2 = a_{21}$ 。Heun 取  $a_{21} = c_2 = 1, b_1 = b_2 = 1/2$ ; 显式中点取  $a_{21} = c_2 = 1/2, b_1 = 0, b_2 = 1$ 。

经典 RK4 的 Butcher 表为

|               |               |               |               |               |
|---------------|---------------|---------------|---------------|---------------|
| 0             |               |               |               |               |
| $\frac{1}{2}$ | $\frac{1}{2}$ |               |               |               |
| $\frac{1}{2}$ | 0             | $\frac{1}{2}$ |               |               |
| 1             | 0             | 0             | 1             |               |
| <hr/>         |               |               |               |               |
|               | $\frac{1}{6}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{6}$ |

每步四次函数评价, 整体四阶。显式  $s$  级 RK 的稳定函数是次数至多  $s$  的多项式, 因此当  $z \rightarrow -\infty$  时不可能保持  $|R(z)| \leq 1$ , 故不存在显式 A-稳定 RK。

**例题 5.2** (由 Butcher 表判阶). 对 Heun 法,  $\sum b_i = 1$ ,  $\sum b_i c_i = 1/2$ , 故至少二阶; 测试方程给  $R(z) = 1 + z + z^2/2$ 。负实轴稳定边界解  $|1 + z + z^2/2| \leq 1$ , 得到  $-2 \leq z \leq 0$ 。阶条件与稳定域是两套问题, 不能互相替代。

嵌入式 RK 用同一组阶段构造两个不同阶近似, 其差估计局部误差并调节

$$h_{\text{new}} \approx \eta h \left( \frac{\text{tol}}{\text{err}} \right)^{1/(p+1)}.$$

接受/拒绝、上下限和安全因子属于实现要点。考试应会解释误差指数  $1/(p+1)$  的来源, 通常不要求背某一特定嵌入式 RK 对的全部系数。

## 5.4 为什么需要多步法

初值问题

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0$$

的一步法只利用  $y_n$  产生  $y_{n+1}$ 。已经计算出的  $f_n = f(t_n, y_n)$  若全部丢弃, 显得浪费。线性多步法 (linear multistep method, LMM) 把若干个历史函数值组合起来, 以一次新的函数评估换取较高阶。代价是: 方法具有“记忆”, 启动值、寄生根和零稳定性都成为不可回避的问题。

**定义 5.1** (线性多步法). 在等距网格  $t_n = t_0 + nh$  上,  $k$  步线性多步法写成

$$\sum_{j=0}^k \alpha_j y_{n+j} = h \sum_{j=0}^k \beta_j f_{n+j}, \quad \alpha_k \neq 0. \quad (5.1)$$

若  $\beta_k = 0$  则为显式法, 否则一般为隐式法。定义特征多项式

$$\rho(\zeta) = \sum_{j=0}^k \alpha_j \zeta^j, \quad \sigma(\zeta) = \sum_{j=0}^k \beta_j \zeta^j.$$

系数整体乘非零常数不改变方法, 故常取  $\alpha_k = 1$ 。一个  $k$  步法需要  $y_0, \dots, y_{k-1}$ ; 这些启动值必须由一步法或精确数据提供, 而且其误差阶不能低到破坏总体精度。

### 易错点

“ $k$  步”不是“ $k$  阶”。步数由历史点个数决定, 阶由局部截断误差决定。例如 BDF2 是二步二阶, 而 Adams–Bashforth 三步法是三阶。

## 5.5 从插值与积分构造 Adams 方法

### 5.5.1 共同出发点

精确解满足

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t, y(t)) dt. \quad (5.2)$$

用历史节点处的  $f$  的插值多项式替代被积函数, 就得到 Adams 族。令  $s = (t - t_n)/h$ , 则积分区间变为  $[0, 1]$ , 系数只是 Lagrange 基函数的积分。

### 5.5.2 Adams–Bashforth: 外推产生显式法

二步情形用  $s = 0, -1$  的数据插值:

$$p(s) = f_n \frac{s+1}{1} + f_{n-1} \frac{s}{-1} = (s+1)f_n - sf_{n-1}.$$

代入式 (5.2) 得

$$y_{n+1} = y_n + h \int_0^1 p(s) ds = y_n + h \left( \frac{3}{2}f_n - \frac{1}{2}f_{n-1} \right).$$

这就是 AB2。一般的  $k$  步 Adams–Bashforth 使用  $f_n, f_{n-1}, \dots, f_{n-k+1}$ , 是  $k$  阶显式法。

**例题 5.3** (AB3 系数的完整计算). 在节点  $s = 0, -1, -2$  上,

$$\ell_0(s) = \frac{(s+1)(s+2)}{2}, \quad \ell_1(s) = -s(s+2), \quad \ell_2(s) = \frac{s(s+1)}{2}.$$

逐项积分:

$$\int_0^1 \ell_0 = \frac{23}{12}, \quad \int_0^1 \ell_1 = -\frac{16}{12}, \quad \int_0^1 \ell_2 = \frac{5}{12}.$$

故

$$y_{n+1} = y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2}).$$

检查常函数: 权重之和  $(23 - 16 + 5)/12 = 1$ , 这是最便宜的自检。

### 5.5.3 Adams–Moulton: 加入新点产生隐式法

若插值节点包括  $s = 1$  的  $f_{n+1}$ , 便得到 Adams–Moulton 法。以  $s = 1, 0$  为节点,

$$p(s) = sf_{n+1} + (1-s)f_n,$$

所以

$$y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n),$$

即梯形法。若用  $s = 1, 0, -1$ , 可得

$$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}),$$

它是三阶法。隐式方程通常用 Newton 或不动点迭代求解; “公式阶数高” 不代表每步代价低。

## 5.6 BDF: 对解插值后再求导

Backward Differentiation Formula 不插值  $f$ , 而是用  $y_{n+1}, y_n, \dots$  插值得到  $p(t)$ , 令  $p'(t_{n+1}) = f_{n+1}$ 。

**例题 5.4** (BDF2 的来源). 令  $s = (t - t_{n+1})/h$ , 三个节点  $t_{n+1}, t_n, t_{n-1}$  对应  $s = 0, -1, -2$ 。Newton 后向插值写成

$$p(s) = y_{n+1} + s \nabla y_{n+1} \frac{s(s+1)}{2} \nabla^2 y_{n+1}.$$

在  $s = 0$  求导并除以  $h$ :

$$p'(t_{n+1}) = \frac{1}{h} \left( \nabla y_{n+1} + \frac{1}{2} \nabla^2 y_{n+1} \right) = \frac{3y_{n+1} - 4y_n + y_{n-1}}{2h}.$$

因此

$$3y_{n+1} - 4y_n + y_{n-1} = 2hf_{n+1}.$$

这里  $\rho(\zeta) = 3\zeta^2 - 4\zeta + 1$ ,  $\sigma(\zeta) = 2\zeta^2$ 。整体除以 3 不影响后续根条件。

BDF 的优势是对刚性问题具有较好的稳定性; 但它不是任意高阶都零稳定。

经典结论是 BDF1–BDF6 零稳定, 六阶以后根条件失败。本章不把该事实当作“背诵替代证明”: 给定具体系数时仍应直接检查  $\rho$  的根。

### 5.6.1 预测—校正并不等于一个隐式法的精确求解

Adams–Bashforth 与 Adams–Moulton 常配成 PECE 流程。例如先用 AB2 预测

$$y_{n+1}^{(0)} = y_n + h \left( \frac{3}{2} f_n - \frac{1}{2} f_{n-1} \right),$$

再用梯形法校正一次:

$$y_{n+1}^{(1)} = y_n + \frac{h}{2} \left[ f(t_{n+1}, y_{n+1}^{(0)}) + f_n \right].$$

这里 P、E、C、E 分别表示 predict、evaluate、correct、evaluate。若只校正有限次, 最终算法是一个显式复合方法, 并不具有“精确解出梯形法隐式方程”的全部稳定性质。

以测试方程  $y' = \lambda y$ 、 $z = h\lambda$  为例, 一次校正给

$$\begin{aligned} y_{n+1}^{(0)} &= \left( 1 + \frac{3}{2}z \right) y_n - \frac{1}{2}zy_{n-1}, \\ y_{n+1}^{(1)} &= y_n + \frac{z}{2} \left( y_{n+1}^{(0)} + y_n \right) \\ &= \left( 1 + z + \frac{3}{4}z^2 \right) y_n - \frac{1}{4}z^2 y_{n-1}. \end{aligned}$$

所以应分析的稳定多项式是

$$\zeta^2 - \left( 1 + z + \frac{3}{4}z^2 \right) \zeta + \frac{1}{4}z^2,$$

而不是梯形法的一步稳定函数。若反复校正到迭代收敛, 且非线性求解误差小于离散误差, 才近似恢复隐式 Adams–Moulton 法。

#### 易错点

题目写“predictor–corrector”时, 必须确认校正次数、每次使用旧值还是新值, 以及是否迭代到收敛。三种实现可能有同样的形式阶, 却有不同的稳定多项式和每步代价。

## 5.7 局部截断误差与阶条件

**定义 5.2** (差分算子与局部截断误差). 对光滑函数  $v$ , 定义

$$L_h[v](t_n) = \sum_{j=0}^k \alpha_j v(t_n + jh) - h \sum_{j=0}^k \beta_j v'(t_n + jh).$$

若把精确解代入 式 (5.1),  $L_h[y](t_n)$  是未除以  $h$  的局部残量。本章称  $\tau_{n+k} = L_h[y](t_n)/h$  为局部截断误差。

在  $t_n$  展开:

$$L_h[v](t_n) = C_0 v(t_n) + C_1 h v'(t_n) + C_2 h^2 v''(t_n) + \cdots,$$

其中

$$C_0 = \sum_{j=0}^k \alpha_j,$$

$$C_q = \frac{1}{q!} \sum_{j=0}^k j^q \alpha_j - \frac{1}{(q-1)!} \sum_{j=0}^k j^{q-1} \beta_j, \quad q \geq 1.$$

**定义 5.3** (阶). 若  $C_0 = C_1 = \cdots = C_p = 0$  且  $C_{p+1} \neq 0$ , 称方法具有  $p$  阶; 此时  $L_h[y] = \mathcal{O}(h^{p+1})$ , 而  $\tau = \mathcal{O}(h^p)$ 。

**命题 5.3** (相容性的三种等价写法). 线性多步法至少一阶相容, 当且仅当

$$\sum \alpha_j = 0, \quad \sum j \alpha_j = \sum \beta_j,$$

也等价于

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1).$$

*证明.* 前两式正是  $C_0 = C_1 = 0$ 。而  $\rho(1) = \sum \alpha_j$ ,  $\rho'(1) = \sum j \alpha_j$ ,  $\sigma(1) = \sum \beta_j$ , 故三种写法等价。□

**例题 5.5** (判阶与误差常数). 对 BDF2 取

$$\alpha = (1, -4, 3), \quad \beta = (0, 0, 2)$$

(索引按  $y_{n-1}, y_n, y_{n+1}$  平移为  $j = 0, 1, 2$ )。于是

$$C_0 = 0, \quad C_1 = (-4 + 6) - 2 = 0,$$

$$C_2 = \frac{-4 + 12}{2} - 4 = 0, \quad C_3 = \frac{-4 + 24}{6} - 4 = -\frac{2}{3} \neq 0.$$

所以是二阶法, 未除以  $h$  的局部残量主项为  $-\frac{2}{3}h^3 y^{(3)}$ 。若先把方法整体除以 3, 误差常数也同步除以 3; 阶数不变。

## 易错点

局部残量  $L_h[y]$  与  $\tau = L_h[y]/h$  相差一个  $h$ 。不同教材可能把二者都称 LTE。答题时必须先声明约定，否则“二阶法的局部误差是  $\mathcal{O}(h^3)$  还是  $\mathcal{O}(h^2)$ ”会出现表面冲突。

## 5.7.1 用指数函数压缩全部阶条件

Taylor 系数逐项计算最稳妥，但还可用一个生成函数核对。对  $v(t) = e^{xt/h}$ ，平移一个网格点等于乘  $e^x$ ，所以差分算子的符号为

$$\rho(e^x) - x\sigma(e^x).$$

方法为  $p$  阶，当且仅当

$$\rho(e^x) - x\sigma(e^x) = \mathcal{O}(x^{p+1}), \quad x \rightarrow 0. \quad (5.3)$$

例如梯形法的  $\rho(\zeta) = \zeta - 1$ 、 $\sigma(\zeta) = (\zeta + 1)/2$ ，故

$$e^x - 1 - \frac{x}{2}(e^x + 1) = -\frac{x^3}{12} + \mathcal{O}(x^4),$$

立即看出它为二阶。这个写法也说明相容性为什么恰是  $\rho(1) = 0, \rho'(1) = \sigma(1)$ ：它们分别消掉式 (5.3) 的常数项和一次项。

注 5.1 (索引平移不改变阶). 把同一个公式从  $y_{n-k}, \dots, y_n$  改写为  $y_n, \dots, y_{n+k}$  会使  $\rho, \sigma$  同时乘一个  $\zeta^r$ 。这不改变非零根及阶，但 Taylor 公式中的  $j$  必须跟着新索引重算。最常见的失分正是“系数来自一种索引， $j^q$  却来自另一种索引”。

## 5.8 零稳定与根条件

## 5.8.1 局部小误差为何可能被放大

把右端暂时置零，误差满足

$$\sum_{j=0}^k \alpha_j e_{n+j} = 0.$$

若  $\rho(\zeta) = 0$  有根  $\zeta$ ，则  $\zeta^n$  是齐次模态；若  $\zeta$  是  $r$  重根，还会出现  $n\zeta^n, \dots, n^{r-1}\zeta^n$ 。因此：

**定义 5.4** (根条件).  $\rho$  的所有根满足  $|\zeta| \leq 1$ ，且单位圆上的根均为单根。

**定义 5.5** (零稳定). 在固定时间区间  $0 \leq nh \leq T$  内，若初始扰动与每步强迫扰动的影响有与  $h$  无关的统一界，则方法称零稳定。等价地，齐次伴随矩阵的幂  $A^n$  在所有  $n \geq 0$  上一致有界。

**定理 5.4** (根条件等价于零稳定). 线性多步法零稳定，当且仅当  $\rho$  满足根条件。

**证明.** 将标量  $k$  阶递推写成一阶伴随系统  $E_{n+1} = AE_n$ 。  $A$  的特征值恰是  $\rho$  的根。若存在  $|\zeta| > 1$ ，则相应模态指数增长，故  $\sup_n \|A^n\| = \infty$ 。若单位圆上有重根，则 Jordan 块产生  $n^{r-1}\zeta^n$ ，同样无界。

反之，若  $|\zeta| < 1$  的 Jordan 块大小任意，则  $n^{r-1}|\zeta|^n$  有界；单位圆上的根均为单根，只给出  $\zeta^n$ ，模长为一。把有限个 Jordan 块合并，得到  $\sup_n \|A^n\| < \infty$ ，即零稳定。  $\square$

**例题 5.6** (相容但不零稳定). 方法

$$y_{n+2} - 2y_{n+1} + y_n = h(f_{n+1} - f_n)$$

满足  $\rho(1) = 0$  且  $\rho'(1) = \sigma(1) = 0$ , 因而相容; 但  $\rho(\zeta) = (\zeta - 1)^2$ , 单位根是二重根. 即使  $f = 0$ , 误差也含  $c_1 + c_2 n$ , 故不零稳定, 更不可能一般收敛。

### 易错点

单位圆上的重根会造成多项式增长, 不能因为“模等于 1”就判稳定。单位圆内的重根可以接受, 因为指数衰减压过多项式因子。

## 5.9 Dahlquist 等价定理: 相容加稳定等于收敛

**定义 5.6** (收敛). 设精确启动值被一致逼近。若对每个固定  $T$ , 当  $h \rightarrow 0$ 、 $nh \leq T$  时

$$\max_{0 \leq j \leq n} |y_j - y(t_j)| \rightarrow 0,$$

则方法收敛。

**引理 5.5** (零稳定给出离散扰动估计). 若方法零稳定, 且

$$\sum_{j=0}^k \alpha_j e_{n+j} = h g_n,$$

则存在与  $h, n$  无关的常数  $C_T$ , 使  $nh \leq T$  时

$$\max_{0 \leq r \leq n+k} |e_r| \leq C_T \left( \max_{0 \leq r < k} |e_r| + h \sum_{\ell=0}^n |g_\ell| \right).$$

**证明.** 写成伴随系统

$$E_{n+1} = A E_n + h b g_n.$$

变参数公式给出

$$E_n = A^n E_0 + h \sum_{\ell=0}^{n-1} A^{n-1-\ell} b g_\ell.$$

由零稳定  $\sup_m \|A^m\| \leq M$ , 取范数即得结论。□

**定理 5.6** (Dahlquist 等价定理). 对满足常规光滑性与 Lipschitz 条件的初值问题, 线性多步法收敛, 当且仅当它相容且零稳定。

**证明. 充分性.** 取  $N$  使  $Nh \leq T$ , 令

$$e_n = y_n - y(t_n), \quad E_{\text{start}} = \max_{0 \leq j < k} |e_j|, \quad M_n = \max_{0 \leq r \leq n+k} |e_r| \quad (0 \leq n \leq N - k).$$

记精确解的未除以  $h$  的残量为  $\tau_n = L_h[y](t_n)$ , 并令

$$\eta_h = \max_{0 \leq n \leq N-k} \frac{|\tau_n|}{h}.$$

相容性给  $\eta_h \rightarrow 0$ . 数值格式减去精确解的残量, 得到

$$\sum_{j=0}^k \alpha_j e_{n+j} = h \sum_{j=0}^k \beta_j [f(t_{n+j}, y_{n+j}) - f(t_{n+j}, y(t_{n+j}))] - \tau_n.$$

把右侧除以  $h$  后应用 引理 5.5. 若  $B = \sum_{j=0}^k |\beta_j|$ , Lipschitz 条件给

$$|f(t_{n+j}, y_{n+j}) - f(t_{n+j}, y(t_{n+j}))| \leq L|e_{n+j}|.$$

这里不能把所有函数差都当作已知强迫: 当方法隐式, 即  $\beta_k \neq 0$  时,  $\ell = n, j = k$  这一项含有当前最高层误差  $|e_{n+k}|$ , 因而含有  $M_n$ . 将该项单独留下, 把  $\ell < n$  的项按相应  $M_\ell$  控制, 并把  $\ell = n, j < k$  的项并入前一层最大值, 可得常数  $C, C_1$  (只依赖方法系数、 $L, k, T$ ) 使

$$M_n \leq C \left[ E_{\text{start}} + (T+h)\eta_h + hC_1 \sum_{\ell=0}^{n-1} M_\ell + hL|\beta_k|M_n \right]. \quad (5.4)$$

空和按零解释;  $n=0$  时较低层项已经包含在  $E_{\text{start}}$  中. 若  $\beta_k = 0$ , 最后一项不存在; 若  $\beta_k \neq 0$ , 取  $h$  足够小使  $ChL|\beta_k| \leq 1/2$ , 把 式 (5.4) 中的当前最高层项吸收到左端. 于是

$$M_n \leq 2C \left[ E_{\text{start}} + (T+h)\eta_h + hC_1 \sum_{\ell=0}^{n-1} M_\ell \right].$$

离散 Grönwall 不等式给出统一估计

$$\max_{0 \leq r \leq N} |e_r| \leq C_T (E_{\text{start}} + \eta_h) \rightarrow 0.$$

这里  $C_T$  可以依赖  $T, L, k, \alpha, \beta$ , 但不依赖  $h, N, n_0$ . 这统一了误差向量、索引范围和常数依赖. 若方法为  $p$  阶且启动误差为  $\mathcal{O}(h^p)$ , 同一估计给出全局误差  $\mathcal{O}(h^p)$ .

**必要性: 相容.** 若方法对所有光滑问题收敛, 先取精确解  $y(t) \equiv 1$ , 代入极限必须有  $\sum \alpha_j = 0$ . 再取  $y(t) = t, f \equiv 1$ , 除以  $h$  后取极限, 得到  $\sum j\alpha_j = \sum \beta_j$ . 故方法相容.

**必要性: 零稳定.** 若根条件失败, 则存在  $|\zeta| > 1$  的指数增长模态, 或单位圆重根产生的多项式增长模态. 给启动值施加趋于零、但沿该模态的扰动; 选择步数  $n \asymp T/h$ , 可使终点扰动不趋零, 甚至无界. 这与收敛要求对启动扰动连续依赖矛盾, 故必须零稳定.  $\square$

### 本节结论

证明时不要只写口号 “consistency + stability = convergence”. 可得分的链条是: 精确解残量  $\rightarrow$  误差递推  $\rightarrow$  根条件给卷积核有界  $\rightarrow$  Lipschitz 项和离散 Grönwall  $\rightarrow$  全局误差.

### 5.9.1 启动值与全局阶

等价定理只给“是否收敛”，精度还取决于启动。若方法为  $p$  阶，则  $\tau_n = \mathcal{O}(h^p)$ 。扰动估计中的累计项为

$$h \sum_{n \leq T/h} \mathcal{O}(h^p) = \mathcal{O}(h^p).$$

因此，在零稳定前提下，只要

$$\max_{0 \leq j < k} |y_j - y(t_j)| = \mathcal{O}(h^p),$$

全局误差也是  $\mathcal{O}(h^p)$ 。如果启动值只有  $\mathcal{O}(h^q)$ 、 $q < p$ ，一般只能保证  $\mathcal{O}(h^q)$ 。

实践中不必为了生成  $k-1$  个启动值永久使用昂贵高阶方法：只需在最初几步用阶数足够的一步 Runge-Kutta 法，之后切换到 LMM。若使用变步长，固定系数公式不再成立，零稳定分析也需重新做。变步长 LMM 不属于当前考纲主线；本书只保留这一边界说明，不把固定步长结论外推到变步长情形。

## 5.10 绝对稳定：测试方程上的另一组根

### 5.10.1 稳定多项式

对测试方程  $y' = \lambda y$ ，令  $z = h\lambda$ 。代入 LMM 得

$$\sum_{j=0}^k (\alpha_j - z\beta_j) y_{n+j} = 0.$$

对应稳定多项式是

$$\pi(\zeta; z) = \rho(\zeta) - z\sigma(\zeta). \quad (5.5)$$

**定义 5.7** (绝对稳定域). 若式 (5.5) 的全部根满足  $|\zeta| \leq 1$ ，且单位圆上的根为单根，则称该  $z$  对方法绝对稳定。所有这类  $z$  构成稳定域  $\mathcal{S}$ 。

稳定域边界常由

$$z(\theta) = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}, \quad 0 \leq \theta \leq 2\pi$$

参数化，但画完边界还必须测试哪一侧是稳定侧。

**例题 5.7** (AB2 在负实轴上的稳定区间). AB2 写为

$$y_{n+2} - y_{n+1} = z \left( \frac{3}{2} y_{n+1} - \frac{1}{2} y_n \right),$$

故

$$\pi(\zeta; z) = \zeta^2 - \zeta - z \left( \frac{3}{2} \zeta - \frac{1}{2} \right).$$

负实轴边界先检查  $\zeta = 1$  得  $z = 0$ ，再检查  $\zeta = -1$ ：

$$2 - z(-2) = 0 \implies z = -1.$$

对实  $z$ , 多项式可写为

$$\zeta^2 - \left(1 + \frac{3}{2}z\right)\zeta + \frac{z}{2}.$$

二次多项式的 Schur 条件在这里化为

$$\left|\frac{z}{2}\right| < 1, \quad 1 - \left(1 + \frac{3}{2}z\right) + \frac{z}{2} = -z > 0, \quad 1 + \left(1 + \frac{3}{2}z\right) + \frac{z}{2} = 2 + 2z > 0.$$

所以  $-1 < z < 0$  时两根严格位于单位圆内。端点按本章同一个根条件检查:

$$z = -1: \quad \pi(\zeta; -1) = (\zeta + 1)\left(\zeta - \frac{1}{2}\right),$$

$$z = 0: \quad \pi(\zeta; 0) = \zeta(\zeta - 1).$$

两处的单位圆根分别为  $-1$  和  $1$ , 且都是单根; 另一根严格在单位圆内。因此按式 (5.5) 后的绝对稳定定义, AB2 在负实轴上的稳定区间是

$$[-1, 0].$$

**例题 5.8** (BDF2 的 A-稳定性). 采用归一化形式

$$\frac{3}{2}y_{n+2} - 2y_{n+1} + \frac{1}{2}y_n = hf_{n+2}.$$

稳定边界由  $|\zeta| = 1$ ,  $\zeta = e^{i\theta}$  给出:

$$z(\theta) = \frac{\rho(\zeta)}{\sigma(\zeta)} = \frac{3}{2} - 2\zeta^{-1} + \frac{1}{2}\zeta^{-2}.$$

其实部为

$$\Re z(\theta) = \frac{3}{2} - 2\cos\theta + \frac{1}{2}\cos 2\theta = (1 - \cos\theta)^2 \geq 0.$$

因此根穿越单位圆的边界位于闭右半平面。取左半平面的测试点, 例如负实轴上充分接近 0 的  $z$ , 由零稳定性和主根连续性知所有根都在单位圆内; 根若要离开, 必须穿过上述边界, 但左半平面没有这样的边界点。故整个左半平面包含于稳定域, BDF2 是 A-稳定的。

注意这并不等于 L-稳定: 还需检查刚性极限中主放大模态是否趋于零。对于多步法, 这个问题仍须通过所有稳定根而非单个  $R(z)$  表述。

### 5.10.2 一步法稳定与多步法稳定的区别

一步法常写成  $y_{n+1} = R(z)y_n$ , 只需检查  $|R(z)| \leq 1$ 。多步法有  $k$  个特征根:

- 与  $z = 0$  时  $\zeta = 1$  连续相接的根称主根;
- 其他根是寄生根, 也必须全部受控;
- 只计算某一个“放大因子”会漏掉寄生模态;
- 边界根若重合, 会出现 Jordan 型多项式增长。

易错点

零稳定检查  $\rho(\zeta) = 0$ ，与步长无关；绝对稳定检查  $\rho(\zeta) - z\sigma(\zeta) = 0$ ，依赖  $z = h\lambda$ 。前者保证  $h \rightarrow 0$  时误差不会由寄生根失控，后者决定给定步长能否衰减测试方程。

5.10.3 边界曲线法的完整使用条件

令  $\zeta = e^{i\theta}$  得到的曲线只记录“某个根位于单位圆上”，它本身不告诉哪一侧稳定，也可能出现自交。可靠流程是：

- 1. 先确认  $z = 0$  时  $\rho$  满足根条件；
- 2. 参数化  $z(\theta) = \rho(e^{i\theta})/\sigma(e^{i\theta})$ ，处理  $\sigma(e^{i\theta}) = 0$  的极点；
- 3. 把复平面按边界分区；
- 4. 每个连通区域选一个测试点，直接求  $\rho(\zeta) - z\sigma(\zeta)$  的全部根；
- 5. 若边界上出现重根，同时检查  $\pi(\zeta; z) = 0$  与  $\partial_\zeta \pi(\zeta; z) = 0$ 。

对实系数方法，边界关于实轴对称，可以只画上半部再镜像。若题目只问负实轴区间，则通常先令  $\zeta = \pm 1$  找端点，再取区间内一点验证，正如 例题 5.7；但这只是针对该截面的简化，不是一般二维稳定域证明。

5.11 方法比较、刚性与考试母题

| 方法族             | 优点                    | 代价与限制              |
|-----------------|-----------------------|--------------------|
| Adams–Bashforth | 显式、每步一次新函数评估、构造直观     | 稳定域较小，不适合强刚性       |
| Adams–Moulton   | 同等历史长度下阶较高，稳定性通常优于 AB | 每步需解隐式方程           |
| BDF             | 对衰减刚性模态稳健，公式只含新的 $f$  | 高阶稳定性受限；BDF7 起不零稳定 |

刚性不是“解变化很快”的同义词。典型刚性系统含时间尺度差异很大的衰减模态；显式法为压住快速模态被迫取极小步长，即使所关心的慢模态很平滑。A-稳定要求整个左半平面包含在  $\mathcal{S}$ 。Dahlquist 第二屏障说明显式 LMM 不可能 A-稳定，A-稳定 LMM 的阶不超过 2。本章把这些作为方法选择的边界，而不替代对具体稳定多项式的计算。

考试级结论

给定二步法

$$y_{n+2} - (1 + a)y_{n+1} + ay_n = h(bf_{n+2} + cf_{n+1} + df_n).$$

- 1. 写出相容性条件；
- 2. 讨论零稳定的  $a$  范围；
- 3. 在相容条件下写出二阶条件；

4. 给出测试方程上的稳定多项式。

解。

$$\rho(\zeta) = \zeta^2 - (1+a)\zeta + a = (\zeta-1)(\zeta-a), \quad \sigma(\zeta) = b\zeta^2 + c\zeta + d.$$

$\rho(1) = 0$  自动成立。相容性第二条件为

$$\rho'(1) = 1 - a = \sigma(1) = b + c + d.$$

根为  $1, a$ , 故零稳定当且仅当  $|a| \leq 1$  且  $a \neq 1$ ; 当  $a = 1$  时单位根  $1$  为二重根。二阶条件  $C_2 = 0$  给出

$$\frac{4-1-a}{2} - (2b+c) = 0, \quad \text{即} \quad 2b+c = \frac{3-a}{2}.$$

测试方程上的稳定多项式为

$$\zeta^2 - (1+a)\zeta + a - z(b\zeta^2 + c\zeta + d).$$

作答时四问分别对应：相容性、根条件、阶条件、绝对稳定；不要混用。

## 5.12 一套可执行的答题流程

### 答题方法

遇到“分析一个线性多步法”的题，按以下顺序：

1. 统一索引，写出  $\alpha_j, \beta_j$  以及  $\rho, \sigma$ ;
2. 用  $\rho(1) = 0, \rho'(1) = \sigma(1)$  检查相容性;
3. 若问阶，计算  $C_q$ ，直到第一个非零项;
4. 分解  $\rho$ ，检查单位圆内外及单位圆重根，判零稳定;
5. 由“相容 + 零稳定”判收敛；若问误差阶，同时检查启动值;
6. 对  $y' = \lambda y$  写  $\rho(\zeta) - z\sigma(\zeta) = 0$ ，检查全部根，而非只找一个;
7. 稳定域边界用  $\zeta = e^{i\theta}$  参数化，再以测试点判定内外侧。

## 5.13 自测与简答

1. 证明 AB2 为二阶，并求第一个非零的  $C_q$ 。
2. 检查 BDF2 的根条件；其第二个根是多少？
3. 构造一个  $\rho$  的所有根都在闭单位圆、但方法不零稳定的例子。
4. 为什么绝对稳定域的定义在单位圆重根处要附加单根条件？
5. 若一个三阶 LMM 使用一阶 Euler 法生成启动值，为什么总体三阶可能丢失？

6. 推导梯形法的稳定函数, 并说明它是否 A-稳定。

### 核对要点

1. 对  $\alpha = (0, -1, 1)$ 、 $\beta = (-\frac{1}{2}, \frac{3}{2}, 0)$ ,  $C_0 = C_1 = C_2 = 0$ , 且按本文定义

$$C_3 = \frac{-1+8}{6} - \frac{1}{2} \left( \frac{3}{2} \right) = \frac{5}{12} \neq 0.$$

因此 AB2 为二阶; 注意索引平移和方法整体归一化必须一致。

2. BDF2 的  $\rho(\zeta) = 3\zeta^2 - 4\zeta + 1$ , 根为  $1, 1/3$ 。
3. 取  $\rho(\zeta) = (\zeta - 1)^2$ ; 单位根重数二导致  $n$  模态。
4. 重根会产生  $n\zeta^n$ , 模长条件本身不能排除多项式增长。
5. 启动误差是全局误差估计中的首项; 低阶启动误差不会被零稳定自动消除。
6. 梯形法  $R(z) = (1 + z/2)/(1 - z/2)$ , 左半平面满足  $|R(z)| \leq 1$ , 故 A-稳定。

### 延伸核对

7. 中点型二步法

$$y_{n+2} - y_n = 2hf_{n+1}$$

有  $\rho(\zeta) = \zeta^2 - 1$ 。它是否零稳定? 为什么这仍不保证对所有  $\Re \lambda < 0$  的步长都绝对稳定?

8. 若  $\rho$  有根  $1, 1/2$ , 方法相容且二阶, 启动误差分别为  $e_0 = 0, e_1 = Ch$ 。预期全局阶是多少?
9. 边界曲线经过同一个  $z_*$  两次时, 为什么不能仅凭曲线朝向判断稳定侧?

**提示。**第 7 题的根  $1, -1$  都是单位圆单根, 所以零稳定; 但代入测试方程后必须检查  $\rho - z\sigma$  的两根, 零稳定不是 A-稳定。第 8 题中零稳定只保证启动扰动不被失控放大, 并不会把  $\mathcal{O}(h)$  自动提升为  $\mathcal{O}(h^2)$ , 故一般只能得到一阶。第 9 题应在边界分出的每个连通区域选测试点求全部根; 自交点还可能对应不同根同时触及单位圆。

### 本节结论

真正需要带走的不是一张系数表, 而是三条互不混淆的链:

插值/积分  $\Rightarrow$  公式,      Taylor 系数  $\Rightarrow$  阶与相容性,

$\rho$  的根  $\Rightarrow$  零稳定与收敛,       $\rho - z\sigma$  的根  $\Rightarrow$  绝对稳定。

## 5.14 边值问题、方法比较与历年迁移

打靶法把未知初始斜率  $s$  当参数, 积分 IVP 后用终点残量  $R(s) = y(b; s) - \beta$  求根; Newton 需要灵敏度方程, 割线只需多次积分。中心差分则直接把二阶边值问题离散成三对角系统。边值问题属于中低频内容, 正文要求会说明两条路线及其代价, 不扩展一般非线性边值理论。

| 方法       | 阶/成本        | 稳定性     | 隐式求解   | 适用       |
|----------|-------------|---------|--------|----------|
| 显式 Euler | 1 阶/1 次 $f$ | 小稳定域    | 无      | 非刚性、教学基线 |
| 后向 Euler | 1 阶         | A/L 稳定  | 每步非线性解 | 强耗散刚性    |
| 梯形       | 2 阶         | A 稳定非 L | 每步非线性解 | 精度高但刚性振荡 |
| 显式 RK4   | 4 阶/4 次 $f$ | 有界稳定域   | 无      | 平滑非刚性    |
| BDF      | 1–6 阶族      | 高阶稳定性受限 | 每步隐式   | 刚性与多步复用  |

历年迁移：2023 秋 Q4 要从二级 RK 系数推导阶和稳定函数；其余五期 ODE/渐近综合题已在真题章逐题核验，不把“振子渐近题”误归为纯数值 ODE。

易错点

不要混淆局部 defect 阶和除  $h$  截断误差阶；不要把一致性当收敛；不要省略多步法单位圆重根必须单；不要把 A-稳定写成“所有  $z$  稳定”；不要把梯形法称 L-稳定；不要用显式法精度阶替代刚性稳定分析。

复习检查

能否由 Taylor 展开判一步法/RK 阶？能否证明离散 Gronwall 整体误差？能否从 Butcher 表写稳定函数？能否对具体多步法写  $\rho, \sigma$ 、判一致性与根条件？能否区分零稳定、绝对稳定、A/L 稳定和刚性，并解释后向 Euler 与梯形的强刚性差异？

答题方法

ODE 题先辨认一步/多步、显式/隐式和所问稳定概念；判阶用 Taylor/阶条件，收敛用相容性加稳定性，刚性比较用测试方程稳定域。

自测

1. 从局部 defect 和 Lipschitz 稳定推导一步法的全局误差界。
2. 写出经典四阶 RK 的 Butcher 条件并解释树条件的作用。
3. 对给定 LMM 计算  $\rho, \sigma$ 、阶、根条件和负实轴稳定区间。
4. 说明零稳定、绝对稳定与 A-稳定分别控制什么。
5. 比较 Adams–Moulton、BDF 与显式 RK 在刚性问题中的取舍。

本节结论

数值 ODE 的核心不是公式清单，而是局部误差、根条件、误差传播和测试方程四条证据链。

## 5.15 Runge–Kutta 方法的树前阶条件

$s$  级显式 Runge–Kutta 写为

$$k_i = f \left( t_n + c_i h, y_n + h \sum_{j < i} a_{ij} k_j \right), \quad y_{n+1} = y_n + h \sum_{i=1}^s b_i k_i.$$

令  $e = (1, \dots, 1)^\top$ , 一致性要求

$$b^\top e = 1.$$

把精确解与数值格式分别 Taylor 展开, 前三阶条件为

$$b^\top e = 1, \quad b^\top c = \frac{1}{2}, \quad b^\top c^{\odot 2} = \frac{1}{3}, \quad b^\top A c = \frac{1}{6}.$$

其中  $c = A e$  常作为内部一致性条件。第三阶出现两个条件, 是因为  $y'''$  中存在不同微分树结构, 不能只凭一个标量矩匹配。

**例题 5.9** (经典 RK4 核验).

|       |     |     |     |     |
|-------|-----|-----|-----|-----|
| 0     |     |     |     |     |
| 1/2   | 1/2 |     |     |     |
| 1/2   | 0   | 1/2 |     |     |
| 1     | 0   | 0   | 1   |     |
| <hr/> |     |     |     |     |
|       | 1/6 | 1/3 | 1/3 | 1/6 |

满足  $b^\top e = 1$ 、 $b^\top c = 1/2$  等四阶条件。考试若只要求判二阶, 核验前两条即可; 若宣称四阶, 则不能只依据方法名称。

## 5.16 稳定函数与 A/L 稳定

对测试方程  $y' = \lambda y$ , RK 方法给

$$y_{n+1} = R(z) y_n, \quad z = h\lambda,$$

$$R(z) = 1 + z b^\top (I - zA)^{-1} e.$$

绝对稳定域为

$$\mathcal{S} = \{z \in \mathbb{C} : |R(z)| \leq 1\}.$$

A 稳定要求整个左半平面包含于  $\mathcal{S}$ 。L 稳定还要求

$$\lim_{z \rightarrow -\infty} R(z) = 0,$$

使极刚模态被强烈阻尼。

显式 Euler  $R(z) = 1 + z$ , 负实轴稳定区间为  $[-2, 0]$ ; 隐式 Euler

$$R(z) = \frac{1}{1 - z}$$

A 稳定且 L 稳定; 梯形法

$$R(z) = \frac{1 + z/2}{1 - z/2}$$

A 稳定但极限为  $-1$ , 不是 L 稳定。

## 5.17 局部误差到全局误差的一步法证明

设一步法

$$y_{n+1} = \Phi_h(t_n, y_n)$$

对状态 Lipschitz:

$$\|\Phi_h(t, u) - \Phi_h(t, v)\| \leq (1 + Lh)\|u - v\|.$$

精确解满足

$$y(t_{n+1}) = \Phi_h(t_n, y(t_n)) + \tau_{n+1}, \quad \|\tau_{n+1}\| \leq Ch^{p+1}.$$

误差递推

$$\|e_{n+1}\| \leq (1 + Lh)\|e_n\| + Ch^{p+1}.$$

离散 Grönwall 给在  $t_n \leq T$  上

$$\|e_n\| \leq e^{LT}\|e_0\| + \frac{C}{L}(e^{LT} - 1)h^p.$$

因此局部  $p+1$  阶产生全局  $p$  阶; 指数因子来自连续依赖, 不是凭空多出。

## 5.18 刚性: 精度允许的步长与稳定允许的步长

考虑

$$y' = -1000(y - \cos t) - \sin t, \quad y(0) = 1,$$

精确解  $y(t) = \cos t$ 。虽然解变化尺度为  $O(1)$ , 齐次扰动满足  $\eta' = -1000\eta$ 。显式 Euler 为稳定必须

$$|1 - 1000h| \leq 1, \quad h \leq 0.002,$$

远小于按  $\cos t$  精度直觉选择的步长。隐式 Euler 对所有  $h > 0$  稳定。刚性是“稳定步长远小于精度步长”的现象, 不能仅由某个大负特征值脱离问题尺度定义。

## 5.19 边值问题: 打靶与有限差分

二阶边值问题

$$-y'' + q(x)y = f(x), \quad y(a) = \alpha, \quad y(b) = \beta$$

可用中心差分

$$-\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + q_i y_i = f_i.$$

若  $q_i \geq 0$ , 离散矩阵严格对角占优/为 M-矩阵, 便于证明唯一性与最大值原理。局部截断误差为  $O(h^2)$ , 稳定逆界再给全局二阶。

打靶法把未知初始斜率  $s = y'(a)$  当参数, 积分 IVP 得

$$F(s) = y(b; s) - \beta,$$

再用 Newton/割线求  $F(s) = 0$ 。若边值问题存在强不稳定模态, 微小  $s$  误差会在积分中指数放大, 打靶可能病态; 有限差分则直接解全局耦合系统。

### 易错点

AB2 按本书闭单位圆且单位圆根单的定义, 负实轴绝对稳定区间为  $[-1, 0]$ :  $z = -1$  的根为  $-1, 1/2$ ,  $z = 0$  的根为  $1, 0$ , 边界单位根均为单根。不能在正文使用闭区间、答案又改用开区间。

### 答题方法

ODE 题先区分四件事: 阶由局部展开决定; 零稳定由  $z = 0$  的第一特征多项式决定; 绝对稳定由测试方程的  $z$ -依赖根决定; 刚性比较稳定与精度步长。任何两个条件都不能互相替代。

### 自测

1. 写出 RK 三阶的四个基本条件并解释为什么有两个三阶条件。
2. 比较隐式 Euler 与梯形法的 A/L 稳定性。
3. 用离散 Grönwall 从局部  $p+1$  阶推出全局  $p$  阶。
4. 对 AB2 在  $z = -1, 0$  处核验根条件。

## 第六章 数值偏微分方程：30 个培训核心与 176 个公式单元

### 考试范围提醒

本章严格沿三节培训板书展开。课堂导言曾列出有限差分、有限元、谱方法和深度学习等方法名称；考试主线只展开板书实际推导的有限差分内容。一般 Poisson 五点/九点误差理论、有限元、谱方法和一般复杂边界处理不在此扩展。

### 6.1 网格、模板、层数与统一参数

在区间  $[a, b]$  上取

$$x_j = a + jh, \quad h = \frac{b-a}{M}, \quad t_n = n\tau, \quad u_j^n \approx u(x_j, t_n).$$

板书同时区分 Cauchy 问题与初边值问题：前者在整线给初值，后者还须在  $x = a, b$  给边界数据。为避免不同方程的无量纲参数混写，本章固定

$$r = \frac{\kappa\tau}{h^2} \quad (\text{常系数热方程}), \quad \lambda = \frac{\tau}{h^2} \quad (\text{Schrödinger 与变系数扩散}),$$

$$\nu = \frac{a\tau}{h} \quad (\text{速度为 } a \text{ 的输运}).$$

常用差分算子为

$$\begin{aligned} \delta_t^+ u_j^n &= \frac{u_j^{n+1} - u_j^n}{\tau}, \quad \delta_t^0 u_j^n = \frac{u_j^{n+1} - u_j^{n-1}}{2\tau}, \\ D_x^+ u_j^n &= \frac{u_{j+1}^n - u_j^n}{h}, \quad D_x^- u_j^n = \frac{u_j^n - u_{j-1}^n}{h}, \quad \delta_x^0 u_j^n = \frac{u_{j+1}^n - u_{j-1}^n}{2h}, \\ \delta_x^2 u_j^n &= \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}. \end{aligned}$$

### 6.1.1 显式/隐式与两层/三层模板

| 格式             | 时间层模板                                                | 类型    | 更新特征                 |
|----------------|------------------------------------------------------|-------|----------------------|
| FTCS           | $\{u_{j-1}^n, u_j^n, u_{j+1}^n; u_j^{n+1}\}$         | 两层、显式 | 由第 $n$ 层直接求第 $n+1$ 层 |
| 后向 Euler       | $\{u_j^n; u_{j-1}^{n+1}, u_j^{n+1}, u_{j+1}^{n+1}\}$ | 两层、隐式 | 每一步解线性系统             |
| Crank–Nicolson | $n, n+1$ 两层均用三点空间模板                                  | 两层、隐式 | 时间居中、二阶              |
| Richardson/DFF | $n-1, n, n+1$ 三层                                     | 三层、显式 | 还需单独提供 $u^1$         |
| leapfrog       | $n-1, n, n+1$ 三层                                     | 三层、显式 | 中心时间、中心空间            |

“显式/隐式”描述每步是否要求解含新层未知量的联立系统; “两层/三层”描述递推涉及多少个时间层; “一步/二步”描述需要多少个历史状态。三层显式格式仍必须用相容的一步法启动。

## 6.2 Taylor 展开、相容、稳定与 Lax 等价

对光滑函数, 板书从

$$\delta_t^+ u = u_t + \frac{\tau}{2} u_{tt} + \frac{\tau^2}{6} u_{ttt} + O(\tau^3),$$

$$\delta_x^2 u = u_{xx} + \frac{h^2}{12} u_{xxxx} + O(h^4)$$

得到各格式的残差。把精确解代入已经除以时间步的差分方程所得残差称局部截断误差; 若采用未除  $\tau$  的递推残差, 阶数整体多一个  $\tau$ , 必须先说明约定。

**定义 6.1** (稳定性). 设初始扰动为  $\xi^0$ , 传播到第  $n$  层为  $\xi^n$ 。若存在与  $h, \tau, n$  无关的  $K, T, \tau_0 > 0$ , 使得在  $n\tau \leq T, \tau \leq \tau_0$  时

$$\|\xi^n\|_X \leq K \|\xi^0\|_X,$$

则称格式在范数  $X$  下稳定。稳定性依赖范数; 一个范数稳定不能自动推出另一范数稳定。

**定理 6.1** (Lax–Richtmyer 等价定理). 对线性适定初值问题及其线性差分逼近, 在指定范数、有限时间区间和指定网格族上, 若格式相容, 则稳定当且仅当收敛。

#### 证明思路

把全局误差写成初值误差与各步局部残差经离散传播算子后的和。稳定性一致控制传播算子, 相容性使每步残差趋零, 因而得到收敛。反向方向由每个数据上的收敛得到离散解算子族逐点有界, 再由一致有界原理得到稳定。线性、适定、相容、范数与网格族缺一不可。

**例题 6.1** (FTCS 的完整 Taylor 误差链). 对  $u_t = \kappa u_{xx}$ , 格式  $\delta_t^+ u_j^n = \kappa \delta_x^2 u_j^n$  给

$$\mathcal{T}_j^n = \delta_t^+ u - \kappa \delta_x^2 u = \frac{\tau}{2} u_{tt} - \frac{\kappa h^2}{12} u_{xxxx} + O(\tau^2 + h^4) = O(\tau + h^2).$$

板书列出的稳定性工具为: 直接矩阵法、Fourier/von Neumann 法、能量法和最大值原理。常系数周期/整线问题优先 Fourier; 边界、变系数问题常用能量或最大值法。

## 6.3 离散 Fourier–Parseval 与矩阵幂判据

### 6.3.1 从网格函数到离散二范数稳定性

对  $u^n = (u_j^n)_{j \in \mathbb{Z}} \in \ell_h^2$ , 定义

$$\|u^n\|_{\ell_h^2}^2 = h \sum_{j \in \mathbb{Z}} |u_j^n|^2, \quad \widehat{u}^n(\xi) = h \sum_{j \in \mathbb{Z}} u_j^n e^{-ijh\xi}, \quad |\xi| \leq \frac{\pi}{h}.$$

等价地, 把  $u_j^n$  延拓为单元  $[x_j - h/2, x_j + h/2)$  上的分片常值网格函数。离散 Fourier 逆变换与 Parseval 恒等式为

$$u_j^n = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} \widehat{u}^n(\xi) e^{ijh\xi} d\xi,$$

$$h \sum_j |u_j^n|^2 = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} |\widehat{u}^n(\xi)|^2 d\xi.$$

若两层标量格式在 Fourier 空间满足

$$\widehat{u}^{n+1}(\xi) = G(h\xi) \widehat{u}^n(\xi),$$

则

$$\|u^n\|_{\ell_h^2}^2 = \frac{1}{2\pi} \int |G(h\xi)|^{2n} |\widehat{u}^0(\xi)|^2 d\xi.$$

因此  $\sup_{\theta, n} |G(\theta)|^n \leq K$  蕴含

$$\|u^n\|_{\ell_h^2} \leq K \|u^0\|_{\ell_h^2}.$$

在固定有限时间  $n\tau \leq T$  的一般定义下,  $|G| \leq 1 + C\tau$  也足够, 因为  $(1 + C\tau)^n \leq e^{CT}$ ; 本章无增长模型通常直接检验  $|G| \leq 1$ 。

**定理 6.2** (Fourier 模态界通过 Parseval 推出离散二范数稳定). 若周期或整线常系数格式满足

$$\widehat{u}^{n+1}(\xi) = G(h\xi) \widehat{u}^n(\xi), \quad \sup_{\theta} |G(\theta)| \leq 1 + C\tau,$$

则对  $n\tau \leq T$ ,

$$\|u^n\|_{\ell_h^2} \leq e^{CT} \|u^0\|_{\ell_h^2}.$$

特别地, 若  $|G| \leq 1$ , 离散二范数不增。

*证明.* 递推给  $\widehat{u}^n(\xi) = G(h\xi)^n \widehat{u}^0(\xi)$ 。由离散 Parseval 恒等式,

$$\begin{aligned} \|u^n\|_{\ell_h^2}^2 &= \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} |G(h\xi)|^{2n} |\widehat{u}^0(\xi)|^2 d\xi \\ &\leq (1 + C\tau)^{2n} \|u^0\|_{\ell_h^2}^2 \leq e^{2CT} \|u^0\|_{\ell_h^2}^2. \end{aligned}$$

取平方根即得结论。这个证明依赖 Parseval 与 Fourier 对角化; 边界或变系数破坏对角化时, 不能只逐模态检查而跳过边界闭合。□

对系统或多步格式, 令  $U^{n+1} = G(\theta)U^n$ , 稳定目标为

$$\sup_{n \geq 0} \sup_{\theta \in [-\pi, \pi]} \|G(\theta)^n\| < \infty.$$

仅有  $\rho(G) \leq 1$  不够: 单位圆特征值必须半单, 相似变换还必须对波数一致良态。

**例题 6.2** (谱半径并不充分).

$$G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad G^n = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}.$$

虽有  $\rho(G) = 1$ , 矩阵幂仍线性增长。

### 6.3.2 二次多项式的闭单位圆根判据

**引理 6.3** (实系数二次式的根条件). 设

$$p(\chi) = \chi^2 + b\chi + c, \quad b, c \in \mathbb{R}.$$

两根都位于闭单位圆内当且仅当

$$|b| \leq 1 + c \leq 2.$$

两根都位于开单位圆内当且仅当  $|b| < 1 + c < 2$ 。闭单位圆判据只控制根的位置; 若单位圆上出现重根, 还须另查半单性。

**完整证明.** 设两根为  $\chi_1, \chi_2$ 。若  $|\chi_i| \leq 1$ , 则  $c = \chi_1\chi_2 \leq 1$ 。当两根为共轭复根时,  $p(1) = |1 - \chi_1|^2 \geq 0$ ,  $p(-1) = |1 + \chi_1|^2 \geq 0$ ; 当两根为实根且位于  $[-1, 1]$  时同样有  $p(\pm 1) \geq 0$ 。故

$$1 + b + c \geq 0, \quad 1 - b + c \geq 0, \quad c \leq 1,$$

即  $|b| \leq 1 + c \leq 2$ 。

反之, 条件给  $p(1) \geq 0$ ,  $p(-1) \geq 0$  以及  $-1 \leq c \leq 1$ 。若根为非实共轭对, 则  $c = |\chi_1|^2 \leq 1$ 。若根为实数而有一根  $> 1$ , 由  $p(1) = (1 - \chi_1)(1 - \chi_2) \geq 0$  可知另一根也  $\geq 1$ , 于是  $c \geq 1$ ; 除  $\chi_1 = \chi_2 = 1$  外均矛盾。根  $< -1$  的情形由  $p(-1) \geq 0$  同理排除。严格不等式恰好排除  $\pm 1$  及  $|c| = 1$  的边界情形。  $\square$

### 6.3.3 参数矩阵族临界频率附近的一致幂有界

**定理 6.4** (板书中的参数矩阵族判据). 设  $\tilde{G}(\sigma) \in \mathbb{C}^{p \times p}$  是  $2\pi$  周期的  $C^s$  矩阵族。对每个  $\sigma_0$ , 假设下列三种情形至少有一种成立:

- (i)  $\tilde{G}(\sigma_0)$  有  $p$  个互异特征值;
- (ii)  $\rho(\tilde{G}(\sigma_0)) < 1$ ;
- (iii) 存在  $s \geq 1$ , 使

$$\tilde{G}^{(j)}(\sigma_0) = \gamma_j I, \quad j = 0, \dots, s-1,$$

且  $\tilde{G}^{(s)}(\sigma_0)$  有  $p$  个互异特征值。

则

$$\sup_{\sigma, n} \left\| \tilde{G}(\sigma)^n \right\| < \infty \iff \rho(\tilde{G}(\sigma)) \leq 1 \text{ 对所有 } \sigma \text{ 成立.}$$

完整证明. 必要性由  $\rho(G^n) \leq \|G^n\|$  立即得到. 对充分性, 在互异谱点附近可连续选择特征向量, 相似变换及其逆在一个小邻域内一致有界; 在  $\rho < 1$  的点附近, 连续性给出统一余量  $\rho \leq q < 1$ , 矩阵幂因而一致有界. 第三种情形由 Taylor 展开

$$\tilde{G}(\sigma_0 + z) = \sum_{j=0}^{s-1} \frac{\gamma_j}{j!} z^j I + \frac{z^s}{s!} H(z), \quad H(0) = \tilde{G}^{(s)}(\sigma_0),$$

化为互异谱矩阵族  $H(z)$  的一致对角化; 前面的标量多项式不改变特征向量. 最后由频率区间的紧性取有限子覆盖. 该导数条件正是处理“临界点处矩阵为标量阵、离开临界点后特征值才分裂”的板书工具.  $\square$

## 6.4 常系数热方程: FTCS、后向 Euler 与 Crank–Nicolson

对  $u_t = \kappa u_{xx}$ , 令  $r = \kappa\tau/h^2$ . FTCS 为

$$u_j^{n+1} = (1 - 2r)u_j^n + r(u_{j-1}^n + u_{j+1}^n).$$

代入  $u_j^n = \hat{u}^n e^{ij\theta}$  得

$$G_{\text{FTCS}} = 1 + r(e^{i\theta} - 2 + e^{-i\theta}) = 1 - 4r \sin^2 \frac{\theta}{2}.$$

故  $|G| \leq 1$  对所有频率成立当且仅当  $0 \leq r \leq 1/2$ ; 此时更新也是非负凸组合, 兼有  $\ell^\infty$  稳定性.

后向 Euler 为

$$u_j^{n+1} - r(u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}) = u_j^n,$$

故

$$G_{\text{BE}} = \frac{1}{1 + 4r \sin^2(\theta/2)}.$$

Crank–Nicolson 为

$$\frac{u_j^{n+1} - u_j^n}{\tau} = \frac{\kappa}{2} \delta_x^2 (u_j^{n+1} + u_j^n), \quad G_{\text{CN}} = \frac{1 - 2r \sin^2(\theta/2)}{1 + 2r \sin^2(\theta/2)}.$$

两者在 Fourier  $\ell^2$  意义下对  $r \geq 0$  无条件稳定. BE 时间一阶且强阻尼; CN 时间二阶, 但  $r$  大时最高频接近  $-1$ , 所以无条件稳定不等于单调或无振荡.

## 6.5 Richardson 与 Du Fort–Frankel 的完整根分析

Richardson 格式

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\tau} = \delta_x^2 u_j^n$$

代入  $u_j^n = \chi^n e^{ij\theta}$  得

$$\chi^2 + 8r \sin^2 \frac{\theta}{2} \chi - 1 = 0, \quad r = \frac{\tau}{h^2}.$$

令  $q = 4r \sin^2(\theta/2) > 0$ , 两根为

$$\chi_{\pm} = -q \pm \sqrt{1+q^2}.$$

其中  $\chi_- < -1$ , 故除零频率外均有增长根, Richardson 不稳定。

Du Fort–Frankel 格式为

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\tau} = \frac{u_{j+1}^n - u_j^{n+1} - u_j^{n-1} + u_{j-1}^n}{h^2},$$

即

$$(1+2r)u_j^{n+1} = (1-2r)u_j^{n-1} + 2r(u_{j-1}^n + u_{j+1}^n).$$

Fourier 代入给完整特征方程

$$(1+2r)\chi^2 - 4r \cos \theta \chi - (1-2r) = 0,$$

或

$$\chi^2 + b\chi + c = 0, \quad b = -\frac{4r \cos \theta}{1+2r}, \quad c = -\frac{1-2r}{1+2r}.$$

**定理 6.5** (DFF 的 Fourier 无条件稳定性). 对任意  $r \geq 0$ , DFF 的两个特征根均在闭单位圆内, 并且相应相伴矩阵族关于波数一致幂有界; 故该格式在 Fourier 意义下无条件稳定。

完整证明. 现在

$$1+c = \frac{4r}{1+2r}, \quad |b| \leq \frac{4r}{1+2r} = 1+c \leq 2.$$

由二次根判据, 两根对任意  $r \geq 0$ 、任意  $\theta$  均在闭单位圆。若出现重根, 则判别式为零, 即

$$1 - 4r^2 \sin^2 \theta = 0;$$

此时重根严格位于单位圆内。单位圆上的根只可能在  $\theta = 0, \pi$ , 且另一根与之不同。结合参数矩阵族判据, 相伴矩阵幂对波数一致有界, 因此 DFF 在 Fourier 意义下无条件稳定。□

稳定不等于相容。精确解代入 DFF, 时间中心差分为

$$\frac{u(t+\tau) - u(t-\tau)}{2\tau} = u_t + \frac{\tau^2}{6} u_{ttt} + O(\tau^4),$$

而右端为

$$u_{xx} + \frac{h^2}{12} u_{xxxx} - \frac{\tau^2}{h^2} u_{tt} + O\left(h^4 + \frac{\tau^4}{h^2}\right).$$

故

$$\mathcal{T}_j^n = \frac{\tau^2}{6} u_{ttt} - \frac{h^2}{12} u_{xxxx} + \frac{\tau^2}{h^2} u_{tt} + O\left(\tau^4 + h^4 + \frac{\tau^4}{h^2}\right),$$

即  $O(\tau^2 + h^2 + \tau^2/h^2)$ 。相容必须满足  $\tau/h \rightarrow 0$ ; 只有再结合稳定性才能沿该网格族得到收敛。

## 6.6 Schrödinger 实虚部系统与严格 Jordan 临界

考虑

$$u_t = -v_{xx}, \quad v_t = u_{xx}, \quad \psi = u + iv,$$

于是  $\psi_t = i\psi_{xx}$ 。顺序半隐式格式为

$$\frac{u_j^{n+1} - u_j^n}{\tau} = -\frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{h^2},$$

$$\frac{v_j^{n+1} - v_j^n}{\tau} = \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2}.$$

此处  $\lambda = \tau/h^2$ 。令  $s = \sin(\sigma/2)$ , 对  $U^n = (\hat{u}^n, \hat{v}^n)^T$ ,

$$G(\sigma) = \begin{pmatrix} 1 & 4\lambda s^2 \\ -4\lambda s^2 & 1 - 16\lambda^2 s^4 \end{pmatrix}, \quad \chi^2 - (2 - 16\lambda^2 s^4)\chi + 1 = 0.$$

令  $q = 4\lambda s^2$ , 迹为  $2 - q^2$ 。根积为一, 闭单位圆根条件给  $|2 - q^2| \leq 2$ , 即  $q \leq 2$ ; 对全频率先得  $\lambda \leq 1/2$ 。

**定理 6.6** (顺序半隐式 Schrödinger 分裂的严格稳定条件). 在矩阵幂对全波数一致有界的稳定定义下, 格式稳定当且仅当

$$\boxed{\lambda < \frac{1}{2}}.$$

完整证明. 当  $0 \leq \lambda < 1/2$  时, 令  $q_{\max} = 4\lambda < 2$ , 并定义

$$\cos \omega = 1 - \frac{q^2}{2}, \quad \sin \omega = q\sqrt{1 - \frac{q^2}{4}}.$$

Cayley-Hamilton 恒等式给

$$G^n = \cos(n\omega)I + \frac{\sin(n\omega)}{\sin \omega}(G - \cos \omega I),$$

且

$$\frac{G - \cos \omega I}{\sin \omega} = \frac{1}{\sqrt{1 - q^2/4}} \begin{pmatrix} q/2 & 1 \\ -1 & -q/2 \end{pmatrix}.$$

因  $q \leq q_{\max} < 2$ , 右端一致有界;  $q = 0$  时  $G = I$ 。故矩阵幂一致有界。

若  $\lambda > 1/2$ , 最高频  $s^2 = 1$  给  $q > 2$ , 一根模大于一。临界  $\lambda = 1/2, s^2 = 1$  时

$$G = \begin{pmatrix} 1 & 2 \\ -2 & -3 \end{pmatrix}, \quad \det(\chi I - G) = (\chi + 1)^2,$$

但  $G \neq -I$ 。写  $G = -I + N$ , 其中  $N \neq 0, N^2 = 0$ , 则

$$G^n = (-1)^n(I - nN),$$

故  $\|G^n\| \sim Cn$ 。临界等号必须排除。 □

## 6.7 输运格式逐项 Fourier 比较与 CFL 必要条件

考虑

$$u_t + au_x = 0, \quad u(x, 0) = u_0(x), \quad u(x, t) = u_0(x - at), \quad \nu = \frac{a\tau}{h}.$$

特征线是  $x - at = C$ 。 $a > 0$  时信息从左向右, 上游在左;  $a < 0$  时信息从右向左, 上游在右。以下先取  $a > 0$ 。

### 6.7.1 顺风、迎风与 FTCS 中心格式

前向时间、前向空间的顺风格式为

$$u_j^{n+1} = (1 + \nu)u_j^n - \nu u_{j+1}^n,$$

$$G_{\text{down}} = 1 + \nu - \nu e^{i\theta}, \quad |G_{\text{down}}|^2 = 1 + 4\nu(1 + \nu) \sin^2 \frac{\theta}{2} > 1$$

(非零频率), 故不稳定。

前向时间、后向空间的迎风格式为

$$u_j^{n+1} = (1 - \nu)u_j^n + \nu u_{j-1}^n,$$

$$G_{\text{up}} = 1 - \nu + \nu e^{-i\theta}, \quad |G_{\text{up}}|^2 = 1 - 4\nu(1 - \nu) \sin^2 \frac{\theta}{2}.$$

故  $0 \leq \nu \leq 1$  时稳定, 并且更新是凸组合。

FTCS 中心格式为

$$u_j^{n+1} = u_j^n - \frac{\nu}{2}(u_{j+1}^n - u_{j-1}^n),$$

$$G_{\text{FTCS}} = 1 - i\nu \sin \theta, \quad |G_{\text{FTCS}}|^2 = 1 + \nu^2 \sin^2 \theta > 1,$$

故一般不稳定。三者的差别来自空间差分方向, 而不是时间差分。

### 6.7.2 Lax-Friedrichs: 稳定与相容分开

$$u_j^{n+1} = \frac{1}{2}(u_{j+1}^n + u_{j-1}^n) - \frac{\nu}{2}(u_{j+1}^n - u_{j-1}^n),$$

$$G_{LF} = \cos \theta - i\nu \sin \theta, \quad |G_{LF}|^2 = 1 - (1 - \nu^2) \sin^2 \theta.$$

故 Fourier 稳定条件为  $|\nu| \leq 1$ 。精确解代入给

$$\mathcal{T} = u_t + au_x - \frac{h^2}{2\tau} u_{xx} + \frac{\tau}{2} u_{tt} + \frac{ah^2}{6} u_{xxx} + O\left(\tau^2 + \frac{h^4}{\tau} + h^4\right),$$

所以相容除  $\tau \rightarrow 0$  外还要求  $h^2/\tau \rightarrow 0$ 。

### 6.7.3 连续依赖域、数值依赖域与 CFL

精确解在  $(x_j, t_n)$  只依赖初始点  $x_j - at_n$ 。若显式一步格式每步最多使用左右各  $m$  个网格点, 则从  $(x_j, t_n)$  回溯一步的数值依赖区间为  $[x_j - mh, x_j + mh]$ 。收敛的必要条件是连续依赖域包含在数值依赖域内:

$$x_j - a\tau \in [x_j - mh, x_j + mh], \quad \text{即} \quad |a|\tau \leq mh.$$

三点一步模板  $m = 1$  时得到  $|\nu| \leq 1$ 。

**定理 6.7** (线性输运格式的 CFL 必要条件). 设显式线性格式每一步只读取  $j - m, \dots, j + m$  的旧层数据, 并且对  $u_t + au_x = 0$  在固定  $T = n\tau$  上对任意有界连续初值一致收敛。则必须有  $|a|\tau \leq mh$ 。

完整证明. 经过  $n$  步, 数值解  $U_j^n$  只依赖初值网格点  $[x_j - nmh, x_j + nmh]$ 。精确解却等于  $u_0(x_j - an\tau)$ 。若  $|a|\tau > mh$ , 则精确足点  $x_j - an\tau$  位于数值依赖区间之外, 且二者距离随  $n$  线性增长。可取两个有界连续初值  $u_0, v_0$ , 使它们在数值依赖区间上完全相同, 而在精确足点附近分别取 0 与 1。由于格式读取到的数据相同, 两组数值解在  $(j, n)$  相等; 但对应精确解相差 1。这与对任意初值的一致收敛矛盾。因此连续依赖域必须包含在数值依赖域内, 即  $|a|n\tau \leq nmh$ , 约去  $n$  得所述条件。  $\square$

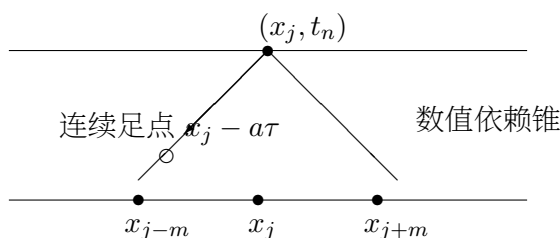


图 6.1: 连续依赖点必须落在数值依赖区间内; 否则不可能对任意初值收敛。

### 6.7.4 leapfrog 的严格临界条件

$$u_j^{n+1} = u_j^{n-1} - \nu(u_{j+1}^n - u_{j-1}^n).$$

代入  $u_j^n = \chi^n e^{ij\theta}$  得

$$\chi^2 + 2i\nu \sin \theta \chi - 1 = 0, \quad \chi_{\pm} = -i\nu \sin \theta \pm \sqrt{1 - \nu^2 \sin^2 \theta}.$$

当  $|\nu| < 1$  时两根对全频率位于单位圆且不在临界频率形成非半单重根; 当  $|\nu| > 1$  时取  $|\sin \theta| = 1$  即有增长根。当  $|\nu| = 1$  且  $|\sin \theta| = 1$  时出现非半单重复单位模根, 相伴矩阵含 Jordan 块。因此矩阵幂一致有界意义下

$$|\nu| < 1.$$

它是三层格式, 还须用相容的一步法提供  $u^1$ 。

### 6.7.5 Lax-Wendroff 的完整 Taylor 推导

从时间 Taylor 展开

$$u(x_j, t_{n+1}) = u + \tau u_t + \frac{\tau^2}{2} u_{tt} + O(\tau^3)$$

出发。由 PDE

$$u_t = -au_x, \quad u_{tt} = -a(u_t)_x = a^2 u_{xx}.$$

再用空间中心近似

$$u_x = \frac{u_{j+1}^n - u_{j-1}^n}{2h} + O(h^2),$$

$$u_{xx} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} + O(h^2),$$

得到

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

其残差为  $O(\tau^2 + h^2)$  (固定 Courant 比), 放大因子为

$$G_{LW} = 1 - i\nu \sin \theta + \nu^2(\cos \theta - 1),$$

$$|G_{LW}|^2 = 1 - 4\nu^2(1 - \nu^2) \sin^4 \frac{\theta}{2}.$$

故  $|\nu| \leq 1$  时稳定。图 6.1 同时解释该条件的依赖域必要性。

## 6.8 变系数非散度型扩散: 显式与 CN 型格式

考虑周期问题

$$u_t = a(x, t)u_{xx}, \quad 0 < a_0 \leq a(x, t), \quad \lambda = \frac{\tau}{h^2}.$$

### 6.8.1 显式格式的最大模与误差递推

令  $\gamma_j^n = \lambda a_j^n$ , 格式为

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

若  $0 \leq \gamma_j^n \leq 1/2$ , 三系数非负且和为一, 故

$$\|u^{n+1}\|_\infty \leq \|u^n\|_\infty, \quad \frac{\tau}{h^2} \|a\|_\infty \leq \frac{1}{2}$$

是统一充分条件。精确解代入产生  $R_j^n = O(\tau + h^2)$ 。设  $e_j^n = u(x_j, t_n) - u_j^n$ , 则

$$e_j^{n+1} = (1 - 2\gamma_j^n)e_j^n + \gamma_j^n(e_{j+1}^n + e_{j-1}^n) + \tau R_j^n.$$

逐层取最大模,

$$\|e^{n+1}\|_\infty \leq \|e^n\|_\infty + \tau \|R^n\|_\infty,$$

从而在  $t_n \leq T$  时

$$\|e^n\|_\infty \leq \|e^0\|_\infty + \tau \sum_{m=0}^{n-1} \|R^m\|_\infty \leq \|e^0\|_\infty + CT(\tau + h^2).$$

### 6.8.2 变系数 Crank–Nicolson 型格式

板书格式为

$$\delta_t^+ u_j^n = \frac{1}{2} a_j^n \delta_x^2 u_j^n + \frac{1}{2} a_j^{n+1} \delta_x^2 u_j^{n+1}.$$

记  $\gamma_j^n = \lambda a_j^n$ , 整理新旧层系数得

$$\begin{aligned} (1 + \gamma_j^{n+1}) u_j^{n+1} - \frac{\gamma_j^{n+1}}{2} (u_{j-1}^{n+1} + u_{j+1}^{n+1}) \\ = (1 - \gamma_j^n) u_j^n + \frac{\gamma_j^n}{2} (u_{j-1}^n + u_{j+1}^n). \end{aligned}$$

以  $t_{n+1/2}$  为中心展开时间平均, 并对两个时层使用中心空间差分, 得到

$$\mathcal{T}_j^{n+1/2} = O(\tau^2 + h^2).$$

常系数  $a$  时

$$G = \frac{1 - 2\lambda a \sin^2(\theta/2)}{1 + 2\lambda a \sin^2(\theta/2)},$$

故 Fourier  $\ell^2$  无条件稳定. 变系数  $\ell^\infty$  分析则不同: 左端矩阵  $A_{n+1}$  是行和为一的 M-矩阵,  $A_{n+1}^{-1} \geq 0$  且  $\|A_{n+1}^{-1}\|_\infty = 1$ ; 若  $0 \leq \gamma_j^n \leq 1$ , 右端矩阵  $B_n$  非负且行和为一, 于是

$$\|u^{n+1}\|_\infty \leq \|A_{n+1}^{-1}\|_\infty \|B_n u^n\|_\infty \leq \|u^n\|_\infty.$$

因此“常系数  $\ell^2$  无条件稳定”和“变系数  $\ell^\infty$  在  $\gamma_j^n \leq 1$  下稳定”是两条不同结论。

## 6.9 散度型变系数扩散: 半网格通量与完整误差链

对

$$u_t = (a(x, t) u_x)_x,$$

令

$$q_{j+1/2}^n = a_{j+1/2}^n \frac{u_{j+1}^n - u_j^n}{h}.$$

守恒格式是

$$\frac{u_j^{n+1} - u_j^n}{\tau} = \frac{q_{j+1/2}^n - q_{j-1/2}^n}{h}.$$

在半网格点作 Taylor 展开:

$$\frac{u(x_{j+1}, t_n) - u(x_j, t_n)}{h} = u_x(x_{j+1/2}, t_n) + \frac{h^2}{24} u_{xxx}(x_{j+1/2}, t_n) + O(h^4),$$

故

$$q_{j+1/2}^n = (au_x)(x_{j+1/2}, t_n) + O(h^2).$$

两个半网格通量再作中心差,

$$\frac{q_{j+1/2}^n - q_{j-1/2}^n}{h} = (au_x)_x(x_j, t_n) + O(h^2).$$

与  $\delta_t^+ u = u_t + O(\tau)$  合并, 得到局部截断误差

$$R_j^n = O(\tau + h^2).$$

令  $\eta_{j\pm 1/2}^n = \tau a_{j\pm 1/2}^n / h^2$ , 数值格式写成

$$u_j^{n+1} = (1 - \eta_{j+1/2}^n - \eta_{j-1/2}^n)u_j^n + \eta_{j+1/2}^n u_{j+1}^n + \eta_{j-1/2}^n u_{j-1}^n.$$

误差满足完全同型的递推

$$e_j^{n+1} = (1 - \eta_{j+1/2}^n - \eta_{j-1/2}^n)e_j^n + \eta_{j+1/2}^n e_{j+1}^n + \eta_{j-1/2}^n e_{j-1}^n + \tau R_j^n.$$

若  $a \geq 0$  且  $\eta_{j+1/2}^n + \eta_{j-1/2}^n \leq 1$ , 则

$$\|e^{n+1}\|_\infty \leq \|e^n\|_\infty + \tau \|R^n\|_\infty,$$

迭代即得

$$\|e^n\|_\infty \leq \|e^0\|_\infty + CT(\tau + h^2).$$

半网格位置和两个通量之间的负号是证明成立的关键。

## 6.10 离散最大值原理、唯一性与比较定理

令内部节点集合为  $I$ 、边界节点集合为  $B$ ,

$$(D_h u)_j = d_{jj}u_j - \sum_{k \in N(j)} d_{jk}u_k, \quad j \in I.$$

假设  $d_{jk} \geq 0$ 、 $d_{jj} \geq \sum_k d_{jk}$ ; 内部网格图与边界相连, 并且每个内部连通分量至少有一个严格对角占优点或边界耦合。

**定理 6.8** (离散最小值原理). 若  $D_h u \geq 0$  于内部, 且  $u \geq 0$  于边界, 则  $u \geq 0$  于全网格。

完整证明. 反设内部有负最小值  $m = u_{j_0} < 0$ 。所有邻点  $u_k \geq m$ , 故

$$(D_h u)_{j_0} = \left( d_{j_0 j_0} - \sum_k d_{j_0 k} \right) m + \sum_k d_{j_0 k} (m - u_k) \leq 0.$$

与  $D_h u \geq 0$  合并知所有项均取等号。于是所有正权邻点也取值  $m$ ; 若本点严格占优则第一项严格为负, 已矛盾。否则负最小值沿正权边传播, 由连通性最终到达边界或严格占优点, 分别与边

界非负或严格不等式矛盾。  $\square$

**推论 6.9** (唯一性). 若  $D_h u = 0$  且边界为零, 分别对  $u$  和  $-u$  应用最小值原理, 得  $u \equiv 0$ 。

**定理 6.10** (离散比较定理). 设  $D_h u = f$ 、边界值为  $g$ ;  $D_h U = F$ 、边界值为  $G$ 。若  $|f| \leq F$ 、 $|g| \leq G$ , 则  $|u| \leq U$ 。

完整证明. 对  $U - u$  与  $U + u$ , 内部有  $D_h(U \mp u) = F \mp f \geq 0$ , 边界有  $G \mp g \geq 0$ 。最大值原理给  $U - u \geq 0$ 、 $U + u \geq 0$ , 即  $-U \leq u \leq U$ 。  $\square$

## 6.11 非线性扩散的中值定理凸组合

对

$$u_t = (f(u))_{xx},$$

显式格式为

$$u_j^{n+1} = u_j^n + \lambda [f(u_{j+1}^n) - 2f(u_j^n) + f(u_{j-1}^n)], \quad \lambda = \frac{\tau}{h^2}.$$

把两个差分别使用中值定理:

$$f(u_{j+1}^n) - f(u_j^n) = f'(\xi_+)(u_{j+1}^n - u_j^n),$$

$$f(u_j^n) - f(u_{j-1}^n) = f'(\xi_-)(u_j^n - u_{j-1}^n).$$

于是

$$u_j^{n+1} = [1 - \lambda f'(\xi_+) - \lambda f'(\xi_-)]u_j^n + \lambda f'(\xi_+)u_{j+1}^n + \lambda f'(\xi_-)u_{j-1}^n.$$

若  $0 \leq f'(z) \leq L$  且  $2\lambda L \leq 1$ , 三系数非负且和为一, 故有最大模稳定与比较性质。模型是  $(f(u))_{xx}$ , 不是  $(f(u_x))_{xx}$ 。

## 6.12 离散内积、范数、分部求和与 Green 公式

在  $x_j = jh$ 、 $j = 0, \dots, M$  上, 为明确端点是否计入, 定义四种离散乘积:

$$\langle u, v \rangle_h = h \sum_{j=1}^{M-1} u_j v_j, \quad [u, v]_h = h \sum_{j=0}^{M-1} u_j v_j,$$

$$(u, v)_h = h \sum_{j=1}^M u_j v_j, \quad [u, v]_h = h \sum_{j=0}^M u_j v_j.$$

相应范数由自乘定义, 例如

$$\|u\|_{0,h}^2 = \langle u, u \rangle_h, \quad \|u\|_{+,h}^2 = [u, u]_h, \quad \|u\|_{\text{full},h}^2 = (u, u)_h.$$

这些记号只是在不同差分算子定义域上避免漏掉端点。

**引理 6.11** (离散分部求和). 若  $q_j$  定义于  $j = 0, \dots, M-1$ ,  $v_j$  定义于  $j = 0, \dots, M$ , 则

$$h \sum_{j=1}^{M-1} (D^- q)_j v_j = -h \sum_{j=0}^{M-1} q_j (D^+ v)_j + q_{M-1} v_M - q_0 v_0.$$

**完整证明.** 展开左端并移动指标:

$$\sum_{j=1}^{M-1} (q_j - q_{j-1}) v_j = \sum_{j=1}^{M-1} q_j v_j - \sum_{j=0}^{M-2} q_j v_{j+1}.$$

在右端补入并减去  $q_0 v_0$ 、 $q_{M-1} v_M$ , 即得到所述恒等式。  $\square$

板书还使用两个直接的离散乘积恒等式:

$$h \sum_{j=0}^{M-1} (D^+ u)_j u_j = \frac{u_M^2 - u_0^2}{2} - \frac{h}{2} \sum_{j=0}^{M-1} |D^+ u_j|^2,$$

$$h \sum_{j=0}^{M-1} (D^+ u)_j u_{j+1} = \frac{u_M^2 - u_0^2}{2} + \frac{h}{2} \sum_{j=0}^{M-1} |D^+ u_j|^2.$$

它们来自  $2a(a-b) = a^2 - b^2 + (a-b)^2$  及其变形。

取  $q = D^+ u$  得含边界通量的离散 Green 公式:

$$\langle \delta_x^2 u, v \rangle_h = -h \sum_{j=0}^{M-1} (D^+ u)_j (D^+ v)_j + (D^+ u)_{M-1} v_M - (D^+ u)_0 v_0.$$

因此:

- (i) 若  $v_0 = v_M = 0$ , 边界通量消失;
- (ii) 若  $u, v$  均周期且端点差分通量匹配, 两端通量相消;
- (iii) 非齐次边界下必须保留两项, 不能直接套齐次公式。

## 6.13 热方程 FTCS 的完整离散能量估计

考虑

$$u_t - au_{xx} = 0, \quad x \in (0, l), \quad u(0, t) = u(l, t) = 0,$$

及 FTCS

$$u^{n+1} - u^n = r \Delta_0 u^n, \quad r = \frac{a\tau}{h^2}, \quad \Delta_0 u_j = u_{j+1} - 2u_j + u_{j-1}.$$

**定理 6.12** (FTCS 能量强制性与衰减). 若  $0 \leq r < 1/2$  且边界齐次, 则下述能量  $E^n$  与  $\|u^n\|_2^2$  等价并单调不减; 等号  $r = 1/2$  时该能量退化。

**完整证明.** 令  $\nabla_0^+ u_j = u_{j+1} - u_j$ , 并与乘子  $u^{n+1} + u^n$  作离散内积:

$$\|u^{n+1}\|_2^2 - \|u^n\|_2^2 = r \langle \Delta_0 u^n, u^{n+1} + u^n \rangle.$$

齐次边界的 Green 公式给

$$= -r \langle \nabla_0^+ u^n, \nabla_0^+ u^{n+1} + \nabla_0^+ u^n \rangle.$$

使用

$$2\langle p, p+q \rangle = \|p+q\|^2 + \|p\|^2 - \|q\|^2$$

并逐项移到等式两边, 得到

$$\begin{aligned} & \left( \|u^{n+1}\|_2^2 - \frac{r}{2} \|\nabla_0^+ u^{n+1}\|_2^2 \right) - \left( \|u^n\|_2^2 - \frac{r}{2} \|\nabla_0^+ u^n\|_2^2 \right) \\ &= -\frac{r}{2} \|\nabla_0^+ (u^{n+1} + u^n)\|_2^2 \leq 0. \end{aligned}$$

因此取

$$E^n = \|u^n\|_2^2 - \frac{r}{2} \|\nabla_0^+ u^n\|_2^2.$$

又由

$$\sum_j |u_{j+1} - u_j|^2 \leq 2 \sum_j (|u_{j+1}|^2 + |u_j|^2) \leq 4 \sum_j |u_j|^2$$

得逆估计  $\|\nabla_0^+ u\|_2^2 \leq 4\|u\|_2^2$ , 所以

$$(1 - 2r) \|u^n\|_2^2 \leq E^n \leq \|u^n\|_2^2.$$

当  $0 \leq r < 1/2$  时能量具有强制性且单调不减。临界  $r = 1/2$  的 Fourier 稳定可单独成立, 但这是一个能量在等号处退化。后向 Euler 则与  $u^{n+1}$  作内积, 直接得到无条件能量衰减。□

## 6.14 有限区间 leapfrog: 显式能量、边界通量与闭合

考虑

$$u_t - u_x = 0, \quad x \in (0, l),$$

速度为  $-1$ , 所以  $x = l$  是入流边界,  $x = 0$  是出流边界。右端给齐次入流数据  $u_M^n = 0$ , 内部格式为

$$u_j^{n+1} = u_j^{n-1} + \gamma(u_{j+1}^n - u_{j-1}^n), \quad \gamma = \frac{\tau}{h}, \quad 1 \leq j \leq M-1.$$

左端数值闭合为

$$u_0^n = \frac{u_1^{n+1} + u_1^{n-1}}{2}.$$

**定理 6.13** (有界区间 leapfrog 的能量稳定性). 若  $0 < \gamma < 1$ 、右端齐次入流且左端采用上述出流闭合, 则下述二层能量与相邻两层内部  $\ell^2$  范数等价, 并满足  $E^n \leq E^{n-1}$ 。

完整证明. 令内部范数

$$\|v\|_I^2 = \sum_{j=1}^{M-1} v_j^2,$$

并定义中心差分双线性型

$$B(v, w) = \sum_{j=1}^{M-1} (v_{j+1} - v_{j-1})w_j.$$

由直接移指标且  $v_M = w_M = 0$ ,

$$B(v, w) + B(w, v) = -v_0 w_1 - w_0 v_1.$$

内部方程乘以  $u_j^{n+1} + u_j^{n-1}$  并求和:

$$\|u^{n+1}\|_I^2 - \|u^{n-1}\|_I^2 = \gamma B(u^n, u^{n+1}) + \gamma B(u^n, u^{n-1}).$$

定义显式二层能量

$$E^n = \|u^{n+1}\|_I^2 + \|u^n\|_I^2 - \gamma B(u^n, u^{n+1}) - \gamma u_0^n u_1^{n+1}.$$

于是

$$\begin{aligned} E^n - E^{n-1} &= \gamma [B(u^n, u^{n-1}) + B(u^{n-1}, u^n)] \\ &\quad - \gamma u_0^n u_1^{n+1} + \gamma u_0^{n-1} u_1^n \\ &= -\gamma u_0^n (u_1^{n+1} + u_1^{n-1}). \end{aligned}$$

代入闭合条件即得

$$E^n - E^{n-1} = -2\gamma (u_0^n)^2 \leq 0.$$

还需证明能量强制性。注意

$$B(u^n, u^{n+1}) + u_0^n u_1^{n+1} = \sum_{j=1}^{M-1} u_{j+1}^n u_j^{n+1} - \sum_{j=2}^{M-1} u_{j-1}^n u_j^{n+1}.$$

Cauchy-Schwarz 与  $2ab \leq a^2 + b^2$  给

$$|B(u^n, u^{n+1}) + u_0^n u_1^{n+1}| \leq \|u^n\|_I^2 + \|u^{n+1}\|_I^2.$$

因此当  $0 < \gamma < 1$ ,

$$(1 - \gamma)(\|u^{n+1}\|_I^2 + \|u^n\|_I^2) \leq E^n \leq (1 + \gamma)(\|u^{n+1}\|_I^2 + \|u^n\|_I^2),$$

且  $E^n \leq E^{n-1}$ , 得到完整能量稳定性。 □

精确解代入闭合, 分别在  $(x_0, t_n)$  对空间和时间展开:

$$u(x_1, t_{n\pm 1}) = u(x_0, t_n) + hu_x \pm \tau u_t + \frac{\tau^2}{2} u_{tt} + O(h^2 + h\tau + \tau^3).$$

两式相加消去奇次时间项, 故边界残量

$$R_0^n = u(x_0, t_n) - \frac{u(x_1, t_{n+1}) + u(x_1, t_{n-1})}{2} = O(h + \tau^2).$$

它不同于内部 leapfrog 的截断误差; 整体精度必须联合内部相容、边界残量与能量稳定性。

## 6.15 板书主线、真题映射与复习检查

| 分析链              | 必须能写出的中间步骤                              | 对应真题或同型题                            |
|------------------|-----------------------------------------|-------------------------------------|
| Fourier–Parseval | 网格 Fourier、Parseval、 $G^n$ 到 $\ell_h^2$ | 热方程与输运稳定题                           |
| 二次根/矩阵族          | $ b  \leq 1 + c \leq 2$ 、导数临界条件、Jordan  | Richardson–DFF、Schrödinger、leapfrog |
| 热方程              | Taylor、FTCS/BE/CN 放大因子、能量强制性            | 2024 秋 Q4；2026 春 Q5                 |
| 输运               | 迎/顺风因子、LF 网格路径、CFL 图、LW 推导              | 2024 春 Q4；2025 春 Q5；2026 春 Q4       |
| 变系数扩散            | 显式/CN 系数、半网格 Taylor、最大模误差               | 培训例题                                |
| 椭圆单调性            | M-矩阵条件、负最小点传播、比较函数                      | 2024 春 Q5；2024 秋 Q5；2026 春 Q5       |
| 能量与边界            | 四类内积、分部求和、Green、显式 $E^n$                | 2025 春 Q5；培训 leapfrog 例题            |

2023 秋 Q5 的三阶色散方程、2024 春 Q4 与 2025 秋 Q5 的波动方程真题仍在真题章完整保留，但不与本章 Schrödinger 分裂混同。一般 Poisson、有限元、谱方法、复杂边界及系统耗散/色散专题不扩为主线。

### 易错点

必须区分  $r = \kappa\tau/h^2$ 、 $\lambda = \tau/h^2$ 、 $\nu = a\tau/h$ ；只查谱半径会漏掉 Jordan 临界；DFF 无条件稳定不等于任意网格路径收敛；LF 的 CFL 之外还有  $h^2/\tau \rightarrow 0$ ；分部求和必须保留边界通量；半网格通量不能写成非散度型中心差分；出流闭合不是物理入流条件。

### 复习检查

能否从 Parseval 推到离散  $L^2$  稳定？能否证明实二次式根判据并说明重根半单性？能否写出矩阵族导数条件？能否完整证明 DFF 无条件稳定并给出相容路径？能否逐项求迎风、顺风、FTCS、LF 的放大因子？能否由依赖域图推出 CFL？能否从 Taylor 推导 LW？能否展开半网格通量并迭代误差？能否整理变系数 CN 的新旧层系数？能否写出四种内积、分部求和和 Green 公式？能否逐步构造热方程与有界区间 leapfrog 的显式能量？

## 6.16 板书范围的状态化阅读

本章主线严格对应最新 24 张照片合并出的 23 个独立黑板状态。审计对象不是“方法名称是否出现”，而是 30 个 `training_core` 下的 176 个公式级单元。最终映射维持

171 fully\_reproduced + 5 equivalently\_reproduced,

无 missing、partial、conflict 或 unreadable。等价复现指符号或推导顺序改变但数学链完整，不是只保留结论。

## 6.17 FTCS、后向 Euler 与 Crank–Nicolson 的统一 Fourier 推导

热方程  $u_t = au_{xx}$ , 记

$$\mu = \frac{a\tau}{h^2}, \quad \delta_{xx}U_j = U_{j+1} - 2U_j + U_{j-1}.$$

FTCS:

$$U_j^{n+1} = U_j^n + \mu \delta_{xx}U_j^n.$$

代入模态  $U_j^n = g^n e^{ij\theta}$  得

$$g_F = 1 - 4\mu \sin^2(\theta/2).$$

要求所有  $\theta$  下  $|g_F| \leq 1$ , 等价于

$$0 \leq \mu \leq \frac{1}{2}.$$

后向 Euler:

$$U_j^{n+1} = U_j^n + \mu \delta_{xx}U_j^{n+1}, \quad g_B = \frac{1}{1 + 4\mu \sin^2(\theta/2)},$$

对  $\mu \geq 0$  无条件稳定。

Crank–Nicolson:

$$U_j^{n+1} = U_j^n + \frac{\mu}{2}(\delta_{xx}U_j^{n+1} + \delta_{xx}U_j^n),$$

$$g_{CN} = \frac{1 - 2\mu \sin^2(\theta/2)}{1 + 2\mu \sin^2(\theta/2)}.$$

其绝对值不超过一, 但当  $\mu$  很大且高频  $\sin^2(\theta/2) \approx 1$  时  $g_{CN} \approx -1$ , 高频会交替而弱阻尼; “无条件稳定”不等于“所有步长都无振荡且准确”。

## 6.18 Richardson 与 Du Fort–Frankel 的三层根

Richardson 中心时间格式为

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\tau} = a \frac{\delta_{xx}U_j^n}{h^2}.$$

模态给

$$g^2 + 8\mu \sin^2(\theta/2)g - 1 = 0.$$

根积为  $-1$ , 若非零频率有一个根模小于一, 另一个必大于一, 因此不稳定。

Du Fort–Frankel 用

$$-2U_j^n \text{ 替换为 } -(U_j^{n+1} + U_j^{n-1})$$

进入空间差分, 得到

$$(1 + 2\mu)U_j^{n+1} = 4\mu \cos \theta U_j^n + (1 - 2\mu)U_j^{n-1}.$$

其二次放大根可对所有  $\mu \geq 0$  控制在闭单位圆内，但一致性展开包含

$$\frac{\tau^2}{h^2} u_{tt}$$

型项。因此除  $\tau, h \rightarrow 0$  外还需  $\tau/h \rightarrow 0$  才逼近原热方程。稳定与一致性必须分开报告。

## 6.19 输运格式的耗散与色散

对  $u_t + cu_x = 0$ ,  $\nu = c\tau/h$ 。当  $c > 0$ , 迎风格式

$$U_j^{n+1} = U_j^n - \nu(U_j^n - U_{j-1}^n)$$

的放大因子

$$g = 1 - \nu + \nu e^{-i\theta},$$

$$|g|^2 = 1 - 2\nu(1 - \nu)(1 - \cos \theta).$$

故  $0 \leq \nu \leq 1$  稳定。其修正方程含正人工扩散

$$u_t + cu_x = \frac{ch}{2}(1 - \nu)u_{xx} + \cdots,$$

解释数值耗散。

Lax-Wendroff 因子为

$$g = 1 - i\nu \sin \theta + \nu^2(\cos \theta - 1),$$

在  $|\nu| \leq 1$  稳定且二阶，但主要误差含三阶导，常表现为色散振荡。高阶不自动保持单调性。

## 6.20 二次根、半单性与一致幂有界

三层格式常得到

$$g^2 - a(\theta)g - b(\theta) = 0.$$

逐频率要求根落在闭单位圆还不够：若单位圆根重合且伴随矩阵不可对角化，则

$$A^n = \lambda^n(I + nN/\lambda)$$

线性增长。因而临界频率必须检查重根是否半单；对参数连续族还要控制接近临界时特征向量条件数，才能获得对网格频率一致的幂界。

Schrödinger 实虚分裂所得放大矩阵与波动方程三层格式结构不同。前者常利用旋转/酉结构，后者依赖离散能量与 CFL；二者不能因都出现  $2 \times 2$  矩阵就互换结论。

## 6.21 离散能量与边界通量

对有限区间, Fourier 模态一般不再对角化边界。定义

$$\langle U, V \rangle_h = h \sum_{j=1}^{J-1} U_j V_j, \quad \|D_+ U\|_h^2 = h \sum_{j=0}^{J-1} \left( \frac{U_{j+1} - U_j}{h} \right)^2.$$

离散分部求和为

$$\langle -\delta_{xx} U, V \rangle_h = \langle D_+ U, D_+ V \rangle_h - \text{边界通量}.$$

齐次 Dirichlet 条件使相应边界项消失; 非齐次或 Neumann 条件必须显式处理, 不能在求和中静默丢掉。

对波动方程中心格式, 乘以时间中心差并求和, 可构造离散动能加势能。CFL 条件保证能量中的交叉项可控制且整体正定, 从而得到

$$E^{n+1/2} \leq C E^{1/2}$$

或守恒式。能量法直接适配有限边界, 也是第三节板书新增内容不可省略的原因。

## 6.22 变系数通量形式与最大值原理

对

$$u_t = (a(x)u_x)_x, \quad a(x) \geq a_0 > 0,$$

自然离散为

$$(au_x)_x(x_j) \approx \frac{1}{h} \left[ a_{j+1/2} \frac{U_{j+1} - U_j}{h} - a_{j-1/2} \frac{U_j - U_{j-1}}{h} \right].$$

半网格系数应由物理通量与界面连续性选择; 强跳跃介质常采用调和平均。显式更新若可写成非负系数凸组合, 便得到离散最大值原理、唯一性与  $l^\infty$  稳定。系数非负条件同时产生步长限制。

### 易错点

谱方法、有限元、完整 Poisson 理论和三阶色散方程若出现在历史真题, 只保留在唯一真题解答与 boundary 映射中, 不回灌为当前培训主线。最新理论主线以 30 core/176 单元为准。

### 自测

1. 推导热方程三种基础格式的放大因子并比较阻尼。
2. 为什么 Richardson 的根积立即暗示不稳定?
3. Du Fort–Frankel 无条件稳定后还需什么网格比例保证一致性?
4. 单位圆重根为什么必须检查半单性?
5. 在有限区间能量估计中, 边界条件具体消去哪一项?

## 第七章 渐近分析

### 7.1 渐近序列、展开与误差

$\{\phi_n\}$  为渐近序列若  $\phi_{n+1} = o(\phi_n)$ 。  $f \sim \sum a_n \phi_n$  表示对每个固定  $N$ ，余项  $f - \sum_{n=0}^N a_n \phi_n = o(\phi_N)$ ；不意味着级数收敛。正则摄动在原域逐项有效，奇异摄动会改变最高阶或边界条件数量，需内外展开与匹配。

### 7.2 量纲分析、无量纲化与定量模型边界

设物理量  $q_1, \dots, q_n$  由  $r$  个基本量纲表示，量纲指数列组成矩阵  $D \in \mathbb{R}^{r \times n}$ 。无量纲单项式  $\Pi = \prod_{j=1}^n q_j^{a_j}$  的充要条件是  $Da = 0$ 。

**定理 7.1** (Buckingham  $\Pi$  定理的线性代数形式). 若量纲矩阵秩为  $s$ ，则存在  $n - s$  个线性无关的无量纲组合  $\Pi_1, \dots, \Pi_{n-s}$ 。任何量纲一致的关系可改写为这些组合间的关系；局部奇异或某些物理量为零的情形需另行检查。

*证明.* 无量纲指数向量正是  $\ker D$ ，由秩—零度定理其维数为  $n - s$ ，选一组基即得独立  $\Pi$  组。改变基本单位相当于沿  $D$  的行空间对  $\log q$  平移；量纲一致关系在这种缩放作用下不变，因而只依赖商空间  $\mathbb{R}^n / \text{row}(D)$  的坐标，也就是零空间基给出的无量纲组合。反过来，每个  $\Pi_j$  的指数向量属于  $\ker D$ ，所以改变基本单位时  $\Pi_j$  的缩放指数为  $Da_j = 0$ ，数值保持不变；因而任意  $F(\Pi_1, \dots, \Pi_{n-s}) = 0$  也保持不变。  $\square$

**例题 7.1** (阻力系数与 Reynolds 数). 球体阻力  $F$  若只依赖密度  $\rho$ 、速度  $U$ 、直径  $L$  和动力黏度  $\mu$ ，变量数为五，基本量纲秩为三，故有两个无量纲组。可取

$$C_D = \frac{F}{\rho U^2 L^2}, \quad \text{Re} = \frac{\rho U L}{\mu},$$

于是量纲分析只给  $C_D = \Phi(\text{Re})$ ，并不能决定  $\Phi$ 。后者仍需方程、渐近极限或实验；这正是定量建模的边界。

模型题应先列变量、单位和量纲矩阵，再求零空间；随后选择使主导项为  $O(1)$  的尺度，无量纲化方程并识别小参数，最后才作主导平衡、展开和有效域判断。只写一组  $\Pi$  而不检查独立性，或把量纲一致性误当作动力学定律，均不是完整答案。

### 7.3 Laplace 方法与 Watson 引理

若  $\phi$  在内点  $x_0$  有唯一非退化最大值、 $\phi''(x_0) < 0$ , 则

$$\int_a^b e^{\lambda\phi(x)} g(x) dx \sim e^{\lambda\phi(x_0)} g(x_0) \sqrt{\frac{2\pi}{\lambda|\phi''(x_0)|}}.$$

完整证明. 把积分分成  $x_0$  小邻域与其补集; 补集指数小一个固定量. 邻域内令  $x = x_0 + s/\sqrt{\lambda}$ , Taylor 展开  $\lambda[\phi(x) - \phi(x_0)] = \phi''(x_0)s^2/2 + O(|s|^3/\sqrt{\lambda})$ , 并用局部 Gaussian 控制作支配收敛, 得到主项.  $\square$

Watson 引理: 若  $f(t) \sim \sum a_n t^{\alpha+n-1}$  于  $0^+$ , 则  $\int_0^\infty e^{-\lambda t} f(t) dt \sim \sum a_n \Gamma(\alpha+n) \lambda^{-(\alpha+n)}$ , 需端点展开、可积控制和远端指数压低。

### 7.4 驻相、最速下降与 WKB

标准术语采用“驻相法”，括注考纲原词“稳定相位”。一维非退化驻点  $\phi'(x_0) = 0, \phi''(x_0) \neq 0$  给

$$\int e^{i\lambda\phi(x)} a(x) dx \sim a(x_0) e^{i\lambda\phi(x_0) + i\pi \operatorname{sgn} \phi''/4} \sqrt{\frac{2\pi}{\lambda|\phi''(x_0)|}}.$$

最速下降把复路径变形到相位虚部常、实部下降的路径; 必须说明解析域和不跨奇点。WKB 对  $\epsilon^2 y'' = Q(x)y$  远离 turning point 给  $y \sim Q^{-1/4} \exp(\pm \epsilon^{-1} \int \sqrt{Q} dx)$ ; 转折点需 Airy 连接。考试主要掌握 WKB 假设、主项推导和转折点失效原因, 连接公式以证明思路为主。

### 7.5 边界层与多尺度

外解满足降阶问题, 内变量  $X = (x - x_b)/\delta(\epsilon)$  由平衡确定; 匹配采用共同极限, 复合解  $y_{\text{comp}} = y_{\text{out}} + y_{\text{in}} - y_{\text{overlap}}$ . 多尺度令  $T_0 = t, T_1 = \epsilon t, \dots$ , 通过消除 secular term 得慢振幅方程, 从而得到大时间一致近似。

### 7.6 渐近符号、渐近序列与唯一性

当  $\epsilon \rightarrow 0$ ,  $f = O(g)$  表示  $|f/g|$  有界,  $f = o(g)$  表示  $f/g \rightarrow 0$ , 而  $f \sim g$  表示  $f/g \rightarrow 1$ 。若  $\phi_{n+1} = o(\phi_n)$ , 则  $\{\phi_n\}$  是渐近序列。

$$f(\epsilon) \sim \sum_{n=0}^{\infty} a_n \phi_n(\epsilon)$$

是指对每个固定  $N$ ,  $f - \sum_{n=0}^N a_n \phi_n = o(\phi_N)$ 。固定  $N$  后取极限是关键; 不能让  $N$  与  $\epsilon^{-1}$  同时任意增长。

**定理 7.2** (渐近展开系数唯一性). 相对于给定渐近序列, 若展开存在, 则系数逐项唯一。

**完整证明.** 若有两组系数, 设  $m$  是第一处不同. 两展开相减并除以  $\phi_m$ :

$$0 = (a_m - b_m) + \sum_{n>m}^N (a_n - b_n) \frac{\phi_n}{\phi_m} + o(1).$$

令极限趋于展开点, 后续比值和余项都趋零, 得  $a_m = b_m$ , 矛盾.  $\square$

渐近级数可以发散. 典型  $n!$  系数使各项先减小后增大, 在最小项附近截断常给最佳代数近似, 余项可为指数小. 超渐近展开与精细 Stokes 切换不属于考试主线, 掌握最小项截断的直观即可.

**例题 7.2** (发散但有效的展开).

$$\int_0^\infty \frac{e^{-t}}{1+\epsilon t} dt \sim \sum_{n=0}^\infty (-1)^n n! \epsilon^n.$$

逐项积分产生  $n!$ , 任意  $\epsilon \neq 0$  时级数发散, 但每个固定截断阶仍满足渐近定义.

## 7.7 主导平衡、正则与奇异摄动

正则摄动可在原变量和原区域逐项展开且最高阶结构不变; 奇异摄动会使最高阶系数消失、边界条件数量减少或极限不一致. 主导平衡先假定尺度, 再比较各项幂次, 只保留至少两项同阶的平衡.

**例题 7.3** (代数根的两种尺度).  $\epsilon x^2 + x - 1 = 0$  有正则根  $x = 1 - \epsilon + O(\epsilon^2)$ . 另一根为  $O(\epsilon^{-1})$ ; 令  $x = X/\epsilon$ , 得  $X^2 + X - \epsilon = 0$ , 从而  $x = -\epsilon^{-1} - 1 + O(\epsilon)$ . 单一正则级数会漏掉大根.

## 7.8 Watson 引理与端点贡献

**定理 7.3** (Watson 引理). 若  $f(t) \sim \sum_{n=0}^\infty a_n t^{n+\alpha-1}$  当  $t \downarrow 0$ ,  $\alpha > 0$ , 并有足够增长控制, 则

$$\int_0^\infty e^{-\lambda t} f(t) dt \sim \sum_{n=0}^\infty a_n \Gamma(n+\alpha) \lambda^{-(n+\alpha)}, \quad \lambda \rightarrow +\infty.$$

**完整证明.** 固定  $N$ , 把积分分成  $[0, \delta]$  和  $[\delta, \infty)$ . 近端代入有限展开, 以  $s = \lambda t$  缩放, 并把上限延到无穷; 误差指数小. 远端由  $e^{-\lambda \delta}$  与增长控制压到任意代数阶以下. 余项缩放后为  $o(\lambda^{-(N+\alpha)})$ .  $\square$

**例题 7.4** (端点 Laplace 积分).

$$I(\lambda) = \int_0^1 e^{-\lambda x} (1+x) dx \sim \lambda^{-1} + \lambda^{-2}.$$

精确式还含  $O(e^{-\lambda})$ , 说明指数小项不会在普通代数级数中出现.

## 7.9 Laplace 方法: 端点与内点

对  $I(\lambda) = \int_a^b e^{\lambda\phi(x)} g(x) dx$ , 先找  $\Re\phi$  的全局最大点并比较端点。若唯一内部最大点  $x_0$  非退化,  $\phi''(x_0) < 0$ , 则

$$I \sim e^{\lambda\phi(x_0)} g(x_0) \sqrt{\frac{2\pi}{\lambda|\phi''(x_0)|}}.$$

若最大值在左端且  $\phi'(a) < 0$ ,

$$I \sim e^{\lambda\phi(a)} \frac{g(a)}{\lambda|\phi'(a)|}.$$

多个同高最大点的贡献要相加; 若振幅在控制点为零或驻点退化, 需展开到首个非零项。

**例题 7.5** (内部最大点).

$$\int_{-\infty}^{\infty} e^{-\lambda(x-1)^2} (2+x) dx \sim 3\sqrt{\pi/\lambda}.$$

这里  $\phi''(1) = -2$ ,  $g(1) = 3$ ; 奇函数修正积分为零, 所以该主项还是精确值。

## 7.10 非驻相衰减与驻相法

若  $\phi'$  不为零, 则

$$\int_a^b e^{i\lambda\phi} g dx = \left[ \frac{e^{i\lambda\phi} g}{i\lambda\phi'} \right]_a^b - \frac{1}{i\lambda} \int_a^b e^{i\lambda\phi} \left( \frac{g}{\phi'} \right)' dx,$$

通常为  $O(\lambda^{-1})$ , 可反复分部积分。

**定理 7.4** (一维非退化驻相主项). 若  $x_0$  是内部唯一驻相点,  $\phi'(x_0) = 0$ ,  $\phi''(x_0) \neq 0$ , 则

$$\int e^{i\lambda\phi(x)} g(x) dx \sim e^{i\lambda\phi(x_0)} e^{i\frac{\pi}{4} \operatorname{sgn} \phi''(x_0)} g(x_0) \sqrt{\frac{2\pi}{\lambda|\phi''(x_0)|}}.$$

### 证明思路

用截断函数分离驻点邻域与非驻相区; 后者分部积分。驻点邻域令  $x - x_0 = \lambda^{-1/2}y$ , 相位保留二次项, 归结为 Fresnel 积分。其平方根支路给  $e^{\pm i\pi/4} \sqrt{2\pi}$ 。

**例题 7.6** (驻相相位因子).

$$\int_{-\infty}^{\infty} e^{i\lambda x^2/2} \chi(x) dx \sim e^{i\pi/4} \sqrt{2\pi/\lambda},$$

其中  $\chi$  在零附近为 1 且紧支撑。相位改成负号时因子变为  $e^{-i\pi/4}$ 。

## 7.11 最速下降法的操作链

对复积分  $\int_{\Gamma} e^{\lambda\phi(z)} g(z) dz$ :

1. 找鞍点  $\phi'(z_s) = 0$ ;
2. 选  $\Im\phi$  常数且  $\Re\phi$  下降的路径;

3. 在不跨越极点和 branch cut 的前提下变形轮廓;
4. 在可达鞍点作二次 Taylor 展开;
5. 选方向与平方根支路, 求 Gaussian 局部贡献, 并比较端点和多鞍点。

“找到鞍点”不等于“该鞍点贡献”; 可达性和 Stokes 切换决定贡献。考试主要要求检查路径可变形性和支路选择, 不要求超渐近层面的 Stokes 乘子理论。

## 7.12 正则摄动、可解性与特征值修正

将  $y = y_0 + \epsilon y_1 + \dots$  代入含参问题逐阶求解。若  $L_0$  有零空间,  $L_0 y_1 = r_1$  可解需  $r_1 \perp \ker L_0^*$ , 这是消除共振、特征值修正和慢幅方程的统一 Fredholm 条件。

对 Hermitian  $A(\epsilon) = A_0 + \epsilon A_1 + \dots$  的简单归一化特征对,

$$\lambda_1 = v_0^* A_1 v_0.$$

它来自一阶方程左乘  $v_0^*$ 。重特征值要先在特征子空间内对扰动压缩矩阵对角化。

**例题 7.7** (简单特征值修正).  $A(\epsilon) = \text{diag}(1, 2) + \epsilon \begin{bmatrix} 3 & 1 \\ 1 & 0 \end{bmatrix}$ 。对  $\lambda_0 = 1$ , 一阶修正为 3; 对  $\lambda_0 = 2$ , 一阶修正为 0。非对角项一阶改变特征向量、二阶才影响特征值。

## 7.13 Poincaré–Lindstedt 与多重尺度

弱非线性振子的正则展开会产生  $t \sin t$  secular 项。Poincaré–Lindstedt 令  $s = \omega t$ ,  $\omega = \omega_0 + \epsilon \omega_1 + \dots$ , 逐阶选频率修正消除共振。多重尺度令  $T_0 = t, T_1 = \epsilon t, \dots$ ,

$$\frac{d}{dt} = D_0 + \epsilon D_1 + \dots, \quad \frac{d^2}{dt^2} = D_0^2 + 2\epsilon D_0 D_1 + \dots.$$

高阶可解性条件给慢振幅和相位方程。2025 春/秋 Q6 与 2026 春 Q6 分别检验阻尼幅值、Rayleigh 振子和周期修正。

## 7.14 WKB、转向点与量子化

对  $\epsilon^2 y'' = q(x)y$ , 设  $y = \exp(S_0/\epsilon + S_1 + \dots)$ 。代入得到

$$(S_0')^2 = q, \quad 2S_0' S_1' + S_0'' = 0.$$

当  $q > 0$ ,

$$y \sim q^{-1/4} \left( C_+ e^{\epsilon^{-1} \int^x \sqrt{q} ds} + C_- e^{-\epsilon^{-1} \int^x \sqrt{q} ds} \right);$$

$q < 0$  时为振荡形式。简单转向点  $q = 0$  需 Airy 统一化; 连接公式只要求理解匹配思路和相位跳跃。两转向点相位匹配产生 Bohr–Sommerfeld 条件。

7.15 边界层、匹配与复合展开

对

$$\epsilon y'' + a(x)y' + b(x)y = f(x)$$

外问题降阶后只能满足一个边界条件。层位置由衰减方向决定；令  $X = (x - x_b)/\delta$ ，平衡  $\epsilon/\delta^2$  与  $a/\delta$  常给  $\delta = O(\epsilon)$ ，一阶项退化时可能是  $O(\sqrt{\epsilon})$ 。匹配要求内解远场与外解在重叠区的再展开一致，复合式

$$y_{\text{comp}} = y_{\text{out}} + Y_{\text{in}} - y_{\text{overlap}}.$$

例题 7.8 (一阶边界层).

$$\epsilon y' + y = 0, \quad y(0) = 1$$

的解为  $e^{-x/\epsilon}$ 。外解为 0；左端内变量  $X = x/\epsilon$  给内解  $e^{-X}$ 。固定  $x > 0$  的所有代数阶都为零，却不能满足边界条件，这就是奇异摄动。

7.16 方法选择、真题迁移与检查

| 方法                   | 控制点/尺度                   | 主项量级                               | 高频错误       |
|----------------------|--------------------------|------------------------------------|------------|
| Watson/端点<br>Laplace | 端点, $O(\lambda^{-1})$    | $\lambda^{-\alpha}$                | 漏 Gamma 系数 |
| 内部 Laplace           | 最大点, $O(\lambda^{-1/2})$ | $e^{\lambda\phi_0} \lambda^{-1/2}$ | 漏多个最大点     |
| 驻相                   | 实驻点, $O(\lambda^{-1/2})$ | $\lambda^{-1/2} e^{\pm i\pi/4}$    | 相位符号       |
| 最速下降                 | 可达复鞍点                    | Gaussian 局部贡献                      | 轮廓/支路不合法   |
| 匹配渐近                 | 内外重叠区                    | 由平衡决定                              | 忘减 overlap |

历史迁移：2023 秋 Q7 是 Sturm–Liouville 一阶修正；2024 春 Q6 是边界层复合式；2025 春/秋 Q6 是多尺度；2026 春 Q6 是闭轨与周期。量纲分析/Buckingham II 只要求会选基本量纲、构造无量纲组并检查独立性，不扩展成一般建模专题。

易错点

不要把渐近展开当收敛级数；不要只找驻点而漏端点；不要在最速下降中无条件移动轮廓；不要对重特征值套简单公式；不要用外解强行满足全部边界；不要写内 + 外而忘减重复项。

复习检查

能否证明渐近系数唯一性和 Watson 引理？能否区分端点/内部 Laplace 与非驻相/驻相？能否写出最速下降五步？能否用 Fredholm 条件求特征值或频率修正？能否推导 WKB 的 eikonal/transport 方程并构造匹配复合解？

### 7.17 渐近展开的运算规则与失效边界

$f(\varepsilon) \sim g(\varepsilon)$  只表示比值趋一, 不表示差的绝对大小。若

$$f(\varepsilon) = \sum_{k=0}^N a_k \phi_k(\varepsilon) + o(\phi_N), \quad \phi_{k+1} = o(\phi_k),$$

则它是关于渐近序列  $\{\phi_k\}$  的 Poincaré 展开。展开通常不要求级数收敛; 截断到最小项附近反而可能给最佳近似。

乘除与复合前必须检查主项不为零、复合函数在有关点足够光滑且误差在所需区域一致。点态展开

$$\frac{1}{1+\varepsilon x} = 1 - \varepsilon x + O(\varepsilon^2 x^2)$$

在  $x = O(1/\varepsilon)$  上不一致, 不能直接在无限区间逐项积分。

### 7.18 Watson 引理的缩放证明

设  $f(t) \sim \sum_{n \geq 0} a_n t^{\lambda_n - 1}$  当  $t \downarrow 0$ , 且积分尾部可控。考虑

$$I(x) = \int_0^\infty e^{-xt} f(t) dt, \quad x \rightarrow \infty.$$

令  $s = xt$ :

$$I(x) = \frac{1}{x} \int_0^\infty e^{-s} f(s/x) ds.$$

形式代入局部展开给

$$I(x) \sim \sum_{n \geq 0} a_n \Gamma(\lambda_n) x^{-\lambda_n}.$$

严格证明把积分分为  $[0, \delta]$  与  $[\delta, \infty)$ : 前段用带余项的局部展开和支配估计, 后段由  $e^{-x\delta}$  指数小控制。不能只说“逐项积分”, 因为展开只在端点附近有效。

### 7.19 Laplace 方法的内点最大值

若  $\phi$  在  $x_0$  有唯一非退化最大值,

$$\phi'(x_0) = 0, \quad \phi''(x_0) < 0,$$

且  $a(x_0) \neq 0$ , 则

$$I(\lambda) = \int_a^b a(x) e^{\lambda \phi(x)} dx.$$

在  $x_0$  附近令

$$\phi(x) = \phi(x_0) - \frac{|\phi''(x_0)|}{2} (x - x_0)^2 + O(|x - x_0|^3),$$

尺度  $x - x_0 = y/\sqrt{\lambda}$  给

$$I(\lambda) \sim a(x_0)e^{\lambda\phi(x_0)}\sqrt{\frac{2\pi}{\lambda|\phi''(x_0)|}}.$$

若最大值在端点且  $\phi'(a) \neq 0$ , 主尺度为  $1/\lambda$ , 系数与内点高斯型不同。

## 7.20 驻相法与非驻相端点展开

对

$$I(\lambda) = \int_a^b e^{i\lambda\phi(x)} g(x) \, dx,$$

假设  $|\phi'(x)| \geq c > 0$ . 定义反复分部积分算子

$$A_0 g = g, \quad A_{j+1} g = \left( \frac{A_j g}{\phi'} \right)'.$$

**定理 7.5** (非驻相积分的有限阶端点渐近展开). 若  $g \in C^{N+1}[a, b]$ 、 $\phi \in C^{N+2}[a, b]$  且  $|\phi'| \geq c > 0$ , 则当  $\lambda \rightarrow +\infty$  时

$$I(\lambda) = \sum_{j=0}^{N-1} \frac{(-1)^j}{(i\lambda)^{j+1}} \left[ e^{i\lambda\phi(x)} \frac{A_j g(x)}{\phi'(x)} \right]_{x=a}^{x=b} + R_N(\lambda), \quad R_N(\lambda) = O(\lambda^{-N-1}).$$

每一阶都同时含左右两个端点; 若某个端点的相应振幅系数为零, 该端点的首个非零贡献自动移到更高阶。

完整证明. 利用

$$e^{i\lambda\phi(x)} = \frac{1}{i\lambda\phi'(x)} \frac{d}{dx} e^{i\lambda\phi(x)}$$

作一次分部积分:

$$I(\lambda) = \frac{1}{i\lambda} \left[ e^{i\lambda\phi} \frac{g}{\phi'} \right]_a^b - \frac{1}{i\lambda} \int_a^b e^{i\lambda\phi} A_1 g \, dx.$$

对余下积分重复这一恒等式  $N-1$  次, 得到所列前  $N$  个端点项以及

$$\tilde{R}_N(\lambda) = \frac{(-1)^N}{(i\lambda)^N} \int_a^b e^{i\lambda\phi(x)} A_N g(x) \, dx.$$

再对最后积分分部一次。由于  $|\phi'| \geq c$  且给定光滑性保证  $A_N g/\phi'$  及其导数在紧区间上有界,

$$\left| \int_a^b e^{i\lambda\phi} A_N g \, dx \right| \leq \frac{C}{\lambda}.$$

因此  $\tilde{R}_N = O(\lambda^{-N-1})$ , 证明完成。若只假设到  $A_N g \in L^1$ , 通常只能得到  $o(\lambda^{-N})$  或较弱余项, 不能无条件写成上述阶数。□

**例题 7.9** (可直接核对的两端点展开). 取  $\phi(x) = x$ 、 $g(x) = 1+x$ 、区间  $[0, 1]$ 。此时  $A_0 g =$

$1+x$ ,  $A_1g=1$ ,  $A_2g=0$ , 故展开在两项后精确终止:

$$\int_0^1 (1+x)e^{i\lambda x} dx = \frac{2e^{i\lambda} - 1}{i\lambda} - \frac{e^{i\lambda} - 1}{(i\lambda)^2}.$$

右端与直接积分所得表达式一致。若改取  $g(x) = x(1-x)$ , 则  $g(0) = g(1) = 0$ , 整个  $O(\lambda^{-1})$  端点项消失, 首项从  $\lambda^{-2}$  开始; 这说明“振幅在端点消失”会改变主阶。

若有非退化驻点  $\phi'(x_0) = 0, \phi''(x_0) \neq 0$ , 主贡献为

$$I(\lambda) \sim a(x_0)e^{i\lambda\phi(x_0)}e^{i\frac{\pi}{4}\text{sgn}\phi''(x_0)}\sqrt{\frac{2\pi}{\lambda|\phi''(x_0)|}}.$$

相位因子来自 Fresnel 高斯积分, 不能按实 Laplace 方法删去。

## 7.21 边界层匹配的完整例子

考虑

$$\varepsilon y'' + y' = 0, \quad y(0) = 0, \quad y(1) = 1.$$

令  $\varepsilon = 0$  的外问题  $y'_0 = 0$ , 由右端条件得

$$y_{\text{out}} = 1.$$

它不能满足左端条件, 故  $x = 0$  附近有边界层。令  $\xi = x/\varepsilon$ , 内方程

$$Y_{\xi\xi} + Y_{\xi} = 0,$$

满足  $Y(0) = 0$  且与外解匹配  $Y(\infty) = 1$ , 故

$$Y_{\text{in}} = 1 - e^{-\xi}.$$

共同部分为 1, 复合展开

$$y_{\text{comp}} = y_{\text{out}} + Y_{\text{in}} - 1 = 1 - e^{-x/\varepsilon}.$$

精确解为

$$y(x) = \frac{1 - e^{-x/\varepsilon}}{1 - e^{-1/\varepsilon}},$$

两者差指数小。边界层位置由降阶外问题缺失哪一个边界条件决定。

## 7.22 多尺度消除久期项

弱振子若普通摄动产生  $\varepsilon t \sin t$  久期项, 展开只在  $t \ll 1/\varepsilon$  有效。引入慢时间

$$T = \varepsilon t, \quad \frac{d}{dt} = \partial_t + \varepsilon \partial_T,$$

令振幅与相位随  $T$  演化，通过消除与齐次解共振的强迫项得到调制方程。Poincaré–Lindstedt 则重标时间  $\tau = \omega(\epsilon)t$ ，通过频率修正消除久期项。两者目的相同，但适用结构与输出不同。

### 答题方法

渐近积分先定位主贡献：端点、内点驻点、鞍点还是奇点；摄动方程先判断是否降阶、是否共振、是否有多个尺度。随后再选 Watson、Laplace、驻相、最速下降、边界层或多尺度，不能只看公式外形。

### 自测

1. 为什么点态 Taylor 展开不能自动在无限区间逐项积分？
2. Watson 引理证明为什么要把积分分成近端与远端？
3. 比较 Laplace 内点最大值与端点最大值的尺度。
4. 完成边界层例子的匹配与复合展开。
5. 久期项说明普通摄动在哪个时间尺度失效？

## 第八章 优化：凸性、对偶、极集与算法

### 8.1 凸性、最优性与梯度下降

可微凸函数满足  $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ ;  $\mu$ -强凸再加  $\mu \|y - x\|^2 / 2$ 。  $L$ -smooth 给下降引理。无约束可微问题一阶必要条件  $\nabla f = 0$ ; 凸时充分; Hessian 正定是严格局部极小的充分条件而非一般必要条件。

若  $f$  凸且  $L$ -smooth, GD 步长  $1/L$  给  $f(x_k) - f^* \leq L \|x_0 - x^*\|^2 / (2k)$ ; 若再  $\mu$ -强凸, 则函数值/距离线性收敛。Newton 解  $\nabla^2 f(x_k) p_k = -\nabla f(x_k)$ , 局部二次率需 Hessian 在最优点正定且 Lipschitz、初值足够近; 全局需线搜索/信赖域。

### 8.2 约束、Lagrange、对偶与 KKT

对  $\min f_0(x)$  s.t.  $f_i(x) \leq 0, h_j(x) = 0, L = f_0 + \sum \lambda_i f_i + \sum \nu_j h_j, \lambda \geq 0$ ; 对偶函数  $g = \inf_x L$  总给弱对偶。凸问题若有 Slater 点且等式仿射, 强对偶和 KKT 的充分必要性成立。KKT 包含原始可行、对偶可行、互补松弛和驻点; 非凸或无资格条件时不能无条件称必要充分。

### 8.3 投影、近端、罚函数与算法边界

闭凸集投影满足  $\langle z - \Pi_C z, y - \Pi_C z \rangle \leq 0$ , 且 firm nonexpansive。投影梯度是  $x_{k+1} = \Pi_C(x_k - \alpha \nabla f(x_k))$ ; 近端梯度把非光滑项  $g$  用  $\text{prox}_{\alpha g}$  处理。罚函数把违反约束加入目标; 障碍法在可行域内部趋近边界。单纯形和内点法掌握基本步骤与最优性证书; 同伦与动态规划只作考纲入口, 不扩展高级实现。

### 8.4 凸集、凸函数、强凸与光滑

集合  $C$  凸是指对任意  $x, y \in C, \theta \in [0, 1], \theta x + (1 - \theta)y \in C$ 。凸锥还对非负缩放封闭。函数  $f$  凸当且仅当其上图集  $\{(x, t) : t \geq f(x)\}$  凸。可微函数在凸开域上凸当且仅当

$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

二阶可微时, 凸等价于 Hessian 半正定。  $\mu$ -强凸要求

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2.$$

$L$ -光滑是梯度 Lipschitz:  $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$ 。

**定理 8.1** (下降引理与共梯度不等式). 若  $f$  的梯度  $L$ -Lipschitz, 则

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2.$$

若  $f$  还凸, 则

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

#### 证明思路

第一式沿线段积分  $f(y) - f(x) = \int_0^1 \nabla f(x + t(y - x))^T(y - x) dt$ , 用 Lipschitz 界积分。第二式对  $g(z) = f(z) - \nabla f(y)^T z$  应用第一式, 并在  $z = x - \nabla f(y)/L$  处比较其全局最小值下界, 再交换  $x, y$  合并。

强凸给唯一极小点和  $f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|^2$ ;  $L$ -光滑给  $f(x) - f^* \leq \|\nabla f(x)\|^2/(2\mu)$  的相关梯度支配界时必须说明强凸。强凸、严格凸和 Hessian 正定并非同义:  $x^4$  严格凸但在零点 Hessian 为零, 也不是全局强凸。

**例题 8.1** (判断凸、强凸与光滑).  $f(x) = \frac{1}{2}x^T Ax - b^T x$ 。若  $A = A^T \succeq 0$  则凸; 若  $A \succeq \mu I$  则  $\mu$ -强凸; 梯度 Lipschitz 常数可取  $L = \lambda_{\max}(A)$ 。条件数  $L/\mu$  决定一阶法速度。

## 8.5 一阶/二阶最优性与次梯度

无约束可微局部极小点必须满足  $\nabla f(x^*) = 0$ ; 二阶必要条件为  $\nabla^2 f(x^*) \succeq 0$ 。若梯度为零且 Hessian 正定, 则为严格局部极小。证明沿任意方向  $d$  研究  $\varphi(t) = f(x^* + td)$  并用一维 Taylor 展开。

凸函数的次微分

$$\partial f(x) = \{g : f(y) \geq f(x) + g^T(y - x), \forall y\}.$$

凸问题的全局最优充要条件是  $0 \in \partial f(x^*)$ 。例如

$$\partial|x| = \begin{cases} \{1\}, & x > 0, \\ [-1, 1], & x = 0, \\ \{-1\}, & x < 0. \end{cases}$$

这直接产生软阈值和  $\ell_1$  KKT 条件。

**例题 8.2** (次梯度最优性). 最小化  $\frac{1}{2}(x - a)^2 + \lambda|x|$ 。当  $x \neq 0$  解  $x - a + \lambda \operatorname{sgn} x = 0$ ; 当  $x = 0$  需  $a \in \lambda[-1, 1]$ 。故

$$x^* = \operatorname{sgn}(a) \max\{|a| - \lambda, 0\}.$$

这也是 2024 春 Q7 的核心软阈值链。

## 8.6 梯度下降的两个基本收敛率

梯度下降  $x_{k+1} = x_k - \alpha \nabla f(x_k)$ 。取  $\alpha = 1/L$ ，下降引理给

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2.$$

**定理 8.2** (凸光滑 GD 的  $O(1/k)$ ). 若  $f$  凸、梯度  $L$ -Lipschitz、极小点  $x^*$  存在, 且  $\alpha = 1/L$ , 则

$$f(x_k) - f^* \leq \frac{L\|x_0 - x^*\|^2}{2k}.$$

完整证明. 由更新和凸性,

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 - \frac{2}{L} \nabla f(x_k)^T (x_k - x^*) + \frac{1}{L^2} \|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - \frac{2}{L} (f(x_k) - f^*) + \frac{1}{L^2} \|\nabla f(x_k)\|^2. \end{aligned}$$

下降引理又给  $\|\nabla f(x_k)\|^2/(2L) \leq f(x_k) - f(x_{k+1})$ 。代入并整理:

$$\frac{2}{L} (f(x_{k+1}) - f^*) \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2.$$

求和并用函数值单调, 得到结论。 □

**定理 8.3** (强凸光滑 GD 的线性收敛). 若  $f$  为  $\mu$ -强凸且  $L$ -光滑, 取  $\alpha = 1/L$ , 则

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*).$$

完整证明. 下降式给  $f(x_{k+1}) \leq f(x_k) - \|\nabla f(x_k)\|^2/(2L)$ 。强凸函数满足  $\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*)$ , 可由强凸下界对  $y$  最小化得到。代入即  $f(x_{k+1}) - f^* \leq (1 - \mu/L)(f(x_k) - f^*)$ , 归纳完成。 □

一般固定步长  $0 < \alpha < 2/L$  保证凸光滑情形下降; 二次型上最优固定步长  $2/(L + \mu)$  的误差谱半径为  $(L - \mu)/(L + \mu)$ 。收敛率必须与函数类、步长、度量 (函数值/距离/梯度) 一起陈述。

**例题 8.3** (二次型 GD 手算).  $f(x) = \frac{1}{2}(4x_1^2 + x_2^2)$ ,  $L = 4, \mu = 1$ 。取  $\alpha = 1/4$ ,  $x_{k+1} = (0, \frac{3}{4}x_{2,k})$ : 第一分量一步消失, 第二分量线性因子  $3/4$ 。通用函数值界因子  $1 - \mu/L = 3/4$  是上界, 实际函数值第二分量因子为  $9/16$ 。

## 8.7 线搜索、Newton、拟 Newton 与信赖域

下降方向  $p_k$  满足  $\nabla f_k^T p_k < 0$ 。Armijo 条件

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k$$

保证充分下降; Wolfe 再要求  $\nabla f(x_k + \alpha p_k)^T p_k \geq c_2 \nabla f_k^T p_k$ ,  $0 < c_1 < c_2 < 1$ 。强 Wolfe 把曲率条件加强为

$$|\nabla f(x_k + \alpha p_k)^T p_k| \leq c_2 |\nabla f_k^T p_k|.$$

仅有函数下降不足以推出全局收敛。

**定理 8.4** (Armijo 回溯与 Wolfe/强 Wolfe 步长存在性). 设  $\nabla f$  在当前水平集邻域内以常数  $L$  Lipschitz, 且  $g_k^\top p_k < 0$ 。若  $0 < c_1 < 1$ , 则充分小的  $\alpha > 0$  都满足 Armijo 条件, 因此以固定因子  $\beta \in (0, 1)$  回溯必在有限步停止。若进一步沿射线  $\phi(\alpha) = f(x_k + \alpha p_k)$  有下界、 $\phi'$  连续且  $0 < c_1 < c_2 < 1$ , 则存在步长同时满足 Armijo 与普通 Wolfe 条件, 也存在步长满足 Armijo 与强 Wolfe 条件。

证明. 下降引理给

$$f(x_k + \alpha p_k) \leq f(x_k) + \alpha g_k^\top p_k + \frac{L}{2} \alpha^2 \|p_k\|^2.$$

只要

$$0 < \alpha \leq \frac{2(1 - c_1)(-g_k^\top p_k)}{L\|p_k\|^2},$$

右端不超过  $f(x_k) + c_1 \alpha g_k^\top p_k$ , 故回溯有限终止。

对 Wolfe 条件, 令  $\psi(\alpha) = \phi(\alpha) - \phi(0) - c_1 \alpha \phi'(0)$ 。有  $\psi'(0) = (1 - c_1)\phi'(0) < 0$ , 而  $\phi$  沿射线有下界, 所以  $\psi$  不可能永远严格下降。取首次使  $\psi$  回到零的点  $\bar{\alpha} > 0$ , 在区间内选取  $\psi$  的极小点  $\alpha_*$ 。此时 Armijo 成立且

$$0 = \psi'(\alpha_*) = \phi'(\alpha_*) - c_1 \phi'(0),$$

从而

$$\phi'(\alpha_*) = c_1 \phi'(0), \quad |\phi'(\alpha_*)| = c_1 |\phi'(0)| < c_2 |\phi'(0)|.$$

因此同一个  $\alpha_*$  不仅满足普通 Wolfe, 也满足强 Wolfe。这里的存在性论证不等于一个数值算法; 实际搜索还需下面的 bracketing/zoom 过程维持区间不变量。□

### 答题方法

**强 Wolfe 的 bracketing/zoom.** 令  $\phi(\alpha) = f(x_k + \alpha p_k)$ ,  $\phi'(0) < 0$ 。

1. 从递增试探步长开始。若 Armijo 失败, 或新函数值不低于前一试探点, 则当前两点围成区间, 进入 zoom。
2. 若试探点满足  $|\phi'(\alpha)| \leq c_2 |\phi'(0)|$ , 立即接受。
3. 若  $\phi'(\alpha) \geq 0$ , 则导数已经换号, 以当前点和前一试探点为区间进入 zoom; 否则继续扩大步长。
4. zoom 在区间内用二次/三次插值或受保护的二分点试探。若 Armijo 失败或函数值过高, 缩短高端; 否则检查强曲率条件, 并根据导数符号调整区间方向。始终保持低端满足 Armijo, 且区间内含一个满足强 Wolfe 的点。

在  $\phi'$  连续、射线有下界并使用受保护的区间收缩时, 区间长度趋零; 结合上面的存在性与区间不变量, 可得到满足强 Wolfe 的步长。若只做无保护高阶插值, 试探点可能反复贴近端点, 不能据此声称算法必然终止。

**例题 8.4** (一次可手算的强 Wolfe 搜索). 取  $f(x) = x^2/2$ ,  $x_0 = 1$ ,  $p = -1$ ,  $c_1 = 0.1$ ,  $c_2 = 0.5$ 。

沿线

$$\phi(\alpha) = \frac{1}{2}(1 - \alpha)^2, \quad \phi'(\alpha) = \alpha - 1.$$

试探  $\alpha = 2$  时,  $\phi(2) = 1/2$  大于 Armijo 右端  $1/2 - 0.2 = 0.3$ , 故区间  $[0, 2]$  被括住。zoom 取中点  $\alpha = 1$ , 此时  $\phi(1) = 0$  满足 Armijo, 且  $|\phi'(1)| = 0 \leq 0.5|\phi'(0)|$ , 故接受。这个例子真实执行了 bracketing 与 zoom, 而不是用一次 Newton/BFGS 更新替代线搜索。

Newton 解

$$\nabla^2 f(x_k)p_k = -\nabla f(x_k), \quad x_{k+1} = x_k + p_k.$$

若  $x^*$  附近 Hessian Lipschitz、 $\nabla^2 f(x^*)$  正定、初值足够近, 则局部二次; 远离极小点 Hessian 可不定, 需阻尼、修正 Hessian 或信赖域。

BFGS 逆 Hessian 更新为

$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T, \quad \rho_k = (y_k^T s_k)^{-1}.$$

若  $H_k \succ 0$  且曲率  $y_k^T s_k > 0$ , 则  $H_{k+1} \succ 0$ 。Wolfe 线搜索可在凸邻域保证曲率条件。局部超线性还需 Hessian 的正则性、迭代收敛到非退化解以及 Dennis–Moré 型逼近条件, 不能写成无条件结论。

**例题 8.5** (Newton 与 BFGS 一步). 对二次型  $f(x) = \frac{1}{2}x^T A x - b^T x$ ,  $A \succ 0$ , 精确 Newton  $p = -A^{-1}(Ax - b)$  一步到达  $A^{-1}b$ 。BFGS 若从  $H_0 = I$  起步不会一般一步到达, 但在精确线搜索下逐步建立割线信息; 曲率条件是保持正定的核心。

信赖域在  $\|p\| \leq \Delta_k$  内近似最小化  $m_k(p) = f_k + g_k^T p + \frac{1}{2}p^T B_k p$ , 用

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}$$

比较实际/预测下降并更新半径。它允许在  $B_k$  不定时寻找负曲率方向, 比“强行解 Newton 方程”更稳健。

## 8.8 投影、近端与复合优化

闭凸集投影  $p = \Pi_C(z)$  的充要条件为

$$(z - p)^T (y - p) \leq 0, \quad \forall y \in C.$$

**完整证明.**  $p$  最小化  $\frac{1}{2}\|y - z\|^2$ 。对任意  $y \in C$ ,  $p + t(y - p) \in C$ ; 单变量函数在  $t = 0$  右导数非负, 得到不等式。反之展开平方, 交叉项非负, 故  $p$  距离最小。□

由两次代入可证明投影 firmly nonexpansive。投影梯度  $x_{k+1} = \Pi_C(x_k - \alpha \nabla f(x_k))$  的固定点条件等价于  $0 \in \nabla f(x) + N_C(x)$ 。

近端算子

$$\text{prox}_{\alpha g}(v) = \arg \min_x \left\{ g(x) + \frac{1}{2\alpha} \|x - v\|^2 \right\}.$$

对复合问题  $F = f + g$ ,  $f$  光滑凸、 $g$  闭凸, 近端梯度

$$x_{k+1} = \text{prox}_{\alpha g}(x_k - \alpha \nabla f(x_k))$$

在  $0 < \alpha \leq 1/L$  下有函数值  $O(1/k)$ 。  $g = \lambda \|\cdot\|_1$  的 prox 是逐分量软阈值。

## 8.9 约束资格、KKT、对偶与 Slater

问题

$$\min f_0(x) \quad \text{s.t.} \quad f_i(x) \leq 0, \quad i = 1, \dots, m, \quad h_j(x) = 0$$

的 Lagrangian 为  $L = f_0 + \sum_i \lambda_i f_i + \sum_j \nu_j h_j$ ,  $\lambda_i \geq 0$ 。KKT 条件:

$$\begin{cases} f_i(x^*) \leq 0, & h_j(x^*) = 0, \\ \lambda_i^* \geq 0, \\ \lambda_i^* f_i(x^*) = 0, \\ \nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) + \sum_j \nu_j^* \nabla h_j(x^*) = 0. \end{cases}$$

在一般非凸问题中, KKT 必要性需要 LICQ、MFCQ 等约束资格; 它通常不充分。凸问题若  $f_i$  凸、等式仿射, 任一满足 KKT 的点全局最优。若还有 Slater 严格可行点, 则强对偶成立并在适当闭性条件下存在乘子。

**定理 8.5** (弱对偶). 对偶函数  $g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$ 。任意原始可行  $x$  和对偶可行  $\lambda \geq 0$  都满足  $g(\lambda, \nu) \leq f_0(x)$ 。

完整证明. 原始可行给  $f_i(x) \leq 0, h_j(x) = 0$ , 对偶可行给  $\sum_i \lambda_i f_i(x) \leq 0$ , 故  $L(x, \lambda, \nu) \leq f_0(x)$ 。又  $g(\lambda, \nu) = \inf_z L(z, \lambda, \nu) \leq L(x, \lambda, \nu)$ , 合并即得。□

**例题 8.6** (KKT 手算)。

$$\min_x \frac{1}{2}(x-2)^2 \quad \text{s.t.} \quad x \leq 1.$$

令  $g(x) = x - 1 \leq 0$ 。KKT:  $x - 2 + \lambda = 0, \lambda \geq 0, x \leq 1, \lambda(x - 1) = 0$ 。无约束解 2 不可行, 故约束活跃,  $x^* = 1, \lambda^* = 1$ 。若把不等式写成  $1 - x \geq 0$  却仍使用同一乘子号, 会得到错误符号。

## 8.10 线性规划、极集与掌握层级

标准型 LP  $\min c^T x$  s.t.  $Ax = b, x \geq 0$  的对偶为  $\max b^T y$  s.t.  $A^T y \leq c$ 。基矩阵  $A_B$  可逆给基本解; 若基本变量非负即基本可行解 (BFS)。

**定理 8.6** (基本可行解等价于顶点). 设  $A \in \mathbb{R}^{m \times n}$  满行秩,  $P = \{x : Ax = b, x \geq 0\}$ 。则  $x \in P$  是基本可行解, 当且仅当  $x$  是多面体  $P$  的顶点。

证明. 若  $x$  是某个基  $B$  对应的 BFS, 且  $x = \frac{1}{2}(y+z), y, z \in P$ , 则所有非基坐标满足  $0 = x_N = (y_N + z_N)/2$ 。由非负性  $y_N = z_N = 0$ , 再由  $A_B y_B = A_B z_B = b$  与  $A_B$  可逆, 得  $y = z = x$ , 故  $x$  是顶点。

反之，设  $x$  是顶点，令  $J = \{j : x_j > 0\}$ 。若列族  $\{A_j : j \in J\}$  线性相关，则存在非零向量  $d$  支持在  $J$  上且  $Ad = 0$ 。取充分小的  $\varepsilon > 0$ ，使  $x \pm \varepsilon d \geq 0$ 。于是两点均在  $P$ ，且

$$x = \frac{1}{2}[(x + \varepsilon d) + (x - \varepsilon d)],$$

与顶点性矛盾。故正分量对应列线性无关，可补成  $m$  个线性无关列作为基；把其余非基变量置零就得到  $x$ ，所以  $x$  是 BFS。□

单纯形从 BFS 沿边换基，用约化成本选入基、最小比值选出基。退化可能零步长并循环；Bland 规则可防循环。两阶段法和修正单纯形只要求掌握可行基的获得与换基逻辑，不扩展复杂退化实现。

对不等式约束  $g_i(x) \leq 0$ ，外二次罚函数

$$\Phi_\rho(x) = f(x) + \frac{\rho}{2} \sum_i [g_i(x)]_+^2$$

通常要令  $\rho \rightarrow \infty$  才逼近可行解； $\ell_1$  型罚  $f(x) + \rho \sum_i [g_i(x)]_+$  在乘子有界等条件下可对有限大  $\rho$  精确，但目标变为非光滑。对严格可行点，对数屏障

$$\Phi_\mu(x) = f(x) - \mu \sum_i \log[-g_i(x)]$$

给出中心路径关系

$$\nabla f(x) + \sum_i \lambda_i \nabla g_i(x) = 0, \quad \lambda_i = -\frac{\mu}{g_i(x)} > 0, \quad \lambda_i [-g_i(x)] = \mu.$$

当  $\mu \downarrow 0$  时互补间隙趋零。中心化 Newton 步解 KKT 线性化系统，并用阻尼保持严格可行；直接令  $\mu = 0$  会失去屏障。罚参数过大和屏障参数过小都会造成病态，必须配合延续和停止准则。

极集  $C^\circ = \{y : \sup_{x \in C} y^T x \leq 1\}$  为闭凸集。闭凸且含原点时双极定理  $C^{\circ\circ} = C$ ；范数单位球的极是对偶范数单位球。2026 春 Q7 完整解答见真题章。

## 8.11 同伦方法与预测—校正路径跟踪

要求解  $F(x) = 0$  时，先选一个容易求解的  $G(x) = 0$ ，构造

$$H(x, t) = (1 - t)G(x) + tF(x), \quad 0 \leq t \leq 1.$$

若在解路径上  $H_x(x(t), t)$  可逆，隐函数定理给局部光滑路径，求导得到

$$H_x(x(t), t)x'(t) + H_t(x(t), t) = 0, \quad x'(t) = -H_x^{-1}H_t.$$

预测步用该切向作 Euler 或高阶外推，校正步在新  $t$  上对  $H(x, t) = 0$  做 Newton。每步都必须同时监控残量、Jacobian 条件数和步长；仅让  $t$  增加而不校正会累积离开解流形的误差。

**例题 8.7** (标量路径). 从  $G(x) = x^2 - 1$  的正根  $x(0) = 1$  连到  $F(x) = x^2 - 2$ 。此时

$$H(x, t) = x^2 - (1 + t), \quad x'(t) = \frac{1}{2x(t)},$$

真路径为  $x(t) = \sqrt{1+t}$ 。在  $t_k$  从  $x_k$  预测  $x_k + \Delta t/(2x_k)$ , 再对  $x^2 - (1 + t_{k+1}) = 0$  做 Newton 校正, 即得到可执行的预测—校正算法。

若路径在折点处  $H_x$  奇异, 直接把  $t$  当参数会失败; 弧长延拓增加一个切向归一化方程, 使  $(x, t)$  同时作为未知量。这是同伦法的适用边界, 不是“减小步长总能解决”的情形。

## 8.12 优化问题的固定点形式

不动点语言在优化中的作用是把最优性条件改写为算子方程, 而不是把非线性方程的任意迭代直接搬来。无约束可微问题满足

$$x^* = x^* - \alpha \nabla f(x^*),$$

所以梯度下降迭代算子为  $T_\alpha = I - \alpha \nabla f$ 。若  $f$  为  $\mu$ -强凸且  $L$ -光滑, 则对适当步长  $0 < \alpha < 2/L$ , 该算子在与谱区间对应的度量下为收缩; 二次函数  $f(x) = \frac{1}{2}x^T Ax - b^T x$  时收缩因子恰由  $\max_{\lambda \in [\mu, L]} |1 - \alpha\lambda|$  控制。

复合凸问题  $\min_x f(x) + g(x)$  的最优性条件

$$0 \in \nabla f(x^*) + \partial g(x^*)$$

等价于近端固定点

$$x^* = \text{prox}_{\alpha g}(x^* - \alpha \nabla f(x^*)).$$

因此“残量”  $G_\alpha(x) = \alpha^{-1}[x - \text{prox}_{\alpha g}(x - \alpha \nabla f(x))]$  比单独的  $\|\nabla f(x)\|$  更适合作为复合问题停止准则。

### 易错点

固定点存在不等于给定迭代收敛; 必须证明收缩、平均非扩张或单调算子条件。对非凸目标,  $G_\alpha(x) = 0$  只给驻点条件, 不能自动推出全局最优。

## 8.13 动态规划与 Bellman 最优性原理

有限时域确定性控制中, 状态满足  $s_{k+1} = T_k(s_k, a_k)$ , 阶段成本为  $c_k(s_k, a_k)$ , 终端成本为  $g(s_N)$ 。定义从时刻  $k$  起的最优剩余成本

$$V_k(s) = \min_{a_k, \dots, a_{N-1}} \left\{ \sum_{j=k}^{N-1} c_j(s_j, a_j) + g(s_N) \right\}.$$

则

$$V_N(s) = g(s), \quad V_k(s) = \min_{a \in \mathcal{A}_k(s)} \{c_k(s, a) + V_{k+1}(T_k(s, a))\}.$$

**定理 8.7** (Bellman 最优性原理). 若可行行动集非空且上式的最小值能取得, 则由反向递推选出的每一步最优行动组成全局最优策略。

*证明.* 对剩余阶段数反向归纳。终端  $k = N$  结论成立。假设  $V_{k+1}$  已等于从下一状态起的最优剩余成本, 则任意时刻  $k$  的策略先选某个  $a$ , 其后成本不小于  $V_{k+1}(T_k(s, a))$ 。对  $a$  取最小得到任何策略的下界; 选取达到该最小值的  $a$ , 再接上归纳假设给出的最优尾策略, 即达到下界。□

**例题 8.8** (两阶段最短路). 终点为  $t$ 。若最后一层状态  $u, v$  到  $t$  的成本分别为 3, 1, 则  $V_1(u) = 3, V_1(v) = 1$ 。起点  $s$  到  $u, v$  的成本分别为 1, 4, 故

$$V_0(s) = \min\{1 + 3, 4 + 1\} = 4.$$

反向递推不仅给最优值, 也记录第一步选  $u$  及随后到  $t$  的策略。

#### 答题方法

动态规划题先写状态、行动、转移和边界, 再写 Bellman 递推; 证明最优性用“任意策略给下界+递推策略达到下界”的归纳。连续状态、高维状态会产生维数灾难, 本书不把近似动态规划扩成当前应用数学考纲主线。

### 8.14 方法比较、真题迁移与终章清单

| 方法     | 主要假设                     | 典型结论            | 每步成本/风险        | 真题        |
|--------|--------------------------|-----------------|----------------|-----------|
| GD     | 凸 $L$ -光滑                | $O(1/k)$        | 1 梯度, 受条件数影响   | 理论主干      |
| GD 强凸  | $\mu$ -强凸、 $L$ -光滑       | $(1 - \mu/L)^k$ | 线性收敛           | 证明链       |
| Newton | Hessian 正定/Lipschitz、近初值 | 局部二次            | 解线性系统          | 算法题       |
| BFGS   | 曲率、线搜索                   | 局部超线性需附加条件      | $O(n^2)$ 存储/更新 | 中频        |
| 投影/近端  | 闭凸集/闭凸非光滑项               | 固定点与 $O(1/k)$   | 投影/prox 可难     | 2024 春 Q7 |
| KKT/对偶 | 资格条件或凸+Slater            | 证书/强对偶          | 活跃集与符号         | 多期高频      |

历史迁移: 2023 秋 Q6 是 BFS-顶点与 LP 强对偶; 2024 春 Q7 是次梯度/软阈值; 2024 秋 Q6/Q7 是强凸水平集和 LP; 2025 春 Q7 是投影与距离平方; 2025 秋 Q7 是  $\ell_1$  残量对偶; 2026 春 Q7 是极集/双极。它们说明本章重心是“定义—资格条件—证明证书”, 而非只背更新式。

## 易错点

不要把严格凸、强凸和 Hessian 处处正定混同；不要写收敛率却省略步长、光滑/强凸条件和度量；不要无资格条件声称 KKT 必要；不要把弱对偶写反；不要显式求 Hessian 逆；不要把单纯形高级实现、同伦或动态规划写成与 KKT/GD 等权的考试主干。

## 复习检查

全书考前必须能手算：均差/Hermite、复化与 Gauss；二分/Newton/割线；LU/QR、Jacobi/GS/SOR、CG、幂法；RK/多步稳定函数；PDE 放大因子/Jordan/CFL；Laplace/匹配；KKT/对偶/Newton 一步。必须能证明：插值余项、压缩、Newton 二次、Cholesky、谱半径判据、CG 主链、Dahlquist 思路、Lax 条件、Laplace 主项、弱对偶/KKT 链。

## 8.15 一阶条件、二阶条件与约束几何

无约束可微问题的局部极小点满足  $\nabla f(x^*) = 0$ 。若二阶可微，则  $\nabla^2 f(x^*) \succeq 0$  是必要条件；若  $\nabla^2 f(x^*) \succ 0$ ，则是严格局部极小的充分条件。半正定不充分，例如  $f(x) = x^3$  在零点 Hessian 为零却非极小。

闭凸集  $C$  上可微凸函数的最优性条件为

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall x \in C.$$

等价写为

$$0 \in \nabla f(x^*) + N_C(x^*).$$

这统一了内点梯度为零、边界法向平衡与投影固定点。

## 8.16 梯度下降的步长与精确二次分析

对二次函数

$$f(x) = \frac{1}{2} x^\top A x - b^\top x, \quad A \succ 0,$$

误差递推

$$e_{k+1} = (I - \alpha A) e_k.$$

收敛当且仅当

$$\rho(I - \alpha A) < 1 \iff 0 < \alpha < \frac{2}{\lambda_{\max}}.$$

使最坏模态  $\max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |1 - \alpha \lambda|$  最小的固定步长为

$$\alpha_* = \frac{2}{\lambda_{\min} + \lambda_{\max}},$$

收敛因子

$$\frac{\kappa - 1}{\kappa + 1}.$$

这给出条件数如何造成狭长等高线上的锯齿收敛。

## 8.17 线搜索、Newton 与拟 Newton

下降方向  $p_k$  满足  $\nabla f(x_k)^\top p_k < 0$ 。Armijo 条件

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^\top p_k$$

保证充分下降；Wolfe 曲率条件控制新梯度沿方向的变化。回溯从初始步长按固定比例缩小，在梯度 Lipschitz 时对下降方向可有限找到 Armijo 步。

BFGS 更新逆 Hessian 近似  $H_k$ ，令

$$s_k = x_{k+1} - x_k, \quad y_k = \nabla f(x_{k+1}) - \nabla f(x_k), \quad \rho_k = (y_k^\top s_k)^{-1},$$

$$H_{k+1} = (I - \rho_k s_k y_k^\top) H_k (I - \rho_k y_k s_k^\top) + \rho_k s_k s_k^\top.$$

若  $H_k \succ 0$  且  $y_k^\top s_k > 0$ ，则  $H_{k+1} \succ 0$ 。合适 Wolfe 线搜索可保证曲率条件。

## 8.18 对偶函数与弱/强对偶

Lagrange 对偶函数

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$$

总是关于  $(\lambda, \nu)$  的凹函数，即使原问题非凸。对任何原始可行  $x$  和对偶可行  $\lambda \geq 0$ ,

$$g(\lambda, \nu) \leq f_0(x),$$

这就是弱对偶。强对偶  $d^* = p^*$  需要额外结构；凸问题的 Slater 条件是常用充分条件。

**例题 8.9.**

$$\min_x \frac{1}{2} \|x\|^2 \quad \text{s.t. } Ax = b.$$

Lagrangian

$$L = \frac{1}{2} \|x\|^2 + \nu^\top (Ax - b).$$

对  $x$  极小给  $x = -A^\top \nu$ ，对偶为

$$\max_\nu -\frac{1}{2} \nu^\top A A^\top \nu - \nu^\top b.$$

若  $A$  满行秩，解  $\nu^* = -(A A^\top)^{-1} b$ ，故

$$x^* = A^\top (A A^\top)^{-1} b,$$

即满足约束的最小范数解。

## 8.19 投影、近端与罚函数的区别

投影法每步严格回到可行集；近端法处理目标中的非光滑项；罚函数把违反约束的代价加入目标。例如二次罚

$$\min_x f(x) + \frac{\rho}{2} \|c(x)\|^2$$

通常要令  $\rho \rightarrow \infty$  才逼近精确等式约束，导致条件数恶化。增广 Lagrangian

$$f(x) + \lambda^\top c(x) + \frac{\rho}{2} \|c(x)\|^2$$

同时更新乘子，可在有限  $\rho$  下改善这一问题。三种方法不能仅因都出现“附加项”而混同。

## 8.20 极集与双极定理的证书

含原点集合  $C$  的极集

$$C^\circ = \{y : \sup_{x \in C} y^\top x \leq 1\}$$

是闭凸集，因为它是闭半空间的交。总有

$$C \subseteq (C^\circ)^\circ.$$

**命题 8.8** (锥的极集公式). 若  $K$  是锥，则在  $K^\circ = \{y : \sup_{x \in K} y^\top x \leq 1\}$  的约定下，

$$K^\circ = \{y : y^\top x \leq 0 \text{ 对所有 } x \in K\}.$$

*证明.* 若存在  $x_0 \in K$  使  $y^\top x_0 > 0$ ，则对任意  $t > 0$ ， $tx_0 \in K$ ，且  $y^\top(tx_0) = ty^\top x_0 \rightarrow \infty$ ，故  $y \notin K^\circ$ 。反之，若  $y^\top x \leq 0$  对所有  $x \in K$ ，则上确界至多为  $0 \leq 1$ ，故  $y \in K^\circ$ 。□

**定理 8.9** (双极定理). 若  $C \subset \mathbb{R}^n$  非空、闭、凸并含原点，则

$$(C^\circ)^\circ = C.$$

一般集合  $A$  的双极是  $\overline{\text{conv}}(A \cup \{0\})$ 。

*证明.* 对任意  $x \in C$  与  $y \in C^\circ$ ，有  $y^\top x \leq 1$ ，故  $C \subseteq (C^\circ)^\circ$ 。反向包含中，若  $C$  闭凸含零且  $z \notin C$ ，强分离定理给出  $a \neq 0$  与常数  $\alpha$ ，使

$$a^\top z > \alpha \geq \sup_{x \in C} a^\top x.$$

因  $0 \in C$ ，有  $\alpha \geq 0$ 。若  $\alpha > 0$ ，令  $y = a/\alpha$ ；若  $\alpha = 0$ ，先按正数缩放  $a$ ，使  $\sup_{x \in C} y^\top x \leq 1 < y^\top z$ 。两种情形都得到  $y \in C^\circ$  而  $z \notin (C^\circ)^\circ$ 。因此  $(C^\circ)^\circ \subseteq C$ 。把  $C$  换成  $\overline{\text{conv}}(A \cup \{0\})$ ，并注意该集合与  $A$  有同一极集，即得一般形式。□

### 答题方法

优化题依次回答：目标与约束的凸性；可行性与存在性；必要/充分最优性条件；算法下降或不动点证书；对偶间隙。只写 KKT 方程而不核验资格条件和充分性，不能构成完整解答。

### 自测

1. 推导 SPD 二次函数梯度下降的最优固定步长。
2. BFGS 为什么需要  $y_k^\top s_k > 0$ ?
3. 弱对偶为何对非凸原问题也成立?
4. 比较投影、近端、二次罚和增广 Lagrangian 的作用对象。
5. 双极定理证明中闭性与凸性用在哪里?

## 第二部分

### 六期正式试卷与完整解答

## 第九章 2023 年秋季正式试卷

### 考试范围提醒

Q1 的打印题面与唯一完整解答见 第 3 章的 Newton 与割线法真题 (第 28 页)。

### 易错点

除特别注明为官方答案的内容外, 本部分解答均为复习用整理稿, 不代表官方评分标准。原题若存在常数、记号或适用范围歧义, 解答会先说明采用的解释。

### 考试范围提醒

本卷共 7 道大题, 以下按原题顺序完整收入。

### 2023 秋 Q2: 显式移位 QR (原卷第 1-2 页)

**Original.** For the explicit shifted QR iteration on an upper Hessenberg  $H_0 = U_0^* A U_0$ : (1) prove every  $H_k$  remains upper Hessenberg; (2) explain the Hessenberg reduction; (3) if  $A$  has distinct eigenvalues and no shift is an eigenvalue, prove the iterates are unreduced; (4) with  $H_k = \begin{bmatrix} G_k & u_k \\ \epsilon_k e^T & \alpha_k \end{bmatrix}$ ,  $e = (0, \dots, 0, 1)^T$ , prove, provided  $\mu_k$  is not an eigenvalue of  $G_k$ ,

$$|\epsilon_{k+1}| \leq \rho_k^2 \|u_k\|_2 |\epsilon_k|^2 + \rho_k |\alpha_k - \mu_k| |\epsilon_k|, \quad \rho_k = \|(G_k - \mu_k I)^{-1}\|_2.$$

### 中文译述

对显式移位 QR: 证明各次迭代保持上 Hessenberg 结构; 解释先把  $A$  化为  $H_0$  的目的; 在互异特征值且移位不命中特征值时证明矩阵不可约; 最后按原卷的  $2 \times 2$  分块证明尾部次对角元  $\epsilon_k$  的给定递推上界。

### 完整解答: Q2

$H_k - \mu I = Q_k R_k$  且  $H_{k+1} = R_k Q_k + \mu I = Q_k^* H_k Q_k$ 。Hessenberg QR 的 Householder/Givens 消元只作用相邻行列,  $R_k Q_k$  仍 Hessenberg。先化 Hessenberg 一次需  $O(n^3)$ , 以后每步  $O(n^2)$ ; 直接对稠密矩阵每步  $O(n^3)$ 。若某次次对角元为零, Hessenberg 分块导致特征多项式分解; 在不碰到特征值的非奇异 QR 步中零/非零次对角模式由 Givens 因子保持, 因此从 distinct-spectrum 的 unreduced 初始相似形出发保持 unreduced。

第四问由分块逆/消元：末行消元向量的范数至多  $\rho_k |\epsilon_k|$ ；更新后的末次对角元由一次二阶耦合项和移位失配项组成，Cauchy-Schwarz 给  $|\epsilon_{k+1}| \leq \rho_k^2 \|u_k\|_2 |\epsilon_k|^2 + \rho_k |\alpha_k - \mu_k| |\epsilon_k|$ 。若取 Rayleigh/Wilkinson 型移位使第二项同阶或更小，即出现快速 deflation。

**关键条件：**  $\mu_k \notin \text{spec}(G_k)$  使分块解存在；结构保持不等于数值舍入下严格非零。

### 2023 秋 Q3: Chebyshev 微分矩阵 (原卷第 2 页)

**Original.** A degree- $n$  polynomial is determined by its values on an  $(n+1)$ -point grid  $\{x_j\}_{j=0}^n$ . Define the differentiation matrix  $D \in \mathbb{R}^{(n+1) \times (n+1)}$  by  $p'(x_i) = \sum_{j=0}^n D_{ij} p(x_j)$ . (1) Prove  $D_{ij} = \ell'_j(x_i)$ , where  $\ell_j$  is the  $j$ th Lagrange basis function. (2) For  $x_j = \cos(j\pi/n)$ , derive explicit formulas for every  $D_{ij}$ .

#### 中文译述

对任意  $(n+1)$  点网格, 证明把多项式函数值映到导数值的微分矩阵满足  $D_{ij} = \ell'_j(x_i)$ ; 当  $x_j = \cos(j\pi/n)$  时推导所有对角与非对角元素的显式公式。

#### 完整解答: Q3

$p(x) = \sum_j p(x_j) \ell_j(x)$ , 逐项求导并令  $x = x_i$  即得。令  $c_0 = c_n = 2, c_j = 1$ , 则对  $i \neq j$ ,

$$D_{ij} = \frac{c_i}{c_j} \frac{(-1)^{i+j}}{x_i - x_j}.$$

由常数导数为零,  $D_{ii} = -\sum_{j \neq i} D_{ij}$ , 化简得  $D_{ii} = -x_i/[2(1-x_i^2)]$  ( $1 \leq i \leq n-1$ ), 端点  $D_{00} = (2n^2+1)/6$ ,  $D_{nn} = -(2n^2+1)/6$ 。直接由节点多项式导数比也可得到非对角式。

### 2023 秋 Q4: 二级 RK (原卷第 2 页)

**Original.** For tableau  $c = (0, 1)^T$ ,  $A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ ,  $b = (1/2, 1/2)^T$ : write  $u_{n+1} = u_n + hF$ ; prove convergence and determine order; find the absolute stability region.

#### 中文译述

对题给二级 Runge-Kutta 表: 写成  $u_{n+1} = u_n + hF(t_n, u_n, h; f)$ ; 在足够光滑与 Lipschitz 条件下证明收敛并确定阶; 给出尽量显式的绝对稳定域。

#### 完整解答: Q4

$K_1 = f(t_n, u_n)$ ,  $K_2 = f(t_n + h, u_n + hK_1)$ ,  $u_{n+1} = u_n + h(K_1 + K_2)/2$ 。Taylor 展开  $K_2 = f + h(f_t + f_y f) + O(h^2)$ , 故一步 defect 为  $O(h^3)$ ; Lipschitz 给一步映射稳定, 离散 Gronwall 得全局二阶。测试方程上  $R(z) = 1 + z + z^2/2$ , 稳定域即  $|1 + z + z^2/2| \leq 1$ ; 负实轴交为  $[-2, 0]$ 。

## 2023 秋 Q5: 三阶色散方程 LF 型格式 (原卷第 2 页)

**Original.** For  $u_t + au_{xxx} = 0$  ( $a$  constant), consider

$$u_m^{n+1} = \frac{1}{2}(u_{m+1}^n + u_{m-1}^n) - \frac{1}{2}akh^{-3}(u_{m+2}^n - 2u_{m+1}^n + 2u_{m-1}^n - u_{m-2}^n),$$

where  $k, h$  are the time step and mesh size. (1) Give the leading-order local truncation term. (2) Analyze stability.

## 中文译述

对原卷给出的  $u_t + au_{xxx} = 0$  Lax-Friedrichs 型五点格式, 求局部截断误差首项, 并用 Fourier 模式分析稳定性与  $k/h^3$  限制。

## 完整解答: Q5

把精确解代入: 空间平均贡献  $\frac{1}{2}h^2u_{xx}$ , 中心三阶差分  $(u_{j+2} - 2u_{j+1} + 2u_{j-1} - u_{j-2})/(2h^3) = u_{xxx} + \frac{1}{4}h^2u_{xxxxx} + O(h^4)$ 。除以时间步后, 主截断项含  $-(h^2/(2k))u_{xx}$  与  $ah^2u_{xxxxx}/4$  (符号按残差定义改变)。Fourier 代入得  $G = \cos \theta - iakh^{-3}(\sin 2\theta - 2\sin \theta)$ , 故  $|G|^2 = \cos^2 \theta + q^2(\sin 2\theta - 2\sin \theta)^2$ , 其中  $q = ak/h^3$ 。由于

$$\sin 2\theta - 2\sin \theta = 2\sin \theta(\cos \theta - 1),$$

条件  $|G| \leq 1$  等价于  $4q^2(1 - \cos \theta)^2 \leq 1$ 。对全波数取上确界得到  $16q^2 \leq 1$ , 故锐 CFL 条件为

$$|ak|/h^3 \leq \frac{1}{4}.$$

端点波数直接代入仍有  $|G| = 1$ 。该模型不在最新版培训主干, 本书只保留分析技法。

## 2023 秋 Q6: 线性规划顶点与强对偶 (原卷第 3 页)

**Original.** For  $\max c^T x$  s.t.  $Ax = b, x \geq 0$ ,  $A$  full row rank, prove basic feasible solutions are exactly vertices; write the dual and prove equality/attainment when an optimum exists.

## 中文译述

对  $\max c^T x$ 、 $Ax = b$ 、 $x \geq 0$  且  $A$  满行秩的线性规划: 证明基本可行解恰为可行域顶点; 写出对偶, 并证明原问题有最优解时对偶也达到最优且两者最优值相等。

## 完整解答: Q6

若 BFS 的基列  $A_B$  可逆而  $x_N = 0$ , 写成两个不同可行点平均会使两点非基分量都为零, 继而  $A_B d_B = 0$ , 故相同, 是顶点。反之若正分量所对应列线性相关, 存在非零  $d$  支撑在这些分量且  $Ad = 0$ ; 取小  $t$  使  $x \pm td \geq 0$ , 则  $x$  非顶点。因此顶点正支撑列独立, 可扩为基, 故是 BFS。

对偶为  $\min b^T y$  s.t.  $A^T y \geq c$ 。弱对偶由  $c^T x \leq y^T A x = b^T y$ 。取原问题最优 BFS 基  $B$ ，令  $y = A_B^{-T} c_B$ ；若某非基列违反  $a_j^T y \geq c_j$ ，沿相应 reduced-cost 正方向可提高目标，矛盾，故  $y$  对偶可行且  $b^T y = c_B^T x_B = c^T x$ ，得到双方达到与等值。

### 2023 秋 Q7: 弱摄动特征值 (原卷第 3 页)

**Original.** For  $u'' + (\lambda + \epsilon f(x))u = 0$ ,  $u(0) = 0, u'(1) = 0$ , obtain  $\lambda$  to accuracy  $O(\epsilon)$ .

#### 中文译述

对  $u'' + (\lambda + \epsilon f(x))u = 0$ ,  $u(0) = 0, u'(1) = 0$ ，在  $0 < \epsilon \ll 1$  下求使误差达到  $O(\epsilon)$  的特征值渐近展开。

#### 完整解答: Q7

令  $\lambda = \lambda_0 + \epsilon \lambda_1 + \dots$ ,  $u = u_0 + \epsilon u_1 + \dots$ 。零阶  $u_0 = \sin(kx)$ ,  $k = (n + \frac{1}{2})\pi$ ,  $\lambda_0 = k^2$ 。一阶  $u_1'' + \lambda_0 u_1 = -(\lambda_1 + f)u_0$ ，同齐次边界。Fredholm 可解条件与  $u_0$  内积为零，故

$$\lambda_1 = -\frac{\int_0^1 f(x) \sin^2(kx) dx}{\int_0^1 \sin^2(kx) dx}.$$

因此  $\lambda_n = k_n^2 - \epsilon \langle f \sin^2(k_n x) \rangle / \langle \sin^2(k_n x) \rangle + O(\epsilon^2)$ ；若题意“accuracy  $O(\epsilon)$ ”只需零阶，则写到余项  $O(\epsilon)$ ，一阶式给更高精度。

## 第十章 2024 年春季正式试卷

### 考试范围提醒

本卷共 7 道大题；Q6、Q7 允许二选一，本书两题均解。

### 2024 春 Q1: Chebyshev 多项式 (原卷第 1 页, 20 分)

**Original.** With  $T_n(x) = \cos(n \arccos x)$ : (i) prove it is degree  $n$ ; (ii) show  $T_n(2x - 1) = T_{2n}(\sqrt{x})$  for  $0 < x < 1$ ; (iii) prove discrete orthogonality on zeros  $x_k = \cos((2k - 1)\pi/(2n))$ .

#### 中文译述

由  $T_n(x) = \cos(n \arccos x)$ : 证明它是  $n$  次多项式; 证明  $T_n(2x - 1) = T_{2n}(\sqrt{x})$ ; 并在  $T_n$  的  $n$  个零点上证明  $\{T_k : 0 \leq k < n\}$  对题给离散内积正交。

#### 完整解答: Q1

$T_{n+1} = 2xT_n - T_{n-1}$ , 由  $T_0 = 1, T_1 = x$  归纳得次数  $n$  且首项系数  $2^{n-1}$ 。令  $\sqrt{x} = \cos \theta$ , 则  $2x - 1 = \cos 2\theta$ , 两边均为  $\cos(2n\theta)$ 。离散正交用积化和差:  $T_j(x_k)T_l(x_k) = \frac{1}{2}[\cos((j - l)\theta_k) + \cos((j + l)\theta_k)]$ ; 对  $\theta_k = (2k - 1)\pi/(2n)$  求和, 非零频率的等比和为零, 因此  $j \neq l$  正交; 范数在  $j = 0$  为  $n$ , 其余为  $n/2$ 。

### 2024 春 Q2: CG 与非 Hermitian Krylov (原卷第 1 页, 15 分)

**Original.** Let  $A$  be SPD and apply CG to  $Ax = b$ . (i) Prove the printed estimate

$$\|x_k - x^*\|_A \leq \left( \frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^k \|x_0 - x^*\|_A,$$

where  $\|x\|_A = \sqrt{x^*Ax}$  and  $\kappa_2(A) = \max \lambda(A) / \min \lambda(A)$ ; any approximation theorem used must be stated. (ii) Describe one Krylov-subspace method for non-Hermitian  $A \neq A^*$ , by clear pseudocode or formulas.

#### 中文译述

对 SPD 系统的 CG, 证明原卷打印的  $A$ -范数条件数误差界, 并清楚陈述所用逼近定理; 再用伪代码或公式描述一种用于非 Hermitian 矩阵的 Krylov 子空间方法。

#### 完整解答: Q2

CG 最小化性质与 Chebyshev 映射给标准界  $\|e_k\|_A \leq 2q^k \|e_0\|_A$ ,  $q = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$ 。原卷公式没有前因子 2; 该无因子形式并非一般标准上界, 故本书保留公式口径差异: 若采用更强归一化/特例可改善常数, 通用证明必须带 2。

GMRES: Arnoldi 构造  $AV_k = V_{k+1}\bar{H}_k$ , 求  $y_k = \arg \min_y \|\beta e_1 - \bar{H}_k y\|_2$ , 置  $x_k = x_0 + V_k y_k$ 。每步正交化成本增长并需存储基; 可 restart, 但会损失最小化历史。

### 2024 春 Q3: Kepler 方程 (原卷第 1-2 页, 15 分)

**Original.** For  $f(x) = x - \epsilon \sin x - \eta$ ,  $0 < |\epsilon| < 1$ : (i) prove a unique real root  $\alpha$  exists and  $\eta - |\epsilon| \leq \alpha \leq \eta + |\epsilon|$ ; (ii) prove  $x_{n+1} = \epsilon \sin x_n + \eta$  converges for every  $x_0$ ; (iii) if  $m\pi < \eta < (m+1)\pi$ , prove Newton converges monotonically when

$$x_0 = \begin{cases} (m+1)\pi, & (-1)^m \epsilon > 0, \\ m\pi, & \text{otherwise.} \end{cases}$$

#### 中文译述

对 Kepler 方程  $x - \epsilon \sin x - \eta = 0$ : 证明唯一实根及  $\eta \pm |\epsilon|$  界; 证明  $x_{n+1} = \epsilon \sin x_n + \eta$  对任意初值收敛; 并按  $m\pi < \eta < (m+1)\pi$  给定的端点初值证明 Newton 单调收敛。

#### 完整解答: Q3

$f'(x) = 1 - \epsilon \cos x \geq 1 - |\epsilon| > 0$ , 且  $f \rightarrow \pm\infty$ , 故唯一根; 由  $\alpha - \eta = \epsilon \sin \alpha$  得界。 $\phi(x) = \epsilon \sin x + \eta$  在全实线 Lipschitz 常数  $|\epsilon| < 1$ , 故任意初值收敛。Newton 单调性: 按  $(-1)^m \epsilon$  选择端点, 使初值一侧  $f(x_0)f''(x_0) > 0$ , 且区间内  $f'$  不变号; 切线交点保持在根同侧并单调有界, 极限代入迭代必为唯一根。

### 2024 春 Q4: 波动一阶系统格式 (原卷第 2 页, 15 分)

**Original.** For  $u_t = v_x$ ,  $v_t = u_x$ , analyze the truncation error and stability of

$$\frac{1}{\tau} \left( u_j^{n+1} - \frac{u_{j+1}^n + u_{j-1}^n}{2} \right) = \frac{v_{j+1}^n - v_{j-1}^n}{2h}, \quad \frac{1}{\tau} \left( v_j^{n+1} - \frac{v_{j+1}^n + v_{j-1}^n}{2} \right) = \frac{u_{j+1}^n - u_{j-1}^n}{2h}.$$

#### 中文译述

对  $u_t = v_x$ ,  $v_t = u_x$  的题给 Lax 平均中心差分系统, 逐式计算截断误差, 并建立  $2 \times 2$  Fourier 放大矩阵判定稳定性。

#### 完整解答: Q4

两式的时间前差与空间中心差给一阶时间、二阶空间; Lax 平均项除以  $\tau$  产生  $O(h^2/\tau)$ , 故在固定 Courant 比下总体一阶。令  $\lambda = \tau/h$ , 代入  $e^{ij\theta}$ , 放大矩阵为  $G = \cos \theta I + i\lambda \sin \theta \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$  (符号不影响模), 特征值  $\cos \theta \pm i\lambda \sin \theta$ , 模平方  $\cos^2 \theta + \lambda^2 \sin^2 \theta$ , 故  $|\lambda| \leq 1$ 。临界等号

矩阵可酉对角化，单位根半单，稳定。

### 2024 春 Q5: 二维中心差分最大值原理 (原卷第 2 页, 15 分)

**Original.** State and prove a maximum principle for the centered discretization of  $u_{xx} + u_{yy} + du_x + eu_y + fu = 0$ ,  $f < 0$ , under suitable assumptions.

#### 中文译述

为  $u_{xx} + u_{yy} + d(x, y)u_x + e(x, y)u_y + f(x, y)u = 0$ ,  $f < 0$  的中心差分，在适当网格与系数条件下写出并证明离散最大值原理。

#### 完整解答: Q5

用中心差分后，若  $|d|h_x \leq 2$ ,  $|e|h_y \leq 2$ ，四邻点系数非负；中心系数为其负和再加  $f < 0$ ，严格为负。若内部取得正最大值  $M$ ，每个邻点不超过  $M$ ，则离散算子在该点严格小于等于  $fM < 0$ ，与齐次方程矛盾。因此正最大值在边界；对  $-u$  得负最小值也在边界，进而唯一性与  $\ell^\infty$  稳定。若系数条件不满足，单调矩阵结构消失，结论不能照搬。

### 2024 春 Q6: 边界层复合展开 (原卷第 2 页, 20 分)

**Original.** For  $\epsilon y'' + \epsilon(1+x)^2 y' - y = x - 1$ ,  $y(0) = \alpha$ ,  $y(1) = -1$ : construct composite expansions for  $\alpha = 1$  and  $\alpha = 0$ , sketch, and justify accuracy for the first case.

#### 中文译述

对  $\epsilon y'' + \epsilon(1+x)^2 y' - y = x - 1$ ,  $y(0) = \alpha$ ,  $y(1) = -1$ : 分别在  $\alpha = 1$ 、 $\alpha = 0$  时构造复合展开并画示意；对第一种情形说明展开的  $\epsilon$  精度。

#### 完整解答: Q6

外层令  $\epsilon = 0$  得  $y_0 = 1 - x$ ，满足  $y_0(1) = 0$  而非  $-1$ ，故  $x = 1$  有宽度  $\sqrt{\epsilon}$  的层；若  $\alpha = 1$ ，左端已匹配。令  $X = (1-x)/\sqrt{\epsilon}$ ，主方程  $Y_{XX} - Y = 0$ ，要求  $Y(0)$  补偿  $-1 - y_0(1) = -1$  且  $Y \rightarrow 0$ ，故  $Y = -e^{-X}$ 。复合主项  $y_c = 1 - x - e^{-(1-x)/\sqrt{\epsilon}}$ 。

$\alpha = 0$  时还需左层  $Z = x/\sqrt{\epsilon}$ ，补偿  $-1$ ，得  $y_c = 1 - x - e^{-x/\sqrt{\epsilon}} - e^{-(1-x)/\sqrt{\epsilon}}$ 。主方程残差含对流项与外解二阶项，代入可得一致误差通常  $O(\sqrt{\epsilon})$ ；若加入下一阶内外项可提升。草图显示  $\alpha = 1$  仅右端快速下降， $\alpha = 0$  两端均有层。

### 2024 春 Q7: 次梯度与软阈值 (原卷第 2 页, 20 分)

**Original.** For convex  $f$ , prove  $\partial f(x)$  nonempty and bounded for  $x \in \text{intdom } f$ ; compute  $\partial \|x\|_1$ ; solve  $\min_x \frac{1}{2} \|x - y\|_2^2 + \lambda \|x\|_1$ .

## 中文译述

对凸函数证明  $x \in \text{intdom } f$  时  $\partial f(x)$  非空且有界；写出  $\|x\|_1$  的完整次微分；求  $\min_x \frac{1}{2}\|x - y\|_2^2 + \lambda\|x\|_1$  的闭式解。

## 完整解答：Q7

在内点取小闭球仍在定义域, 凸函数局部 Lipschitz; 支撑超平面定理给 epigraph 在  $(x, f(x))$  的非竖直支撑, 从而次梯度非空, 局部 Lipschitz 常数界住全部次梯度。 $\partial\|x\|_1$  的第  $i$  分量为  $\text{sign}(x_i)$  若非零, 为  $[-1, 1]$  若为零。目标可分离; 最优性  $0 \in x_i - y_i + \lambda\partial|x_i|$  给  $x_i^* = \text{sign}(y_i)(|y_i| - \lambda)_+$ 。

## 第十一章 2024 年秋季正式试卷

### 考试范围提醒

本卷共 7 道大题，以下按原题顺序完整收入。

### 2024 秋 Q1: SVD 最佳低秩 (原卷第 1 页, 15 分)

**Original.** Find the rank- $k$  best approximation to  $A \in \mathbb{C}^{m \times n}$  in spectral and Frobenius norms.

#### 中文译述

给定  $A \in \mathbb{C}^{m \times n}$  和  $k \leq n \leq m$ ，分别在谱范数与 Frobenius 范数下构造不劣于任意秩  $k$  矩阵的最佳秩  $k$  逼近。

#### 完整解答: Q1

若  $A = U \operatorname{diag}(\sigma_1, \dots) V^*$ ，取  $A_k = U_k \Sigma_k V_k^*$ 。对任意  $\operatorname{rank} B \leq k$ ，在  $\operatorname{span}\{v_1, \dots, v_{k+1}\}$  中存在单位  $z$  使  $Bz = 0$ ，故  $\|A - B\|_2 \geq \|Az\| \geq \sigma_{k+1}$ ，而截断 SVD 达到。Frobenius 情形由酉不变性和奇异值平方和的尾部最小性得  $\|A - A_k\|_F = (\sum_{j>k} \sigma_j^2)^{1/2}$ 。

### 2024 秋 Q2: Gauss-Seidel 收敛 (原卷第 1 页, 20 分)

**Original.** Derive  $x^{k+1} = (D - L)^{-1} U x^k + (D - L)^{-1} b$ ; prove convergence for strict diagonal dominance and for SPD  $A$ .

#### 中文译述

从分量 Gauss-Seidel 写出  $(D - L)^{-1} U$  迭代矩阵；分别证明严格对角占优和对称正定时  $A$  可逆且迭代从任意初值收敛到  $A^{-1}b$ 。

#### 完整解答: Q2

逐行把新分量放左、旧分量放右即得矩阵式。严格对角占优时，若  $B_{GS} z = \lambda z$ ，取使  $|z_i|$  最大的最先/合适指标，从  $\lambda a_{ii} z_i = -\lambda \sum_{j<i} a_{ij} z_j - \sum_{j>i} a_{ij} z_j$  推出  $|\lambda| < 1$ ，故收敛；同一占优也由 Levy-Desplanques 给可逆。

SPD 时 GS 是对二次函数  $\frac{1}{2} x^T A x - b^T x$  的循环坐标精确最小化，每次更新使能量下降，下降量控制步长；能量下有界、步长趋零，任一聚点满足全部坐标梯度为零，SPD 唯一解迫使全列收敛。亦可用矩阵分裂定理证明  $\rho(B_{GS}) < 1$ 。

## 2024 秋 Q3: Hilbert 矩阵谱半径 (原卷第 1 页, 20 分)

**Original.** Prove  $H_n = (1/(i+j-1))$  is positive definite,  $\rho(H_n)$  strictly increases, and  $\rho(H_n) \rightarrow \pi$ .

## 中文译述

对  $H_n = (1/(i+j-1))$ : 证明正定; 证明谱半径随  $n$  严格增加; 证明当  $n \rightarrow \infty$  时谱半径趋于  $\pi$ 。

## 完整解答: Q3

$H_{ij} = \int_0^1 x^{i-1} x^{j-1} dx$ , 故  $c^T H_n c = \int_0^1 |\sum c_i x^{i-1}|^2 dx > 0$ .  $H_n$  是  $H_{n+1}$  主子矩阵, Rayleigh 商给非降; 若相等,  $H_{n+1}$  的最大特征向量末分量为零, 但末行与正 Perron 向量内积非零, 矛盾, 故严格增。Hilbert 不等式给  $\|H_n\|_2 < \pi$ ; 取近似极值序列 (例如离散化  $x^{-1/2}$  的截断向量) 使 Rayleigh 商趋于  $\pi$ , 故极限为  $\pi$ 。

**限制:** 最后一步依赖 Hilbert 不等式及其常数锐性; 原卷允许引用时应明确陈述。

## 2024 秋 Q4: 热方程三层隐式格式 (原卷第 1 页, 10 分)

**Original.** Apply, for  $u_t - a^2 u_{xx} = 0$  (Cauchy or periodic problem),

$$(1+\theta) \frac{u_j^{n+1} - u_j^n}{k} - \theta \frac{u_j^n - u_j^{n-1}}{k} = a^2 \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2}, \quad \theta \geq 0.$$

Give its truncation error and stability, and determine the value of  $\theta$  yielding the highest truncation order.

## 中文译述

对热方程题给含参数  $\theta \geq 0$  的三层隐式格式, 求截断误差、判定稳定性, 并确定使截断阶最高的  $\theta$ 。

## 完整解答: Q4

在  $t_n$  展开左端:  $(1+\theta)(u^{n+1} - u^n)/k - \theta(u^n - u^{n-1})/k = u_t + \frac{1}{2}(1+2\theta)ku_{tt} + O(k^2)$ ; 右端新层中心差为  $a^2[u_{xx} + ku_{xxt} + h^2u_{xxx}/12 + \dots]$ 。用 PDE  $u_{tt} = a^2u_{xxt}$ , 令  $\frac{1}{2}(1+2\theta) = 1$  得  $\theta = 1/2$  消去  $O(k)$ , 最高为时间二阶、空间二阶。

令  $s = 4\sin^2(\xi h/2)$ 、 $r = a^2 k/h^2$ , 并代入  $u_j^n = \zeta^n e^{ij\xi h}$ , 得到

$$(1+\theta+rs)\zeta^2 - (1+2\theta)\zeta + \theta = 0.$$

Jury 条件依次为

$$|\theta| \leq 1+\theta+rs, \quad rs \geq 0, \quad 2+4\theta+rs > 0,$$

故对全部  $\theta \geq 0, r \geq 0$  无条件稳定。 $s = 0$  时两根为 1 与  $\theta/(1+\theta)$ ；单位根为单根，因而没有 Jordan 增长。特别  $\theta = 1/2$  同时给最高阶。

### 2024 秋 Q5：一维离散最大值与稳定界（原卷第 1-2 页，15 分）

**Original.** On  $x_0 < \cdots < x_M$ , let  $Lu_j = -(a_j u_{j-1} - b_j u_j + c_j u_{j+1}) + q_j u_j$  for  $1 \leq j \leq M-1$ , where  $a_j, b_j, c_j > 0, q_j \geq 0$ , and  $a_j + c_j \leq b_j$ . (i) If  $Lu_j \leq 0$ , prove that  $u$  cannot attain a positive interior maximum unless  $u_j \equiv C$ . (ii) If  $d_j = b_j - a_j - c_j + q_j > 0$  and  $Lu_j = \phi_j, u_0 = u_M = 0$ , prove  $\|u\|_\infty \leq \max_j |\phi_j|/d_j$ .

#### 中文译述

在  $a_j, b_j, c_j > 0, q_j \geq 0, a_j + c_j \leq b_j$  下，证明  $Lu_j \leq 0$  时内部不能取得正最大值（除非常数）；再在  $d_j > 0$  和零边界下证明题给  $\ell^\infty$  解界。

#### 完整解答：Q5

在正最大点  $i, u_{i\pm 1} \leq u_i = M$ ，故  $Lu_i \geq (b_i - a_i - c_i + q_i)M = d_i M \geq 0$ ；若假设  $Lu_i \leq 0$ ，等号迫使相邻点也等于  $M$ ，连锁到边界/全网格，得到常数。对方程取  $i$  为  $|u|$  最大点，并对  $u$  或  $-u$  应用同一估计： $d_i \|u\|_\infty \leq |\phi_i|$ ，故  $\|u\|_\infty \leq \max_i |\phi_i|/d_i$ （更精确为  $\max_i |\phi_i|/d_i$ ）。

### 2024 秋 Q6：强凸水平集（原卷第 2 页，10 分）

**Original.** If  $f \in C^2(S)$  is convex on closed convex  $S$  and  $\nabla^2 f \succeq mI$  on the sublevel  $L(f(x_0))$ , prove this sublevel is bounded, closed and convex.

#### 中文译述

设  $S$  非空闭凸、 $f \in C^2(S)$  凸，且在次水平集  $L(f(x_0))$  上有  $d^T \nabla^2 f(x) d \geq m \|d\|^2$ ；证明该次水平集有界、闭且凸。

#### 完整解答：Q6

连续性与  $S$  闭给水平集闭；凸性给其凸。对  $x$  在该水平集连接  $x_0$  的线段仍在水平集，Taylor 积分式给  $f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{m}{2} \|x - x_0\|^2$ 。左侧不超过  $f(x_0)$ ，故二次项被线性项控制，推出  $\|x - x_0\| \leq 2 \|\nabla f(x_0)\|/m$ （零梯度时只有  $x_0$ ），因此有界。

### 2024 秋 Q7：线性规划手算（原卷第 2 页，10 分）

**Original.** For

$$\max 3x_1 + 2x_2 + 15x_3, \quad x_1 + x_2 + x_3 \leq 12, \quad 2x_1 + x_2 + 5x_3 \leq 18, \quad x_1, x_2, x_3 \geq 0,$$

(i) write the standard LP form; (ii) compute the optimal solution and optimal value.

## 中文译述

把原卷的三变量线性规划写成标准形，并计算其最优解和最优目标值。

## 完整解答：Q7

加  $s_1, s_2 \geq 0$ :  $x_1 + x_2 + x_3 + s_1 = 12$ 、 $2x_1 + x_2 + 5x_3 + s_2 = 18$ 。因  $x_3$  单位第二资源收益高，考察顶点。取  $x_1 = x_2 = 0$  得  $x_3 \leq 18/5$ ，值 54；若两约束等号并令一个变量零： $x_1 = 0$  给  $x_3 = 2, x_2 = 10$ ，值 50； $x_2 = 0$  给  $x_3 = -2$  不可行； $x_3 = 0$  给  $x_1 = 6, x_2 = 6$ ，值 30。轴顶点  $x_2 = 12$  值 24、 $x_1 = 9$  值 27。因此最优  $x^* = (0, 0, 18/5)$ 、最优值 54。

## 第十二章 2025 年春季正式试卷

### 考试范围提醒

Q1、Q2 的打印题面与唯一完整解答分别见 第 2 章的周期复化梯形真题（第 13 页）、第 3 章的重根 Newton 真题（第 29 页）。

### 考试范围提醒

本卷共 7 道大题；Q6、Q7 允许二选一，本书两题均解。

### 2025 春 Q3: Sherman–Morrison 与 $O(n)$ 结构（原卷第 1 页，20 分）

**Original.** Prove Sherman–Morrison. For  $A_{jk} = 4n\delta_{jk} + \cos(t_j - s_k)$ , give  $O(n)$  methods for  $Ab$ ,  $A^{-1}$  and  $|A|$ .

#### 中文译述

先证明 Sherman–Morrison 公式；再对  $A_{jk} = 4n\delta_{jk} + \cos(t_j - s_k)$  和任意  $b$ ，分别给出计算  $Ab$ 、 $A^{-1}$  与  $|A|$  的  $O(n)$  方法。

#### 完整解答：Q3

直接乘  $(A + uv^T)[A^{-1} - A^{-1}uv^T A^{-1}/(1 + v^T A^{-1}u)] = I$  即证。由  $\cos(t - s) = \cos t \cos s + \sin t \sin s$ ，写  $A = 4nI + UV^T$ ， $U = [\cos t, \sin t]$ 、 $V = [\cos s, \sin s]$ ，秩至多 2。 $Ab = 4nb + U(V^T b)$ ， $O(n)$ 。Woodbury:  $A^{-1} = (4n)^{-1}I - (4n)^{-2}U[I_2 + (4n)^{-1}V^T U]^{-1}V^T$ ；表示/作用于向量为  $O(n)$ ，无需输出稠密  $n^2$  元素。矩阵行列式引理  $\det A = (4n)^n \det[I_2 + (4n)^{-1}V^T U]$ ，累加四个内积即可。

### 2025 春 Q4: 抛物型一阶系统隐式格式（原卷第 1–2 页，15 分）

**Original.** For  $u_t = -v_{xx}$ ,  $v_t = u_{xx}$ , analyze truncation error and stability of

$$\frac{u_j^{n+1} - u_j^n}{\tau} = -\frac{1}{2h^2}(\delta_x^2 v_j^n + \delta_x^2 v_j^{n+1}), \quad \frac{v_j^{n+1} - v_j^n}{\tau} = \frac{1}{2h^2}(\delta_x^2 u_j^n + \delta_x^2 u_j^{n+1}),$$

where  $\delta_x^2 w_j = w_{j+1} - 2w_j + w_{j-1}$ .

## 中文译述

对  $u_t = -v_{xx}, v_t = u_{xx}$  的题给新旧时间层空间二阶差分平均格式, 分析截断误差并用矩阵 Fourier 方法判定稳定性。

## 完整解答: Q4

两式等价于对空间中心差矩阵  $L_h$  使用梯形法: 若  $w = (u, v)^T$ ,  $J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ , 则  $(w^{n+1} - w^n)/\tau = JL_h(w^{n+1} + w^n)/2$ 。Taylor 给时间二阶、空间二阶。Fourier 下  $L_h$  特征值  $-\lambda_h \leq 0$ , 放大矩阵  $G = (I + \frac{\tau}{2}\lambda_h J)^{-1}(I - \frac{\tau}{2}\lambda_h J)$ 。因  $J^* = -J$ , 这是 Cayley 变换且酉, 故  $\|G\|_2 = 1$ , 无条件  $\ell^2$  稳定。

## 2025 春 Q5: Lax-Wendroff 能量与边界 (原卷第 2 页, 15 分)

**Original.** For  $u_t + au_x = 0$ ,  $\nu = a\tau/h$ , use

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

(i) On the real line prove

$$\|u^{n+1}\|_2^2 = \|u^n\|_2^2 - \nu^2(1-\nu^2) \left( \frac{1}{2} \|\delta_x^+ u^n\|_2^2 - \langle \delta_x^+ u^n, \delta_x^- u^n \rangle \right),$$

with the norms and differences defined in the question. (ii) For  $a > 0$  on  $(0, 1)$  and  $u_0^n = 0$ , give a simple stable numerical boundary condition at  $x = 1$ .

## 中文译述

对输运方程的 Lax-Wendroff 格式: 在全实线证明题给  $\ell^2$  恒等式; 当  $a > 0$ 、左端齐次边界时, 为右端点给出一个使格式稳定的简单数值边界条件。

## 完整解答: Q5

把 LW 写为  $u^{n+1} = u^n - \nu\delta_0 u^n + (\nu^2/2)\delta_+ \delta_- u^n$ , 平方求和; 用  $\langle \delta_0 u, u \rangle = 0$ 、 $\langle \delta_+ \delta_- u, u \rangle = -\|\delta_+ u\|^2$  以及差分恒等式整理, 即得原卷给出的等式。 $|\nu| \leq 1$  时右侧非增 (交叉项可化为非负组合), 故 Cauchy 问题稳定。

$a > 0$  时右端是出流边界, 不应强加物理数据; 简单闭合可取零阶外推  $u_{M+1}^n = u_M^n$  (或一阶线性外推), 配合能量边界通量非增。入流端  $u_0^n = 0$  才是必需边界条件。

## 2025 春 Q6: 立方阻尼多尺度 (原卷第 2 页, 20 分)

**Original.** For  $y'' + y + \epsilon(y')^3 = 0$ ,  $y(0) = 1$ ,  $y'(0) = 0$ , construct a multiple-scale large-time approximation and discuss validity.

## 中文译述

对含立方阻尼的振子  $y'' + y + \epsilon(y')^3 = 0$  及给定初值, 用多尺度法构造大时间有效近似, 并说明该近似的有效时间尺度与误差理由。

## 完整解答: Q6

令  $T = \epsilon t$ ,  $y_0 = A(T) \cos(t + \phi)$ 。能量平均:  $E = (y'^2 + y^2)/2 = A^2/2$ ,  $dE/dt = -\epsilon(y')^4$ , 周期平均  $\langle (y')^4 \rangle = 3A^4/8$ , 故  $AA_T = -3A^4/8$ , 即  $A_T = -3A^3/8$ ,  $\phi_T = 0$ 。  $A(0) = 1$  给  $A(T) = (1 + 3T/4)^{-1/2}$ ,  $y(t) \approx (1 + 3\epsilon t/4)^{-1/2} \cos t$ , 再加一阶三次谐波可提高点态精度。消除了 secular term, 主近似在  $t = O(1/\epsilon)$  一致有效; 更长时间或振幅趋零时需重新估计相对误差。

## 2025 春 Q7: 投影与距离平方 (原卷第 2 页, 20 分)

**Original.** For a nonempty closed convex  $K$ , prove the projection variational inequality and prove  $\Theta(z) = \frac{1}{2} \|z - \Pi_K z\|^2$  is convex  $C^1$  with gradient  $z - \Pi_K z$ .

## 中文译述

对非空闭凸集  $K$ : 证明投影  $\Pi_K(z)$  满足题给变分不等式; 再证明  $\Theta(z) = \frac{1}{2} \|z - \Pi_K z\|^2$  是凸的连续可微函数且  $\nabla \Theta(z) = z - \Pi_K z$ 。

## 完整解答: Q7

对  $p = \Pi_K z$  和  $d \in K$ , 函数  $\|z - [p + t(d - p)]\|^2$  在  $t = 0^+$  导数非负, 得  $\langle z - p, d - p \rangle \leq 0$ 。令  $p_z, p_w$  分别为投影, 两个变分不等式相加得 firm nonexpansiveness:  $\langle p_z - p_w, z - w \rangle \geq \|p_z - p_w\|^2$ 。用上下界  $\Theta(z + h) \leq \frac{1}{2} \|z + h - p_z\|^2$  与投影最优性反向估计, 余项由投影 Lipschitz 控制为  $O(\|h\|^2)$ , 故  $\nabla \Theta(z) = z - p_z$ 。凸性可由梯度单调性  $\langle (z - p_z) - (w - p_w), z - w \rangle \geq 0$  得到。

## 第十三章 2025 年秋季正式试卷

### 考试范围提醒

Q2 的打印题面与唯一完整解答见第 2 章的 Chebyshev 极小与 Gauss–Chebyshev 真题 (第 14 页)。

### 考试范围提醒

本卷共 7 道大题, 以下按原题顺序完整收入。

### 2025 秋 Q1: Newton–Schulz 迭代 (原卷第 1 页, 10 分)

**Original.** For  $X_{k+1} = 2X_k - X_kAX_k$ , prove quadratic convergence if  $\|I - AX_0\| < 1$ , and prove such  $X_0$  exists iff  $A$  is nonsingular.

#### 中文译述

对  $X_{k+1} = 2X_k - X_kAX_k$ : 证明  $\|I - AX_0\| < 1$  时  $A$  非奇异且  $X_k$  二次收敛到  $A^{-1}$ ; 证明存在这样的  $X_0$  当且仅当  $A$  非奇异。

#### 完整解答: Q1

令  $E_k = I - AX_k$ , 则  $E_{k+1} = I - A(2X_k - X_kAX_k) = (I - AX_k)^2 = E_k^2$ , 故  $\|E_k\| \leq \|E_0\|^{2^k}$ 。  
 $\|I - AX_0\| < 1$  使  $AX_0 = I - E_0$  可逆, 故  $A$  可逆; 且  $X_k - A^{-1} = -A^{-1}E_k$ , 二次收敛。  
若  $A$  可逆取  $X_0 = \alpha A^T$ ,  $0 < \alpha < 2/\lambda_{\max}(AA^T)$ , 则  $\|I - \alpha AA^T\|_2 < 1$ ; 反向已证。

### 2025 秋 Q3: 广义特征值与 QZ (原卷第 1 页, 20 分)

**Original.** For the pencil  $\mu Ax = \lambda Bx$ : give an example with more than  $n$  eigenvalues; prove simultaneous unitary equivalence to upper triangular  $(S, T)$ ; recover eigenvalues; outline Hessenberg–triangular reduction.

#### 中文译述

对广义特征值问题  $\mu Ax = \lambda Bx$ : 给出特征值超过  $n$  个的例子; 证明存在酉等价把  $(A, B)$  同时化为上三角; 说明如何读出特征值; 并以  $n = 4$  为例构造 Hessenberg–triangular 预化简。

#### 完整解答: Q3

若  $A = B = 0$ , 任意射影对  $(\lambda, \mu)$  都是特征值, 超过  $n$ ; 这说明需 regular pencil 才有有限计数。广义 Schur 定理给酉  $Q, Z$  使  $Q^*AZ = S, Q^*BZ = T$  均上三角; 可由对矩阵铅笔作归纳: 取一个广义特征向量作为  $Z$  首列、相应左向量构造  $Q$ , 消去首行列后递归。对 regular pencil, 特征值由对角对  $(s_{ii}, t_{ii})$  给射影比  $\lambda_i/\mu_i = s_{ii}/t_{ii}$ ,  $t_{ii} = 0$  表示无穷特征值。Hessenberg-triangular reduction 交替用 Householder: 先对  $B$  作 QR 使其三角, 再用左变换消  $A$  的列下部, 同时用右变换恢复  $B$  的三角结构, 逐列推进; 之后 QZ 迭代保持结构。

#### 2025 秋 Q4: 二维 Du Fort-Frankel (原卷第 1 页, 15 分)

**Original.** Construct DFF for  $u_t = u_{xx} + u_{yy}$  and discuss consistency and stability.

##### 中文译述

为二维扩散方程  $u_t = u_{xx} + u_{yy}$  构造 Du Fort-Frankel 格式, 并分别讨论稳定性及在网格联合趋零路径下的一致性。

##### 完整解答: Q4

令  $r_x = \tau/h_x^2, r_y = \tau/h_y^2$ , 用  $2u_{ij}^n \approx u_{ij}^{n+1} + u_{ij}^{n-1}$  替换两个 Laplacian 的中心项, 整理得

$$(1 + 2r_x + 2r_y)u_{ij}^{n+1} = (1 - 2r_x - 2r_y)u_{ij}^{n-1} + 2r_x(u_{i+1,j}^n + u_{i-1,j}^n) + 2r_y(u_{i,j+1}^n + u_{i,j-1}^n).$$

Fourier 得二次根条件, 可验证对非负  $r_x, r_y$  稳定。Taylor 代入出现通常热方程误差外还含  $\tau^2/h_x^2, \tau^2/h_y^2$  型项, 因此一致需  $\tau/h_x \rightarrow 0, \tau/h_y \rightarrow 0$ ; “无条件稳定”不等于任意路径都收敛。

#### 2025 秋 Q5: 波动三层空间平均格式 (原卷第 2 页, 15 分)

**Original.** For  $u_{tt} = u_{xx}$ , analyze stability of

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\tau^2} = \frac{\delta_x^2 u_j^{n+1}}{4h^2} + \frac{\delta_x^2 u_j^n}{2h^2} + \frac{\delta_x^2 u_j^{n-1}}{4h^2}.$$

##### 中文译述

对波动方程题给三层格式 (空间二阶差分在  $n+1, n, n-1$  层权重分别为  $1/4, 1/2, 1/4$ ), 建立二步特征方程并完整判定稳定性及临界重根。

##### 完整解答: Q5

对 Fourier 模式令  $s = \sin^2(\theta/2)$ 、 $\lambda = \tau/h$ , 代入得到关于  $\zeta$  的对称二次式  $(1 + \lambda^2 s)\zeta^2 + (-2 + 2\lambda^2 s)\zeta + (1 + \lambda^2 s) = 0$ 。除以首项并令  $\zeta = e^{i\omega}$ , 得到

$$\cos \omega = \frac{1 - \lambda^2 s}{1 + \lambda^2 s} \in [-1, 1]$$

对任意有限  $\lambda$  和  $s \in [0, 1]$  成立。因此所有非零波数的两根均在单位圆上且互异，格式在离散能量范数下无条件稳定。 $s = 0$  给双根  $\zeta = 1$ ；对应零波数解为常数加线性时间项，它表示波动方程的空间常数位移与常数速度模态。若只用位移的逐步累有界定义，必须排除不相容的零模初速度；若采用包含离散速度的自然能量范数，则该物理解不构成数值不稳定。这里不能机械地把“双根”一律判为不稳定。

### 2025 秋 Q6: Rayleigh 振子 (原卷第 2 页, 20 分)

**Original.** For  $y'' + y + \epsilon[(y')^3/3 - y'] = 0$ ,  $y(0) = 0, y'(0) = 2a$ , construct a large-time approximation, state accuracy, and sketch phase orbits.

#### 中文译述

对 Rayleigh 振子及  $y(0) = 0, y'(0) = 2a$ : 构造小  $\epsilon$  下的大时间近似; 说明精度; 对若干不同  $a$  画  $y-y'$  相图并解释轨道向极限环靠近的现象。

#### 完整解答: Q6

令  $y = A(T) \sin(t + \phi)$ ,  $T = \epsilon t$ 。平均  $dE/dt = -\epsilon[(y')^4/3 - (y')^2]$ , 并用  $\langle (y')^2 \rangle = A^2/2$ ,  $\langle (y')^4 \rangle = 3A^4/8$ , 得  $A_T = A(4 - A^2)/8, \phi_T = 0$ 。由  $A(0) = 2a, A^2(T) = 4/[1 + (a^{-2} - 1)e^{-T}]$ :  $a < 1$  时趋增、 $a > 1$  时趋减、 $a = 1$  已在半径约 2 的极限环上; 相图轨道均趋向该闭轨。在  $t = O(1/\epsilon)$  上误差为  $O(\epsilon)$ , 更长时间需高阶平均。

### 2025 秋 Q7: $\ell_1$ 残量问题对偶 (原卷第 2 页, 20 分)

**Original.** For  $\min_{x \in \mathbb{R}^n} \|Ax - b\|_2 + \gamma \|x\|_1$  with  $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, \gamma > 0$ : (a) derive the Lagrange dual of  $\min_{x,y} \|y\|_2 + \gamma \|x\|_1$  subject to  $Ax - b = y$ ; (b) if  $Ax^* - b \neq 0$  and  $r = (Ax^* - b)/\|Ax^* - b\|_2$ , show

$$\|A^T r\|_\infty \leq \gamma, \quad r^T Ax^* + \gamma \|x^*\|_1 = 0;$$

(c) show that  $\|a_i\|_2 < \gamma$  implies  $x_i^* = 0$ .

#### 中文译述

对  $\min_x \|Ax - b\|_2 + \gamma \|x\|_1$ : 推导引入  $y = Ax - b$  后的 Lagrange 对偶; 在非零残量下证明题给两条最优性关系; 证明列范数  $\|a_i\|_2 < \gamma$  时必有  $x_i^* = 0$ 。

#### 完整解答: Q7

取  $L = \|y\|_2 + \gamma \|x\|_1 + \lambda^T (Ax - b - y)$ 。关于  $y, x$  的下确界有限分别等价于  $\|\lambda\|_2 \leq 1, \|A^T \lambda\|_\infty \leq \gamma$ , 故对偶为  $\max\{-b^T \lambda : \|\lambda\|_2 \leq 1, \|A^T \lambda\|_\infty \leq \gamma\}$ 。残量  $y^* = Ax^* - b \neq 0$  时置  $r = y^*/\|y^*\|$ ;  $y$  的次梯度与  $x$  的 KKT 条件给  $\lambda = r, -A^T r \in \gamma \partial \|x^*\|_1$ 。于是  $\|A^T r\|_\infty \leq \gamma$ , 并由  $x^*$  内积得  $r^T Ax^* + \gamma \|x^*\|_1 = 0$ 。若  $x_i^* \neq 0$ , KKT 又给  $|a_i^T r| = \gamma$ ; 但  $|a_i^T r| \leq \|a_i\|_2 \|r\|_2 < \gamma$ , 矛盾。

## 第十四章 2026 年春季正式试卷

### 考试范围提醒

Q1、Q2 的打印题面与唯一完整解答分别见 第 3 章的无导数 Newton 型真题（第 29 页）、第 2 章的 Legendre、插值与 Gauss 真题（第 16 页）。

### 考试范围提醒

本卷共 7 道大题；Q6、Q7 允许二选一，本书两题均解。

### 2026 春 Q3：矩阵最小二乘 $AXB \approx C$ （原卷第 1 页，20 分）

**Original.** Given  $A \in \mathbb{R}^{\ell \times m}$ ,  $B \in \mathbb{R}^{n \times p}$ ,  $C \in \mathbb{R}^{\ell \times p}$ , consider  $\min_X \|AXB - C\|_F$ . (a) Prove that the minimum is zero iff  $\text{rank}(A) = \text{rank}([A \ C])$  and  $\text{rank}(B) = \text{rank}\begin{pmatrix} B \\ C \end{pmatrix}$ . (b) Prove that the minimizer is unique iff  $\text{rank}(A) = m$  and  $\text{rank}(B) = n$ . (c) Give a numerical method for the unique minimizer. (d) Does it also minimize  $\|AXB - C\|_2$ ? Explain.

#### 中文译述

对  $\min_X \|AXB - C\|_F$ ：用秩条件刻画最小值为零；刻画极小点唯一；给出数值算法；并判断该  $X^*$  是否也最小化谱范数，说明理由。

#### 完整解答：Q3

$AXB = C$  可解当且仅当  $\text{col } C \subset \text{col } A$  且  $\text{row } C \subset \text{row } B$ ，等价原卷两个增广秩条件。线性映射  $X \mapsto AXB$  单射当且仅当  $A$  满列秩且  $B$  满行秩，即  $\text{rank } A = m, \text{rank } B = n$ ；此时平方损失严格凸于映射像，极小点唯一。

用 thin QR/SVD:  $A = U_A \Sigma_A V_A^T$ ,  $B = U_B \Sigma_B V_B^T$ ，则  $X^* = A^\dagger C B^\dagger$ （在满秩条件下唯一），不形成 Kronecker 大矩阵。它一般不同时最小化谱范数。反例取

$$A = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad B = (1 \ 0), \quad C = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

此时  $AXB = \begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix}$ 。Frobenius 极小点为  $x = 0$ ，其残量谱范数为  $(1 + \sqrt{5})/2$ ；取  $x = 1$  时残量为  $\begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix}$ ，谱范数为  $\sqrt{2} < (1 + \sqrt{5})/2$ 。因此不能用“范数等价”推出同一极小点；只有增加特殊结构条件时才可能共同最优。

## 2026 春 Q4: 三种隐式输运格式 (原卷第 1 页, 15 分)

**Original.** For  $u_t + u_x = 0$ , analyze stability of

$$(a) \quad \frac{u_j^{n+1} - u_j^n}{\tau} + \frac{u_j^{n+1} - u_{j-1}^{n+1}}{h} = 0,$$

$$(b) \quad \frac{u_j^{n+1} - u_j^n}{\tau} + \frac{u_{j+1}^{n+1} - u_j^{n+1}}{h} = 0, \quad (c) \quad \frac{u_j^{n+1} - u_j^n}{\tau} + \frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2h} = 0.$$

## 中文译述

对  $u_t + u_x = 0$ , 分别分析后向时间配后向、前向和中心空间差分且空间项取在  $n+1$  层的三种隐式格式的稳定性。

## 完整解答: Q4

令  $\nu = \tau/h$ 。代入 Fourier:

$$G_a = \frac{1}{1 + \nu(1 - e^{-i\theta})}, \quad G_b = \frac{1}{1 + \nu(e^{i\theta} - 1)}, \quad G_c = \frac{1}{1 + i\nu \sin \theta}.$$

对  $\nu > 0$ , (a) 分母模平方  $[1 + \nu(1 - \cos \theta)]^2 + \nu^2 \sin^2 \theta \geq 1$ , 无条件稳定; (c) 亦  $|G_c| \leq 1$ 。对 (b), 分母实部  $1 - \nu(1 - \cos \theta)$ , 存在波数使模小于 1, 故一般不稳定 (顺风)。这与特征来向一致。

## 2026 春 Q5: 最大值原理与后向 Euler (原卷第 1-2 页, 15 分)

**Original.** State and prove the maximum principle of a centered finite-difference discretization of

$$u_{xx} + u_{yy} + du_x + eu_y + fu = 0, \quad f < 0,$$

under suitable assumptions. Apply it to prove stability of backward Euler for

$$u_t - (u_{xx} + u_{yy} + du_x + eu_y + fu) = 0, \quad (x, y) \in (0, 1)^2,$$

with homogeneous Dirichlet boundary conditions.

## 中文译述

对含  $d, e, f$  且  $f < 0$  的二维椭圆算子写出并证明中心差分最大值原理; 再把它用于  $(0, 1)^2$  上齐次 Dirichlet 抛物问题的后向 Euler, 证明稳定性。

## 完整解答: Q5

最大值原理见 2024 春 Q5。后向 Euler 写  $(I - \tau L_h)u^{n+1} = u^n$ 。在中心差分单调系数条件下,  $I - \tau L_h$  的中心系数严格占优、非对角系数非正, 是 M-矩阵。若  $u^{n+1}$  取得正内部最大值, 离散方程给该最大值不超过  $\|u^n\|_\infty$ ; 对负最小值同理, 故  $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$ , 迭代得无条件  $\ell^\infty$  稳定。

2026 春 Q6: 非线性振子闭轨与周期 (原卷第 2 页, 20 分)

**Original.** For  $\ddot{x} - 2x - x^2 + x^3 = 0$ : find equilibria; show the orbit from  $(2 + \epsilon, 0)$  is closed for small nonzero  $\epsilon$  and approximate its period through  $O(\epsilon^2)$ .

中文译述

对  $\ddot{x} - 2x - x^2 + x^3 = 0$ : 求全部静态解; 证明充分小非零  $\epsilon$  时由  $(2 + \epsilon, 0)$  出发的相轨闭合, 并把周期近似求到  $O(\epsilon^2)$ 。

完整解答: Q6

平衡解满足  $x(x-2)(x+1) = 0$ , 故  $x = -1, 0, 2$ 。写  $x = 2 + z$ , 方程化为  $z'' + 6z + 5z^2 + z^3 = 0$ 。势能在  $z = 0$  二阶导为 6, 故这是严格局部极小; 小能量水平集为围绕原点的闭曲线, 因此初值  $(\epsilon, 0)$  产生闭轨。

令  $s = \omega t$ , 并设  $\omega = \sqrt{6}(1 + \omega_2 \epsilon^2 + O(\epsilon^3))$ ,  $z = \epsilon z_1 + \epsilon^2 z_2 + \epsilon^3 z_3 + \dots$ , 其中  $z_1 = \cos s$ 。二阶方程为

$$z_2'' + z_2 = -\frac{5}{6}z_1^2 = -\frac{5}{12}(1 + \cos 2s),$$

故可取  $z_2 = -5/12 + (5/36)\cos 2s$ , 再用齐次项调整  $z(0) = \epsilon$ 。三阶方程的  $\cos s$  系数由  $2\omega_2 z_1'' + (5/3)z_1 z_2 + (1/6)z_1^3$  给出; 消去共振得到

$$\omega_2 = -\frac{49}{216}.$$

这也与一般公式  $\omega/\omega_0 = 1 + [3\beta/(8\omega_0^2) - 5\alpha^2/(12\omega_0^4)]\epsilon^2 + \dots$  ( $\omega_0^2 = 6, \alpha = 5, \beta = 1$ ) 一致。因此

$$T = \frac{2\pi}{\sqrt{6}} \left( 1 + \frac{49}{216}\epsilon^2 + O(\epsilon^3) \right).$$

2026 春 Q7: polar 与双极定理 (原卷第 2 页, 20 分)

**Original.** Prove polar convexity; find the polar of a cone, a norm unit ball, and the simplex  $\{x : 1^T x = 1, x \geq 0\}$ ; prove  $(C^\circ)^\circ = C$  for closed convex  $C \ni 0$ .

中文译述

按极集定义: 证明任意  $C$  的极集凸; 分别求锥、一般范数单位球和概率单纯形的极集; 最后证明闭凸且含原点时的双极定理  $(C^\circ)^\circ = C$ 。

完整解答: Q7

$C^\circ = \cap_{x \in C} \{y : y^T x \leq 1\}$  是半空间交, 故凸闭。若  $K$  是锥, 缩放迫使  $y^T x \leq 0$  对所有  $x \in K$ , 故为负对偶锥。范数单位球的极是对偶范数单位球  $\{y : \|y\|_* \leq 1\}$ 。概率单纯形上  $\sup_x y^T x = \max_i y_i$ , 故极为  $\{y : \max_i y_i \leq 1\}$ 。

总有  $C \subset (C^\circ)^\circ$ 。若  $z \notin C$ , 闭凸分离给  $a^T z > \sup_{x \in C} a^T x$ 。因  $0 \in C$ , 右侧非负; 适当缩放  $a$  使  $\sup_C a^T x \leq 1$  而  $a^T z > 1$ , 于是  $a \in C^\circ$  但  $z \notin (C^\circ)^\circ$ 。故反包含成立。

# 第十五章 跨年度题型索引与答题策略

| 模块     | 代表真题                                                                      | 主动作            |
|--------|---------------------------------------------------------------------------|----------------|
| 逼近/积分  | 2024 春 Q1; 2025 春 Q1; 2025 秋 Q2; 2026 春 Q2                                | 证明 + 误差        |
| 非线性    | 2023 秋 Q1; 2024 春 Q3; 2025 春 Q2; 2026 春 Q1                                | 局部率            |
| 线性代数   | 2023 秋 Q2-Q3; 2024 春 Q2; 2024 秋 Q1-Q3; 2025 春 Q3; 2025 秋 Q1/Q3; 2026 春 Q3 | 算法 + 界         |
| ODE/渐近 | 2023 秋 Q4/Q7; 2024 春 Q6; 2025 春 Q6; 2025 秋 Q6; 2026 春 Q6                  | 尺度 + 一致性       |
| PDE    | 六期均出现, 最新范围仅以培训 core 为主                                                   | Taylor+Fourier |
| 优化     | 2023 秋 Q6; 2024 春 Q7; 2024 秋 Q6-Q7; 2025 春 Q7; 2025 秋 Q7; 2026 春 Q7       | KKT/几何         |

六期试卷共 42 道顶层大题。对 2024 春 CG 前因子、2025 秋 Chebyshev 极小条件和 2026 春振子周期常数等需要特别说明的地方, 正文保留数学辨析和采用的解释。

# 索引

Fourier 稳定性, 2

Gauss 求积, 2

GMRES, 2

KKT 条件, 97

Laplace 方法, 2

优化, 97

共轭梯度, 2

共轭梯度法, 33

凸优化, 2

匹配渐近展开, 87

后向稳定, 2

对偶, 2

插值, 2

数值偏微分方程, 2, 67

数值线性代数, 2, 33

条件数, 2

渐近分析, 2, 87

离散最大值原理, 67

离散能量, 2

离散能量法, 67

稳定性, 67

线性多步法, 2

绝对稳定, 2

边界层, 2

零稳定, 2

非线性方程, 2