

矩阵计算讲义*

潘建瑜

jypan@math.ecnu.edu.cn

2023 年 2 月

*本讲义仅供《矩阵计算》课堂教学使用

<http://math.ecnu.edu.cn/~jypan/Teaching/MatrixComp/>

纸上得来终觉浅，绝知此事要躬行。

学而不思则罔，思而不学则殆。

知识，只有当它靠积极的思考得来
而不是凭记忆得来的时候，才是真正的知识。



目 录

第零讲 引言	1
0.1 计算数学介绍	2
0.2 数值线性代数	3
第一讲 线性代数基础	6
1.1 线性空间与内积空间	6
1.1.1 线性空间	6
1.1.2 内积空间	9
1.1.3 正交与正交补	9
1.2 矩阵与投影	11
1.2.1 矩阵的秩	11
1.2.2 特征值与特征向量	13
1.2.3 特征值的粗略估计	15
1.2.4 不变子空间	17
1.2.5 投影变换	18
1.3 向量范数与矩阵范数	21
1.3.1 向量范数	21
1.3.2 矩阵范数	24
1.3.3 谱半径与范数	27
1.3.4 最佳逼近与正交投影	28
1.4 矩阵标准型	29
1.4.1 Jordan 标准型	29
1.4.2 Schur 分解	30
1.5 几类特殊矩阵	33
1.5.1 对称正定矩阵	33
1.5.2 对角占优矩阵	34
1.5.3 不可约矩阵	35
1.5.4 其它常见特殊矩阵	37
1.6 Kronecker 积	40
1.7 课后习题	42

第二讲 线性方程组直接方法	47
2.1 LU 分解与 Gauss 消去法	48
2.1.1 LU 分解	48
2.1.2 LU 分解的实现	50
2.1.3 Gauss 消去法	53
2.1.4 选主元 LU 分解	54
2.1.5 矩阵求逆	58
2.1.6 分块 LU 分解	58
2.2 特殊方程组的求解	60
2.2.1 对称正定线性方程组	60
2.2.2 对称不定线性方程组	64
2.2.3 三对角线性方程组	65
2.2.4 带状线性方程组	67
2.2.5 Toeplitz 线性方程组	68
2.3 扰动分析	73
2.3.1 矩阵条件数	73
2.3.2 δx 与 \hat{x} 的关系	74
2.3.3 δx 与 x_* 的关系	74
2.3.4 δx 与残量的关系	77
2.3.5 相对扰动分析	78
2.4 误差分析	80
2.4.1 LU 分解的舍入误差分析	80
2.4.2 Gauss 消去法的舍入误差分析	80
2.5 解的改进	82
2.5.1 高精度运算	82
2.5.2 矩阵元素缩放或平衡 (Scaling or equilibration)	82
2.5.3 迭代改进法	82
2.6 课后习题	84
第三讲 线性最小二乘问题	86
3.1 问题介绍	87
3.1.1 超定方程组	87
3.1.2 欠定方程组	87
3.2 几类重要的矩阵变换	89



3.2.1	初等矩阵变换	89
3.2.2	Gauss 变换	90
3.2.3	Householder 变换	90
3.2.4	Givens 变换	94
3.2.5	正交变换的舍入误差分析	96
3.3	QR 分解	97
3.3.1	QR 分解的存在性与唯一性	97
3.3.2	基于 MGS 的 QR 分解	99
3.3.3	基于 Householder 变换的 QR 分解	100
3.3.4	列主元 QR 分解	103
3.3.5	基于 Givens 变换的 QR 分解	104
3.3.6	QR 分解的稳定性	105
3.4	奇异值分解	107
3.4.1	奇异值, 奇异向量和奇异值分解	107
3.4.2	奇异值基本性质	110
3.4.3	奇异值更多性质	111
3.4.4	奇异值扰动分析	115
3.5	线性最小二乘问题的求解方法	116
3.5.1	正规方程	116
3.5.2	QR 分解法	118
3.5.3	奇异值分解法	119
3.6	广义逆与最小二乘	121
3.6.1	广义逆	121
3.6.2	广义逆基本性质	122
3.6.3	广义逆的计算	123
3.6.4	广义逆与线性最小二乘	123
3.6.5	左逆和右逆	124
3.7	最小二乘扰动分析	125
3.8	推广与应用	126
3.8.1	最小二乘问题的推广	126
3.8.2	最小二乘问题的应用	127
3.9	课后习题	130



第四讲 非对称特征值问题	133
4.1 幂迭代法	134
4.1.1 算法介绍	134
4.1.2 收敛性分析	134
4.1.3 位移策略	135
4.2 反迭代法	136
4.2.1 算法介绍	136
4.2.2 Rayleigh 商迭代	136
4.3 正交迭代法	138
4.4 QR 迭代法	140
4.4.1 QR 迭代与幂迭代的关系	140
4.4.2 QR 迭代与反迭代的关系	141
4.4.3 QR 迭代与正交迭代的关系	141
4.4.4 QR 迭代的收敛性	142
4.4.5 带位移的 QR 迭代法	143
4.5 带位移的隐式 QR 迭代法	145
4.5.1 上 Hessenberg 矩阵	145
4.5.2 隐式 QR 迭代	147
4.5.3 位移的选取	150
4.5.4 收缩	154
4.6 特征向量的计算	155
4.7 广义特征值问题	156
4.7.1 广义特征值基本理论	156
4.7.2 广义 Schur 分解	156
4.7.3 QZ 迭代法	157
4.8 应用	158
4.8.1 多项式求根	158
4.9 课后习题	160
第五讲 对称特征值问题	162
5.1 Jacobi 迭代法	163
5.2 Rayleigh 商迭代法	168
5.3 对称 QR 迭代法	171
5.4 分而治之法	175



5.5	对分法和反迭代法	183
5.6	奇异值分解	186
5.6.1	二对角化	186
5.6.2	Golub-Kahan SVD 算法	188
5.6.3	dqds 算法	189
5.6.4	Jacobi 算法	192
5.7	扰动分析	194
5.7.1	特征值与 Rayleigh 商	194
5.7.2	对称矩阵特征值的扰动分析	196
5.7.3	对称矩阵特征向量的扰动	197
5.7.4	Rayleigh 商逼近	200
5.7.5	相对扰动分析	201
5.8	应用	204
5.8.1	SVD 与图像压缩	204
5.9	课后习题	205
第六讲	线性方程组定常迭代法	207
6.1	定常迭代法	209
6.2	矩阵分裂迭代法	210
6.2.1	Jacobi 迭代法	210
6.2.2	Gauss-Seidel 迭代法	211
6.2.3	SOR 迭代法	212
6.2.4	SSOR 迭代法	213
6.2.5	AOR 迭代法	214
6.2.6	Richardson 迭代法	214
6.2.7	分块迭代法	215
6.3	应用: Poisson 方程求解	216
6.3.1	一维 Poisson 方程	216
6.3.2	二维 Poisson 方程	217
6.3.3	求解方法小结	221
6.4	收敛性分析	223
6.4.1	向量与矩阵序列收敛基本概念	223
6.4.2	定常迭代法的收敛性	224
6.4.3	二维离散 Poisson 方程情形	227



6.4.4	不可约对角占优矩阵	229
6.4.5	对称正定矩阵	231
6.4.6	相容次序矩阵	233
6.5	加速方法	237
6.5.1	外推技术	237
6.5.2	Chebyshev 多项式加速	238
6.6	交替方向法与 HSS 迭代法	244
6.6.1	多步迭代法	244
6.6.2	交替方向法	244
6.6.3	HSS 方法	245
6.7	Poisson 方程快速求解方法	247
6.7.1	快速 Fourier 变换	247
6.7.2	离散 Sine 变换	248
6.7.3	Poisson 方程与 DST	249
6.8	课后习题	251
第七讲	Krylov 子空间迭代法	255
7.1	Krylov 子空间	256
7.1.1	Arnoldi 过程与 Lanczos 过程	256
7.1.2	Krylov 子空间方法一般格式	259
7.1.3	常用的 Krylov 子空间迭代算法	260
7.2	GMRES 方法	261
7.2.1	算法描述	261
7.2.2	具体实施细节	262
7.2.3	GMRES 方法的中断	265
7.2.4	带重启的 GMRES 方法	265
7.3	共轭梯度法	267
7.3.1	算法基本过程	267
7.3.2	实用迭代格式	268
7.4	收敛性分析	274
7.4.1	CG 的收敛性	274
7.4.2	CG 的超收敛性	275
7.4.3	GMRES 的收敛性	276
7.5	预处理方法	280



7.5.1 预处理方法介绍	280
7.5.2 预处理 CG 方法	281
7.5.3 预处理 GMRES 方法	284
7.5.4 预处理子构造	286
7.6 课后习题	289
第八讲 特征值问题的迭代解法	290
8.1 投影算法	290
8.2 Rayleigh-Ritz 方法	291
8.3 Lanczos 方法	293
8.4 Arnoldi 方法	295
8.5 非对称 Lanczos 方法	296
附录 A IEEE 浮点运算标准	299
A.1 浮点数与定点数	299
A.2 IEEE 中的浮点数的表示方法	299
A.3 IEEE 中的浮点数运算	303
A.4 浮点运算舍入误差分析	305
A.5 课后习题	308
附录 B 数值计算中的误差	309
B.1 误差与有效数字	309
B.1.1 基本算术运算的误差估计	311
B.1.2 函数求值的误差估计	311
B.2 误差分析	312
B.2.1 定量分析	312
B.2.2 定性分析	312
B.3 数值稳定性	313
B.3.1 数学问题的稳定性	313
B.3.2 病态问题与条件数	313
B.3.3 算法的稳定性	314
B.3.4 数值计算注意事项	315
B.4 课后习题	316



附录 C 高性能计算 – 科学计算软件介绍	317
C.1 科学计算发展	317
C.1.1 数值分析经典论文	318
C.1.2 Longer list of papers	319
C.2 矩阵运算的复杂度	321
C.3 矩阵乘积的快速算法	321
C.4 数值线性代数程序库	323
C.4.1 BLAS	323
C.4.2 LAPACK	323
C.4.3 ARPACK	324
C.4.4 其它	324
C.5 交互式数学软件	324
C.5.1 MATLAB	324
C.5.2 Mathematica	325
C.5.3 Maple	326
C.5.4 SageMath	326
C.6 集群管理	326
参考文献	327



算法目录

2.1	Gauss 消去法	48
2.2	LU 分解	51
2.3	LU 分解 (用 A 存储 L 和 U)	52
2.4	LU 分解 (IKJ 型)	53
2.5	Gauss 消去法	53
2.6	回代求解 $Ly = b$ 和 $Ux = y$	54
2.7	向后回代求解 $Ux = y$ (列存储方式)	54
2.8	部分选主元 LU 分解	57
2.9	Cholesky 分解	61
2.10	改进的平方根法	63
2.11	追赶法	67
2.12	带状矩阵的 LU 分解	68
2.13	求解 Yule-Walker 方程组的 Levinson-Durbin 算法	70
2.14	求解对称正定 Toeplitz 线性方程组的 Levinson-Durbin 算法	71
2.15	通过迭代改进解的精度	83
3.1	计算 Householder 向量	93
3.2	Givens 变换	95
3.3	Gram-Schmidt 正交化过程	97
3.4	基于 MGS 的 QR 分解	100
3.5	基于 Householder 变换的 QR 分解	102
3.6	基于 Givens 变换的 QR 分解	105
4.1	幂迭代法 (Power Iteration)	134
4.2	带位移的反迭代法 (Inverse Iteration)	136
4.3	Rayleigh 商迭代法 (Rayleigh Quotient Iteration (RQI))	136
4.4	正交迭代法 (Orthogonal Iteration)	138
4.5	QR 迭代法 (QR Iteration)	140
4.6	带位移的 QR 迭代法 (QR Iteration with shift)	143
4.7	上 Hessenberg 化 (Upper Hessenberg Reduction)	146
5.1	Jacobi 迭代算法	164
5.2	经典 Jacobi 迭代算法	165
5.3	循环 Jacobi 迭代算法 (逐行扫描)	166
5.4	Rayleigh 商迭代算法 (RQI, Rayleigh Quotient Iterations)	168
5.5	计算对称三对角矩阵的特征值和特征向量的分而治之法 (函数形式)	177

5.6	修正的 Newton 算法	180
5.7	计算矩阵 $D + uu^T$ 的特征值和特征向量的稳定算法	181
5.8	对分法: 计算 A 在 $[a, b)$ 中的所有特征值	183
5.9	带位移的 LR 算法	189
5.10	qds 算法的单步 ($B_i \rightarrow B_{i+1}$)	190
5.11	dqds 算法的单步 ($B_i \rightarrow B_{i+1}$)	191
5.12	单边 Jacobi 旋转的单步	192
5.13	单边 Jacobi: 计算 $A = U\Sigma V^T$	192
6.1	Jacobi 迭代法	211
6.2	Gauss-Seidel 迭代法	211
6.3	求解线性方程组的 SOR 迭代法	212
6.4	SSOR 迭代法	213
6.5	求解二维离散 Poisson 方程的 Jacobi 迭代法	219
6.6	求解二维离散 Poisson 方程的红黑排序 G-S 迭代法	219
6.7	求解二维离散 Poisson 方程的 SOR 迭代法	220
6.8	Chebyshev 加速方法	242
6.9	二维离散 Poisson 方程的快速方法	249
7.1	Arnoldi 过程 (MGS)	256
7.2	Lanczos 过程	258
7.3	Krylov 子空间迭代算法	259
7.4	GMRES 方法基本框架	261
7.5	实用 GMRES 方法	264
7.6	GMRES(k): 带重启的 GMRES 方法	266
7.7	共轭梯度法 (CG)	272
7.8	两边预处理 CG 方法	282
7.9	PCG: 预处理 CG 方法	282
7.10	基于 P -内积的 CG 方法	284
7.11	右预处理 GMRES 方法	285
8.1	幂迭代: 计算最大特征值	291
8.2	Rayleigh Ritz 算法	292
8.3	Lanczos 算法	293
8.4	Arnoldi 算法	296
8.5	非对称 Lanczos 算法	297
A.1	Horner 法则	306
A.2	带舍入误差的 Horner 法则	307



第零讲 引言

本讲义主要介绍矩阵计算 (或数值线性代数) 中基本问题的数值求解方法, 具体包括:

- 线性方程组的直接解法
- 线性最小二乘问题的数值解法
- 非对称矩阵的特征值计算
- 对称矩阵特征值计算
- 矩阵奇异值分解
- 线性方程组的定常迭代法
- 线性方程组的 Krylov 子空间迭代法

主要参考资料

- G. Golub and C. F. van Loan, Matrix Computations, 4th Edition, 2013.
- J. W. Demmel, Applied Numerical Linear Algebra, 1997.
- L. N. Trefethen and D. Bau, III, Numerical Linear Algebra, 1997.
- 徐树方, 矩阵计算的理论与方法, 1995.

二十世纪最优秀的十大算法, Top 10 Algorithms

“We tried to assemble the 10 algorithms with the greatest influence on the development and practice of science and engineering in the 20th century”
— Editors

1. Metropolis Algorithm for Monte Carlo / Monte Carlo method (1946)
2. Simplex Method for Linear Programming (1947)
3. Krylov Subspace Iteration Methods (1950)
4. The Decompositional Approach to Matrix Computations (1951)
5. The Fortran Optimizing Compiler (1957)
6. QR Algorithm for Computing Eigenvalues (1959-61)
7. Quicksort Algorithm for Sorting (1962)
8. Fast Fourier Transform (1965)
9. Integer Relation Detection Algorithm (1977)
10. Fast Multipole Method (1987)

(下划线: 属于数值线性代数; 波浪线: 与数值线性代数密切相关.)

- ▷ J. Dongarra and F. Sullivan, Guest Editors Introduction to the top 10 algorithms, Computing in Science & Engineering, 2 (2000), 22–23. <https://doi.org/10.1109/MCISE.2000.814652>
- ▷ B. A. Cipra, The Best of the 20th Century: Editors Name Top 10 Algorithms, SIAM News, Volume 33, Number 4, 2000. <https://archive.siam.org/pdf/news/637.pdf>

Nicholas J. Higham 版本 (2016)

1. Newton and quasi-Newton methods
2. Matrix factorizations (LU, Cholesky, QR)
3. Singular value decomposition, QR and QZ algorithms
4. Monte-Carlo methods
5. Fast Fourier Transform (1965)
6. Krylov Subspace Iteration Methods
7. JPEG
8. PageRank
9. Simplex method
10. Kalman filter

▷ N. Higham, The Top 10 Algorithms in Applied Mathematics, 2016. <https://nhigham.com/2016/03/29/the-top-10-algorithms-in-applied-mathematics/>

0.1 计算数学介绍

1947 年, Von Neumann 和 Goldstine 在 Bulletin of the AMS (美国数学会通报) 上发表了题为 “Numerical inverting of matrices of high order” (高阶矩阵的数值求逆) 的著名论文 [123], 开启了现代计算数学的研究. 半个多世纪以来, 伴随着计算机技术的不断进步, 计算数学得到了蓬勃发展, 并逐渐成为了一个独立和重要的学科.

通俗地讲, 科学计算就是通过数学建模将实际问题转化为数学问题, 然后对数学问题进行离散和数值求解, 从而得到原问题的近似解, 同时对得到的近似解进行误差估计, 以确保近似解的可靠性. 科学计算利用先进的计算能力认识 and 解决复杂的科学工程问题, 它融建模、算法、软件研制和计算模拟为一体, 是计算机实现其在高科技领域应用的必不可少的纽带和工具. 计算已不仅仅只是作为验证理论模型正确性的手段, 大量的事例表明它已成为重大科学发现的一种重要手段 [140]. 科学计算已经和理论与实验研究一起, 成为当今世界科学技术创新的主要方式 [136], 也是当前公认的从事现代科学研究的第三种方法.

高性能科学计算在国家安全和科技创新等方面的重要作用也日益受到世界各国的重视. 欧美等国家已投入巨资, 大力发展先进的计算方法, 研制大型的实用软件. 2005 年 6 月, 美国总统信息技术咨询委员会 (President's Information Technology Advisory Committee) 在给美国总统提交的报告《计算科学: 确保美国竞争力》(Computational Science: Ensuring America's Competitiveness) 中明确阐述了计算科学的重要性. 报告认为, 虽然计算本身也是一门学科, 但是其有促进其他学科发展的作用, 21 世纪科学上最重要的、经济上最有前途的前沿研究都有可能通过先进的计算技术和计算科学而得到解决. 报告还认为, 在迅猛发展的高性能计算技术推动下, 计算科学将是 21 世纪确保国家核心竞争能力的战略技术之一, 而科学计算则是计算科学中最主要的内容.

科学计算的核心是计算数学 [140]. 计算数学, 也称数值分析或计算方法, 主要研究各种数学问题的有效数值计算方法及其相关的数学理论, 包括算法的设计与分析 (收敛性, 稳定性, 复杂性等). 其研究内容有数值代数 (线性和非线性), 数值逼近 (插值, 函数逼近和数据拟合), 数值积分与数值微分, 微分方程数值解 (常微分方程, 偏微分方程), 最优化理论与方法等.



有关计算数学和数值线性代数的定义可以参考 Golub 教授的 History of numerical linear algebra: A personal view [51], 或 Trefethen 教授的 Numerical analysis [120].

计算数学的主要任务

- 算法设计: 构造求解各种数学问题的高效数值方法.
- 算法分析: 研究数值方法的收敛性、稳定性、复杂性、计算精度等.
- 算法实现: 编程实现、软件开发.

一个好的数值方法一般需满足以下几点:

- 有可靠的理论分析, 即收敛性稳定性等有数学理论保证.
- 有良好的计算复杂性 (时间和空间).
- 易于在计算机上编程实现.
- 要有具体的数值试验来证明算法是行之有效的.

关于术语“数值方法”或“数值算法”, 一般情况下我们将不加区分地使用.

数值方法一般都是近似方法, 求出的解是有误差的, 因此误差分析也非常重要.

0.2 数值线性代数

If any other mathematical topic is as fundamental to the mathematical sciences as calculus and differential equations, it is numerical linear algebra.

— Trefethen & Bau III [119], 1997.

- 数值代数, 包含数值线性代数和数值非线性代数, 是计算数学的基础 [140].
- **数值线性代数**, 也称**矩阵计算**, 基本问题:

- 线性方程组求解

$$Ax = b, \quad A \in \mathbb{R}^{n \times n} \text{ 非奇异.}$$

- 线性最小二乘问题

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2, \quad A \in \mathbb{R}^{m \times n}.$$

- 矩阵特征值问题

$$Ax = \lambda x, \quad A \in \mathbb{R}^{n \times n}, \lambda \in \mathbb{C}, x \in \mathbb{C}^n, x \neq 0.$$

- 矩阵奇异值问题

$$A^T Ax = \sigma^2 x, \quad A \in \mathbb{R}^{m \times n}, \sigma \geq 0, x \in \mathbb{R}^n, x \neq 0.$$

- 其它问题: 广义特征值问题, 二次特征值问题, 非线性特征值问题, 矩阵方程, 特征值反问题, 张量计算, 随机数值线性代数, 等等.



- 矩阵计算常用方法 (技术, 技巧或工具)
 - 矩阵分解: LU 分解, Cholesky 分解, QR 分解, Schur 分解, SVD 分解等¹
 - 矩阵分裂: 定常迭代法, 预处理子构造等
 - 矩阵降维: 子空间迭代法

🔪 问题的特殊结构对算法的设计具有非常重要的影响.

🔪 动手编写程序对理解算法非常有帮助.

🔪 在实际应用中, 要充分利用现有的优秀程序库.

4 一些记号

- 实部与虚部: $\operatorname{Re}(z)$ 表示复数 z 的实部, $\operatorname{Im}(z)$ 表示复数 z 的虚部;
- 矩阵: 通常用大写字母表示, 如 A, B, X ;
- 向量: 小写字母, 如 x, y, z ;
- 标量: 小写字母或希腊字母, 如 a, b, c, α, β ;
- 下标: 用 a_{ij} , $A(i, j)$ 或 $[A + B]_{i,j}$ 表示矩阵的元素;
- 冒号的作用: (采用 MATLAB 中的表示方法)
 - $a:h:b \rightarrow$ 等差序列: a 为首项, h 为公差, 最后一项 $\leq b$, 若 $h = 1$ 时, 可简写成 $a:b$
 - $x(2:5) \rightarrow$ 子向量: $[x(2), x(3), x(4), x(5)]$
 - $A(i_1:i_2, j_1:j_2) \rightarrow$ 子矩阵: A 的第 i_1 至第 i_2 行与第 j_1 至第 j_2 列组成的子矩阵
 - $A(i_1:i_2, :) \rightarrow$ 子矩阵: A 的第 i_1 至第 i_2 行组成的子矩阵
 - $A(:, j_1:j_2) \rightarrow$ 子矩阵: A 的第 j_1 至第 j_2 列组成的子矩阵.

🔪 本讲义中, 为了便于算法的 MATLAB 实现, 我们通常会用 MATLAB 的方式来编写算法的伪代码.

5 本讲义中常用的数学记号

符号	含义
\mathbb{R}	实数域
\mathbb{R}_+	所有正实数组成的集合
\mathbb{R}^n	n 维实向量空间
\mathbb{C}	复数域
\mathbb{C}^n	n 维复向量空间
$C[a, b]$	$[a, b]$ 上的连续函数空间
$C^p[a, b]$	$[a, b]$ 上的 p 次连续可导函数空间
\bar{x}	共轭

¹The Big Six Matrix Factorizations, Nick Higham, 2022.



符号	含义
x^{\top}, A^{\top}	向量或矩阵的转置
x^*, A^*	向量或矩阵的共轭转置
$\text{rank}(A)$	矩阵的秩
$\det(A)$	矩阵的行列式
$\lambda(A)$	矩阵的特征值
$\rho(A)$	矩阵的谱半径
$A \otimes B$	矩阵的 Kronecker 乘积
$\kappa(A)$	矩阵的条件数
$\dim(\mathcal{S})$	线性空间的维数
$\text{span}\{x_1, x_2, \dots, x_k\}$	由 x_1, x_2, \dots, x_k 张成的线性空间



第一讲 线性代数基础

这里介绍本讲义所涉及的线性代数基础知识,特别是线性空间和矩阵的相关基础知识.

本讲主要参考文献

- ▶ R. A. Horn and C. R. Johnson, *Matrix Analysis*, 1985. 2nd edition, 2013. [71]
- ▶ R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, 1991. [72]
- ▶ 戴华, 矩阵论, 2001. [137]


1.1 线性空间与内积空间

1.1.1 线性空间

线性空间是线性代数最基本的概念之一,它是定义在某个数域上并满足一定条件的集合. 我们首先给出数域的概念.

定义 1.1 (数域) 设 \mathbb{F} 是包含 0 和 1 的一个数集, 如果 \mathbb{F} 中的任意两个数的和, 差, 积, 商 (除数不为 0) 仍然在 \mathbb{F} 中, 则称 \mathbb{F} 为一个数域.

例 1.1 常见的数域有: 有理数域 \mathbb{Q} , 实数域 \mathbb{R} 和复数域 \mathbb{C} .

 本讲义只考虑实数域 \mathbb{R} 和复数域 \mathbb{C} .

定义 1.2 (线性空间) 设 S 是一个非空集合, \mathbb{F} 是一个数域 (\mathbb{C} 或 \mathbb{R}). 在 S 上定义一种代数运算, 称为**加法**, 记为 “+” (即对任意 $x, y \in S$, 都存在唯一的 $z \in S$, 使得 $z = x + y$), 并定义一个从 $\mathbb{F} \times S$ 到 S 的代数运算, 称为**数乘**, 记为 “ \cdot ” (即对任意 $\alpha \in \mathbb{F}$ 和任意 $x \in S$, 都存在唯一的 $y \in S$, 使得 $y = \alpha \cdot x$). 如果这两个运算满足下面的规则, 则称 $(S, +, \cdot)$ 是数域 \mathbb{F} 上的一个**线性空间** (通常简称 S 是数域 \mathbb{F} 上的一个线性空间):

- 加法四条规则

- (1) **交换律**: $x + y = y + x$, $\forall x, y \in S$;
- (2) **结合律**: $(x + y) + z = x + (y + z)$, $\forall x, y, z \in S$;
- (3) **零元素**: 存在一个元素 0 , 使得 $x + 0 = x$, $\forall x \in S$;
- (4) **逆运算**: 对任意 $x \in S$, 都存在**负元素** $y \in S$, 使得 $x + y = 0$, 记 $y = -x$;

- 数乘四条规则


- (1) **单位元**: $1 \cdot x = x$, $1 \in \mathbb{F}, \forall x \in S$;

- (2) **结合律**: $\alpha \cdot (\beta \cdot x) = (\alpha\beta) \cdot x, \quad \forall \alpha, \beta \in \mathbb{F}, x \in \mathbb{S};$
 (3) **分配律**: $(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x, \quad \forall \alpha, \beta \in \mathbb{F}, x \in \mathbb{S};$
 (4) **分配律**: $\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y, \quad \forall \alpha \in \mathbb{F}, x, y \in \mathbb{S}.$

为了表示方便, 通常省略数乘符号, 即将 $\alpha \cdot x$ 写成 αx .

例 1.2 常见的线性空间:

- $\mathbb{R}^n \rightarrow$ 所有 n 维实向量组成的集合, 是 \mathbb{R} 上的线性空间.
- $\mathbb{C}^n \rightarrow$ 所有 n 维复向量组成的集合, 是 \mathbb{C} 上的线性空间.
- $\mathbb{R}^{m \times n} \rightarrow$ 所有 $m \times n$ 阶实矩阵组成的集合, 是 \mathbb{R} 上的线性空间.
- $\mathbb{C}^{m \times n} \rightarrow$ 所有 $m \times n$ 阶复矩阵组成的集合, 是 \mathbb{C} 上的线性空间.
- $\mathbb{P}_n \rightarrow$ 所有次数不超过 n 的多项式组成的集合.
- $C[a, b] \rightarrow$ 区间 $[a, b]$ 上所有连续函数组成的集合.
- $C^p[a, b] \rightarrow$ 区间 $[a, b]$ 上所有 p 次连续可微函数组成的集合.

 为了表述方便, 线性空间的元素通常称为**向量**.

线性相关性和维数

设 \mathbb{S} 是数域 \mathbb{F} 上的一个线性空间, x_1, x_2, \dots, x_k 是 \mathbb{S} 中的一组向量. 如果存在 k 个不全为零的数 $\alpha_1, \alpha_2, \dots, \alpha_k \in \mathbb{F}$, 使得

$$\alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k = 0,$$

则称 x_1, x_2, \dots, x_k **线性相关**, 否则就是**线性无关**.

设 x_1, x_2, \dots, x_k 是 \mathbb{S} 中的一组向量. 如果 $x \in \mathbb{S}$ 可以表示为

$$x = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k,$$

其中 $\alpha_1, \alpha_2, \dots, \alpha_k \in \mathbb{F}$, 则称 x 可以由 x_1, x_2, \dots, x_k **线性表示**, 或者称 x 是 x_1, x_2, \dots, x_k 的**线性组合**, $\alpha_1, \alpha_2, \dots, \alpha_k$ 称为**线性表出系数**.

设向量组 $\{x_1, x_2, \dots, x_m\}$, 如果存在其中的 r ($r \leq m$) 个线性无关向量 $x_{i_1}, x_{i_2}, \dots, x_{i_r}$, 使得所有向量都可以由它们线性表示, 则称 $x_{i_1}, x_{i_2}, \dots, x_{i_r}$ 为向量组 $\{x_1, x_2, \dots, x_m\}$ 的一个**极大线性无关组**, 并称这组向量的**秩**为 r , 记为 $\text{rank}(\{x_1, x_2, \dots, x_m\}) = r$.

设 x_1, x_2, \dots, x_n 是 \mathbb{S} 中的一组线性无关向量. 如果 \mathbb{S} 中的任意一个向量都可以由 x_1, x_2, \dots, x_n 线性表示, 则称 x_1, x_2, \dots, x_n 是 \mathbb{S} 的一组**基**, 并称 \mathbb{S} 是 n 维的, 即 \mathbb{S} 的**维数**为 n , 记为 $\dim(\mathbb{S}) = n$. 如果 \mathbb{S} 中可以找到任意多个线性无关向量, 则称 \mathbb{S} 是**无限维**的.

子空间

设 \mathbb{S} 是一个线性空间, \mathbb{W} 是 \mathbb{S} 的一个非空子集. 如果 \mathbb{W} 关于 \mathbb{S} 上的加法和数乘也构成一个线性空间, 则称 \mathbb{W} 为 \mathbb{S} 的一个**线性子空间**, 简称**子空间**.



例 1.3 设 S 是一个线性空间, 则由零向量组成的子集 $\{0\}$ 是 S 的一个子空间, 称为零子空间. 另外, S 本身也是 S 的子空间. 这两个特殊的子空间称为 S 的**平凡子空间**, 其他子空间都是**非平凡子空间**.

定理 1.1 (子空间的判别) 设 S 是数域 F 上的一个线性空间, W 是 S 的一个非空子集, 则 W 是 S 的一个子空间的充要条件是 W 关于加法和数乘封闭, 即

- (1) 对任意 $x, y \in W$, 有 $x + y \in W$;
- (2) 对任意 $\alpha \in F$ 和任意 $x \in W$, 有 $\alpha x \in W$.

设 S_1, S_2 是线性空间 S 的两个子空间, 则它们的**和**定义为

$$S_1 + S_2 \triangleq \{x + y : x \in S_1, y \in S_2\}.$$

容易证明 $S_1 + S_2$ 也是 S 的子空间. 下面是关于子空间的维数的一个重要性质.

定理 1.2 (维数公式) 设 S_1, S_2 是线性空间 S 的两个有限维子空间, 则 $S_1 + S_2$ 和 $S_1 \cap S_2$ 也都是 S 的子空间, 且

$$\dim(S_1 + S_2) + \dim(S_1 \cap S_2) = \dim(S_1) + \dim(S_2).$$

直和

设 S_1, S_2 是线性空间 S 的两个子空间, 如果 $S_1 + S_2$ 中的任一元素 x 都可以唯一表示成

$$x = x_1 + x_2, \quad x_1 \in S_1, x_2 \in S_2,$$

则称 $S_1 + S_2$ 为**直和**, 记为 $S_1 \oplus S_2$.

关于子空间的直和的判定, 有下面的结论.


定理 1.3 设 S_1, S_2 是线性空间 S 的两个子空间, 则下面的论述等价:

- (1) $S_1 + S_2$ 是直和;
- (2) $S_1 + S_2$ 中的零元素表示方法唯一, 即若 $0 = x_1 + x_2, x_1 \in S_1, x_2 \in S_2$, 则 $x_1 = x_2 = 0$;
- (3) $S_1 \cap S_2 = \{0\}$;
- (4) $\dim(S_1) + \dim(S_2) = \dim(S_1 + S_2)$.

定理 1.4 设 S_1 是线性空间 S 的一个子空间, 则存在 S 的另一个子空间 S_2 , 使得

$$S = S_1 \oplus S_2.$$

我们称 S_2 是 S_1 关于 S 的**补空间**. 显然 S_1 也是 S_2 的**补空间**, 因此它们是互补的.

 **思考:** 补空间是否唯一?



1.1.2 内积空间

内积空间就是带有内积运算的线性空间.

定义 1.3 (内积空间) 设 \mathbb{S} 是数域 \mathbb{F} (\mathbb{C} 或 \mathbb{R}) 上的一个线性空间, 定义一个从 $\mathbb{S} \times \mathbb{S}$ 到 \mathbb{F} 的代数运算, 记为 “ (\cdot, \cdot) ”, 即对任意 $x, y \in \mathbb{S}$, 都存在唯一的 $f \in \mathbb{F}$, 使得 $f = (x, y)$. 如果该运算满足

- (1) $(y, x) = \overline{(x, y)}, \quad \forall x, y \in \mathbb{S};$
- (2) $(x + y, z) = (x, z) + (y, z), \quad \forall x, y, z \in \mathbb{S};$
- (3) $(\alpha x, y) = \alpha(x, y), \quad \forall \alpha \in \mathbb{F}, x, y \in \mathbb{S};$
- (4) $(x, x) \geq 0$, 等号当且仅当 $x = 0$ 时成立;

则称 (\cdot, \cdot) 为 \mathbb{S} 上的一个 **内积 (inner product)**, 定义了内积的线性空间称为 **内积空间**.

🔴 内积有时也称为 **标量积 (scalar product)**.

🔴 定义在实数域 \mathbb{R} 上的内积空间称为**欧氏空间 (Euclidean space)**, 定义在复数域 \mathbb{C} 上的内积空间称为**酉空间**.

🔴 $\overline{(x, y)}$ 表示 (x, y) 的共轭. 当 $\mathbb{F} = \mathbb{R}$ 时, 条件 (1) 即为 $(y, x) = (x, y)$.

🔴 内积可以看作是从线性空间 \mathbb{S} 到数域 \mathbb{F} 的**二元函数**.

例 1.4 设 (\cdot, \cdot) 是 \mathbb{S} 上的一个内积, 则容易验证:

$$(x, \alpha y) = \bar{\alpha}(x, y), \quad \forall \alpha \in \mathbb{F}, x, y \in \mathbb{S}.$$

例 1.5 在 \mathbb{C}^n 上定义内积

$$(x, y) \triangleq y^* x = \sum_{i=1}^n x_i \bar{y}_i,$$

则 \mathbb{C}^n 构成一个内积空间. 类似的, \mathbb{R}^n 上可以定义内积

$$(x, y) \triangleq y^T x = \sum_{i=1}^n x_i y_i.$$

这种方式定义的内积称为 **欧几里得内积 (Euclidean inner product)**, 或 **点积 (dot product)**, 或 **标准内积 (standard inner product)**, 这也是 $\mathbb{C}^n/\mathbb{R}^n$ 上的常用内积 [71, page 15].

例 1.6 对任意 $A, B \in \mathbb{R}^{m \times n}$, 定义

$$(A, B) \triangleq \text{tr}(B^T A),$$

其中 $\text{tr}(\cdot)$ 表示矩阵的**迹**, 即对角线元素之和, 则可以证明 (A, B) 是一个内积, 因此 $\mathbb{R}^{m \times n}$ 构成一个欧氏空间. (留作练习)

1.1.3 正交与正交补

有了内积以后, 我们就可以定义正交.



定义 1.4 (正交) 设 S 是内积空间, $x, y \in S$, 如果 $(x, y) = 0$, 则称 x 与 y **正交**, 记为 $x \perp y$;
 设 S_1 是 S 的子空间, $x \in S$, 如果对任意 $y \in S_1$ 都有 $(x, y) = 0$, 则称 x 与 S_1 **正交**, 记为 $x \perp S_1$;
 设 S_1, S_2 是 S 的两个子空间, 如果对任意 $x \in S_1$, 都有 $x \perp S_2$, 则称 S_1 与 S_2 **正交**, 记为 $S_1 \perp S_2$.

定理 1.5 设 S_1, S_2 是内积空间 S 的两个子空间, 如果 $S_1 \perp S_2$, 则 $S_1 + S_2$ 是直和.

(留作课外自习)

定义 1.5 (正交补) 设 S_1 是内积空间 S 的一个子空间, 则 S_1 的**正交补**定义为

$$S_1^\perp \triangleq \{x \in S : x \perp S_1\},$$

即 S 中所有与 S_1 正交的元素组成的集合.

容易验证, S_1^\perp 也是 S 的一个子空间. 另外, 我们还可以得到下面的结论.

定理 1.6 设 S_1 是内积空间 S 的一个有限维子空间, 则 S_1^\perp 存在唯一, 且

$$S = S_1 \oplus S_1^\perp.$$


例 1.7 设 $x_1, x_2, \dots, x_n \in \mathbb{R}^n$. 若 x_1, x_2, \dots, x_n 线性无关, 则 $\{x_1, x_2, \dots, x_n\}$ 构成 \mathbb{R}^n 的一组基. 进一步, 如果 x_1, x_2, \dots, x_n 相互正交, 即

$$(x_i, x_j) = x_j^\top x_i = 0, \quad i, j = 1, 2, \dots, n,$$

则称它们是一组**正交基**. 如果还满足

$$(x_i, x_i) = x_i^\top x_i = 1, \quad i = 1, 2, \dots, n,$$

则称它们是一组**标准正交基**或**规范正交基**. 特别地, 记 e_i 为单位矩阵的第 i 列, 则 $\{e_1, e_2, \dots, e_n\}$ 构成 \mathbb{R}^n 的一组标准正交基, 这组基通常称为**自然基**.

 任何一组基都可以通过 Gram-Schmidt 正交化过程构造出一组标准正交基.



1.2 矩阵与投影

注记

为了讨论方便, 如果没有特别指出, 本节仅考虑实数情形, 对于复数情形, 我们可以得到类似的结论.

1.2.1 矩阵的秩

设 $A \in \mathbb{R}^{m \times n}$, 则称 A 的列向量组的秩为 A 的**列秩**, 称 A 的行向量组的秩为 A 的**行秩**. 可以验证, 矩阵 A 的列秩与行秩是相等的. 因此我们统一称它们为矩阵 A 的**秩**, 记为 $\text{rank}(A)$.

定理 1.7 设 $A \in \mathbb{R}^{m \times n}$, 则 $\text{rank}(A) = k$ ($0 \leq k \leq \min\{m, n\}$) 的充要条件是 A 存在非奇异的 k 阶子矩阵, 且所有 $k+1$ 阶子矩阵都奇异. (留作课外自习, 可参见 [135])

关于矩阵的秩, 我们有下列的基本性质.

定理 1.8 设 $A, B \in \mathbb{R}^{m \times n}$, 则

- $\text{rank}(A) \leq \min\{m, n\}$;
- $\text{rank}(A^T) = \text{rank}(A)$;
- $\text{rank}(A^T A) = \text{rank}(A A^T) = \text{rank}(A)$;
- $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$;
- 对任意非奇异矩阵 $P \in \mathbb{R}^{m \times m}$ 和 $Q \in \mathbb{R}^{n \times n}$, 有

$$\text{rank}(PA) = \text{rank}(AQ) = \text{rank}(PAQ) = \text{rank}(A).$$

下面是关于矩阵的秩的一些常用性质.

定理 1.9 (秩分解) 设 $\text{rank}(A) = \ell$, 则存在非奇异矩阵 $P \in \mathbb{R}^{m \times m}$ 和 $Q \in \mathbb{R}^{n \times n}$ 使得

$$A = P \begin{bmatrix} I_\ell & 0 \\ 0 & 0 \end{bmatrix} Q.$$

进一步, $\text{rank}(A) = \text{rank}(B)$ 当且仅当存在非奇异矩阵 $P \in \mathbb{R}^{m \times m}$ 和 $Q \in \mathbb{R}^{n \times n}$ 使得 $A = PBQ$.

推论 1.10 (满秩分解) 设 $\text{rank}(A) = \ell$, 则存在非奇异矩阵 $F \in \mathbb{R}^{m \times \ell}$ 和 $G \in \mathbb{R}^{\ell \times n}$ 使得

$$A = FG.$$

定理 1.11 设 $A \in \mathbb{R}^{m \times k}$, $B \in \mathbb{R}^{k \times n}$, 则

$$\text{rank}(A) + \text{rank}(B) - k \leq \text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}.$$

证明. (1) 易知

$$\begin{bmatrix} I_m & -A \\ 0 & I_k \end{bmatrix} \begin{bmatrix} A & 0 \\ I_k & B \end{bmatrix} = \begin{bmatrix} 0 & -AB \\ I_k & B \end{bmatrix}.$$



所以

$$\begin{aligned}\operatorname{rank}(A) + \operatorname{rank}(B) &= \operatorname{rank} \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \\ &\leq \operatorname{rank} \begin{pmatrix} A & 0 \\ I_k & B \end{pmatrix} = \operatorname{rank} \begin{pmatrix} 0 & -AB \\ I_k & B \end{pmatrix} = \operatorname{rank}(AB) + k.\end{aligned}$$

(2) 显然 AB 的列向量都是 A 的列向量的线性组合, 所以 $\operatorname{rank}(AB) \leq \operatorname{rank}(A)$. 同理, AB 的行向量都是 B 的行向量的线性组合, 所以 $\operatorname{rank}(AB) \leq \operatorname{rank}(B)$. \square

推论 1.12 设 $A \in \mathbb{R}^{m \times k}$, $B \in \mathbb{R}^{k \times m}$, $k \leq m$. 若 A 和 B 都是满秩矩阵, 则

$$\operatorname{rank}(AB) = \operatorname{rank}(BA) = \operatorname{rank}(A) = \operatorname{rank}(B) = k.$$

张成的线性空间

设 $x_1, x_2, \dots, x_k \in \mathbb{R}^n$, 记

$$\operatorname{span}\{x_1, x_2, \dots, x_k\} \triangleq \{ \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_k x_k : \alpha_1, \alpha_2, \dots, \alpha_k \in \mathbb{R} \},$$

则 $\operatorname{span}\{x_1, x_2, \dots, x_k\}$ 构成 \mathbb{R}^n 的一个线性子空间, 称为由 x_1, x_2, \dots, x_k **张成的线性空间**. 特别地, 记 $\operatorname{span}(A)$ 为由 A 的所有列向量张成的线性空间.

矩阵 A 相关的四个子空间

设 $A \in \mathbb{R}^{m \times n}$, 则 A 可以看作是从 \mathbb{R}^n 到 \mathbb{R}^m 的一个线性变换 (或线性映射, 线性算子)¹, 即

$$A: x \rightarrow Ax$$

我们分别称

$$\operatorname{Ker}(A) \triangleq \{ x \in \mathbb{R}^n : Ax = 0 \} \subseteq \mathbb{R}^n$$

和

$$\operatorname{Ran}(A) \triangleq \{ y \in \mathbb{R}^m : y = Ax, x \in \mathbb{R}^n \} \subseteq \mathbb{R}^m$$

为 A 的**零空间 (核)** 和**像空间 (列空间, 值域)**, 称

$$\operatorname{Ker}(A^T) \triangleq \{ y \in \mathbb{R}^m : A^T y = 0 \} \subseteq \mathbb{R}^m$$

和

$$\operatorname{Ran}(A^T) \triangleq \{ x \in \mathbb{R}^n : x = A^T y, y \in \mathbb{R}^m \} \subseteq \mathbb{R}^n$$

为 A 的**左零空间**和**行空间**. 可以证明, $\operatorname{Ker}(A)$ 和 $\operatorname{Ran}(A^T)$ 是 \mathbb{R}^n 的线性子空间, $\operatorname{Ran}(A)$ 和 $\operatorname{Ker}(A^T)$ 是 \mathbb{R}^m 的线性子空间, 且 $\operatorname{Ran}(A) = \operatorname{span}(A)$.

定理 1.13 设 $A \in \mathbb{R}^{m \times n}$, 则有

- $\dim(\operatorname{Ran}(A)) = \dim(\operatorname{Ran}(A^T)) = \operatorname{rank}(A)$;

¹严格地讲, 是某个线性变换在某组基下的矩阵



- $\dim(\text{Ker}(A)) + \dim(\text{Ran}(A^T)) = n$;
- $\text{Ran}(A^T A) = \text{Ran}(A^T)$, $\text{Ker}(A^T A) = \text{Ker}(A)$.

例 1.8 设 $A \in \mathbb{R}^{m \times n}$, 则

$$\text{Ran}(A)^\perp = \text{Ker}(A^T).$$

证明. 首先证明 $\text{Ker}(A^T) \subseteq \text{Ran}(A)^\perp$. 设 $y \in \text{Ker}(A^T)$, 即 $A^T y = 0$. 设 z 是 $\text{Ran}(A)$ 中的任意一个向量, 则存在 $x \in \mathbb{R}^n$, 使得 $z = Ax$. 于是


$$z^T y = (Ax)^T y = x^T (A^T y) = 0, \quad \forall z \in \text{Ran}(A),$$

即 $y \in \text{Ran}(A)^\perp$. 所以 $\text{Ker}(A^T) \subseteq \text{Ran}(A)^\perp$.

另一方面, 设 $y \in \text{Ran}(A)^\perp$, 对任意 $z \in \text{Ran}(A)$, 都有 $y^T z = 0$. 又 $AA^T y \in \text{Ran}(A)$, 所以

$$(A^T y)^T (A^T y) = y^* (AA^T y) = 0.$$

因此 $A^T y = 0$, 即 $y \in \text{Ker}(A^T)$. 所以 $\text{Ran}(A)^\perp \subseteq \text{Ker}(A^T)$. 由此可知, 结论成立. \square

 类似地, 有 $\text{Ker}(A)^\perp = \text{Ran}(A^T)$.

 结论在复数域也成立: 如果 $A \in \mathbb{C}^{m \times n}$, 则 $\text{Ran}(A)^\perp = \text{Ker}(A^*)$, $\text{Ker}(A)^\perp = \text{Ran}(A^*)$.

例 1.9 设 $A \in \mathbb{R}^{m \times n}$, 则由例 1.8 可知

$$\text{Ker}(A) \oplus \text{Ran}(A^T) = \mathbb{R}^n, \quad \text{Ker}(A^T) \oplus \text{Ran}(A) = \mathbb{R}^m.$$


例 1.10 (矩阵的秩与齐次线性方程组基础解系) 设 $A \in \mathbb{R}^{m \times n}$ 的秩为 $k \leq \min\{m, n\}$, 则齐次线性方程组 $Ax = 0$ 的基础解系所含解的个数为 $n - k$, 也即 $\dim(\text{Ker}(A)) = n - k$.

1.2.2 特征值与特征向量

定义 1.6 (特征多项式和特征向量) 设 $A \in \mathbb{R}^{n \times n}$. 若存在 $\lambda \in \mathbb{C}$ 和非零向量 $x, y \in \mathbb{C}^n$, 满足

$$Ax = \lambda x, \quad y^* A = \lambda y^*,$$

则称 λ 为 A 的**特征值**, x 为 A 对应于 λ 的 (右) **特征向量**, y 为 A 对应于 λ 的**左特征向量**, 并称 (λ, x) 为 A 的一个**特征对 (eigenpair)**.

 **思考:** 从定义 1.6 能否判断矩阵 A 是否一定存在特征值和特征向量?

矩阵特征值也可以通过特征多项式来定义.

定义 1.7 (特征多项式和特征值) 设 $A \in \mathbb{R}^{n \times n}$, 记 $p_A(\lambda) \triangleq \det(A - \lambda I)$. 易知 $p_A(\lambda)$ 是关于 λ 的 n 次多项式, 我们称之为 A 的**特征多项式**, 其在复数域中的零点称为 A 的**特征值**.

我们知道, n 次多项式在复数域中一定存在 n 个零点 (不考虑重复零点), 因此根据定义 1.7, 一个 n 阶矩阵一定存在 n 个特征值.

下面是关于特征多项式的一个重要性质.

定理 1.14 (Cayley-Hamilton) 设 $p_A(\lambda)$ 是 $A \in \mathbb{R}^{n \times n}$ 的特征多项式, 则 $p_A(A) = 0$.

由 Cayley-Hamilton 定理 1.14 可知, 总存在多项式 $p(t)$ 使得 $p(A) = 0$. 这种特殊多项式称为**零化多项式**.

定义 1.8 (零化多项式和最小多项式) 设 $A \in \mathbb{R}^{n \times n}$, 如果多项式 $p(t)$ 满足 $p(A) = 0$, 则称其为 A 的**零化多项式**, 其中次数最低的首一 (即首项系数为 1) 多项式称为 A 的**最小多项式**.

最小多项式是矩阵的一个重要概念, 在线性代数中有着重要的应用. 容易证明, 最小多项式是存在唯一的, 而且次数不超过 n . 计算最小多项式通常是非常困难的, 一种方法是通过 Jordan 标准型来计算, 见定理 1.46.

关于特征值的几点说明

- 只有当 A 是方阵时, 特征值与特征向量才有定义.
- 实矩阵的特征值与特征向量也有可能是复的.
- 一个 n 阶矩阵总是存在 n 个特征值 (其中可能有相等的), 通常记为 $\lambda_1, \lambda_2, \dots, \lambda_n$.
- 所有特征值组成的集合称为矩阵的**谱**, 通常记为 $\sigma(A)$, 即

$$\sigma(A) \triangleq \{ \lambda_1, \lambda_2, \dots, \lambda_n \}.$$

- 特征值有代数重数 (所对应的特征多项式零点的重数) 和几何重数 (所对应的特征空间的维数), 几何重数不超过代数重数.
- 相似变换不改变矩阵的特征值.
- 合同变换 (congruence transformation) 不改变矩阵的惯性指数 (即正特征值、负特征值和零特征值的个数).



思考: 设 $A \in \mathbb{R}^{n \times n}$, 则 A^T 与 A 的特征值和特征向量是什么关系?

设 A 非奇异, 则 A^{-1} 与 A 的特征值和特征向量是什么关系?

更一般地, 设 $p(t)$ 是一个多项式, 则 $p(A)$ 与 A 的特征值和特征向量是什么关系?

下面给出特征值的一些常用性质.



定理 1.15 设 $A \in \mathbb{R}^{n \times n}$, 则有

$$\lambda_1 \lambda_2 \cdots \lambda_n = \det(A), \quad \lambda_1 + \lambda_2 + \cdots + \lambda_n = \operatorname{tr}(A),$$

其中 $\det(A)$ 表示 A 的行列式, $\operatorname{tr}(A)$ 表示 A 的迹 (对角线元素之和), 即

$$\operatorname{tr}(A) \triangleq a_{11} + a_{22} + \cdots + a_{nn}.$$

推论 1.16 若 A 与 B 相似, 则 $\operatorname{tr}(A) = \operatorname{tr}(B)$, 即相似矩阵具有相同的迹.

定义 1.9 设 $A \in \mathbb{R}^{n \times n}$. 若存在一个非奇异矩阵 $X \in \mathbb{C}^{n \times n}$, 使得

$$X^{-1}AX = \Lambda, \quad (1.1)$$

其中 $\Lambda \in \mathbb{C}^{n \times n}$ 是对角矩阵, 则称 A 是可对角化的, 矩阵 Λ 的对角线元素即为 A 的特征值, 分解 (1.1) 称为矩阵 A 的特征值分解或谱分解.

并非所有矩阵都可以对角化, 比如 $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ 是无法对角化的.

定理 1.17 设 $A \in \mathbb{R}^{n \times n}$, 则

- (1) A 可对角化当且仅当 A 有 n 个线性无关的特征向量;
- (2) A 可对角化当且仅当 A 的所有特征值的代数重数与几何重数都相等;
- (3) 若 A 有 n 个互不相等的特征值, 则 A 可对角化.

例 1.11 设 $A \in \mathbb{R}^{n \times n}$ 是对称矩阵, 则 A 可对角化, 而且可以正交对角化, 即存在正交矩阵 $Q \in \mathbb{R}^{n \times n}$ 使得

$$A = Q\Lambda Q^T.$$

酉矩阵与正交矩阵: 设 $Q \in \mathbb{C}^{n \times n}$, 若 $Q^*Q = I$, 即所有列向量都相互正交, 则称 Q 为酉矩阵. 如果 Q 是实矩阵, 则称为正交矩阵.

根据多项式零点关于多项式系数的连续性, 我们可以得到下面的结论.

定理 1.18 矩阵的特征值关于矩阵元素是连续的, 即当矩阵的元素发生变化时, 其特征值的变化是连续的.

1.2.3 特征值的粗略估计

矩阵特征值在科学与工程计算中应用非常广泛, 但直接计算特征值通常比较困难, 特别是当矩阵规模非常大时. 本小节给出两个估计特征值所在范围的方法. 我们这里假定 A 是复矩阵.



Bendixson 估计方法

设 $A \in \mathbb{C}^{n \times n}$, 我们记

$$H \triangleq \frac{1}{2}(A + A^*), \quad S \triangleq \frac{1}{2}(A - A^*).$$

易知 $A = H + S$, 且

$$H^* = H, \quad S^* = -S,$$

即 H 是 Hermite 的, S 是 Skew-Hermite (斜 Hermite) 的. 我们分别称 H 和 S 为 A 的 Hermite 部分和 Skew-Hermite 部分.

定理 1.19 (Bendixson 定理) 设 $A \in \mathbb{C}^{n \times n}$, 则

$$\lambda_{\min}(H) \leq \operatorname{Re}(\lambda(A)) \leq \lambda_{\max}(H),$$

$$\lambda_{\min}(-\mathbf{i}S) \leq \operatorname{Im}(\lambda(A)) \leq \lambda_{\max}(-\mathbf{i}S),$$

其中 $\operatorname{Re}(\cdot)$ 和 $\operatorname{Im}(\cdot)$ 分别表示实部和虚部, \mathbf{i} 是虚部单位.

这个定理告诉我们, 一个矩阵的特征值的实部的取值范围由其 Hermite 部分确定, 而虚部则由其 Skew-Hermite 部分确定.

Gerschgorin 圆盘估计方法

设 $A = [a_{ij}] \in \mathbb{C}^{n \times n}$, 定义集合

$$\mathcal{D}_i \triangleq \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j \neq i}^n |a_{ij}| \right\}, \quad i = 1, 2, \dots, n. \quad (1.2)$$

我们称 \mathcal{D}_i 为 A 的 **Gerschgorin 圆盘** (也称**盖尔圆盘**).

定理 1.20 (Gerschgorin 圆盘定理) 设 $A \in \mathbb{C}^{n \times n}$, 则 A 的所有特征值都包含在 A 的 Gerschgorin 圆盘的并集中, 即 $\sigma(A) \subset \bigcup_{i=1}^n \mathcal{D}_i$.

证明. 设 λ 是 A 的特征值, 对应的非零特征向量为 $x = [x_1, x_2, \dots, x_n]^\top \in \mathbb{C}^n$, 即 $Ax = \lambda x$. 不失一般性, 设 $\|x\|_\infty = |x_i|$, 则 $|x_i| > 0$. 考察 $Ax = \lambda x$ 的第 i 个方程可得

$$\lambda x_i - a_{ii} x_i = \sum_{j=1, j \neq i}^n a_{ij} x_j.$$

因此

$$|\lambda - a_{ii}| = \frac{1}{|x_i|} \cdot \left| \sum_{j=1, j \neq i}^n a_{ij} x_j \right| \leq \sum_{j=1, j \neq i}^n |a_{ij}| \cdot \frac{|x_j|}{|x_i|} \leq \sum_{j=1, j \neq i}^n |a_{ij}|.$$

所以 $\lambda \in \mathcal{D}_i$. □

将 A 的非对角线元素换成 τa_{ij} , 其中 $0 \leq \tau \leq 1$, 并利用特征值关于矩阵元素的连续性, 我们就可以得到下面的结论.



定理 1.21 设 $A \in \mathbb{C}^{n \times n}$, 如果 $\bigcup_{i=1}^n D_i$ 可分解成两个不相交的子集 S 和 T , 即

$$\bigcup_{i=1}^n D_i = S \cup T \quad \text{且} \quad S \cap T = \emptyset,$$

并假定 S 由 k 的圆盘组成, 而 T 由其它 $n - k$ 个圆盘组成, 则 S 中恰好包含 A 的 k 个特征值 (重特征值按重数计算), 而 T 中则包含 A 的其它 $n - k$ 个特征值.

1.2.4 不变子空间

定义 1.10 设 $A \in \mathbb{R}^{n \times n}$, 子空间 $S \subseteq \mathbb{R}^n$. 若 $AS \subseteq S$, 即对任意 $x \in S$, 都有 $Ax \in S$, 则称 S 为 A 的一个不变子空间.

一类特殊的不变子空间就是由特征向量所张成的子空间.

定理 1.22 设 x_1, x_2, \dots, x_m 是 A 的一组线性无关的特征向量, 则 $\text{span}\{x_1, x_2, \dots, x_m\}$ 是 A 的一个 m 维不变子空间.

下面的结论对矩阵特征值计算非常重要.

定理 1.23 设 $A \in \mathbb{R}^{n \times n}$, $X \in \mathbb{R}^{n \times k}$ 且 $\text{rank}(X) = k$, 则 $\text{span}(X)$ 是 A 的不变子空间的充要条件是存在一个矩阵 $B \in \mathbb{R}^{k \times k}$ 使得

$$AX = XB,$$

此时, B 的特征值都是 A 的特征值.

(板书)

证明. 设 $X = [x_1, x_2, \dots, x_k]$, 由 $\text{rank}(X) = k$ 可知向量组 $\{x_1, x_2, \dots, x_k\}$ 构成子空间 $\text{span}(X)$ 的一组基.

首先证明**必要性**. 设 $\text{span}(X)$ 是 A 的不变子空间, 则 $Ax_j \in \text{span}(X)$. 所以有

$$Ax_j = b_{1j}x_1 + b_{2j}x_2 + \dots + b_{kj}x_k, \quad j = 1, 2, \dots, k,$$

其中 $b_{ij} \in \mathbb{R}$ 是线性表出系数. 将上式写成矩阵形式即为

$$AX = XB \quad \text{其中} \quad B = [b_{ij}] \in \mathbb{R}^{k \times k}.$$

其次证明**充分性**. 设存在矩阵 $B \in \mathbb{R}^{k \times k}$, 使得 $AX = XB$. 则 Ax_j 为 x_1, x_2, \dots, x_k 的线性组合. 又 $\{x_1, x_2, \dots, x_k\}$ 为 $\text{span}(X)$ 的一组基, 所以对任意 $x \in \text{span}(X)$ 都有 $Ax \in \text{span}(X)$, 即 $\text{span}(X)$ 是 A 的一个不变子空间.

下面证明 B 的特征值都是 A 的特征值. 将 X 扩充成一个非奇异的方阵, 即存在矩阵 $\tilde{X} \in \mathbb{R}^{n \times (n-k)}$, 使得 $Y = [X, \tilde{X}] \in \mathbb{R}^{n \times n}$ 非奇异. 将 Y^{-1} 写成分块形式: $Y^{-1} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$, 其中 $Z_1 \in \mathbb{R}^{k \times n}$, $Z_2 \in \mathbb{R}^{(n-k) \times n}$. 由等式 $Y^{-1}Y = I_{n \times n}$ 可得 $Z_1X = I_{k \times k}$, $Z_2X = 0$. 又 $AX = XB$, 所以

$$Y^{-1}AY = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} A[X, \tilde{X}] = \begin{bmatrix} Z_1AX & Z_1A\tilde{X} \\ Z_2AX & Z_2A\tilde{X} \end{bmatrix} = \begin{bmatrix} Z_1XB & Z_1A\tilde{X} \\ Z_2XB & Z_2A\tilde{X} \end{bmatrix} = \begin{bmatrix} B & Z_1A\tilde{X} \\ 0 & Z_2A\tilde{X} \end{bmatrix}.$$



因此 B 的特征值都是 $Y^{-1}AY$ 的特征值. 由于 A 与 $Y^{-1}AY$ 相似, 所以它们具有相同的特征值.

由此, 定理结论成立. \square

推论 1.24 设 $A \in \mathbb{R}^{n \times n}$, $X \in \mathbb{R}^{n \times k}$ 且 $\text{rank}(X) = k$. 若存在一个矩阵 $B \in \mathbb{R}^{k \times k}$ 使得 $AX = XB$, 则 (λ, v) 是 B 的一个特征对当且仅当 (λ, Xv) 是 A 的一个特征对. (留作课外自习)

1.2.5 投影变换

设 S_1 和 S_2 是内积空间 S 的两个子空间, 且 $S = S_1 \oplus S_2$, 则 S 中的任意向量 x 都可唯一表示为

$$x = x_1 + x_2, \quad x_1 \in S_1, \quad x_2 \in S_2.$$

我们称 x_1 为 x 沿 S_2 到 S_1 上的**投影**, 记为 $x|_{S_1}$.

需要指出的是, 由于 S_1 的补空间不唯一, 因此在讨论投影时一定要明确给定 S_2 .

例 1.12 设 $S_1 = \text{span}\{e_1\}$, $S_2 = \text{span}\{e_2\}$, $\tilde{S}_2 = \text{span}\{e\}$, 其中 $e_1 = [1, 0]^T$, $e_2 = [0, 1]^T$, $e = [1, 1]^T$. 于是有

$$\mathbb{R}^2 = S_1 \oplus S_2 = S_1 \oplus \tilde{S}_2.$$

向量 $x = [2, 3]^T$ 沿 S_2 到 S_1 上的投影是 $[2, 0]^T$, 而它沿 \tilde{S}_2 到 S_1 上的投影是 $[-1, 0]^T$.

定义线性变换 $P: S \rightarrow S$ 如下:

$$Px = x|_{S_1}, \quad \forall x \in S.$$

称 P 是从 S 沿 S_2 到 S_1 上的**投影变换** (也称**投影算子**), 对应的矩阵称为**投影矩阵**.

几点注记

- 对于给定的子空间 S_1 和 S_2 (构成直和 $S = S_1 \oplus S_2$), 投影变换是唯一的.
- 线性变换在不同的基下对应不同的变换矩阵. 在不加特别指出时, 本讲义中如果线性空间是 \mathbb{R}^n 或 $\mathbb{R}^{n \times n}$, 我们采用自然基, 即 $\{e_1, e_2, \dots, e_n\}$ 和 $\{e_{ij}\}_{i,j=1}^n$.
- 为了书写方便, 我们这里使用 P 既表示投影变换也表示其对应的投影矩阵.

设 P 是从 S 沿 S_2 到 S_1 上的投影变换, 则对任意 $x \in S_1$ 都有 $Px = x$. 因此, $S_1 \subseteq \text{Ran}(P)$. 又由定义可知 $\text{Ran}(P) \subseteq S_1$, 所以

$$S_1 = \text{Ran}(P).$$

类似地, 我们也可以验证

$$S_2 = \text{Ker}(P).$$

于是存在直和分解


$$S = \text{Ran}(P) \oplus \text{Ker}(P).$$

若 $S = \mathbb{R}^n$, 则立即可以得到下面的结论.



引理 1.25 设 $P \in \mathbb{R}^{n \times n}$ 是一个投影矩阵, 则

$$\mathbb{R}^n = \text{Ran}(P) \oplus \text{Ker}(P). \quad (1.3)$$

 **思考:** 对于一般的矩阵 $A \in \mathbb{R}^{n \times n}$, 结论 $\mathbb{R}^n = \text{Ran}(A) \oplus \text{Ker}(A)$ 是否成立?

下面的性质表明, 投影矩阵由其像空间和零空间所唯一确定.

定理 1.26 设 $\mathbb{R}^n = \mathbb{S}_1 \oplus \mathbb{S}_2$, 则存在唯一的投影矩阵 P , 使得

$$\text{Ran}(P) = \mathbb{S}_1, \quad \text{Ker}(P) = \mathbb{S}_2,$$

即对任意向量 $x \in \mathbb{R}^n$, 有

$$Px \in \mathbb{S}_1, \quad x - Px \in \mathbb{S}_2.$$

例 1.13 若 $\mathbb{S}_1 = \mathbb{R}^n$, 则 $\mathbb{S}_2 = \{0\}$, 所对应的投影矩阵即为单位矩阵 I .

反之, 若 $\mathbb{S}_1 = \{0\}$, 则 $\mathbb{S}_2 = \mathbb{R}^n$, 此时所对应的投影矩阵即为零矩阵.

引理 1.27 设 $P \in \mathbb{R}^{n \times n}$ 是一个投影矩阵, 则

(1) $I - P$ 也是一个投影矩阵, 且 $\text{Ker}(P) = \text{Ran}(I - P)$;

(2) P^T 也是一个投影矩阵.

(留作练习)

下面给出投影矩阵的判别定理. 首先, 根据定义, P 是沿 \mathbb{S}_2 到 \mathbb{S}_1 的投影变换的充要条件是: 对任意 $x \in \mathbb{S}_1$ 有 $Px = x$, 而对任意 $x \in \mathbb{S}_2$ 有 $Px = 0$.

定理 1.28 矩阵 $P \in \mathbb{R}^{n \times n}$ 是投影矩阵的充要条件是 $P^2 = P$, 即 P 是幂等矩阵.


(板书)

证明. 必要性: 设 P 是投影矩阵, 则对任意 $x \in \mathbb{R}^n$, 都有

$$P^2x = P(Px) = Px.$$

因此 $P^2 = P$.

充分性: 设 $P^2 = P$. 我们只需证明 $\text{Ran}(P) + \text{Ker}(P) = \mathbb{R}^n$. 显然 $\text{Ran}(P) + \text{Ker}(P) \subseteq \mathbb{R}^n$, 因此只要证明 $\mathbb{R}^n \subseteq \text{Ran}(P) + \text{Ker}(P)$. 对任意 $x \in \mathbb{R}^n$, 有 $x = Px + (x - Px)$. 由 $P(x - Px) = Px - P^2x = 0$ 可知 $x - Px \in \text{Ker}(P)$. 因此 $\mathbb{R}^n \subseteq \text{Ran}(P) + \text{Ker}(P)$. 所以结论 $\text{Ran}(P) + \text{Ker}(P) = \mathbb{R}^n$ 成立. \square

 **思考:** (1) 在证明充分性时, 为什么只需证明 $\text{Ran}(P) + \text{Ker}(P) = \mathbb{R}^n$?
(2) 该结论在 $\mathbb{C}^{n \times n}$ 中是否成立?

设 \mathbb{S}_1 和 \mathbb{S}_2 是 \mathbb{R}^n 的两个 m 维子空间且 $\mathbb{R}^n = \mathbb{S}_1 \oplus \mathbb{S}_2^\perp$, 则存在唯一的投影变换 P , 使得

$$\text{Ran}(P) = \mathbb{S}_1, \quad \text{Ker}(P) = \mathbb{S}_2^\perp.$$


此时, 我们称 P 是 S_1 上与 S_2 正交的投影变换.

令 v_1, v_2, \dots, v_m 和 w_1, w_2, \dots, w_m 分别是 S_1 和 S_2 的一组基, 则 P 可以由这两组基来表示.

定理 1.29 设 $P \in \mathbb{R}^{n \times n}$ 是 S_1 上与 S_2 正交的投影矩阵, 则

$$P = V(W^\top V)^{-1}W^\top, \quad (1.4)$$

其中 $V = [v_1, v_2, \dots, v_m]$, $W = [w_1, w_2, \dots, w_m]$. (留作练习)

 提示: 先证明 P 是投影矩阵, 再证明 $\text{Ran}(P) = S_1$, $\text{Ker}(P) = S_2^\perp$.

 虽然投影矩阵 P 由 S_1 和 S_2 唯一确定, 但其矩阵表示形式 (1.4) 并不唯一 (W 和 V 不唯一).

设 S_1 是内积空间 S 的一个子空间, 则由定理 1.6 可知 $S = S_1 \oplus S_1^\perp$. 因此, 任意 $x \in S$ 都可唯一分解成

$$x = x_1 + x_2, \quad x_1 \in S_1, \quad x_2 \in S_1^\perp.$$

我们称 x_1 称为 x 在 S_1 中的**正交投影**.

若 P 是从 S 沿 S_1^\perp 到 S_1 上的投影变换, 则称 P 为子空间 S_1 上的**正交投影变换** (也称**正交投影算子**, **orthogonal projector**, 对应的矩阵称为**正交投影矩阵**), 记为 P_{S_1} . 如果 P 不是正交投影变换, 则称为**斜投影变换** (**oblique projector**).

由定理 1.29 可立即得到下面的结论.

推论 1.30 设 P 是子空间 S_1 上的正交投影变换, $\{v_1, v_2, \dots, v_m\}$ 是 S_1 的一组标准正交基, 则

$$P = VV^\top. \quad (1.5)$$

定理 1.31 投影矩阵 $P \in \mathbb{R}^{n \times n}$ 是正交投影矩阵的充要条件 $P^\top = P$.

(留作练习)

 **思考:** 正交投影矩阵 P 的特征值可能取值有哪些?




1.3 向量范数与矩阵范数

1.3.1 向量范数

定义 1.11 (向量范数) 若函数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ (或 $f: \mathbb{C}^n \rightarrow \mathbb{R}$) 满足

- (1) $f(x) \geq 0, \forall x \in \mathbb{R}^n$ (或 \mathbb{C}^n), 等号当且仅当 $x = 0$ 时成立;
- (2) $f(\alpha x) = |\alpha| \cdot f(x), \forall x \in \mathbb{R}^n$ (或 \mathbb{C}^n), $\alpha \in \mathbb{R}$ (或 \mathbb{C});
- (3) $f(x+y) \leq f(x) + f(y), \forall x, y \in \mathbb{R}^n$ (或 \mathbb{C}^n);

则称 $f(x)$ 为 \mathbb{R}^n (或 \mathbb{C}^n) 上的 **范数 (norm)**, 通常记作 $\|x\|$.

 如果 f 只满足 $f(x) \geq 0$, 以及 (2) 和 (3), 则称为 **半范数 (seminorm)**.

例 1.14 \mathbb{R}^n 和 \mathbb{C}^n 上常见的向量范数:

- 1-范数: $\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$;
- 2-范数: $\|x\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2}$;
- ∞ -范数: $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$;
- p -范数: $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad p \geq 1$.

 容易验证, 1-范数、2-范数和 ∞ -范数就是 p -范数在 $p = 1, 2, \infty$ 的特殊情形.

 p -范数也称为 Hölder 范数 [143] 或 ℓ_p 范数.

定理 1.32 (范数的连续性) 设 $\|\cdot\|$ 是 \mathbb{R}^n (或 \mathbb{C}^n) 上的一个向量范数, 则 $f(x) \triangleq \|x\|$ 是 \mathbb{R}^n (或 \mathbb{C}^n) 上的连续函数. (留作课外自习, 利用范数的三角不等式)

证明. 由范数的定义可知:

$$|f(x) - f(y)| = ||x| - |y|| \leq \|x - y\|.$$

□

定义 1.12 (范数的等价性) 设 $\|\cdot\|_\alpha$ 与 $\|\cdot\|_\beta$ 是 \mathbb{R}^n (或 \mathbb{C}^n) 上的两个向量范数, 若存在正常数 c_1, c_2 , 使得

$$c_1 \|x\|_\alpha \leq \|x\|_\beta \leq c_2 \|x\|_\alpha$$

对任意 $x \in \mathbb{R}^n$ (或 \mathbb{C}^n) 都成立, 则称 $\|\cdot\|_\alpha$ 与 $\|\cdot\|_\beta$ 是等价的.

定理 1.33 \mathbb{R}^n (或 \mathbb{C}^n) 上的所有向量范数都是等价的, 特别地, 有

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2,$$

$$\|x\|_\infty \leq \|x\|_1 \leq n \|x\|_\infty,$$

$$\|x\|_{\infty} \leq \|x\|_2 \leq \sqrt{n} \|x\|_{\infty}.$$

(板书, 以 \mathbb{R}^n 为例, 只证明等价性, 三个不等式留作练习)

证明. 只需证明任意向量范数 $\|\cdot\|$ 都与 $\|x\|_{\infty}$ 等价即可, 即存在正常数 c_1, c_2 使得

$$c_1 \|x\|_{\infty} \leq \|x\| \leq c_2 \|x\|_{\infty}, \quad \forall x \in \mathbb{R}^n.$$

考虑函数 $f(x) \triangleq \|x\|$, 则 $f(x)$ 连续且非负. 定义集合

$$S \triangleq \{x \in \mathbb{R}^n : \|x\|_{\infty} = 1\},$$

则 S 是 \mathbb{R}^n 中的有界闭集, 所以 $f(x)$ 在 S 上存在最小值和最大值, 分别记为 c_1 和 c_2 . 由于 $f(x) = 0$ 当且仅当 $x = 0$, 所以 c_1 和 c_2 都大于 0.

对任意非零向量 $x \in \mathbb{R}^n$, 有 $\frac{x}{\|x\|_{\infty}} \in S$, 所以

$$c_1 \leq f\left(\frac{x}{\|x\|_{\infty}}\right) \leq c_2, \quad \text{即} \quad c_1 \leq \frac{\|x\|}{\|x\|_{\infty}} \leq c_2.$$


所以结论成立. □

 更一般地, 根据 Jensen 不等式 [71]:

$$\left(\sum_{i=1}^n x_i^p\right)^{1/p} > \left(\sum_{i=1}^n x_i^q\right)^{1/q}, \quad x_i > 0, \quad 0 < p < q,$$

我们有

$$\|x\|_p \geq \|x\|_q, \quad \forall 1 \leq p \leq q.$$

 Jensen 不等式:


$$\left(\sum_{i=1}^n x_i^p\right)^{1/p} > \left(\sum_{i=1}^n x_i^q\right)^{1/q}, \quad x_i > 0, \quad 0 < p < q.$$

证明. 直接计算可得

$$\begin{aligned} \frac{\left(\sum_{i=1}^n x_i^q\right)^{1/q}}{\left(\sum_{i=1}^n x_i^p\right)^{1/p}} &= \left(\sum_{i=1}^n \frac{x_i^q}{\left(\sum_{j=1}^n x_j^p\right)^{q/p}}\right)^{1/q} \\ &= \left(\sum_{i=1}^n \left(\frac{x_i^p}{\sum_{j=1}^n x_j^p}\right)^{q/p}\right)^{1/q} \\ &< \left(\sum_{i=1}^n \frac{x_i^p}{\sum_{j=1}^n x_j^p}\right)^{1/q} = 1. \end{aligned}$$

□



 事实上, 有限维赋范线性空间上的所有范数都是等价的 [146].

定理 1.34 (Cauchy-Schwarz 不等式) 设 (\cdot, \cdot) 是 \mathbb{C}^n (或 \mathbb{R}^n) 上的内积, 则对任意 $x, y \in \mathbb{C}^n$ (或 \mathbb{R}^n), 有

$$|(x, y)|^2 \leq (x, x) \cdot (y, y) \quad \text{或} \quad |(x, y)| \leq \|x\| \cdot \|y\|,$$

且等号成立的充要条件是 x 与 y 线性相关, 其中 $\|\cdot\|$ 为内积导出范数, 即 $\|x\| = \sqrt{(x, x)}$.

(板书)

证明. 若 $y = 0$, 则结论显然成立.

假设 $y \neq 0$, 则对任意 $\alpha \in \mathbb{C}$ 有

$$0 \leq (x - \alpha y, x - \alpha y) = (x, x) - \bar{\alpha}(x, y) - \alpha((y, x) - \bar{\alpha}(y, y)).$$

由于 $y \neq 0$, 所以 $(y, y) > 0$. 取 $\bar{\alpha} = \frac{(y, x)}{(y, y)}$, 代入上式可得

$$0 \leq (x, x) - \frac{(y, x)}{(y, y)}(x, y).$$

由于 $(y, x) = \overline{(x, y)}$, 所以上式即为

$$|(x, y)|^2 \leq (x, x) \cdot (y, y).$$


下面考虑等号成立的条件.

充分性. 如果 x 与 y 线性相关, 则通过直接验证即可知等号成立.

必要性. 假设等号成立. 如果 $y = 0$, 则显然 x 与 y 线性相关. 现假定 $y \neq 0$. 取 $\alpha = \frac{(x, y)}{(y, y)}$, 则

$$(x - \alpha y, x - \alpha y) = (x, x) - \frac{|(x, y)|^2}{(y, y)} = 0,$$

即 $x - \alpha y = 0$. 所以 x 与 y 线性相关. □

 注记: Cauchy-Schwarz 不等式在有的文献中也称为 **Cauchy-Bunyakovski 不等式** 或 **Cauchy-Schwarz-Bunyakovski 不等式** [101]. 中学数学中的 Cauchy 不等式

$$(a_1 b_1 + a_2 b_2 + \cdots + a_n b_n)^2 \leq (a_1^2 + a_2^2 + \cdots + a_n^2)(b_1^2 + b_2^2 + \cdots + b_n^2), \quad (a_i, b_i \in \mathbb{R})$$

是 Cauchy-Schwarz 不等式的特例.

更一般地, 我们有下面的 Holder 不等式.

定理 1.35 (Holder 不等式) 设 (\cdot, \cdot) 是 \mathbb{C}^n (或 \mathbb{R}^n) 上的内积, 则对任意 $x, y \in \mathbb{C}^n$ (或 \mathbb{R}^n), 有

$$|(x, y)| \leq \|x\|_p \cdot \|y\|_q,$$

其中 $p, q > 0$, 且 $\frac{1}{p} + \frac{1}{q} = 1$.

(留作课外自习)



内积导出范数

设 S 是内积空间, 对任意 $x \in S$, 定义

$$\|x\| \triangleq (x, x)^{\frac{1}{2}}, \quad (1.6)$$

则可以验证, $\|x\|$ 是 S 上的范数. 这就是由内积导出的范数.

 任意一个内积都可以导出一个相应的范数.

例 1.15 \mathbb{R}^n 上由标准内积导出的范数为

$$\|x\| = (x, x)^{\frac{1}{2}} = \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}.$$

这就是 2-范数.

例 1.16 (极化恒等式) 设 $\|\cdot\|$ 是 \mathbb{R}^n 上由内积 (\cdot, \cdot) 导出的范数, 则有

$$(x, y) = \frac{1}{4} (\|x + y\|^2 - \|x - y\|^2).$$

证明. 由题意可知

$$\|x + y\|^2 = (x + y, x + y) = (x, x) + 2(x, y) + (y, y),$$

$$\|x - y\|^2 = (x - y, x - y) = (x, x) - 2(x, y) + (y, y).$$

两式相减即可得结论成立. □

1.3.2 矩阵范数

定义 1.13 (矩阵范数) 若函数 $f: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ (或 $f: \mathbb{C}^{m \times n} \rightarrow \mathbb{R}$) 满足

- (1) $f(A) \geq 0$, $\forall A \in \mathbb{R}^{m \times n}$ (或 $\mathbb{C}^{m \times n}$) 且等号当且仅当 $A = 0$ 时成立;
- (2) $f(\alpha A) = |\alpha| f(A)$, $\forall A \in \mathbb{R}^{m \times n}$ (或 $\mathbb{C}^{m \times n}$), $\alpha \in \mathbb{R}$ (或 \mathbb{C});
- (3) $f(A + B) \leq f(A) + f(B)$, $\forall A, B \in \mathbb{R}^{m \times n}$ (或 $\mathbb{C}^{m \times n}$);

则称 $f(X)$ 为 $\mathbb{R}^{m \times n}$ (或 $\mathbb{C}^{m \times n}$) 上的矩阵范数, 通常记作 $\|X\|$.

设 f 是 $\mathbb{R}^{n \times n}$ (或 $\mathbb{C}^{n \times n}$) 上的矩阵范数, 如果 f 还满足

- (4) $f(AB) \leq f(A)f(B)$, $\forall A, B \in \mathbb{R}^{n \times n}$ (或 $\mathbb{C}^{n \times n}$),

则称 f 是相容的矩阵范数.

注记

在本讲义中, 如果不加特别指出, 所使用的矩阵范数都是指相容的矩阵范数.

设 $\|\cdot\|$ 是 $\mathbb{R}^{m \times n}$ (或 $\mathbb{C}^{m \times n}$) 上的矩阵范数, 若对任意 $A \in \mathbb{R}^{m \times n}$ (或 $\mathbb{C}^{m \times n}$) 和任意 $x \in \mathbb{R}^n$ (或



\mathbb{C}^n), 有

$$\|Ax\|_\alpha \leq \|A\| \|x\|_\beta,$$


则称矩阵范数 $\|\cdot\|$ 与向量范数 $\|\cdot\|_\alpha$ 和 $\|\cdot\|_\beta$ **相容**, 这里的 $\|\cdot\|_\alpha$ 和 $\|\cdot\|_\beta$ 分别为 \mathbb{R}^m 和 \mathbb{R}^n (或 \mathbb{C}^m 和 \mathbb{C}^n) 上的向量范数.

一类常用的矩阵范数是由向量范数导出的算子范数.

引理 1.36 (算子范数) 设 $\|\cdot\|$ 是 \mathbb{R}^n (或 \mathbb{C}^n) 上的向量范数, 则

$$\|A\| \triangleq \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

是 $\mathbb{R}^{n \times n}$ (或 $\mathbb{C}^{n \times n}$) 上的矩阵范数, 称为 **算子范数**, 有时也称为 **诱导范数** 或 **导出范数**. (板书)

 相应地, 可以定义 $\mathbb{R}^{m \times n}$ (或 $\mathbb{C}^{m \times n}$) 上的算子范数, 此时涉及 \mathbb{R}^m 和 \mathbb{R}^n (或 \mathbb{C}^m 和 \mathbb{C}^n) 上的向量范数.

例 1.17 设 $A \in \mathbb{R}^{n \times n}$ (或 $\mathbb{C}^{n \times n}$), 常见的矩阵范数有:

- p -范数 (算子范数)

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}, \quad p \geq 1.$$

- Frobenius 范数, 简称 F -范数


$$\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2};$$

(留作课外自习, 验证满足 4 条性质)

引理 1.37 可以证明:

- (1) 矩阵 1-范数 (列范数): $\|A\|_1 = \max_{1 \leq j \leq n} \left(\sum_{i=1}^n |a_{ij}| \right);$
- (2) 矩阵 ∞ -范数 (行范数): $\|A\|_\infty = \max_{1 \leq i \leq n} \left(\sum_{j=1}^n |a_{ij}| \right);$
- (3) 矩阵 2-范数: $\|A\|_2 = \sqrt{\rho(A^*A)};$
- (4) 矩阵 F -范数: $\|A\|_F = \sqrt{\text{tr}(A^*A)}$

(板书, 以 ∞ -范数和 2-范数为例, 其他留作练习)

 这里 $\rho(\cdot)$ 表示矩阵的**谱半径**, 定义如下:

$$\rho(A) \triangleq \max_{\lambda \in \sigma(A)} |\lambda|,$$

其中 $\sigma(A)$ 表示 A 的谱, 即所有特征值组成的集合.

计算 2-范数时需要谱半径, 因此通常比计算 1-范数和 ∞ -范数更困难. 但在某些情况下可以用下面的范数等价性来估计一个矩阵的 2-范数.



定理 1.38 (矩阵范数的等价性) $\mathbb{R}^{n \times n}$ (或 $\mathbb{C}^{n \times n}$) 上的所有范数都是等价的, 特别地, 有

$$\frac{1}{\sqrt{n}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{n} \|A\|_1,$$

$$\frac{1}{\sqrt{n}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{n} \|A\|_\infty.$$

(留作练习)

除此之外, 我们还有下面的性质.

引理 1.39 设 $A \in \mathbb{R}^{n \times n}$ (或 $\mathbb{C}^{n \times n}$), 则 $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$, 且

$$\max_{1 \leq i, j \leq n} \{|a_{ij}|\} \leq \|A\|_2 \leq n \max_{1 \leq i, j \leq n} \{|a_{ij}|\}.$$

(留作练习)

矩阵范数的更多性质

设 $A \in \mathbb{R}^{n \times n}$ (或 $\mathbb{C}^{n \times n}$), $x \in \mathbb{R}^n$ (或 \mathbb{C}^n).


- (1) 对任意矩阵范数 $\|\cdot\|$, 有 $\|A^k\| \leq \|A\|^k$;
- (2) 对任意算子范数 $\|\cdot\|$, 有 $\|Ax\| \leq \|A\| \cdot \|x\|$, $\|AB\| \leq \|A\| \cdot \|B\|$;
- (3) $\|Ax\|_2 \leq \|A\|_F \cdot \|x\|_2$, $\|AB\|_F \leq \|A\|_F \cdot \|B\|_F$;
- (4) F -范数不是算子范数;
- (5) $\|\cdot\|_2$ 和 $\|\cdot\|_F$ 是酉不变范数, 即对任意酉矩阵 (或正交矩阵) U, V , 有

$$\|UA\|_2 = \|AV\|_2 = \|UAV\|_2 = \|A\|_2,$$

$$\|UA\|_F = \|AV\|_F = \|UAV\|_F = \|A\|_F$$

- (6) $\|A^*\|_2 = \|A\|_2$, $\|A^*\|_1 = \|A\|_\infty$;
- (7) 若 A 是正规矩阵, 则 $\|A\|_2 = \rho(A)$.

(留作课外自习)

 在数据处理和机器学习等学科中经常会用到下面的范数:

- 向量 ℓ_0 范数:

$$\|x\|_0 \triangleq x \text{ 中非零元素的个数}, \quad x \in \mathbb{R}^n \text{ (或 } \mathbb{C}^n \text{)}.$$

需要指出的是, 上式定义的 ℓ_0 范数并不满足向量范数定义中的条件 (2). 该范数主要用于衡量向量的稀疏性, 是压缩感知和稀疏优化中的研究对象.

- 矩阵 核范数 (Nuclear Norm):

$$\|A\|_* \triangleq \sum \sigma_i, \quad \text{其中 } \sigma_i \text{ 为 } A \text{ 的所有奇异值}, A \in \mathbb{R}^{m \times n} \text{ (或 } \mathbb{C}^{m \times n} \text{)}.$$



(关于矩阵奇异值的定义见 3.12) 根据奇异值的性质, 核范数也可以定义为

$$\|A\|_* \triangleq \operatorname{tr}(\sqrt{A^T A}).$$

矩阵直和

设 $A_i \in \mathbb{R}^{n_i \times n_i}$ (或 $\mathbb{C}^{n_i \times n_i}$), $i = 1, 2, \dots, k$, 定义直和

$$\bigoplus_{i=1}^k A_i = A_1 \oplus A_2 \oplus \cdots \oplus A_k \triangleq \begin{bmatrix} A_1 & & \\ & A_2 & \\ & & \ddots \\ & & & A_k \end{bmatrix},$$

即以 A_i 为对角块的块对角矩阵. 可以验证

$$\left\| \bigoplus_{i=1}^k A_i \right\|_p = \max_{1 \leq i \leq k} \|A_i\|_p, \quad p = 1, 2, \infty. \quad (1.7)$$

1.3.3 谱半径与范数

定理 1.40 (谱半径与范数的关系) 设 $A \in \mathbb{R}^{n \times n}$ (或 $\mathbb{C}^{n \times n}$), 则

- (1) 对任意算子范数, 有 $\rho(A) \leq \|A\|$;
- (2) 反之, 对任意 $\varepsilon > 0$, 都存在一个算子范数 $\|\cdot\|_\varepsilon$, 使得 $\|A\|_\varepsilon \leq \rho(A) + \varepsilon$, 其中范数 $\|\cdot\|_\varepsilon$ 依赖于 A 和 ε . 所以, 若 $\rho(A) < 1$, 则存在算子范数 $\|\cdot\|_\varepsilon$, 使得 $\|A\|_\varepsilon < 1$;

(板书)

证明. (1) 设 λ 是 A 的一个特征值, 对应的特征向量为 $x \neq 0$, 则由 $Ax = \lambda x$ 可得

$$|\lambda| \cdot \|x\| = \|\lambda x\| = \|Ax\| \leq \|A\| \cdot \|x\|.$$

故 $|\lambda| \leq \|A\|$. 所以

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda| \leq \|A\|.$$

(2) 用构造法证明. 设 A 的 Jordan 标准型为 J , 即

$$S^{-1}AS = J = \bigoplus_{i=1}^p J_i, \quad \text{其中} \quad J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}.$$

令 $D = \operatorname{diag}(1, \varepsilon, \varepsilon^2, \dots, \varepsilon^{n-1})$, 则

$$(SD)^{-1}A(SD) = D^{-1}JD = \bigoplus_{i=1}^p J_i^\varepsilon, \quad \text{其中} \quad J_i^\varepsilon = \begin{bmatrix} \lambda_i & \varepsilon & & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda_i \end{bmatrix}.$$



定义 $\|x\|_\varepsilon \triangleq \|(SD)^{-1}x\|_\infty$. 可以证明 $\|\cdot\|_\varepsilon$ 构成一个向量范数 (见习题 1.6). 由此可构造算子范数

$$\begin{aligned}\|A\|_\varepsilon &\triangleq \sup_{x \neq 0} \frac{\|Ax\|_\varepsilon}{\|x\|_\varepsilon} = \sup_{x \neq 0} \frac{\|(SD)^{-1}Ax\|_\infty}{\|(SD)^{-1}x\|_\infty} \\ &= \sup_{y \neq 0} \frac{\|(SD)^{-1}A(SD)y\|_\infty}{\|y\|_\infty} \\ &= \left\| \bigoplus_{i=1}^k J_i^\varepsilon \right\|_\infty = \max_{1 \leq i \leq p} \|J_i^\varepsilon\|_\infty \leq \max_{1 \leq i \leq p} \{|\lambda_i|\} + \varepsilon = \rho(A) + \varepsilon.\end{aligned}$$

□

事实上, 定理 1.40 中的结论 (1) 对任意矩阵范数都成立, 见习题 1.9.

1.3.4 最佳逼近与正交投影

下面是关于正交投影变换的一个常用性质, 可以直接通过 2-范数的定义证明.

定理 1.41 设 $P \in \mathbb{R}^{n \times n}$ (或 $\mathbb{C}^{n \times n}$) 是正交投影矩阵, 则 $\|P\|_2 = 1$, 且对 $\forall x \in \mathbb{R}^n$ (或 \mathbb{C}^n), 有

$$\|x\|_2^2 = \|Px\|_2^2 + \|(I - P)x\|_2^2.$$

正交投影可用于描述最佳逼近问题的解.

定理 1.42 设 S 是 \mathbb{R}^n (或 \mathbb{C}^n) 的子空间, $z \in \mathbb{R}^n$ (或 \mathbb{C}^n) 是一个给定的向量, 则最佳逼近问题

$$\min_{x \in S} \|x - z\|_2$$

的唯一解为

$$x_* = P_S z.$$

即 S 中距离 z 最近 (在 2-范数意义下) 的向量是 z 在 S 中的正交投影.

(留作练习)

上述定理中的 2-范数可以推广到一般的能量范数.

推论 1.43 设 $A \in \mathbb{R}^{n \times n}$ (或 $\mathbb{C}^{n \times n}$) 对称正定 (或 Hermite 正定), S 是 \mathbb{R}^n (或 \mathbb{C}^n) 的子空间, 给定 $z \in \mathbb{R}^n$ (或 \mathbb{C}^n), 则 x_* 是最佳逼近问题

$$\min_{x \in S} \|x - z\|_A$$

的解的充要条件是

$$x_* \in S \quad \text{且} \quad A(x_* - z) \perp S.$$

此处能量范数 $\|\cdot\|_A$ 的定义为: $\|x - z\|_A \triangleq \sqrt{(x - z)^* A (x - z)}$.

(留作练习)



1.4 矩阵标准型

1.4.1 Jordan 标准型

在计算矩阵的特征值时, 一个基本的思想是通过相似变换, 将其转化为一个形式尽可能简单的矩阵, 使得其特征值更易于计算. **Jordan 标准型**则是矩阵在相似变化下的最简形式.

定理 1.44 设 $A \in \mathbb{C}^{n \times n}$ (或 $\mathbb{R}^{n \times n}$) 有 p 个互不相同的特征值, 则存在非奇异矩阵 $X \in \mathbb{C}^{n \times n}$, 使得

$$X^{-1}AX = \begin{bmatrix} J_1 & & \\ & J_2 & \\ & & \ddots \\ & & & J_p \end{bmatrix} \triangleq J, \quad (1.8)$$

其中 J_i 的维数等于 λ_i 的代数重数, 且具有下面的结构

$$J_i = \begin{bmatrix} J_i^{(1)} & & \\ & J_i^{(2)} & \\ & & \ddots \\ & & & J_i^{(\nu_i)} \end{bmatrix}, \quad J_i^{(k)} = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{bmatrix}.$$

这里的 ν_i 为 λ_i 的几何重数, $J_i^{(k)}$ 称为 (对应于 λ_i 的) **Jordan 块**, J 就称为 A 的 **Jordan 标准型**.

该定理可以通过 λ -矩阵来证明 (见高等代数教材), 也可以通过后面的 Schur 分解来证明.

除了 Jordan 块的排列次序外, Jordan 标准型是唯一确定的.

可以证明, 对于每一个 Jordan 块 $J_i^{(k)}$, 都存在一个列满秩矩阵 $X_i^{(k)}$ 使得

$$AX_i^{(k)} = X_i^{(k)} J_i^{(k)}.$$

Jordan 标准型的基本性质

- Jordan 块的个数等于 A 的线性无关的特征向量的个数;
- A 可对角化的充要条件是每个 Jordan 块都是 1×1 的, 此时 X 的列向量就是 A 的特征向量.

根据 Jordan 标准型和特征值的连续性, 我们可以得到下面的结论.

推论 1.45 所有可对角化矩阵组成的集合在所有矩阵组成的集合中是稠密的, 即任何一个矩阵都可以通过可对角化矩阵来逼近.

Jordan 标准型的一个重要应用是可以用来计算矩阵的最小多项式.

定理 1.46 设 $\lambda_1, \lambda_2, \dots, \lambda_p$ 为 $A \in \mathbb{C}^{n \times n}$ 的互不相等的特征值, 则 A 的最小多项式为

$$p(\lambda) = \prod_{i=1}^q (\lambda - \lambda_i)^{r_i},$$

其中 r_i 是与 λ_i 所对应的最大 Jordan 块的维数.

1.4.2 Schur 分解

Jordan 标准型在理论研究中非常有用, 但数值计算比较困难. 下面我们介绍一个比较实用的矩阵分解, 即 **Schur 分解**.

定理 1.47 设 $A \in \mathbb{C}^{n \times n}$ (或 $\mathbb{R}^{n \times n}$), 则存在酉矩阵 $U \in \mathbb{C}^{n \times n}$ 使得

$$U^*AU = \begin{bmatrix} \lambda_1 & r_{12} & \cdots & r_{1n} \\ 0 & \lambda_2 & \cdots & r_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} \triangleq R \quad \text{或} \quad A = URU^*, \quad (1.9)$$

其中 $\lambda_1, \lambda_2, \dots, \lambda_n$ 是 A 的特征值 (可以按任意顺序排列).

(板书)

证明. 我们对 n 使用归纳法.

当 $n = 1$ 时, 结论显然成立.

假设结论对阶数为 $n - 1$ 的矩阵都成立. 考虑 n 阶矩阵 $A \in \mathbb{C}^{n \times n}$. 设 λ 是 A 的一个特征值, 其对应的单位特征向量为 $x \in \mathbb{C}^n$. 构造一个以 x 为第一列的酉矩阵 $X = [x, \tilde{X}]$. 于是

$$X^*AX = \begin{bmatrix} x^* \\ \tilde{X}^* \end{bmatrix} A \begin{bmatrix} x & \tilde{X} \end{bmatrix} = \begin{bmatrix} x^*Ax & x^*A\tilde{X} \\ \tilde{X}^*Ax & \tilde{X}^*A\tilde{X} \end{bmatrix}.$$

因为 $x^*Ax = \lambda x^*x = \lambda$, 且 $\tilde{X}^*Ax = \tilde{X}^*(\lambda x) = \lambda \tilde{X}^*x = 0$, 故

$$X^*AX = \begin{bmatrix} \lambda & x^*A\tilde{X} \\ 0 & \tilde{X}^*A\tilde{X} \end{bmatrix} \triangleq \begin{bmatrix} \lambda & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix},$$

其中 $\tilde{A}_{22} \in \mathbb{C}^{(n-1) \times (n-1)}$. 根据归纳假设, 存在酉矩阵 $\tilde{U} \in \mathbb{C}^{(n-1) \times (n-1)}$, 使得 $\tilde{U}^*\tilde{A}_{22}\tilde{U} = \tilde{R} \in \mathbb{C}^{(n-1) \times (n-1)}$ 是一个上三角矩阵. 令

$$U = X \begin{bmatrix} 1 & 0 \\ 0 & \tilde{U} \end{bmatrix},$$

则有

$$\begin{aligned} U^*AU &= \begin{bmatrix} 1 & 0 \\ 0 & \tilde{U}^* \end{bmatrix} X^*AX \begin{bmatrix} 1 & 0 \\ 0 & \tilde{U} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \tilde{U}^* \end{bmatrix} \begin{bmatrix} \lambda & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{U} \end{bmatrix} = \begin{bmatrix} \lambda & \tilde{A}_{12}\tilde{U} \\ 0 & \tilde{U}^*\tilde{A}_{22}\tilde{U} \end{bmatrix} = \begin{bmatrix} \lambda & \tilde{A}_{12}\tilde{U} \\ 0 & \tilde{R} \end{bmatrix} \triangleq R. \end{aligned}$$

由于 \tilde{R} 是上三角矩阵, 故 R 也是一个上三角矩阵, 其对角线元素即为 A 的特征值.

由归纳法可知, 定理结论成立. □



关于 Schur 分解的几点说明

- 该结论告诉我们, 任意一个矩阵都可以西三角化.
- 三角矩阵可以说是一般矩阵在酉相似变化下的最简形式.
- 定理中的 U 和 R 不是唯一的.
- R 的对角线元素可以按任意顺序排列, 特别地, 可以按模从大到小排列.

推论 1.48 设 $A \in \mathbb{C}^{n \times n}$, 则

- (1) A 是正规矩阵当且仅当 R 是对角矩阵, 即 A 可酉对角化当且仅当 A 是正规矩阵;
- (2) A 是 Hermite 矩阵当且仅当 R 是实对角矩阵.

众所周知, 当 A 是实矩阵时, 其特征值和特征向量仍可能是复的. 但在实际计算一个实矩阵的特征值时, 希望尽可能地避免复数运算. 这时, 我们就需要用到下面的 **实 Schur 分解** (或 **拟 Schur 分解**).

定理 1.49 设 $A \in \mathbb{R}^{n \times n}$, 则存在正交矩阵 $Q \in \mathbb{R}^{n \times n}$, 使得

$$Q^T A Q = T, \quad (1.10)$$

其中 $T \in \mathbb{R}^{n \times n}$ 是 **拟上三角矩阵**, 即 T 是块上三角的, 且对角块为 1×1 或 2×2 的块矩阵. 若对角块是 1×1 的, 则其就是 A 的一个特征值, 若对角块是 2×2 的, 则其特征值是 A 的一对共轭复特征值. (板书)

证明. 同样可以使用归纳法. 当 $n = 1$ 时, 结论显然成立.

假定结论对所有不超过 $n - 1$ 阶的矩阵都成立. 考虑 n 阶实矩阵 A . 设 λ 是 A 的一个特征值. 若 λ 是实的, 则存在一个对应的实特征向量, 后面的证明与定理 1.47 的证明类似.

若 λ 是复数 (虚部不为 0), 设其对应的单位复特征向量为 u . 由于

$$\bar{\lambda} \bar{u} = \overline{\lambda u} = \overline{A u} = \bar{A} \bar{u} = A \bar{u},$$

故 $(\bar{\lambda}, \bar{u})$ 也是 A 的一个特征对, 且 u 和 \bar{u} 线性无关. 令

$$\tilde{u} = \frac{1}{2}(u + \bar{u}), \quad \tilde{v} = \frac{1}{2i}(u - \bar{u}),$$

即 \tilde{u}, \tilde{v} 分别为 u 的实部与虚部, 于是 $\tilde{u} \in \mathbb{R}^n, \tilde{v} \in \mathbb{R}^n$. 由定理 1.22 可知, $\text{span}\{\tilde{u}, \tilde{v}\} = \text{span}\{u, \bar{u}\}$ 是 A 的一个不变子空间. 将 $\{\tilde{u}, \tilde{v}\}$ 进行正交化 (利用 Gram-Schmidt 正交化过程): 存在列正交矩阵 $\tilde{Q} \in \mathbb{R}^{n \times 2}$ 和非奇异上三角矩阵 $\tilde{R} \in \mathbb{R}^{2 \times 2}$, 使得 $[\tilde{u}, \tilde{v}] = \tilde{Q} \tilde{R}$. 则 $\text{span}\{\tilde{Q}\} = \text{span}\{\tilde{u}, \tilde{v}\}$ 也是 A 的不变子空间. 由定理 1.23 可知, 存在矩阵 $B \in \mathbb{R}^{2 \times 2}$ 使得 $A \tilde{Q} = \tilde{Q} B$. 将 \tilde{Q} 扩充成一个正交矩阵, 即存在矩阵 $\hat{Q} \in \mathbb{R}^{n \times (n-2)}$ 使得 $[\tilde{Q}, \hat{Q}]$ 是正交矩阵. 于是有

$$\begin{bmatrix} \tilde{Q} & \hat{Q} \end{bmatrix}^T A \begin{bmatrix} \tilde{Q} & \hat{Q} \end{bmatrix} = \begin{bmatrix} \tilde{Q}^T A \tilde{Q} & \tilde{Q}^T A \hat{Q} \\ \hat{Q}^T A \tilde{Q} & \hat{Q}^T A \hat{Q} \end{bmatrix} = \begin{bmatrix} B & \tilde{Q}^T A \hat{Q} \\ 0 & \hat{Q}^T A \hat{Q} \end{bmatrix},$$

其中 $\hat{Q}^T A \hat{Q} \in \mathbb{R}^{(n-2) \times (n-2)}$. 对 $\hat{Q}^T A \hat{Q}$ 使用归纳假设, 即可证明定理结论成立. \square

易知, 若拟上三角矩阵 T 是上三角矩阵, 则 A 的特征值都是实的. 反之, 结论也成立.



推论 1.50 设 $A \in \mathbb{R}^{n \times n}$ 的特征值都是实的, 则存在正交矩阵 $Q \in \mathbb{R}^{n \times n}$ 和上三角矩阵 $R \in \mathbb{R}^{n \times n}$ 使得

$$Q^T A Q = R,$$

其中 R 的对角线元素即为 A 的特征值.





1.5 几类特殊矩阵

1.5.1 对称正定矩阵

定义 1.14 设 $A \in \mathbb{C}^{n \times n}$.

- 若对所有向量 $x \in \mathbb{C}^n$ 有 $\operatorname{Re}(x^*Ax) \geq 0$, 则称 A 是半正定的;
- 若对所有非零向量 $x \in \mathbb{C}^n$ 有 $\operatorname{Re}(x^*Ax) > 0$, 则称 A 是正定的;
- 若 A 是 Hermite 的且半正定, 则称 A 为 Hermite 半正定;
- 若 A 是 Hermite 的且正定, 则称 A 为 Hermite 正定;
- 若 $A \in \mathbb{R}^{n \times n}$ 是对称的且半正定, 则称 A 为对称半正定;
- 若 $A \in \mathbb{R}^{n \times n}$ 是对称的且正定, 则称 A 为对称正定.

 若对所有向量 $x \in \mathbb{C}^n$ 有 $x^*Ax \in \mathbb{R}$, 则 $A^* = A$.

 本讲义中, 正定和半正定矩阵不要求是对称或 Hermite.

定理 1.51 设 $A \in \mathbb{C}^{n \times n}$, 则 A 正定 (或半正定) 的充要条件是矩阵 $H = \frac{1}{2}(A + A^*)$ 正定 (或半正定). (留作练习)


如果 A 是实矩阵, 则只需在实数域中考虑即可.

定理 1.52 设 $A \in \mathbb{R}^{n \times n}$, 则 A 正定 (或半正定) 的充要条件是对任意非零向量 $x \in \mathbb{R}^n$ 有 $x^T Ax > 0$ (或 $x^T Ax \geq 0$). (留作练习)

对称正定矩阵的基本性质

设 $A \in \mathbb{C}^{n \times n}$.

- (1) A Hermite 正定当且仅当 A Hermite 且所有特征值都是正的;
- (2) A Hermite 正定当且仅当存在酉矩阵 U 使得 $A = U\Lambda U^*$, 其中 Λ 为对角矩阵且对角线均为正实数;
- (3) A Hermite 正定当且仅当 S^*AS 对称正定, 其中 $S \in \mathbb{C}^{n \times n}$ 是一个任意的非奇异矩阵;
- (4) 若 A Hermite 正定, 则 A 的任意主子矩阵都 Hermite 正定;
- (5) 若 A Hermite 正定, 则 A 的所有对角线元素都是正的, 且 $\max_{i \neq j} \{|a_{ij}|\} < \max_i \{a_{ii}\}$, 即绝对值最大的元素出现在对角线上.

 以上结论在实数域中也成立.

平方根

如果 A 是 Hermite (半) 正定矩阵, 则可以定义其平方根, 即存在唯一的 Hermite (半) 正定矩阵 B , 使得 $B^2 = A$. 事实上, 我们有下面更一般的结论 [71].



定理 1.53 设 $A \in \mathbb{C}^{n \times n}$ 是一个 Hermite 半正定矩阵, k 是一个给定的正整数. 则存在唯一的 Hermite 半正定矩阵 $B \in \mathbb{C}^{n \times n}$ 使得

$$B^k = A.$$

同时, 我们还有下面的性质:

- (1) $BA = AB$, 且存在一个多项式 $p(t)$ 使得 $B = p(A)$;
- (2) $\text{rank}(B) = \text{rank}(A)$, 因此, 若 A 是正定的, 则 B 也正定;
- (3) 如果 A 是实矩阵的, 则 B 也是实矩阵.

特别地, 当 $k = 2$ 时, 称 B 为 A 的平方根, 记为 $A^{\frac{1}{2}}$.

(留作课外自习)

Hermite 正定矩阵与内积

Hermite 正定矩阵与内积之间有如下的一一对应关系.

定理 1.54 设 (\cdot, \cdot) 是 \mathbb{C}^n 上的一个内积, 则存在一个 Hermite 正定矩阵 $A \in \mathbb{C}^{n \times n}$ 使得

$$(x, y) = y^* Ax.$$

反之, 若 $A \in \mathbb{C}^{n \times n}$ 是 Hermite 正定矩阵, 则

$$f(x, y) \triangleq y^* Ax$$

是 \mathbb{C}^n 上的一个内积.

(留作练习)

 定理 1.54 的结论在实数域中也成立.

1.5.2 对角占优矩阵

定义 1.15 设 $A \in \mathbb{C}^{n \times n}$, 若

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$$

对所有 $i = 1, 2, \dots, n$ 都成立, 且至少有一个不等式严格成立, 则称 A 为弱行对角占优, 简称弱对角占优或对角占优. 若所有不等式都严格成立, 则称 A 是严格行对角占优, 简称严格对角占优.

 类似地, 可以定义列对角占优和严格列对角占优.

定理 1.55 若 $A \in \mathbb{C}^{n \times n}$ 是严格对角占优矩阵, 则 A 非奇异.

(留作课外自习)

证明. 我们使用反证法. 假设 A 奇异, 即 $Ax = 0$ 存在非零解, 不妨设为 $x = [x_1, x_2, \dots, x_n]^T$. 不失一般性, 设下标 k 满足 $\|x\|_\infty = |x_k|$, 则 $|x_k| > 0$. 考察 $Ax = 0$ 的第 k 个方程:

$$a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n = 0.$$



可得

$$|a_{kk}| = \frac{1}{|x_k|} \left| \sum_{j=1, j \neq i}^n a_{kj} x_j \right| \leq \sum_{j=1, j \neq i}^n |a_{kj}| \cdot \frac{|x_j|}{|x_k|} \leq \sum_{j=1, j \neq i}^n |a_{kj}|,$$

这与 A 严格对角占优矛盾. 所以 A 非奇异. \square


1.5.3 不可约矩阵

定义 1.16 设 $A \in \mathbb{C}^{n \times n}$, 如果存在置换矩阵 P 使得 PAP^T 为块上三角矩阵, 即

$$PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad A_{11} \in \mathbb{C}^{k \times k}, \quad A_{22} \in \mathbb{C}^{(n-k) \times (n-k)} \quad (1.11)$$

其中 $1 \leq k < n$, 则称 A 是 **可约的**, 否则就称 A 为 **不可约的**.

 如果 A 是 1×1 矩阵, 则 A 不可约当且仅当 A 非零.

 如果 $A \in \mathbb{C}^{n \times n}$ 是可约的, 则 A 至少有 $n-1$ 个零.

不可约的意义

若 A 可约, 即存在置换矩阵 P , 使得

$$PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

则线性方程组 $Ax = b$ 等价于 $PAP^T Px = Pb$. 记 $y \triangleq Px$, $f \triangleq Pb$, 则

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}.$$


因此, 原方程组就转化为下面两个更小规模的子方程组

$$\begin{cases} A_{22}y_2 = f_2, \\ A_{11}y_1 = f_1 - A_{12}y_2. \end{cases}$$

显然, 求解这两个方程组比求解原方程组所需的运算量更少.

对于特征值问题, A 的特征值就是 A_{11} 和 A_{22} 的特征值的并, 因此也只需考虑子矩阵 A_{11} 和 A_{22} 的特征值即可.

如果 A_{22} 或 A_{11} 仍然是可约的, 则可以转化为更小规模的子问题, 直到不可约为止. **因此我们在讨论某些算法的性质时, 一般只需考虑不可约情形.**

 由于 PAP^T 保持 A 的对角线元素仍然在对角线上, 因此由定理 1.58 可知, 主对角线元素是否为零并不影响矩阵的可约性.

推论 1.56 设 $A = [a_{ij}] \in \mathbb{C}^{n \times n}$ 不可约, 则 $A + D$ 也不可约, 其中 D 是任意对角矩阵. 特别地, $B \triangleq A - \text{diag}(a_{11}, a_{22}, \dots, a_{nn})$ 不可约.

如果 A 可约, 即存在置换矩阵 P 使得 PAP^T 具有 (1.11) 的形式, 则对任意正整数 k , 有

$$PA^k P^T = (PAP^T)^k = \begin{bmatrix} A_{11}^k & \tilde{A}_{12}^{(k)} \\ 0 & A_{22}^k \end{bmatrix}. \quad (1.12)$$

于是我们有下面的结论.

定理 1.57 设 $A \in \mathbb{C}^{n \times n}$. 若 A 可约, 则 A^k 也可约. 反之, 若存在一个正整数 k , 使得 A^k 是不可约的, 则 A 也不可约.

需要指出的是, 不可约矩阵的幂不一定是不可约的. 如 $A = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$ 不可约, 但 $A^2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ 可约.

下面的定理给出了一个矩阵不可约的充要条件.

定理 1.58 设 $A = [a_{ij}] \in \mathbb{C}^{n \times n}$, 指标集 $\mathbb{Z}_n = \{1, 2, \dots, n\}$, 则 A 可约的充要条件是存在非空指标集 $S, T \subset \mathbb{Z}_n$ 满足 $S \oplus T = \mathbb{Z}_n$, 使得

$$a_{ij} = 0, \quad i \in S, \quad j \in T.$$

(板书)

证明. 必要性. 设 A 可约, 即存在置换矩阵 P 使得 $PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$. 设 PA 是将 A 的第 i_1, i_2, \dots, i_k 行置换到前 k 行, 则 PAP^T 是将 PA 的第 i_1, i_2, \dots, i_k 列置换到前 k 列. 令 $T = \{i_1, i_2, \dots, i_k\}$, $S = \mathbb{Z} \setminus T$, 则有

$$a_{ij} = 0, \quad i \in S, \quad j \in T.$$

充分性. 设存在非空集合 S 和 T 满足 $S \oplus T = \mathbb{Z}_n$, 使得对任意 $i \in S, j \in T$ 都有 $a_{ij} = 0$. 设 $T = \{i_1, i_2, \dots, i_k\}$, 构造置换矩阵 P , 其作用是将矩阵 A 的第 i_1, i_2, \dots, i_k 行置换到前 k 行, 则 PAP^T 的第 $k+1$ 行至第 n 行和前 k 列组成的子矩阵是 0 矩阵, 故 PAP^T 具有 (1.11) 结构, 即 A 可约. \square

下面是判断矩阵是否可约的另一个充要条件.

定理 1.59 设 $A = [a_{ij}] \in \mathbb{C}^{n \times n}$, 指标集 $\mathbb{Z}_n = \{1, 2, \dots, n\}$, 则 A 可约的充要条件是存在两个相异的正整数 $k, l \in \mathbb{Z}_n$, 使得对任意指标序列 $\{i_1, i_2, \dots, i_r\} \subseteq \mathbb{Z}_n$, 都有

$$a_{ki_1} a_{i_1 i_2} \cdots a_{i_r l} = 0.$$



这里 $r > 0$ 是任意正整数.

(留作课外自习)

定理 1.59 的等价描述

矩阵 $A = [a_{ij}] \in \mathbb{C}^{n \times n}$ 不可约的充要条件是: 对任意两个相异正整数 $k, l \in \mathbb{Z}_n$, 都存在一个指标序列 $\{i_1, i_2, \dots, i_m\} \subseteq \mathbb{Z}_n$ 使得

$$a_{ki_1} a_{i_1 i_2} \cdots a_{i_m l} \neq 0.$$

我们知道, 严格对角占优矩阵是非奇异的. 如果 A 是不可约对角占优矩阵, 则可以同样证明 A 是非奇异的.

定理 1.60 设 $A \in \mathbb{C}^{n \times n}$ 是不可约的弱对角占优矩阵, 则 A 非奇异.

(留作练习)

 关于严格对角占优矩阵和不可约弱对角占优矩阵的非奇异性, 也可以使用 Gerschgorin 圆盘定理来证明, 参见 [138].

一个有意思的现象

通常, 如果某个元素全不为零的矩阵具有某种性质, 则这个性质往往能推广到不可约矩阵.

1.5.4 其它常见特殊矩阵

- 带状矩阵 (banded matrix): $a_{ij} \neq 0$ 当且仅当 $-b_u \leq i - j \leq b_l$, 其中 b_u 和 b_l 为非负整数, 分别称为**上带宽**和**下带宽**, $b_u + b_l + 1$ 称为 A 的**带宽** (bandwidth);
- 上 Hessenberg 矩阵 (upper Hessenberg matrix): $a_{ij} = 0$ for $i - j > 1$;

$$\begin{bmatrix} * & * & * & \cdots & * \\ * & * & * & \cdots & * \\ & * & * & \cdots & * \\ & & \ddots & \ddots & \vdots \\ & & & * & * \end{bmatrix}$$

- 下 Hessenberg 矩阵 (lower Hessenberg matrix): $a_{ij} = 0$ for $i - j < -1$;
- Toeplitz 矩阵:

$$T = \begin{bmatrix} t_0 & t_{-1} & \cdots & t_{-n+1} \\ t_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & \cdots & t_1 & t_0 \end{bmatrix}$$

- 循环矩阵 (circulant matrix):

$$C = \begin{bmatrix} c_0 & c_{n-1} & c_{n-2} & \cdots & c_1 \\ c_1 & c_0 & c_{n-1} & \cdots & c_2 \\ c_2 & c_1 & c_0 & \cdots & c_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{n-1} & c_{n-2} & c_{n-3} & \cdots & c_0 \end{bmatrix}$$

- Hankel 矩阵:

$$H = \begin{bmatrix} h_0 & h_1 & \cdots & h_{n-2} & h_{n-1} \\ h_1 & \ddots & \ddots & \ddots & h_n \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ h_{n-2} & \ddots & \ddots & \ddots & h_{2n-3} \\ h_{n-1} & h_n & \cdots & h_{2n-3} & h_{2n-2} \end{bmatrix}$$

- 几个能够直接写出逆的矩阵:

$$\begin{bmatrix} a_1 b_1 & & & & \\ a_2 b_1 & a_2 b_2 & & & \\ a_3 b_1 & a_3 b_2 & a_3 b_3 & & \\ \vdots & \vdots & \vdots & \ddots & \\ a_n b_1 & a_n b_2 & a_n b_3 & \cdots & a_n b_n \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{b_1 a_1} & & & & \\ -\frac{1}{b_2 a_1} & \frac{1}{b_2 a_2} & & & \\ & -\frac{1}{b_3 a_2} & \frac{1}{b_3 a_3} & & \\ & & \ddots & \ddots & \\ & & & -\frac{1}{b_n a_{n-1}} & \frac{1}{b_n a_n} \end{bmatrix}$$

特例:

$$\begin{bmatrix} 1 & & & & \\ 1 & 1 & & & \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & 1 & \end{bmatrix}^{-1} = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & & -1 & 1 \end{bmatrix}$$

$$\begin{bmatrix} b_1 & b_1 & b_1 & \cdots & b_1 \\ b_1 & b_1 + b_2 & b_1 + b_2 & \cdots & b_1 + b_2 \\ b_1 & b_1 + b_2 & b_1 + b_2 + b_3 & \cdots & b_1 + b_2 + b_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_1 & b_1 + b_2 & b_1 + b_2 + b_3 & \cdots & b_1 + b_2 + \cdots + b_n \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{b_1} + \frac{1}{b_2} & -\frac{1}{b_2} & & & \\ -\frac{1}{b_2} & \frac{1}{b_2} + \frac{1}{b_3} & -\frac{1}{b_3} & & \\ & \ddots & \ddots & & \\ & & & -\frac{1}{b_{n-1}} & \frac{1}{b_{n-1}} + \frac{1}{b_n} - \frac{1}{b_n} \\ & & & -\frac{1}{b_n} & \frac{1}{b_n} \end{bmatrix}$$



特例:


$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & \cdots & 2 \\ 1 & 2 & 3 & \cdots & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & \cdots & n \end{bmatrix}^{-1} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}$$



1.6 Kronecker 积

定义 1.17 设 $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times q}$, 则 A 与 B 的 **Kronecker 积** 定义为

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{mp \times nq}.$$

 Kronecker 积也称为**直积** (direct product), 或**张量积** (tensor product).

任意两个矩阵都存在 Kronecker 积, 且 $A \otimes B$ 和 $B \otimes A$ 是同阶矩阵. 通常 $A \otimes B \neq B \otimes A$, 但它们之间存在下面的关系式.

定理 1.61 设 $A \in \mathbb{C}^{m \times m}$, $B \in \mathbb{C}^{n \times n}$, 则存在 $m+n$ 阶置换矩阵 P 使得

$$P^T(A \otimes B)P = B \otimes A.$$

定理 1.62 矩阵的 Kronecker 积有以下性质:

- (1) $(\alpha A) \otimes B = A \otimes (\alpha B) = \alpha(A \otimes B), \quad \forall \alpha \in \mathbb{R};$
- (2) $(A \otimes B)^T = A^T \otimes B^T, \quad (A \otimes B)^* = A^* \otimes B^*;$
- (3) $(A \otimes B) \otimes C = A \otimes (B \otimes C);$
- (4) $(A + B) \otimes C = A \otimes C + B \otimes C;$
- (5) $A \otimes (B + C) = A \otimes B + A \otimes C;$
- (6) 混合积: $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$
- (7) $(A_1 \otimes A_2 \otimes \cdots \otimes A_k)(B_1 \otimes B_2 \otimes \cdots \otimes B_k)$
 $= (A_1 B_1) \otimes (A_2 B_2) \otimes \cdots \otimes (A_k B_k);$
- (8) $(A_1 \otimes B_1)(A_2 \otimes B_2) \cdots (A_k \otimes B_k) = (A_1 A_2 \cdots A_k) \otimes (B_1 B_2 \cdots B_k);$
- (9) $\text{rank}(A \otimes B) = \text{rank}(A) \text{rank}(B);$

推论 1.63 设 $A = Q_1 \Lambda_1 Q_1^{-1} \in \mathbb{C}^{m \times m}$, $B = Q_2 \Lambda_2 Q_2^{-1} \in \mathbb{C}^{n \times n}$, 则

$$A \otimes B = (Q_1 \otimes Q_2)(\Lambda_1 \otimes \Lambda_2)(Q_1 \otimes Q_2)^{-1}.$$

Kronecker 积的特征值

定理 1.64 设 $A \in \mathbb{R}^{m \times m}$, $B \in \mathbb{R}^{n \times n}$, 并设 (λ, x) 和 (μ, y) 分别是 A 和 B 的一个特征对, 则 $(\lambda\mu, x \otimes y)$ 是 $A \otimes B$ 的一个特征对. 由此可知, $B \otimes A$ 与 $A \otimes B$ 具有相同的特征值.



推论 1.65 设 $A \in \mathbb{C}^{m \times m}$, $B \in \mathbb{C}^{n \times n}$, 则 $A \otimes I_n + I_m \otimes B$ 的特征值为 $\lambda_i + \mu_j$, 其中 λ_i 和 μ_j 分别为 A 和 B 的特征值.

定理 1.66 设 $A \in \mathbb{R}^{m \times m}$, $B \in \mathbb{R}^{n \times n}$, 则

- (1) $\text{tr}(A \otimes B) = \text{tr}(A)\text{tr}(B)$;
- (2) $\det(A \otimes B) = \det(A)^n \det(B)^m$;
- (3) 若 A 和 B 都非奇异, 则 $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

Kronecker 积与向量的乘积

定理 1.67 设矩阵 $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{m \times n}$, 记 $\text{vec}(X)$ 为 X 按列拉成的 mn 维列向量, 即

$$\text{vec}(X) = [x_1^T, x_2^T, \dots, x_n^T]^T,$$

则有

$$\text{vec}(AX) = (I \otimes A)\text{vec}(X), \quad \text{vec}(XB) = (B^T \otimes I)\text{vec}(X),$$

以及

$$(A \otimes B)\text{vec}(X) = \text{vec}(BXA^T).$$

 我们称 $\text{vec}(X)$ 为**向量化算子**或**拉直算子**. 该结论对节省运算量和存储量都有好处.

Kronecker 积与矩阵方程

Kronecker 积一个重要应用是可以将某些矩阵方程转化成一般的代数方程.

定理 1.68 矩阵方程

$$AX + XB = D$$

等价于代数方程

$$(I \otimes A + B^T \otimes I)\text{vec}(X) = \text{vec}(D).$$

1.7 课后习题

练习 1.1 证明: $(A, B) \triangleq \text{tr}(B^T A)$ 是 $\mathbb{R}^{n \times n}$ 上的内积.

练习 1.2 (定理 1.33) 证明不等式:

- (1) $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2$,
- (2) $\|x\|_\infty \leq \|x\|_1 \leq n \|x\|_\infty$,
- (3) $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty$.

练习 1.3 证明:

$$(1) \|A\|_1 = \max_{1 \leq j \leq n} \left(\sum_{i=1}^n |a_{ij}| \right); \quad (2) \|A\|_2 = \max_{x \neq 0, y \neq 0} \frac{y^T A x}{\|x\|_2 \|y\|_2}.$$

练习 1.4 证明:

- (1) $\frac{1}{\sqrt{n}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{n} \|A\|_1$;
- (2) $\frac{1}{\sqrt{n}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{n} \|A\|_\infty$;
- (3) $\frac{1}{\sqrt{n}} \|A\|_F \leq \|A\|_2 \leq \|A\|_F$;
- (4) $\frac{1}{\sqrt{n}} \|A\|_F \leq \|A\|_1 \leq \sqrt{n} \|A\|_F$.

练习 1.5 (引理 1.39) 设 $A \in \mathbb{R}^{n \times n}$. 证明: $\|A\|_2^2 \leq \|A\|_1 \cdot \|A\|_\infty$, 且

$$\max_{1 \leq i, j \leq n} \{|a_{ij}|\} \leq \|A\|_2 \leq n \max_{1 \leq i, j \leq n} \{|a_{ij}|\}.$$

练习 1.6 设 $\|\cdot\|$ 是 \mathbb{R}^m 空间上的一个向量范数, $A \in \mathbb{R}^{m \times n}$, 且 $\text{rank}(A) = n$. 证明: $\|x\|_A \triangleq \|Ax\|$ 是一个向量范数. 特别地, 如果 $A \in \mathbb{R}^m$ 非奇异, 则 $\|x\|_A \triangleq \|Ax\|$ 是一个向量范数.

练习 1.7 设 $\|\cdot\|$ 是 \mathbb{R}^n 空间上的一个向量范数. 证明: $\|A^{-1}\|^{-1} = \min_{\|x\|=1} \|Ax\|$.

练习 1.8 设 $\|\cdot\|_\alpha$ 是定义在 $\mathbb{C}^{n \times n}$ 上的一个矩阵范数. 证明: 存在 \mathbb{C}^n 上的向量范数 $\|\cdot\|_\beta$, 该范数与 $\|\cdot\|_\alpha$ 相容, 即

$$\|Ax\|_\beta \leq \|A\|_\alpha \|x\|_\beta, \quad \forall A \in \mathbb{C}^{n \times n}, x \in \mathbb{C}^n.$$

练习 1.9 设 $A \in \mathbb{C}^{n \times n}$. 试证明对任意矩阵范数都有 $\rho(A) \leq \|A\|$.

练习 1.10 设 (\cdot, \cdot) 是 \mathbb{C}^n 上的内积. 证明: $\|x\| \triangleq \sqrt{(x, x)}$ 是 \mathbb{C}^n 上的一个向量范数.

练习 1.11 设 $J = \begin{bmatrix} \lambda & \varepsilon & & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda \end{bmatrix}$, 其中 $\lambda \geq 0, \varepsilon \geq 0$. 证明 $\|J\|_2 \leq \lambda + \varepsilon$.

练习 1.12 设 $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$ 且 $x \neq 0$. 证明

$$\left\| A \left(I - \frac{xx^T}{x^T x} \right) \right\|_F^2 = \|A\|_F^2 - \frac{\|Ax\|_2^2}{x^T x}.$$



练习 1.13 (定理 1.29) 设 $P \in \mathbb{R}^{n \times n}$ 是 \mathbb{S}_1 上与 \mathbb{S}_2 正交的投影矩阵. 证明:

$$P = V(W^T V)^{-1} W^T,$$

其中 $V = [v_1, v_2, \dots, v_m]$ 和 $W = [w_1, w_2, \dots, w_m]$ 的列向量组分别构成 \mathbb{S}_1 和 \mathbb{S}_2 的一组基.

练习 1.14 (定理 1.31) 证明: 投影矩阵 $P \in \mathbb{R}^{n \times n}$ 是正交投影矩阵的充要条件 $P^T = P$

练习 1.15 (定理 1.42) 设 \mathbb{S} 是 \mathbb{R}^n 的一个子空间, $z \in \mathbb{R}^n$ 是一个给定的向量, 则最佳逼近问题

$$\min_{x \in \mathbb{S}_1} \|x - z\|_2$$

的唯一解为

$$x_* = P_{\mathbb{S}} z.$$

即 \mathbb{S} 中距离 z 最近 (在 2-范数意义下) 的向量是 z 在 \mathbb{S} 中的正交投影.

练习 1.16 (推论 1.43) 设 $A \in \mathbb{R}^{n \times n}$ 对称正定, \mathbb{S} 是 \mathbb{R}^n 的一个子空间, 给定 $z \in \mathbb{R}^n$, 则 x_* 是最佳逼近问题

$$\min_{x \in \mathbb{S}} \|x - z\|_A$$

的解的充要条件是

$$x_* \in \mathbb{S} \quad \text{且} \quad A(x_* - z) \perp \mathbb{S}.$$

此处 $\|\cdot\|_A$ 的定义为: $\|x - z\|_A \triangleq \sqrt{(x - z)^T A (x - z)}$.

练习 1.17 (定理 1.51) 证明: $A \in \mathbb{C}^{n \times n}$ 正定 (半正定) 的充要条件是矩阵 $H = \frac{1}{2}(A + A^*)$ 正定 (半正定).

练习 1.18 (定理 1.52) 证明: $A \in \mathbb{R}^{n \times n}$ 正定 (半正定) 的充要条件是对任意非零向量 $x \in \mathbb{R}^n$ 有 $x^T A x > 0$ ($x^T A x \geq 0$).

练习 1.19 设 $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times m}$, 证明: $\text{tr}(AB) = \text{tr}(BA)$.

练习 1.20 (定理 1.60) 设 $A \in \mathbb{C}^{n \times n}$ 是不可约对角占优矩阵, 证明: A 非奇异.

练习 1.21 (Sherman-Morrison 公式) 设 $A \in \mathbb{R}^{n \times n}$ 非奇异, $x, y \in \mathbb{R}^n$.

证明: 若 $y^T A^{-1} x \neq 1$, 则 $A - xy^T$ 可逆, 且

$$(A - xy^T)^{-1} = A^{-1} - \frac{A^{-1} x y^T A^{-1}}{y^T A^{-1} x - 1}.$$

练习 1.22 证明下面的结论:

- (1) 设 $A, B \in \mathbb{R}^{n \times n}$ 可交换, 即 $AB = BA$. 若 A 是对角矩阵且对角线元素互不相等, 则 B 也是对角矩阵.
- (2) 设 $A, B \in \mathbb{R}^{n \times n}$ 可交换, 即 $AB = BA$. 若 A 是块对角矩阵且对角块为标量矩阵, 即 $A = \lambda_1 I_{n_1} \oplus \lambda_2 I_{n_2} \oplus \dots \oplus \lambda_p I_{n_p}$, 其中 $n_1 + n_2 + \dots + n_p = n$, 且 λ_i 互不相等, 则 B 是具有相应分块结构的块对角矩阵, 即 $B = B_1 \oplus B_2 \oplus \dots \oplus B_p$, 其中 $B_i \in \mathbb{R}^{n_i \times n_i}$, $i = 1, 2, \dots, p$.
- (3) 设 $A, B \in \mathbb{R}^{n \times n}$ 都是对称矩阵, 则 $AB = BA$ 的充要条件是存在正交矩阵 Q 使得

$$Q A Q^T = \Lambda_A, \quad Q B Q^T = \Lambda_B,$$



其中 Λ_A 和 Λ_B 分别表示由 A 和 B 的特征值构成的对角矩阵.

练习 1.23 判断下列矩阵是否可约:

$$(1) \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad (2) \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad (3) \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 2 \\ -3 & 0 & 0 \end{bmatrix}.$$

练习 1.24*(定理 1.54) 设 (\cdot, \cdot) 是 \mathbb{C}^n 上的一个内积. 证明: 存在一个 Hermite 正定矩阵 $A \in \mathbb{C}^{n \times n}$ 使得 $(x, y) = y^* A x$ 对任意 $x, y \in \mathbb{C}^n$ 都成立. 反之, 若 $A \in \mathbb{C}^{n \times n}$ 是 Hermite 正定矩阵, 则 $f(x, y) \triangleq y^* A x$ 是 \mathbb{C}^n 上的一个内积.

练习 1.25* 设 $B \in \mathbb{R}^{m \times n}$ ($m \leq n$) 是满秩矩阵, $C \in \mathbb{R}^{m \times m}$ 是对称半正定矩阵. 证明:

$$B^T(BB^T)^{-1}B - B^T(C + BB^T)^{-1}B$$

是对称半正定的.

练习 1.26* 设 P 是置换矩阵, 则 P 可以表示为一系列初等置换矩阵 (即交换单位矩阵的两行所得到的矩阵) 的乘积. 试问在什么条件下, 有 $P^T = P$.

..... 以下为可选题

练习 1.27 设 $A \in \mathbb{R}^{m \times n}$ ($m < n$) 是满秩矩阵, $Z \in \mathbb{R}^{n \times (n-m)}$ 是由 $\ker(A)$ 的一组基构成的矩阵. 证明:

$$\text{Ran}(A^T) = \text{Ker}(Z^T).$$

练习 1.28 设 $A, B \in \mathbb{R}^{n \times n}$, 证明:

- (1) 若 $AB = 0$, 则 $\text{rank}(A) + \text{rank}(B) \leq n$.
- (2) 若 $A^2 = A$, 则 $\text{rank}(A) + \text{rank}(I - A) = n$.
- (3) 若 $A^2 = I$, 则 $\text{rank}(I + A) + \text{rank}(I - A) = n$.

练习 1.29 设 $A, B \in \mathbb{R}^{n \times n}$ 都是正交矩阵, 且 $\det(A) = -\det(B)$. 证明: $A + B$ 奇异.

练习 1.30 证明: 定义在 $\mathbb{R}^{n \times n}$ 上的算子范数和 F -范数都是相容范数.

练习 1.31 证明:

- (1) 对任意的算子范数 $\|\cdot\|$, 有 $\|I\| = 1$;
- (2) 对任意的相容范数 $\|\cdot\|$, 有 $\|I\| \geq 1$.

练习 1.32 证明: $\|A\| \triangleq \max_{1 \leq i, j \leq n} |a_{ij}|$ 是矩阵范数, 但不是相容范数.

练习 1.33 设 $A \in \mathbb{R}^{n \times n}$ 是正交矩阵. 证明: $\det(A) = \pm 1$.

练习 1.34 设 $A \in \mathbb{R}^{n \times n}$. 证明: $\text{rank}(A) = 1$ 的充要条件是存在非零向量 $a, b \in \mathbb{R}^n$ 使得 $A = ab^T$.

练习 1.35 设 A_k 是 $A \in \mathbb{R}^{n \times n}$ 的一个 k 阶子矩阵. 证明: $\|A_k\|_p \leq \|A\|_p$.

练习 1.36 设 $A \in \mathbb{R}^{n \times n}$, $U, V \in \mathbb{R}^{n \times n}$ 是正交矩阵. 证明:

$$\|UAV\|_2 = \|A\|_2, \quad \|UAV\|_F = \|A\|_F.$$



练习 1.37 设 $A, B \in \mathbb{R}^{n \times n}$, 矩阵

$$C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}, \quad D = \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix}.$$

证明: $\|C\|_2 = \max\{\|A\|_2, \|B\|_2\}$, $\|D\|_2 = \max\{\|A\|_2, \|B\|_2\}$.

练习 1.38 设 $P \in \mathbb{R}^{n \times n}$ 是一个投影矩阵, 且秩为 r , 计算 P 的所有特征值.

练习 1.39 (引理 1.27) 设 $P \in \mathbb{R}^{n \times n}$ 是一个投影矩阵. 证明: $\text{Ker}(P) = \text{Ran}(I - P)$.

练习 1.40 设 P 是从 \mathbb{S} 沿 \mathbb{S}_1 到 \mathbb{S}_2 上的投影变换, $V = [v_1, v_2, \dots, v_m]$ 构成 \mathbb{S}_1 的一组基, $W = [w_1, w_2, \dots, w_{n-m}]$ 构成 \mathbb{S}_2 的一组基. 证明:

$$P = \begin{bmatrix} W & 0 \end{bmatrix} \begin{bmatrix} W & V \end{bmatrix}^{-1}.$$

练习 1.41 设 $A \in \mathbb{R}^{m \times n}$ ($m < n$) 是满秩矩阵, $Z \in \mathbb{R}^{n \times (n-m)}$ 是由 $\text{Ker}(A)$ 的一组基构成的矩阵. 证明:

- (1) $P_Z \triangleq Z(Z^\top Z)^{-1}Z^\top$ 是 $\text{Ker}(A)$ 上的正交投影算子;
- (2) $P_A \triangleq A^\top(AA^\top)^{-1}A$ 是 $\text{Ran}(A^\top)$ 上的正交投影算子;
- (3) $P_Z = I - P_A$.

练习 1.42 设 Jordan 块矩阵

$$J = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

证明:

$$J^k = \begin{bmatrix} \lambda^k & c_1^k \lambda^{k-1} & \dots & c_{n-1}^k \lambda^{k-(n-1)} \\ & \ddots & \ddots & \vdots \\ & & \ddots & c_1^k \lambda^{k-1} \\ & & & \lambda^k \end{bmatrix}, \quad \text{其中 } c_p^k \triangleq \binom{k}{p} = \begin{cases} \frac{k!}{p!(k-p)!}, & p \leq k; \\ 0, & p > k. \end{cases}$$

练习 1.43 设 $A \in \mathbb{R}^{n \times n}$ 非奇异, $X, Y \in \mathbb{R}^{n \times m}$ ($n \geq m$).

证明: 若 $Y^\top A^{-1}X - I$ 非奇异, 则 $A - XY^\top$ 可逆, 且

$$(A - XY^\top)^{-1} = A^{-1} - A^{-1}X(Y^\top A^{-1}X - I)^{-1}Y^\top A^{-1}.$$

练习 1.44* 设 $A = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$, 其中 $B, D \in \mathbb{C}^{n \times n}$ 均为上三角矩阵, 且 B 和 D 的对角线元素互不相

等. 证明: 存在矩阵 S , 使得 $S^{-1}AS = \begin{bmatrix} B & 0 \\ 0 & D \end{bmatrix}$. (提示: 可设 $S = \begin{bmatrix} I & \tilde{S} \\ 0 & I \end{bmatrix}$)

练习 1.45 设 $x, y, z \in \mathbb{R}^n$, 试证明:

- (1) $[x \otimes z, y \otimes z] = [x, y] \otimes z$;
- (2) $[x \otimes y, x \otimes z] = x \otimes [y, z]$.



练习 1.46 设 $a \in \mathbb{R}$, 证明:

$$\begin{bmatrix} 1 & -a & & & \\ & 1 & -a & & \\ & & \ddots & \ddots & \\ & & & \ddots & -a \\ & & & & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & a & a^2 & \cdots & a^{n-1} \\ & 1 & a & \ddots & \vdots \\ & & \ddots & \ddots & a^2 \\ & & & \ddots & a \\ & & & & 1 \end{bmatrix}.$$



第二讲 线性方程组直接方法

Linear algebra — in particular, *the solution of linear systems of equations* — lies at the heart of most calculations in scientific computing.

— Dongarra & Eijkhout [32], 2000.

考虑线性方程组

$$Ax = b, \quad A \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n. \quad (2.1)$$

线性方程组的求解有着非常广泛的应用背景, 科学计算中的很多问题最后都可能归结为求解一个或多个线性方程组. 从纯数学角度来看, 这个问题已经得到了完美的解决, 因为它的解可以通过行列式直接表示出来, 即 Cramer 法则. 但在实际计算中, 由于运算量增长速度太快, 当 n 较大时, 用 Cramer 法则解线性方程组是不可行的. 另外, 由于实际计算中的舍入误差, 可能会导致一系列非常严重的问题.

一个从纯数学角度看似非常简单的问题, 实际计算时可能会非常困难, 有时甚至可能是一个无法解决的问题.

一般来说, 求解线性方程组的数值方法可以分为两类: **直接法**与**迭代法**. 本讲介绍直接法, 即 Gauss 消去法. 直接法具有良好的稳定性和健壮性, 因此在工程计算中很受欢迎. 但由于运算量是 $\mathcal{O}(n^3)$, 对于大规模问题, 所需时间会很长 (这里 n 表示未知量的个数). 目前, Gauss 消去法是求解中小规模线性方程组或某些具有特殊结构的大规模稀疏线性方程组的首选方法.

Gauss 消去法的思想可以追溯到公元一世纪左右的《九章算术》, Newton, Gauss, Lagrange 等数学家都对该方法做出了贡献, 相关历史可以参见 [61].

关于线性方程组直接法的相关参考文献

- ▶ G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th, 2013. [56]
- ▶ J.W. Demmel, *Applied Numerical Linear Algebra*, 1997. [30]
- ▶ L. N. Trefethen and D. Bau, III, *Numerical Linear Algebra*, 1997. [119]
- ▶ I. S. Duff, A. M. Erisman and J. K. Reid, *Direct Methods for Sparse Matrices*, 2nd, 2017. [37]
- ▶ T. A. Davis, *Direct Methods for Sparse Linear Systems*, SIAM, 2006. [29]

后面两个文献主要是介绍大规模稀疏线性方程组的直接解法.

在本讲中, 我们总是假定系数矩阵 A 是非奇异的, 即线性方程组 (2.1) 的解存在且唯一. 另外, 为了讨论方便, 我们只考虑实数情形, 对于复系数线性方程组, 其求解方法是类似的.

2.1 LU 分解与 Gauss 消去法

2.1.1 LU 分解

Gauss 消去法本质上就是对系数矩阵 A 进行 **LU 分解**, 即将 A 分解成两个矩阵的乘积

$$A = LU, \quad (2.2)$$

其中 L 是单位下三角矩阵, U 为非奇异上三角矩阵, 然后再对两个三角方程进行求解: 假定矩阵 A 存在 LU 分解 (2.2), 则方程组 (2.1) 就转化为求解下面两个三角方程组

$$\begin{cases} Ly = b, \\ Ux = y. \end{cases}$$

显然, 上式中的两个三角方程组都非常容易求解.

基于 LU 分解的 Gauss 消去法描述如下:

算法 2.1. Gauss 消去法

- 1: 对 A 进行 LU 分解: $A = LU$, 其中 L 为单位下三角矩阵, U 为非奇异上三角矩阵;
- 2: 利用向前回代, 求解 $Ly = b$, 即得 $y = L^{-1}b$;
- 3: 利用向后回代, 求解 $Ux = y$, 即得 $x = U^{-1}y = (LU)^{-1}b = A^{-1}b$.

我们知道, 当系数矩阵 A 非奇异时, 方程组 (2.1) 总是存在唯一解. 但是, 并不是每个非奇异矩阵都存在 LU 分解.

定理 2.1 (LU 分解的存在性和唯一性) 矩阵 $A \in \mathbb{R}^{n \times n}$ 存在 LU 分解 (即存在单位下三角矩阵 L 和非奇异上三角矩阵 U 使得 $A = LU$) 的充要条件是 A 的所有顺序主子矩阵都非奇异. 进一步, 若 A 存在 LU 分解, 则分解是唯一的.

(板书)

证明. 必要性: 设 A_{11} 是 A 的 k 阶顺序主子矩阵, 其中 $1 \leq k \leq n$. 将 $A = LU$ 写成分块形式, 即

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix} = \begin{bmatrix} L_{11}U_{11} & L_{11}U_{12} \\ L_{21}U_{11} & L_{21}U_{12} + L_{22}U_{22} \end{bmatrix}.$$

可得 $A_{11} = L_{11}U_{11}$. 由于 L_{11} 和 U_{11} 均非奇异, 所以 A_{11} 也非奇异.

充分性: 用归纳法.

当 $n = 1$ 时, 结论显然成立.

假设结论对所有 $n - 1$ 阶矩阵都成立, 即对任意 $n - 1$ 阶矩阵, 如果其所有的顺序主子矩阵都非奇异, 则存在 LU 分解.

考虑 n 阶的矩阵 A , 写成分块形式

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

其中 $A_{11} \in \mathbb{R}^{(n-1) \times (n-1)}$ 是 A 的 $n - 1$ 阶顺序主子矩阵. 由归纳假设可知, A_{11} 存在 LU 分解, 即



存在单位下三角矩阵 L_{11} 和非奇异上三角矩阵 U_{11} 使得

$$A_{11} = L_{11}U_{11}.$$

令

$$L_{21} = A_{21}U_{11}^{-1}, \quad U_{12} = L_{11}^{-1}A_{12}, \quad U_{22} = A_{22} - L_{21}U_{12},$$

则

$$\begin{bmatrix} L_{11} & 0 \\ L_{21} & 1 \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix} = \begin{bmatrix} L_{11}U_{11} & L_{11}U_{12} \\ L_{21}U_{11} & U_{22} + L_{21}U_{12} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = A.$$

因此可得 A 的 LU 分解 $A = LU$, 其中 $L \triangleq \begin{bmatrix} L_{11} & 0 \\ L_{21} & 1 \end{bmatrix}$ 为单位下三角矩阵, $U \triangleq \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}$ 为非奇异的 (上三角矩阵 (U 的非奇异性可由 A 的非奇异性可得)).

由归纳法可知, 结论成立.

下面证明 **唯一性**. 设 A 存在两个不同的 LU 分解:

$$A = LU = \tilde{L}\tilde{U},$$

其中 L 和 \tilde{L} 为单位下三角矩阵, U 和 \tilde{U} 为非奇异上三角矩阵. 则有

$$L^{-1}\tilde{L} = U\tilde{U}^{-1},$$

该等式左边为下三角矩阵, 右边为上三角矩阵, 所以只能是对角矩阵. 又单位下三角矩阵的逆仍然是单位下三角矩阵, 所以 $L^{-1}\tilde{L}$ 的对角线元素全是 1, 故

$$L^{-1}\tilde{L} = I,$$

即 $\tilde{L} = L, \tilde{U} = U$. 唯一性得证. \square

记 D 为 U 的对角线部分, 则 $A = LD\tilde{U}$, 其中 $\tilde{U} = D^{-1}U$ 是单位上三角矩阵. 因此我们就有下面的 **LDU 分解**.

推论 2.2 (LDU 分解) 设 $A \in \mathbb{R}^{n \times n}$ 的所有顺序主子矩阵都非奇异, 则 A 存在 LDU 分解, 即存在单位下三角矩阵 L 和单位上三角矩阵 U , 以及非奇异对角矩阵 D , 使得 $A = LDU$, 其中 L, U, D 都是唯一的. 反之, A 存在 LDU 分解, 则 A 的所有顺序主子矩阵都非奇异.

对角占优情形

一般的非奇异矩阵不一定存在 LU 分解, 但如果 A 是对角占优的, 则一定存在 LU 分解.

定理 2.3 [56] 设 $A \in \mathbb{R}^{n \times n}$ 非奇异且列对角占优, 则 A 存在 LU 分解且 L 中的元素的绝对值都不超过 1. (留作练习, 数学归纳法)

对角占优矩阵在 LU 分解中保持对角占优非常重要. 如果是严格对角占优的, 则可以给出 $\|A^{-1}\|_1$ 的一个估计.



定理 2.4 [56] 设 $A \in \mathbb{R}^{n \times n}$ 严格列对角占优, 记

$$\delta \triangleq \min_{1 \leq j \leq n} \left(|a_{jj}| - \sum_{i=1}^n |a_{ij}| \right) > 0,$$

则有

$$\|A^{-1}\|_1 \leq \delta^{-1}.$$

(留作练习)

该定理中的 δ 是衡量对角占优程度的一个重要指标: 如果 δ 很小 (接近于 0), 则表示对角占优性很弱. 在实际计算中, 由于舍入误差的影响, 很有可能会失去对角占优性, 从而导致 LU 分解失败.

2.1.2 LU 分解的实现

矩阵的 LU 分解可以通过初等行变换来实现. 给定矩阵

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

- 第一步: 假定 $a_{11} \neq 0$, 构造矩阵

$$L_1 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & 0 & 0 & \cdots & 1 \end{bmatrix}, \quad \text{其中 } l_{i1} = \frac{a_{i1}}{a_{11}}, i = 2, 3, \dots, n.$$

易知 L_1 的逆为

$$L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -l_{21} & 1 & 0 & \cdots & 0 \\ -l_{31} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -l_{n1} & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

用 L_1^{-1} 左乘 A , 并将所得到的矩阵记为 $A^{(1)}$, 则

$$A^{(1)} = L_1^{-1}A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{bmatrix}.$$

即左乘 L_1^{-1} 后, A 的第一列中除第一个元素外其它都变为 0.

- 第二步: 类似地, 我们可以将上面的操作作用在 $A^{(1)}$ 的子矩阵 $A^{(1)}(2:n, 2:n)$ 上, 将其第一



列除第一个元素外都变为 0. 也就是说, 假定 $a_{22}^{(1)} \neq 0$, 构造矩阵

$$L_2 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & l_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & l_{n2} & 0 & \cdots & 1 \end{bmatrix}, \quad \text{其中 } l_{i2} = \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}, i = 3, 4, \dots, n.$$

用 L_2^{-1} 左乘 $A^{(1)}$, 并将所得到的矩阵记为 $A^{(2)}$, 则

$$A^{(2)} = L_2^{-1}A^{(1)} = L_2^{-1}L_1^{-1}A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}.$$

- 依此类推, 假定 $a_{kk}^{(k-1)} \neq 0$ ($k = 3, 4, \dots, n-1$), 则我们可以构造一系列的矩阵 L_3, L_4, \dots, L_{n-1} , 使得

$$L_{n-1}^{-1} \cdots L_2^{-1}L_1^{-1}A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \cdots & a_{nn}^{(n-1)} \end{bmatrix}$$

为一个上三角矩阵. 我们将这个上三角矩阵记为 U , 并记

$$L \triangleq L_1L_2 \cdots L_{n-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{bmatrix}, \quad (2.3)$$

则可得

$$A = LU,$$

这就是 A 的 LU 分解.

将上述过程写成算法, 描述如下:

算法 2.2. LU 分解

- ```

1: Set $L = I, U = 0$ % 将 L 设为单位矩阵, U 设为零矩阵
2: for $k = 1$ to $n - 1$ do
3: for $j = k$ to n do
4: $u_{kj} = a_{kj}$ % 更新 U 的第 k 行
5: end for

```



```

6: for $i = k + 1$ to n do
7: $l_{ik} = a_{ik}/a_{kk}$ % 计算 l_{ik}
8: for $j = k + 1$ to n do
9: $a_{ij} = a_{ij} - l_{ik}u_{kj}$ % 更新 $A(i, k + 1 : n)$
10: end for
11: end for
12: end for

```

### LU 分解的运算量

由算法 2.2 可知, LU 分解的运算量 (含加减乘除) 为

$$\sum_{i=1}^{n-1} \left( \sum_{j=i+1}^n 1 + \sum_{j=i+1}^n \sum_{k=i+1}^n 2 \right) = \sum_{i=1}^{n-1} (n-i+2(n-i)^2) = \frac{2}{3}n^3 + \mathcal{O}(n^2).$$

评价算法的一个主要指标是执行时间, 但这依赖于计算机硬件和编程技巧等, 因此直接给出算法执行时间是不太现实的. 所以我们通常是统计算法中算术运算 (加减乘除) 的次数. 在矩阵计算中, 大多仅仅涉及加减乘除和开方运算. 一般情况下, 加减运算次数与乘法运算次数具有相同的量级, 而除法运算和开方运算次数具有更低的量级.

为了尽可能地减少运算量, 在实际计算中, 数, 向量和矩阵做乘法运算时的先后执行次序为: 先计算数与向量的乘法, 然后计算矩阵与向量的乘法, 最后才计算矩阵与矩阵的乘法. 比如计算  $\alpha ABx$ , 其中  $\alpha$  是数,  $A, B$  是矩阵,  $x$  是向量, 如果按照从左往右计算的话, 则运算量为  $\mathcal{O}(n^3)$ , 但是如果先计算  $\alpha x$ , 然后计算  $B(\alpha x)$ , 最后再计算  $A(B(\alpha x))$  的话, 运算量则为  $\mathcal{O}(n^2)$ , 相差一个量级.

### 矩阵 $L$ 和 $U$ 的存储

当  $A$  的第  $i$  列 (严格下三角部分) 被用于计算  $L$  的第  $i$  列后, 在后面的计算中不再被使用. 而  $A$  的第  $i$  行 (上三角部分) 更新后就是  $U$  的第  $i$  行. 因此, 为了节省存储空间, 我们可以在计算过程中将  $L$  的第  $i$  列存放在  $A$  的第  $i$  列 (严格下三角部分,  $L$  的对角线全部为 1, 不需要存储), 将  $U$  的第  $i$  行存放在  $A$  的第  $i$  行 (上三角部分), 这样就不需要另外分配空间存储  $L$  和  $U$ . 计算结束后,  $A$  的上三角部分为  $U$ , 其严格下三角部分为  $L$  的绝对下三角部分. 此时算法可以描述为:

#### 算法 2.3. LU 分解 (用 $A$ 存储 $L$ 和 $U$ )

```

1: for $k = 1$ to $n - 1$ do
2: for $i = k + 1$ to n do
3: $a_{ik} = a_{ik}/a_{kk}$
4: for $j = k + 1$ to n do

```





```

5: $a_{ij} = a_{ij} - a_{ik}a_{kj}$
6: end for
7: end for
8: end for

```

根据指标的循环次序, 算法 2.3 也称为 KIJ 型 LU 分解. 在实际计算中, 我们一般不建议使用这个算法. 因为对于指标  $k$  的每次循环, 都需要更新  $A$  的第  $k+1$  至第  $n$  行. 这种反复读取数据的做法会使得计算效率大大降低.

### IKJ 型 LU 分解

如果数据是按行存储的, 如 C/C++, 为了提高计算效率, 我们可以采用下面的 IKJ 型 LU 分解算法.

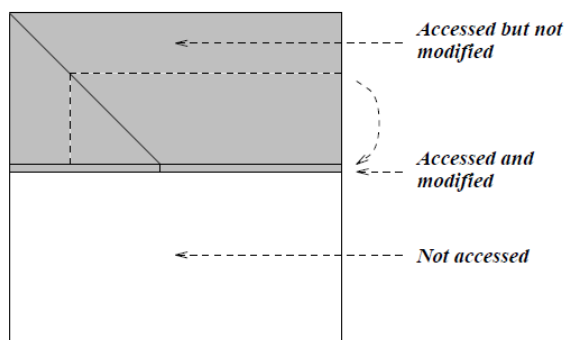
#### 算法 2.4. LU 分解 (IKJ 型)

```

1: for $i = 2$ to n do
2: for $k = 1$ to $i - 1$ do
3: $a_{ik} = a_{ik} / a_{kk}$
4: for $j = k + 1$ to n do
5: $a_{ij} = a_{ij} - a_{ik}a_{kj}$
6: end for
7: end for
8: end for

```

IKJ 型 LU 分解算法可以用下图来描述.



**思考:** 如果数据是按列存储的, 如 FORTRAN 或 MATLAB, 则怎样设计算法比较好?

### 2.1.3 Gauss 消去法

假定矩阵  $A$  存在 LU 分解  $A = LU$ , 则方程组  $Ax = b$  就转化为求解下面两个三角方程组

$$Ly = b, \quad Ux = y.$$

这两个方程组都非常容易求解, 于是 Gauss 消去法描述如下:

#### 算法 2.5. Gauss 消去法

```

1: 将 A 进行 LU 分解: $A = LU$, 其中 L 为单位下三角矩阵, U 为非奇异上三角矩阵;
2: 利用向前回代, 求解 $Ly = b$, 即得 $y = L^{-1}b$;
3: 利用向后回代, 求解 $Ux = y$, 即得 $x = U^{-1}y = (LU)^{-1}b = A^{-1}b$.

```

得到  $A$  的 LU 分解后, 我们最后需要用回代法求解两个三角方程组, 计算过程描述如下.

**算法 2.6.** 回代求解  $Ly = b$  和  $Ux = y$ 

```

1: $y_1 = b_1/l_{11}$ % 向前回代求解 $Ly = b$
2: for $i = 2 : n$ do
3: for $j = 1 : i - 1$ do
4: $b_i = b_i - l_{ij}y_j$
5: end for
6: $y_i = b_i/l_{ii}$
7: end for
8: $x_n = y_n/u_{nn}$ % 向后回代求解 $Ux = y$
9: for $i = n - 1 : -1 : 1$ do
10: for $j = n : -1 : i + 1$ do
11: $y_i = y_i - u_{ij}x_j$
12: end for
13: $x_i = y_i/u_{ii}$
14: end for

```

如果数据是按列存储的, 则采用列存储方式效率会高一些. 下面是以  $Ux = y$  为例, 描述按列存储时的回代求解过程.

**算法 2.7.** 向后回代求解  $Ux = y$  (列存储方式)

```

1: for $k = n : -1 : 1$ do
2: $x_k = y_k/u_{kk}$
3: for $i = k - 1 : -1 : 1$ do
4: $y_i = y_i - x_k u_{ik}$
5: end for
6: end for

```

这两个算法的运算量均为  $n^2 + \mathcal{O}(n)$ , 加上 LU 分解的运算量, Gauss 消去法的总运算量为  $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ .

可以证明, 以上两个算法都是向后稳定的 (componentwise backward stable) [68].

**2.1.4 选主元 LU 分解**

在 LU 分解过程中, 我们称  $a_{kk}^{(k-1)}$  为主元. 如果  $a_{kk}^{(k-1)} = 0$ , 则算法就无法进行下去. 即使  $a_{kk}^{(k-1)}$  不为零, 但如果  $|a_{kk}^{(k-1)}|$  的值很小, 由于舍入误差的原因, 也可能会给计算结果带来很大的误差. 此时我们就需要通过选主元来解决这个问题.



**例 2.1** 用 LU 分解求解线性方程组  $Ax = b$ , 其中  $A = \begin{bmatrix} 0.02 & 61.3 \\ 3.43 & -8.5 \end{bmatrix}$ ,  $b = \begin{bmatrix} 61.5 \\ 25.8 \end{bmatrix}$ , 要求在运算过程中保留 3 位有效数字.

(板书)

**解.** 根据 LU 分解待定系数法, 设

$$A = LU = \begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}.$$

直接比较等式两边可得

$$u_{11} = a_{11} = 0.02, \quad u_{12} = a_{12} = 61.3,$$

$$l_{21} = a_{21}/u_{11} = 3.43/0.02 \approx 172,$$

$$u_{22} = a_{22} - l_{21}u_{12} \approx -8.5 - 1.05 \times 10^4 \approx -1.05 \times 10^4,$$


解方程组  $Ly = b$  可得

$$y_1 = 61.5, \quad y_2 = b_2 - l_{21}y_1 \approx -1.06 \times 10^4.$$

解方程组  $Ux = y$  可得

$$x_2 = y_2/u_{22} \approx 1.01, \quad x_1 = (y_1 - u_{12}x_2)/u_{11} \approx -0.413/u_{11} \approx -20.7$$

□

 事实上, 精确解为  $x_1 = 10.0$  和  $x_2 = 1.00$ . 我们发现  $x_1$  的数值解误差非常大. 出现这个问题的原因就是  $|a_{11}|$  太小, 用它做主元时会放大舍入误差.

选主元是通过利用置换矩阵 (也称排列矩阵) 对行进行交换来实现的. 首先介绍置换矩阵的一些基本性质.

**引理 2.5** 设  $P \in \mathbb{R}^{n \times n}$  为置换矩阵,  $A \in \mathbb{R}^{n \times n}$  为任意矩阵, 则

- (1)  $PA$  相当于将  $A$  的行进行置换;  $AP$  相当于将  $A$  的列进行置换;
- (2)  $P^{-1} = P^T$ , 即  $P$  是正交矩阵;
- (3)  $\det(P) = \pm 1$ ;
- (4) 置换矩阵的乘积仍然是置换矩阵.

**定理 2.6 (选主元 LU 分解的存在性)** 设  $A \in \mathbb{R}^{n \times n}$  非奇异, 则存在置换矩阵  $P_L, P_R$ , 以及单位下三角矩阵  $L$  和非奇异上三角矩阵  $U$ , 使得

$$P_L A P_R = LU.$$

(板书)

**证明.** 用归纳法.

当  $n = 1$  时, 取  $P_L = P_R = L = 1, U = A$  即可.



假设结论对所有  $n-1$  阶矩阵都成立.

考虑  $n$  阶非奇异矩阵  $A \in \mathbb{R}^{n \times n}$ , 易知  $A$  至少存在一个非零元, 取置换矩阵  $P_1$  和  $P_2$  使得

$$P_1 A P_2 = \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

其中  $a_{11} \neq 0$ ,  $A_{22} \in \mathbb{R}^{(n-1) \times (n-1)}$ . 令

$$u_{11} = a_{11}, \quad U_{12} = A_{12}, \quad L_{21} = A_{21}/a_{11}, \quad U_{22} = A_{22} - L_{21}U_{12}.$$

则  $u_{11} \neq 0$ , 且有

$$\begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = P_1 A P_2.$$

两边取行列式可得

$$0 \neq \det(P_1 A P_2) = \det \left( \begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \right) \cdot \det \left( \begin{bmatrix} u_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix} \right) = a_{11} \cdot \det(U_{22}).$$

所以  $\det(U_{22}) \neq 0$ , 即  $U_{22} \in \mathbb{R}^{(n-1) \times (n-1)}$  非奇异. 由归纳假设可知, 存在置换矩阵  $\tilde{P}_L \in \mathbb{R}^{(n-1) \times (n-1)}$  和  $\tilde{P}_R \in \mathbb{R}^{(n-1) \times (n-1)}$ , 使得

$$\tilde{P}_L U_{22} \tilde{P}_R = \tilde{L}_{22} \tilde{U}_{22},$$

其中  $\tilde{L}_{22}$  为单位下三角矩阵,  $\tilde{U}_{22}$  为非奇异上三角矩阵. 令


$$P_L = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_L \end{bmatrix} P_1, \quad P_R = P_2 \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_R \end{bmatrix},$$


则有

$$\begin{aligned} P_L A P_R &= \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_L \end{bmatrix} P_1 A P_2 \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_R \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_L \end{bmatrix} \begin{bmatrix} 1 & 0 \\ L_{21} & I \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_R \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ \tilde{P}_L L_{21} & \tilde{P}_L \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & \tilde{P}_L^T \tilde{L}_{22} \tilde{U}_{22} \tilde{P}_R^T \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_R \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ \tilde{P}_L L_{21} & \tilde{P}_L \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_L^T \tilde{L}_{22} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \\ 0 & \tilde{U}_{22} \tilde{P}_R^T \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{P}_R \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ \tilde{P}_L L_{21} & \tilde{L}_{22} \end{bmatrix} \begin{bmatrix} u_{11} & U_{12} \tilde{P}_R \\ 0 & \tilde{U}_{22} \end{bmatrix} \triangleq LU, \end{aligned}$$

其中  $L$  为单位下三角矩阵,  $U$  为非奇异上三角矩阵, 即  $A$  存在 LU 分解.

因此, 由数学归纳法可知, 结论成立. □

 定理 2.6 也可以利用定理 2.1 的结论来证明, 见习题 2.3.

 事实上, 定理 2.6 中的  $P_L$  和  $P_R$  只有一个必需的.



### 置换矩阵的选取

**问题:** 第  $k$  步时, 如何选取置换矩阵  $P_L^{(k)}$  和  $P_R^{(k)}$ ?

**选法一.** 选取  $P_L^{(k)}$  和  $P_R^{(k)}$  使得主元为剩下的矩阵中绝对值最大, 这种选取方法称为“全主元 Gauss 消去法”, 简称 GECP (Gaussian elimination with complete pivoting);

**选法二.** 选取  $P_L^{(k)}$  和  $P_R^{(k)}$  使得主元为第  $k$  列中第  $k$  到第  $n$  个元素中, 绝对值最大, 这种选取方法称为“部分选主元 Gauss 消去法”, 简称 GEPP (Gaussian elimination with partial pivoting), 此时  $P_R^{(k)} = I$ , 因此也称为列主元 Gauss 消去法.

- 🔴 (1) GECP 比 GEPP 更稳定, 但工作量太大, 在实际应用中通常使用 GEPP 算法.
- (2) GEPP 算法能保证  $L$  所有的元素的绝对值都不超过 1.

### 算法 2.8. 部分选主元 LU 分解

```

1: $p = 1 : n$ % 用于记录置换矩阵
2: for $k = 1$ to $n - 1$ do
3: $[a_{\max}, l] = \max_{k \leq i \leq n} |a_{ik}|$ % 选列主元, 其中 l 表示主元所在的行
4: if $l \neq k$ then
5: for $j = 1$ to n do
6: $a_{tmp} = a_{kj}, a_{kj} = a_{lj}, a_{lj} = a_{tmp}$ % 交换 A 的第 k 行与第 l 行
7: end for
8: $p_{tmp} = p(k), p(k) = p(l), p(l) = p_{tmp}$ % 更新置换矩阵
9: end if
10: for $i = k + 1$ to n do
11: $a_{ik} = a_{ik} / a_{kk}$
12: for $j = k + 1$ to n do
13: $a_{ij} = a_{ij} - a_{ik} a_{kj}$
14: end for
15: end for
16: end for

```

相应的 MATLAB 程序见 [LE\\_PLU.m](#).

**例 2.2** 用部分选主元 LU 分解求解线性方程组  $Ax = b$ , 其中  $A = \begin{bmatrix} 0.02 & 61.3 \\ 3.43 & -8.5 \end{bmatrix}$ ,  $b = \begin{bmatrix} 61.5 \\ 25.8 \end{bmatrix}$ , 要求在运算过程中保留 3 位有效数字.

(板书)



**解.** 由于  $|a_{21}| > |a_{11}|$ , 根据部分选主元 LU 分解算法, 我们需要将第一行与第二行交换, 即取  $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ , 然后计算  $\tilde{A} = PA = \begin{bmatrix} 3.43 & -8.5 \\ 0.02 & 61.3 \end{bmatrix}$  的 LU 分解, 即设

$$\tilde{A} = LU = \begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}.$$

直接比较等式两边可得

$$\begin{aligned} u_{11} &= \tilde{a}_{11} = 3.43, \quad u_{12} = \tilde{a}_{12} = -8.5, \\ l_{21} &= \tilde{a}_{21}/u_{11} \approx 5.83 \times 10^{-3}, \\ u_{22} &= \tilde{a}_{22} - l_{21}u_{12} \approx 61.3 + 0.0496 \approx 61.3, \end{aligned}$$

即

$$PA \approx \begin{bmatrix} 1.00 & 0 \\ 5.83 \times 10^{-3} & 1.00 \end{bmatrix} \begin{bmatrix} 3.43 & -8.50 \\ 0 & 61.3 \end{bmatrix}.$$

解方程组  $Ly = Pb$  可得

$$y_1 = 25.8, \quad y_2 \approx 61.2.$$

解方程组  $Ux = y$  可得

$$x_2 = y_2/u_{22} \approx 0.998, \quad x_1 = (y_1 - u_{12}x_2)/u_{11} \approx 10.0.$$

与精确解  $x_1 = 10, x_2 = 1$  相比, 数值解具有 3 位有效数字. □


### 2.1.5 矩阵求逆

我们可以通过部分选主元 LU 分解来计算矩阵的逆. 设  $PA = LU$ , 则

$$A^{-1} = U^{-1}L^{-1}P,$$

等价于求解下面  $2n$  个三角线性方程组

$$Ly_i = Pe_i, \quad Ux_i = y_i, \quad i = 1, 2, \dots, n.$$

 **思考:** 也可以分别计算  $L^{-1}$  和  $U^{-1}$ , 然后相乘. 哪种方法划算?

### 2.1.6 分块 LU 分解

为了提高计算效率, 实际计算中通常采用 **分块 LU 分解**, 即

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & A_{22} & \cdots & A_{2p} \\ \vdots & & \ddots & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pp} \end{bmatrix} = \begin{bmatrix} I_1 & & & \\ L_{21} & I_2 & & \\ \vdots & \ddots & \ddots & \\ L_{p1} & \cdots & L_{p,p-1} & I_p \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} & \cdots & U_{1p} \\ & U_{22} & \cdots & U_{2p} \\ & & \ddots & \vdots \\ & & & U_{pp} \end{bmatrix} = LU,$$

其中  $A_{ii}$  是非奇异的方阵.

与定理 2.1 相类似, 对于分块 LU 分解, 我们有下面的结论.



**定理 2.7 (分块 LU 分解的存在性和唯一性)** 矩阵  $A$  存在唯一分块 LU 分解的充要条件是  $A$  的所有顺序分块主子矩阵都非奇异. (留作练习)

### 几点注记

- ▶ 分块 LU 分解不是 LU 分解, 若  $A$  存在 LU 分解, 则一定存在分块 LU 分解, 反之不成立.
- ▶ 分块 LU 分解与普通 LU 分解的总运算量是一样的, 但分块 LU 分解可以借助 3 级 BLAS 运算, 因此效率更高.
- ▶ 在计算分块 LU 分解过程中需要求解一系列小规模线性方程组 (以  $A_{ii}$  为系数矩阵), 可以采用普通 (选主元) LU 分解方法求解.



## 2.2 特殊方程组的求解

如果系数矩阵具有一定的特殊结构或性质, 则可以充分利用这些特殊结构或性质来设计高效的数值算法. 本节考虑以下特殊方程组的求解:

- 对称正定矩阵
- 对称不定矩阵
- 三对角矩阵
- 带状矩阵
- Toeplitz 矩阵

### 2.2.1 对称正定线性方程组

考虑线性方程组

$$Ax = b$$

其中  $A \in \mathbb{R}^{n \times n}$  对称正定的.

**定理 2.8 (Cholesky 分解)** 设  $A \in \mathbb{R}^{n \times n}$  对称正定, 则存在唯一的对角线元素为正的下三角矩阵  $L$ , 使得

$$A = LL^T.$$

该分解称为 **Cholesky 分解**.

(板书)

**证明.** 首先证明**存在性**, 用数学归纳法.

当  $n = 1$  时, 由  $A$  的对称正定性可知  $a_{11} > 0$ . 取  $l_{11} = \sqrt{a_{11}}$  即可.

假定结论对所有不超过  $n - 1$  阶的对称正定矩阵都成立. 设  $A \in \mathbb{R}^{n \times n}$  是  $n$  阶对称正定, 则  $A$  可分解为

$$A = \begin{bmatrix} a_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix} = \begin{bmatrix} \sqrt{a_{11}} & 0 \\ \frac{1}{\sqrt{a_{11}}} A_{12}^T & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{A}_{22} \end{bmatrix} \begin{bmatrix} \sqrt{a_{11}} & 0 \\ \frac{1}{\sqrt{a_{11}}} A_{12}^T & I \end{bmatrix}^T,$$

其中  $\tilde{A}_{22} = A_{22} - A_{12}^T A_{12} / a_{11}$ . 由于  $A$  对称正定, 所以  $\begin{bmatrix} 1 & 0 \\ 0 & \tilde{A}_{22} \end{bmatrix}$  也对称正定, 故  $\tilde{A}_{22}$  是  $n - 1$  阶对称正定矩阵. 根据归纳假设, 存在唯一的对角线元素为正的下三角矩阵  $\tilde{L}$ , 使得  $\tilde{A}_{22} = \tilde{L}\tilde{L}^T$ . 令

$$L = \begin{bmatrix} \sqrt{a_{11}} & 0 \\ \frac{1}{\sqrt{a_{11}}} A_{12}^T & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{L} \end{bmatrix} = \begin{bmatrix} \sqrt{a_{11}} & 0 \\ \frac{1}{\sqrt{a_{11}}} A_{12}^T & \tilde{L} \end{bmatrix}.$$

易知,  $L$  是对角线元素均为正的下三角矩阵, 且

$$LL^T = \begin{bmatrix} \sqrt{a_{11}} & 0 \\ \frac{1}{\sqrt{a_{11}}} A_{12}^T & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{L} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{L}^T \end{bmatrix} \begin{bmatrix} \sqrt{a_{11}} & 0 \\ \frac{1}{\sqrt{a_{11}}} A_{12}^T & I \end{bmatrix}^T = A.$$

由归纳法可知, 对任意对称正定实矩阵  $A$ , 都存在一个对角线元素为正的下三角矩阵  $L$ , 使得

$$A = LL^T.$$





唯一性可以采用反证法, 留做练习.  $\square$

 该定理也可以通过 LU 分解的存在唯一性来证明.

### Cholesky 分解的实现

我们利用待定系数法来计算对称正定矩阵的 Cholesky 分解. 设  $A = LL^T$ , 即

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & \cdots & l_{n1} \\ & l_{22} & \cdots & l_{n2} \\ & & \ddots & \vdots \\ & & & l_{nn} \end{bmatrix}.$$

直接比较等式两边的元素可得

$$a_{ij} = \sum_{k=1}^n l_{ik}l_{jk} = l_{jj}l_{ij} + \sum_{k=1}^{j-1} l_{ik}l_{jk}, \quad i, j = 1, 2, \dots, n.$$

先计算  $l_{11} = \sqrt{a_{11}}$ , 然后计算  $L$  第 1 列的其他元素:

$$l_{i1} = \frac{a_{i1}}{l_{11}}, \quad i = 2, 3, \dots, n.$$

依此类推, 可逐次计算出  $L$  的第  $2, 3, \dots, n$  列. 具体算法描述如下.

#### 算法 2.9. Cholesky 分解

```

1: for $j = 1$ to n do
2: $l_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 \right)^{1/2}$ % 先计算对角线元素
3: for $i = j + 1$ to n do
4: $l_{ij} = \frac{1}{l_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk} \right)$ % 计算 L 的第 j 列
5: end for
6: end for

```

#### 关于 Cholesky 算法的几点说明

- 与 LU 分解一样, 可以利用  $A$  的下三角部分来存储  $L$ ;
- Cholesky 分解算法的运算量为  $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ , 大约为 LU 分解的一半;
- Cholesky 分解算法是稳定的 (稳定性与全主元 Gauss 消去法相当), 故不需要选主元.
- 利用 Cholesky 分解求解对称正定线性方程组的算法也称为平方根法.

**例 2.3** 已知矩阵  $A = \begin{bmatrix} 4 & 2 & 8 & 0 \\ 2 & 10 & 10 & 9 \\ 8 & 10 & 21 & 6 \\ 0 & 9 & 6 & 34 \end{bmatrix}$ , 计算  $A$  的 Cholesky 分解.

(板书)

**解.** 设  $A$  的 Cholesky 分解为  $A = LL^T$ , 即

$$A = \begin{bmatrix} 4 & 2 & 8 & 0 \\ 2 & 10 & 10 & 9 \\ 8 & 10 & 21 & 6 \\ 0 & 9 & 6 & 34 \end{bmatrix} = \begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ l_{31} & l_{32} & l_{33} & \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ l_{31} & l_{32} & l_{33} & \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix}^T.$$

比较等式两边的第一行可得

$$l_{11}^2 = 4 \implies l_{11} = 2, \quad (l_{11} > 0)$$

$$l_{11}l_{21} = 2 \implies l_{21} = 1,$$

$$l_{11}l_{31} = 8 \implies l_{31} = 4,$$

$$l_{11}l_{41} = 0 \implies l_{41} = 0.$$

比较等式两边的第二行可得

$$l_{21}^2 + l_{22}^2 = 10 \implies l_{22}^2 = 9 \implies l_{22} = 3, \quad (l_{22} > 0)$$

$$l_{21}l_{31} + l_{22}l_{32} = 10 \implies l_{32} = (10 - 4)/3 = 2,$$

$$l_{21}l_{41} + l_{22}l_{42} = 9 \implies l_{42} = (9 - 0)/3 = 3.$$

比较等式两边的第三行可得

$$l_{31}^2 + l_{32}^2 + l_{33}^2 = 21 \implies l_{33}^2 = 21 - 16 - 4 = 1 \implies l_{33} = 1 \quad (l_{33} > 0)$$

$$l_{31}l_{41} + l_{32}l_{42} + l_{33}l_{43} = 6 \implies l_{43} = (6 - 0 - 6)/1 = 0.$$

比较等式两边的第四行可得

$$l_{41}^2 + l_{42}^2 + l_{43}^2 + l_{44}^2 = 34 \implies l_{44}^2 = 34 - 0 - 9 - 0 = 25 \implies l_{44} = 5. \quad (l_{44} > 0)$$

所以  $A$  的 Cholesky 分解为

$$A = \begin{bmatrix} 2 & & & \\ 1 & 3 & & \\ 4 & 2 & 1 & \\ 0 & 3 & 0 & 5 \end{bmatrix} \begin{bmatrix} 2 & & & \\ 1 & 3 & & \\ 4 & 2 & 1 & \\ 0 & 3 & 0 & 5 \end{bmatrix}^T.$$

□



改进的 Cholesky 分解算法 (LDL<sup>T</sup> 分解)

为了避免开方运算, 我们可以将  $A$  分解为:  $A = LDL^T$ , 即

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & & \ddots & \\ l_{n1} & \cdots & l_{n,n-1} & 1 \end{bmatrix} \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{bmatrix} \begin{bmatrix} 1 & l_{21} & \cdots & l_{n1} \\ & 1 & \cdots & l_{n2} \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}.$$

通过待定系数法可得

$$a_{ij} = \sum_{k=1}^n l_{ik} d_k l_{jk} = d_j l_{ij} + \sum_{k=1}^{j-1} l_{ik} d_k l_{jk}, \quad j = 1, 2, \dots, n, \quad i = j+1, j+2, \dots, n.$$

因此我们也可以依次计算出  $D$  和  $L$  的各个元素, 这就是 **LDL<sup>T</sup> 分解**.


基于 LDL<sup>T</sup> 分解求解对称正定线性方程组的算法称为**改进的平方根法**, 其优点是不需要计算平方根.

## 算法 2.10. 改进的平方根法

```

1: % 先计算 LDLT 分解
2: for j = 1 to n do
3: $d_j = a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 d_k$
4: for i = j + 1 to n do
5: $l_{ij} = (a_{ij} - \sum_{k=1}^{j-1} l_{ik} d_k l_{jk}) / d_j$
6: end for
7: end for
8: % 解方程组: $Ly = b$ 和 $DL^T x = y$
9: $y_1 = b_1$
10: for i = 2 to n do
11: $y_i = b_i - \sum_{k=1}^{i-1} l_{ik} y_k$
12: end for
13: $x_n = y_n / d_n$
14: for i = n - 1 to 1 do
15: $x_i = y_i / d_i - \sum_{k=i+1}^n l_{ki} x_k$
16: end for

```

 对于稀疏对称正定线性方程组, [50] 中详细讨论了各种具体的实施细节.



### 2.2.2 对称不定线性方程组

设  $A \in \mathbb{R}^{n \times n}$  是非奇异的对称不定矩阵. 若  $A$  存在 LU 分解, 即  $A = LU$ , 则可写成

$$A = LDL^T,$$

其中  $D$  是由  $U$  的对角线元素构成的对角矩阵. 然而, 当  $A$  不定时, 其 LU 分解不一定存在. 若采用选主元 LU 分解, 则其对称性将被破坏. 为了保持对称性, 在选主元时必须对行和列进行同样的置换, 即选取置换矩阵  $P$ , 使得

$$PAP^T = LDL^T. \quad (2.4)$$

通常称 (2.4) 为对称矩阵的  $LDL^T$  分解. 但遗憾的是, 这样的置换矩阵可能不一定存在, 即分解 (2.4) 不一定存在.

**例 2.4** 设对称矩阵

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

由于  $A$  的对角线元素都是 0, 对任意置换矩阵  $P$ , 矩阵  $PAP^T$  的对角线元素仍然都是 0. 因此, 矩阵  $A$  不存在  $LDL^T$  分解 (2.4).

#### 块 $LDL^T$ 分解

由于对称不定矩阵不一定存在  $LDL^T$  分解, 于是人们提出了块  $LDL^T$  分解, 不再要求  $D$  是对角矩阵, 它可以是**拟对角矩阵** (即块对角矩阵, 而且对角块至多是 2 阶的).

人们发现, 对于一个对称非奇异矩阵  $A \in \mathbb{R}^{n \times n}$ , 总存在置换矩阵  $P$  使得 (见习题 2.9)

$$PAP^T = \begin{bmatrix} B & E^T \\ E & C \end{bmatrix},$$

其中  $B \in \mathbb{R}$  或  $B \in \mathbb{R}^{2 \times 2}$ , 且非奇异. 因此可以对  $PAP^T$  进行块对角化, 即

$$PAP^T = \begin{bmatrix} I & 0 \\ EB^{-1} & I \end{bmatrix} \begin{bmatrix} B & 0 \\ 0 & C - EB^{-1}E^T \end{bmatrix} \begin{bmatrix} I & B^{-1}E^T \\ 0 & I \end{bmatrix},$$

其中  $C - EB^{-1}E^T$  是 Schur 补.

不断重复以上过程, 就可以得到  $A$  的**块  $LDL^T$  分解**:

$$PAP^T = L\tilde{D}L^T,$$

其中  $\tilde{D}$  是拟对角矩阵.

与选主元 LU 分解类似, 我们需要考虑块  $LDL^T$  分解的选主元策略, 即如何选取置换矩阵  $P$ . 一方面使得分解可以进行下去, 另一方面提高算法的稳定性. Kahan 于 1965 年首先考虑了选主元块  $LDL^T$  分解. 目前常用的策略有

- 全主元策略 (Bunch-Parlett 策略): Bunch 和 Parlett [22] 于 1971 年提出了全主元策略来选取置换矩阵, 并证明了其稳定性, 指出其向后误差上界与全主元 Gauss 消去法几乎一样 [20]. 但需要进行  $n^3/6$  次比较运算, 代价比较昂贵.



- 部分选主元策略 (Bunch-Kaufman 策略): 为了减少比较运算, Bunch 和 Kaufman [21] 于 1977 年提出了部分选主元策略. 这种策略在每次选主元时, 只需搜索两列, 因此将比较运算复杂度降低到了  $\mathcal{O}(n^2)$  量级. 而且这种选主元策略具有较满意的向后稳定性 [67, 68], 运算量为  $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ , 因此被广泛使用. 著名的程序库 LAPACK 一开始就采用了这种策略. 具体实施也可以参见 [68], 或 [31] (分块版本).
- Rook 策略: 采用 Bunch-Kaufman 策略得到的矩阵  $L$  的元素不能得到很好的控制. 为了克服由此带来的不稳定性, Ashcraft, Grimes 和 Lewis [4] 于 1998 年提出采用对称的 Rook 选主元策略, 以潜在的高比较运算量, 可以将矩阵  $L$  的元素控制在 1 左右. 这种对称的 Rook 选主元策略最先用在 Gauss 消去法中 [92]. Rook 策略整体上与 Bunch-Kaufman 策略类似, 但在选主元时加了一层迭代, 从而能够提供更高的精度. Cheng 在其博士论文 [25] 中对这两种策略进行了数值测试, 给出了两种例子, 分别说明 Rook 策略和 Bunch-Kaufman 策略各有优势. LAPACK 从 3.5.0 版本开始增加了 Rook 策略.

目前大部分软件都采用部分选主元块  $\text{LDL}^T$  分解来求解对称线性方程组. 关于 Aasen 算法和块  $\text{LDL}^T$  分解比较也可参见 [4, 12].

以上算法是针对稠密的对称不定线性方程组. 对于稀疏的对称不定矩阵, 可以采用 Duff 和 Reid 的多波前法 (Multifrontal) [38, 39], 或者 Liu 提出的稀疏阈值算法 [89].

### Aasen 算法

Aasen[1] 于 1971 年提出了下面的分解

$$PAP^T = LTL^T, \quad (2.5)$$

其中  $P$  为置换矩阵,  $L$  为单位下三角矩阵,  $T$  为对称三对角矩阵. 分解 (2.5) 本质上与部分选主元 LU 分解是一样的, 具体实施细节可参见 [68, 143].

2011 年, Rozložník, Shklarski 和 Toledo [103] 给出了基于现代化计算机结构和 BLAS-3 的分块三对角化算法. 该算法本质上是 Aasen 算法的分块形式.

Aasen 算法不如的块  $\text{LDL}^T$  分解使用广泛. 2016 年 12 月, LAPACK 发行了更新版本 3.7.0, 由 Yamazaki 加入了 Aasen 算法 (<http://www.netlib.org/lapack/lapack-3.7.0.html>).

### 2.2.3 三对角线性方程组

考虑三对角线性方程组  $Ax = f$ , 其中  $A$  是三对角矩阵:

$$A = \begin{bmatrix} b_1 & c_1 & & & \\ a_1 & \ddots & \ddots & & \\ & \ddots & \ddots & c_{n-1} & \\ & & a_{n-1} & b_n & \end{bmatrix}.$$

我们假定

$$|b_1| > |c_1| > 0, \quad |b_n| > |a_{n-1}| > 0, \quad |b_i| \geq |a_{i-1}| + |c_i|, \quad i = 2, \dots, n-1. \quad (2.6)$$

且

$$a_i c_i \neq 0, \quad i = 1, \dots, n-1. \quad (2.7)$$

即  $A$  是不可约弱对角占优的. 此时, 我们可以得到下面的三角分解

$$A = \begin{bmatrix} b_1 & c_1 & & \\ a_1 & \ddots & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ & & a_{n-1} & b_n \end{bmatrix} = \begin{bmatrix} \alpha_1 & & & \\ a_1 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & a_{n-1} & \alpha_n \end{bmatrix} \begin{bmatrix} 1 & \beta_1 & & \\ & 1 & \ddots & \\ & & \ddots & \beta_{n-1} \\ & & & 1 \end{bmatrix} \triangleq LU. \quad (2.8)$$

由待定系数法, 我们可以得到递推公式:

$$\begin{aligned} \alpha_1 &= b_1, \quad \beta_1 = c_1/\alpha_1 = c_1/b_1, \\ \begin{cases} \alpha_i = b_i - a_{i-1}\beta_{i-1}, \\ \beta_i = c_i/\alpha_i = c_i/(b_i - a_{i-1}\beta_{i-1}), \end{cases} & i = 2, 3, \dots, n-1 \\ \alpha_n &= b_n - a_{n-1}\beta_{n-1}. \end{aligned}$$

为了使得算法能够顺利进行下去, 我们需要证明  $\alpha_i \neq 0$ .

**定理 2.9** 设三对角矩阵  $A$  满足条件 (2.6) 和 (2.7). 则  $A$  非奇异, 且

- (1)  $|\alpha_1| = |b_1| > 0$ ;
- (2)  $0 < |\beta_i| < 1, i = 1, 2, \dots, n-1$ ;
- (3)  $0 < |c_i| \leq |b_i| - |a_{i-1}| < |\alpha_i| < |b_i| + |a_{i-1}|, i = 2, 3, \dots, n$ ;

(板书)

**证明.** 由于  $A$  是不可约且弱对角占优, 所以  $A$  非奇异. (见定理 1.60)

结论 (1) 是显然的.

下面我们证明结论 (2) 和 (3).

由于  $0 < |c_1| < |b_1|$ , 且  $\beta_1 = c_1/b_1$ , 所以  $0 < |\beta_1| < 1$ . 又  $\alpha_2 = b_2 - a_1\beta_1$ , 所以

$$|\alpha_2| \geq |b_2| - |a_1| \cdot |\beta_1| > |b_2| - |a_1| \geq |c_2| > 0, \quad (2.9)$$

$$|\alpha_2| \leq |b_2| + |a_1| \cdot |\beta_1| < |b_2| + |a_1|. \quad (2.10)$$

再由结论 (2.9) 和  $\beta_2$  的计算公式可知  $0 < |\beta_2| < 1$ . 类似于 (2.9) 和 (2.10), 我们可以得到

$$|\alpha_3| \geq |b_3| - |a_2| \cdot |\beta_2| > |b_3| - |a_2| \geq |c_3| > 0,$$

$$|\alpha_3| \leq |b_3| + |a_2| \cdot |\beta_2| < |b_3| + |a_2|.$$

依此类推, 我们就可以证明结论 (2) 和 (3). □

由定理 2.9 可知, 分解 (2.8) 是存在的. 因此, 原方程就转化为求解  $Ly = f$  和  $Ux = y$ . 由此便可得求解三对角线性方程组的 **追赶法** 也称为 **Thomas 算法** (1949), 其运算量大约为  $8n - 6$ .



**算法 2.11.** 追赶法

```

1: $\alpha_1 = b_1$
2: $\beta_1 = c_1/b_1$
3: $y_1 = f_1/b_1$
4: for $i = 2$ to $n - 1$ do
5: $\alpha_i = b_i - a_{i-1}\beta_{i-1}$
6: $\beta_i = c_i/\alpha_i$
7: $y_i = (f_i - a_{i-1}y_{i-1})/\alpha_i$
8: end for
9: $\alpha_n = b_n - a_{n-1}\beta_{n-1}$
10: $y_n = (f_n - a_{n-1}y_{n-1})/\alpha_n$
11: $x_n = y_n$
12: for $i = n - 1$ to 1 do
13: $x_i = y_i - \beta_i x_{i+1}$
14: end for

```

具体计算时, 由于求解  $Ly = f$  与矩阵 LU 分解是同时进行的, 因此,  $\alpha_i$  可以不用存储.

由于  $|\beta_i| < 1$ , 因此在回代求解  $x_i$  时, 误差可以得到有效控制.

需要指出的是, 我们也可以考虑下面的分解

$$A = \begin{bmatrix} b_1 & c_1 & & \\ a_1 & \ddots & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ & & a_{n-1} & b_n \end{bmatrix} = \begin{bmatrix} 1 & & & \\ \gamma_1 & 1 & & \\ & \ddots & \ddots & \\ & & \gamma_{n-1} & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 & c_1 & & \\ & \alpha_2 & \ddots & \\ & & \ddots & c_{n-1} \\ & & & \alpha_n \end{bmatrix}. \quad (2.11)$$

但此时  $|\gamma_i|$  可能大于 1. 比如  $\gamma_1 = a_1/b_1$ , 因此当  $|b_1| < |a_1|$  时,  $|\gamma_1| > 1$ . 所以在回代求解时, 误差可能得不到有效控制. 另外一方面, 计算  $\gamma_i$  时也可能会产生较大的舍入误差 (大数除以小数). 但如果  $A$  是列对角占优, 则可以保证  $|\gamma_i| < 1$ .

如果  $A$  是 (行) 对角占优, 则采用分解 (2.8); 如果  $A$  是列对角占优, 则采用分解 (2.11).

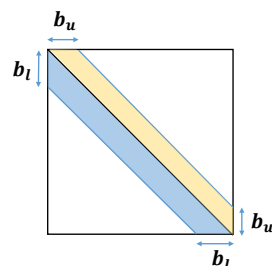
**2.2.4 带状线性方程组**

设系数矩阵  $A \in \mathbb{R}^{n \times n}$  是带状矩阵, 其下带宽为  $b_l$ , 上带宽为  $b_u$ , 即

当  $i > j + b_l$  或  $i < j - b_u$  时, 有  $a_{ij} = 0$ .

其形状如右图所示.

对于带状矩阵, 不带选主元的 LU 分解算法如下.



**算法 2.12.** 带状矩阵的 LU 分解

```

1: for $k = 1$ to $n - 1$ do
2: for $i = k + 1$ to $\min(b_l, n)$ do
3: $a_{ik} = a_{ik} / a_{kk}$
4: for $j = k + 1$ to $\min(k + b_u, n)$ do
5: $a_{ij} = a_{ij} - a_{ik}a_{kj}$
6: end for
7: end for
8: end for

```

如果  $A$  存在 LU 分解, 则可以验证,  $L$  和  $U$  也是带状矩阵.

**定理 2.10** 设  $A \in \mathbb{R}^{n \times n}$  是带状矩阵, 其下带宽为  $b_l$ , 上带宽为  $b_u$ . 若  $A$  存在不带选主元的 LU 分解  $A = LU$ , 则  $L$  为下带宽  $b_l$  的带状矩阵,  $U$  为上带宽  $b_u$  的带状矩阵, 求解  $Ax = b$  的总运算量大约为  $2nb_lb_u + n(3b_l + 2b_u + 1)$ . (留作课外自习)

若采用部分选主元的 LU 分解, 则  $L$  和  $U$  会有所变化, 但仍具有一定的特殊结构.

**定理 2.11** 设  $A \in \mathbb{R}^{n \times n}$  是带状矩阵, 其下带宽为  $b_l$ , 上带宽为  $b_u$ . 设  $PA = LU$  是  $A$  的部分选主元 LU 分解, 则  $U$  为上带宽不超过  $b_l + b_u$  的带状矩阵,  $L$  为下带宽为  $b_l$  的“基本带状矩阵”, 即  $L$  每列的非零元素不超过  $b_l + 1$  个. (留作课外自习)

**2.2.5 Toeplitz 线性方程组**

设  $T_n \in \mathbb{R}^{n \times n}$  是 Toeplitz 矩阵, 即

$$T_n = \begin{bmatrix} t_0 & t_{-1} & \cdots & t_{-n+1} \\ t_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_{-1} \\ t_{n-1} & \cdots & t_1 & t_0 \end{bmatrix}.$$

可知 Toeplitz 矩阵是反向对称 (persymmetric) 矩阵, 即关于东北-西南对角线对称. 记  $J_n$  为  $n$  阶反向单位矩阵, 即

$$J_n = \begin{bmatrix} & & & 1 \\ & & 1 & \\ & \ddots & & \\ 1 & & & \end{bmatrix}.$$

易知  $J_n^T = J_n^{-1} = J_n$ .

**引理 2.12** 矩阵  $A \in \mathbb{R}^{n \times n}$  是反向对称矩阵当且仅当

$$A = J_n A^T J_n \quad \text{或} \quad J_n A = A^T J_n.$$





若  $A$  可逆, 则可得

$$A^{-1} = J_n^{-1} (A^T)^{-1} J_n^{-1} = J_n (A^{-1})^T J_n,$$

即反向对称矩阵的逆也是反向对称矩阵.

 Toeplitz 矩阵的逆是反向对称矩阵, 但不一定是 Toeplitz 矩阵.

一般情况下, Toeplitz 矩阵的乘积不再是一个 Toeplitz 矩阵, 但如果是两个上三角或下三角 Toeplitz 矩阵相乘, 则乘积仍然是 Toeplitz 矩阵.

**定理 2.13** 设  $T, S \in \mathbb{R}^{n \times n}$  是上三角 Toeplitz 矩阵, 则  $TS$  也是上三角 Toeplitz 矩阵. 进一步, 若  $T$  非奇异, 则  $T^{-1}$  也是上三角 Toeplitz 矩阵.

### Yule-Walker 方程组

我们先考虑一类特殊右端项的线性方程组. 设  $T_n$  对称正定, 考虑线性方程组

$$T_n x = -r_n, \quad (2.12)$$

其中  $r_n = [t_1, t_2, \dots, t_{n-1}, t_n]^T$ . 这类线性方程组称为 **Yule-Walker 方程组**, 其中  $t_n$  为任意给定的实数.

由于  $T_n$  对称正定, 所以  $t_0 > 0$ . 因此我们可以对  $T_n$  的对角线元素进行单位化. 不失一般性, 我们假定  $T_n$  的对角线元素为 1, 即

$$T_n = \begin{bmatrix} 1 & t_1 & \cdots & t_{n-1} \\ t_1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_1 \\ t_{n-1} & \cdots & t_1 & 1 \end{bmatrix}.$$

由于方程组右端项的特殊性, 我们可以通过递推来求解.

设  $x^{(k)}$  是  $T_k x = -r_k$  的解, 下面导出  $T_{k+1} x = -r_{k+1}$  的解  $x^{(k+1)}$ . 记

$$x^{(k+1)} = \begin{bmatrix} z^{(k)} \\ \alpha_k \end{bmatrix},$$

则  $T_{k+1} x^{(k+1)} = -r_{k+1}$  可写为

$$\begin{bmatrix} T_k & J_k r_k \\ r_k^T J_k & 1 \end{bmatrix} \begin{bmatrix} z^{(k)} \\ \alpha_k \end{bmatrix} = - \begin{bmatrix} r_k \\ t_{k+1} \end{bmatrix}.$$

因此可得

$$z^{(k)} = T_k^{-1} (-r_k - \alpha_k J_k r_k) = x^{(k)} - \alpha_k T_k^{-1} J_k r_k, \quad (2.13)$$

$$\alpha_k = -t_{k+1} - r_k^T J_k z^{(k)}. \quad (2.14)$$

由于  $T_k$  是反向对称矩阵, 故  $T_k^{-1} J_k = J_k T_k^{-1}$ . 所以可得

$$z^{(k)} = x^{(k)} - \alpha_k T_k^{-1} J_k r_k = x^{(k)} - \alpha_k J_k T_k^{-1} r_k = x^{(k)} + \alpha_k J_k x^{(k)}.$$



代入 (2.14) 可得

$$(1 + r_k^T x^{(k)}) \alpha_k = -t_{k+1} - r_k^T J_k x^{(k)}.$$

又

$$\begin{bmatrix} I & J_k x^{(k)} \\ 0 & 1 \end{bmatrix}^T \begin{bmatrix} T_k & J_k r_k \\ r_k^T J_k & 1 \end{bmatrix} \begin{bmatrix} I & J_k x^{(k)} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} T_k & 0 \\ 0 & 1 + r_k^T x^{(k)} \end{bmatrix},$$

由  $T_{k+1}$  的对称正定性可知  $1 + r_k^T x^{(k)} > 0$ , 故可得  $x^{(k+1)}$  的计算公式

$$\alpha_k = \frac{-t_{k+1} - r_k^T J_k x^{(k)}}{1 + r_k^T x^{(k)}}, \quad z^{(k)} = x^{(k)} + \alpha_k J_k x^{(k)}. \quad k = 1, 2, \dots \quad (2.15)$$

运算量为  $\mathcal{O}(k)$ . 因此, 我们就可以从一阶 Yule-Walker 方程出发, 利用递推公式 (2.15) 计算  $T_n x = -r_n$  的解. 总的运算量大约为  $3n^2$ .


为了减少运算量, 我们引入一个变量  $\beta_k \triangleq 1 + r_k^T x^{(k)}$ , 则

$$\begin{aligned} \beta_{k+1} &= 1 + r_{k+1}^T x^{(k+1)} \\ &= 1 + [r_k^T, t_{k+1}] \begin{bmatrix} x^{(k)} + \alpha_k J_k x^{(k)} \\ \alpha_k \end{bmatrix} \\ &= 1 + r_k^T x^{(k)} + \alpha_k (t_{k+1} + r_k^T J_k x^{(k)}) \\ &= (1 - \alpha_k^2) \beta_k. \end{aligned}$$

于是可得求解 Yule-Walker 方程组的 **Levinson-Durbin 算法** [36, 85] (由 Levinson 和 Durbin 分别于 1946 年和 1960 年独立提出 [78]), 总运算量大约为  $2n^2$ .

#### 算法 2.13. 求解 Yule-Walker 方程组的 Levinson-Durbin 算法

- 1: 输入数据:  $t = [t_1, t_2, \dots, t_n]$     % 注: 这里假定  $t_0 = 1$
- 2:  $x(1) = -t_1, \beta = 1, \alpha = -t_1$
- 3: **for**  $k = 1$  **to**  $n - 1$  **do**
- 4:     $\beta = (1 - \alpha^2) \beta$
- 5:     $\alpha = -\frac{1}{\beta} \left( t_{k+1} - \sum_{i=1}^k t_i x(k+1-i) \right)$
- 6:     $x(1:k) = x(1:k) + \alpha x(k:-1:1)$
- 7:     $x(k+1) = \alpha$
- 8: **end for**

 由 (2.15) 可知, 只需  $T_{k+1}$  非奇异, 就能确保  $1 + r_k^T x^{(k)} \neq 0$ , 算法就能顺利进行下去. 所以 Levinson-Durbin 算法有意义的充要条件是  $T_n$  的所有顺序主子式非零, 此时我们称  $T_n$  是**强正则** (strongly regular) 的.



## 一般右端项的对称正定 Toeplitz 线性方程组

考虑一般右端项的方程组

$$T_n x = b,$$

其中  $T_n$  是对称正定 Toeplitz 矩阵,  $b = [b_1, b_2, \dots, b_n]^T$ . 与求解 Yule-Walker 方程组类似, 我们利用递推方法来求解.

假定  $x^{(k)}$  和  $y^{(k)}$  分别是方程组

$$T_k x = [b_1, b_2, \dots, b_k]^T \triangleq b^{(k)}$$

和

$$T_k y = -[t_1, t_2, \dots, t_k]^T$$

的解. 设  $x^{(k+1)} = \begin{bmatrix} z^{(k)} \\ \mu_k \end{bmatrix}$  是  $T_{k+1} x = b^{(k+1)}$  的解, 则可得

$$\begin{bmatrix} T_k & J_k r_k \\ r_k^T J_k & 1 \end{bmatrix} \begin{bmatrix} z^{(k)} \\ \mu_k \end{bmatrix} = \begin{bmatrix} b^{(k)} \\ b_{k+1} \end{bmatrix}.$$

通过计算可得

$$\begin{aligned} z^{(k)} &= T_k^{-1} b^{(k)} - \mu_k T_k^{-1} J_k r_k = x^{(k)} - \mu_k J_k T_k^{-1} r_k = x^{(k)} + \mu_k J_k y^{(k)}, \\ \mu_k &= \frac{b_{k+1} - r_k^T J_k x^{(k)}}{1 + r_k^T y^{(k)}}. \end{aligned}$$

所以, 我们可以先计算  $T_k x = b^{(k)}$  和  $T_k x = -r_k$  的解, 然后利用上述公式得到  $T_{k+1} x = b^{(k+1)}$  的解, 这就是 Levinson-Durbin 算法, 该算法的总运算量大约为  $4n^2$ .

**算法 2.14.** 求解对称正定 Toeplitz 线性方程组的 Levinson-Durbin 算法

- 1: 输入数据:  $t = [t_1, t_2, \dots, t_{n-1}]$  和  $b = [b_1, b_2, \dots, b_n]$     % 这里假定  $t_0 = 1$
- 2:  $y(1) = -t_1, x(1) = b_1, \beta = 1, \alpha = -t_1$
- 3: **for**  $k = 1$  to  $n - 1$  **do**
- 4:     $\beta = (1 - \alpha^2)\beta$
- 5:     $\mu = \frac{1}{\beta} \left( b_{k+1} - \sum_{i=1}^k t_i x(k+1-i) \right)$
- 6:     $x(1:k) = x(1:k) + \mu y(k:-1:1)$
- 7:     $x(k+1) = \mu$
- 8:    **if**  $k < n - 1$  **then**
- 9:         $\alpha = -\frac{1}{\beta} \left( t_{k+1} + \sum_{i=1}^k t_i y(k+1-i) \right)$
- 10:         $y(1:k) = y(1:k) + \alpha y(k:-1:1)$
- 11:         $y(k+1) = \alpha$
- 12:    **end if**
- 13: **end for**



在数学与工程的许多应用中都会出现 Toeplitz 线性方程组, 如样条插值, 时间序列分析, Markov 链, 排队论, 信号与图像处理等. Levinson-Durbin 算法最早用于求解线性预测问题中的 Toeplitz 线性方程组 (即 Yule-Walker 方程组) [85]

$$T_{n+1} \begin{bmatrix} x \\ 1 \end{bmatrix} = \begin{bmatrix} 0_n \\ E_n \end{bmatrix},$$

其中  $x \in \mathbb{R}^n$  和  $E_n \in \mathbb{R}$  是未知量,  $0_n$  表示长度为  $n$  的零向量. 显然该线性方程组等价于求解方程组 (2.12). 在线性预测模型中,  $T_{n+1}$  表示相关矩阵 (correlation matrix),  $E_n$  表示 (均方根) 预测误差 (root mean square prediction error).

通过 Levinson-Durbin 算法可以直接得到  $T_n$  的  $\text{LDL}^T$  分解 [78], 因此对称正定 Toeplitz 矩阵的  $\text{LDL}^T$  分解运算量为  $\mathcal{O}(n^2)$ .



## 2.3 扰动分析

在实际应用中, 所给的数据 (系数矩阵  $A$  和右端项  $b$ ) 往往是通过实验或观测等方式获得的, 因此通常是带有误差的, 这也导致最后求得的数值解也是有误差的. 本节就讨论原始数据误差对最后数值解的影响.

### 2.3.1 矩阵条件数

首先介绍一个重要概念, 即矩阵条件数.

**定义 2.1** 考虑线性方程组  $Ax = b$ , 如果  $A$  或  $b$  的微小变化会导致解的巨大变化, 则称此线性方程组是**病态**的, 反之则是**良态**的.

**例 2.5** 考虑线性方程组  $Ax = b$ , 其中  $A = \begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix}$ ,  $b = [2, 2]^T$ , 可求得解为  $x = [2, 0]^T$ . 如果  $b$  的第二个元素出现细微的偏差, 变为  $b = [2, 2.0001]^T$ , 则解就变为  $x = [1, 1]^T$ . 由此可见, 当右端项出现细微变化时, 解会出现很大的变化, 因此该线性方程组是病态的.

线性方程组是否病态主要取决于其系数矩阵. 怎样来判断一个矩阵是否病态? 目前比较常用的一个指标就是**矩阵条件数**.

**定义 2.2** 设  $A$  非奇异,  $\|\cdot\|$  是任一算子范数, 则称

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

为  $A$  的**条件数**.


 常用的矩阵条件数有

$$\kappa_2(A) \triangleq \|A^{-1}\|_2 \|A\|_2, \quad \kappa_1(A) \triangleq \|A^{-1}\|_1 \|A\|_1, \quad \kappa_\infty(A) \triangleq \|A^{-1}\|_\infty \|A\|_\infty.$$

  $\kappa_2(A)$  也称为**谱条件数**, 当  $A$  对称时, 有

$$\kappa_2(A) = \frac{\max_{1 \leq i \leq n} |\lambda_i|}{\min_{1 \leq i \leq n} |\lambda_i|}.$$

一般情况下, 如果没有特别指出, 则  $\kappa(A)$  指的是矩阵  $A$  的谱条件数.

 条件数是衡量一个矩阵是否病态的主要指标. 当矩阵条件数比较大时, 我们就称这个矩阵是**病态** (或者**坏条件**) 的. 由 (2.16) 可知, 如果矩阵是病态的, 则近似解的误差受数据扰动的影响就可能很大.

**例 2.6 Hilbert 矩阵**是一个典型的病态矩阵, 其定义如下:

$$H_n = [h_{ij}]_{n \times n}, \quad \text{其中 } h_{ij} = \frac{1}{i+j-1}.$$

可以验证  $H_n$  是对称正定的, 但随着  $n$  的增长, 其条件数会快速增长, 见下表:

**表 2.1.** Hilbert 矩阵的条件数

| $n$             | 2       | 3       | 4       | 5       | 6       | 7       | 8       | 9       | 10      |
|-----------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| $\kappa_2(H_n)$ | 1.9e+01 | 5.2e+02 | 1.6e+04 | 4.8e+05 | 1.5e+07 | 4.8e+08 | 1.5e+10 | 4.9e+11 | 1.6e+13 |

### 2.3.2 $\delta x$ 与 $\hat{x}$ 的关系

设  $x_*$  是精确解,  $\hat{x}$  是通过数值计算得到的近似解. 假定  $\hat{x}$  满足线性方程组

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

下面讨论  $\delta x \triangleq \hat{x} - x_*$  的大小, 即**向后误差分析**.

**定理 2.14** 设  $\|\cdot\|$  是任一向量范数 (当该范数作用在矩阵上时就是相应的导出范数), 则  $\delta x$  与  $\hat{x}$  满足下面的关系式

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \|A^{-1}\| \|A\| \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \|\hat{x}\|} \right).$$

当  $\delta b = 0$  时, 有

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \kappa(A) \frac{\|\delta A\|}{\|A\|}, \quad (2.16)$$

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

(板书)

**证明.** 由等式  $(A + \delta A)\hat{x} = b + \delta b = Ax_* + \delta b$  可知  $A(\hat{x} - x_*) = -\delta A\hat{x} + \delta b$ , 即

$$\delta x = A^{-1}(-\delta A\hat{x} + \delta b).$$

所以

$$\|\delta x\| \leq \|A^{-1}\| \cdot (\|\delta A\| \cdot \|\hat{x}\| + \|\delta b\|), \quad (2.17)$$

即

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \|A^{-1}\| \cdot \|A\| \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \cdot \|\hat{x}\|} \right).$$

若  $\delta b = 0$ , 则可得

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \kappa(A) \frac{\|\delta A\|}{\|A\|}.$$

□

 由 (2.16) 可知, 如果矩阵是病态的, 则近似解的误差受数据扰动的影响就可能会很大.

### 2.3.3 $\delta x$ 与 $x_*$ 的关系



**引理 2.15** 设  $\|\cdot\|$  是任一算子范数,  $B \in \mathbb{R}^{n \times n}$ . 若  $\|B\| < 1$ , 则  $I - B$  可逆, 且有

$$(I - B)^{-1} = \sum_{k=0}^{\infty} B^k \quad \text{和} \quad \|(I - B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$

(板书)

**证明.** 由  $\|B\| < 1$  可知  $\rho(B) < 1$ , 所以  $I - B$  的特征值都具有正实部, 故  $I - B$  非奇异.

下面证明  $(I - B)^{-1} = \sum_{k=0}^{\infty} B^k$ . 首先证明级数  $\sum_{k=0}^{\infty} B^k$  收敛, 即其每个分量所对应的级数都收敛. 记  $b_{ij}^{(k)}$  为  $B^k$  的第  $(i, j)$  元素. 由范数的等价性可知, 存在常数  $c$  使得对任意矩阵  $X \in \mathbb{R}^{n \times n}$  都有  $\|X\|_F \leq c\|X\|$ . 所以

$$|b_{ij}^{(k)}| \leq \|B^k\|_F \leq c\|B^k\| \leq c\|B\|^k.$$

注意, 这里的常数  $c$  与  $B$  和  $k$  都无关. 由条件  $\|B\| < 1$  可知, 级数  $\sum_{k=0}^{\infty} c\|B\|^k$  收敛, 所以级数  $\sum_{k=0}^{\infty} b_{ij}^{(k)}$  也收敛, 即  $\sum_{k=0}^{\infty} B^k$  收敛.

因为  $\lim_{k \rightarrow \infty} \|B^k\| = 0$ , 且  $(I - B)(I + B + B^2 + \cdots + B^k) = I - B^{k+1}$ , 两边取极限可得

$$(I - B) \sum_{k=0}^{\infty} B^k = \lim_{k \rightarrow \infty} (I - B^{k+1}) = I,$$

即

$$(I - B)^{-1} = \sum_{k=0}^{\infty} B^k,$$

且

$$\|(I - B)^{-1}\| = \left\| \sum_{k=0}^{\infty} B^k \right\| \leq \sum_{k=0}^{\infty} \|B^k\| \leq \sum_{k=0}^{\infty} \|B\|^k = \frac{1}{1 - \|B\|}.$$

□

由  $(A + \delta A)\hat{x} = b + \delta b$  可得

$$\begin{aligned} \delta x &= (A + \delta A)^{-1}(b + \delta b - Ax_* - \delta Ax_*) \\ &= (I + A^{-1}\delta A)^{-1}A^{-1}(-\delta Ax_* + \delta b). \end{aligned}$$

假定  $\|\delta A\|$  很小, 满足  $\|A^{-1}\delta A\| \leq \|A^{-1}\| \|\delta A\| < 1$ , 则由引理 2.15 可得

$$\begin{aligned} \frac{\|\delta x\|}{\|x_*\|} &\leq \|(I + A^{-1}\delta A)^{-1}\| \|A^{-1}\| \left( \|\delta A\| + \frac{\|\delta b\|}{\|x_*\|} \right) \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\delta A\|} \left( \|\delta A\| + \frac{\|\delta b\|}{\|x_*\|} \right) \\ &= \frac{\|A^{-1}\| \|A\|}{1 - \|A^{-1}\| \|A\| \frac{\|\delta A\|}{\|A\|}} \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|A\| \|x_*\|} \right) \\ &\leq \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}} \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right) \end{aligned}$$



当  $\|\delta A\| \rightarrow 0$  时, 我们可得

$$\frac{\|\delta x\|}{\|x_*\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}} \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right) \rightarrow \kappa(A) \frac{\|\delta b\|}{\|b\|}.$$

**定理 2.16** 设  $A \in \mathbb{R}^{n \times n}$  非奇异且  $\|A^{-1}\| \|\delta A\| < 1$ , 则

$$\frac{\|\delta x\|}{\|x_*\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}} \left( \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right). \quad (2.18)$$

如果  $\|\delta A\| = 0$ , 则

$$\frac{1}{\kappa(A)} \frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x_*\|} \leq \kappa(A) \frac{\|\delta b\|}{\|b\|}. \quad (2.19)$$

(板书)

**证明.** 只需证明 (2.19) 中的左边一个不等式即可. 由于  $\delta A = 0$ , 所以  $A\delta x = \delta b$ . 两边取范数, 然后同除  $\|x_*\|$  可得

$$\frac{\|A\| \cdot \|\delta x\|}{\|x_*\|} \geq \frac{\|A\delta x\|}{\|A^{-1}b\|} \geq \frac{\|\delta b\|}{\|A^{-1}\| \cdot \|b\|}.$$

所以结论成立.  $\square$

**定理 2.17** 设  $A \in \mathbb{R}^{n \times n}$  非奇异, 则有

$$\min \left\{ \frac{\|\delta A\|_2}{\|A\|_2} : A + \delta A \text{ 奇异} \right\} = \frac{1}{\kappa_2(A)}$$

(板书)

**证明.** 记  $d \triangleq \min \{ \|\delta A\|_2 : A + \delta A \text{ 奇异} \}$ , 只需证明  $d = \frac{1}{\|A^{-1}\|_2}$ .

先证明  $d \geq \frac{1}{\|A^{-1}\|_2}$ . 若  $\|\delta A\|_2 < \|A^{-1}\|_2^{-1}$ , 则

$$\|A^{-1}\delta A\|_2 \leq \|A^{-1}\|_2 \cdot \|\delta A\|_2 < 1.$$

由引理 2.15 可知  $I + A^{-1}\delta A$  非奇异. 因此  $A + \delta A = A(I + A^{-1}\delta A)$  也非奇异, 这表明使得  $A + \delta A$  奇异的  $\delta A$  必须满足  $\|\delta A\|_2 \geq \|A^{-1}\|_2^{-1}$ , 即

$$d \geq \frac{1}{\|A^{-1}\|_2}.$$

下面证明  $d \leq \frac{1}{\|A^{-1}\|_2}$ , 即证明存在  $\delta A$  满足  $\|\delta A\|_2 = \|A^{-1}\|_2^{-1}$  使得  $A + \delta A$  奇异. 由范数的定义可知

$$\|A^{-1}\|_2 = \max_{\|x\|_2=1} \|A^{-1}x\|_2,$$

故存在  $x$  满足  $\|x\|_2 = 1$  使得

$$\|A^{-1}\|_2 = \|A^{-1}x\|_2.$$

令  $y = A^{-1}x / \|A^{-1}x\|_2$ , 则  $\|y\|_2 = 1$ , 且

$$\|xy^T\|_2 = \max_{\|z\|_2=1} \|xy^T z\|_2 = \max_{\|z\|_2=1} |y^T z| \cdot \|x\|_2 = \max_{\|z\|_2=1} |y^T z|.$$

由于  $|y^T z| \leq \|y\|_2 \cdot \|z\|_2 = 1$ , 且当  $z = y$  时有  $|y^T z| = 1$ , 所以  $\|xy^T\|_2 = 1$ . 构造

$$\delta A = -\frac{xy^T}{\|A^{-1}\|_2},$$





则

$$\|\delta A\|_2 = \frac{\|xy^T\|_2}{\|A^{-1}\|_2} = \frac{1}{\|A^{-1}\|_2}.$$

下面证明  $A + \delta A$  奇异. 我们只需证明以  $A + \delta A$  为系数矩阵的齐次线性方程组有非零解. 由于  $\|A^{-1}x\|_2 = \|A^{-1}\|_2$ , 容易验证

$$(A + \delta A)y = A \frac{A^{-1}x}{\|A^{-1}x\|_2} - \frac{xy^T}{\|A^{-1}\|_2} y = \frac{x}{\|A^{-1}\|_2} - \frac{x}{\|A^{-1}\|_2} = 0,$$

即  $A + \delta A$  奇异, 所以  $d \leq \frac{1}{\|A^{-1}\|_2}$ .

综上所述可得

$$d = \min \{ \|\delta A\|_2 : A + \delta A \text{ 奇异} \} = \frac{1}{\|A^{-1}\|_2}.$$

□

 定理 2.17 的结论对所有  $p$ -范数都成立, 参见 [48, 77].

 度量

$$\text{dist}_p(A) \triangleq \min \left\{ \frac{\|\delta A\|_p}{\|A\|_p} : A + \delta A \text{ 奇异} \right\} = \frac{1}{\kappa_p(A)},$$

表示  $A$  距离奇异矩阵集合的相对距离.

### 2.3.4 $\delta x$ 与残量的关系

这是研究线性方程组的扰动理论的一个较实用的方法.

记残量 (残差) 为  $r = b - A\hat{x}$ , 则有

$$\delta x = \hat{x} - x_* = \hat{x} - A^{-1}b = A^{-1}(A\hat{x} - b) = -A^{-1}r,$$

所以可得

$$\|\delta x\| \leq \|A^{-1}\| \|r\|.$$

这个估计式的优点是不用去估计  $\delta A$  和  $\delta b$  的大小. 由于在实际计算中,  $r$  通常是可以计算的, 因此该估计式比较实用.

**定理 2.18** 设  $A \in \mathbb{R}^{n \times n}$  非奇异,  $\|\cdot\|$  为任一算子范数. 记  $r = b - A\hat{x}$ , 则

- (1) 若存在  $\hat{A}$  满足  $\hat{A}\hat{x} = b$ , 则  $\|\hat{A} - A\| \geq \frac{\|r\|}{\|\hat{x}\|}$ ;
- (2) 存在  $\delta A$  满足  $\|\delta A\| = \frac{\|r\|}{\|\hat{x}\|}$ , 使得  $(A + \delta A)\hat{x} = b$ .

(板书)

**证明.** (1) 由  $\hat{A}\hat{x} = b$  可知

$$(\hat{A} - A)\hat{x} = b - A\hat{x} = r.$$

所以有

$$\|r\| = \|(\hat{A} - A)\hat{x}\| \leq \|\hat{A} - A\| \cdot \|\hat{x}\|,$$

即

$$\|\hat{A} - A\| \geq \frac{\|r\|}{\|\hat{x}\|}.$$

(2) 以 2-范数为例, 取  $\delta A = \frac{r\hat{x}^\top}{\|\hat{x}\|_2^2}$  即可. □

### 2.3.5 相对扰动分析

前面给出了解的误差  $\delta x$  的界是与条件数  $\kappa(A)$ ,  $\delta A$  和  $\delta b$  成比例的. 在许多情况下, 这个界是令人满意的. 但有时会相差很大, 这个界就不能很好的反映实际计算中解的误差.

**例 2.7** 设  $A = \begin{bmatrix} \gamma & 0 \\ 0 & 1 \end{bmatrix}$ ,  $b = \begin{bmatrix} \gamma \\ 1 \end{bmatrix}$ , 其中  $\gamma > 1$ . 则  $Ax = b$  的精确解为  $x_* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ , 任何合理的直接法求得的解的误差都很小. 但系数矩阵的谱条件数为  $\kappa_2(A) = \gamma$ , 当  $\gamma$  很大时,  $\kappa_2(A)$  也很大, 因此误差界 (2.16) 和 (2.18) 可以是很大.

针对这个问题, 我们按分量进行分析. 记

$$\delta A = \begin{bmatrix} \delta a_{11} & \\ & \delta a_{22} \end{bmatrix}, \quad \delta b = \begin{bmatrix} \delta b_1 \\ \delta b_2 \end{bmatrix},$$

并设  $|\delta a_{ij}| \leq \varepsilon |a_{ij}|$ ,  $|\delta b_i| \leq \varepsilon |b_i|$ . 则

$$\delta x = \begin{bmatrix} \hat{x}_1 - x_1 \\ \hat{x}_2 - x_2 \end{bmatrix} = \begin{bmatrix} \frac{\delta b_1 + b_1}{\delta a_{11} + a_{11}} - 1 \\ \frac{\delta b_2 + b_2}{\delta a_{22} + a_{22}} - 1 \end{bmatrix} = \begin{bmatrix} \frac{\delta b_1 + \gamma}{\delta a_{11} + \gamma} - 1 \\ \frac{\delta b_2 + 1}{\delta a_{22} + 1} - 1 \end{bmatrix} = \begin{bmatrix} \frac{\delta b_1 - \delta a_{11}}{\delta a_{11} + \gamma} \\ \frac{\delta b_2 - \delta a_{22}}{\delta a_{22} + 1} \end{bmatrix}.$$

故

$$\|\delta x\|_\infty \leq \frac{2\varepsilon}{1 - \varepsilon}.$$

如果  $\delta b = 0$ , 则

$$\|\delta x\|_\infty \leq \frac{\varepsilon}{1 - \varepsilon}.$$

这个界与 (2.16) 或 (2.18) 相差约  $\gamma$  倍.

### 相对条件数

为了得到更好误差界, 我们引入**相对条件数**  $\kappa_{cr}(A)$ , 即

$$\kappa_{cr}(A) \triangleq \| |A^{-1}| |A| \|,$$

有时也称为 Bauer 条件数或 Skeel 条件数.

假定  $\delta A$  和  $\delta b$  满足  $|\delta A| \leq \varepsilon |A|$  和  $|\delta b| \leq \varepsilon |b|$ . 则由  $(A + \delta A)\hat{x} = b + \delta b$  可得

$$\begin{aligned} |\delta x| &= |A^{-1}(-\delta A\hat{x} + \delta b)| \\ &\leq |A^{-1}| (|\delta A|\hat{x} + |\delta b|) \\ &\leq |A^{-1}| (\varepsilon |A|\hat{x} + \varepsilon |b|) \\ &= \varepsilon |A^{-1}| (|A|\hat{x} + |b|). \end{aligned} \tag{2.20}$$



若  $\delta b = 0$ , 则有

$$\|\delta x\| = \|\delta x\| \leq \varepsilon \| |A^{-1}| |A| |\hat{x}| \| \leq \varepsilon \| |A^{-1}| |A| \| \|\hat{x}\|,$$

即

$$\frac{\|\delta x\|}{\|\hat{x}\|} \leq \| |A^{-1}| |A| \| \varepsilon = \kappa_{cr}(A) \varepsilon. \quad (2.21)$$

相对条件数有下面的性质

**引理 2.19** 设  $A \in \mathbb{R}^{n \times n}$  非奇异,  $D \in \mathbb{R}^{n \times n}$  为非奇异对角矩阵, 则

$$\kappa_{cr}(DA) = \kappa_{cr}(A).$$

**定理 2.20** 设  $A \in \mathbb{R}^{n \times n}$  非奇异, 使得  $|\delta A| \leq \varepsilon |A|$ ,  $|\delta b| \leq \varepsilon |b|$  成立, 且满足

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

的最小的  $\varepsilon > 0$  称为按分量的相对向后误差, 其表达式为

$$\varepsilon = \max_{1 \leq i \leq n} \frac{|r_i|}{(|A| |\hat{x}| + |b|)_i},$$

其中  $r = b - A\hat{x}$ .

更多关于数值计算的稳定性和矩阵扰动分析方面的知识, 可以参考 [68, 116, 141].

## 2.4 误差分析

### 2.4.1 LU 分解的舍入误差分析

关于 LU 分解的舍入误差分析, 我们有下面的结果.

**定理 2.21** [68] 假定  $A \in \mathbb{R}^{n \times n}$  的所有顺序主子式都不为 0, 则带舍入误差的 LU 分解可表示为

$$A = LU + E,$$

其中误差  $E$  满足

$$|E| \leq \gamma_n |L| \cdot |U|.$$

这里  $\gamma_n = \frac{n\varepsilon_u}{1 - n\varepsilon_u}$ ,  $\varepsilon_u$  表示机器精度.

### 2.4.2 Gauss 消去法的舍入误差分析

**引理 2.22** [68] 设  $\hat{y}$  和  $\hat{x}$  分别是由向前回代算法 ?? 和向后回代算法 2.7 计算得到的数值解, 则

$$\begin{aligned}(L + \delta L)\hat{y} &= b, & |\delta L| &\leq \gamma_n |L| \\ (U + \delta U)\hat{x} &= \hat{y}, & |\delta U| &\leq \gamma_n |U|.\end{aligned}$$

该引理表明,  $\hat{y}$  和  $\hat{x}$  只有很小的误差, 因此向前回代算法和向后回代算法都是稳定的. 于是

$$\begin{aligned}b &= (L + \delta L)\hat{y} \\ &= (L + \delta L)(U + \delta U)\hat{x} \\ &= (LU + L \cdot \delta U + \delta L \cdot U + \delta L \cdot \delta U)\hat{x} \\ &= (A - E + L \cdot \delta U + \delta L \cdot U + \delta L \cdot \delta U)\hat{x}.\end{aligned}$$

记  $\delta A = -E + L \cdot \delta U + \delta L \cdot U + \delta L \cdot \delta U$ , 则  $\hat{x}$  是扰动方程  $(A + \delta A)x = b$  精确解, 且

$$\begin{aligned}|\delta A| &= |-E + L \cdot \delta U + \delta L \cdot U + \delta L \cdot \delta U| \\ &\leq |E| + |L| \cdot |\delta U| + |\delta L| \cdot |U| + |\delta L| \cdot |\delta U| \\ &\leq (3\gamma_n + \gamma_n^2)|L| \cdot |U| \leq \gamma_{3n}|L| \cdot |U|,\end{aligned}$$

其中  $\gamma_{3n} = \frac{3n\varepsilon_u}{1 - 3n\varepsilon_u}$ . 两边取范数后可得

$$\|\delta A\| \leq 3n\varepsilon_u \|L\| \cdot \|U\|$$

对 1-范数,  $\infty$ -范数和  $F$ -范数成立 (2-范数不成立).

根据算法向后稳定性的定义, 要说明带选主元 Gauss 消去法是向后稳定的, 必须要求  $\|\delta A\|$  是“小”的, 即

$$\|\delta A\| = \mathcal{O}(\varepsilon_u) \|A\|.$$

数值试验表明, 部分选主元 Gauss 消去法几乎总是保持

$$\|L\| \cdot \|U\| \approx \|A\|.$$



记

$$\rho_n \triangleq \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}$$

为部分选主元 Gauss 消去法的**增长因子**, 其中  $a_{ij}^{(k)}$  是部分选主元 Gauss 消去法过程第  $k$  步时  $a_{ij}$  的值. 由于  $\|L\|_\infty \leq n$ ,  $\|U\|_\infty \leq n\rho_n\|A\|_\infty$ , 因此 [68]

**定理 2.23** 设  $\hat{x}$  是由部分选主元 Gauss 法得到的数值解, 则  $\hat{x}$  满足

$$(A + \delta A)\hat{x} = b, \quad \|\delta A\|_\infty \leq n^2 \gamma_{3n} \rho_n \|A\|_\infty. \quad (2.22)$$

所以若  $\rho_n$  比较小或随着  $n$  变大时增长比较缓慢, 则当  $n$  不是很大时, 部分选主元 Gauss 消去法是向后稳定的. 遗憾的是, 理论上无法保证  $\rho_n$  比较小 [68, page 166].

**定理 2.24** 部分选主元 Gauss 消去法能保证  $\rho_n \leq 2^{n-1}$ , 且这个界是可以达到的.

事实上, (2.22) 中的界几乎总是远远大于真正的  $\|\delta A\|$ .

在绝大多数情况下, 部分选主元 Gauss 消去法是向后稳定的, 但理论上也存在失败的例子.

全主元 Gauss 消去法是数值稳定的. 在大部分实际应用中, 部分选主元 Gauss 消去法与全主元 Gauss 消去法具有同样的数值稳定性.

## 2.5 解的改进

当矩阵  $A$  是病态时, 即使残量  $r = b - A\hat{x}$  很小, 所求得的数值解  $\hat{x}$  仍可能带有较大的误差. 此时需要通过一些方法来提高解的精度.

### 2.5.1 高精度运算

在计算中, 尽可能采用高精度的运算. 比如, 原始数据是单精度的, 但在计算时都采用双精度运算, 或者更高精度的运算. 但更高精度的运算会带来更大的开销.

### 2.5.2 矩阵元素缩放或平衡 (Scaling or equilibration)

在求解线性方程组时, 先对系数矩阵元素进行适当的缩放是一种很常用的技术手段.

一方面, 如果  $A$  的元素在数量级上相差很大, 则在计算过程中很可能会出现大数与小数的加减运算, 这样就可能会引入更多的舍入误差. 为了避免由于这种情况而导致的舍入误差, 我们可以在求解之前先对矩阵元素进行缩放 (Scaling), 即在矩阵两边同时乘以两个适当的对角矩阵. 在对矩阵进行缩放时, 需要  $\mathcal{O}(n^2)$  运算量, 通常不会产生较大的舍入误差.

另一方面, 由前面的扰动分析可知, 矩阵条件数对数值计算的误差有着很大的影响, 是衡量问题是否病态的一个重要指标. 在实际计算中, 如果遇到病态问题, 我们希望能够通过一些简单的方法降低其条件数. 其中一类简单有效的方法就是在矩阵两边同时乘以一个对角矩阵, 即寻找对角矩阵  $D_r$  和  $D_c$ , 使得  $D_r^{-1}AD_c^{-1}$  的条件数达到最小 (或者尽可能小). 显然, 寻找这样的对角矩阵是非常困难的, 目前仍然是一个开放问题 [111].

一种比较可行的实用方案是使得缩放后的矩阵的每行或每列具有相同的  $p$ -范数. 比如取  $D_r = \text{diag}(d_1, d_2, \dots, d_n)$ , 其中  $d_i = \sum_{j=1}^n |a_{ij}|$ , 这样  $D_r^{-1}A$  的每行的 1-范数都一样, 但没法使得所有列也具有相同的范数.

如果矩阵

$$D_r^{-1}AD_c^{-1}$$

的所有行和所有列都具有相同 (或近似相同) 的范数, 则称为  $A$  的**均衡化** (equilibration) [111]. 有学者提出了一种迭代方法对  $A$  进行均衡化 [79]. 更多信息可参见相关资料, 如 [37].

### 2.5.3 迭代改进法

设近似解  $\hat{x}$ , 残量  $r = b - A\hat{x}$ . 当  $\hat{x}$  没达到精度要求时, 可以考虑方程组  $Az = r$ . 设  $z$  是该方程组的精确解, 则

$$A(\hat{x} + z) = A\hat{x} + Az = (b - r) + r = b,$$

因此  $\hat{x} + z$  就是原方程组的精确解.

在实际计算中, 我们可能得到的是近似解  $\hat{z}$ , 但通常  $\|r - A\hat{z}\|$  应该比较小, 特别地, 比  $\|r\|$  更小. 因此  $\hat{x} + \hat{z}$  应该比  $\hat{x}$  更接近精确解.

如果新的近似解  $\hat{x} + \hat{z}$  仍不满足精度要求, 则可重复以上过程.



这就是通过迭代来提高解的精度.

**算法 2.15.** 通过迭代改进解的精度

- 1: 设  $PA = LU$ ,  $\hat{x}$  是  $Ax = b$  的近似解
- 2: **while** 近似解  $\hat{x}$  不满足精度要求, **do**
- 3:     计算  $r = b - A\hat{x}$
- 4:     求解  $Ly = Pr$ , 即  $y = L^{-1}Pr$
- 5:     求解  $Uz = y$ , 即  $z = U^{-1}y$
- 6:     令  $\hat{x} = \hat{x} + z$
- 7: **end while**

由于每次迭代只需计算一次残量和求解两个三角线性方程组, 因此运算量为  $O(n^2)$ . 所以相对来讲还是比较经济的.

- ✎ 为了提高计算精度, 在计算残量  $r$  时最好使用原始数据  $A$ , 而不是  $P^T LU$ , 因此对  $A$  做 PLU 分解时需要保留矩阵  $A$ , 不能被  $L$  和  $U$  覆盖.
- ✎ 实际计算经验表明, 当  $A$  病态不是很严重时, 即  $\varepsilon_u \kappa_\infty(A) < 1$ , 迭代法可以有效改进解的精度, 最后达到机器精度. 但  $\varepsilon_u \kappa_\infty(A) \geq 1$  时, 一般没什么效果. 这里  $\varepsilon_u$  表示机器精度.

## 2.6 课后习题

练习 2.1 设  $A = [a_{ij}] \in \mathbb{R}^{n \times n}$ , 且  $a_{11} \neq 0$ , 经过第一步 LU 分解后得到  $A^{(2)} = \begin{bmatrix} a_{11} & * \\ 0 & A_{22} \end{bmatrix}$ .

证明: (1) 若  $A$  对称, 则  $A_{22}$  也对称;

(2) 若  $A$  对称正定, 则  $A_{22}$  也对称正定;

(3) 若  $A$  严格行对角占优, 则  $A_{22}$  也是严格行对角占优.

练习 2.2 (定理 2.3) 设  $A \in \mathbb{R}^{n \times n}$  非奇异且列对角占优. 证明:  $A$  存在 LU 分解且  $L$  中的元素的绝对值都不超过 1.

练习 2.3 设  $A \in \mathbb{R}^{n \times n}$  非奇异. 证明: 存在置换矩阵  $P$ , 使得  $PA$  的所有顺序主子矩阵都非奇异.

练习 2.4 设矩阵  $A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 12 \end{bmatrix}$ , 计算  $A$  的 LU 分解和 PLU 分解.

练习 2.5 计算  $A = \begin{bmatrix} 4 & 2 & 4 \\ 2 & 37 & 8 \\ 4 & 8 & 14 \end{bmatrix}$  的 Cholesky 分解, 并求解  $Ax = b$ , 其中  $b = \begin{bmatrix} 6 \\ -9 \\ 7 \end{bmatrix}$ .

练习 2.6 设矩阵  $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & a \\ 0 & a & 2 \end{bmatrix}$ , 问: 当  $a$  取何值时,  $A$  存在 Cholesky 分解?

练习 2.7 证明 Cholesky 分解的唯一性.

练习 2.8 设  $A \in \mathbb{R}^{n \times n}$  对称非奇异, 且存在分解  $A = LDM^T$ , 其中  $L, M \in \mathbb{R}^{n \times n}$  是单位下三角矩阵,  $D \in \mathbb{R}^{n \times n}$  是对角矩阵. 证明:  $L = M$ .

练习 2.9 设  $A$  对称且非零, 试证明: 存在置换矩阵  $P$  使得

$$PAP^T = \begin{bmatrix} B & E^T \\ E & C \end{bmatrix},$$

其中  $B \in \mathbb{R}$  或  $B \in \mathbb{R}^{2 \times 2}$ , 且非奇异.

练习 2.10 设  $\lambda \neq 0$ , 矩阵  $A = \begin{bmatrix} \lambda & 2\lambda \\ 1 & 1 \end{bmatrix}$ , 当  $\lambda$  取何值时,  $\kappa_\infty(A)$  达到最小.

练习 2.11 设  $A \in \mathbb{R}^{m \times n}$ , 其中  $m \geq n$ , 证明:  $\|A^T A\|_2 = \|A\|_2^2$ .

当  $m = n$  时, 证明:  $\kappa_2(A^T A) = (\kappa_2(A))^2$ .

## ..... 以下为可选题 .....

练习 2.12 设  $A = [a_{ij}] \in \mathbb{R}^{n \times n}$  对称正定, 证明:  $a_{ij}^2 < a_{ii}a_{jj}$ .

练习 2.13\* (定理 2.4) 设  $A \in \mathbb{R}^{n \times n}$  严格列对角占优, 记

$$\delta \triangleq \min_{1 \leq j \leq n} \left( |a_{jj}| - \sum_{i=1}^n |a_{ij}| \right).$$





证明:  $\|A^{-1}\|_1 \leq \delta^{-1}$ .

练习 2.14 (验证等式 (2.3)) 证明:

$$L = L_1 L_2 \cdots L_{n-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{bmatrix}.$$

练习 2.15 将  $A \in \mathbb{R}^{n \times n}$  写成分块形式

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

其中  $A_{11} \in \mathbb{R}^{k \times k}$  ( $1 \leq k \leq n$ ) 非奇异. 我们称矩阵  $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$  为  $A$  中  $A_{11}$  的 Schur 补 (通常简称 **Schur 补**).

(1) 假设  $A$  存在 LU 分解, 证明: 对于不选主元的 Gauss 消去法, 第  $k$  步后,  $A_{22}$  已被  $S$  覆盖.

(2) 假设  $A_{21} = A_{12}^T$ , 且  $A_{11}$  和  $-A_{22}$  都正定, 证明  $A$  非奇异.

练习 2.16\* 假定已知  $A$  的 LU 分解:  $A = LU$ , 试设计算法计算  $A^{-1}$  的第  $(i, j)$  个元素.

练习 2.17\* 设  $A, B \in \mathbb{R}^{n \times n}$  是两个上三角矩阵,  $\alpha \in \mathbb{R}$  是给定常数, 且  $AB - \alpha I$  非奇异. 试设计求解  $(AB - \alpha I)x = f$  的算法, 使得运算量为  $\mathcal{O}(n^2)$ .

### ..... 以下为实践题 .....

练习 2.18 设  $L \in \mathbb{R}^{n \times n}$  是非奇异下三角矩阵,  $B \in \mathbb{R}^{m \times n}$ , 统计以下运算的计算量 (加减乘除, 只需给出最高次项):

(1) 计算  $L^{-1}$ ; (2) 计算  $L^2$ ; (3) 计算  $LL^T$ ; (4) 计算  $BL$ .

练习 2.19 写出按列存储方式下三角方程组  $Ly = b$  的求解算法, 并编写相应的 MATLAB 程序.

练习 2.20 利用列主元 LU 分解, 给出计算矩阵的逆的实用算法, 并编写相应的 MATLAB 程序.

练习 2.21 根据算法 2.11, 编写求解对角占优三对角线性方程组的追赶法程序.

练习 2.22 带状矩阵的 LU 分解. 设  $A$  是  $n$  阶带状矩阵, 上带宽为  $L < n$ , 下带宽为  $M < n$ , 编写一个函数, 计算  $A$  的 LU 分解 (不带选主元), 并统计运算量.

练习 2.23 设  $A = \text{tridiag}(a, b, a) \in \mathbb{R}^{n \times n}$  是对称正定三对角 Toeplitz 矩阵, 试设计  $A$  的 LDL<sup>T</sup> 分解算法.



## 第三讲 线性最小二乘问题

**最小二乘问题 (Least Squares)** 包括线性最小二乘问题, 总体最小二乘问题, 等式约束最小二乘问题, 刚性加权最小二乘问题等等. 它在统计学, 最优化问题, 材料与结构力学, 信号与图像处理等方面都有着广泛的应用, 是计算数学的一个重要研究分支, 也是一个活跃的研究领域.

本讲主要介绍求解**线性最小二乘问题**的三种常用方法: 正规方程法 (也称法方程法), QR 分解法和 SVD 分解法. 一般来说, 正规方程法是最快的, 特别是当  $A$  的条件数较小时, 正规方程法几乎与其他方法一样精确. 而 SVD 分解法是最慢的, 但结果最可靠.

为了方便起见, 我们有时将线性最小二乘问题 (3.1) 简称为**最小二乘问题**.

### 关于最小二乘问题的相关资料

- ▶ Åke Björck, *Numerical Methods for Least Squares Problems*, 1996 [16]
- ▶ 魏木生, 李莹, 赵建立, *广义最小二乘问题的理论与计算* (第二版), 2020 [142]

### 3.1 问题介绍

考虑**线性最小二乘问题**

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2, \quad (3.1)$$

其中  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$ . 问题 (3.1) 的解称为**最小二乘解**.

- 当  $m = n$  且  $A$  非奇异时, 这就是一个线性方程组, 解为  $x_* = A^{-1}b$ ;
- 当  $m > n$  时, 约束个数大于未知量个数, 此时我们称问题 (3.1) 为**超定**的;
- 当  $m < n$  时, 未知量个数大于约束个数, 此时我们称问题 (3.1) 为**欠定** (或亚定) 的.

 为了讨论方便, 本讲假定  $A$  是满秩的.

#### 3.1.1 超定方程组

当  $m > n$  时, 线性方程组  $Ax = b$  的解可能不存在. 此时一般考虑求解最小二乘问题 (3.1). 记

$$J(x) \triangleq \|Ax - b\|_2^2.$$

易知  $J(x)$  是关于  $x$  的二次函数, 而且是凸函数 (当  $A$  满秩时,  $J(x)$  的 Hesse 阵是正定的). 因此, 由凸函数的性质可知,  $x_*$  是问题 (3.1) 的解当且仅当  $x_*$  是  $J(x)$  的稳定点. 令其一阶导数为零, 可得

$$A^T Ax - A^T b = 0.$$

于是将最小二乘问题转化为一个线性方程组, 这就是后面的正规方程.

 如果  $A$  不是满秩, 则  $A^T A$  半正定, 此时解不唯一.

#### 3.1.2 欠定方程组

若  $m < n$ , 则线性方程组  $Ax = b$  存在无穷多个解 (假定  $A$  满秩). 这时我们通常寻求最小范数解, 即所有解中范数最小的解. 于是原问题就转化为下面的**约束优化问题**

$$\min_{Ax=b} \frac{1}{2} \|x\|_2^2 \quad (3.2)$$

对应的 **Lagrange 函数**为

$$\mathcal{L}(x, \lambda) = \frac{1}{2} \|x\|_2^2 + \lambda^T (Ax - b),$$

其中  $\lambda = [\lambda_1, \lambda_2, \dots, \lambda_m]^T$  是 **Lagrange 乘子**. 此时优化问题 (3.2) 的解就是  $\mathcal{L}(x, \lambda)$  的**鞍点**, 即下面方程组的解:

$$\frac{\partial \mathcal{L}}{\partial x} = x + A^T \lambda = 0, \quad \frac{\partial \mathcal{L}}{\partial \lambda} = Ax - b = 0.$$

写成矩阵形式为

$$\begin{bmatrix} I & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}.$$

如果  $A$  满秩, 即  $\text{rank}(A) = m$ , 则系数矩阵非奇异 (见练习 3.1), 上述方程组存在唯一解.

本讲主要讨论超定线性最小二乘问题的求解.



### 3.2 几类重要的矩阵变换


矩阵计算的一个基本思想就是把复杂的问题转化为等价的且易于求解的问题. 完成这个转化的基本工具就是**矩阵变换**, 除了三类初等变换以外, 在矩阵计算中常用的矩阵变换还有: Gauss 变换, Householder 变换和 Givens 变换, 其中 Gauss 变换主要用于矩阵的 LU 分解, Householder 变换和 Givens 变换是正交变换, 主要用于计算线性最小二乘问题、矩阵特征值和奇异值问题等.

#### 3.2.1 初等矩阵变换

本科高等代数教材中介绍了三类初等矩阵变换, 这里将其推广到更一般的情形. 我们称

$$E(u, v, \tau) = I - \tau uv^* \quad (3.3)$$

为**初等矩阵变换**或**初等矩阵**, 其中  $u, v \in \mathbb{C}^n$  是非零向量,  $\tau$  是一个非零复数. 因此, 初等矩阵是单位矩阵的一个秩 1 修正.

 高等代数中介绍的三类初等矩阵变换分别是 (以行变换为例): (1) 交换两行; (2) 某行乘以一个非零常数; (3) 某行乘以一个常数后加到另外一行. 这三类矩阵变换所对应的初等矩阵可表示为 (简单示例)

$$E_1 = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 0 & & 1 \\ & & & \ddots & \\ & 1 & & & 0 \\ & & & & & \ddots \\ & & & & & & 1 \end{bmatrix}, E_2 = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & c & \\ & & & & 1 \\ & & & & & \ddots \\ & & & & & & 1 \end{bmatrix}, E_3 = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & c & \\ & & & & 1 \\ & & & & & \ddots \\ & & & & & & 1 \end{bmatrix}.$$

可以验证, 它们都可以表示为 (3.3) 形式, 留作练习.

下面的定理给出了初等矩阵的基本性质.

**定理 3.1** 设  $E(u, v, \tau)$  是一个初等矩阵, 我们有

(1)  $\det(E(u, v, \tau)) = 1 - \tau v^* u$ ;

(2) 若  $1 - \tau v^* u \neq 0$ , 则  $E(u, v, \tau)$  非奇异, 且

$$(E(u, v, \tau))^{-1} = E(u, v, \gamma), \quad \text{其中 } \gamma = \frac{\tau}{\tau v^* u - 1}.$$

(3) 对任意非零向量  $x, y \in \mathbb{C}^n$ , 存在  $u, v \in \mathbb{C}^n$  和  $\tau \in \mathbb{C}$ , 使得

$$E(u, v, \tau)x = y.$$

(板书)

**证明.** (1) 易知

$$\begin{bmatrix} I & 0 \\ v^* & 1 \end{bmatrix} \begin{bmatrix} I - \tau uv^* & -\tau u \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I & 0 \\ -v^* & 1 \end{bmatrix} = \begin{bmatrix} I & -\tau u \\ 0 & 1 - \tau v^* u \end{bmatrix}.$$

由行列式的乘法可知

$$\det \left( \begin{bmatrix} I & 0 \\ v^* & 1 \end{bmatrix} \right) \cdot \det \left( \begin{bmatrix} I - \tau uv^* & -\tau u \\ 0 & 1 \end{bmatrix} \right) \cdot \det \left( \begin{bmatrix} I & 0 \\ -v^* & 1 \end{bmatrix} \right) = \det \left( \begin{bmatrix} I & -\tau u \\ 0 & 1 - \tau v^* u \end{bmatrix} \right).$$



所以

$$\det(E(u, v, \tau)) = \det(I - \tau uv^*) = 1 - \tau v^* u.$$

(2) 若  $1 - \tau v^* u \neq 0$ , 则  $\det(E(u, v, \tau)) \neq 0$ , 所以  $E(u, v, \tau)$  非奇异. 通过直接计算可知

$$\begin{aligned} E(u, v, \tau)E(u, v, \gamma) &= I - \tau uv^* - \gamma(1 - \tau v^* u)uv^* \\ &= I - \tau uv^* - \frac{\tau}{\tau v^* u - 1}(1 - \tau v^* u)uv^* \\ &= I. \end{aligned}$$

(3) 留作练习. □

更一般地, 我们有下面的结论.

### Sylvester 降幂公式

设  $A \in \mathbb{C}^{m \times n}$ ,  $B \in \mathbb{C}^{n \times m}$ , 其中  $m \geq n$ , 则有

$$\det(\lambda I - AB) = \lambda^{m-n} \det(\lambda I - BA).$$

因此  $AB$  和  $BA$  具有相同的非零特征值.

**定理 3.2** 设  $A \in \mathbb{C}^{n \times n}$ , 则  $A$  非奇异当且仅当  $A$  可以分解成若干个初等矩阵的乘积.

### 3.2.2 Gauss 变换

设  $l_j = [0, \dots, 0, l_{j+1,j}, \dots, l_{n,j}]^T, j = 1, 2, \dots, n$ , 则 **Gauss 变换** 定义为

$$L(l_j) \triangleq E(l_j, e_j, -1) = I + l_j e_j^T = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{j+1,j} & 1 & \\ & & \vdots & & \ddots \\ & & l_{n,j} & & & 1 \end{bmatrix}.$$

显然, Gauss 变换也属于初等矩阵变换. 向量  $l_j$  称为 **Gauss 向量** [56]. 由定理 3.1 可知

$$\det(L(l_j)) = 1, \quad (L(l_j))^{-1} = E(l_j, e_j, 1) = E(-l_j, e_j, -1) = L(-l_j).$$

 Gauss 变换有时也成为**初等下三角矩阵**, 主要用于矩阵的 LU 分解.

### 3.2.3 Householder 变换

**定义 3.1** 我们称矩阵

$$H = I - \frac{2}{v^* v} v v^* = I - \frac{2}{\|v\|_2^2} v v^*, \quad 0 \neq v \in \mathbb{C}^n, \quad (3.4)$$




为 **Householder 矩阵** (或 **Householder 变换**, 或 **Householder 反射**), 向量  $v$  称为 **Householder 向量**. 我们通常将矩阵 (3.4) 记为  $H(v)$ .

 Householder 矩阵也是初等矩阵.

 有时也将 Householder 变换定义为

$$H = I - 2vv^*, \quad v \in \mathbb{C}^n \text{ 且 } \|v\|_2 = 1.$$

 Householder 变换有时也称为**初等 Hermite 矩阵**.

从几何上看, 一个 Householder 变换是一个关于超平面  $\text{span}\{v\}^\perp$  的反射. 由于  $\mathbb{C}^2 = \text{span}\{v\} \oplus \text{span}\{v\}^\perp$ , 因此对任意一个向量  $x \in \mathbb{C}^n$ , 都可写为

$$x = \frac{v^*x}{v^*v}v + y \triangleq \alpha v + y,$$

其中  $\alpha v \in \text{span}\{v\}$ ,  $y \in \text{span}\{v\}^\perp$ . 于是

$$Hx = x - \frac{2}{v^*v}vv^*x = x - 2\alpha v = -\alpha v + y,$$

即  $Hx$  与  $x$  在  $\text{span}\{v\}^\perp$  方向有着相同的分量, 而在  $v$  方向的分量正好相差一个符号. 也就是说,  $Hx$  是  $x$  关于超平面  $\text{span}\{v\}^\perp$  的镜面反射, 见图 3.1. 因此, Householder 变换也称为 Householder 反射.

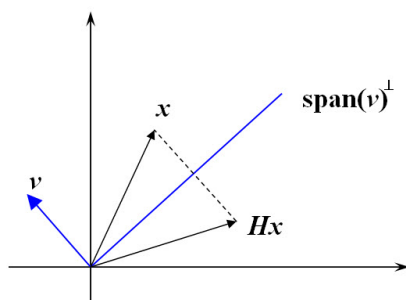


图 3.1. Householder 变换的几何意义

下面是关于 Householder 矩阵的几个基本性质.

**定理 3.3** 设  $H \in \mathbb{C}^{n \times n}$  是一个 Householder 矩阵, 则

- (1)  $H^* = H$ , 即  $H$  是 Hermite 的;
- (2)  $H^*H = I$ , 即  $H$  是酉矩阵;
- (3)  $H^2 = I$ , 所以  $H^{-1} = H$ ;
- (4)  $\det(H) = -1$ ;
- (5)  $H$  有两个互异的特征值:  $\lambda = 1$  和  $\lambda = -1$ , 其中  $\lambda = 1$  的代数重数为  $n - 1$ .

Householder 矩阵的一个非常重要的应用就是可以将一个向量除第一个元素以外的所有元素都化为零. 我们首先给出一个引理.

**引理 3.4** 设  $x, y \in \mathbb{C}^n$  为任意两个互异的向量, 则存在一个 Householder 矩阵  $H(v)$  使得  $y = H(v)x$  的充要条件是  $\|x\|_2 = \|y\|_2$  且  $x^*y \in \mathbb{R}$ .

(板书)

**证明.** 若  $\|x\|_2 = \|y\|_2$  且  $x^*y \in \mathbb{R}$ , 则  $y^*y = x^*x$  且  $x^*y = y^*x$ . 于是

$$\|x - y\|_2^2 = (x - y)^*(x - y) = x^*x - y^*x - x^*y + y^*y = 2(x^*x - y^*x).$$

令  $v = x - y$ , 则有

$$H(v)x = x - \frac{2(x - y)(x - y)^*x}{\|x - y\|_2^2} = x - \frac{2(x - y)(x^*x - y^*x)}{2(x^*x - y^*x)} = y,$$


即存在 Householder 矩阵  $H(v)$  使得  $y = H(v)x$ .

反之, 如果存在 Householder 矩阵  $H$  使得  $y = Hx$ , 由于  $H$  是 Hermite 的, 所以  $x^*y = x^*Hx \in \mathbb{R}$ . 又因为  $H$  是酉矩阵, 所以  $\|y\|_2 = \|Hx\|_2 = \|x\|_2$ .  $\square$

 如果  $x, y$  都是实向量, 则条件  $x^*y \in \mathbb{R}$  自然成立, 此时充要条件就是  $\|x\|_2 = \|y\|_2$ .

由引理 3.4, 我们可以立即得到下面的结论.

**定理 3.5** 设  $x = [x_1, x_2, \dots, x_n]^T \in \mathbb{R}^n$  是一个非零向量, 则存在 Householder 矩阵  $H(v)$  使得  $H(v)x = \alpha e_1$ , 其中  $\alpha = \|x\|_2$  (或  $\alpha = -\|x\|_2$ ),  $e_1 = [1, 0, \dots, 0]^T \in \mathbb{R}^n$ .

 在后面的讨论中, 我们将定理的向量  $v$  称为  $x$  对应的 Householder 向量.

设  $x = [x_1, x_2, \dots, x_n]^T \in \mathbb{R}^n$  是一个实的非零向量, 下面讨论如何计算定理 3.5 中 Householder 矩阵  $H(v)$  所对应的 Householder 向量  $v$ . 由引理 3.4 的证明过程可知

$$v = x - \alpha e_1 = [x_1 - \alpha, x_2, \dots, x_n]^T.$$

在实际计算中, 为了尽可能地减少舍入误差, 我们通常避免两个相近的数做减运算, 否则就会损失有效数字. 因此, 我通常取

$$\alpha = -\text{sign}(x_1) \cdot \|x\|_2. \quad (3.5)$$

事实上, 我们也可以取  $\alpha = \text{sign}(x_1)\|x\|_2$ , 但此时为了减少舍入误差, 我们需要通过下面的公式来计算  $v$  的第一个分量  $v_1$

$$\alpha = \text{sign}(x_1)\|x\|_2, \quad v_1 = x_1 - \alpha = \frac{x_1^2 - \|x\|_2^2}{x_1 + \alpha} = \frac{-(x_2^2 + x_3^2 + \dots + x_n^2)}{x_1 + \alpha}. \quad (3.6)$$

在  $v_1$  的两种计算方法 (3.5) 和 (3.6) 中,  $\alpha$  的取值都与  $x_1$  的符号有关. 但在某些应用中, 我们需要确保  $\alpha$  非负, 此时我们可以将这两种方法结合起来使用, 即:

$$v_1 = \begin{cases} x_1 - \alpha, & \text{if } \text{sign}(x_1) < 0 \\ \frac{-(x_2^2 + x_3^2 + \dots + x_n^2)}{x_1 + \alpha}, & \text{otherwise} \end{cases}$$





无论怎样选取  $\alpha$ , 我们都有  $H = I - \beta vv^*$ , 其中

$$\beta = \frac{2}{v^*v} = \frac{2}{(x_1 - \alpha)^2 + x_2^2 + \cdots + x_n^2} = \frac{2}{2\alpha^2 - 2\alpha x_1} = -\frac{1}{\alpha v_1}.$$

 **思考:** 如果  $x$  是复向量, 有没有相应的结论?

在实数域中计算 Householder 向量  $v$  的算法如下, 总运算量大约为  $2n$ , 如果加上赋值运算的话则为  $3n$ .


### 算法 3.1. 计算 Householder 向量

```
% Given $x \in \mathbb{R}^n$, compute $v \in \mathbb{R}^n$ such that $Hx = \|x\|_2 e_1$, where $H = I - \beta vv^*$ (House.m)
1: function [β, v] = House(x)
2: $n = \text{length}(x)$ (here $\text{length}(x)$ denotes the dimension of x)
3: $\sigma = x_2^2 + x_3^2 + \cdots + x_n^2$
4: $v = x$
5: if $\sigma = 0$ then
6: if $x_1 < 0$ then
7: $v_1 = 2x_1, \beta = 2/v_1^2$
8: else
9: $v_1 = 0, \beta = 0$
10: end if
11: else
12: $\alpha = \sqrt{x_1^2 + \sigma}$ % $\alpha = \|x\|_2$
13: if $x_1 < 0$ then
14: $v_1 = x_1 - \alpha$
15: else
16: $v_1 = -\sigma/(x_1 + \alpha)$
17: end if
18: $\beta = 2/(v_1^2 + \sigma)$
19: end if
```

可以证明, 上述算法具有很好的数值稳定性 [128], 即

$$\|\tilde{H} - H\|_2 = \mathcal{O}(\varepsilon_u),$$

其中  $\tilde{H}$  是由上述算法计算得到的近似矩阵,  $\varepsilon_u$  是机器精度.

 在实际计算时, 我们可以将向量  $v$  单位化, 使得  $v_1 = 1$ . 这样, 我们就无需为  $v$  另外分配空间, 而是将  $v(2:n)$  存放在  $x(2:n)$  中, 因为经过 Householder 变换后, 向量  $x$  除第一个分量外, 其它都为零.



**思考：**这里要求  $v_1 \neq 0$ , 那么什么情况下  $v_1 = 0$ ?

 为了避免可能产生的溢出, 我们也可以事先将  $x$  单位化, 即令  $x = x/\|x\|_2$

### Householder 变换与矩阵的乘积

设  $A \in \mathbb{R}^{m \times n}$ ,  $H = I - \beta vv^* \in \mathbb{R}^m$ , 则

$$HA = (I - \beta vv^*)A = A - \beta v(v^*A).$$

因此, 在做 Householder 变换时, 并不需要生成 Householder 矩阵, 只需要 Householder 向量即可. 上面矩阵相乘的总运算量大约为  $4mn$ .

关于其他类型的 Householder 变换, 可以参见 [35].


### 3.2.4 Givens 变换

为简单起见, 我们这里讨论实数域中的 Givens 变换. 设  $\theta \in [0, 2\pi]$ , 我们称矩阵

$$G(i, j, \theta) = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & c & \cdots & s \\ & & \vdots & \ddots & \vdots \\ & & -s & \cdots & c \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad (i \leq j)$$

为 **Givens 变换** (或 **Givens 旋转**, 或 **Givens 矩阵**), 其中  $c = \cos(\theta)$ ,  $s = \sin(\theta)$ . 即将单位矩阵的  $(i, i)$  和  $(j, j)$  位置上的元素用  $c$  代替, 而  $(i, j)$  和  $(j, i)$  位置上的元素分别用  $s$  和  $-s$  代替, 所得到的矩阵就是  $G(i, j, \theta)$ .

**定理 3.6**  $G(i, j, \theta)$  是正交矩阵, 且  $\det(G(i, j, \theta)) = 1$ .

 Givens 变换不是一个初等矩阵变换, 事实上, 它是单位矩阵的一个秩 2 修正, 即

$$G(i, j, \theta) = I + [e_i, e_j] \begin{bmatrix} c-1 & s \\ -s & c-1 \end{bmatrix} [e_i, e_j]^T.$$

 如果是定义在复数域上的 Givens 变换, 则  $c = e^{i\alpha} \cos \theta$ ,  $s = e^{i\beta} \sin \theta$ ,  $0 \leq \alpha, \beta, \theta < 2\pi$ .

**例 3.1** 设  $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$ , 则存在一个 Givens 变换  $G = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \in \mathbb{R}^{2 \times 2}$  使得  $G^T x = \begin{bmatrix} r \\ 0 \end{bmatrix}$ , 其中  $c, s$  和  $r$  的值如下:

$$c = \frac{x_1}{r}, \quad s = \frac{x_2}{r}, \quad r = \sqrt{x_1^2 + x_2^2},$$

也就是说, 通过 Givens 变换, 我们可以将向量  $x \in \mathbb{R}^2$  的第二个分量化为 0.

事实上, 对任意一个向量  $x \in \mathbb{R}^n$ , 都可以通过 Givens 变换将其任意一个位置上的分量化为 0. 更进一步, 我们也可以通过若干个 Givens 变换, 将  $x$  中除第一个分量外的所有元素都化为 0.



**算法 3.2.** Givens 变换

```

% Given $x = [x_1, x_2]^T \in \mathbb{R}^2$, compute c and s such that $Gx = [r, 0]^T$ where $r = \|x\|_2$ (Givens.m)
1: function $[c, s] = \text{givens}(x_1, x_2)$
2: if $x_2 = 0$ then
3: $c = \text{sign}(x_1), \quad s = 0$
4: else
5: if $|x_2| > |x_1|$ then
6: $\tau = \frac{x_1}{x_2}, \quad s = \frac{\text{sign}(x_2)}{\sqrt{1 + \tau^2}}, \quad c = s\tau$
7: else
8: $\tau = \frac{x_2}{x_1}, \quad c = \frac{\text{sign}(x_1)}{\sqrt{1 + \tau^2}}, \quad s = c\tau$
9: end if
10: end if

```

**例 3.2** 设  $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} r \cos \alpha \\ r \sin \alpha \end{bmatrix} \in \mathbb{R}^2$ , 其中  $r = \sqrt{x_1^2 + x_2^2}$ ,  $\alpha = \arctan \frac{x_2}{x_1}$ , 则

$$G^T x = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}^T \begin{bmatrix} r \cos \alpha \\ r \sin \alpha \end{bmatrix} = \begin{bmatrix} r(\cos \theta \cos \alpha - \sin \theta \sin \alpha) \\ r(\sin \theta \cos \alpha + \cos \theta \sin \alpha) \end{bmatrix} = \begin{bmatrix} r \cos(\alpha + \theta) \\ r \sin(\alpha + \theta) \end{bmatrix},$$

也就是说,  $G^T x$  相当于将向量  $x$  按逆时针旋转  $\theta$  角度. 因此当  $\theta = -\alpha$  时,  $Gx = \begin{bmatrix} r \\ 0 \end{bmatrix}$ .

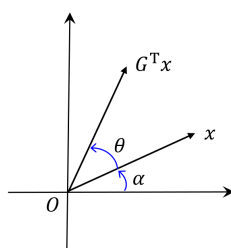


图 3.2. Givens 变换的几何意义

**Givens 变换与矩阵的乘积**

设  $A \in \mathbb{R}^{m \times n}$ ,  $G = G(i, j, \theta) \in \mathbb{R}^m$ , 则  $G^T A$  只会影响其第  $i$  行和第  $j$  行的元素, 也就是说, 只需对  $A$  的第  $i$  行和第  $j$  行做 Givens 变换即可, 运算量大约为  $6n$ .

同样地, 如果是右乘 Givens 变换, 即  $AG$ , 则只会影响其第  $i$  列和第  $j$  列的元素, 运算量大约为  $6m$ .



任何一个正交矩阵都可以写成若干个 Householder 矩阵或 Givens 矩阵的乘积 (见习题 3.5 和 3.20), 所以正交矩阵所对应的线性变换可以看作是反射变换和旋转变换的推广, 因此它不会改变向量的长度与 (不同向量之间的) 角度.

### 3.2.5 正交变换的舍入误差分析

**引理 3.7** 设  $P \in \mathbb{R}^{n \times n}$  是一个精确的 Householder 或 Givens 变换,  $\tilde{P}$  是其浮点运算近似, 则

$$\mathfrak{fl}(\tilde{P}A) = P(A + E), \quad \mathfrak{fl}(A\tilde{P}) = (A + F)P,$$

其中  $\|E\|_2 = \mathcal{O}(\varepsilon_u) \cdot \|A\|_2$ ,  $\|F\|_2 = \mathcal{O}(\varepsilon_u) \cdot \|A\|_2$ .

这说明对一个矩阵做 Householder 变换或 Givens 变换是向后稳定的.

**定理 3.8** 考虑对矩阵  $A$  做一系列的正交变换, 则有

$$\mathfrak{fl}(\tilde{P}_k \cdots \tilde{P}_1 A \tilde{Q}_1 \cdots \tilde{Q}_k) = P_k \cdots P_1 (A + E) Q_1 \cdots Q_k,$$

其中  $\|E\|_2 = \mathcal{O}(\varepsilon_u) \cdot (k\|A\|_2)$ . 这说明整个计算过程是向后稳定的.

一般地, 假设  $X$  是一个非奇异的线性变换,  $\tilde{X}$  是其浮点运算近似. 当  $X$  作用到  $A$  上时, 我们有

$$\mathfrak{fl}(\tilde{X}A) = XA + E = X(A + X^{-1}E) \triangleq X(A + F),$$

其中  $\|E\|_2 = \mathcal{O}(\varepsilon_u) \cdot \|XA\|_2 \leq \mathcal{O}(\varepsilon_u) \cdot \|X\|_2 \cdot \|A\|_2$ , 故

$$\|F\|_2 = \|X^{-1}E\|_2 \leq \mathcal{O}(\varepsilon_u) \cdot \|X^{-1}\|_2 \cdot \|X\|_2 \cdot \|A\|_2 = \mathcal{O}(\varepsilon_u) \cdot \kappa_2(X) \cdot \|A\|_2,$$

因此, 舍入误差将被放大  $\kappa_2(X)$  倍. 当  $X$  是正交变换时,  $\kappa_2(X)$  达到最小值 1, 这就是为什么在浮点运算中尽量使用正交变换的原因.



### 3.3 QR 分解

QR 分解是将一个矩阵分解一个正交矩阵 (酉矩阵) 和一个三角矩阵的乘积. QR 分解被广泛应用于线性最小二乘问题的求解和矩阵特征值的计算.

#### 3.3.1 QR 分解的存在性与唯一性

**定理 3.9 (QR 分解)** [71] 设  $A \in \mathbb{C}^{m \times n}$  ( $m \geq n$ ). 则存在一个单位列正交矩阵  $Q \in \mathbb{C}^{m \times n}$  (即  $Q^*Q = I_{n \times n}$ ) 和一个上三角矩阵  $R \in \mathbb{C}^{n \times n}$ , 使得

$$A = QR. \quad (3.7)$$

若  $A$  列满秩, 则存在一个具有正对角线元素的上三角矩阵  $R$  使得 (3.7) 成立, 且此时 QR 分解唯一, 即  $Q$  和  $R$  都唯一.

**证明.** 设  $A = [a_1, a_2, \dots, a_n] \in \mathbb{C}^{m \times n}$ . 若  $A$  列满秩, 即  $\text{rank}(A) = n$ . 则 QR 分解 (3.7) 就是对  $A$  的列向量组进行 Gram-Schmidt 正交化过程的矩阵描述 (见算法 3.3).

#### 算法 3.3. Gram-Schmidt 正交化过程

```

1: $r_{11} = \|a_1\|_2$
2: $q_1 = a_1/r_{11}$
3: for $j = 2$ to n do
4: $q_j = a_j$
5: for $i = 1$ to $j - 1$ do
6: $r_{ij} = (a_j, q_i)$ % 计算内积, 用于正交化
7: $q_j = q_j - r_{ij}q_i$
8: end for
9: $r_{jj} = \|q_j\|_2$
10: $q_j = q_j/r_{jj}$
11: end for

```

由算法 3.3 可知

$$a_1 = r_{11}q_1, \quad a_j = r_{1j}q_1 + r_{2j}q_2 + \cdots + r_{jj}q_j = [q_1, q_2, \dots, q_j] \begin{bmatrix} r_{1j} \\ r_{2j} \\ \vdots \\ r_{jj} \end{bmatrix}, \quad j = 2, 3, \dots, n.$$

记  $Q = [q_1, q_2, \dots, q_n]$ ,  $R = [r_{ij}]_{n \times n}$ , 其中

$$r_{ij} = \begin{cases} q_i^* a_j, & \text{for } i \leq j \\ 0, & \text{for } i > j \end{cases} \quad (3.8)$$



于是 Gram-Schmidt 正交化过程可表示为

$$[a_1, a_2, \dots, a_n] = [q_1, q_2, \dots, q_n] \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & r_{n-1,n} \\ & & & r_{nn} \end{bmatrix}, \quad \text{即} \quad A = QR.$$

如果  $A$  不是列满秩, 我们可以通过下面的方式做类似的正交化过程:

- 如果  $a_1 = 0$ , 则令  $q_1 = 0$ ; 否则令  $q_1 = a_1 / \|a_1\|_2$ ;
- 对于  $j = 2, 3, \dots, n$ , 计算  $\tilde{q}_j = a_j - \sum_{i=1}^{j-1} (q_i^* a_j) q_i$ . 如果  $\tilde{q}_j = 0$ , 则表明  $a_j$  可以由  $a_1, a_2, \dots, a_{j-1}$  线性表出, 令  $q_j = 0$ . 否则令  $q_j = \tilde{q}_j / \|\tilde{q}_j\|_2$ .

于是我们有

$$A = QR,$$

其中  $Q = [q_1, q_2, \dots, q_n]$  列正交 (但不是单位列正交), 其列向量要么是单位向量, 要么就是零向量.  $R = [r_{ij}]_{n \times n}$  的定义同 (3.8). 需要注意的是, 如果  $Q$  的某一列  $q_k = 0$ , 那么  $R$  中对应的第  $k$  行就全部为 0.

设  $\text{rank}(A) = l < n$ , 则  $Q$  有  $l$  个非零列, 设为  $q_{i_1}, q_{i_2}, \dots, q_{i_l}$ . 它们形成  $\mathbb{C}^m$  中的一个单位正交向量组, 所以我们可以将其扩展成  $\mathbb{C}^m$  中的一组标准正交基, 即

$$q_{i_1}, q_{i_2}, \dots, q_{i_l}, \tilde{q}_1, \dots, \tilde{q}_{m-l}.$$

然后我们用  $\tilde{q}_1$  替换  $Q$  中的第一个零列, 用  $\tilde{q}_2$  替换  $Q$  中的第二个零列, 依此类推, 将  $Q$  中的所有零列都替换掉. 将最后得到的矩阵记为  $\tilde{Q}$ , 则  $\tilde{Q} \in \mathbb{C}^{m \times n}$  单位列正交, 且

$$\tilde{Q}R = QR.$$

这是由于  $\tilde{Q}$  中的新添加的列向量正好与  $R$  中的零行相对应. 所以我们有 QR 分解

$$A = \tilde{Q}R.$$

下面证明 **满秩矩阵 QR 分解的存在唯一性**.

存在性: 由于  $A$  列满秩, 由 Gram-Schmidt 正交化过程 (算法 3.3) 可知, 存在上三角矩阵  $R = [r_{ij}]_{n \times n}$  满足  $r_{jj} > 0$ , 使得  $A = QR$ , 其中  $Q$  单位列正交.

唯一性: 假设  $A$  存在 QR 分解

$$A = Q_1 R_1 = Q_2 R_2,$$

其中  $Q_1, Q_2 \in \mathbb{C}^{m \times n}$  单位列正交,  $R_1, R_2 \in \mathbb{C}^{n \times n}$  为具有正对角元素的上三角矩阵. 则有

$$Q_1 = Q_2 R_2 R_1^{-1}. \quad (3.9)$$

于是

$$1 = \|Q_1\|_2 = \|Q_2 R_2 R_1^{-1}\|_2 = \|R_2 R_1^{-1}\|_2.$$

又  $R_1, R_2$  均为上三角矩阵, 所以  $R_2 R_1^{-1}$  也是上三角矩阵, 且其对角线元素为  $R_2(i, i) / R_1(i, i)$ ,



$i = 1, 2, \dots, n$ . 因此

$$\frac{R_2(i, i)}{R_1(i, i)} \leq \rho(R_2 R_1^{-1}) \leq \|R_2 R_1^{-1}\|_2 \leq 1, \quad i = 1, 2, \dots, n.$$


同理可证  $R_1(i, i)/R_2(i, i) \leq 1$ . 所以

$$R_1(i, i) = R_2(i, i), \quad i = 1, 2, \dots, n.$$

又  $\|Q_1\|_F^2 = \text{tr}(Q_1^* Q_1) = n$ , 所以由 (3.9) 可知

$$\|R_2 R_1^{-1}\|_F^2 = \|Q_2 R_2 R_1^{-1}\|_F^2 = \|Q_1\|_F^2 = n.$$

由于  $R_2 R_1^{-1}$  的对角线元素都是 1, 所以  $R_2 R_1^{-1}$  只能是单位矩阵, 即  $R_2 = R_1$ . 因此  $Q_2 = A R_2^{-1} = A R_1^{-1} = Q_1$ , 即  $A$  的 QR 分解是唯一的.  $\square$


 有时也将 QR 分解定义为: 存在酉矩阵  $Q \in \mathbb{C}^{m \times m}$  使得

$$A = QR,$$

其中  $R = \begin{bmatrix} R_{11} \\ 0 \end{bmatrix} \in \mathbb{C}^{m \times n}$  是上三角矩阵.

 如果  $A$  是实矩阵, 则上面证明中的运算都可以在实数下进行, 因此  $Q$  和  $R$  都可以是实矩阵.

 如果  $A$  是非奇异的方阵, 则 QR 分解也可以用来求解线性方程组  $Ax = b$ .

 基于 GS 正交化的 QR 分解算法 3.3 的运算量大约为  $2mn^2$ .

下面是 QR 分解的一个应用.

**推论 3.10 (满秩分解)** 设  $A \in \mathbb{C}^{m \times n}$  且  $\text{rank}(A) = r \leq \min\{m, n\}$ , 则存在满秩矩阵  $F \in \mathbb{C}^{m \times r}$  和  $G \in \mathbb{C}^{r \times n}$ , 使得

$$A = FG.$$

下面给出 QR 分解的具体实现方法, 分别基于 MGS 正交化过程, Householder 变换和 Givens 变换.

### 3.3.2 基于 MGS 的 QR 分解

在证明 QR 分解的存在性时, 我们利用了 Gram-Schmidt 正交化过程. 但由于数值稳定性方面的原因, 在实际计算中, 我们一般不采用 Gram-Schmidt 正交化过程, 取而代之的是 **修正的 Gram-Schmidt 正交化过程** (modified Gram-Schmidt process, **MGS**), 即对正交化过程做如下修改:

- Gram-Schmidt 正交化过程的第  $j$  步:
  - (1) 计算  $r_{ij} = (a_j, q_i)$ ,  $i = 1, 2, \dots, j-1$ ;
  - (2) 计算  $\tilde{q}_j = a_j - r_{1j}q_1 - r_{2j}q_2 - \dots - r_{j-1,j}q_{j-1}$ ;
  - (3) 计算  $r_{jj} = \|\tilde{q}_j\|$ ,  $q_j = \tilde{q}_j/r_{jj}$ ;
- MGS 正交化过程的第  $j$  步:



- (1) 令  $\tilde{q}_j = a_j$ ;
- (2) 计算  $r_{ij} = (\tilde{q}_j, q_i)$ ,  $\tilde{q}_j = \tilde{q}_j - r_{ij}q_i$ ,  $i = 1, 2, \dots, j-1$ ;
- (3) 计算  $r_{jj} = \|\tilde{q}_j\|$ ,  $q_j = \tilde{q}_j/r_{jj}$ ;

可以证明, 数学上这两个算法完全等价, 即  $r_{ij}$  和  $q_j$  都一样. 但在数值上, Gram-Schmidt 正交化过程是不稳定的, 而 MGS 是向后稳定的 [94].

#### 算法 3.4. 基于 MGS 的 QR 分解

```
% Given $A \in \mathbb{C}^{m \times n}$, compute $Q = [q_1, \dots, q_n] \in \mathbb{R}^{m \times n}$ and $R \in \mathbb{R}^{n \times n}$ such that $A = QR$
1: Set $R = [r_{ij}] = 0_{n \times n}$ (the $n \times n$ zero matrix)
2: if $a_1 = 0$ then
3: $q_1 = 0$
4: else
5: $r_{11} = \|a_1\|_2$
6: $q_1 = a_1/\|a_1\|_2$
7: end if
8: for $j = 2$ to n do
9: $q_j = a_j$
10: for $i = 1$ to $j-1$ do % MGS, 注意与 GS 的区别
11: $r_{ij} = (q_j, q_i)$
12: $q_j = q_j - r_{ij}q_i$
13: end for
14: if $q_j \neq 0$ then
15: $r_{jj} = \|q_j\|_2$
16: $q_j = q_j/r_{jj}$
17: end if
18: end for
```

🔪 本算法的运算量大约为  $2mn^2$ .

🔪 由 MGS 得到的 QR 分解中,  $Q \in \mathbb{R}^{m \times n}$ ,  $R \in \mathbb{R}^{n \times n}$ .

### 3.3.3 基于 Householder 变换的 QR 分解

由定理 3.5 可知, 通过 Householder 变换, 我们可以将任何一个非零变量  $x \in \mathbb{R}^n$  转化成  $\|x\|_2 e_1$ , 即除第一个元素外, 其它都为零. 下面我们就考虑通过 Householder 变换来实现矩阵的 QR 分解.





我们以  $m = n$  为例. 设矩阵  $A \in \mathbb{R}^{n \times n}$ , 构造 Householder 变换  $H_1 \in \mathbb{R}^{n \times n}$ , 使得

$$H_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} = \begin{bmatrix} r_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

于是

$$H_1 A = \left[ \begin{array}{c|ccc} r_1 & \tilde{a}_{12} & \cdots & \tilde{a}_{1n} \\ \hline 0 & & & \\ \vdots & & \tilde{A}_2 & \\ 0 & & & \end{array} \right],$$

其中  $\tilde{A}_2 \in \mathbb{R}^{(n-1) \times (n-1)}$ . 同样地, 我们可以构造一个 Householder 变换  $\tilde{H}_2 \in \mathbb{R}^{(n-1) \times (n-1)}$ , 将  $\tilde{A}_2$  的第一列中除第一个元素外的所有元素都化为 0, 即

$$\tilde{H}_2 \tilde{A}_2 = \left[ \begin{array}{c|ccc} r_2 & \tilde{a}_{23} & \cdots & \tilde{a}_{2n} \\ \hline 0 & & & \\ \vdots & & \tilde{A}_3 & \\ 0 & & & \end{array} \right].$$

令

$$H_2 = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{H}_2 \end{bmatrix}.$$

则  $H_2 \in \mathbb{R}^{n \times n}$ , 且

$$H_2 H_1 A = \left[ \begin{array}{cc|ccc} r_1 & \tilde{a}_{12} & \tilde{a}_{13} & \cdots & \tilde{a}_{1n} \\ 0 & r_2 & \tilde{a}_{23} & \cdots & \tilde{a}_{2n} \\ \hline 0 & 0 & & & \\ \vdots & \vdots & & \tilde{A}_3 & \\ 0 & 0 & & & \end{array} \right].$$

不断重复上述过程. 这样, 我们就得到一系列的矩阵

$$H_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & \tilde{H}_k \end{bmatrix}, \quad k = 1, 2, 3, 4, \dots, n-1$$

使得

$$H_{n-1} \cdots H_2 H_1 A = \begin{bmatrix} r_1 & \tilde{a}_{12} & \cdots & \tilde{a}_{1n} \\ 0 & r_2 & \cdots & \tilde{a}_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & r_n \end{bmatrix} \triangleq R.$$

由于 Householder 变换都是正交矩阵, 因此  $H_1, H_2, \dots, H_{n-1}$  也都是正交矩阵. 令

$$Q = (H_{n-1} \cdots H_2 H_1)^{-1} = H_1^{-1} H_2^{-1} \cdots H_{n-1}^{-1} = H_1 H_2 \cdots H_{n-1},$$

则  $Q$  也是正交矩阵, 且

$$A = (H_{n-1} \cdots H_2 H_1)^{-1} R = QR.$$



以上就是基于 Householder 变换的 QR 分解的具体实现过程. 最后所得到的上三角矩阵  $R$  就存放在  $A$  的上三角部分. 矩阵  $Q$  可通过下面的算法实现

$$\begin{cases} Q = I_n, \\ Q = QH_k, \quad k = 1, 2, \dots, n-1. \end{cases}$$

如果  $m > n$ , 我们仍然可以通过上面的过程进行 QR 分解, 只是最后我们得到一个正交矩阵  $Q \in \mathbb{R}^{m \times m}$  和一个上三角矩阵  $R \in \mathbb{R}^{m \times n}$ , 使得  $A = QR$ .

如果不需要生成  $Q$ , 则基于 Householder 变换的 QR 分解的总运算量大约为  $2mn^2 - 2n^3/3$ .

如果保留了每一步的 Householder 向量, 则  $Q$  也可以通过下面的**向后累积方法**来计算:

$$\begin{aligned} Q &= I_n, \\ Q &= H_k Q, \quad k = n-1, n-2, \dots, 1. \end{aligned}$$

这样做的好处是一开始  $Q$  会比较稀疏, 随着迭代的进行,  $Q$  才会慢慢变满. 而前面的计算方法, 第一步就将  $Q$  变成了一个满矩阵. 采用这种方法计算  $Q$  的运算量大约为  $4m^2n - 4mn^2 + 4n^3/3$ .

如果将  $Q$  写成下面的形式

$$Q = I + WY^T,$$

则可以采用分块形式来计算  $W$  和  $Y$ , 虽然运算量会稍有增长, 但大多数运算是矩阵乘法, 因此可以尽可能多地采用 3 级 BLAS 运算, 效率可能会更高. 详情可参见 [56].

### 算法 3.5. 基于 Householder 变换的 QR 分解

% Given  $A \in \mathbb{R}^{m \times n}$ , compute  $Q$  and  $R$  such that  $A = QR$  where  $Q \in \mathbb{R}^{m \times m}$  and  $R \in \mathbb{R}^{m \times n}$

% The upper triangular part of  $R$  is stored in the upper triangular part of  $A$

```
1: Set $Q = I_{m \times m}$
2: for $k = 1$ to n do
3: $x = A(k:m, k)$
4: $[\beta, v_k] = \text{House}(x)$
5: $A(k:m, k:n) = (I_{m-k+1} - \beta v_k v_k^T) A(k:m, k:n)$
6: $\quad = A(k:m, k:n) - \beta v_k (v_k^T A(k:m, k:n))$
7: $Q(:, k:m) = Q(:, k:m) (I_{m-k+1} - \beta v_k v_k^T)$
8: $\quad = Q(:, k:m) - \beta (Q(:, k:m) v_k) v_k^T$
9: end for
```

上面的算法只是关于利用 Householder 变换来实现 QR 分解的一个简单描述, 并没有考虑运算量问题. 在实际计算时, 我们通常会保留所有的 Householder 向量. 由于第  $k$  步中  $\tilde{H}_k$  所对应的 Householder 向量  $v_k$  的长度为  $m - k + 1$ , 因此我们先把  $v_k$  单位化, 使得  $v_k$  的第一元素为 1, 这样就只要存储  $v_k(2:end)$ , 共  $m - k$  个元素. 这样, 我们就可以把所有的 Householder



向量存放在  $A$  的严格下三角部分, 而  $A$  的上三角部分仍然存放  $R$ .

✎ 计算  $Q$  可采用向后累积计算方法.

✎ 我们也可以考虑分块 Householder QR 分解, 以便充分利用 3 级 BLAS 运算, 提高计算效率.

✎ 算法 3.5 针对的是实矩阵, 如果是复矩阵则要适当做些修改.

### 3.3.4 列主元 QR 分解

如果  $A$  不是满秩, 记  $l \triangleq \text{rank}(A) < n$ , 则存在一个置换矩阵  $P$ , 使得  $AP$  的前  $l$  列是线性无关的. 因此我们可以对  $AP$  进行 QR 分解, 于是我们可以得到下面的结论.

**定理 3.11 (列主元 QR 分解)** 设  $A \in \mathbb{C}^{m \times n}$  ( $m \geq n$ ), 且  $\text{rank}(A) = l < n$ . 则存在置换矩阵  $P$  和正交矩阵  $Q \in \mathbb{C}^{m \times m}$ , 使得

$$AP = Q \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix}_{m \times n},$$

其中  $R_{11} \in \mathbb{C}^{l \times l}$  是非奇异上三角矩阵, 且对角线元素满足  $r_{11} \geq r_{22} \geq \cdots \geq r_{ll} > 0$ .

✎ 上述结论也可简化为

$$AP = Q_1 \begin{bmatrix} R_{11} & R_{12} \end{bmatrix},$$

其中  $Q_1 \in \mathbb{C}^{m \times l}$  单位列正交 (上述结论中  $Q$  的前  $l$  列).

列主元 QR 分解的实现过程与 QR 分解基本类似, 只是在第  $k$  步时, 需要选列主元, 同时可能需要做一个列交换.

假设经过  $k-1$  步后, 我们得到下面的分解

$$AP^{(k-1)} = Q^{(k-1)} \begin{bmatrix} R_{11}^{(k-1)} & R_{12}^{(k-1)} \\ 0 & R_{22}^{(k-1)} \end{bmatrix} \triangleq Q^{(k-1)} R^{(k-1)}, \quad \text{即} \quad (Q^{(k-1)})^T AP^{(k-1)} = R^{(k-1)},$$

其中  $P^{(k-1)}$  是置换矩阵,  $Q^{(k-1)}$  是正交矩阵,  $R_{11}^{(k-1)} \in \mathbb{R}^{(k-1) \times (k-1)}$  是非奇异上三角矩阵.

下面考虑第  $k$  步:

- (1) 首先计算  $R_{22}^{(k-1)}$  的所有列的范数, 如果范数都为 0, 则  $R_{22}^{(k-1)} = 0$ , 此时必有  $k-1 = l$ , 算法结束.
- (2) 当  $k \leq l$  时,  $R_{22}^{(k-1)} \neq 0$ , 记其范数最大的列为第  $i_k$  列 (如果有相等的, 取其中一个即可). 若  $i_k \neq k$ , 则交换  $R^{(k-1)}$  的第  $k$  列与第  $i_k$  列, 并记相应的置换矩阵为  $P_k$ .
- (3) 由于列交换不会影响到  $R^{(k-1)}$  的前  $k-1$  列, 因此列交换后的矩阵可记为

$$R^{(k-1)} P_k \triangleq \begin{bmatrix} R_{11}^{(k-1)} & \tilde{R}_{12}^{(k-1)} \\ 0 & \tilde{R}_{22}^{(k-1)} \end{bmatrix}.$$

构造  $\tilde{R}_{22}^{(k-1)}$  的第 1 列所对应的 Householder 变换  $\tilde{H}_k$ , 并令  $H_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & \tilde{H}_k \end{bmatrix}$ ,  $P^{(k)} =$



$P^{(k-1)}P_k$ , 于是

$$H_k \left( Q^{(k-1)} \right)^T A P^{(k)} = H_k R^{(k-1)} P_k = \begin{bmatrix} R_{11}^{(k-1)} & \tilde{R}_{12}^{(k-1)} \\ 0 & \tilde{H}_k \tilde{R}_{22}^{(k-1)} \end{bmatrix} \triangleq R^{(k)},$$

其中  $\tilde{H}_k \tilde{R}_{22}^{(k-1)}$  的第一列除第一个元素外, 其余都是零, 且该元素的值等于  $\tilde{R}_{22}^{(k-1)}$  的第 1 列的范数. 记  $Q^{(k)} \triangleq Q^{(k-1)} H_k^T$ , 则

$$A P^{(k)} = Q^{(k)} R^{(k)} = \begin{bmatrix} R_{11}^{(k)} & R_{12}^{(k)} \\ 0 & R_{22}^{(k)} \end{bmatrix},$$

其中  $R_{11}^{(k)} \in \mathbb{R}^{k \times k}$  为非奇异上三角矩阵.

依此类推, 直到第  $l$  步, 我们就可以得到  $A$  的列主元 QR 分解, 其中  $R_{11}$  的对角线元素非负且按降序排列是由列主元的选取方法和 Householder 变换的性质得到的.

### 3.3.5 基于 Givens 变换的 QR 分解

我们同样可以利用 Givens 变换来做 QR 分解.

设  $A \in \mathbb{R}^{n \times n}$ , 首先构造一个 Givens 变换  $G_{21}$ , 作用在  $A$  的最前面的两行上, 使得

$$G_{21} \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{bmatrix} = \begin{bmatrix} \tilde{a}_{11} \\ 0 \\ a_{31} \\ \vdots \\ a_{n1} \end{bmatrix}.$$

由于  $G_{21}$  只改变矩阵的第 1 行和第 2 行的值, 所以其它行保存不变. 然后再构造一个 Givens 变换  $G_{31}$ , 作用在  $G_{21}A$  的第 1 行和第 3 行, 将其第一列的第三个元素化为零. 由于  $G_{31}$  只改变矩阵的第 1 行和第 3 行的值, 所以第二行的零元素维持不变. 以此类推, 我们可以构造一系列的 Givens 变换  $G_{41}, G_{51}, \dots, G_{n1}$ , 使得  $G_{n1} \cdots G_{21}A$  的第一列中除第一个元素外, 其它元素都化为零, 即

$$G_{n1} \cdots G_{21}A = \begin{bmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{bmatrix}.$$

下面我们可以对第二列进行类似的处理. 构造 Givens 变换  $G_{32}, G_{42}, \dots, G_{n2}$ , 将第二列的第 3 至第  $n$  个元素全化为零, 同时保持第一列不变.

以此类推, 我们对其他列也做类似的处理. 最后, 通过构造  $\frac{1}{2}n(n-1)$  个 Givens 变换, 将  $A$  转化成一个上三角矩阵  $R$ , 即

$$R = G_{n,n-1} \cdots G_{21}A.$$

令  $Q = (G_{n,n-1} \cdots G_{21})^T$ . 由于 Givens 变换是正交矩阵, 所以  $Q$  也是正交矩阵. 于是, 我们就得到矩阵  $A$  的 QR 分解

$$A = QR.$$

与 Householder 变换一样, 在进行 Givens 变换时, 我们不需要显式地写出 Givens 矩阵.



- ✎ 对于稠密矩阵而言, 基于 Givens 变换的 QR 分解的运算量比 Householder 变换要多很多.
- ✎ 当需要连续应用一系列 Givens 变换时, 我们可以使用快速 Givens 变换 (见 [16]). 但即便如此, 其速度仍然要慢于 Householder 变换. 因此基于 Givens 变换的 QR 分解主要用于当矩阵的非零下三角元素相对较少时的情形 (比如对上 Hessenberg 矩阵进行 QR 分解), 或者稀疏矩阵 (可以尽可能地保留矩阵的稀疏性).
- ✎ 另外, 由于每次 Givens 变换只影响矩阵的两行或者两列, 因此非常适合并行计算.
- ✎ 如果  $A \in \mathbb{R}^{m \times n}$ , 其中  $m > n$ , 我们仍然可以通过 Givens 变换进行 QR 分解.

下面是基于 Givens 变换的 QR 分解的算法描述.

#### 算法 3.6. 基于 Givens 变换的 QR 分解

```
% Given $A \in \mathbb{R}^{m \times n}$, compute Q and R such that $A = QR$ where $Q \in \mathbb{R}^{m \times m}$ and $R \in \mathbb{R}^{m \times n}$
% The upper triangular part of R is stored in the upper triangular part of A
1: Set $Q = I_{m \times m}$
2: for $k = 1$ to n do
3: for $i = k + 1$ to m do
4: $[c, s] = \text{givens}(a_{kk}, a_{ik})$
5: $\begin{bmatrix} A(k, k:n) \\ A(i, k:n) \end{bmatrix} = G \begin{bmatrix} A(k, k:n) \\ A(i, k:n) \end{bmatrix}$ where $G = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$
6: $[Q(1:m, k), Q(1:m, i)] = [Q(1:m, k), Q(1:m, i)]G^T$
7: end for
8: end for
```

#### 3.3.6 QR 分解的稳定性

基于 Householder 变换和 Givens 变换的 QR 分解都具有很好的数值稳定性. 详细分析可以参考 [128] 和 [68]. 基于 MGS 的 QR 分解也是向后稳定的, 参见 [94].

当需要计算矩阵  $Q$  时, 基于 MGS 的 QR 分解的运算量相对较少. 因此当  $A$  的列向量具有很好的线性无关性时, 我们可以使用 MGS 来计算 QR 分解. 但需要注意的是, MGS 得到的  $Q$  不是方阵, 除非  $A$  是方阵.

但是, 由于舍入误差的原因, 最后得到的矩阵  $Q$  会带有一定的误差, 可能会导致  $Q$  失去正交性. Björck [15] 证明了, 通过 MGS 计算的矩阵  $Q$  满足

$$Q^T Q = I + E_{MGS} \quad \text{其中} \quad \|E_{MGS}\|_2 \approx \varepsilon_u \kappa_2(A).$$

而通过 Householder 变换计算的矩阵  $Q$  满足

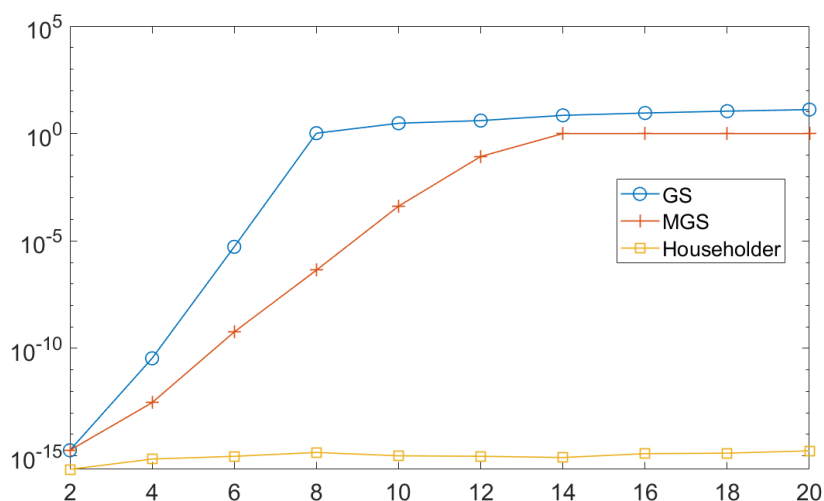
$$Q^T Q = I + E_H \quad \text{其中} \quad \|E_H\|_2 \approx \varepsilon_u.$$

因此, 如果正交性至关重要, 则当  $A$  的列向量接近线性相关时, 建议使用 Householder 变换.



**例 3.3** 编写程序, 分别用 GS, MGS 和 Householder 变换计算  $n$  阶 Hilbert 矩阵  $H$  的 QR 分解, 并比较三种算法的稳定性, 即观察  $\|\tilde{Q}\tilde{R} - H\|_2$  和  $\|\tilde{Q}^T\tilde{Q} - I\|_2$  的值, 其中  $\tilde{Q}$  和  $\tilde{R}$  是计算出来的 QR 分解矩阵因子. 试验结果如下. (QR\_3methods.m)

| $n$ | GS                             |                                  | MGS                            |                                  | Householder                    |                                  |
|-----|--------------------------------|----------------------------------|--------------------------------|----------------------------------|--------------------------------|----------------------------------|
|     | $\ \tilde{Q}\tilde{R} - H\ _2$ | $\ \tilde{Q}^T\tilde{Q} - I\ _2$ | $\ \tilde{Q}\tilde{R} - H\ _2$ | $\ \tilde{Q}^T\tilde{Q} - I\ _2$ | $\ \tilde{Q}\tilde{R} - H\ _2$ | $\ \tilde{Q}^T\tilde{Q} - I\ _2$ |
| 2   | 0.00e+00                       | 1.81e-15                         | 0.00e+00                       | 1.81e-15                         | 1.24e-16                       | 2.36e-16                         |
| 4   | 3.93e-17                       | 3.45e-11                         | 5.55e-17                       | 2.98e-13                         | 2.46e-16                       | 7.08e-16                         |
| 8   | 7.48e-17                       | 1.03e+00                         | 6.59e-17                       | 4.38e-07                         | 2.57e-16                       | 1.44e-15                         |
| 16  | 7.82e-17                       | 9.00e+00                         | 6.70e-17                       | 9.98e-01                         | 2.79e-16                       | 1.27e-15                         |



### 3.4 奇异值分解

奇异值分解 (Singular Value Decomposition, SVD) 分解是矩阵计算中非常有用的工具之一, 在图像处理、机器学习、数据科学等领域有着非常重要的应用.

#### 3.4.1 奇异值, 奇异向量和奇异值分解

设  $A \in \mathbb{C}^{m \times n}$  ( $m \geq n$ ), 则  $A^*A \in \mathbb{C}^{n \times n}$  和  $AA^* \in \mathbb{C}^{m \times m}$  都是 Hermite 半正定矩阵, 且它们具有相同的非零特征值 (都是正实数).

**定理 3.12 (SVD)** [56] 设  $A \in \mathbb{C}^{m \times n}$  ( $m \geq n$ ), 则存在酉矩阵  $U \in \mathbb{C}^{m \times m}$  和  $V \in \mathbb{C}^{n \times n}$  使得

$$U^*AV = \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \quad \text{或} \quad A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^*, \quad (3.10)$$

其中  $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n) \in \mathbb{R}^{n \times n}$ , 且  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$ . 分解 (3.10) 称为  $A$  的**奇异值分解 (SVD)**, 而  $\sigma_1, \sigma_2, \dots, \sigma_n$  则称为  $A$  的**奇异值**.

(板书)

**证明.** 首先假设  $A \neq 0$ , 否则只需令  $\Sigma = 0$  即可,  $U$  和  $V$  可以是任意酉矩阵.

下面我们对  $m$  和  $n$  用归纳法来证明.

当  $n = 1, m \geq 1$  时, 我们取  $\Sigma = \|A\|_2, V = 1, U \in \mathbb{C}^{m \times m}$  是第一列为  $u_1 = A/\|A\|_2$  的酉矩阵, 于是

$$A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^*$$

即为  $A$  的奇异值分解.

假设  $\mathbb{C}^{(m-1) \times (n-1)}$  中的矩阵都存在奇异值分解, 下面证明  $A \in \mathbb{C}^{m \times n}$  也存在有奇异值分解. 由 2-范数的定义可知, 存在向量  $v \in \mathbb{C}^n$  满足  $\|v\|_2 = 1$  使得  $\|A\|_2 = \|Av\|_2$ . 令

$$u = \frac{1}{\sigma} Av \in \mathbb{C}^m, \quad \text{其中 } \sigma = \|A\|_2,$$

则  $\|u\|_2 = 1$ . 我们将  $v$  和  $u$  都扩充成酉矩阵, 即存在  $\tilde{U} \in \mathbb{C}^{m \times (m-1)}$  和  $\tilde{V} \in \mathbb{C}^{n \times (n-1)}$ , 使得  $[u, \tilde{U}] \in \mathbb{C}^{m \times m}$  和  $[v, \tilde{V}] \in \mathbb{C}^{n \times n}$  都是酉矩阵. 于是

$$\tilde{U}^*Av = \tilde{U}^*(\sigma u) = 0, \quad u^*Av = u^*(\sigma u) = \sigma.$$

所以

$$\tilde{A} \triangleq [u, \tilde{U}]^* A [v, \tilde{V}] = \begin{bmatrix} u^*Av & u^*A\tilde{V} \\ \tilde{U}^*Av & \tilde{U}^*A\tilde{V} \end{bmatrix} = \begin{bmatrix} \sigma & u^*A\tilde{V} \\ 0 & \tilde{U}^*A\tilde{V} \end{bmatrix}.$$

又

$$\sigma = \|A\|_2 = \|\tilde{A}\|_2 = \|\tilde{A}^*\|_2 \geq \|\tilde{A}^*e_1\|_2 = \left\| \begin{bmatrix} \sigma & u^*A\tilde{V} \end{bmatrix}^* \right\|_2 = \sqrt{\sigma^2 + \|u^*A\tilde{V}\|_2^2},$$

所以  $\|u^*A\tilde{V}\|_2 = 0$ , 即  $u^*A\tilde{V} = 0$ . 于是

$$[u, \tilde{U}]^* A [v, \tilde{V}] = \begin{bmatrix} \sigma & 0 \\ 0 & A_1 \end{bmatrix},$$





其中  $A_1 = \tilde{U}^* A \tilde{V} \in \mathbb{C}^{(m-1) \times (n-1)}$ . 由归纳假设可知,  $A_1$  存在奇异值分解, 设为

$$A_1 = U_1 \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} V_1^*,$$

其中  $U_1 \in \mathbb{C}^{(m-1) \times (m-1)}$  和  $V_1 \in \mathbb{C}^{(n-1) \times (n-1)}$  都是酉矩阵,  $\Sigma_1 \in \mathbb{R}^{(n-1) \times (n-1)}$  是对角矩阵, 且其对角线元素按降序排列. 令

$$U = \begin{bmatrix} u, \tilde{U} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & U_1 \end{bmatrix}, \quad V = \begin{bmatrix} v, \tilde{V} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V_1 \end{bmatrix},$$

则  $U \in \mathbb{C}^{m \times m}$  和  $V \in \mathbb{C}^{n \times n}$  都是酉矩阵, 且


$$\begin{aligned} U^* A V &= \begin{bmatrix} 1 & 0 \\ 0 & U_1 \end{bmatrix}^* \begin{bmatrix} u, \tilde{U} \end{bmatrix}^* A \begin{bmatrix} v, \tilde{V} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V_1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & U_1 \end{bmatrix}^* \begin{bmatrix} \sigma & 0 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V_1 \end{bmatrix} \\ &= \begin{bmatrix} \sigma & 0 \\ 0 & U_1^* A_1 V_1 \end{bmatrix} = \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}, \end{aligned} \quad (3.11)$$


其中  $\Sigma = \begin{bmatrix} \sigma & 0 \\ 0 & \Sigma_1 \end{bmatrix}$ . 又

$$\sigma^2 = \|A\|_2^2 = \|U^* A V\|_2^2 = \left\| \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} \right\|_2^2 = \rho \left( \begin{bmatrix} \sigma^2 & 0 \\ 0 & \Sigma_1^2 \end{bmatrix} \right),$$

所以  $\sigma$  不小于  $\Sigma_1$  中的所有对角线元素, 即  $\Sigma$  的对角线元素也是按降序排列. 因此, (3.11) 这就是  $A$  的奇异值分解.

由归纳法可知, 定理的结论成立. □

 该定理也可以通过 Hermite 半正定矩阵的特征值分解来证明.

 如果  $A \in \mathbb{R}^{m \times n}$  是实矩阵, 则  $U, V$  也都可以是实矩阵 [71].

由 (3.10) 可知,

$$A^* A = V \Sigma^* \Sigma V^*, \quad A A^* = U \begin{bmatrix} \Sigma \Sigma^* & 0 \\ 0 & 0 \end{bmatrix} U^*.$$

所以  $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$  是  $A^* A$  和  $A A^*$  的特征值. 因此,  $A$  的奇异值就是  $A^* A$  的特征值的平方根.

### 奇异向量

矩阵  $U = [u_1, u_2, \dots, u_m]$  和  $V = [v_1, v_2, \dots, v_n]$  的列向量分别称为  $A$  的左奇异向量和右奇异向量, 即存在关系式

$$\begin{aligned} A v_i &= \sigma_i u_i, \quad i = 1, 2, \dots, n, \\ A^* u_i &= \sigma_i v_i, \quad i = 1, 2, \dots, n, \\ A^* u_i &= 0, \quad i = n+1, n+2, \dots, m. \end{aligned}$$





**细 SVD (降阶 SVD)**

由定理 3.12 可知

$$A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^* = \sigma_1 u_1 v_1^* + \sigma_2 u_2 v_2^* + \cdots + \sigma_n u_n v_n^* = \sum_{i=1}^n \sigma_i u_i v_i^*.$$

记  $U_n = [u_1, u_2, \dots, u_n] \in \mathbb{C}^{m \times n}$ , 则  $U_n$  是单位列正交矩阵 (即  $U_n^* U_n = I_{n \times n}$ ), 且

$$A = U_n \Sigma V^*. \quad (3.12)$$

这就是所谓的**细 SVD** (或瘦 SVD, **thin SVD**, skinny SVD) [56] 或**降阶 SVD** (**reduced SVD**) [119], 有的文献将 (3.12) 称为奇异值分解.

**压缩 SVD**

设  $l = \triangleq \text{rank}(A) \leq n$ , 则有

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_l > 0, \quad \sigma_{l+1} = \cdots = \sigma_n = 0.$$

我们称

$$A_l = \sum_{i=1}^l \sigma_i u_i v_i^*$$

为  $A$  的**压缩 SVD** (condensed SVD).

**截断 SVD**

设  $k < n$ , 我们称

$$A_k = \sigma_1 u_1 v_1^* + \sigma_2 u_2 v_2^* + \cdots + \sigma_k u_k v_k^* = \sum_{i=1}^k \sigma_i u_i v_i^*$$

为  $A$  的**截断 SVD** (**truncated SVD**).

由于压缩 SVD 和截断 SVD 在计算和存储方面存在优势, 因此实际应用中经常使用的是压缩 SVD 或截断 SVD.

**奇异值的应用**

- 矩阵计算: 矩阵范数, 矩阵条件数, 最小二乘问题, 广义逆, 矩阵和张量的低秩分解, 等等.
- 工程应用: 信号处理, 图像压缩, 机器学习, 主成分分析, 数据降维, 等等.

**Six Great Theorems of Linear Algebra**

- **Dimension Theorem:** All bases for a vector space have the same number of vectors.
- **Counting Theorem:** Dimension of column space + dimension of nullspace = number of columns.
- **Rank Theorem:** Dimension of column space = dimension of rowspace. This is the rank.



- **Fundamental Theorem:** The row space and nullspace of  $A$  are orthogonal complements in  $\mathbb{R}^n$ .
- **SVD:** There are orthonormal bases ( $v$ 's and  $u$ 's for the row and column spaces) so that  $Av_i = \sigma_i u_i$ .
- **Spectral Theorem:** If  $A^T = A$  there are orthonormal  $q$ 's so that  $Aq_i = \lambda_i q_i$  and  $A = Q\Lambda Q^T$ .

— *Introduction to Linear Algebra*, 5th, G. Strang, 2016.

### 3.4.2 奇异值基本性质

下面是关于奇异值的一些基本性质:

**定理 3.13** 设  $A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^*$  是  $A \in \mathbb{C}^{m \times n}$  ( $m \geq n$ ) 的奇异值分解, 则下面结论成立:

- (1)  $A^*A$  的特征值是  $\sigma_i^2$ , 对应的特征向量是  $v_i, i = 1, 2, \dots, n$ ;
- (2)  $AA^*$  的特征值是  $\sigma_i^2$  和  $m - n$  个零, 对应的特征向量是  $u_i, i = 1, 2, \dots, m$ ;
- (3)  $\|A\|_2 = \sigma_1, \|A\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}$ ;
- (4) 若  $\text{rank}(A) = r \leq n$ , 则

$$\text{Ran}(A) = \text{span}\{u_1, u_2, \dots, u_r\}, \quad \text{Ker}(A) = \text{span}\{v_{r+1}, v_{r+2}, \dots, v_n\};$$

- (5) 设  $x \in \mathbb{C}^n$  且  $\|x\|_2 = 1$ , 则  $\sigma_n \leq \|Ax\|_2 \leq \sigma_1$ ;
- (6) (酉不变性) 设  $X \in \mathbb{C}^{m \times m}$  和  $Y \in \mathbb{C}^{n \times n}$  是酉矩阵, 则  $\sigma_i(X^*AY) = \sigma_i(A)$ .

(留作练习)

**定理 3.14** 设  $A = U\Sigma V^*$  是  $A \in \mathbb{C}^{n \times n}$  的奇异值分解, 则下面结论成立:

- (1)  $|\det(A)| = \sigma_1 \sigma_2 \cdots \sigma_n$ ;
- (2) 若  $A$  非奇异, 则  $\|A^{-1}\|_2 = \sigma_n^{-1}, \kappa_2(A) = \sigma_1/\sigma_n$ ;
- (3) 若  $A$  是 Hermite 的, 且  $A = U\Lambda U^*$  是  $A$  的酉特征值分解, 即  $U^*U = I, \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ . 设  $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$ , 则  $A = U\Sigma V$  是  $A$  的奇异值分解, 其中  $\sigma_i = |\lambda_i|, v_i = \text{sign}(\lambda_i)u_i$ , 若  $\lambda_i = 0$ , 则取  $v_i = u_i$ ;
- (4) 矩阵  $H = \begin{bmatrix} 0 & A^* \\ A & 0 \end{bmatrix}$  的特征值是  $\pm\sigma_i$ , 对应的单位特征向量为  $\frac{1}{\sqrt{2}} \begin{bmatrix} v_i \\ \pm u_i \end{bmatrix}$ . ( $A \in \mathbb{C}^{m \times n}$  时也有类似结论)

(留作练习)

**例 3.4** 由奇异值的性质可知, 矩阵的谱条件数取决于最大奇异值和最小奇异值的比值, 如果最大奇异值与最小奇异值比较接近, 则谱条件数就比较小. 那么, 谱条件数与特征值有没有直接关系? 我们知道, 如果矩阵  $A$  对称正定, 则其谱条件数就是最大特征值与最小特征值的比值. 但是,



对于一般的矩阵, 则没有这个性质. 如下面的矩阵:

$$A = \begin{bmatrix} 1 & -\frac{1}{2} & \cdots & -\frac{1}{2} \\ & 1 & \ddots & \vdots \\ & & \ddots & -\frac{1}{2} \\ & & & 1 \end{bmatrix} \in \mathbb{R}^{n \times n},$$

即对角线全是 1, 严格上三角部分全是  $-\frac{1}{2}$ , 其它为 0. 易知  $A$  的所有特征值都是 1, 但其谱条件数却随着矩阵规模的增长而快速变大, 见下表:

| $n$           | 10      | 20      | 30      | 40      | 50      | 60      | 70      | 80      |
|---------------|---------|---------|---------|---------|---------|---------|---------|---------|
| $\kappa_2(A)$ | 6.3e+01 | 7.6e+03 | 6.8e+05 | 5.3e+07 | 3.9e+09 | 2.7e+11 | 1.8e+13 | 1.2e+15 |

另外, 通过直接计算可知

$$A^{-1} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1.5}{2} & \frac{1.5^2}{2} & \cdots & \frac{1.5^{n-2}}{2} \\ & 1 & \frac{1}{2} & \frac{1.5}{2} & \ddots & \vdots \\ & & 1 & \ddots & \ddots & \frac{1.5^2}{2} \\ & & & \ddots & \frac{1}{2} & \frac{1.5}{2} \\ & & & & 1 & \frac{1}{2} \\ & & & & & 1 \end{bmatrix}.$$

因此,  $\kappa_1(A)$  和  $\kappa_\infty(A)$  都是矩阵维数  $n$  的幂函数.

### 3.4.3 奇异值更多性质

下面是关于矩阵  $A$  的一个低秩逼近.

**定理 3.15** 设  $A = U_n \Sigma V^*$  是  $A \in \mathbb{C}^{m \times n}$  的细奇异值分解. 令  $A_k = \sum_{i=1}^k \sigma_i u_i v_i^*$ , 则  $A_k$  是

$$\min_{B \in \mathbb{C}^{m \times n}, \text{rank}(B)=k} \|A - B\|_2 \quad (3.13)$$

的一个解, 且

$$\|A - A_k\|_2 = \sigma_{k+1}.$$

此时, 我们称  $A_k$  是  $A$  的一个秩  $k$  逼近.

(板书)

**证明.** 设  $B \in \mathbb{C}^{m \times n}$  且  $\text{rank}(B) = k$ , 则

$$\dim(\text{Ker}(B)) + \dim(\text{span}\{V_{k+1}\}) = (n - k) + (k + 1) = n + 1 > n,$$

其中  $V_{k+1} = [v_1, v_2, \dots, v_{k+1}] \in \mathbb{C}^{n \times (k+1)}$ . 所以  $\text{Ker}(B)$  与  $\text{span}\{V_{k+1}\}$  有非零公共元素. 令  $0 \neq x \in \text{Ker}(B) \cap \text{span}\{V_{k+1}\}$ , 不失一般性, 我们假设  $\|x\|_2 = 1$ . 故存在  $y \in \mathbb{C}^{k+1}$  满足  $\|y\|_2 = 1$  使得  $x = V_{k+1}y$ . 于是

$$\|A - B\|_2^2 \geq \|(A - B)x\|_2^2 = \|Ax\|_2^2 = \|U_n \Sigma V^* V_{k+1} y\|_2^2$$



$$= \left\| \Sigma \begin{bmatrix} I_{k+1} \\ 0 \end{bmatrix} y \right\|_2^2 = \left\| \begin{bmatrix} \Sigma_1 y \\ 0 \end{bmatrix} \right\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 |y_i|^2 \geq \sigma_{k+1}^2,$$

其中  $\Sigma_1 = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_{k+1})$ . 这里我们利用了性质  $\sum_{i=1}^{k+1} |y_i|^2 = \|y\|_2^2 = 1$ . 所以

$$\min_{B \in \mathbb{C}^{m \times n}, \text{rank}(B)=k} \|A - B\|_2 \geq \sigma_{k+1}.$$

又  $\text{rank}(A_k) = k$ , 且

$$\|A - A_k\|_2 = \left\| \sum_{i=k+1}^n \sigma_i u_i v_i^* \right\|_2 = \left\| U_n \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & \ddots \\ & & \sigma_{k+1} & \\ & & & \ddots \\ & & & & \sigma_n \end{bmatrix} V^* \right\|_2 = \sigma_{k+1},$$

所以

$$\min_{B \in \mathbb{C}^{m \times n}, \text{rank}(B)=k} \|A - B\|_2 = \sigma_{k+1},$$

且  $A_k$  是问题 (3.13) 的一个解. □

定理中的 (3.13) 式也可以改写为

$$\min_{B \in \mathbb{C}^{m \times n}, \text{rank}(B) \leq k} \|A - B\|_2 \quad (3.14)$$

对于 Frobenius 范数, 我们有类似的结论, 见习题 3.11.

**定理 3.16 (Weyl)** [113, page 67] 设  $A, B \in \mathbb{C}^{m \times n}$  ( $m \geq n$ ), 且  $\text{rank}(B) = k$ . 则有

$$\max_{x \in \text{Ker}(B), \|x\|_2=1} \|Ax\|_2 \geq \sigma_{k+1}(A), \quad (3.15)$$

和

$$\min_{x \in \text{Ker}(B), \|x\|_2=1} \|Ax\|_2 \leq \sigma_{n-k}(A). \quad (3.16)$$

因此,

$$\sigma_1(A - B) \geq \sigma_{k+1}(A), \quad \sigma_n(A - B) \leq \sigma_{n-k}(A) \quad (3.17)$$

且

$$\sigma_{i+j-1}(A) \leq \sigma_i(B) + \sigma_j(A - B), \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, n - i + 1. \quad (3.18)$$

(板书)

**证明.** 不等式 (3.15) 也可以写为: 设  $\mathcal{L}$  是  $\mathbb{C}^n$  中的任意一个  $n - k$  维子空间, 则有

$$\max_{x \in \mathcal{L}, \|x\|_2=1} \|Ax\|_2 \geq \sigma_{k+1}(A).$$

该不等式的证明与定理 3.15 的证明类似, 不再赘述.

下面证明结论 (3.16). 令  $\tilde{V}_{k+1} = [v_{n-k}, v_{n-k+1}, \dots, v_n] \in \mathbb{C}^{n \times (k+1)}$ . 类似地, 存在向量  $y \in$



$\mathbb{C}^{k+1}$  满足  $\|y\|_2 = 1$ , 使得  $x = \tilde{V}_{k+1}y \in \text{Ker}(B)$ . 于是

$$\|Ax\|_2^2 = \|U_n \Sigma V^* \tilde{V}_{k+1}y\|_2^2 = \left\| \Sigma \begin{bmatrix} 0 \\ I_{k+1} \end{bmatrix} y \right\|_2^2 = \left\| \begin{bmatrix} 0 \\ \tilde{\Sigma}_{k+1} \end{bmatrix} y \right\|_2^2 = \sum_{i=1}^{k+1} \sigma_{n-k-1+i}^2 |y_i|^2 \leq \sigma_{n-k}^2,$$

其中  $\tilde{\Sigma}_{k+1} = \text{diag}(\sigma_{n-k}, \sigma_{n-k+1}, \dots, \sigma_n)$ . 所以

$$\min_{x \in \text{Ker}(B), \|x\|_2=1} \|Ax\|_2 \leq \sigma_{n-k}(A).$$

不等式 (3.17) 可由 (3.15), (3.16) 以及定理 3.13 中的性质 5 得到.

下面证明不等式 (3.18). 首先证明  $i = j = 1$  时结论成立. 事实上, 我们有

$$\sigma_1(A) = \|A\|_2 = \|B + (A - B)\|_2 \leq \|B\|_2 + \|A - B\|_2 = \sigma_1(B) + \sigma_1(A - B).$$

令  $C = A - B$ , 并设  $B_{i-1}$  和  $C_{j-1}$  分别是  $B$  和  $C$  的秩  $i-1$  和秩  $j-1$  逼近. 则

$$\sigma_1(B - B_{i-1}) = \|B - B_{i-1}\|_2 = \sigma_i(B), \quad \sigma_1(C - C_{j-1}) = \|C - C_{j-1}\|_2 = \sigma_j(C).$$


所以

$$\begin{aligned} \sigma_i(B) + \sigma_j(C) &= \sigma_1(B - B_{i-1}) + \sigma_1(C - C_{j-1}) \\ &\geq \sigma_1(A - B_{i-1} - C_{j-1}). \end{aligned}$$

又  $\text{rank}(B_{i-1} + C_{j-1}) \leq i + j - 2$ , 所以由不等式 (3.17) 可知

$$\sigma_i(B) + \sigma_j(C) = \sigma_1(A - B_{i-1} - C_{j-1}) \geq \sigma_{i+j-1}(A).$$

□

 从该结论可知, 如果通过对  $A$  进行秩  $k$  ( $1 \leq k \leq \frac{n}{2}$ ) 修正来改善其谱条件数, 则最佳情况是降低到  $\frac{\sigma_{n-k}(A)}{\sigma_{k+1}}$ .

根据定理 3.16, 我们可以得到矩阵奇异值的最小最大定理.

**定理 3.17** 设  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$  是矩阵  $A \in \mathbb{C}^{m \times n}$  ( $m \geq n$ ) 的奇异值, 则

$$\sigma_k(A) = \min_{\dim(\mathcal{L})=n-k+1} \max_{x \in \mathcal{L}, \|x\|_2=1} \|Ax\|_2, \quad k = 1, 2, \dots, n$$

和

$$\sigma_k(A) = \max_{\dim(\mathcal{L})=k} \min_{x \in \mathcal{L}, \|x\|_2=1} \|Ax\|_2, \quad k = 1, 2, \dots, n,$$

其中  $\mathcal{L}$  表示  $\mathbb{C}^n$  的一个子空间.

(留作练习, 也可利用 Hermite 矩阵的特征值最小最大定理)

**引理 3.18 (交错不等式)** [72, page 149] 设  $A \in \mathbb{C}^{m \times n}$ ,  $A_r$  是由  $A$  去除  $r$  列 (或  $r$  行) 后得到的子矩阵, 则

$$\sigma_i(A) \geq \sigma_i(A_r) \geq \sigma_{i+r}(A), \quad i = 1, 2, \dots, \min\{m, n\}.$$

这里, 当下标  $k$  大于矩阵的维数时, 我们令  $\sigma_k = 0$ . 更进一步, 如果  $B \in \mathbb{C}^{(m-r) \times (n-s)}$  是  $A$  的子



矩阵, 则

$$\sigma_i(A) \geq \sigma_i(B) \geq \sigma_{i+r+s}(A).$$

(留作课外自习)

**证明.** 只需证明  $r = 1$  的情形, 其余情形可以通过递推实现.

假设  $A_1$  是由  $A$  中去除第  $k$  列后的子矩阵. 令  $e_k \in \mathbb{C}^n$  为单位矩阵的第  $k$  列, 即第  $k$  个元素是 1, 其它均为 0. 由定理 3.17 可知

$$\sigma_k(A) = \min_{\dim(\mathcal{L})=n-k+1} \max_{x \in \mathcal{L}, \|x\|_2=1} \|Ax\|_2 \geq \min_{\dim(\mathcal{L})=n-k+1} \max_{x \in \mathcal{L}, \|x\|_2=1, x \perp e_k} \|Ax\|_2.$$

如果  $x \perp e_k$ , 则  $x_k = 0$ , 所以

$$Ax = A_1 \tilde{x},$$

其中  $\tilde{x} = [x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n]^T \in \mathbb{C}^{n-1}$ . 故条件 “ $x \in \mathcal{L}$ ,  $\|x\|_2 = 1$ ,  $x \perp e_k$ ” 就等价于 “ $x \in \tilde{\mathcal{L}}$ ,  $\|\tilde{x}\|_2 = 1$ ”, 其中  $\tilde{\mathcal{L}} \subset \mathbb{C}^{n-1}$  是由  $\mathcal{L}$  中的向量去除第  $k$  个分量后组成的集合. 因此  $\dim(\tilde{\mathcal{L}}) = \dim(\mathcal{L}) = n - k + 1$  或  $\dim(\tilde{\mathcal{L}}) = \dim(\mathcal{L}) - 1 = n - k$ . 于是

$$\min_{\dim(\mathcal{L})=n-k+1} \max_{x \in \mathcal{L}, \|x\|_2=1, x \perp e_k} \|Ax\|_2 \geq \min_{\dim(\tilde{\mathcal{L}})=n-k+1 \text{ 或 } \dim(\tilde{\mathcal{L}})=n-k} \max_{\tilde{x} \in \tilde{\mathcal{L}}, \|\tilde{x}\|_2=1} \|A_1 \tilde{x}\|_2.$$

又

$$\min_{\dim(\tilde{\mathcal{L}})=n-k+1} \max_{\tilde{x} \in \tilde{\mathcal{L}}, \|\tilde{x}\|_2=1} \|A_1 \tilde{x}\|_2 \geq \min_{\dim(\tilde{\mathcal{L}})=n-k} \max_{\tilde{x} \in \tilde{\mathcal{L}}, \|\tilde{x}\|_2=1} \|A_1 \tilde{x}\|_2,$$

且  $A_1 \in \mathbb{C}^{m \times (n-1)}$ , 所以

$$\begin{aligned} \sigma_k(A) &\geq \min_{\dim(\mathcal{L})=n-k+1} \max_{x \in \mathcal{L}, \|x\|_2=1, x \perp e_k} \|Ax\|_2 \\ &\geq \min_{\dim(\tilde{\mathcal{L}})=(n-1)-k+1} \max_{\tilde{x} \in \tilde{\mathcal{L}}, \|\tilde{x}\|_2=1} \|A_1 \tilde{x}\|_2 \\ &= \sigma_k(A_1). \end{aligned}$$

同理可得

$$\begin{aligned} \sigma_{k+1}(A) &= \max_{\dim(\mathcal{L})=k+1} \min_{x \in \mathcal{L}, \|x\|_2=1} \|Ax\|_2 \\ &\leq \max_{\dim(\mathcal{L})=k+1} \min_{x \in \mathcal{L}, \|x\|_2=1, x \perp e_k} \|Ax\|_2 \\ &\leq \max_{\dim(\tilde{\mathcal{L}})=k} \min_{\tilde{x} \in \tilde{\mathcal{L}}, \|\tilde{x}\|_2=1} \|A_1 \tilde{x}\|_2 \\ &= \sigma_k(A_1). \end{aligned}$$

我们注意到, 在前面的证明中, 并没有要求  $m \geq n$ . 因此对于  $m < n$  的情形, 上面的结论仍然成立.

如果  $A_1$  是由  $A$  中去除第  $k$  行后的子矩阵. 由于  $A^*$  与  $A$  具有相同的奇异值, 因此只需将上面的讨论运用到  $A^*$  上即可.  $\square$

**引理 3.19** [72, page 170] 设  $A \in \mathbb{C}^{n \times n}$ ,  $1 \leq k \leq n$ , 则对任意的单位列正交矩阵  $U_k \in \mathbb{C}^{n \times k}$  和



$V_k \in \mathbb{C}^{n \times k}$ , 有

$$\sigma_i(U_k^* A V_k) \leq \sigma_i(A), \quad i = 1, 2, \dots, k. \quad (3.19)$$

因此,

$$|\det(U_k^* A V_k)| \leq \sigma_1(A) \sigma_2(A) \cdots \sigma_k(A). \quad (3.20)$$

(板书)

**证明.** 我们将  $U_k$  和  $V_k$  都扩充成酉矩阵  $U \in \mathbb{C}^{m \times m}$  和  $V \in \mathbb{C}^{n \times n}$ , 则  $U_k^* A V_k$  是  $U^* A V$  的子矩阵. 由引理 3.18 可知,

$$\sigma_i(U_k^* A V_k) \leq \sigma_i(U^* A V).$$

又  $U^* A V$  与  $A$  有相同的奇异值, 故结论 (3.19) 成立.

利用定理 3.14 中的性质 1, 即可得到不等式 (3.20). □

**定理 3.20 (Weyl 不等式, 1949)** [72, page 171] 设  $A \in \mathbb{C}^{n \times n}$ , 其奇异值和特征值分别为  $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$  和  $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ . 则

$$|\lambda_1 \lambda_2 \cdots \lambda_k| \leq \sigma_1 \sigma_2 \cdots \sigma_k, \quad k = 1, 2, \dots, n$$

且当  $k = n$  时, 等号成立.

(板书)

**证明.** 设  $A = U R U^*$  是  $A$  的 Schur 分解, 其中  $U \in \mathbb{C}^{n \times n}$  是酉矩阵,  $R \in \mathbb{C}^{n \times n}$  是上三角矩阵, 且其对角线元素依次为  $\lambda_1, \lambda_2, \dots, \lambda_n$ . 取  $U$  的前  $k$  列组成一个单位列正交矩阵  $U_k$ , 则  $R_k \triangleq U_k^* A U_k$  为  $U^* A U = R$  的  $k$  阶顺序主子矩阵, 即  $R_k$  也是上三角矩阵, 且其对角线元素依次为  $\lambda_1, \lambda_2, \dots, \lambda_k$ . 所以由引理 3.19 可得

$$|\lambda_1 \lambda_2 \cdots \lambda_k| = |\det(R_k)| = |\det(U_k^* A U_k)| \leq \sigma_1 \sigma_2 \cdots \sigma_k.$$

当  $k = n$  时,

$$|\lambda_1 \lambda_2 \cdots \lambda_n| = |\det(A)| = \sigma_1 \sigma_2 \cdots \sigma_n. \quad \square$$

### 3.4.4 奇异值扰动分析

下面的定理是关于奇异值的一个扰动性质.

**定理 3.21** [113, page 69] 设  $A, E \in \mathbb{C}^{m \times n}$  ( $m \geq n$ ). 则

$$|\sigma_i(A + E) - \sigma_i(A)| \leq \|E\|_2, \quad i = 1, 2, \dots, n.$$

(留作练习, 可直接利用不等式 (3.18))

下面是定理 3.21 的一个推论, 也是 SVD 的一个重要应用.

**推论 3.22** 设  $A \in \mathbb{C}^{n \times n}$ ,  $\|\cdot\|$  是任意一个相容矩阵范数, 则对任意的  $\varepsilon > 0$ , 总存在一个矩阵  $A_\varepsilon$  使得  $\|A - A_\varepsilon\| \leq \varepsilon$ , 其中  $A_\varepsilon$  具有互不相同的特征值.

由推论 3.22 可知, 可对角化矩阵在所有矩阵组成的集合中是稠密的.



### 3.5 线性最小二乘问题的求解方法

#### 3.5.1 正规方程

这里我们考虑超定线性最小二乘问题的求解.

**定理 3.23** 设  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ). 则  $x_* \in \mathbb{R}^n$  是线性最小二乘问题 (3.1) 的解当且仅当残量  $r = b - Ax_*$  与  $\text{Ran}(A)$  (值域) 正交, 即  $x_*$  是下面的**正规方程**的解

$$A^T(b - Ax) = 0 \quad \text{或} \quad A^T Ax = A^T b. \quad (3.21)$$

(板书)

**证明.** 充分性: 设  $x_*$  是正规方程 (3.21) 的解. 对任意向量  $y \in \mathbb{R}^n$ , 由  $(b - Ax_*) \perp \text{Ran}(A)$  可知

$$\begin{aligned} \|Ay - b\|_2^2 &= \|(Ax_* - b) + A(y - x_*)\|_2^2 \\ &= \|Ax_* - b\|_2^2 + \|A(y - x_*)\|_2^2 \\ &\geq \|Ax_* - b\|_2^2. \end{aligned}$$

因此,  $x_*$  是线性最小二乘问题 (3.1) 的解.

必要性: 设  $x_*$  是线性最小二乘问题 (3.1) 的解. 用反证法, 假定  $z \triangleq A^T(b - Ax_*) \neq 0$ . 取  $y = x_* + \alpha z$ , 其中  $\alpha > 0$ , 则有

$$\|Ay - b\|_2^2 = \|Ax_* - b + \alpha Az\|_2^2 = \|Ax_* - b\|_2^2 - 2\alpha \|z\|_2^2 + \alpha^2 \|Az\|_2^2.$$

由于  $\|z\|_2 > 0$ , 当  $\alpha$  充分小时, 有  $2\|z\|_2^2 > \alpha \|Az\|_2^2$ , 即上式右端小于  $\|Ax_* - b\|_2^2$ . 这与  $x_*$  是最小二乘解相矛盾. 所以  $z = 0$ , 即  $A^T(b - Ax_*) = 0$ .  $\square$

由定理 3.23 可知, 求线性最小二乘问题 (3.1) 的解等价于求正规方程 (3.21) 的解. 由于

$$A^T b \in \text{Ran}(A^T) = \text{Ran}(A^T A),$$


因此正规方程  $A^T Ax = A^T b$  是相容 (consistent) 的, 即**最小二乘解总是存在的**. 当  $A$  非奇异时, 这个解也是唯一的.

**定理 3.24** 设  $A \in \mathbb{R}^{m \times n}$  ( $m > n$ ). 则  $A^T A$  对称正定当且仅当  $A$  是列满秩的, 即  $\text{rank}(A) = n$ . 此时, 线性最小二乘问题 (3.1) 的解是唯一的, 其表达式为

$$x = (A^T A)^{-1} A^T b.$$

当  $A$  列满秩时, 我们就可以使用 Cholesky 分解来求解正规方程, 总的运算量大约为  $mn^2 + \frac{1}{3}n^3 + mn + \mathcal{O}(n^2)$ , 其中大部分的运算量 ( $mn^2$ ) 是用来计算  $A^T A$  (由于  $A^T A$  对称, 因此只需计算其下三角部分).

 通过直接求解正规方程来求解最小二乘问题, 运算量小, 而且简单直观.

 但由于  $A^T A$  的条件数是  $A$  的条件数的平方, 因此对于病态情形 (即  $A$  的条件数比较大), 不建议使用该方法.





**例 3.5** 下面的例子说明, 计算  $A^T A$  可能会损失计算精度: 设

$$A = \begin{bmatrix} 1 & 1 & 1 \\ \varepsilon & & \\ & \varepsilon & \\ & & \varepsilon \end{bmatrix},$$

则

$$A^T A = \begin{bmatrix} 1 + \varepsilon^2 & 1 & 1 \\ 1 & 1 + \varepsilon^2 & 1 \\ 1 & 1 & 1 + \varepsilon^2 \end{bmatrix}.$$

记  $\varepsilon_u$  为机器精度, 则当  $\varepsilon_u < \varepsilon < \sqrt{\varepsilon_u}$  时有  $\varepsilon^2 < \varepsilon_u$ , 由于舍入误差的原因, 通过浮点运算计算得到的  $A^T A$  是奇异的. 但我们注意到  $A$  是满秩的.

### 最小二乘解的几何含义

根据定理 3.23, 我们可以把  $b$  写成

$$b = Ax_* + r, \quad \text{其中 } r \perp \text{Ran}(A). \quad (3.22)$$

所以  $Ax_*$  就是  $b$  在  $\text{Ran}(A)$  上的正交投影, 见图 3.3.

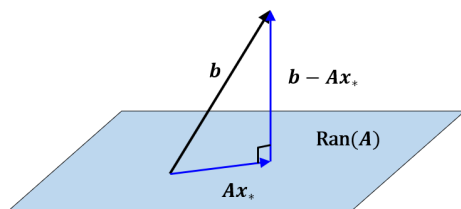



图 3.3. 最小二乘解的几何描述

 最小二乘解可能并不唯一, 但分解 (3.22) 总是唯一的.

### 最小二乘与鞍点问题

由定理 3.23 可知, 线性最小二乘问题 (3.1) 等价于

$$A^T r = 0, \quad r = b - Ax,$$

即

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}. \quad (3.23)$$

这就是线性最小二乘问题 (3.1) 的**增广方程** (augmented system). 事实上, 方程组 (3.23) 是下面方程组的一种特殊情形

$$\begin{bmatrix} B & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix},$$

其中  $B \in \mathbb{R}^{m \times m}$  对称半正定. 这就是通常所说的**鞍点问题**. 这个方程组存在唯一解当且仅当  $A$  列满秩且矩阵  $[B, A]$  行满秩 (见练习 3.2).

如果  $B$  对称正定, 则  $r = B^{-1}(f - Ax)$ , 代入第二个方程可得

$$A^T B^{-1} A x = A^T B^{-1} f - g.$$

这就是**广义的正规方程**. 其所对应的**广义线性最小二乘问题**是

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - f\|_{B^{-1}}^2 + g^T x,$$

其中范数  $\|\cdot\|_{B^{-1}}$  的定义是  $\|x\|_{B^{-1}}^2 = x^T B^{-1} x$ , 这里要求  $B$  是对称正定的.

### 3.5.2 QR 分解法

这里假定  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ) 是满秩的. 设  $A$  的 QR 分解为  $A = QR$ , 我们用三种不同的方法来推导线性最小二乘问题的解.

- 将  $Q$  的扩充成一个正交矩阵, 记为  $[Q, \hat{Q}] \in \mathbb{R}^{m \times m}$ . 于是有

$$\begin{aligned} \|Ax - b\|_2^2 &= \|[Q, \hat{Q}]^T (Ax - b)\|_2^2 = \|[Q, \hat{Q}]^T (QRx - b)\|_2^2 \\ &= \left\| \begin{bmatrix} Rx - Q^T b \\ -\hat{Q}^T b \end{bmatrix} \right\|_2^2 = \|Rx - Q^T b\|_2^2 + \|\hat{Q}^T b\|_2^2 \geq \|\hat{Q}^T b\|_2^2, \end{aligned}$$

等号成立当且仅当  $Rx = Q^T b$ . 所以最小二乘解为


$$x_* = R^{-1} Q^T b.$$

- 由于  $Q$  的列向量构成  $\text{Ran}(A)$  的一组单位正交基, 所以  $QQ^T$  和  $I - QQ^T$  分别是  $\text{Ran}(A)$  和  $\text{Ran}(A)^\perp$  上的正交投影矩阵. 又  $Ax \in \text{Ran}(A)$ , 所以

$$\begin{aligned} \|Ax - b\|_2^2 &= \|QQ^T(Ax - b)\|_2^2 + \|(I - QQ^T)(Ax - b)\|_2^2 \\ &= \|QRx - QQ^T b\|_2^2 + \|(I - QQ^T)b\|_2^2 \\ &= \|Rx - Q^T b\|_2^2 + \|(I - QQ^T)b\|_2^2 \\ &\geq \|(I - QQ^T)b\|_2^2, \end{aligned}$$

等号成立当且仅当  $Rx = Q^T b$ . 所以最小二乘解为

$$x_* = R^{-1} Q^T b.$$

 事实上, 由图 3.3 可知,  $x_*$  满足


$$Ax_* = QQ^T b,$$


即  $QRx_* = QQ^T b$ , 由此可得  $x_* = R^{-1} Q^T b$ .



- 解正规方程. 由定理 3.23 可知, 最小二乘解为

$$x_* = (A^T A)^{-1} A^T b = (R^T Q^T Q R)^{-1} R^T Q^T b = (R^T R)^{-1} R^T Q^T b = R^{-1} Q^T b.$$

 用 QR 分解来求最小二乘解的运算量大约为  $2mn^2$  (如果采用 Householder 变换的话, 运算量大约为  $2mn^2 - \frac{2}{3}n^3$ ). 当  $m \gg n$  时, 大约为正规方程的两倍. 当  $m = n$  时, 几乎相同.

 通常 QR 算法比较稳定, 是求解最小二乘问题的首选方法, 特别是当  $A$  条件数较大 (病态) 时.

### 3.5.3 奇异值分解法


设  $A \in \mathbb{R}^{m \times n}$  列满秩,  $A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T$  是  $A$  的奇异值分解. 令  $U_n$  为  $U$  的前  $n$  列组成的矩阵, 即  $U = [U_n, \tilde{U}]$ , 则

$$\begin{aligned} \|Ax - b\|_2^2 &= \left\| U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T x - b \right\|_2^2 = \left\| \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T x - [U_n, \tilde{U}]^T b \right\|_2^2 \\ &= \left\| \begin{bmatrix} \Sigma V^T x - U_n^T b \\ -\tilde{U}^T b \end{bmatrix} \right\|_2^2 \\ &= \|\Sigma V^T x - U_n^T b\|_2^2 + \|\tilde{U}^T b\|_2^2 \geq \|\tilde{U}^T b\|_2^2, \end{aligned}$$

等号当且仅当  $\Sigma V^T x - U_n^T b = 0$  时成立, 即

$$x = (\Sigma V^T)^{-1} U_n^T b = V \Sigma^{-1} U_n^T b.$$

这就是线性最小二乘问题 (3.1) 的解.


 相比于正规方程和 QR 分解, 用 SVD 求解最小二乘问题具有更高的健壮性, 但由于需要计算系数矩阵的 SVD, 运算量远超正规方程和 QR 分解. 所以只有当系数矩阵秩亏或者接近秩亏时才使用 (此时 QR 分解法可能会失效).

**例 3.6** 分别用三种方法求解最小二乘问题, 比较运算时间.

(LS\_3methods.m)

三种方法的运算时间如下 (以秒为单位):

| $n$  | 正规方程法  | QR 分解法 | 奇异值分解法  |
|------|--------|--------|---------|
| 500  | 0.0050 | 0.0220 | 0.0370  |
| 1000 | 0.0160 | 0.0340 | 0.1440  |
| 1500 | 0.0490 | 0.1330 | 0.5530  |
| 2000 | 0.0870 | 0.2070 | 1.4840  |
| 2500 | 0.1910 | 0.4430 | 3.1160  |
| 3000 | 0.2500 | 0.6950 | 5.9600  |
| 3500 | 0.4640 | 1.2130 | 10.0500 |
| 4000 | 0.4940 | 1.5700 | 14.8750 |
| 4500 | 0.6690 | 2.1680 | 20.6410 |
| 5000 | 1.0720 | 2.9350 | 28.6360 |

 这里的计算时间可能受计算机系统和 MATLAB 矩阵运算优化影响, 并不一定能准确反映各种方法的实际运算量.



### 3.6 广义逆与最小二乘

满秩的正方矩阵存在逆, 一个很自然的问题就是, 对于不满秩的矩阵或者非正方矩阵, 是不是可以定义类似的逆?

#### 3.6.1 广义逆

广义逆的概念最早由 Moore [91] 于 1920 年提出, 他给出的定义如下: 设  $A \in \mathbb{C}^{m \times n}$ , 若  $X \in \mathbb{C}^{n \times m}$  满足

$$AX = P_{\text{Ran}(A)}, \quad XA = P_{\text{Ran}(X)}, \quad (3.24)$$

即  $AX$  和  $XA$  分别为  $\text{Ran}(A)$  和  $\text{Ran}(X)$  上的正交投影算子, 则称  $X$  是  $A$  的广义逆.

1955 年, Penrose [100] 利用下面四个矩阵方程给出了广义逆的另一个定义, 这也是当前常用的广义逆定义方式.

**定义 3.2** 设  $A \in \mathbb{C}^{m \times n}$ , 若  $X \in \mathbb{C}^{n \times m}$  满足

$$AXA = A \quad (3.25)$$


$$XAX = X \quad (3.26)$$

$$(AX)^* = AX \quad (3.27)$$

$$(XA)^* = XA. \quad (3.28)$$

则称  $X$  为  $A$  的**广义逆**, 记为  $A^\dagger$ .

方程组 (3.24) 和 (3.25)-(3.28) 分别称为 Moore 方程组和 Penrose 方程组. 可以证明, 以上两种定义是等价的, 因此广义逆也称为 **Moore-Penrose 逆**, 简称 **MP 逆**.

-  (1) 需要指出的是, 广义逆对所有矩阵都有定义, 并不要求是正方矩阵.
- (2) 若  $A \in \mathbb{C}^{n \times n}$  非奇异, 则  $A^\dagger = A^{-1}$ .
- (3) 在有的文献中, 广义逆也称为**伪逆**.

**定理 3.25** [142] 设  $A \in \mathbb{C}^{m \times n}$ , 则满足矩阵方程组 (3.25)-(3.28) 的矩阵  $X \in \mathbb{C}^{n \times m}$  存在且唯一.

(板书)

**证明. 存在性:** 可以通过 SVD 构造.

设  $\text{rank}(A) = r > 0$ , 且  $A$  的 SVD 为

$$A = U \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} V^*, \quad \Sigma_1 = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r) \in \mathbb{R}^{r \times r}.$$

令

$$X = V \begin{bmatrix} \Sigma_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^*.$$

容易验证,  $X$  满足矩阵方程 (3.25)-(3.28).

**唯一性:** 假设存在  $X$  和  $Y$  都满足矩阵方程 (3.25)-(3.28). 则

$$X = XAX = X(AX)^* = XX^*A^* = XX^*(AY)^* = X(AX)^*(AY)^* = (XAX)(AY) = XAY.$$

另一方面,


$$Y = YAY = (YA)^*Y = A^*Y^*Y = (AXA)^*Y^*Y = (XA)^*(YA)^*Y = XAYAY = XAY.$$

所以  $Y = X$ . □

### 3.6.2 广义逆基本性质

**定理 3.26** 设  $A \in \mathbb{C}^{m \times n}$ , 则

- (1)  $(A^\dagger)^\dagger = A$ ;
- (2)  $(A^T)^\dagger = (A^\dagger)^T$ ,  $(A^*)^\dagger = (A^\dagger)^*$ ;
- (3)  $\text{rank}(A) = \text{rank}(A^\dagger) = \text{rank}(A^\dagger A)$ ;
- (4)  $(AA^*)^\dagger = (A^*)^\dagger A^\dagger$ ,  $(A^*A)^\dagger = A^\dagger (A^*)^\dagger$ ;
- (5)  $(AA^*)^\dagger AA^* = AA^\dagger$ ,  $(A^*A)^\dagger A^*A = A^\dagger A$ ;
- (6)  $A^\dagger = (A^*A)^\dagger A^* = A^*(AA^*)^\dagger$ ,  
特别地, 若  $A$  列满秩, 则  $A^\dagger = (A^*A)^{-1}A^*$ , 若  $A$  行满秩, 则  $A^\dagger = A^*(AA^*)^{-1}$ ;
- (7) 若  $U, V$  是酉矩阵, 则  $(UAV)^\dagger = V^*A^\dagger U^*$ .

 一般来说, 当  $A, B$  是方阵时,

- $(AB)^\dagger \neq B^\dagger A^\dagger$ ;
- $AA^\dagger \neq A^\dagger A$ ;
- $(A^k)^\dagger \neq (A^\dagger)^k$ ;
- $A$  和  $A^\dagger$  的非零特征值并不是互为倒数.

**定理 3.27** 设  $A \in \mathbb{C}^{m \times n}$ , 则

$$\begin{aligned} \text{Ran}(AA^\dagger) &= \text{Ran}(AA^*) = \text{Ran}(A), \\ \text{Ran}(A^\dagger A) &= \text{Ran}(A^*A) = \text{Ran}(A^*) = \text{Ran}(A^\dagger), \\ \text{Ker}(AA^\dagger) &= \text{Ker}(AA^*) = \text{Ker}(A^*) = \text{Ker}(A^\dagger), \\ \text{Ker}(A^\dagger A) &= \text{Ker}(A^*A) = \text{Ker}(A). \end{aligned}$$

**推论 3.28 (广义逆与正交投影)** 设  $A \in \mathbb{C}^{m \times n}$ , 则

$$P_A = AA^\dagger, \quad P_{A^T} = A^\dagger A,$$

其中  $P_A$  和  $P_{A^T}$  分别表示  $\text{Ran}(A)$  和  $\text{Ran}(A^T)$  上的正交投影变换.



特别地, 如果  $A \in \mathbb{C}^{n \times n}$  是正交投影, 则

$$A^\dagger = A.$$

### 3.6.3 广义逆的计算

#### 利用 SVD

我们可以利用  $A$  的奇异值分解来计算  $A^\dagger$ , 但运算量较大. 因为奇异值分解通常与  $AA^*$  或  $A^*A$  的特征值分解有关.

#### 利用满秩分解

**定理 3.29** 设  $A \in \mathbb{R}^{m \times n}$ .

- (1) 若  $A$  是列满秩矩阵, 则  $A^\dagger = (A^*A)^{-1}A^*$ ;
- (2) 若  $A$  是行满秩矩阵, 则  $A^\dagger = A^*(AA^*)^{-1}$ ;
- (3) 若  $A$  的秩是  $r \leq \min\{m, n\}$ , 且其满秩分解为  $A = FG$ , 其中  $F \in \mathbb{R}^{m \times r}$ ,  $G \in \mathbb{R}^{r \times n}$ , 则

$$A^\dagger = G^\dagger F^\dagger = G^*(GG^*)^{-1}(F^*F)^{-1}F^*.$$

#### 利用 QR 分解

**定理 3.30** 设  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ) 是列满秩矩阵, 其 QR 分解为  $A = QR$ , 其中  $Q \in \mathbb{R}^{m \times n}$ ,  $R \in \mathbb{R}^{n \times n}$ , 则

$$A^\dagger = R^{-1}Q^*.$$


#### 其他算法

其他比较重要的算法有: [Greville 递推算法](#), [Cline 算法](#)等.

### 3.6.4 广义逆与线性最小二乘

**定理 3.31** [142, page 85] 设  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ), 则线性最小二乘问题 (3.1) 的解为

$$x = A^\dagger b + (I - P_{\text{Ran}(A^\top)})z, \quad \forall z \in \mathbb{R}^n. \quad (3.29)$$

 通常, 线性最小二乘问题的解 (3.29) 不唯一. 但当  $A$  列满秩时,  $P_{\text{Ran}(A^\top)} = I$ , 此时解唯一.

**定理 3.32** [142, page 85] 设  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ). 记线性最小二乘问题 (3.1) 的解集为  $\mathcal{S}$ , 则

$$\min_{x \in \mathcal{S}} \|x\|_2 \quad (3.30)$$

存在唯一解, 即线性最小二乘问题 (3.1) 存在唯一的最小范数解.



### 3.6.5 左逆和右逆

在工程计算中,有时也会用到矩阵的左逆和右逆.

**定义 3.3** 设  $A \in \mathbb{C}^{m \times n}$ , 如果存在矩阵  $A_{\text{left}}^{-1} \in \mathbb{C}^{n \times m}$  使得  $A_{\text{left}}^{-1}A = I_n$ , 则称  $A_{\text{left}}^{-1}$  是  $A$  的左逆. 类似地, 如果存在矩阵  $A_{\text{right}}^{-1} \in \mathbb{C}^{n \times m}$  使得  $AA_{\text{right}}^{-1} = I_m$ , 则称  $A_{\text{right}}^{-1}$  是  $A$  的右逆.

**定理 3.33** 设  $A \in \mathbb{C}^{m \times n}$ , 则  $A$  存在左逆的充要条件是  $A$  列满秩;  $A$  存在右逆的充要条件是  $A$  行满秩.

易知, 左逆和右逆一般是不唯一的. 比如

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \text{则} \quad A_{\text{left}}^{-1} = \begin{bmatrix} 1 & 0 & x \\ 0 & 1 & y \end{bmatrix}, \quad \text{其中 } x, y \text{ 可以取任意值.}$$





## 3.7 最小二乘扰动分析

**定理 3.34** 设  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ) 且  $\text{rank}(A) = n$ . 设  $x$  是线性最小二乘问题 (3.1) 的解,  $\tilde{x}$  极小化  $\|(A + \delta A)\tilde{x} - (b + \delta b)\|_2$ , 则

$$\frac{\|\tilde{x} - x\|_2}{\|x\|_2} \leq \varepsilon \cdot \left\{ \frac{2\kappa_2(A)}{\cos \theta} + \kappa_2^2(A) \tan \theta \right\} + \mathcal{O}(\varepsilon^2),$$

其中  $\kappa_2(A) = \sigma_1(A)/\sigma_n(A)$ ,  $\theta$  为  $b$  与  $\text{Ran}(A)$  的夹角,

$$\varepsilon \triangleq \max \left\{ \frac{\|\delta A\|_2}{\|A\|_2}, \frac{\|\delta b\|_2}{\|b\|_2} \right\},$$

并假定  $\varepsilon \cdot \kappa_2(A) < 1$  (确保  $A + \delta A$  满秩, 从而  $\tilde{x}$  唯一确定).

我们记

$$\kappa_{LS} \triangleq \frac{2\kappa_2(A)}{\cos \theta} + \kappa_2^2(A) \tan \theta,$$

这就是最小二乘问题的条件数. 当  $\theta = 0$  时,  $b \in \text{Ran}(A)$ , 此时  $\kappa_{LS} = 2\kappa_2(A)$ ; 当  $\theta = \pi/2$  时,  $b \perp \text{Ran}(A)$ , 此时最小二乘解为  $x = 0$ , 而  $\kappa_{LS} = \infty$ ; 当  $0 < \theta < \pi/2$  时,  $\kappa_{LS} = \mathcal{O}(\kappa_2^2(A))$ .

定义残量  $r = b - Ax$ ,  $\tilde{r} = (b + \delta b) - (A + \delta A)\tilde{x}$ , 我们有下面的性质 [68]

$$\frac{\|\tilde{r} - r\|_2}{\|r\|_2} \leq \varepsilon \cdot (1 + 2\kappa_2(A)).$$

当我们使用 QR 分解或 SVD 分解求解最小二乘问题时, 由于采用的是正交变换, 它们都是数值稳定的. 而正规方程涉及求解方程组  $A^T A x = A^T b$ , 其精度依赖于条件数  $\kappa_2(A^T A) = \kappa_2^2(A)$ , 因为其误差是以  $\kappa_2^2(A)$  倍数增长. 因此当  $A$  的条件数较大时, 正规方程法的精度会大大降低.



### 3.8 推广与应用

#### 3.8.1 最小二乘问题的推广

##### 正则化

在求解超定线性方程组时, 我们极小化  $\|Ax - b\|_2^2$ . 而对于欠定线性方程组, 由于解不唯一, 我们往往还需要极小化  $\|x\|_2^2$ . 两者合起来就是

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \frac{\alpha}{2} \|x\|_2^2, \quad (3.31)$$

其中  $\alpha > 0$  是正则化参数. 对应的目标函数记为

$$J(x) = \|Ax - b\|_2^2 + \alpha \|x\|_2^2.$$

当  $\alpha > 0$  时,  $J(x)$  是一个严格凸的二次函数, 因此存在唯一的最小值点. 令  $J(x)$  关于  $x$  的一阶导数为零, 则可得

$$(A^T A + \alpha I)x = A^T b.$$

由于  $A^T A + \alpha I$  对称正定, 故非奇异. 所以问题 (3.31) 的唯一解为

$$x = (A^T A + \alpha I)^{-1} A^T b.$$

##### 加权正则化

一类应用更广泛的问题是下面的加权正则化问题

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \frac{\alpha}{2} \|Wx\|_2^2, \quad (3.32)$$

其中  $\alpha > 0$  是正则化参数,  $W \in \mathbb{R}^{m \times n}$  是广义加权矩阵, 可以是非负对角矩阵, 对称正定矩阵, 也可以是一般矩阵 (如一阶差分算子, 二阶差分算子等). 需要指出的是,  $W$  不一定要是方阵.

经过类似的推导, 可知问题 (3.32) 的解满足

$$(A^T A + \alpha W^T W)x = A^T b.$$

如果  $A^T A + \alpha W^T W$  非奇异, 则存在唯一解

$$x = (A^T A + \alpha W^T W)^{-1} A^T b.$$

##### 约束最小二乘问题

考虑带有约束的最小二乘问题

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} \|Ax - b\|_2^2 \\ \text{s.t.} \quad & Bx = f \end{aligned} \quad (3.33)$$

其中  $Bx = f$  是约束条件,  $B \in \mathbb{R}^{p \times n}$  和  $f \in \mathbb{R}^p$  都是给定的. 对应的 Lagrange 函数为

$$J(x) = \frac{1}{2} \|Ax - b\|_2^2 + \lambda^T (Bx - f).$$



分别对  $x$  和  $\lambda$  求一阶导数, 并令其等于零, 可得

$$\begin{bmatrix} A^T A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} A^T b \\ f \end{bmatrix}.$$

如果  $A^T A$  非奇异, 且  $B$  行满秩, 则存在唯一解

$$\begin{aligned} \lambda &= [B(A^T A)^{-1} B^T]^{-1} [B(A^T A)^{-1} A^T b - f] \\ x &= (A^T A)^{-1} (A^T b - B^T \lambda). \end{aligned}$$

### 3.8.2 最小二乘问题的应用

#### 多项式数据拟合

最小二乘的一个重要应用就是低次多项式数据拟合: 已知平面上  $n$  个点  $\{(t_i, f_i)\}_{i=1}^n$ , 寻找一个低次多项式来拟合这些数据.

设拟合多项式为

$$p(x) = a_0 + a_1 t + a_2 t^2 + \cdots + a_m t^m.$$

通常  $m \ll n$ . 将上述  $n$  个点  $\{(t_i, f_i)\}_{i=1}^n$  代入可得

$$\begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^m \\ 1 & t_2 & t_2^2 & \cdots & t_2^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^m \end{bmatrix}_{n \times (m+1)} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} \quad \text{或} \quad Ax = f,$$

其中  $A \in \mathbb{R}^{n \times (m+1)}$ ,  $x = [a_0, a_1, \dots, a_m]^T$ . 由于  $m \ll n$ , 该方程组是超定的, 解通常是不存在的. 因此, 我们寻找一个近似解, 使得残量  $\|f - Ax\|_2$  最小, 即求解最小二乘问题

$$\min_{x \in \mathbb{R}^{m+1}} \|f - Ax\|_2^2$$

#### 线性预测

预测一个时间序列的未来走向, 其中一个常用方法就是线性预测 (linear prediction). 假定一个时间序列在  $t_k$  时刻的值  $f_k$  线性依赖于其前  $m$  个时刻的值  $f_{k-1}, f_{k-2}, \dots, f_{k-m}$ , 即

$$f_k = a_1 f_{k-1} + a_2 f_{k-2} + \cdots + a_m f_{k-m}. \quad (3.34)$$

现在已经测得该时间序列的前  $n$  个值  $f_i, i = 0, 1, 2, \dots, n-1$ , 如何预测其未来的取值. 这里  $n \gg m$ . 将现有的数据代入关系式 (3.34) 可得

$$\begin{bmatrix} f_{m-1} & f_{m-2} & f_{m-3} & \cdots & f_0 \\ f_m & f_{m-1} & f_{m-2} & \cdots & f_1 \\ \vdots & \vdots & \vdots & & \vdots \\ f_{n-2} & f_{n-3} & f_{n-4} & \cdots & f_{n-m-1} \end{bmatrix}_{(n-m) \times m} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} f_m \\ f_{m+1} \\ \vdots \\ f_{n-1} \end{bmatrix} \quad \text{或} \quad Ax = f,$$

其中  $A \in \mathbb{R}^{(n-m) \times m}$ ,  $x = [a_1, a_2, \dots, a_m]^T$ . 这也是一个超定问题, 其解也是通过求解下面的



最小二乘问题来获取

$$\min_{x \in \mathbb{R}^m} \|f - Ax\|_2^2$$

### 信号恢复

在获取数字信号时, 由于各种各样的原因, 最后得到的信号总会带有一定的噪声, 即

$$b = x + \eta,$$

其中  $x$  是真实的信号,  $b$  是观察到的信号,  $\eta$  是噪声.

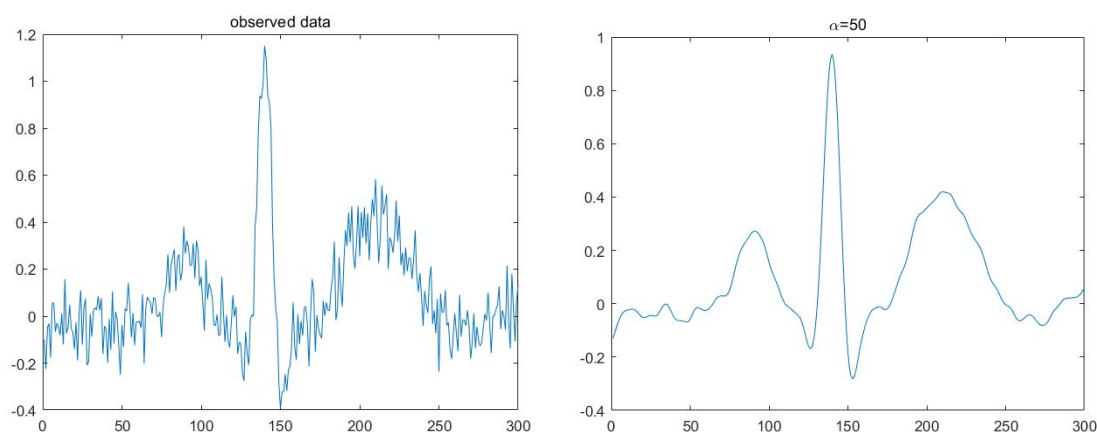
**去噪**是数字信号和图像处理中的一个基本问题, 其中一个有效方法就是加权最小二乘法, 即

$$\min_x \frac{1}{2} \|x - b\|_2^2 + \frac{1}{2} \alpha \|Dx\|_2^2,$$

其中  $D$  是离散的二阶导算子或 TV 算子.

#### 例 3.7 数字信号去噪.

(LS\_denoise.m)



### 图像恢复

除了带噪声以外, 在获取数字图像时也经常会由于各种原因 (设备仪器, 拍摄环境等) 导致图像模糊. 通常数字图像的获取模型为

$$f = \mathcal{B}(x) + \eta,$$

其中  $x$  是真实图像,  $f$  是观察到的图像,  $\mathcal{B}(\cdot)$  是卷积算子 (代表模糊机制),  $\eta$  是噪声.

由于问题本身是不适定的, 因此求解时需要进行正则化, 常用的模型有 **Tikhonov 正则化**模型:

$$\min \| \mathcal{B}(x) - f \|_2^2 + \mu^2 \|x\|_2^2$$

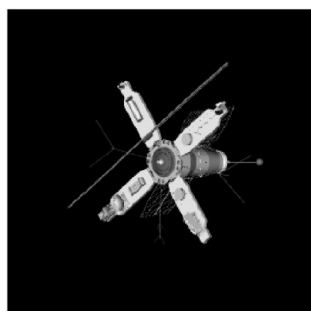


或**加权正则化**模型:

$$\min \| \mathcal{B}(x) - f \|_2^2 + \mu^2 \| Dx \|_2^2,$$

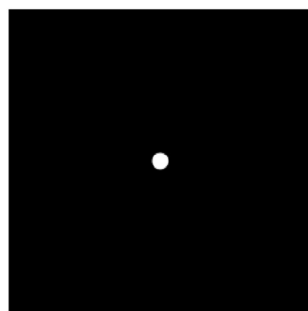
其中  $D$  是广义加权矩阵, 可以是非负对角矩阵或 TV 算子等.

**例 3.8** 数字图像去噪和去模糊.



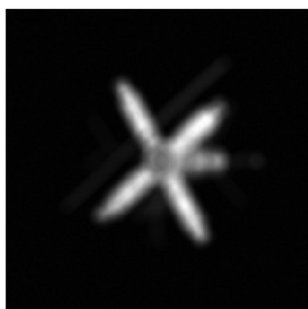
(a)

(a) 真实图像



(b)

(b) 模糊机制 (out-of-focus)



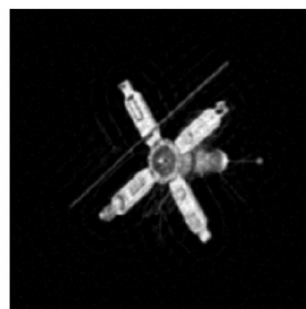
(a)

(a) 获取到的图像 (带有噪声和模糊)



(b)

(b) 恢复后的图像 (基于 Tikhonov 正则化模型)



(c)

(c) 恢复后的图像 (基于加权正则化模型)

## 3.9 课后习题

练习 3.1 设  $A \in \mathbb{R}^{m \times n}$ , 其中  $m \leq n$ . 试证明: 矩阵  $\begin{bmatrix} I & A^\top \\ A & 0 \end{bmatrix}$  非奇异的充要条件是  $\text{rank}(A) = m$ .

练习 3.2 设  $B \in \mathbb{R}^{m \times m}$  对称半正定,  $A \in \mathbb{R}^{m \times n}$ , 其中  $m \geq n$ . 试证明: 矩阵  $\begin{bmatrix} B & A \\ A^\top & 0 \end{bmatrix}$  非奇异的充要条件是  $A$  列满秩且矩阵  $[B, A]$  行满秩 (即  $A$  列满秩且  $\text{Ker}(B) \cap \text{Ker}(A^\top) = \{0\}$ ).

练习 3.3 设  $\tau \neq 0$ , 向量  $u, v \in \mathbb{R}^n$  均不为零, 求矩阵  $E(u, v, \tau) = I - \tau uv^\top$  的特征值.

练习 3.4\* 设  $A \in \mathbb{C}^{m \times n}$ ,  $B \in \mathbb{C}^{n \times m}$ , 证明:  $AB$  与  $BA$  具有相同的非零特征值.

练习 3.5 设  $A \in \mathbb{R}^{n \times n}$  是正交矩阵, 则  $A$  可表示成至多  $n$  个 Householder 变换的乘积.

练习 3.6 设  $G = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \in \mathbb{R}^{2 \times 2}$  是 Givens 变换, 由练习 3.5 可知,  $G$  可以表示为 2 个 Householder 变换的乘积, 试给出这两个 Householder 变换所对应的 Householder 向量.

练习 3.7 设  $x = [x_1, x_2, \dots, x_n]^\top \in \mathbb{R}^n$  是一个非零向量,  $H$  是 Householder 矩阵, 满足  $Hx = \alpha e_1$ . 证明:  $H$  的第一列与  $x$  平行.

(注: 该结论提供了一种构造指定第一列的正交矩阵或酉矩阵的实用方法)

练习 3.8 设  $H_k \in \mathbb{R}^{k \times k}$  是 Householder 变换, 其中  $k < n$ .

证明:  $H_n = \begin{bmatrix} I_{n-k} & 0 \\ 0 & H_k \end{bmatrix}$  是  $n$  阶 Householder 变换.

练习 3.9 设  $R = \begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix} \in \mathbb{R}^{2 \times 2}$ , 求一个 Givens 变换  $G$ , 使得  $G^\top R G = \begin{bmatrix} r_{22} & r_{12} \\ 0 & r_{11} \end{bmatrix}$ .

练习 3.10 设  $R \in \mathbb{R}^{n \times n}$  是一个上三角矩阵, 且对角线元素互不相同. 证明: 存在正交矩阵  $Q$ , 使得  $Q^\top R Q$  为上三角矩阵, 且对角线元素为  $R$  的对角线元素的降序排列.

练习 3.11\* 设矩阵  $A$  的奇异值分解为  $A = U_n \Sigma V^* = \sum_{i=1}^n \sigma_i u_i v_i^* \in \mathbb{C}^{m \times n}$ , 证明

$$\min_{B \in \mathbb{C}^{m \times n}, \text{rank}(B)=k} \|A - B\|_F = \|A - A_k\|_F = \sqrt{\sigma_{k+1}^2 + \sigma_{k+2}^2 + \dots + \sigma_n^2}.$$

其中  $A_k = \sum_{i=1}^k \sigma_i u_i v_i^*$ . (提示: 利用定理 5.28, 并构造矩阵  $\begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}$  和  $\begin{bmatrix} 0 & B \\ B^* & 0 \end{bmatrix}$ )

练习 3.12\* 设  $A \in \mathbb{R}^{n \times n}$ , 证明  $\|A\|_F^2 \geq \sum_{i=1}^n |\lambda_i(A)|^2$ .

(提示: 利用  $F$ -范数与迹的关系, 以及 Schur 分解)

练习 3.13 设  $A \in \mathbb{C}^{n \times n}$  有  $n$  个互不相同的非零特征值, 其 Schur 分解为  $A = URU^*$ , 其中  $U$  为酉矩阵,  $R$  为上三角矩阵. 设矩阵  $B \in \mathbb{C}^{n \times n}$  满足  $AB = BA$ . 证明:  $U^* B U$  是上三角矩阵.

练习 3.14 用 Householder 变换计算矩阵  $A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & -1 & -1 \\ 2 & -4 & 10 \end{bmatrix}$  的 QR 分解.



练习 3.15 求解最小二乘问题:  $\min \|Ax - b\|_2$ , 其中  $A = \begin{bmatrix} 2 & 0 \\ -2 & 3 \\ 0 & -1 \\ 1 & -2 \end{bmatrix}$ ,  $b = \begin{bmatrix} 2 \\ 0 \\ 2 \\ 2 \end{bmatrix}$ .

练习 3.16\* 设  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ) 且  $\text{rank}(A) = n$ . 计算矩阵  $\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix}$  的谱条件数. (用  $A$  的奇异值表示)

..... 以下为可选题 .....

练习 3.17 设  $A \in \mathbb{R}^{m \times n}$ , 证明:  $\text{Ran}(A^T) = \text{Ran}(A^T A)$ .

练习 3.18 试证明三类初等矩阵都可以写成 (3.3) 中的形式. (以第 89 页上的  $E_1, E_2, E_3$  为例)

练习 3.19 证明定理 3.1 中的结论 (3), 即:

对任意非零向量  $x, y \in \mathbb{C}^n$ , 存在  $u, v \in \mathbb{C}^n$  和  $\tau \in \mathbb{C}$ , 使得

$$E(u, v, \tau)x = y.$$

练习 3.20 设  $A \in \mathbb{R}^{n \times n}$  是正交矩阵, 则  $A$  可表示成至多  $\frac{1}{2}n(n-1)$  个 Givens 变换和 1 个对角线为  $\pm 1$  的对角矩阵的乘积. (更进一步, 该对角矩阵除了最后一个对角线元素是  $\pm 1$  外, 其他对角线元素都是 1)

练习 3.21\* 试给出定理 3.5 在复数空间上的相对应的结论, 并证明.

练习 3.22 证明定理 3.13 和定理 3.14. (奇异值的相关性质)

练习 3.23\* 证明奇异值的最小最大定理, 即定理 3.17.

练习 3.24 证明奇异值的扰动性质, 即定理 3.21.

练习 3.25 试给出加权正则化问题 (3.32) 存在唯一解的充要条件.

练习 3.26\* 设  $A \in \mathbb{R}^{n \times n}$  是一个对角加边矩阵

$$A = \begin{bmatrix} a_1 & b_2 & b_3 & \cdots & b_n \\ c_2 & a_2 & & & \\ c_3 & & a_3 & & \\ \vdots & & & \ddots & \\ c_n & & & & a_n \end{bmatrix}.$$

试给出用 Givens 变换计算  $A$  的 QR 分解的详细算法, 使得运算量为  $\mathcal{O}(n^2)$ .

练习 3.27\* 设  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ), 如何求解下面的数据拟合问题

$$\min_{x \in \mathbb{R}^n} \|b - Ax\|_1.$$

练习 3.28\* 设  $x = [x_1, x_2, \dots, x_n]^T \in \mathbb{C}^n$  是一个非零复向量, 问: 是否存在 Householder 矩阵  $H(v)$  使得  $H(v)x = \alpha e_1$ ? 若存在, 如何计算  $v$ ? (提示: 这里  $\alpha$  可以是复数)



练习 3.29\* 如果矩阵在内存中是按行排列的, 怎么实现 QR 分解比较高效?

练习 3.30\* 设  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ), 如果  $A$  不是满秩的, 如何求解相应的最小二乘问题?

练习 3.31 设  $x, y \in \mathbb{R}^n$  且  $y^\top x = 0$  (即  $x, y$  正交). 证明

$$(x + y)^\dagger = x^\dagger + y^\dagger.$$

推广到矩阵情形: 设  $X, Y \in \mathbb{R}^{m \times n}$  且  $Y^\top X = 0$  (即  $X$  与  $Y$  正交). 证明

$$(X + Y)^\dagger = X^\dagger + Y^\dagger.$$

进一步推广到多个情形: 设  $X_i \in \mathbb{R}^{m \times n}$  且  $X_i^\top X_j = 0, i \neq j$  (即  $X_i$  与  $X_j$  相互正交). 证明

$$(X_1 + X_2 + \cdots + X_k)^\dagger = X_1^\dagger + X_2^\dagger + \cdots + X_k^\dagger.$$

### ..... 以下为实践题 .....

练习 3.32 设  $A = [a_1, a_2, \dots, a_n] \in \mathbb{R}^{m \times n}$ , 下面是另外一种修正 Gram-Schmidt 正交化过程 [56], 试指出与 Gram-Schmidt 正交化过程的区别.

```

1: for $i = 1$ to n do
2: $r_{ii} = \|a_i\|_2$
3: $q_i = a_i / r_{ii}$
4: for $j = i + 1$ to n do
5: $r_{ij} = q_i^\top a_j$
6: $a_j = a_j - r_{ij} q_i$
7: end for
8: end for

```

练习 3.33 编写 Householder 变换函数: 对任意给定的变量  $x \in \mathbb{R}^n$ , 输出  $v$  使得  $H(v)x = \|x\|_2 e_1$ , 其中  $H(v) = I - 2vv^\top$ , 函数原型: `v=House2(x)`

练习 3.34 编写 Householder 变换函数: 对任意给定的变量  $x \in \mathbb{R}^n$ , 输出  $v$  使得  $H(v)x = \|x\|_2 e_1$ , 其中  $H(v) = I - \beta vv^\top$  且  $v$  的第一个分量为 1. 函数原型: `[beta,v]=House3(x)`

练习 3.35 编写基于 Householder 变换的 QR 分解, 函数原型: `[Q,R]=QR_Householder(A)`

练习 3.36 设  $A = \begin{bmatrix} R \\ S \end{bmatrix}$ , 其中  $R \in \mathbb{R}^{n \times n}$  是上三角矩阵,  $S \in \mathbb{R}^{m \times n}$  是稠密矩阵, 试描述  $A$  的基于 Householder 变换的 QR 算法. 要求算法过程中始终保持  $R$  中的零元.

练习 3.37 设  $A = R + uv^\top$ , 其中  $R \in \mathbb{R}^{n \times n}$  是上三角矩阵,  $u, v \in \mathbb{R}^n$  为非零列向量, 试给出计算  $A$  的 QR 分解的有效算法. (提示: 使用 Givens 变换, 算法总运算量大约为  $\mathcal{O}(n^2)$ )

练习 3.38 构造并实现列主元 QR 分解的算法.

练习 3.39 构造并实现矩阵的 LQ 分解算法, 即  $A = LQ$ , 其中  $L$  为下三角矩阵,  $Q$  为正交矩阵.





## 第四讲 非对称特征值问题

设  $A$  是一个非对称的稠密矩阵, 本讲主要讨论如何计算  $A$  的全部特征值和特征向量. 为了讨论方便, 本讲同样只讨论实矩阵情形.

记  $A \in \mathbb{R}^{n \times n}$  的特征值为  $\lambda_1, \lambda_2, \dots, \lambda_n$ . 本讲中我们总是假定

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n| \geq 0,$$

即  $A$  的特征值按模降序排列.

### 关于稠密矩阵特征值计算的相关参考资料

- J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, 1965. [128] (有中文翻译)
- B. N. Parlett, *The Symmetric Eigenvalue Problem*, 2nd Eds., 1998. [96]
- G. W. Stewart, *Matrix Algorithms, Vol II: Eigensystems*, 2001. [114]
- G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2013. [56]
- Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, 2000. [7]
- ▷ G. H. Golub and H. A. van der Vorst, [Eigenvalue computation in the 20th century](#), 2000. [55]

## 4.1 幂迭代法

### 4.1.1 算法介绍

**幂迭代法**是计算特征值和特征向量的一种简单易用的算法. 幂迭代法虽然简单, 但它却建立了计算特征值和特征向量的算法的一个基本框架.

#### 算法 4.1. 幂迭代法 (Power Iteration)

```

1: Choose an initial guess $x^{(0)}$ with $\|x^{(0)}\|_2 = 1$
2: set $k = 0$
3: while not convergence do
4: $y^{(k+1)} = Ax^{(k)}$
5: $x^{(k+1)} = y^{(k+1)} / \|y^{(k+1)}\|_2$ % 单位化, 防止溢出
6: $\mu_{k+1} = (Ax^{(k+1)}, x^{(k+1)})$ % 计算内积
7: $k = k + 1$
8: end while

```

### 4.1.2 收敛性分析

下面讨论幂迭代的收敛性. 假设

- (1)  $A \in \mathbb{R}^{n \times n}$  是可对角化的, 即  $A = V\Lambda V^{-1}$ , 其中  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \in \mathbb{C}^{n \times n}$ ,  $V = [v_1, v_2, \dots, v_n] \in \mathbb{C}^{n \times n}$ , 且  $\|v_i\|_2 = 1, i = 1, 2, \dots, n$ .
- (2) 同时, 我们还假设  $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|$ .

由于  $V \in \mathbb{C}^{n \times n}$  非奇异, 所以它的列向量组构成  $\mathbb{C}^n$  的一组基. 因此迭代初始向量  $x^{(0)}$  可表示为

$$x^{(0)} = \alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_n v_n = V[\alpha_1, \alpha_2, \dots, \alpha_n]^T.$$

我们假定  $\alpha_1 \neq 0$ , 即  $x^{(0)}$  不属于  $\text{span}\{v_2, v_3, \dots, v_n\}$  (由于  $x^{(0)}$  是随机选取的, 从概率意义上讲, 这个假设通常是成立的). 于是我们可得

$$A^k x^{(0)} = (V\Lambda V^{-1})^k V \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = V\Lambda^k \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = V \begin{bmatrix} \alpha_1 \lambda_1^k \\ \alpha_2 \lambda_2^k \\ \vdots \\ \alpha_n \lambda_n^k \end{bmatrix} = \alpha_1 \lambda_1^k V \begin{bmatrix} 1 \\ \frac{\alpha_2}{\alpha_1} \left(\frac{\lambda_2}{\lambda_1}\right)^k \\ \vdots \\ \frac{\alpha_n}{\alpha_1} \left(\frac{\lambda_n}{\lambda_1}\right)^k \end{bmatrix}.$$

又  $|\lambda_i/\lambda_1| < 1, i = 2, 3, \dots, n$ , 所以

$$\lim_{k \rightarrow \infty} \left(\frac{\lambda_i}{\lambda_1}\right)^k = 0, \quad i = 2, 3, \dots, n.$$

故当  $k$  趋向于无穷大时, 向量


$$\left[1, \frac{\alpha_2}{\alpha_1} \left(\frac{\lambda_2}{\lambda_1}\right)^k, \dots, \frac{\alpha_n}{\alpha_1} \left(\frac{\lambda_n}{\lambda_1}\right)^k\right]^T, \quad k = 0, 1, 2, \dots$$



收敛到  $e_1 = [1, 0, \dots, 0]^T$ . 所以向量  $x^{(k)} = A^k x^{(0)} / \|A^k x^{(0)}\|_2$  收敛到  $\pm v_1$ , 即  $A$  的对应于 (模) 最大的特征值  $\lambda_1$  的特征向量. 而  $\mu_k = (Ax^{(k)}, x^{(k)})$  则收敛到  $v_1^* A v_1 = \lambda_1$ .

显然, 幂迭代的收敛快慢取决于  $|\lambda_2/\lambda_1|$  的大小,  $|\lambda_2/\lambda_1|$  越小, 收敛越快.

通过上面的分析可知, 幂迭代只能用于计算矩阵的模最大的特征值和其相应的特征向量. 如果  $A$  的模最大的特征值是唯一的, 则称其为**主特征值**. 当  $|\lambda_2/\lambda_1|$  接近于 1 时, 收敛速度会非常慢. 同时, 如果  $A$  的模最大特征值不唯一, 比如一对共轭复特征值, 则幂迭代就可能就会失效.

 如果需要计算其他特征值, 比如模第二大特征值  $\lambda_2$ , 则可以在模最大特征值  $\lambda_1$  计算出来后, 采用**收缩 (Deflation)**技术: 构造酉矩阵  $U$ , 使得

$$U^* A U = \begin{bmatrix} \lambda_1 & A_{12} \\ 0 & A_{22} \end{bmatrix}.$$

然后将幂迭代作用到  $A_{22}$  上, 就可以求出  $\lambda_2$ . 以此类推, 可以依次求出所有特征值 (这里假定特征值互不相同).

 **思考:** 上面的收缩技术中的  $U$  怎么选取?

### 4.1.3 位移策略

前面已经提到, 幂迭代法的收敛速度取决于  $|\lambda_2/\lambda_1|$  的大小. 当它的值接近于 1 时, 收敛速度会非常缓慢. 因此, 为了加快幂迭代法的收敛速度, 我们希望  $|\lambda_2/\lambda_1|$  的值越小越好.


一个简单易用的方法就是使用**位移策略**, 即将计算  $A$  的特征值转化为计算  $A - \sigma I$  的特征值, 即对  $A$  做一个移位. 这里  $\sigma$  是一个给定的数, 称  $\sigma$  为**位移** (shift). 为了使得幂迭代作用到  $A - \sigma I$  时具有更快的收敛速度, 我们要求  $\sigma$  满足下面的两个条件:

- (1)  $\lambda_1 - \sigma$  是  $A - \sigma I$  的模最大的特征值;
- (2)  $\max_{2 \leq i \leq n} \left| \frac{\lambda_i - \sigma}{\lambda_1 - \sigma} \right|$  尽可能地小.

其中第一个条件保证最后所求得特征值是我们所要的, 第二个条件用于加快幂迭代的收敛速度.

显然, 在实际应用中,  $\sigma$  的取值并不是一件容易的事.

**例 4.1** 设  $A = X \Lambda X^{-1}$ , 其中  $\Lambda$  为对角矩阵, 分别用幂迭代法和带位移的幂迭代法计算  $A$  的主特征值. (见 [Eig\\_Power\\_shift.m](#))

 位移策略在特征值计算中非常重要, 特别是在反迭代法和 QR 迭代法中.

## 4.2 反迭代法

### 4.2.1 算法介绍

如果我们将幂迭代法作用在  $A^{-1}$  上, 则可求出  $A$  的模最小的特征值. 事实上, 结合这种思想和位移策略, 我们就可以计算矩阵的任意一个特征值.

#### 算法 4.2. 带位移的反迭代法 (Inverse Iteration)

```

1: Choose a scalar σ and an initial vector $x^{(0)}$ with $\|x^{(0)}\|_2 = 1$
2: set $k = 0$
3: while not convergence do
4: $y^{(k+1)} = (A - \sigma I)^{-1}x^{(k)}$
5: $x^{(k+1)} = y^{(k+1)} / \|y^{(k+1)}\|_2$
6: $\mu_{k+1} = (Ax^{(k+1)}, x^{(k+1)})$
7: $k = k + 1$
8: end while

```

该算法称为**反迭代法**. 显然, 在反迭代法中,  $\mu_k$  收敛到距离  $\sigma$  最近的特征值, 而  $x^{(k)}$  则收敛到其对应的特征向量.

设距离  $\sigma$  最近的特征值为  $\lambda_k$ , 则算法的收敛速度取决于

$$\max_{1 \leq i \leq n} \left| \frac{\lambda_k - \sigma}{\lambda_i - \sigma} \right|$$

的大小. 显然,  $\sigma$  约接近于  $\lambda_k$ , 其值越小, 即算法收敛越快. 若  $\sigma \approx \lambda_k$ , 则迭代几步就可以了.

反迭代法的另一个优点是, 只要选取合适的位移  $\sigma$ , 就可以计算  $A$  的任意一个特征值.

但反迭代法的缺点也很明显: 每步迭代需要解一个线性方程组  $(A - \sigma I)y^{(k+1)} = x^{(k)}$ , 这就需要对  $A - \sigma I$  做一次 LU 分解. 另外, 与幂迭代一样, 反迭代法一次只能求一个特征值.

### 4.2.2 Rayleigh 商迭代

在反迭代法中, 有一个很重要的问题, 就是位移  $\sigma$  的选取.

显然, 选取  $\sigma$  的基本原则是使得其与所求的特征值越靠近越好. 这里需要指出的是, 在每步迭代中, 位移  $\sigma$  可以不一样, 即在迭代过程中可以选取不同的位移  $\sigma$ .

由于在反迭代法 4.2 中,  $\mu_k$  是收敛到所求的特征值的, 所以我们可以选取  $\mu_k$  作为第  $k$  步的位移. 这时所得到的反迭代法就称为 **Rayleigh 商迭代** (Rayleigh Quotient Iteration)<sup>1</sup>, 简记为 RQI.

#### 算法 4.3. Rayleigh 商迭代法 (Rayleigh Quotient Iteration (RQI))

```

1: Choose an initial vector $x^{(0)}$ with $\|x^{(0)}\|_2 = 1$
2: set $k = 0$

```

<sup>1</sup>关于 Rayleigh 商的定义见 (5.13)



```
3: compute $\sigma = (x^{(0)}, Ax^{(0)})$
4: while not converge do
5: $y^{(k+1)} = (A - \sigma I)^{-1}x^{(k)}$
6: $x^{(k+1)} = y^{(k+1)} / \|y^{(k+1)}\|_2$
7: $\mu_{k+1} = (Ax^{(k+1)}, x^{(k+1)})$
8: $\sigma = \mu_{k+1}$
9: $k = k + 1$
10: end while
```

一般来说, 如果 Rayleigh 商迭代收敛到  $A$  的一个单特征值, 则至少是二次收敛的, 即具有局部二次收敛性. 如果  $A$  是对称的, 则能达到局部三次收敛.

在 Rayleigh 商迭代中, 由于每次迭代的位移是不同的, 因此每次迭代需要求解一个不同的线性方程组, 这使得运算量大大增加.

**例 4.2** 设  $A = X\Lambda X^{-1}$ , 其中  $\Lambda$  为对角矩阵, 用 Rayleigh 商迭代计算  $A$  的特征值.

(见 [Fig\\_Rayleigh.m](#))



### 4.3 正交迭代法


幂迭代和反迭代都只能同时计算一个特征对. 如果想同时计算多个特征对, 我们可以采用多个初始向量进行迭代. 而**正交迭代**算法就是基于这种思想, 它能够计算  $A$  的一个不变子空间, 从而可以同时计算出多个特征值.

#### 算法 4.4. 正交迭代法 (Orthogonal Iteration)

```

1: Choose an $n \times p$ column orthogonal matrix Z_0
2: set $k = 0$
3: while not convergence do
4: compute $Y_{k+1} = AZ_k$
5: $Y_{k+1} = Z_{k+1} \hat{R}_{k+1}$ % QR 分解
6: $k = k + 1$
7: end while

```

 正交迭代方法有时也称为**子空间迭代方法** (subspace iteration method) 和**同步迭代方法** (simultaneous iteration method).

在算法中使用 QR 分解是为了保持  $Z_k$  的列正交性, 使得其列向量组构成子空间  $\text{span}\{A^i Z_0\}$  的一组正交基. 一方面提高算法的数值稳定性, 另一方面避免所有列都收敛到最大特征值所对应的特征向量.

下面我们分析该算法的收敛性质. 假设  $A$  是对角化的, 即  $A = V\Lambda V^{-1}$ , 其中  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ , 且  $|\lambda_1| \geq \dots \geq |\lambda_p| > |\lambda_{p+1}| \geq \dots \geq |\lambda_n|$ . 则可得

$$\text{span}\{Z_k\} = \text{span}\{Y_k\} = \text{span}\{AZ_{k-1}\}, \quad k = 1, 2, \dots,$$

由此可知

$$\text{span}\{Z_k\} = \text{span}\{A^k Z_0\} = \text{span}\{V\Lambda^k V^{-1} Z_0\}.$$

我们注意到

$$\Lambda^k V^{-1} Z_0 = \lambda_p^k \begin{bmatrix} (\lambda_1/\lambda_p)^k & & & \\ & \ddots & & \\ & & 1 & \\ & & & \ddots \\ & & & & (\lambda_n/\lambda_p)^k \end{bmatrix} V^{-1} Z_0 \triangleq \lambda_p^k \begin{bmatrix} W_p^{(k)} \\ W_{n-p}^{(k)} \end{bmatrix}.$$

由于当  $i > p$  时有  $|\lambda_i/\lambda_p| < 1$ , 所以当  $k$  趋于无穷大时,  $W_{n-p}^{(k)}$  趋向于 0. 令  $V = [V_p, V_{n-p}]$ , 则

$$V\Lambda^k V^{-1} Z_0 = \lambda_p^k [V_p, V_{n-p}] \begin{bmatrix} W_p^{(k)} \\ W_{n-p}^{(k)} \end{bmatrix} = \lambda_p^k (V_p W_p^{(k)} + V_{n-p} W_{n-p}^{(k)}).$$



所以当  $k \rightarrow \infty$  时, 有

$$\begin{aligned}\operatorname{span}\{Z_k\} &= \operatorname{span}\{V \Lambda^k V^{-1} Z_0\} = \operatorname{span}\{V_p W_p^{(k)} + V_{n-p} W_{n-p}^{(k)}\} \\ &\rightarrow \operatorname{span}\{V_p W_p^{(k)}\} = \operatorname{span}\{V_p\},\end{aligned}$$

即  $\operatorname{span}\{Z_k\}$  趋向于  $A$  的一个  $p$  维不变子空间  $\operatorname{span}\{V_p\}$ .

**定理 4.1** 给定正整数  $p$  ( $1 \leq p \leq n$ ), 考虑算法 4.4. 假设  $A$  是可对角化的, 且  $|\lambda_1| \geq \cdots \geq |\lambda_p| > |\lambda_{p+1}| \geq \cdots \geq |\lambda_n|$ . 则  $\operatorname{span}\{Z_k\}$  收敛到  $A$  的一个  $p$  维不变子空间.

当  $A$  不可对角化时, 利用 Jordan 标准型, 我们可以得到同样的结论, 见 [125, 126].

在正交迭代中, 如果我们取  $Z_0 = I$ , 则可得到一类特殊的正交迭代法. 此时, 在一定条件下, 正交迭代会收敛到  $A$  的 Schur 标准型.



## 4.4 QR 迭代法

**QR 迭代法**的基本思想是通过不断的正交相似变换, 使得  $A$  趋向于一个上三角形式 (或拟上三角形式). 算法形式非常简单, 描述如下:

### 算法 4.5. QR 迭代法 (QR Iteration)

```

1: Set $A_1 = A$ and $k = 1$
2: while not convergence do
3: $A_k = Q_k R_k$ % QR 分解
4: compute $A_{k+1} = R_k Q_k$
5: $k = k + 1$
6: end while

```

在该算法中, 我们有

$$A_{k+1} = R_k Q_k = (Q_k^T Q_k) R_k Q_k = Q_k^T (Q_k R_k) Q_k = Q_k^T A_k Q_k.$$

由这个递推关系可得

$$A_{k+1} = Q_k^T A_k Q_k = Q_k^T Q_{k-1}^T A_{k-1} Q_{k-1} Q_k = \cdots = Q_k^T Q_{k-1}^T \cdots Q_1^T A Q_1 \cdots Q_{k-1} Q_k.$$

记  $\tilde{Q}_k = Q_1 \cdots Q_{k-1} Q_k = [\tilde{q}_1^{(k)}, \tilde{q}_2^{(k)}, \dots, \tilde{q}_n^{(k)}]$ , 则

$$A_{k+1} = \tilde{Q}_k^T A \tilde{Q}_k, \quad (4.1)$$

即  $A_{k+1}$  与  $A$  正交相似.

### 4.4.1 QR 迭代与幂迭代的关系

记  $\tilde{R}_k = R_k R_{k-1} \cdots R_1$ , 则有

$$\tilde{Q}_k \tilde{R}_k = \tilde{Q}_{k-1} (Q_k R_k) \tilde{R}_{k-1} = \tilde{Q}_{k-1} (A_k) \tilde{R}_{k-1} = \tilde{Q}_{k-1} (\tilde{Q}_{k-1}^T A \tilde{Q}_{k-1}) \tilde{R}_{k-1} = A \tilde{Q}_{k-1} \tilde{R}_{k-1},$$

由此递推下去, 即可得

$$\tilde{Q}_k \tilde{R}_k = A^{k-1} \tilde{Q}_1 \tilde{R}_1 = A^{k-1} Q_1 R_1 = A^k. \quad (4.2)$$

故

$$\tilde{Q}_k \tilde{R}_k e_1 = A^k e_1,$$

这说明 QR 迭代与幂迭代有关.

假设  $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$ , 则当  $k$  充分大时,  $A^k e_1$  收敛到  $A$  的模最大特征值  $\lambda_1$  所对应的特征向量, 故  $\tilde{Q}_k$  的第一列  $\tilde{q}_1^{(k)}$  也收敛到  $\lambda_1$  所对应的特征向量. 因此, 当  $k$  充分大时,  $A \tilde{q}_1^{(k)} \rightarrow \lambda_1 \tilde{q}_1^{(k)}$ , 此时由 (4.1) 可知,  $A_{k+1}$  的第一列

$$A_{k+1}(:, 1) = \tilde{Q}_k^T A \tilde{q}_1^{(k)} \rightarrow \lambda_1 \tilde{Q}_k^T \tilde{q}_1^{(k)} = \lambda_1 e_1,$$

即  $A_{k+1}$  的第一列的第一个元素收敛到  $\lambda_1$ , 而其它元素都趋向于 0, 收敛速度取决于  $|\lambda_2/\lambda_1|$  的大小.





## 4.4.2 QR 迭代与反迭代的关系

下面观察  $\tilde{Q}_k$  的最后一列. 由 (4.1) 可知

$$A\tilde{Q}_k = \tilde{Q}_k A_{k+1} = \tilde{Q}_k Q_{k+1} R_{k+1} = \tilde{Q}_{k+1} R_{k+1},$$

所以有

$$\tilde{Q}_{k+1} = A\tilde{Q}_k R_{k+1}^{-1}.$$

由于  $\tilde{Q}_{k+1}$  和  $\tilde{Q}_k$  都是正交矩阵, 上式两边转置后求逆, 可得

$$\tilde{Q}_{k+1} = \left(\tilde{Q}_{k+1}^T\right)^{-1} = \left((R_{k+1}^{-1})^T \tilde{Q}_k^T A^T\right)^{-1} = (A^T)^{-1} \tilde{Q}_k R_{k+1}^T.$$

观察等式两边矩阵的最后一列, 可得

$$\tilde{q}_n^{(k+1)} = c_1 (A^T)^{-1} \tilde{q}_n^{(k)},$$

其中  $c_1$  为某个常数. 依此类推, 可知

$$\tilde{q}_n^{(k+1)} = c(A^T)^{-k} \tilde{q}_n^{(1)},$$

其中  $c$  为某个常数. 假设  $A$  的特征值满足  $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n| > 0$ , 则  $\lambda_n^{-1}$  是  $(A^T)^{-1}$  的模最大的特征值. 由幂迭代的收敛性可知,  $\tilde{q}_n^{(k+1)}$  收敛到  $(A^T)^{-1}$  的模最大特征值  $\lambda_n^{-1}$  所对应的特征向量, 即当  $k$  充分大时, 有

$$(A^T)^{-1} \tilde{q}_n^{(k+1)} \rightarrow \lambda_n^{-1} \tilde{q}_n^{(k+1)}.$$

所以

$$A^T \tilde{q}_n^{(k+1)} \rightarrow \lambda_n \tilde{q}_n^{(k+1)}.$$

由 (4.1) 可知,  $A_{k+1}^T$  的最后一列为

$$A_{k+1}^T(:, n) = \tilde{Q}_k^T A^T \tilde{q}_n^{(k)} \rightarrow \lambda_n \tilde{Q}_k^T \tilde{q}_n^{(k)} = \lambda_n e_n,$$

即  $A_{k+1}$  的最后一行的最后一个元素收敛到  $\lambda_n$ , 而其它元素都趋向于 0, 收敛速度取决于  $|\lambda_n/\lambda_{n-1}|$  的大小.

## 4.4.3 QR 迭代与正交迭代的关系

下面的定理给出了 QR 迭代法与正交迭代法 (取  $Z_0 = I$ ) 之间的关系.

**定理 4.2** 设正交迭代法 4.4 和 QR 算法 4.5 中所涉及的 QR 分解都是唯一的. 设  $A_k$  是由 QR 迭代法 4.5 生成的矩阵,  $Z_k$  是由正交迭代法 4.4 (取  $Z_0 = I$ ) 生成的矩阵, 则有

$$A_{k+1} = Z_k^T A Z_k.$$

(板书)

**证明.** 我们用归纳法证明该结论.

当  $k=0$  时,  $A_1 = A$ ,  $Z_0 = I$ . 结论显然成立.

设  $A_k = Z_{k-1}^T A Z_{k-1}$ . 由于  $Z_{k-1}$  是正交矩阵, 我们有

$$Z_k \hat{R}_k = Y_k = A Z_{k-1} = (Z_{k-1} Z_{k-1}^T) A Z_{k-1} = Z_{k-1} A_k = (Z_{k-1} Q_k) R_k,$$



即  $Z_k \hat{R}_k$  和  $(Z_{k-1} Q_k) R_k$  都是  $Y_k$  的 QR 分解. 由 QR 分解的唯一性可知,  $Z_k = Z_{k-1} Q_k$ ,  $\hat{R}_k = R_k$ . 所以

$$Z_k^T A Z_k = (Z_{k-1} Q_k)^T A (Z_{k-1} Q_k) = Q_k^T A_k Q_k = Q_k^T (Q_k R_k) Q_k = R_k Q_k = A_{k+1},$$

即  $A_{k+1} = Z_k^T A Z_k$ . 由归纳法可知, 定理结论成立.  $\square$

#### 4.4.4 QR 迭代的收敛性

**定理 4.3** 设  $A = V \Lambda V^{-1} \in \mathbb{R}^{n \times n}$ , 其中  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ , 且  $|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|$ . 若  $V^{-1}$  的所有顺序主子矩阵都非奇异 (即  $V^{-1}$  存在 LU 分解), 则  $A_k$  的对角线以下的元素均收敛到 0.

(板书)

**证明.** 设  $V = Q_v R_v$  是  $V$  的 QR 分解,  $V^{-1} = L_v U_v$  是  $V^{-1}$  的 LU 分解, 其中  $L_v$  是单位下三角矩阵. 则

$$A^k = V \Lambda^k V^{-1} = Q_v R_v \Lambda^k L_v U_v = Q_v R_v (\Lambda^k L_v \Lambda^{-k}) \Lambda^k U_v.$$

注意到矩阵  $\Lambda^k L_v \Lambda^{-k}$  是一个下三角矩阵, 且其  $(i, j)$  位置上的元素为

$$\left( \Lambda^k L_v \Lambda^{-k} \right) (i, j) = \begin{cases} 0, & i < j, \\ 1, & i = j, \\ L_v(i, j) \lambda_i^k / \lambda_j^k, & i > j. \end{cases}$$

由于当  $i > j$  时有  $|\lambda_i / \lambda_j| < 1$ , 故当  $k$  充分大时,  $\lambda_i^k / \lambda_j^k$  趋向于 0. 所以我们可以把  $\Lambda^k L_v \Lambda^{-k}$  写成

$$\Lambda^k L_v \Lambda^{-k} = I + E_k,$$

其中  $E_k$  满足  $\lim_{k \rightarrow \infty} E_k = 0$ . 于是

$$A^k = Q_v R_v (I + E_k) \Lambda^k U_v = Q_v (I + R_v E_k R_v^{-1}) R_v \Lambda^k U_v. \quad (4.3)$$

对矩阵  $I + R_v E_k R_v^{-1}$  做 QR 分解:  $I + R_v E_k R_v^{-1} = Q_{E_k} R_{E_k}$ . 由于  $E_k \rightarrow 0$ , 所以  $Q_{E_k} \rightarrow I$ ,  $R_{E_k} \rightarrow I$ . 将其代入 (4.3) 可得

$$A^k = Q_v Q_{E_k} R_{E_k} R_v \Lambda^k U_v = Q_v Q_{E_k} D_k (D_k^{-1} R_{E_k} R_v \Lambda^k U_v), \quad (4.4)$$

其中  $D_k$  是对角矩阵, 其对角线元素的模均为 1, 它使得上三角矩阵  $D_k^{-1} R_{E_k} R_v \Lambda^k U_v$  的对角线元素均为正. 这样, (4.4) 就构成  $A^k$  的 QR 分解. 又由 (4.2) 可知  $A^k = \tilde{Q}_k \tilde{R}_k$ , 根据 QR 分解的唯一性, 我们可得

$$\tilde{Q}_k = Q_v Q_{E_k} D_k, \quad \tilde{R}_k = D_k^{-1} R_{E_k} R_v \Lambda^k U_v.$$

所以由 (4.1) 可知

$$\begin{aligned} A_{k+1} &= \tilde{Q}_k^T A \tilde{Q}_k \\ &= (Q_v Q_{E_k} D_k)^T V \Lambda V^{-1} Q_v Q_{E_k} D_k \\ &= D_k^T Q_{E_k}^T Q_v^T Q_v R_v \Lambda R_v^{-1} Q_v^{-1} Q_v Q_{E_k} D_k \\ &= D_k^T Q_{E_k}^T R_v \Lambda R_v^{-1} Q_{E_k} D_k. \end{aligned}$$



由于  $Q_{E_k} \rightarrow I$ , 所以当  $k \rightarrow \infty$  时,  $A_{k+1}$  收敛到一个上三角矩阵. 收敛速度取决于

$$\max_{1 \leq i < n} \left| \frac{\lambda_{i+1}}{\lambda_i} \right|$$

的大小.

□ 需要指出的是, 由于  $D_k$  的元素不一定收敛, 故  $A_{k+1}$  对角线以上 (不含对角线) 的元素不一定收敛, 但这不妨碍  $A_{k+1}$  的对角线元素收敛到  $A$  的特征值 (即  $A_{k+1}$  的对角线元素是收敛的).

**例 4.3** QR 迭代法演示. 设

$$A = X \begin{bmatrix} 9 & & & \\ & 5 & & \\ & & 3 & \\ & & & 1 \end{bmatrix} X^{-1},$$

其中  $X$  是由 MATLAB 随机生成的非奇异矩阵.

在迭代过程中, 对于  $A_k$  的下三角部分中元素, 如果其绝对值小于某个阈值  $tol$ , 则直接将其设为 0, 即

$$a_{ij}^{(k)} = 0 \quad \text{if } i > j \text{ and } |a_{ij}^{(k)}| < tol.$$

这里我们取  $tol = 10^{-6} \max_{1 \leq i, j \leq n} \{|a_{ij}^{(k)}|\}$ .

(Fig\_QR.m)

#### 4.4.5 带位移的 QR 迭代法

为了加快 QR 迭代的收敛速度, 我们可以采用位移策略和反迭代思想.

**算法 4.6.** 带位移的 QR 迭代法 (QR Iteration with shift)

```

1: Set $A_1 = A$ and $k = 1$
2: while not convergence do
3: Choose a shift σ_k
4: $A_k - \sigma_k I = Q_k R_k$ % QR 分解
5: Compute $A_{k+1} = R_k Q_k + \sigma_k I$
6: $k = k + 1$
7: end while

```

与不带位移的 QR 迭代一样, 我们有

$$A_{k+1} = R_k Q_k + \sigma_k I = (Q_k^T Q_k) R_k Q_k + \sigma_k I = Q_k^T (A_k - \sigma_k I) Q_k + \sigma_k I = Q_k^T A_k Q_k$$

所以, 带位移的 QR 算法中所得到的矩阵  $A_k$  仍然与  $A_1 = A$  正交相似.

在带位移的 QR 迭代法中, 一个很重要的问题就是位移  $\sigma_k$  的选取. 在前面的分析中我们已经知道,  $A_{k+1}(n, n)$  将收敛到  $A$  的模最小的特征值, 且收敛速度取决于模最小特征值与模第二小特征值之间的比值. 显然, 若  $\sigma_k$  就是  $A$  的一个特征值, 则  $A_k - \sigma_k I$  的模最小特征值为 0, 故 QR 算



法迭代一步就收敛. 此时

$$A_{k+1} = R_k Q_k + \sigma_k I = \begin{bmatrix} A_{k+1}^{(n-1) \times (n-1)} & * \\ 0 & \sigma_k \end{bmatrix}.$$

如果需要计算  $A$  的其它特征值, 则可对子矩阵  $A_{k+1}^{(n-1) \times (n-1)}$  使用带位移的 QR 迭代法.

通常, 如果  $\sigma_k$  与  $A$  的某个特征值非常接近, 则收敛速度通常会很快. 由于  $A_k(n, n)$  收敛到  $A$  的一个特征值, 所以在实际使用中, 一个比较直观的位移选择策略是  $\sigma_k = A_k(n, n)$ . 事实上, 这样的位移选取方法通常会使得 QR 迭代有二次收敛速度.

**例 4.4** 带位移的 QR 迭代法演示. 所有数据和设置与例 4.3 相同, 在迭代过程中, 取  $\sigma_k = A_k(n, n)$ . 如果  $A_k(n, n)$  已经收敛, 则取  $\sigma_k = A_k(n-1, n-1)$ . (Eig\_QR\_shift.m)



## 4.5 带位移的隐式 QR 迭代法

QR 迭代法中需要考虑的另一个重要问题就是运算量: 每一步迭代都需要做一次 QR 分解和矩阵乘积, 运算量为  $\mathcal{O}(n^3)$ . 即使每计算一个特征值只需迭代一步, 则计算所有特征值也需要  $\mathcal{O}(n^4)$  的运算量. 这是令人无法忍受的. 下面我们就想办法将总运算量从  $\mathcal{O}(n^4)$  减小到  $\mathcal{O}(n^3)$ .

为了实现这个目标, 我们需要利用 Hessenberg 矩阵. 具体步骤如下: 首先通过相似变化将  $A$  转化成一个上 Hessenberg 矩阵, 然后再对这个 Hessenberg 矩阵实施**隐式 QR 迭代**. 所谓隐式 QR 迭代, 就是在 QR 迭代中, 我们不需要进行显式的 QR 分解. 这样就可以将 QR 迭代的每一步运算量从  $\mathcal{O}(n^3)$  降低到  $\mathcal{O}(n^2)$ . 从而将总的运算量降低到  $\mathcal{O}(n^3)$ .

### 4.5.1 上 Hessenberg 矩阵

设  $H = [h_{ij}] \in \mathbb{R}^{n \times n}$ , 若当  $i > j + 1$  时, 有  $h_{ij} = 0$ , 则称  $H$  为 **上 Hessenberg 矩阵**.

**定理 4.4** 设  $A \in \mathbb{R}^{n \times n}$ , 则存在正交矩阵  $Q \in \mathbb{R}^{n \times n}$ , 使得  $Q A Q^T$  是上 Hessenberg 矩阵.

下面我们以一个  $5 \times 5$  的矩阵  $A$  为例, 给出具体的**上 Hessenberg 化**过程, 所采用的工具仍然是 Householder 变换.

**第一步:** 令  $Q_1 = \text{diag}(I_{1 \times 1}, H_1)$ , 其中  $H_1$  是对应于向量  $A(2:5, 1)$  的 Householder 矩阵. 于是可得

$$Q_1 A = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \end{bmatrix}.$$

由于用  $Q_1^T$  右乘  $Q_1 A$  时, 不会改变  $Q_1 A$  第一列元素的值, 故

$$A_1 \triangleq Q_1 A Q_1^T = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \end{bmatrix}.$$

**第二步:** 令  $Q_2 = \text{diag}(I_{2 \times 2}, H_2)$ , 其中  $H_2$  是对应于向量  $A_1(3:5, 2)$  的 Householder 矩阵, 则用  $Q_2$  左乘  $A_1$  时, 不会改变  $A_1$  的第一列元素的值. 用  $Q_2^T$  右乘  $Q_2 A_1$  时, 不会改变  $Q_2 A_1$  前两列元素的值. 因此,

$$Q_2 A_1 = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{bmatrix} \quad \text{和} \quad A_2 \triangleq Q_2 A_1 Q_2^T = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{bmatrix}.$$



**第三步:** 令  $Q_3 = \text{diag}(I_{3 \times 3}, H_3)$ , 其中  $H_3$  是对应于向量  $A_2(4:5, 3)$  的 Householder 矩阵, 则有


$$Q_3 A_2 = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix} \quad \text{和} \quad A_3 \triangleq Q_3 A_2 Q_3^T = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}.$$

这时, 我们就将  $A$  转化成一个上 Hessenberg 矩阵, 即  $Q A Q^T = A_3$  其中  $Q = Q_3 Q_2 Q_1$  是正交矩阵,  $A_3$  是上 Hessenberg 矩阵.

下面是将任意一个矩阵转化成上 Hessenberg 矩阵的算法.

#### 算法 4.7. 上 Hessenberg 化 (Upper Hessenberg Reduction)

- 1: Set  $Q = I$
- 2: **for**  $k = 1$  to  $n - 2$  **do**
- 3:     compute Householder matrix  $H_k = I - \beta_k v_k v_k^T$  with respect to  $A(k+1:n, k)$
- 4:      $A(k+1:n, k:n) = H_k A(k+1:n, k:n)$   
 $\quad\quad\quad = A(k+1:n, k:n) - \beta_k v_k (v_k^T A(k+1:n, k:n))$
- 5:      $A(1:n, k+1:n) = A(1:n, k+1:n) H_k^T$   
 $\quad\quad\quad = A(1:n, k+1:n) - \beta_k A(1:n, k+1:n) v_k v_k^T$
- 6:      $Q(k+1:n, :) = H_k Q(k+1:n, :)$   
 $\quad\quad\quad = Q(k+1:n, :) - \beta_k v_k (v_k^T Q(k+1:n, :))$
- 7: **end for**

 在实际计算时, 我们不需要显式地形成 Householder 矩阵  $H_k$ .

上述算法的运算量大约为  $\frac{14}{3}n^3 + \mathcal{O}(n^2)$ . 如果不需要计算特征向量, 则正交矩阵  $Q$  也不用计算, 此时运算量大约为  $\frac{10}{3}n^3 + \mathcal{O}(n^2)$ .

上 Hessenberg 矩阵的一个很重要的性质就是在 QR 迭代中能保持形状不变.

**定理 4.5** 设  $A \in \mathbb{R}^{n \times n}$  是非奇异上 Hessenberg 矩阵, 其 QR 分解为  $A = QR$ , 则  $\tilde{A} \triangleq RQ$  也是上 Hessenberg 矩阵.

(板书)

**证明.** 设  $A = QR$  是  $A$  的 QR 分解, 则  $Q = AR^{-1}$ . 由于  $R$  是一个上三角矩阵, 所以  $R^{-1}$  也是一个上三角矩阵. 因此  $Q$  的第  $j$  列是  $A$  的前  $j$  列的线性组合. 又  $A$  是上 Hessenberg 矩阵, 所以  $Q$  也是一个上 Hessenberg 矩阵.

相类似地, 我们很容易验证  $RQ$  也是一个上 Hessenberg 矩阵. 所以结论成立.  $\square$



若  $A$  是奇异的, 也可以通过选取适当的  $Q$ , 使得定理 4.5 中的结论成立. 事实上, 由基于 Gram-Schmidt 过程的 QR 分解可知, 如果  $A$  是奇异的上 Hessenberg 矩阵, 则  $Q$  仍可取为上 Hessenberg 矩阵.

由这个性质可知, 如果  $A \in \mathbb{R}^{n \times n}$  是上 Hessenberg 矩阵, 则 QR 迭代中的每一个  $A_k$  都是上 Hessenberg 矩阵. 这样, 在进行 QR 分解时, 运算量可大大降低.

Hessenberg 矩阵还有一个重要性质, 就是在 QR 迭代过程中能保持下次对角线元素非零.

**定理 4.6** 设  $A \in \mathbb{R}^{n \times n}$  是非奇异的上 Hessenberg 矩阵, 且下次对角线元素均非零, 即  $a_{i+1,i} \neq 0$ ,  $i = 1, 2, \dots, n-1$ . 设其 QR 分解为  $A = QR$ , 则  $\tilde{A} \triangleq RQ$  的下次对角线元素也都非零.

(留作练习)

易知, 如果上 Hessenberg 矩阵  $A$  存在某个下次对角线元素为零, 则  $A$  一定可约, 此时计算  $A$  的特征值就转化为计算两个更小规模的矩阵的特征值. 因此, 我们只需考虑下次对角线均非零的情形.

需要指出的是, 如果上 Hessenberg 矩阵  $A$  存在某个下次对角线元素为零, 则  $A$  一定可约, 但反之却不成立, 即如果  $A$  是可约的上 Hessenberg 矩阵,  $A$  的下次对角线元素仍可能均非零.

**推论 4.7** 设  $A \in \mathbb{R}^{n \times n}$  是非奇异的上 Hessenberg 矩阵, 且下次对角线元素均非零, 则在带位移的 QR 迭代中, 所有的  $A_k$  的下次对角线元素均非零.

### 4.5.2 隐式 QR 迭代

在 QR 迭代中, 我们需要先做 QR 分解  $A_k = Q_k R_k$ , 然后再计算  $A_{k+1} = R_k Q_k$ . 但事实上, 我们可以将这个过程进行简化, 即在不对  $A_k$  进行 QR 分解的前提下, 直接计算出  $A_{k+1}$ . 这就是**隐式 QR 迭代**.

我们这里考虑不可约的上 Hessenberg 矩阵, 即  $A$  的下次对角线元素都不为 0. 事实上, 若  $A$  是可约的, 则  $A$  就是一个块上三角矩阵, 这时  $A$  的特征值计算问题就转化成计算两个对角块的特征值问题. 隐式 QR 迭代的理论基础就是下面的**隐式 Q 定理**.

**定理 4.8 (Implicit Q Theorem)** 设  $H = Q^T A Q \in \mathbb{R}^{n \times n}$  是一个不可约上 Hessenberg 矩阵, 其中  $Q \in \mathbb{R}^{n \times n}$  是正交矩阵, 则  $Q$  的第 2 至第  $n$  列均由  $Q$  的第一列所唯一确定 (可相差一个符号).

(板书)

**证明.** 设  $H = Q^T A Q$  和  $G = V^T A V$  都是不可约上 Hessenberg 矩阵, 其中  $Q = [q_1, q_2, \dots, q_n]$ ,  $V = [v_1, v_2, \dots, v_n]$  都是正交矩阵, 且  $q_1 = v_1$ . 下面我们只需证明  $q_i = v_i$  或  $q_i = -v_i$ ,  $i = 2, 3, \dots, n$ , 即证明

$$W \triangleq V^T Q = \text{diag}(1, \pm 1, \dots, \pm 1).$$



记  $W = [w_1, w_2, \dots, w_n]$ ,  $H = [h_{ij}]$ , 则有

$$GW = GV^T Q = (V^T AV)V^T Q = V^T AQ = V^T Q(Q^T AQ) = V^T QH = WH,$$

即

$$Gw_i = \sum_{j=1}^{i+1} h_{ji} w_j, \quad i = 1, 2, \dots, n-1.$$

所以

$$h_{i+1,i} w_{i+1} = Gw_i - \sum_{j=1}^i h_{ji} w_j.$$

因为  $q_1 = v_1$ , 所以  $w_1 = [1, 0, \dots, 0]^T$ . 又  $G$  是上 Hessenberg 矩阵, 利用归纳法, 我们可以证明  $w_i$  的第  $i+1$  到第  $n$  个元素均为 0. 所以  $W$  是一个上三角矩阵. 又  $W$  是正交矩阵, 所以  $W = \text{diag}(1, \pm 1, \dots, \pm 1)$ . 由此, 定理结论成立.  $\square$

由于  $Q_k$  的其它列都是由  $Q_k$  的第一列唯一确定 (至多相差一个符号), 所以我们只要找到一个正交矩阵  $\tilde{Q}_k$  使得其第一列与  $Q_k$  的第一列相等, 且  $\tilde{Q}_k^T A_k \tilde{Q}_k$  为上 Hessenberg 矩阵, 则由隐式 Q 定理可知  $\tilde{Q}_k = WQ_k$ , 其中  $W = \text{diag}(1, \pm 1, \dots, \pm 1)$ . 于是

$$\tilde{Q}_k^T A_k \tilde{Q}_k = W^T Q_k^T A_k Q_k W = W^T A_{k+1} W.$$

又  $W^T A_{k+1} W$  与  $A_{k+1}$  相似, 且对角线元素相等, 而其它元素也至多相差一个符号, 所以不会影响  $A_{k+1}$  的收敛性, 即下三角元素收敛到 0, 对角线元素收敛到  $A$  的特征值. 换句话说, 在 QR 迭代法中, 如果我们用  $\tilde{Q}_k^T A_k \tilde{Q}_k$  代替  $Q_k A_k Q_k^T$ , 即直接令  $A_{k+1} = \tilde{Q}_k^T A_k \tilde{Q}_k$ , 则其收敛性与原 QR 迭代法没有任何区别! 这就是隐式 QR 迭代的基本思想.

由于  $A$  是上 Hessenberg 矩阵, 因此在计算中可以使用 Givens 变换, 即  $\tilde{Q}_k$  是一系列 Givens 变换的乘积. 下面我们举一个例子, 具体说明如何利用隐式 Q 定理, 由  $A_1$  得到  $A_2$ .

**例 4.5** 设  $A \in \mathbb{R}^{5 \times 5}$  是一个不可约上 Hessenberg 矩阵, 即

$$A_1 = A = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}.$$

**第一步:** 构造一个 Givens 变换  $G_1$ , 其转置如下

$$G_1^T \triangleq G(1, 2, \theta_1) = \begin{bmatrix} c_1 & s_1 & & & \\ -s_1 & c_1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}.$$

这里  $G_1$  的第一列  $[c_1, s_1, 0, \dots, 0]^T$  就是  $A_1 - \sigma_1 I$  的第一列  $[a_{11} - \sigma_1, a_{21}, 0, \dots, 0]^T$  的单





位化后的列向量, 其中  $\sigma_1$  是位移. 于是有

$$G_1^T A = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix} \quad \text{和} \quad A^{(1)} \triangleq G_1^T A G_1 = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ + & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}.$$

与  $A_1$  相比较,  $A^{(1)}$  在  $(3,1)$  位置上多出一个非零元, 我们把它记为 “+”, 并称之为 **bulge**. 在下面的计算过程中, 我们的目标就是将其 “赶” 出矩阵, 从而得到一个新的上 Hessenberg 矩阵, 即  $A_2$ .

**第二步:** 为了消去这个 bulge, 我们可以构造 Givens 变换

$$G_2^T \triangleq G(2, 3, \theta_2) = \begin{bmatrix} 1 & & & & \\ & c_2 & s_2 & & \\ & -s_2 & c_2 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix} \quad \text{使得} \quad G_2^T A^{(1)} = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}.$$

为了保持与原矩阵的相似性, 需要再右乘  $G_2$ , 所以

$$A^{(2)} \triangleq G_2^T A^{(1)} G_2 = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & + & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}.$$

此时, bulge 从  $(3,1)$  位置被 “赶” 到  $(4,2)$  位置.

**第三步:** 与第二步类似, 构造 Givens 变换

$$G_3^T \triangleq G(3, 4, \theta_3) = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & c_3 & s_3 & \\ & & -s_3 & c_3 & \\ & & & & 1 \end{bmatrix} \quad \text{使得} \quad G_3^T A^{(2)} = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}.$$

这时

$$A^{(3)} \triangleq G_3^T A^{(2)} G_3 = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & + & * & * \end{bmatrix}.$$

于是, bulge 又从  $(4,2)$  位置又被 “赶” 到  $(5,3)$  位置.



**第四步:** 再次构造 Givens 变换

$$G_4^T \triangleq G(4, 5, \theta_4) = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & c_4 & s_4 \\ & & & -s_4 & c_4 \end{bmatrix} \quad \text{使得} \quad G_4^T A^{(3)} = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}$$

于是

$$A^{(4)} \triangleq G_4^T A^{(3)} G_4 = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}.$$

现在, bulge 已经被“赶”出矩阵, 且

$$A^{(4)} = G_4^T G_3^T G_2^T G_1^T A_1 G_1 G_2 G_3 G_4 = \tilde{Q}_1^T A_1 \tilde{Q}_1,$$

其中  $\tilde{Q}_1 = G_1 G_2 G_3 G_4$ . 通过直接计算可知,  $\tilde{Q}_1$  的第一列为  $[c_1, s_1, 0, 0, 0]^T$ , 即  $A_1 - \sigma_1 I$  的第一列的单位化. 根据隐式 Q 定理,  $A_2 \triangleq A^{(4)} = \tilde{Q}_1^T A_1 \tilde{Q}_1$  就是我们所需要的矩阵.

如果  $A \in \mathbb{R}^{n \times n}$  是上 Hessenberg 矩阵, 则使用上面的算法, 带位移的隐式 QR 迭代中每一步的运算量为  $6n^2 + \mathcal{O}(n)$ , 如果需要计算  $Q$  的话, 则总运算量为  $12n^2 + \mathcal{O}(n)$ .

### 4.5.3 位移的选取

在带位移的 QR 迭代法中, 位移的选取非常重要. 通常, 位移越离某个特征值越近, 收敛速度就越快. 由习题 4.5 可知, 如果位移  $\sigma$  与某个特征值非常接近, 则  $A_k(n, n) - \sigma$  就非常接近于 0. 这说明  $A_k(n, n)$  通常会首先收敛到  $A$  的一个特征值, 所以  $\sigma = A_k(n, n)$  是一个不错的选择. 但是, 如果这个特征值是复数, 这种位移选取方法就可能失效.

下面我们介绍一种针对共轭复特征值的位移选取方法, 即双位移策略.

设  $\sigma \in \mathbb{C}$  是  $A$  的某个复特征值  $\lambda$  的一个很好的近似, 则其共轭  $\bar{\sigma}$  也应该是  $\bar{\lambda}$  的一个很好的近似. 因此我们可以考虑**双位移**策略, 即先以  $\sigma$  为位移迭代一次, 然后再以  $\bar{\sigma}$  为位移迭代一次, 如此不断交替进行迭代. 这样就有

$$\begin{aligned} A_1 - \sigma I &= Q_1 R_1, \\ A_2 &= R_1 Q_1 + \sigma I, \\ A_2 - \bar{\sigma} I &= Q_2 R_2, \\ A_3 &= R_2 Q_2 + \bar{\sigma} I. \end{aligned} \tag{4.5}$$

容易验证

$$A_3 = Q_2^* A_2 Q_2 = Q_2^* Q_1^* A_1 Q_1 Q_2 = Q^* A_1 Q,$$

其中  $Q = Q_1 Q_2$ . 我们注意到  $\sigma$  可能是复的, 所以  $Q_1$  和  $Q_2$  都可能是复矩阵. 但我们却可以选取适当的  $Q_1$  和  $Q_2$ , 使得  $Q = Q_1 Q_2$  是实正交矩阵.



**引理 4.9** 在双位移 QR 迭代 (4.5) 中, 我们可以选取酉矩阵  $Q_1$  和  $Q_2$  使得  $Q = Q_1 Q_2$  是实矩阵.

(板书)

**证明.** 由于

$$Q_2 R_2 = A_2 - \bar{\sigma} I = R_1 Q_1 + (\sigma - \bar{\sigma}) I,$$

所以

$$\begin{aligned} Q_1 Q_2 R_2 R_1 &= Q_1 (R_1 Q_1 + (\sigma - \bar{\sigma}) I) R_1 \\ &= Q_1 R_1 Q_1 R_1 + (\sigma - \bar{\sigma}) Q_1 R_1 \\ &= (A_1 - \sigma I)^2 + (\sigma - \bar{\sigma})(A_1 - \sigma I) \\ &= A_1^2 - (\sigma + \bar{\sigma}) A_1 + \bar{\sigma} \sigma I, \\ &= A_1^2 - 2\operatorname{Re}(\sigma) A_1 + |\sigma|^2 I \in \mathbb{R}^{n \times n}. \end{aligned}$$

又  $Q_1 Q_2$  是酉矩阵,  $R_2 R_1$  是上三角矩阵, 故  $(Q_1 Q_2)(R_2 R_1)$  是实矩阵  $A_1^2 - 2\operatorname{Re}(\sigma) A_1 + |\sigma|^2 I = (A_1 - \sigma I)(A_1 - \bar{\sigma} I)$  的 QR 分解. 所以  $Q_1 Q_2$  和  $R_2 R_1$  都可以是实矩阵.  $\square$

由这个引理可知, 存在  $Q_1$  和  $Q_2$ , 使得  $Q = Q_1 Q_2$  是实正交矩阵, 从而  $A_3 = Q^T A_1 Q$  也是实矩阵. 这时, 在迭代过程 (4.5) 中, 我们无需计算  $A_2$ , 可直接由  $A_1$  计算出  $A_3$ .

具体计算过程仍然是根据隐式 Q 定理: 我们只要找到一个实正交矩阵  $Q$ , 使得其第一列与  $A_1^2 - 2\operatorname{Re}(\sigma) A_1 + |\sigma|^2 I$  的第一列平行, 并且  $A_3 = Q^T A_1 Q$  是上 Hessenberg 矩阵即可.

由于  $A_1^2 - 2\operatorname{Re}(\sigma) A_1 + |\sigma|^2 I$  的第一列为

$$\begin{bmatrix} a_{11}^2 + a_{12}a_{21} - 2\operatorname{Re}(\sigma)a_{11} + |\sigma|^2 \\ a_{21}(a_{11} + a_{22} - 2\operatorname{Re}(\sigma)) \\ a_{21}a_{32} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (4.6)$$

所以  $Q$  的第一列是上述向量的单位化. 其它过程可以通过隐式 QR 迭代来实现. 但此时的“bulge”是一个  $2 \times 2$  的小矩阵. 因此, 在双位移隐式 QR 迭代过程中, 我们需要使用 Householder 变换.

需要指出的是, 双位移 QR 迭代法中的运算都是实数运算.

下面我们举一个例子, 具体说明如何在实数运算下实现双位移隐式 QR 迭代法.

**例 4.6** 设  $A \in \mathbb{R}^{6 \times 6}$  是一个不可约上 Hessenberg 矩阵, 即

$$A_1 = A = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & * & * \end{bmatrix}.$$



**第一步:** 构造一个正交矩阵  $H_1 = \begin{bmatrix} \tilde{H}_1^\top & 0 \\ 0 & I_{3 \times 3} \end{bmatrix}$ , 其中  $\tilde{H}_1 \in \mathbb{R}^{3 \times 3}$ , 使得其第一列与向量 (4.6)

平行. 于是有

$$H_1^\top A = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ + & * & * & * & * & * \\ 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & * & * \end{bmatrix} \quad \text{和} \quad A^{(1)} \triangleq H_1^\top A H_1 = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ + & * & * & * & * & * \\ + & + & * & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & * & * \end{bmatrix}.$$

与  $A_1$  相比较,  $A^{(1)}$  在  $(3, 1)$ ,  $(4, 1)$  和  $(4, 2)$  位置上出现 **bulge**. 在下面的计算过程中, 我们的目标就是要把它们“赶”出矩阵, 从而得到一个新的上 Hessenberg 矩阵.

**第二步:** 令  $H_2 = \begin{bmatrix} I_{1 \times 1} & 0 & 0 \\ 0 & \tilde{H}_2^\top & 0 \\ 0 & 0 & I_{2 \times 2} \end{bmatrix}$ , 其中  $\tilde{H}_2 \in \mathbb{R}^{3 \times 3}$  是对应于  $A(2:4, 1)$  的 Householder 变换, 使得

$$H_2^\top A^{(1)} = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & + & * & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & * & * \end{bmatrix} \quad \text{和} \quad A^{(2)} \triangleq H_2^\top A^{(1)} H_2 = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & + & * & * & * & * \\ 0 & + & + & * & * & * \\ 0 & 0 & 0 & 0 & * & * \end{bmatrix}.$$

这时, 我们将 bulge 向右下角方向“赶”了一个位置.

**第三步:** 与第二步类似, 令  $H_3 = \begin{bmatrix} I_{2 \times 2} & 0 & 0 \\ 0 & \tilde{H}_3^\top & 0 \\ 0 & 0 & I_{1 \times 1} \end{bmatrix}$ , 其中  $\tilde{H}_3 \in \mathbb{R}^{3 \times 3}$  是对应于  $A(3:5, 2)$  的 Householder 变换, 使得

$$H_3^\top A^{(2)} = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \\ 0 & 0 & + & * & * & * \\ 0 & 0 & 0 & 0 & * & * \end{bmatrix} \quad \text{和} \quad A^{(3)} \triangleq H_3^\top A^{(2)} H_3 = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \\ 0 & 0 & + & * & * & * \\ 0 & 0 & + & + & * & * \end{bmatrix}.$$

此时, bulge 又被向右下角方向“赶”了一个位置.

**第四步:** 令  $H_4 = \begin{bmatrix} I_{3 \times 3} & 0 \\ 0 & \tilde{H}_4^\top \end{bmatrix}$ , 其中  $\tilde{H}_4 \in \mathbb{R}^{3 \times 3}$  是对应于  $A(4:6, 3)$  的 Householder 变换,



使得

$$H_4^T A^{(3)} = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & + & * & * \end{bmatrix} \quad \text{和} \quad A^{(4)} \triangleq H_4^T A^{(3)} H_4 = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & + & * & * \end{bmatrix}.$$

**第五步:** 此时, 只需构造一个 Givens 变换  $G_5 = \begin{bmatrix} I_{4 \times 4} & 0 \\ 0 & G(4, 5, \theta)^T \end{bmatrix}$ , 使得

$$G_5^T A^{(4)} = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & * & * \end{bmatrix} \quad \text{和} \quad A^{(5)} \triangleq G_5^T A^{(4)} G_5 = \begin{bmatrix} * & * & * & * & * & * \\ * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & * & * \end{bmatrix}.$$

现在, bulge 已经被全部消除, 且

$$A^{(5)} = Q^T A Q,$$

其中  $Q = H_1 H_2 H_3 H_4 G_5$ . 通过直接计算可知,  $Q$  的第一列即为  $H_1$  的第一列, 也就是向量 (4.6) 的单位化. 根据隐式 Q 定理, 可以直接令  $A_3 \triangleq A^{(5)} = Q^T A Q$ .


最后要考虑位移  $\sigma$  的具体选取问题. 在单位 QR 迭代法中, 如果  $A$  的特征值都是实数的话, 我们可以取  $\sigma = A_k(n, n)$ . 推广到复共轭特征值上, 我们可以取  $A_k$  的右下角矩阵

$$\begin{bmatrix} A_k(n-1, n-1) & A_k(n-1, n) \\ A_k(n, n-1) & A_k(n, n) \end{bmatrix}$$

的复共轭特征值作为双位移. 这样选取的位移就是 **Francis 位移**.

 如果上述矩阵的两个特征值都是实的, 则选取其中模较小的特征值做单位 QR 迭代.

一般来说, 采用 Francis 位移的 QR 迭代法会使得迭代矩阵的右下角收敛到一个上三角矩阵 (两个实特征值) 或一个  $2 \times 2$  的矩阵 (一对复共轭特征值), 即  $A_k(n, n-1)$  或  $A_k(n-1, n-2)$  趋向于 0. 带 Francis 位移的隐式 QR 迭代法通常具有二次渐进收敛性, 在实际计算中, 计算一个特征值大约平均迭代两步.

 有时也可能是中间某个下次对角线元素 (不是最后两个) 首先收敛到 0, 此时可以转化为计算两个子矩阵 (对角块) 的特征值. 因此在迭代中要判别所有下次对角线元素是否满足精度要求.

但需要指出的是, QR 迭代并不是对所有矩阵都收敛. 例如:

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

对于上面的矩阵, 采用 Francis 位移的 QR 迭代法无效.

另外, 也可以考虑多重位移策略, 参见 [125].

#### 4.5.4 收缩

收缩 (deflation) 技术是实用 QR 迭代中的一个非常重要概念.

在隐式 QR 迭代过程中, 当矩阵  $A_{k+1}$  的某个下次对角线元素  $a_{i+1,i}$  很小时, 我们可以将其设为 0. 由于  $A_{k+1}$  是上 Hessenberg 矩阵, 这时  $A_{k+1}$  就可以写成分块上三角形式, 其中两个对角块都是上 Hessenberg 矩阵. 因此我们可以将隐式 QR 迭代作用在这两个规模相对较小的矩阵上, 从而可以大大节约运算量.



## 4.6 特征向量的计算

设  $A$  的特征值都是实的,  $R = Q^T A Q$  是其 Schur 标准型. 若  $Ax = \lambda x$ , 则  $Ry = \lambda y$ , 其中  $y = Q^T x$  或  $x = Qy$ . 故只需计算  $R$  对应于  $\lambda$  的特征向量  $y$  即可.

因为  $R$  的对角线元素即为  $A$  的特征值, 不妨设  $\lambda = R(i, i)$ . 假定  $\lambda$  是单重特征值, 则方程  $(R - \lambda I)y = 0$  即为

$$\begin{bmatrix} R_{11} - \lambda I & R_{12} & R_{13} \\ 0 & 0 & R_{23} \\ 0 & 0 & R_{33} - \lambda I \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = 0,$$

即

$$(R_{11} - \lambda I)y_1 + R_{12}y_2 + R_{13}y_3 = 0, \quad (4.7)$$

$$R_{23}y_3 = 0, \quad (4.8)$$

$$(R_{33} - \lambda I)y_3 = 0, \quad (4.9)$$

其中  $R_{11} \in \mathbb{R}^{(i-1) \times (i-1)}$ ,  $R_{33} \in \mathbb{R}^{(n-i) \times (n-i)}$ . 由于  $\lambda$  是单重特征值, 故  $R_{33} - \lambda I$  非奇异, 因此  $y_3 = 0$ . 令  $y_2 = 1$ , 则可得

$$y_1 = (R_{11} - \lambda I)^{-1} R_{12}.$$

因此计算特征向量  $y$  只需求解一个上三角线性方程组.

若  $\lambda$  是多重特征值, 则计算方法类似. 但如果  $A$  有复特征值, 则需要利用实 Schur 标准型.



## 4.7 广义特征值问题

### 4.7.1 广义特征值基本理论

设  $A, B \in \mathbb{R}^{n \times n}$ , 若存在  $\lambda \in \mathbb{C}$  和非零向量  $x \in \mathbb{C}^n$  使得

$$Ax = \lambda Bx, \quad (4.10)$$

则称  $\lambda$  为**矩阵对** (或**矩阵束**, **matrix pair**, **matrix pencil**)  $(A, B)$  的特征值,  $x$  为相应的特征向量. 计算矩阵对  $(A, B)$  的特征值和特征向量就是**广义特征值问题**. 当  $B = I$  时, 广义特征值问题就退化为标准特征值问题. 当  $B$  非奇异时, 广义特征值问题就等价于标准特征值问题

$$B^{-1}Ax = \lambda x \quad \text{或} \quad AB^{-1}y = \lambda y,$$

其中  $y = Bx$ .

容易看出,  $\lambda$  是  $(A, B)$  的一个特征值当且仅当

$$\det(A - \lambda B) = 0. \quad (4.11)$$

若 (4.11) 对所有  $\lambda \in \mathbb{C}$  都成立, 则称矩阵对  $(A, B)$  是**奇异矩阵对**, 否则称为**正则矩阵对**.

当  $B$  非奇异时, 特征方程 (4.11) 是一个  $n$  次多项式, 因此恰好有  $n$  个特征值. 当  $B$  奇异时, 特征方程 (4.11) 的次数低于  $n$ , 因此方程的解的个数小于  $n$ . 但是, 注意到  $\lambda \neq 0$  是  $(A, B)$  的特征值当且仅当  $\mu = \frac{1}{\lambda}$  是  $(B, A)$  的特征值. 因此, 当  $B$  奇异时,  $\mu = 0$  是  $(B, A)$  的特征值, 于是我们自然地把  $\lambda = \frac{1}{\mu} = \infty$  当作是  $(A, B)$  的特征值. 所以, 广义特征值不是分布在  $\mathbb{C}$  上, 而是分布在  $\mathbb{C} \cup \{\infty\}$  上.

容易验证, 若  $U, V$  非奇异, 则矩阵对  $(U^*AU, U^*BV)$  的特征值与  $(A, B)$  是一样的. 因此我们称这种变换为**矩阵对的等价变换**. 如果  $U, V$  是酉矩阵, 则称为**酉等价变换**.

### 4.7.2 广义 Schur 分解

**广义 Schur 分解** 是矩阵对在酉等价变化下的最简形式.

**定理 4.10 (广义 Schur 分解)** 设  $A, B \in \mathbb{C}^{n \times n}$ , 则存在酉矩阵  $Q, Z \in \mathbb{C}^{n \times n}$ , 使得

$$Q^*AZ = R_A, \quad Q^*BZ = R_B, \quad (4.12)$$

其中  $R_A, R_B \in \mathbb{C}^{n \times n}$  都是上三角矩阵. 此时矩阵对  $(A, B)$  的特征值为  $R_A$  和  $R_B$  的对角线元素的比值, 即

$$\lambda_i = \frac{R_A(i, i)}{R_B(i, i)}, \quad i = 1, 2, \dots, n.$$

当  $R_B(i, i) = 0$  时, 对应的特征值  $\lambda_i = \infty$ .

(留作课外自习)

**证明.** 若  $B$  非奇异, 则可设  $B^{-1}A$  的 Schur 分解为

$$Z^*B^{-1}AZ = R,$$

其中  $R \in \mathbb{C}^{n \times n}$  是上三角矩阵,  $Z \in \mathbb{C}^{n \times n}$  是酉矩阵. 令  $BZ$  的 QR 分解为

$$BZ = QT,$$





其中  $T \in \mathbb{C}^{n \times n}$  是上三角矩阵,  $Q \in \mathbb{C}^{n \times n}$  是酉矩阵. 则  $Q^* B Z = T$ , 且

$$Q^* A Z = T R$$

是上三角矩阵. 令  $S = T R$  即可.

若  $B$  奇异, 则根据矩阵特征值的连续性, 存在序列  $\{B_k\}$ , 使得  $B_k$  非奇异, 且收敛到  $B$ . 由上面的证明可知, 存在酉矩阵  $Q_k$  和  $Z_k$ , 使得  $Q_k^* B_k Z_k$  和  $Q_k^* A Z_k$  都是上三角矩阵. 由于  $\{Q_k\}$  和  $\{Z_k\}$  都是有界的, 因此存在收敛子列. 记它们的极限分别为  $Q$  和  $Z$ , 由矩阵乘积的连续性可知,  $Q$  和  $Z$  都是酉矩阵, 且  $Q^* B Z$  和  $Q^* A Z$  都是上三角矩阵.  $\square$

与实 Schur 分解类似, 当  $A, B$  都是实矩阵时, 我们有相应的 **广义实 Schur 分解**, 具体证明过程可参见相关资料.

**定理 4.11 (广义实 Schur 分解)** 设  $A, B \in \mathbb{R}^{n \times n}$ , 则存在正交矩阵  $Q, Z \in \mathbb{R}^{n \times n}$ , 使得

$$Q^T A Z = T_A, \quad Q^T B Z = T_B, \quad (4.13)$$

其中  $T_A, T_B \in \mathbb{R}^{n \times n}$  都是拟上三角矩阵.

(留作课外自习)

### 4.7.3 QZ 迭代法

**QZ 迭代法** 是用于计算  $(A, B)$  的广义 Schur 分解的方法, 是 QR 方法的自然推广, 本质上可以看作是将 QR 方法作用到矩阵  $AB^{-1}$  上, 在具体实施时需要做一些优化, 以提高执行效率. QZ 方法的详细推导和实现过程可参见相关资料, 如 [80, 114, 144].



## 4.8 应用

### 4.8.1 多项式求根

考虑  $n$  次多项式

$$p_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0. \quad (4.14)$$

这里假定  $a_i$  都是实数, 且  $a_n \neq 0$ . 由代数学基本定理可知,  $p_n(x)$  在复数域中有且仅有  $n$  的零点 (其中重根按重数计).

当  $n = 1$  时, 可直接求解. 当  $n = 2$  时, 可以通过求根公式求解. 当  $n = 3$  时, 也存在相应的求根公式, 但不那么直观, 需要一定的构造技巧才能导出. 当  $n = 4$  时, 可以从三次方程求根公式中导出相应的求根公式.

而当  $n \geq 5$  时, Abel 证明了不存在求根公式, 这意味着只能用数值方法 (迭代法) 求解. 从理论上讲, 任何求解非线性方程的迭代法都可以用来计算多项式的零点, 如著名的 Newton 法. 计算出一个零点后, 通过收缩技术, 将原问题转化为  $n - 1$  多项式零点问题, 然后继续用迭代法求解. 如此不断重复, 直到求出所有的零点. 但由于舍入误差等原因, 对于高次多项式, 随着计算过程的推进, 计算误差会越来越大, 最后的计算结果往往无法令人满意.

在 MATLAB 中, 命令 `roots` 可以计算出多项式的所有零点, 其使用的方法就是计算矩阵特征值的 QR 迭代法.

首先将多项式 (4.14) 转化为首项系数为 1 的多项式, 记为

$$q_n(x) = x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0.$$

多项式  $q_n(x)$  可以看作是某个  $n$  阶矩阵的特征值多项式, 如:

$$A = \begin{bmatrix} 0 & & & -c_0 \\ 1 & 0 & & -c_1 \\ & \ddots & \ddots & \vdots \\ & & 1 & -c_{n-1} \end{bmatrix}. \quad (4.15)$$

我们称矩阵  $A$  为  $q_n(x)$  的友矩阵. 这样, 计算多项式  $q_n(x)$  的零点问题就转化为计算  $A$  的特征值问题.

由于  $A$  已经是上 Hessenberg 矩阵, 因此隐式 QR 迭代法的第一步 (上 Hessenberg 化) 就不用做了. 虽然  $A$  上三角部分大多是零, 而且分布也很有规律, 但无论是单位移 QR 迭代还是双位移 QR 迭代, 经过一步迭代后, 这些零元素都会消失, 因此总运算量仍然是  $\mathcal{O}(n^3)$ .

于是, 如何利用  $A$  的特殊结构, 降低 QR 方法的运算量和存储量, 一直是颇受关注的问题. 最近, 陆续有学者 [5, 14, 24, 130] 提出了快速 QR 方法, 将运算量降为  $\mathcal{O}(n^2)$ , 存储量也降为  $\mathcal{O}(n)$ . 主要思想是将  $A$  写成一个酉矩阵与秩一矩阵之差:

$$A = \begin{bmatrix} 0 & & 1 \\ 1 & 0 & 0 \\ & \ddots & \ddots & \vdots \\ & & 1 & 0 \end{bmatrix} - \begin{bmatrix} c_0 + 1 \\ c_1 \\ \vdots \\ c_{n-1} \end{bmatrix} [0, 0, \dots, 1] \triangleq U - xy^T.$$



易知, 经过一次 QR 迭代后, 仍可以写成一个酉矩阵与秩一矩阵之差. 基于这种观察, 就可以设计出快速的 QR 迭代法, 详细过程可参见 [5, 14, 24, 130] 或 [144].



## 4.9 课后习题

练习 4.1 设  $A \in \mathbb{C}^{n \times n}$ . 证明:

- (1) 若  $A$  是上三角矩阵, 且  $A^*A = AA^*$ , 则  $A$  必定是对角矩阵;
- (2)  $A$  是正规矩阵的充要条件是  $A$  酉相似于一个对角矩阵;
- (3) 设  $\lambda_1, \lambda_2, \dots, \lambda_n$  是  $A$  的特征值, 则  $A$  是正规矩阵的充要条件是  $\sum_{i=1}^n |\lambda_i|^2 = \|A\|_F^2$ .

练习 4.2 设  $\lambda_1, \lambda_2 \in \mathbb{C}$  是  $A \in \mathbb{C}^{n \times n}$  的两个互不相同的特征值,  $x \in \mathbb{C}^n$  是  $\lambda_1$  的特征向量,  $y \in \mathbb{C}^n$  是  $\lambda_2$  的左特征向量. 证明:  $y^*x = 0$ .

练习 4.3\* 设  $A \in \mathbb{C}^{n \times n}$  可对角化, 其特征值为  $\lambda_1, \lambda_2, \dots, \lambda_n$ , 证明:

$$\sum_{i=1}^n |\lambda_i|^2 = \min_{\det(S) \neq 0} \|S^{-1}AS\|_F^2.$$

练习 4.4 设  $A = QR = UG$  是非奇异矩阵  $A \in \mathbb{C}^{n \times n}$  的两个 QR 分解, 其中  $Q, U \in \mathbb{C}^{n \times n}$  是酉矩阵,  $R, G \in \mathbb{C}^{n \times n}$  是上三角矩阵. 证明: 存在对角矩阵  $W = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_n) \in \mathbb{C}^{n \times n}$  满足  $|\alpha_i| = 1$ , 使得

$$Q = UW, \quad R = W^{-1}G.$$

练习 4.5 设  $H = [h_{ij}] \in \mathbb{R}^{n \times n}$  是上 Hessenberg 矩阵, 其 QR 分解为  $H = QR$ , 其中  $R = [r_{ij}] \in \mathbb{R}^{n \times n}$  是上三角矩阵且对角线元素均非负. 证明:

$$r_{kk} \geq |h_{k+1,k}|, \quad k = 1, 2, \dots, n-1.$$

因此, (1) 若  $H$  不可约, 则  $r_{kk} > 0, k = 1, 2, \dots, n-1$ ; (2) 若  $H$  不可约且奇异, 则  $r_{nn} = 0$ .  
(提示: 观察  $H$  的 QR 分解过程, 借助 Givens 变换)

练习 4.6\* (定理 4.6) 设  $A \in \mathbb{R}^{n \times n}$  是非奇异上 Hessenberg 矩阵且下次对角线元素均非零, 即  $a_{i+1,i} \neq 0, i = 1, 2, \dots, n-1$ . 设其 QR 分解为  $A = QR$ , 则  $\tilde{A} \triangleq RQ$  的下次对角线元素也都非零.

练习 4.7 用 Householder 变换, 通过相似变换将矩阵  $A$  化为上 Hessenberg 型, 其中

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & -1 & -1 & 3 \\ 2 & -4 & 7 & 3 \\ 1 & 4 & -3 & 6 \end{bmatrix}.$$

练习 4.8 考虑用 Householder 变换将矩阵上 Hessenberg 化的算法, 给出具体的乘法运算次数和加减运算次数.

练习 4.9\* 设  $A \in \mathbb{C}^{m \times m}, B \in \mathbb{C}^{n \times n}, C \in \mathbb{C}^{m \times n}$ , 考虑矩阵方程

$$AX - XB = C.$$

- (1) 证明: 当  $A$  和  $B$  没有共同特征值时, 矩阵方程存在唯一解.

(提示: 利用 Kronecker 积, 上述矩阵方程等价于  $(I_n \otimes A - B^T \otimes I_m) \text{vec}(X) = \text{vec}(C)$ , 其中  $\text{vec}(\cdot)$  表示将矩阵按列排列得到的向量)

- (2) 当  $A$  和  $B$  没有共同特征值时, 给出求解算法.



(提示: 利用 Schur 分解, 将  $A$  和  $B$  转化为相应的上三角矩阵)

..... 以下为可选题 .....

练习 4.10 设  $J = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} \in \mathbb{C}^{m \times m}$ , 计算其特征值  $\lambda$  对应的左、右特征向量.

练习 4.11 设

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ & A_{22} & \cdots & A_{2k} \\ & & \ddots & \vdots \\ & & & A_{kk} \end{bmatrix},$$

其中  $A_{ii}$  都是方阵. 证明:  $A$  的特征值即为对角块  $A_{11}, A_{22}, \dots, A_{kk}$  的特征值的并.

练习 4.12\* 设  $H \in \mathbb{R}^{n \times n}$  是不可约上 Hessenberg 矩阵, 证明: 存在对角矩阵  $D$  使得  $D^{-1}HD$  的次对角元素都为 1. 若  $H$  的次对角元素都小于 1 或都大于 1, 估计  $D$  的谱条件数  $\kappa_2(D)$ .

练习 4.13 证明矩阵

$$A = \begin{bmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & -c_{n-2} \\ 0 & 0 & \cdots & 1 & -c_{n-1} \end{bmatrix}$$

的特征多项式是

$$p(\lambda) = \lambda^n + c_{n-1}\lambda^{n-1} + \cdots + c_1\lambda + c_0.$$

并利用这个结果给出计算一个多项式所有零点的实用算法.

..... 以下为实践题 .....

练习 4.14 编写函数, 实现矩阵的上 Hessenberg 化, 即算法 4.7.

练习 4.15 编写函数, 实现带 Francis 位移的隐式 QR 迭代算法的单个迭代步, 即从  $A_k$  到  $A_{k+1}$ , 这里假定  $A_k$  是上 Hessenberg 矩阵, 且下次对角线上的元素都非零.

练习 4.16 编写函数, 实现计算上 Hessenberg 矩阵  $A$  的带 Francis 位移的隐式 QR 迭代算法. 只计算特征值, 收缩技术可以通过递归方式实现.

练习 4.17 编写函数, 实现计算友矩阵 (4.15) 的带 Francis 位移的快速 QR 迭代算法.




## 第五讲 对称特征值问题

设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵. 在计算  $A$  的特征值和特征值向量时, 我们可以充分利用  $A$  的对称结构, 一方面尽可能地减少运算量, 另一方面也能构造出更加快速高效的算法.

本讲主要介绍以下方法.

- **Jacobi 迭代法**: 较古老的方法, 收敛速度较慢, 但能达到很高的计算精度, 且非常适合并行.
- **Rayleigh 商迭代法**: 利用 Rayleigh 商作为动态位移的反迭代算法, 一般具有全局线性收敛和局部三次收敛性.
- **对称 QR 迭代法**: 针对对称矩阵的带位移隐式 QR 迭代算法. 如果只需计算一个对称三对角矩阵的所有特征值, 则该算法是目前最快的方法, 运算量为  $\mathcal{O}(n^2)$ . 如果需要计算所有的特征值和特征向量, 则运算量约为  $6n^3$ .
- **分而治之法**: 同时计算对称矩阵的特征值和特征向量的一种快速算法. 基本思想是将大矩阵分解成小矩阵, 然后利用递归思想求解, 是目前求解对称矩阵的所有特征值和特征向量的最快方法之一.
- **对分法和反迭代法**: 对分法主要用于求解对称矩阵在某个区间中的特征值, 反迭代法用于计算特征向量.

 除了 Jacobi 迭代和 Rayleigh 商迭代外, 其余算法都需要先将对称矩阵三对角化. 这个过程大约需花费  $\frac{4}{3}n^3$  的工作量, 如果需要计算特征向量的话, 则总运算量约为  $\frac{8}{3}n^3$ .

## 5.1 Jacobi 迭代法

该算法的基本思想是通过一系列的 **Jacobi 旋转**  $J_k$ , 将  $A$  正交相似于一个对角矩阵, 即

$$A^{(0)} = A, \quad A^{(k+1)} = J_k A^{(k)} J_k^T, \quad k = 0, 1, \dots,$$

且  $A^{(k)}$  收敛到一个对角矩阵, 其中  $J_k$  为 **Jacobi 旋转**, 通常选取  $J_k$  为 Givens 变换, 即

$$J_k = G(i_k, j_k, \theta_k) = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & \cos \theta_k & \sin \theta_k & \\ & & & -\sin \theta_k & \cos \theta_k & \\ & & & & & 1 & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{bmatrix}$$

易知, 在  $A^{(k)}$  两边分别左乘  $J_k$  和右乘  $J_k^T$  时, 只会修改  $A^{(k)}$  的第  $i_k$  和第  $j_k$  行, 以及第  $i_k$  和第  $j_k$  列.

由于  $A^{(k)}$  是对称矩阵, 由下面的引理可知, 通过选取适当的  $\theta_k$ , 可以将  $A^{(k)}(i_k, j_k)$  和  $A^{(k)}(j_k, i_k)$  同时化为 0.

**引理 5.1** 设  $A \in \mathbb{R}^{2 \times 2}$  是对称矩阵, 则存在 Givens 变换  $G \in \mathbb{R}^{2 \times 2}$ , 使得  $GAG^T$  为对角矩阵.

(板书)

**证明.** 设

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad G = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix},$$

则

$$\begin{aligned} GAG^T &= \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}^T \\ &= \begin{bmatrix} a \cos^2 \theta + c \sin^2 \theta + b \sin 2\theta & \frac{1}{2}(c-a) \sin 2\theta + b \cos 2\theta \\ \frac{1}{2}(c-a) \sin 2\theta + b \cos 2\theta & a \sin^2 \theta + c \cos^2 \theta - b \sin 2\theta \end{bmatrix} \end{aligned}$$

令  $\frac{1}{2}(c-a) \sin 2\theta + b \cos 2\theta = 0$ , 可得

$$\frac{a-c}{2b} = \cot 2\theta = \frac{1 - \tan^2 \theta}{2 \tan \theta}.$$

解得

$$\tan \theta = \frac{\text{sign}(\tau)}{|\tau| + \sqrt{1 + \tau^2}}, \quad \tau = \frac{a-c}{2b}.$$

故引理结论成立. □

为了使得  $A^{(k)}$  收敛到一个对角矩阵, 其非对角线元素必须趋向于 0. 记  $\text{off}(A)$  为所有非对角



线元素的平方和, 即

$$\text{off}(A) = \sum_{i \neq j} a_{ij}^2 = \|A\|_F^2 - \sum_{i=1}^n a_{ii}^2,$$

我们的目标就是使得  $\text{off}(A)$  尽快趋于 0.

**引理 5.2** 设  $A = [a_{ij}] \in \mathbb{R}^{n \times n}$  是对称矩阵,  $\hat{A} = [\hat{a}_{ij}] = JAJ^T$ ,  $J = G(i, j, \theta)$ , 其中  $\theta$  的选取使得  $\hat{a}_{ij} = \hat{a}_{ji} = 0$ , 则

$$\text{off}(\hat{A}) = \text{off}(A) - 2a_{ij}^2, \quad i \neq j.$$

(板书)

**证明.** 记  $A = [a_1, a_2, \dots, a_n]$ . 令  $\tilde{A} = JA = [\tilde{a}_{ij}]_{n \times n}$ . 由于  $J$  是正交阵, 故

$$\|Ja_k\|_2 = \|a_k\|_2, \quad k = 1, 2, \dots, n.$$

又  $J$  左乘  $a_k$  时, 只影响其第  $i$  和第  $j$  个元素的值, 故由  $\|Ja_i\|_2 = \|a_i\|_2$  和  $\|Ja_j\|_2 = \|a_j\|_2$  可得

$$\tilde{a}_{ii}^2 + \tilde{a}_{ji}^2 = a_{ii}^2 + a_{ji}^2, \quad \tilde{a}_{ij}^2 + \tilde{a}_{jj}^2 = a_{ij}^2 + a_{jj}^2. \quad (5.1)$$

同理, 由  $\hat{A} = \tilde{A}J^T$  可得

$$\hat{a}_{ii}^2 + \hat{a}_{ij}^2 = \tilde{a}_{ii}^2 + \tilde{a}_{ij}^2, \quad \hat{a}_{ji}^2 + \hat{a}_{jj}^2 = \tilde{a}_{ji}^2 + \tilde{a}_{jj}^2. \quad (5.2)$$

又  $\hat{a}_{ij} = \hat{a}_{ji} = 0$ , 故

$$\hat{a}_{ii}^2 + \hat{a}_{jj}^2 = a_{ii}^2 + a_{jj}^2 + a_{ij}^2 + a_{ji}^2 = a_{ii}^2 + a_{jj}^2 + 2a_{ij}^2.$$

由于  $JAJ^T$  只影响  $A$  的第  $i, j$  行和第  $i, j$  列, 故对角线元素中只有  $a_{ii}$  和  $a_{jj}$  受影响. 所以

$$\sum_{k=1}^n \hat{a}_{kk}^2 = \sum_{k=1}^n a_{kk}^2 + 2a_{ij}^2,$$

故

$$\text{off}(\hat{A}) = \|\hat{A}\|_2^2 - \sum_{k=1}^n \hat{a}_{kk}^2 = \|A\|_2^2 - \sum_{k=1}^n a_{kk}^2 - 2a_{ij}^2 = \text{off}(A) - 2a_{ij}^2,$$

即引理结论成立. □

由此可知,  $\text{off}(A^{(k)})$  总是不断减小的. 下面给出 Jacobi 迭代算法.

#### 算法 5.1. Jacobi 迭代算法

- 1: Given a symmetric matrix  $A \in \mathbb{R}^{n \times n}$
- 2: **if** eigenvectors are desired **then**
- 3:     set  $J = I$  and  $flag = 1$
- 4: **end if**
- 5: **while** not converge **do**
- 6:     choose an index pair  $(i, j)$  such that  $a_{ij} \neq 0$
- 7:      $\tau = (a_{ii} - a_{jj}) / (2a_{ij})$
- 8:      $t = \text{sign}(\tau) / (|\tau| + \sqrt{1 + \tau^2})$    % 计算  $\tan \theta$





```

9: $c = 1/\sqrt{1+t^2}$, $s = ct$ % 计算 $\cos \theta$ 和 $\sin \theta$
10: $A = G(i, j, \theta)AG(i, j, \theta)^\top$
11: if $flag = 1$ then
12: $J = G(i, j, \theta)J$
13: end if
14: end while

```

该算法涉及到  $a_{ij}$  的选取问题, 一种直观的选取方法就是使得  $a_{ij}$  为所有非对角线元素中绝对值最大的一个, 这就是**经典 Jacobi 迭代算法**.

### 算法 5.2. 经典 Jacobi 迭代算法

```

1: Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$
2: if eigenvectors are desired then
3: set $J = I$ and $flag = 1$
4: end if
5: while $\text{off}(A) > tol$ do
6: choose (i, j) such that $|a_{ij}| = \max_{k \neq l} |a_{kl}|$ % 选取绝对值最大的元素
7: $\tau = (a_{ii} - a_{jj})/(2a_{ij})$
8: $t = \text{sign}(\tau)/(|\tau| + \sqrt{1 + \tau^2})$
9: $c = 1/\sqrt{1+t^2}$, $s = ct$ % 计算 $\cos \theta$ 和 $\sin \theta$
10: $A = G(i, j, \theta)AG(i, j, \theta)^\top$
11: if $flag = 1$ then
12: $J = G(i, j, \theta)J$
13: end if
14: end while

```

可以证明, 经典 Jacobi 算法至少是线性收敛的.

**定理 5.3** 对于经典 Jacobi 算法 5.2, 有

$$\text{off}(A^{(k+1)}) \leq \left(1 - \frac{1}{N}\right) \text{off}(A^{(k)}), \quad N = \frac{n(n-1)}{2}.$$

故  $k$  步迭代后, 有

$$\text{off}(A^{(k)}) \leq \left(1 - \frac{1}{N}\right)^k \text{off}(A^{(0)}) = \left(1 - \frac{1}{N}\right)^k \text{off}(A).$$

(板书)

**证明.** 由于在经典 Jacobi 算法 5.2 中,  $|a_{ij}| = \max_{k \neq l} |a_{kl}|$ , 故  $\text{off}(A^{(k)}) \leq n(n+1) \left(a_{ij}^{(k)}\right)^2$ , 即

$$2 \left(a_{ij}^{(k)}\right)^2 \geq \frac{1}{N} \text{off}(A^{(k)}), \quad N = \frac{n(n-1)}{2}.$$



所以由引理 5.2 可知

$$\text{off}(A^{(k+1)}) = \text{off}(A^{(k)}) - \left(a_{ij}^{(k)}\right)^2 \leq \left(1 - \frac{1}{N}\right) \text{off}(A^{(k)}).$$

□

事实上, 经典 Jacobi 算法最终是 (渐进) 二次收敛的 [30, 96]

**定理 5.4** 经典 Jacobi 算法 5.2 是  $N$  步 (渐进) 二次收敛的, 即对足够大的  $k$ , 有

$$\text{off}(A^{(k+N)}) = O\left(\text{off}^2(A^{(k)})\right).$$

由于在经典 Jacobi 算法中, 每一步都要寻找绝对值最大的非对角元, 比较费时, 因此实用性较差. 我们可以通过逐行扫描来选取  $(i, j)$ , 这就是**循环 Jacobi 迭代算法**.

#### 算法 5.3. 循环 Jacobi 迭代算法 (逐行扫描)

```

1: Given a symmetric matrix $A \in \mathbb{R}^{n \times n}$
2: if eigenvectors are desired then
3: set $J = I$ and $flag = 1$
4: end if
5: while $\text{off}(A) > tol$ do
6: for $i = 1$ to $n - 1$ do
7: for $j = i + 1$ to n do
8: if $a_{ij} \neq 0$ then
9: $\tau = (a_{ii} - a_{jj}) / (2a_{ij})$
10: $t = \text{sign}(\tau) / (|\tau| + \sqrt{1 + \tau^2})$
11: $c = 1 / \sqrt{1 + t^2}$
12: $s = c \cdot t$
13: $A = G(i, j, \theta)^T A G(i, j, \theta)$
14: if $flag = 1$ then
15: $J = J \cdot G(i, j, \theta)$
16: end if
17: end if
18: end for
19: end for
20: end while

```

循环 Jacobi 也具有 (渐进) 二次收敛性 [128, page 270].

#### Jacobi 迭代法的优缺点

- 优点: 能够达到很高的计算精度 (特别是小特征值); 同时非常适合并行计算.



- 缺点: 计算速度较慢; 矩阵稀疏性得不到充分的利用.



## 5.2 Rayleigh 商迭代法

在反迭代算法中,  $x^{(k)}$  的 Rayleigh 商是特征值  $\lambda_k$  的近似, 因此我们可以把它作为位移, 于是就得到下面的 **Rayleigh 商迭代法**.

### 算法 5.4. Rayleigh 商迭代算法 (RQI, Rayleigh Quotient Iterations)

```

1: Given an initial guess $x^{(0)}$ with $\|x^{(0)}\|_2 = 1$
2: compute the Rayleigh quotient $\rho_0 = (x^{(0)}, Ax^{(0)})$
3: set $k = 1$
4: while not converge do
5: $\sigma = \rho_{k-1}$
6: $y^{(k)} = (A - \sigma I)^{-1}x^{(k-1)}$
7: $x^{(k)} = y^{(k)} / \|y^{(k)}\|_2$
8: $\rho_k = (x^{(k)}, Ax^{(k)})$
9: $k = k + 1$
10: end while

```

关于 Rayleigh 商迭代的收敛性, 我们有下面的结论.

**定理 5.5** 如果特征值是单重的, 则当误差足够小时, Rayleigh 商迭代法中每步迭代所得的正确数字的位数增至三倍, 即 Rayleigh 商迭代是局部三次收敛的.

(板书)

**证明.** 设  $A = Q\Lambda Q^T$ , 令  $\hat{x}^{(k)} = Q^T x^{(k)}$ , 则在 Rayleigh 商迭代算法中

$$\rho_k = (x^{(k)})^T A x^{(k)} = (\hat{x}^{(k)})^T Q^T A Q \hat{x}^{(k)} = (\hat{x}^{(k)})^T \Lambda \hat{x}^{(k)}.$$

令  $\hat{y}^{(k)} = Q^T y^{(k)}$ , 则

$$\hat{y}^{(k)} = Q^T (A - \rho_{k-1} I)^{-1} x^{(k)} = (Q^T A Q - \rho_{k-1} I)^{-1} \hat{x}^{(k-1)} = (\Lambda - \rho_{k-1} I)^{-1} \hat{x}^{(k-1)},$$

即, “以初始向量  $x^{(0)}$  对  $A$  做 Rayleigh 商迭代” 等价于 “以初始向量  $\hat{x}^{(0)}$  对  $\Lambda$  做 Rayleigh 商迭代”, 即它们有相同的收敛性. 因此, 不失一般性, 我们可以假定  $A = \Lambda$  为对角阵, 此时  $A$  的特征向量为  $e_i, i = 1, 2, \dots, n$ .

我们假定  $x^{(k)}$  收敛到  $e_1$ . 令  $d_k = x^{(k)} - e_1$ , 则  $\|d_k\|_2 \rightarrow 0$ . 为了证明算法具有局部三次收敛, 我们需要证明: 当  $\varepsilon_k = \|d_k\|_2$  充分小时, 有  $\varepsilon_{k+1} = \|d_{k+1}\|_2 = \|x^{(k+1)} - e_1\|_2 = \mathcal{O}(\varepsilon_k^3)$ .

我们注意到

$$1 = (x^{(k)})^T x^{(k)} = (e_1 + d_k)^T (e_1 + d_k) = 1 + 2d_k(1) + d_k^T d_k = 1 + 2d_k(1) + \varepsilon_k^2,$$

其中  $d_k(1)$  表示  $d_k$  的第一个元素. 故  $d_k(1) = -\varepsilon_k^2/2$ . 所以

$$\rho_k = (x^{(k)})^T \Lambda x^{(k)} = (e_1 + d_k)^T \Lambda (e_1 + d_k) = e_1^T \Lambda e_1 + 2e_1^T \Lambda d_k + d_k^T \Lambda d_k \triangleq \lambda_1 - \eta,$$

其中  $\eta = -(2e_1^T \Lambda d_k + d_k^T \Lambda d_k) = -2\lambda_1 d_k(1) - d_k^T \Lambda d_k = \lambda_1 \varepsilon_k^2 - d_k^T \Lambda d_k$ . 于是

$$|\eta| \leq |\lambda_1| \varepsilon_k^2 + \|\Lambda\|_2 \cdot \|d_k\|_2^2 \leq 2\|\Lambda\|_2 \varepsilon_k^2.$$



由 Rayleigh 商算法 5.4 可知

$$\begin{aligned}
 y^{(k+1)} &= (\Lambda - \rho_k I)^{-1} x^{(k)} \\
 &= \left[ \frac{x^{(k)}(1)}{\lambda_1 - \rho_k}, \frac{x^{(k)}(2)}{\lambda_2 - \rho_k}, \dots, \frac{x^{(k)}(n)}{\lambda_n - \rho_k} \right]^\top \\
 &= \left[ \frac{1 + d_k(1)}{\lambda_1 - \rho_k}, \frac{d_k(2)}{\lambda_2 - \rho_k}, \dots, \frac{d_k(n)}{\lambda_n - \rho_k} \right]^\top \\
 &= \left[ \frac{1 - \varepsilon_k^2/2}{\eta}, \frac{d_k(2)}{\lambda_2 - \lambda_1 + \eta}, \dots, \frac{d_k(n)}{\lambda_n - \lambda_1 + \eta} \right]^\top \\
 &= \frac{1 - \varepsilon_k^2/2}{\eta} \left[ 1, \frac{d_k(2)\eta}{(1 - \varepsilon_k^2/2)(\lambda_2 - \lambda_1 + \eta)}, \dots, \frac{d_k(n)\eta}{(1 - \varepsilon_k^2/2)(\lambda_n - \lambda_1 + \eta)} \right]^\top \\
 &\triangleq \frac{1 - \varepsilon_k^2/2}{\eta} \cdot (e_1 + \hat{d}_{k+1}).
 \end{aligned}$$

其中

$$\hat{d}_{k+1} = \left[ 0, \frac{d_k(2)\eta}{(1 - \varepsilon_k^2/2)(\lambda_2 - \lambda_1 + \eta)}, \dots, \frac{d_k(n)\eta}{(1 - \varepsilon_k^2/2)(\lambda_n - \lambda_1 + \eta)} \right]^\top.$$

因为  $\lambda_1$  是单重特征值, 所以

$$\text{gap}(\lambda_1, \Lambda) \triangleq \min_{i \neq 1} |\lambda_i - \lambda_1| > 0,$$

故对于  $i = 2, 3, \dots, n$ , 当  $\varepsilon_k$  足够小时有

$$|\lambda_i - \lambda_1 + \eta| \geq |\lambda_i - \lambda_1| - |\eta| \geq \text{gap}(\lambda_1, \Lambda) - |\eta| \geq \text{gap}(\lambda_1, \Lambda) - 2\|\Lambda\|_2 \varepsilon_k^2 > 0.$$

于是我们有

$$\|\hat{d}_{k+1}\|_2 \leq \frac{\|d_k\|_2 |\eta|}{(1 - \varepsilon_k^2/2)(\text{gap}(\lambda_1, \Lambda) - |\eta|)} \leq \frac{2\|\Lambda\|_2 \varepsilon_k^3}{(1 - \varepsilon_k^2/2)(\text{gap}(\lambda_1, \Lambda) - |\eta|)},$$

即  $\|\hat{d}_{k+1}\|_2 = \mathcal{O}(\varepsilon_k^3)$ . 又

$$1 - \|\hat{d}_{k+1}\|_2 \leq \|e_1 + \hat{d}_{k+1}\|_2 \leq 1 + \|\hat{d}_{k+1}\|_2,$$

即

$$\left| 1 - \|e_1 + \hat{d}_{k+1}\|_2 \right| \leq \|\hat{d}_{k+1}\|_2.$$

由于

$$x^{(k+1)} = \frac{y^{(k+1)}}{\|y^{(k+1)}\|_2} = \frac{e_1 + \hat{d}_{k+1}}{\|e_1 + \hat{d}_{k+1}\|_2},$$

所以

$$\begin{aligned}
 \|d_{k+1}\|_2 &= \|x^{(k+1)} - e_1\|_2 = \frac{\|(1 - \|e_1 + \hat{d}_{k+1}\|_2) e_1 + \hat{d}_{k+1}\|_2}{\|e_1 + \hat{d}_{k+1}\|_2} \\
 &\leq \frac{|1 - \|e_1 + \hat{d}_{k+1}\|_2| + \|\hat{d}_{k+1}\|_2}{\|e_1 + \hat{d}_{k+1}\|_2} \leq \frac{2\|\hat{d}_{k+1}\|_2}{\|e_1 + \hat{d}_{k+1}\|_2}.
 \end{aligned}$$

又  $\|\hat{d}_{k+1}\|_2 = \mathcal{O}(\varepsilon_k^3)$ , 故  $\varepsilon_{k+1} = \|d_{k+1}\|_2 = \mathcal{O}(\varepsilon_k^3)$ . □



🚩 RQI 算法具有局部三次收敛性, 但无法确定收敛到哪个特征向量 (特征值), 因此可以其他算法的加速手段, 即先使用其他算法 (比如幂迭代) 计算出所需特征值的近似值, 然后再使用 RQI 算法加速.

🚩 实际计算时, 判断  $(\rho_k, x^{(k)})$  是否收敛可以观察残量  $r_k = (A - \rho_k)x^{(k)}$  是否趋于零.

下面是关于 RQI 算法的全局收敛性, 可参见文献 [96].

**定理 5.6** 在 RQI 算法中, 设  $r_k = (A - \rho_k)x^{(k)}$ , 则有


$$\|r_{k+1}\| \leq \|r_k\|,$$

其中等号成立当且仅当  $\rho_{k+1} = \rho_k$  且  $x^{(k)}$  是  $(A - \rho_k)^2$  的特征向量.



### 5.3 对称 QR 迭代法

将带位移的隐式 QR 方法运用到对称矩阵, 就得到**对称 QR 迭代法**. 由于此时  $A$  是对称的, 所以上 Hessenberg 化后就转化为一个对称三对角矩阵, 相应的过程就称为**对称三对角化**.

 任何一个对称矩阵  $A \in \mathbb{R}^{n \times n}$  都可以通过正交变换转化成一个对称三对角矩阵  $T$ . 这个过程可以通过 Householder 变换或 Givens 变换来实现.

#### 对称 QR 迭代算法基本步骤

1. 对称三对角化: 利用 Householder 变换, 将  $A$  化为对称三对角矩阵, 即寻找正交矩阵  $Q$  使得  $T = QAQ^T$  为对称三对角矩阵;
2. 使用带(单)位移的隐式 QR 迭代算法计算  $T$  的特征值与特征值向量;
3. 计算  $A$  的特征向量.

#### 对称 QR 迭代算法的运算量


- 三对角化需要  $4n^3/3 + \mathcal{O}(n^2)$ , 如果需要计算特征向量, 则运算量为  $8n^3/3 + \mathcal{O}(n^2)$ ;
- 对  $T$  做带位移的隐式 QR 迭代, 每次迭代的运算量为  $6n$ ;
- 计算  $T$  的特征值时, 假定每个平均迭代 2 步, 则计算所有特征值的运算量为  $12n^2$ ; (每求出一个特征值后, 通过 deflation, 矩阵规模也会相应减小, 因此运算量可能会更少)
- 若要计算  $T$  的所有特征值和特征向量, 则运算量为  $6n^3 + \mathcal{O}(n^2)$ ;
- 若只要计算  $A$  的所有特征值, 运算量为  $4n^3/3 + \mathcal{O}(n^2)$ ;
- 若需要计算  $A$  的所有特征值和特征向量, 则运算量为  $26n^3/3 + \mathcal{O}(n^2)$ ;

#### 位移的选取 — Wilkinson 位移

位移的选取直接影响到算法的收敛速度. 我们可以通过下面的方式来选取位移. 设

$$A_k = \begin{bmatrix} a_1^{(k)} & b_1^{(k)} & & & \\ b_1^{(k)} & \ddots & \ddots & & \\ & \ddots & \ddots & b_{n-1}^{(k)} & \\ & & & b_{n-1}^{(k)} & a_n^{(k)} \end{bmatrix},$$

一种简单的位移选取策略就是令  $\sigma_k = a_n^{(k)}$ . 这种位移选取方法几乎对所有的矩阵都有三次渐进收敛速度, 但也存在不收敛的例子, 故我们需要对其做一些改进.

 事实上,  $a_n^{(k)}$  就是收敛到特征向量的迭代向量的 Rayleigh 商: 如果在 QR 迭代算法中使用位移  $\sigma_k = a_n^{(k)}$ , 而在 Rayleigh 商迭代算法 5.4 中取  $x_0 = e_n = [0, \dots, 0, 1]^T$ , 则利用 QR 迭代与反迭代之间的关系可以证明: QR 迭代算法中的  $\sigma_k$  与 Rayleigh 商迭代算法中的  $\rho_k$  相等.

一种有效的位移是 **Wilkinson 位移**, 即取子矩阵  $\begin{bmatrix} a_{n-1}^{(k)} & b_{n-1}^{(k)} \\ b_{n-1}^{(k)} & a_n^{(k)} \end{bmatrix}$  的最接近  $a_n^{(k)}$  的特征值作为

位移. 通过计算可得 Wilkinson 位移为

$$\sigma = a_n^{(k)} + \delta - \text{sign}(\delta) \sqrt{\delta^2 + \left(b_{n-1}^{(k)}\right)^2}, \quad \text{其中} \quad \delta = \frac{1}{2}(a_{n-1}^{(k)} - a_n^{(k)}).$$

出于稳定性方面的考虑, 我们通常用下面的计算公式

$$\sigma = a_n^{(k)} - \frac{\left(b_{n-1}^{(k)}\right)^2}{\delta + \text{sign}(\delta) \sqrt{\delta^2 + \left(b_{n-1}^{(k)}\right)^2}}.$$

**定理 5.7** [56, 96] 采用 Wilkinson 位移的 QR 迭代是整体收敛的, 且至少是线性收敛. 事实上, 几乎对所有的对称矩阵都是渐进三次收敛的.

**例 5.1** 带 Wilkinson 位移的隐式 QR 迭代算法收敛性演示.

(Eig\_TriQR.m)

```
T =
-1.1495e+00 1.9345e-01 0 0 0
 1.9345e-01 -5.7144e-01 -3.5163e+00 0 0
 0 -3.5163e+00 1.4138e+00 -1.2639e+00 0
 0 0 -1.2639e+00 -2.0125e-01 4.3216e+00
 0 0 0 4.3216e+00 1.9285e+00

iter = 1
T =
-1.1606e+00 2.0488e-01 0 0 0
 2.0488e-01 -3.1370e+00 -1.0240e+00 0 0
 0 -1.0240e+00 1.2439e+00 -3.5560e+00 0
 0 0 -3.5560e+00 -1.1053e+00 3.1938e-01
 0 0 0 3.1938e-01 5.5790e+00

iter = 2
T =
-1.1748e+00 2.6621e-01 0 0 0
 2.6621e-01 -3.3005e+00 -6.4187e-01 0 0
 0 -6.4187e-01 -3.1162e+00 -1.8413e+00 0
 0 0 -1.8413e+00 3.4052e+00 1.3658e-03
 0 0 0 1.3658e-03 5.6064e+00

iter = 3
T =
-1.1990e+00 3.4962e-01 0 0 0
 3.4962e-01 -3.3676e+00 -6.3718e-01 0 0
 0 -6.3718e-01 -3.4941e+00 -3.6853e-01 0
```





```

 0 0 -3.6853e-01 3.8743e+00 1.4108e-10
 0 0 0 1.4108e-10 5.6064e+00
iter = 4
T =
-1.2569e+00 4.9647e-01 0 0 0
 4.9647e-01 -3.4224e+00 -6.4282e-01 0 0
 0 -6.4282e-01 -3.3999e+00 -7.0835e-06 0
 0 0 -7.0835e-06 3.8929e+00 0
 0 0 0 0 5.6064e+00
iter = 5
T =
-1.3717e+00 6.9691e-01 0 0 0
 6.9691e-01 -3.4212e+00 -6.3796e-01 0 0
 0 -6.3796e-01 -3.2863e+00 -5.2850e-20 0
 0 0 -5.2850e-20 3.8929e+00 0
 0 0 0 0 5.6064e+00
iter = 6
T =
-1.2373e+00 -5.2803e-01 0 0 0
-5.2803e-01 -3.9958e+00 8.9840e-02 0 0
 0 8.9840e-02 -2.8461e+00 0 0
 0 0 0 3.8929e+00 0
 0 0 0 0 5.6064e+00
iter = 7
T =
-1.1937e+00 3.9668e-01 0 0 0
 3.9668e-01 -4.0455e+00 -6.4555e-05 0 0
 0 -6.4555e-05 -2.8400e+00 0 0
 0 0 0 3.8929e+00 0
 0 0 0 0 5.6064e+00
iter = 8
T =
-1.1695e+00 -2.9629e-01 0 0 0
-2.9629e-01 -4.0697e+00 1.2962e-14 0 0
 0 1.2962e-14 -2.8400e+00 0 0
 0 0 0 3.8929e+00 0
 0 0 0 0 5.6064e+00
iter = 9

```



$T =$ 

|             |             |             |            |            |
|-------------|-------------|-------------|------------|------------|
| -1.1396e+00 | -2.2223e-17 | 0           | 0          | 0          |
| -2.2223e-17 | -4.0996e+00 | 0           | 0          | 0          |
| 0           | 0           | -2.8400e+00 | 0          | 0          |
| 0           | 0           | 0           | 3.8929e+00 | 0          |
| 0           | 0           | 0           | 0          | 5.6064e+00 |



## 5.4 分而治之法

分而治之 (Divide-and-Conquer) 算法是由 Cuppen [28] 于 1981 年首次提出, 但直到 1995 年才出现稳定的实现方式 [64]. 该算法是目前计算维数大于 25 的矩阵的所有特征值和特征向量的最快算法. 下面我们介绍该算法.

考虑不可约对称三对角矩阵

$$\begin{aligned}
 T &= \left[ \begin{array}{ccc|ccc} a_1 & b_1 & & & & \\ b_1 & \ddots & & & & \\ & \ddots & a_{m-1} & b_{m-1} & & \\ & & b_{m-1} & a_m & & \\ \hline & & & b_m & a_{m+1} & b_{m+1} \\ & & & & b_{m+1} & \ddots \\ & & & & & \ddots \\ & & & & & & b_{n-1} \\ & & & & & & b_{n-1} & a_n \end{array} \right] \\
 &= \left[ \begin{array}{ccc|ccc} a_1 & b_1 & & & & \\ b_1 & \ddots & & & & \\ & \ddots & a_{m-1} & b_{m-1} & & \\ & & b_{m-1} & a_m - b_m & & \\ \hline & & & & a_{m+1} - b_m & b_{m+1} \\ & & & & b_{m+1} & \ddots \\ & & & & & \ddots \\ & & & & & & b_{n-1} \\ & & & & & & b_{n-1} & a_n \end{array} \right] + \left[ \begin{array}{ccc|ccc} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ \hline & & & b_m & & \\ & & & b_m & b_m & \\ & & & & & \end{array} \right] \\
 &= \left[ \begin{array}{c|c} T_1 & 0 \\ \hline 0 & T_2 \end{array} \right] + b_m vv^T,
 \end{aligned}$$

其中  $v = [0, \dots, 0, 1, 1, 0, \dots, 0]^T$ . 假定  $T_1$  和  $T_2$  的特征值分解已经计算出来了, 即  $T_1 = Q_1 \Lambda_1 Q_1^T$ ,  $T_2 = Q_2 \Lambda_2 Q_2^T$ , 下面考虑  $T$  的特征值分解.

首先介绍一个引理.

**引理 5.8** 设  $x, y \in \mathbb{R}^n$ , 则  $\det(I + xy^T) = 1 + y^T x$ .

(留作练习)

我们首先考虑  $T$  的特征值与  $T_1$  和  $T_2$  的特征值之间的关系.

$$\begin{aligned}
 T &= \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix} + b_m vv^T \\
 &= \begin{bmatrix} Q_1 \Lambda_1 Q_1^T & 0 \\ 0 & Q_2 \Lambda_2 Q_2^T \end{bmatrix} + b_m vv^T \\
 &= \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix} \left( \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} + b_m uu^T \right) \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix}^T,
 \end{aligned}$$



其中

$$u = \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix}^T, \quad v = \begin{bmatrix} Q_1^T \text{ 的最后一列} \\ Q_2^T \text{ 的第一列} \end{bmatrix}.$$

令  $\alpha = b_m$ ,  $D = \text{diag}(\Lambda_1, \Lambda_2) = \text{diag}(d_1, d_2, \dots, d_n)$ , 并假定  $d_1 \geq d_2 \geq \dots \geq d_n$ . 则  $T$  的特征值与  $D + \alpha uu^T$  的特征值相同.

下面计算  $D + \alpha uu^T$  的特征值. 设  $\lambda$  是  $D + \alpha uu^T$  的一个特征值, 若  $D - \lambda I$  非奇异, 则

$$\det(D + \alpha uu^T - \lambda I) = \det(D - \lambda I) \cdot \det(I + \alpha(D - \lambda I)^{-1}uu^T).$$

故  $\det(I + \alpha(D - \lambda I)^{-1}uu^T) = 0$ . 又由引理 5.8 可知

$$\det(I + \alpha(D - \lambda I)^{-1}uu^T) = 1 + \alpha u^T(D - \lambda I)^{-1}u = 1 + \alpha \sum_{i=1}^n \frac{u_i^2}{d_i - \lambda} \triangleq f(\lambda).$$

故求  $A$  的特征值等价于求**特征方程 (secular equation)**  $f(\lambda) = 0$  的根. 由于

$$f'(\lambda) = \alpha \sum_{i=1}^n \frac{u_i^2}{(d_i - \lambda)^2},$$

当所有的  $d_i$  都互不相同, 且所有的  $u_i$  都不为零时,  $f(\lambda)$  在  $\lambda \neq d_i$  处都是严格单调的 (见下图).

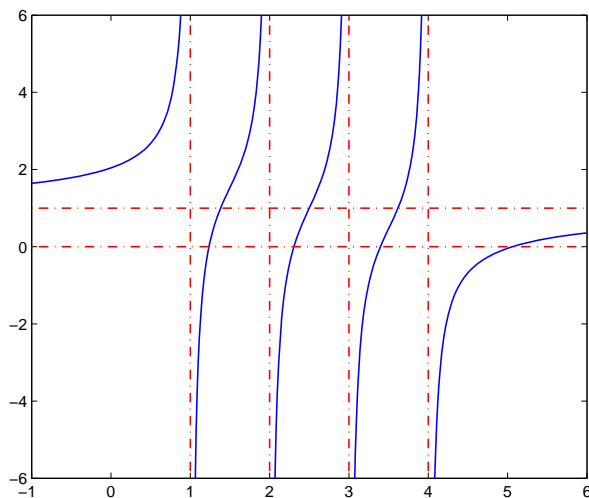


图 5.1.  $f(\lambda) = 1 + 0.5 \left( \frac{1}{4-\lambda} + \frac{1}{3-\lambda} + \frac{1}{2-\lambda} + \frac{1}{1-\lambda} \right)$  的图像

所以  $f(\lambda)$  在每个区间  $(d_{i+1}, d_i)$  内都有一个根, 共  $n - 1$  个, 另一个根在  $(d_1, \infty)$  (若  $\alpha > 0$ ) 或  $(-\infty, d_n)$  (若  $\alpha < 0$ ) 中. 由于  $f(\lambda)$  在每个区间  $(d_{i+1}, d_i)$  内光滑且严格单调递增 ( $\alpha > 0$ ) 或递减 ( $\alpha < 0$ ), 所以在实际计算中, 可以使用对分法, 牛顿类方法, 或有理逼近等算法来求解. 通常都能很快收敛, 一般只需迭代几步即可. 因此, 计算一个特征值的运算量约为  $\mathcal{O}(n)$ , 计算  $D + \alpha uu^T$  的所有特征值的运算量约为  $\mathcal{O}(n^2)$ .

当所有特征值计算出来后, 我们可以利用下面的引理来计算特征向量.

**引理 5.9** 设  $D \in \mathbb{R}^{n \times n}$  为对角矩阵,  $u \in \mathbb{R}^n$ ,  $\alpha \in \mathbb{R}$ , 若  $\lambda$  是  $D + \alpha uu^T$  的特征值, 且  $\lambda \neq d_i$ ,  $i = 1, 2, \dots, n$ , 则  $(D - \lambda I)^{-1}u$  是其对应的特征向量.



(板书)

**证明.** 由引理 5.8 可知

$$\begin{aligned} 0 &= \det(D + \alpha uu^T - \lambda I) = \det(D - \lambda I) \cdot \det(I + \alpha(D - \lambda I)^{-1}uu^T) \\ &= \det(D - \lambda I) \cdot (1 + \alpha u^T(D - \lambda I)^{-1}u), \end{aligned}$$

故  $1 + \alpha u^T(D - \lambda I)^{-1}u = 0$ , 即  $\alpha u^T(D - \lambda I)^{-1}u = -1$ . 直接计算可得

$$\begin{aligned} (D + \alpha uu^T)((D - \lambda I)^{-1}u) &= (D - \lambda I + \lambda I + \alpha uu^T)(D - \lambda I)^{-1}u \\ &= u + \lambda(D - \lambda I)^{-1}u + (\alpha u^T(D - \lambda I)^{-1}u)u \\ &= u + \lambda(D - \lambda I)^{-1}u - u \\ &= \lambda(D - \lambda I)^{-1}u, \end{aligned}$$


即引理结论成立. □

**算法 5.5.** 计算对称三对角矩阵的特征值和特征向量的分而治之法 (函数形式)

```

1: function [Q, Λ] = dc_eig(T) % T = QΛQT
2: if T is of 1 × 1 then
3: Q = 1, Λ = T
4: return
5: end if
6: form T = $\begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix} + b_m vv^T$
7: [Q1, Λ1] = dc_eig(T1)
8: [Q2, Λ2] = dc_eig(T2)
9: form D + αuuT from Λ1, Λ2, Q1, Q2
10: compute the eigenvalues Λ and eigenvectors \hat{Q} of D + αuuT
11: compute the eigenvectors of T with $Q = \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix} \cdot \hat{Q}$
12: end

```

 在分而治之法中, 特征值和特征向量是同时计算的.

### 实施细节

在实际使用分而治之算法时, 我们需要考虑以下几个细节问题:


- (1) 如何减小运算量;
- (2) 如何求解特征方程  $f(\lambda) = 0$ ;
- (3) 如何稳定地计算特征向量.



**(1) 如何减小运算量 — 收缩技巧 (deflation)**

分而治之算法的计算复杂性分析: 用  $t(n)$  表示对  $n$  阶矩阵调用函数 `dc_eig` 的运算量, 则

$$\begin{aligned} t(n) &= 2t(n/2) && \text{递归调用 dc\_eig 两次} \\ &+ \mathcal{O}(n^2) && \text{计算 } D + \alpha uu^T \text{ 的特征值和特征向量} \\ &+ c \cdot n^3 && \text{计算 } Q. \end{aligned}$$

 如果计算  $Q$  时使用的是稠密矩阵乘法, 则  $c = 2$ ;  
若不计  $\mathcal{O}(n^2)$  项, 则由递归公式  $t(n) = 2t(n/2) + c \cdot n^3$  可得  $t(n) \approx c \cdot 4n^3/3$ .

但事实上, 由于**收缩 (deflation)**现象的存在, 常数  $c$  通常比 1 小得多.

在前面的算法描述过程中, 我们假定  $d_i$  互不相等且  $u_i$  不能为零. 事实上, 容易证明当  $d_i = d_{i+1}$  或  $u_i = 0$  时,  $d_i$  即为  $D + \alpha uu^T$  的特征值, 这种现象我们称为**收缩 (deflation)**. 在实际计算时, 当  $d_i - d_{i+1}$  或  $|u_i|$  小于一个给定的阈值时, 我们就近似认为  $d_i$  为  $D + \alpha uu^T$  的特征值, 即出现收缩现象.

在实际计算中, 收缩现象会经常发生, 而且非常频繁, 所以我们可以而且应该利用这种优点加快分而治之算法的速度 [28, 106].

由于主要的计算量集中在计算  $Q$ , 即算法最后一步的矩阵乘积. 如果  $u_i = 0$ , 则  $d_i$  为特征值, 其对应的特征向量为  $e_i$ , 即  $\hat{Q}$  的第  $i$  列为  $e_i$ , 故计算  $Q$  的第  $i$  列时不需要做任何的计算. 当  $d_i = d_{i+1}$  时, 也存在一个类似的简化.

**(2) 特征方程求解**

通常我们可以使用牛顿法来计算特征方程  $f(\lambda) = 0$  的解.

当  $d_i \neq d_{i+1}$  且  $u_i \neq 0$  时, 我们用牛顿法计算  $f(\lambda)$  在  $(d_{i+1}, d_i)$  中的零点  $\lambda_i$ . 如果  $|u_i|$  小于给定的阈值时, 我们可直接将  $d_i$  作为特征值  $\lambda_i$  的一个近似. 但当  $u_i$  很小 (却大于给定的阈值) 时, 此时  $f(\lambda)$  在区间  $[d_{i+1}, d_i]$  中的大部分处的斜率几乎为 0 (见图 5.2). 这时, 如果任取  $[d_{i+1}, d_i]$  中的一个点作为迭代初始点, 经过一次牛顿迭代后, 迭代解可能会跑到区间  $[d_{i+1}, d_i]$  的外面, 造成不收敛.



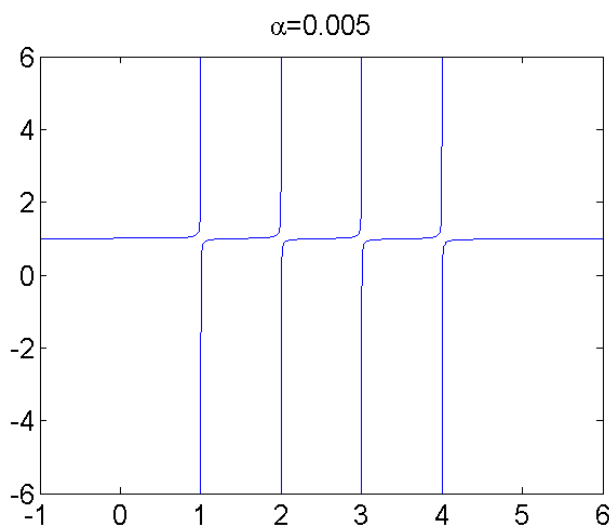


图 5.2.  $f(\lambda) = 1 + 0.005 \left( \frac{1}{4-\lambda} + \frac{1}{3-\lambda} + \frac{1}{2-\lambda} + \frac{1}{1-\lambda} \right)$  的图像

这时需要采用修正的牛顿法. 假设我们已经计算出  $\lambda_i$  的一个近似  $\tilde{\lambda}$ , 下面我们需要从  $\tilde{\lambda}$  出发, 利用牛顿迭代计算下一个近似, 直至收敛. 我们知道牛顿法的基本原理是使用  $f(\lambda)$  在点  $\tilde{\lambda}$  的切线来近似  $f(\lambda)$ , 并将切线的零点作为下一个近似, 即用切线来近似曲线  $f(\lambda)$ .

当  $u_i$  很小时, 这种近似方法可能会出现問題. 这时, 我们需要寻求用其它简单函数  $h(\lambda)$  (不一定是直线) 来近似  $f(\lambda)$ , 然后用  $h(\lambda)$  的零点作为  $f(\lambda)$  零点的近似, 并不断迭代下去, 直至收敛.

关于  $h(\lambda)$  的选取, 通常需要满足以下几个要求: (1)  $h(\lambda)$  必须容易构造; (2)  $h(\lambda)$  的零点容易计算; (3)  $h(\lambda)$  尽可能地与  $f(\lambda)$  相近.

当然, 这样的  $h(\lambda)$  有多种不同的构造方法. 这里介绍其中的一种方法. 假定我们需要计算的是  $f(\lambda)$  在  $(d_{i+1}, d_i)$  中的零点  $\lambda_i$ . 因为  $d_i$  和  $d_{i+1}$  是  $f(\lambda)$  的奇点, 所以我们令

$$h(\lambda) = \frac{c_1}{d_i - \lambda} + \frac{c_2}{d_{i+1} - \lambda} + c_3,$$

其中  $c_1, c_2, c_3$  是待定的参数. 显然,  $h(\lambda)$  的零点很容易计算, 只需求解一个一元二次方程即可, 运算量与牛顿法相差无几. 选取参数  $c_1, c_2, c_3$  的基本原则是使得  $h(\lambda)$  在  $\tilde{\lambda}$  附近尽可能地接近  $f(\lambda)$ . 又  $f(\lambda)$  可写为

$$f(\lambda) = 1 + \alpha \sum_{k=1}^n \frac{u_k^2}{d_k - \lambda} = 1 + \alpha \left( \sum_{k=1}^i \frac{u_k^2}{d_k - \lambda} + \sum_{k=i+1}^n \frac{u_k^2}{d_k - \lambda} \right) \triangleq 1 + \alpha (\Psi_1(\lambda) + \Psi_2(\lambda)).$$

当  $\lambda \in (d_{i+1}, d_i)$  时,  $\Psi_1(\lambda)$  为所有正项的和,  $\Psi_2(\lambda)$  为所有负项的和, 因此它们都可以较精确地计算. 但如果把它们加在一起时可能会引起对消, 从而可能失去相对精度. 因此我们将  $h(\lambda)$  相应地写成

$$h(\lambda) = 1 + \alpha (h_1(\lambda) + h_2(\lambda)),$$

其中

$$h_1(\lambda) = \frac{c_1}{d_i - \lambda} + \hat{c}_1, \quad h_2(\lambda) = \frac{c_2}{d_{i+1} - \lambda} + \hat{c}_2$$

满足

$$\begin{aligned} h_1(\tilde{\lambda}) &= \Psi_1(\tilde{\lambda}), & h'_1(\tilde{\lambda}) &= \Psi'_1(\tilde{\lambda}), \\ h_2(\tilde{\lambda}) &= \Psi_2(\tilde{\lambda}), & h'_2(\tilde{\lambda}) &= \Psi'_2(\tilde{\lambda}). \end{aligned}$$

即  $h_1(\lambda)$  和  $h_2(\lambda)$  分别是  $\Psi_1(\lambda)$  和  $\Psi_2(\lambda)$  的一次 Hermite 插值函数. 通过直接计算可得

$$\begin{cases} c_1 = \Psi'_1(\tilde{\lambda})(d_i - \tilde{\lambda})^2, & \hat{c}_1 = \Psi_1(\tilde{\lambda}) - \Psi'_1(\tilde{\lambda})(d_i - \tilde{\lambda}), \\ c_2 = \Psi'_2(\tilde{\lambda})(d_{i+1} - \tilde{\lambda})^2, & \hat{c}_2 = \Psi_2(\tilde{\lambda}) - \Psi'_2(\tilde{\lambda})(d_{i+1} - \tilde{\lambda}). \end{cases} \quad (5.3)$$

所以

$$h(\lambda) = 1 + \alpha(\hat{c}_1 + \hat{c}_2) + \alpha \left( \frac{c_1}{d_i - \lambda} + \frac{c_2}{d_{i+1} - \lambda} \right). \quad (5.4)$$

这就是迭代函数.

#### 算法 5.6. 修正的 Newton 算法

- 1: set  $k = 0$
- 2: choose an initial guess  $\lambda_0 \in [d_{i+1}, d_i]$
- 3: **while** not convergence **do**
- 4:   let  $\tilde{\lambda} = \lambda_k$  and compute  $c_1, c_2, \hat{c}_1, \hat{c}_2$  from (5.3)
- 5:   set  $k = k + 1$
- 6:   compute the solution  $\lambda_k$  of  $h(\lambda)$  defined by (5.4)
- 7: **end while**

### (3) 计算特征向量的稳定算法

设  $\lambda_i$  是  $D + \alpha uu^T$  的特征值, 则根据引理 5.9, 可利用公式  $(D - \lambda_i I)^{-1}u$  来计算其对应的特征向量. 但遗憾的是, 当相邻的两个特征值非常接近时, 这个公式可能不稳定. 如果  $\lambda_i$  与  $\lambda_{i+1}$  非常接近, 则它们都很接近  $d_{i+1}$  (这里假定  $\lambda_i \in (d_{i+1}, d_i)$ ,  $\lambda_{i+1} \in (d_{i+2}, d_{i+1})$ ), 此时计算  $d_{i+1} - \lambda_i$  和  $d_{i+1} - \lambda_{i+1}$  可能会存在对消情形, 从而有可能会损失有效数字, 产生较大的相对误差. 这样就导致  $(D - \lambda_i I)^{-1}u$  与  $(D - \lambda_{i+1} I)^{-1}u$  的计算精度下降, 从而使得特征向量之间的正交性也因此而失去.

下面的定理可以解决这个问题. 详情参见 [62, 139].

**定理 5.10 (Löwner)** 设对角阵  $D = \text{diag}(d_1, d_2, \dots, d_n)$  满足  $d_1 > d_2 > \dots > d_n$ . 若矩阵  $\hat{D} = D + \hat{u}\hat{u}^T$  的特征值  $\lambda_1, \lambda_2, \dots, \lambda_n$  满足交错性质

$$\lambda_1 > d_1 > \lambda_2 > d_2 > \dots > \lambda_n > d_n, \quad (5.5)$$

则向量  $\hat{u}$  的分量满足

$$|\hat{u}_i| = \left( \frac{\prod_{k=1}^n (\lambda_k - d_i)}{\prod_{k=1, k \neq i}^n (d_k - d_i)} \right)^{1/2}. \quad (5.6)$$

(留作课外自习)





**证明.** 由引理 5.8 可知  $\hat{D}$  的特征多项式为 (假设  $\lambda \neq d_i, i = 1, 2, \dots, n$ )

$$\begin{aligned}
 p(\lambda) &= \det(\hat{D} - \lambda I) = \det(D - \lambda I + \hat{u}\hat{u}^\top) \\
 &= \det(D - \lambda I) \cdot \det(I + (D - \lambda I)^{-1}\hat{u}\hat{u}^\top) \\
 &= \det(D - \lambda I) \cdot (1 + \hat{u}^\top(D - \lambda I)^{-1}\hat{u}) \\
 &= \left( \prod_{k=1}^n (d_k - \lambda) \right) \cdot \left( 1 + \sum_{i=1}^n \frac{\hat{u}_i^2}{d_i - \lambda} \right) \\
 &= \prod_{k=1}^n (d_k - \lambda) + \sum_{i=1}^n \left( \prod_{k=1, k \neq i}^n (d_k - \lambda) \hat{u}_i^2 \right). \quad (5.7)
 \end{aligned}$$

由于等式 (5.7) 两边都是关于  $\lambda$  的连续函数, 所以当  $\lambda = d_i$  时, 等式 (5.7) 仍然成立.

又  $\lambda_1, \lambda_2, \dots, \lambda_n$  是  $\hat{D}$  的特征值, 所以特征多项式也可以写成  $p(\lambda) = \prod_{k=1}^n (\lambda_k - \lambda)$ , 即

$$\det(\hat{D} - \lambda I) = \prod_{k=1}^n (\lambda_k - \lambda).$$

取  $\lambda = d_i$ , 则由  $\det(\hat{D} - \lambda I)$  的两个表达式可得


$$\prod_{k=1, k \neq i}^n (d_k - d_i) \hat{u}_i^2 = \prod_{k=1}^n (\lambda_k - d_i),$$

即

$$\hat{u}_i^2 = \frac{\prod_{k=1}^n (\lambda_k - d_i)}{\prod_{k=1, k \neq i}^n (d_k - d_i)}.$$

由交错性质可知, 上式右边是正的, 故定理结论成立.  $\square$

设  $\lambda_1, \lambda_2, \dots, \lambda_n$  是  $D + \alpha uu^\top$  的特征值, 且满足交错性质 (5.5). 假定  $\alpha > 0$ , 则根据定理 5.10 可知, 向量  $\sqrt{\alpha}u$  的分量满足 (5.6).

 **思考:** 思考: 如果  $\alpha < 0$ , 结论会是怎样?

因此, 我们可以采用公式 (5.6) 来计算特征向量. 这样就尽可能地避免了出现分母很小的情形.

下面是计算矩阵  $D + uu^\top$  的特征值和特征向量的稳定算法.

**算法 5.7.** 计算矩阵  $D + uu^\top$  的特征值和特征向量的稳定算法

- 1: Compute the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_n$  by solving  $f(\lambda) = 0$
- 2: Compute  $\hat{u}_i$  by Löwner Theorem so that  $\lambda_1, \lambda_2, \dots, \lambda_n$  are the exact eigenvalues of  $D + \hat{u}\hat{u}^\top$
- 3: Compute the eigenvectors of  $D + \hat{u}\hat{u}^\top$  by Lemma 5.9.

通过分析可以说明上述算法计算出来的  $D + \hat{u}\hat{u}^\top$  的特征值和特征值向量是非常精确的, 这意味着该算法是数值稳定的. 同时,  $D + \hat{u}\hat{u}^\top$  与原矩阵  $D + uu^\top$  具有相同的特征值和特征向量, 见习题 5.3.



### 箭型分而治之法

分而治之算法于 1981 年被首次提出,但直到 1995 年才由 Gu 和 Eisenstat 给出了一种快速稳定的实现方式,称为**箭型分而治之法** (**Arrowhead Divide-and-Conquer**, ADC). 他们做了大量的数值试验,在试验中,当矩阵规模不超过 6 时,就采用对称 QR 迭代来计算特征值和特征向量. 在对特征方程求解时,他们采用的是修正的有理逼近法. 数值结果表明,ADC 算法的计算精度可以与其他算法媲美,而计算速度通常比对称 QR 迭代快 5 至 10 倍,比 Cuppen 的分而治之法快 2 倍. 详细介绍见 [63, 64].



## 5.5 对分法和反迭代法

对分法 (Bisection) 的基本思想是利用惯性定理来计算所需的部分特征值.

**定义 5.1** 设  $A$  为对称矩阵, 则其惯性定义为

$$\text{Inertia}(A) = (\nu, \zeta, \pi)$$

其中  $\nu, \zeta, \pi$  分别表示  $A$  的负特征值, 零特征值和正特征值的个数.

**定理 5.11 (Sylvester 惯性定理)** 设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵,  $X \in \mathbb{R}^{n \times n}$  非奇异, 则  $X^T A X$  与  $A$  有相同的惯性.

利用 LU 分解可得  $A - zI = LDL^T$ , 其中  $L$  为非奇异下三角矩阵,  $D$  为对角阵, 则

$$\text{Inertia}(A - zI) = \text{Inertia}(D).$$

由于  $D$  是对角矩阵, 所以  $\text{Inertia}(D)$  很容易计算.

设  $\alpha \in \mathbb{R}$ , 记  $\text{Negcount}(A, \alpha)$  为小于  $\alpha$  的  $A$  的特征值的个数, 即

$$\text{Negcount}(A, \alpha) = \#(\lambda(A) < \alpha).$$

设  $\alpha_1 < \alpha_2$ , 则  $A$  在区间  $[\alpha_1, \alpha_2)$  中的特征值个数为

$$\text{Negcount}(A, \alpha_2) - \text{Negcount}(A, \alpha_1).$$

如果  $\alpha_2 - \alpha_1 < \text{tol}$  (其中  $\text{tol} \ll 1$  为事先给定的阈值), 且  $A$  在  $[\alpha_1, \alpha_2)$  中有特征值, 则我们可将  $[\alpha_1, \alpha_2)$  中的任意一个值作为  $A$  在该区间中的特征值的近似.

由此我们可以给出下面的对分法.

**算法 5.8.** 对分法: 计算  $A$  在  $[a, b)$  中的所有特征值

```

1: Let tol be a given threshold
2: compute $n_a = \text{Negcount}(A, a)$
3: compute $n_b = \text{Negcount}(A, b)$
4: if $n_a = n_b$ then
5: return % 此时 $[a, b)$ 中没有 A 的特征值
6: end if
7: put (a, n_a, b, n_b) onto worklist
8: % worklist 中的元素是“四元素对”, 即由四个数组成的数对
9: while worklist not empty do
10: remove $(\text{low}, n_{\text{low}}, \text{up}, n_{\text{up}})$ from the worklist
11: % $(\text{low}, n_{\text{low}}, \text{up}, n_{\text{up}})$ 是 worklist 中的任意一个元素
12: if $(\text{up} - \text{low}) < \text{tol}$ then
13: print "There are $n_{\text{up}} - n_{\text{low}}$ eigenvalues in $[\text{low}, \text{up})$ "
14: else

```



```

15: compute $mid = (low + up)/2$
16: compute $n_{mid} = \text{Negcount}(A, mid)$
17: if ($n_{mid} > n_{low}$) then
18: put ($low, n_{low}, mid, n_{mid}$) onto worklist
19: end if
20: if ($n_{up} > n_{mid}$) then
21: put (mid, n_{mid}, up, n_{up}) onto worklist
22: end if
23: end if
24: end while

```

显然, 对分法的主要运算量集中在计算  $\text{Negcount}(A, z)$ . 通常是事先将  $A$  转化成对称三对角矩阵, 这样计算  $A - zI$  的  $LDL^T$  分解就非常简单:

$$\begin{aligned}
 A - zI &= \begin{bmatrix} a_1 - z & b_1 & & \\ b_1 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ & & b_{n-1} & a_n - z \end{bmatrix} \\
 &= \begin{bmatrix} 1 & & & \\ l_1 & \ddots & & \\ & \ddots & \ddots & \\ & & l_{n-1} & 1 \end{bmatrix} \begin{bmatrix} d_1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & d_n \end{bmatrix} \begin{bmatrix} 1 & l_1 & & \\ & \ddots & \ddots & \\ & & \ddots & l_{n-1} \\ & & & 1 \end{bmatrix} \triangleq LDL^T.
 \end{aligned}$$

利用待定系数法, 可以得到下面的递推公式

$$d_1 = a_1 - z, \quad d_i = (a_i - z) - \frac{b_{i-1}^2}{d_{i-1}}, \quad i = 2, 3, \dots, n. \quad (5.8)$$


用上面的公式计算  $d_i$  的运算量约为  $4n$ .

注意这里没有选主元, 但针对对称三对角矩阵, 该算法是非常稳定的, 即使当  $d_i$  有可能很小时, 算法依然很稳定.

**定理 5.12** [30] 利用公式 (5.8) 计算所得的  $d_i$  与精确计算  $\hat{A}$  的  $\hat{d}_i$  有相同的符号, 故有相同的惯性. 这里  $\hat{A}$  与  $A$  非常接近, 即

$$\hat{A}(i, i) = a_i, \quad \hat{A}(i, i+1) = b_i(1 + \varepsilon_i),$$

其中  $|\varepsilon_i| \leq 2.5\varepsilon + \mathcal{O}(\varepsilon^2)$ , 这里的  $\varepsilon$  为机器精度.

 由于单独调用一次  $\text{Negcount}$  的运算量为  $4n$ , 故计算  $k$  个特征值的总运算量约为  $\mathcal{O}(kn)$ .

当特征值计算出来后, 我们可以使用带位移的逆迭代来计算对应的特征向量. 通常只需迭代 1 至 2 次即可, 由于  $A$  是三对角矩阵, 故计算每个特征向量的运算量为  $\mathcal{O}(n)$ . 整个合起来就构成 **对分法和逆迭代**.



当特征值紧靠在一起时, 计算出来的特征向量可能会失去正交性, 此时需要进行再正交化, 可通过 MGS 的 QR 分解来实现.



## 5.6 奇异值分解

奇异值分解 (SVD) 具有十分广泛的应用背景, 因此, 如何更好更快地计算一个给定矩阵的 SVD 是科学与工程计算领域中的一个热门研究课题, 吸引了众多专家进行这方面的研究, 也涌现出了许多奇妙的方法. 本章主要介绍计算 SVD 的常用算法.

对任意矩阵  $A \in \mathbb{R}^{m \times n}$ , 其奇异值与对称矩阵  $A^T A$ ,  $AA^T$  和  $\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$  的特征值是密切相关的, 故理论上计算对称特征值的算法都可以用于计算奇异值. 但在实际计算中, 我们通常可以利用 SVD 的特殊结构使得算法更加有效和准确.

与计算对称矩阵的特征值类似, 计算一个矩阵  $A$  的奇异值分解的算法通常分为以下几个步骤 (Jacobi 算法除外):

1. 将  $A$  二对角化:  $B = U_1^T A V_1$ , 其中  $B$  为上二对角矩阵,  $U_1, V_1$  为正交阵;
2. 计算  $B$  的 SVD:  $B = U_2 \Sigma V_2^T$ , 其中  $\Sigma$  为对角阵,  $U_2, V_2$  为正交阵;
3. 合并得到  $A$  的 SVD:  $A = U_1 B V_1^T = (U_1 U_2) \Sigma (V_1 V_2)^T$ .

 Cleve Moler's movie about SVD <https://www.youtube.com/watch?v=R9UoFyqJca8>

### 5.6.1 二对角化

我们知道, 对称矩阵可以通过一系列 Householder 变换转化为对称三对角矩阵. 对于一般矩阵  $A \in \mathbb{R}^{m \times n}$ , 我们也可以通过 Householder 变换, 将其转化为二对角矩阵, 即计算正交矩阵  $U_1$  和  $V_1$  使得

$$U_1^T A V_1 = B, \quad (5.9)$$

其中  $B$  是一个实 (上) 二对角矩阵. 这个过程就称为**二对角化**.

 需要注意的是, 与对称矩阵的对称三对角化不同,  $A$  与  $B$  是不相似的.

设  $A \in \mathbb{R}^{m \times n}$ , 二对角化过程大致如下:

- (1) 首先确定一个 Householder 矩阵  $H_1 \in \mathbb{R}^{m \times m}$ , 使得  $H_1 A$  的第一列除第一个元素外, 其它分量都为零, 即

$$H_1 A = \begin{bmatrix} * & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix}.$$

- (2) 再确定一个 Householder 矩阵  $\tilde{H}_1 \in \mathbb{R}^{(n-1) \times (n-1)}$ , 把  $H_1 A$  的第一行的第 3 至第  $n$  个元素化



为零, 即


$$H_1 A \begin{bmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{bmatrix} = \begin{bmatrix} * & * & 0 & \cdots & 0 \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix}.$$

(3) 重复上面的过程, 直到把  $A$  最终化为二对角矩阵.

有了分解 (5.9) 以后, 我们可得

$$A^T A = (U_1 B V_1^T)^T U_1 B V_1^T = V_1 B^T B V_1^T,$$

即  $V_1^T A^T A V_1 = B^T B$ . 由于  $B^T B$  是对称三对角的, 所以这就相当于将  $A^T A$  三对角化.

 整个二对角化过程的运算量约为  $4mn^2 + 4m^2n - 4n^3/3$ , 若不需要计算  $U_1$  和  $V_1$ , 则运算量约为  $4mn^2 - 4n^3/3$ .

## 二对角矩阵的奇异值分解

设  $B \in \mathbb{R}^{n \times n}$  是一个二对角矩阵

$$B = \begin{bmatrix} a_1 & b_1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & b_{n-1} & \\ & & & & a_n \end{bmatrix}. \quad (5.10)$$

我们假定二对角矩阵的元素都非负, 即  $a_k, b_k \geq 0$ , 否则可以通过在左右分别乘以一个正交矩阵来实现 (见习题 5.7). 下面三种方法均可将计算  $B$  的 SVD 转化成计算对称三对角矩阵的特征分解:

- (1) 构造**增广矩阵**  $T = \begin{bmatrix} 0 & B^T \\ B & 0 \end{bmatrix}$ , 取置换矩阵  $P = [e_1, e_{n+1}, e_2, e_{n+2}, \dots, e_n, e_{2n}]$ , 则  $T_{ps} = P^T T P$  是对称三对角矩阵, 且主对角线元素全为 0, 次对角线元素为  $a_1, b_1, a_2, b_2, \dots, a_{n-1}, b_{n-1}, a_n$ . 设  $(\lambda_i, x_i)$  是  $T_{ps}$  的一个特征对, 则

$$\lambda_i = \pm \sigma_i, \quad P x_i = \frac{1}{\sqrt{2}} \begin{bmatrix} v_i \\ \pm u_i \end{bmatrix},$$

其中  $\sigma_i$  为  $B$  一个奇异值,  $u_i$  和  $v_i$  分别为对应的左和右奇异向量.

- (2) 令  $T_{BB^T} = BB^T$ , 则

$$T_{BB^T} = \begin{bmatrix} a_1^2 + b_1^2 & a_2 b_1 & & & \\ a_2 b_1 & \ddots & \ddots & & \\ & \ddots & \ddots & a_{n-1}^2 + b_{n-1}^2 & a_n b_{n-1} \\ & & & a_n b_{n-1} & a_n^2 \end{bmatrix}.$$


$T_{BB^T}$  的特征值为  $B$  的奇异值的平方, 且  $T_{BB^T}$  的特征向量为  $B$  的左奇异向量.



(3) 令  $T_{B^T B} = B^T B$ , 则

$$T_{B^T B} = \begin{bmatrix} a_1^2 & a_1 b_1 & & \\ a_1 b_1 & a_2^2 + b_1^2 & \ddots & \\ & \ddots & \ddots & a_{n-1} b_{n-1} \\ & & a_{n-1} b_{n-1} & a_n^2 + b_{n-1}^2 \end{bmatrix}.$$

$T_{B^T B}$  的特征值为  $B$  的奇异值的平方, 且  $T_{B^T B}$  的特征向量为  $B$  的右奇异向量.

 Golub & Kahan (1965) 采用前面两种做法. Reinsch (1970) 采用第三种做法, 其巧妙之处是, 只需对  $B$  做 Givens 变换即可, 类似隐式 QR 迭代.

理论上, 我们可以直接使用 QR 迭代、分而治之法或带反迭代的对分法, 计算三对角矩阵  $T_{ps}$ ,  $T_{BB^T}$  和  $T_{B^T B}$  的特征值和特征向量. 但一般来说, 这种做法并不是最佳的, 原因如下:

- (1) 对  $T_{ps}$  做 QR 迭代并不划算, 因为 QR 迭代计算所有的特征值和特征向量, 而事实上只要计算正的特征值即可;
- (2) 直接构成  $T_{BB^T}$  或  $T_{B^T B}$  是数值不稳定的. 事实上, 这样做可能会使得  $B$  的小奇异值的精度丢失一半.

下面是一些计算奇异值分解的比较实用的算法.

1. **Golub-Kahan SVD 算法**: 由 Golub 和 Kahan [52] 于 1965 年提出, 并在 1970 年进行了完善 [53], 是一种十分稳定且高效的计算 SVD 的算法, 也是最早的实用 SVD 算法. 主要思想是将带位移的对称 QR 迭代算法隐式地用到  $B^T B$  上, 在该算法中, 并不需要显式地把  $B^T B$  计算出来. 该算法也通常就称为 SVD 算法, 是一个基本且实用的算法, 目前仍然是计算小规模矩阵奇异值分解的常用算法. 关于这个算法的详细描述, 也可以参见 [56, 144].
2. **dqds 算法**: 由 Fernando 和 Parlett [43] 于 1994 年提出, 是计算二对角矩阵所有奇异值的最快算法, 而且能达到很高的相对精度, 包括奇异值很小的情形. 该算法主要基于对  $B^T B$  的 Cholesky 迭代, 可以看作是 LR 迭代算法的改进. 由于 LR 迭代算法在一定条件下与对称 QR 算法是等价的, 因此该算法也可以看作是 QR 迭代的变形.
3. **分而治之法**: 该算法是计算维数  $n \geq 25$  的矩阵的所有奇异值和奇异向量的最快算法, 但不能保证小奇异值的相对精度, 即  $\sigma_i$  的相对精度为  $\mathcal{O}(\varepsilon)\sigma_1$ , 而不是  $\mathcal{O}(\varepsilon)\sigma_i$ .
4. **对分法和反迭代**: 主要用于计算某个区间内的奇异值及对应的奇异向量, 能保证较高的相对精度.
5. **Jacobi 迭代**: 可隐式地对  $AA^T$  或  $A^T A$  实施对称 Jacobi 迭代, 能保证较高的相对精度. 最近, Z. Drmač 和 K. Veselić [33, 34] 改进了最初的 Jacobi 算法, 使其变成一个速度快、精度高的实用算法.

在这里, 我们介绍 Golub-Kahan SVD 算法, dqds 算法和 Jacobi 迭代.

### 5.6.2 Golub-Kahan SVD 算法

该算法主要思想是将带位移的对称 QR 迭代算法隐式地用到  $B^T B$  上, 而无需将  $B^T B$  显式地计算出来. Golub-Kahan SVD 算法有时也简称 SVD 算法, 其基本框架是:





- 将矩阵  $A$  二对角化, 得到上二对角矩阵  $B$ ;
- 用隐式 QR 迭代计算  $B^T B$  的特征值分解, 即

$$B^T B = Q \Lambda Q^T, \quad \Lambda = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2). \quad (5.11)$$

- 计算  $BQ$  的列主元 QR 分解, 即

$$(BQ)P = UR, \quad (5.12)$$

其中  $P$  是置换矩阵,  $U$  是正交矩阵,  $R$  是上三角矩阵.

由 (5.11) 可知

$$(BQ)^T BQ = \Lambda,$$

因此  $BQ$  是列正交矩阵 (但不是单位列正交). 再由 (5.12) 可知  $R = U^T(BQ)P$  也是列正交矩阵. 又  $R$  是上三角矩阵, 所以  $R$  必定是对角矩阵. 令  $V = QP$ , 则由 (5.12) 可知

$$U^T B V = R.$$

这就是二对角矩阵  $B$  的奇异值分解.

算法的具体实现可参见 [56, 144].

### 5.6.3 dqds 算法

我们首先介绍针对实对称正定矩阵的 LR 算法, 该算法思想与 QR 迭代算法类似, 但提出时间更早.

#### 算法 5.9. 带位移的 LR 算法

```

1: Let T_0 be a given real symmetric positive definite matrix
2: set $i = 0$
3: while not converge do
4: choose a shift τ_i^2 satisfying $\tau_i^2 < \min\{\lambda(T_i)\}$
5: compute B_i such that $T_i - \tau_i^2 I = B_i^T B_i$ % Cholesky factorization
6: $T_{i+1} = B_i B_i^T + \tau_i^2 I$
7: $i = i + 1$
8: end while

```

LR 迭代算法在形式上与 QR 迭代算法非常类似. 事实上, 对于不带位移的 LR 迭代算法, 我们可以证明, 两步 LR 迭代等价于一步 QR 迭代.

**引理 5.13** 设  $\tilde{T}$  是不带位移的 LR 算法迭代两步后生成的矩阵,  $\hat{T}$  是不带位移的 QR 算法迭代一步后生成的矩阵, 则  $\tilde{T} = \hat{T}$ .



- (1) LR 算法中要求  $T_0$  对称正定, 但并不一定是三对角矩阵;  
 (2) 由该引理可知, QR 算法与 LR 算法有相同的收敛性.

下面我们介绍 dqds (differential quotient difference with shifts) 算法. 该算法是针对三对角的对称正定矩阵  $B^T B$ , 其中  $B$  是二对角矩阵. 在数学上, dqds 算法与 LR 算法是等价的, 但在该算法中, 我们是直接通过  $B_i$  来计算  $B_{i+1}$ , 从而避免计算中间矩阵  $T_{i+1}$ , 这样也就尽可能地避免了由于计算  $B_i B_i^T$  而可能带来的数值不稳定性.

下面推导如何从  $B_i$  直接计算  $B_{i+1}$ . 设

$$B_i = \begin{bmatrix} a_1 & b_1 & & & \\ & a_2 & \ddots & & \\ & & \ddots & b_{n-1} & \\ & & & a_n & \\ & & & & a_n \end{bmatrix}, \quad B_{i+1} = \begin{bmatrix} \tilde{a}_1 & \tilde{b}_1 & & & \\ & \tilde{a}_2 & \ddots & & \\ & & \ddots & \tilde{b}_{n-1} & \\ & & & \tilde{a}_n & \\ & & & & \tilde{a}_n \end{bmatrix}.$$

为了书写方便, 我们记  $b_0 = b_n = \tilde{b}_0 = \tilde{b}_n = 0$ . 由 LR 算法 5.9 可知

$$B_{i+1}^T B_{i+1} + \tau_{i+1}^2 I = B_i B_i^T + \tau_i^2 I.$$

比较等式两边矩阵的对角线和上对角线元素, 可得

$$\begin{aligned} \tilde{a}_k^2 + \tilde{b}_{k-1}^2 + \tau_{i+1}^2 &= a_k^2 + b_k^2 + \tau_i^2, \quad k = 1, 2, \dots, n \\ \tilde{a}_k \tilde{b}_k &= a_{k+1} b_k \quad \text{或} \quad \tilde{a}_k^2 \tilde{b}_k^2 = a_{k+1}^2 b_k^2, \quad k = 1, 2, \dots, n-1. \end{aligned}$$

记  $\delta = \tau_{i+1}^2 - \tau_i^2$ ,  $p_k = a_k^2$ ,  $q_k = b_k^2$ ,  $\tilde{p}_k = \tilde{a}_k^2$ ,  $\tilde{q}_k = \tilde{b}_k^2$ , 则可得下面的 **qds 算法**.

#### 算法 5.10. qds 算法的单步 ( $B_i \rightarrow B_{i+1}$ )

- 1:  $\delta = \tau_{i+1}^2 - \tau_i^2$
- 2: **for**  $k = 1$  to  $n - 1$  **do**
- 3:    $\tilde{p}_k = p_k + q_k - \tilde{q}_{k-1} - \delta$
- 4:    $\tilde{q}_k = q_k \cdot (p_{k+1} / \tilde{p}_k)$
- 5: **end for**
- 6:  $\tilde{p}_n = p_n - \tilde{q}_{n-1} - \delta$

qds 算法中的每个循环仅需 5 个浮点运算, 所以运算量较少.

为了提高算法的精确性, 我们引入一个辅助变量  $d_k \triangleq p_k - \tilde{q}_{k-1} - \delta$ , 则

$$\begin{aligned} d_k &= p_k - \tilde{q}_{k-1} - \delta \\ &= p_k - \frac{q_{k-1} p_k}{\tilde{p}_{k-1}} - \delta \\ &= p_k \cdot \frac{\tilde{p}_{k-1} - q_{k-1}}{\tilde{p}_{k-1}} - \delta \\ &= p_k \cdot \frac{p_{k-1} - \tilde{q}_{k-2} - \delta}{\tilde{p}_{k-1}} - \delta \\ &= \frac{p_k}{\tilde{p}_{k-1}} \cdot d_{k-1} - \delta. \end{aligned}$$



于是就可得到 **dqds 算法**:

**算法 5.11.** dqds 算法的单步 ( $B_i \rightarrow B_{i+1}$ )

```

1: $\delta = \tau_{i+1}^2 - \tau_i^2$
2: $d_1 = p_1 - \delta$
3: for $k = 1$ to $n - 1$ do
4: $\tilde{p}_k = d_k + q_k$
5: $t = p_{k+1} / \tilde{p}_k$
6: $\tilde{q}_k = q_k \cdot t$
7: $d_{k+1} = d_k \cdot t - \delta$
8: end for
9: $\tilde{p}_n = d_n$

```

dqds 算法的运算量与 dqs 差不多, 但更精确. 这里我们只列出相应的结果.

**引理 5.14** 设二对角矩阵  $B$  的对角元和上对角元分别为  $a_1, a_2, \dots, a_n$  和  $b_1, b_2, \dots, b_{n-1}$ ,  $\tilde{B}$  的对角元和上对角元分别为  $\tilde{a}_i = a_i \eta_i$ ,  $\tilde{b}_i = b_i \xi_i$ . 则  $\tilde{B} = D_1 B D_2$ , 其中

$$D_1 = \text{diag} \left( \eta_1, \frac{\eta_2 \eta_1}{\xi_1}, \frac{\eta_3 \eta_2 \eta_1}{\xi_2 \xi_1}, \dots, \frac{\eta_n \eta_{n-1} \cdots \eta_1}{\xi_{n-1} \xi_{n-2} \cdots \xi_1} \right),$$

$$D_2 = \text{diag} \left( 1, \frac{\xi_1}{\eta_1}, \frac{\xi_2 \xi_1}{\eta_2 \eta_1}, \dots, \frac{\xi_{n-1} \xi_{n-2} \cdots \xi_1}{\eta_{n-1} \eta_{n-2} \cdots \eta_1} \right).$$

**定理 5.15** 设  $B$  和  $\tilde{B}$  的定义如引理 5.14. 若存在  $\tau > 1$  使得  $\tau^{-1} \leq \eta_i \leq \tau$ ,  $\tau^{-1} \leq \xi_i \leq \tau$ , 即  $\varepsilon = \tau - 1$  是  $B$  与  $\tilde{B}$  的相应元素之间的相对误差的一个上界. 设  $B$  与  $\tilde{B}$  的奇异值分别为  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$  和  $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \dots \geq \tilde{\sigma}_n$ , 则

$$|\tilde{\sigma}_i - \sigma_i| \leq (\tau^{4n-2} - 1) \sigma_i.$$

若  $\sigma_i \neq 0$ ,  $\varepsilon = \tau - 1 \ll 1$ , 则上式可记为

$$\frac{|\tilde{\sigma}_i - \sigma_i|}{\sigma_i} \leq \tau^{4n-2} - 1 = (4n - 2)\varepsilon + \mathcal{O}(\varepsilon^2).$$

下面的定理显示了 dqds 算法的高精度性质.


**定理 5.16** 以浮点运算对  $B$  应用 dqds 算法的单步, 得到矩阵  $\tilde{B}$ , 该过程等价于

1. 对  $B$  的每个元素作一个小的相对扰动 (不超过  $1.5\varepsilon$ ), 得到  $\tilde{B}$ ;
2. 对  $\tilde{B}$  应用精确的 dqds 算法的单步, 得到  $\bar{B}$ ;
3. 对  $\bar{B}$  的每个元素作一个小的相对扰动 (不超过  $\varepsilon$ ), 得到  $\tilde{B}$ .

由定理 5.15 可知,  $B$  和  $\tilde{B}$  的奇异值满足高的相对精度.

关于 dqds 算法中位移的选取, 以及如何判断收敛性, 可以参见 [43].



 dqds 算法是求解二对角矩阵所有奇异值的主要方法之一, 2000 年就被加入到 LAPACK 中 [2, 98]. 关于 dqds 的最新改进可以参见 [86].

### 5.6.4 Jacobi 算法

本节讨论对矩阵  $M = A^T A$  实施隐式的 Jacobi 算法来计算  $A$  的奇异值.

我们知道, Jacobi 算法的每一步就是对矩阵作 Jacobi 旋转, 即  $A^T A \rightarrow J^T A^T A J$ , 其中  $J$  的选取将某两个非对角元化为 0. 在实际计算中, 我们只需计算  $AJ$ , 故该算法称为**单边 Jacobi 旋转**.

#### 算法 5.12. 单边 Jacobi 旋转的单步

```
% 对 $M = A^T A$ 作 Jacobi 旋转, 将 $M(i, j), M(j, i)$ 化为 0
1: Compute $m_{ii} = (A^T A)_{ii}, m_{ij} = (A^T A)_{ij}, m_{jj} = (A^T A)_{jj}$
2: if m_{ij} is not small enough then
3: $\tau = (m_{ii} - m_{jj}) / (2 \cdot m_{ij})$
4: $t = \text{sign}(\tau) / (|\tau| + \sqrt{1 + \tau^2})$
5: $c = 1 / \sqrt{1 + t^2}$
6: $s = c \cdot t$
7: $A = AG(i, j, \theta)$ % $G(i, j, \theta)$ 为 Givens 变换
8: if eigenvectors are desired then
9: $J = J \cdot G(i, j, \theta)$
10: end if
11: end if
```

在上面算法的基础上, 我们可以给出完整的单边 Jacobi 算法.

#### 算法 5.13. 单边 Jacobi: 计算 $A = U \Sigma V^T$

```
1: while $A^T A$ is not diagonal enough do
2: for $i = 1$ to $n - 1$ do
3: for $j = i + 1$ to n do
4: 调用单边 Jacobi 旋转
5: end for
6: end for
7: end while
8: compute $\sigma_i = \|A(:, i)\|_2, i = 1, 2, \dots, n$
9: $U = [u_1, \dots, u_n]$ with $u_i = A(:, i) / \sigma_i$
10: $V = J$
```

Jacobi 算法的特点:



- 不需要双对角化, 这样可以避免双对角化引入的误差;
- 可达到相对较高的计算精度;
- 速度较慢. (目前已有快速的改进算法, 参见 [33, 34])

**定理 5.17** 设  $A = DX \in \mathbb{R}^{n \times n}$ , 其中  $D$  为非奇异对角阵,  $X$  非奇异. 设  $\hat{A}$  是按浮点运算单边 Jacobi 旋转  $m$  次后所得到的矩阵. 若  $A$  和  $\hat{A}$  的奇异值分别为  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$  和  $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \dots \geq \hat{\sigma}_n$ , 则

$$\frac{|\hat{\sigma}_i - \sigma_i|}{\sigma_i} \leq \mathcal{O}(m\varepsilon)\kappa(X).$$

故  $X$  的条件数越小, 计算矩阵  $A$  的奇异值时相对误差越小.



## 5.7 扰动分析

设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵, 则有下面的谱分解.

**定理 5.18** 设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵. 则存在一个正交矩阵  $Q$  使得

$$A = Q\Lambda Q^T$$

其中  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  是一个实对角矩阵.

这里的  $\lambda_i$  就是  $A$  的特征值, 我们假设  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ . 令  $Q = [q_1, q_2, \dots, q_n]$ , 则  $q_i$  就是  $\lambda_i$  对应的单位正交特征向量.

关于对称矩阵特征值问题的扰动理论, 这里只做一些简单介绍, 若要深入了解这方面的信息, 可以参考 [71, 72, 116, 141].

### 5.7.1 特征值与 Rayleigh 商

**定义 5.2** 设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵, 向量  $x \in \mathbb{R}^n$  非零, 则  $x$  关于  $A$  的 **Rayleigh 商** 定义为:

$$\rho(x, A) = \frac{x^T A x}{x^T x}. \quad (5.13)$$

有时简记为  $\rho(x)$ .

下面是关于 Rayleigh 商的一些基本性质:

- (1)  $\rho(\alpha x) = \rho(x)$ ,  $\forall \alpha \in \mathbb{R}, \alpha \neq 0$ ;
- (2)  $\rho(q_i) = \lambda_i, i = 1, 2, \dots, n$ ;
- (3) 设  $x = \alpha_1 q_1 + \alpha_2 q_2 + \dots + \alpha_n q_n$ , 则

$$\rho(x) = \frac{\alpha_1^2 \lambda_1 + \alpha_2^2 \lambda_2 + \dots + \alpha_n^2 \lambda_n}{\alpha_1^2 + \alpha_2^2 + \dots + \alpha_n^2};$$

- (4)  $\lambda_n \leq \rho(x) \leq \lambda_1, |\rho(x)| \leq \|A\|_2$ .

实对称矩阵的特征值与 Rayleigh 商之间的一个基本性质是 Courant-Fischer 极小极大定理.

**定理 5.19 (Courant-Fischer)** 设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵, 其特征值为  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ , 则有

$$\lambda_k = \max_{U \in \mathbb{S}_k^n} \min_{x \in U, x \neq 0} \frac{x^T A x}{x^T x} = \min_{V \in \mathbb{S}_{n-k+1}^n} \max_{x \in V, x \neq 0} \frac{x^T A x}{x^T x},$$

其中  $\mathbb{S}_i^n$  表示  $\mathbb{R}^n$  中所有  $i$  维子空间构成的集合. 当

$$U = \text{span}\{q_1, \dots, q_k\}, \quad V = \text{span}\{q_k, \dots, q_n\}, \quad x = q_k$$

时, 上式中的等号成立.

(板书)

**证明.** 设  $U \in \mathbb{S}_k^n$  和  $V \in \mathbb{S}_{n-k+1}^n$  分别为  $\mathbb{R}^n$  中任意的  $k$  和  $n-k+1$  维子空间. 由于

$$\dim U + \dim V = n + 1 > n,$$



可得

$$\mathbb{U} \cap \mathbb{V} \neq \{0\}.$$

故存在非零向量  $\tilde{x} \in \mathbb{U} \cap \mathbb{V}$ . 所以有

$$\min_{x \in \mathbb{U}, x \neq 0} \rho(x) \leq \rho(\tilde{x}) \leq \max_{x \in \mathbb{V}, x \neq 0} \rho(x).$$

由  $\mathbb{U}$  和  $\mathbb{V}$  的任意性可知,

$$\max_{\mathbb{U} \in \mathbb{S}_k^n} \min_{x \in \mathbb{U}, x \neq 0} \rho(x) \leq \min_{\mathbb{V} \in \mathbb{S}_{n-k+1}^n} \max_{x \in \mathbb{V}, x \neq 0} \rho(x). \quad (5.14)$$

取  $\mathbb{U} = \text{span}\{q_1, \dots, q_k\}$ , 则  $\mathbb{U}$  中的任意向量都可写成  $x = \alpha_1 q_1 + \dots + \alpha_k q_k$ , 此时

$$\rho(x) = \frac{x^T A x}{x^T x} = \frac{\alpha_1^2 \lambda_1 + \dots + \alpha_k^2 \lambda_k}{\alpha_1^2 + \dots + \alpha_k^2} \geq \frac{\sum_{i=1}^k \alpha_i^2 \lambda_k}{\sum_{i=1}^k \alpha_i^2} = \lambda_k,$$

即

$$\max_{\mathbb{U} \in \mathbb{S}_k^n} \min_{x \in \mathbb{U}, x \neq 0} \rho(x) \geq \lambda_k. \quad (5.15)$$

同理, 取  $\mathbb{V} = \text{span}\{q_k, \dots, q_n\}$ , 则  $\mathbb{V}$  中的任意向量都可写成  $x = \alpha_k q_k + \dots + \alpha_n q_n$ , 此时

$$\rho(x) = \frac{x^T A x}{x^T x} = \frac{\alpha_k^2 \lambda_k + \dots + \alpha_n^2 \lambda_n}{\alpha_k^2 + \dots + \alpha_n^2} \leq \frac{\sum_{i=k}^n \alpha_i^2 \lambda_k}{\sum_{i=k}^n \alpha_i^2} = \lambda_k,$$

即

$$\min_{\mathbb{V} \in \mathbb{S}_{n-k+1}^n} \max_{x \in \mathbb{V}, x \neq 0} \rho(x) \leq \lambda_k. \quad (5.16)$$

由 (5.14), (5.15), (5.16) 可知, 定理结论成立.  $\square$

当  $k=1$  和  $k=n$  时, 就可以得到下面的定理.

**定理 5.20 (Rayleigh-Ritz)** 设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵, 其特征值为  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ , 则有

$$\lambda_1 = \max_{x \in \mathbb{R}^n, x \neq 0} \frac{x^T A x}{x^T x}, \quad \lambda_n = \min_{x \in \mathbb{R}^n, x \neq 0} \frac{x^T A x}{x^T x}.$$

由极小极大定理, 我们可以得到下面的特征值分隔定理.

**定理 5.21 (分隔定理)** 设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵,  $B = Q^T A Q$ , 其中  $Q \in \mathbb{R}^{n \times (n-1)}$  满足  $Q^T Q = I_{n-1}$ . 再设  $A$  和  $B$  的特征值分别为

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \quad \text{和} \quad \tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_{n-1},$$

则有

$$\lambda_1 \geq \tilde{\lambda}_1 \geq \lambda_2 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_{n-1} \geq \lambda_n.$$

特别地, 在定理 5.21 中, 取  $Q = [e_1, \dots, e_{i-1}, e_{i+1}, \dots, e_n]$ , 则可以得到下面的结论.

**推论 5.22** 设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵,  $\tilde{A}$  是  $A$  的一个  $n-1$  阶主子矩阵,  $A$  和  $\tilde{A}$  的特征值分别为

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \quad \text{和} \quad \tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_{n-1},$$



则有

$$\lambda_1 \geq \tilde{\lambda}_1 \geq \lambda_2 \geq \tilde{\lambda}_2 \cdots \geq \tilde{\lambda}_{n-1} \geq \lambda_n.$$

反复应用上面的推论, 即可得到下面的结论.

**推论 5.23** 设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵,  $\tilde{A}$  是  $A$  的一个  $k$  阶主子矩阵 ( $1 \leq k \leq n-1$ ),  $A$  和  $\tilde{A}$  的特征值分别为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \quad \text{和} \quad \tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_k,$$

则有

$$\lambda_i \geq \tilde{\lambda}_i \geq \lambda_{n-k+i}, \quad i = 1, 2, \dots, k.$$

### 5.7.2 对称矩阵特征值的扰动分析

设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵, 扰动矩阵  $E \in \mathbb{R}^{n \times n}$  也是对称矩阵, 下面讨论  $A + E$  的特征值与  $A$  的特征值之间的关系.

由极小极大定理, 我们可以证明下面的性质.

**定理 5.24** 设  $A \in \mathbb{R}^{n \times n}$  和  $B = A + E \in \mathbb{R}^{n \times n}$  都是对称矩阵, 其特征值分别为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \quad \text{和} \quad \tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_n.$$

假定  $E$  的最大和最小特征值分别为  $\mu_1$  和  $\mu_n$ , 则有

$$\lambda_i + \mu_1 \geq \tilde{\lambda}_i \geq \lambda_i + \mu_n, \quad i = 1, 2, \dots, n.$$

(板书)

**证明.** 由 Courant-Fischer 定理 5.19 和 Rayleigh-Ritz 定理 5.20 可知

$$\begin{aligned} \tilde{\lambda}_i &= \min_{V \in \mathbb{S}_{n-i+1}^n} \max_{x \in V, x \neq 0} \frac{x^T B x}{x^T x} \\ &= \min_{V \in \mathbb{S}_{n-i+1}^n} \max_{x \in V, x \neq 0} \left( \frac{x^T A x}{x^T x} + \frac{x^T E x}{x^T x} \right) \\ &\leq \min_{V \in \mathbb{S}_{n-i+1}^n} \max_{x \in V, x \neq 0} \left( \frac{x^T A x}{x^T x} + \mu_1 \right) \\ &= \min_{V \in \mathbb{S}_{n-i+1}^n} \max_{x \in V, x \neq 0} \frac{x^T A x}{x^T x} + \mu_1 \\ &= \lambda_i + \mu_1. \end{aligned}$$

同理可得

$$\begin{aligned} \tilde{\lambda}_i &= \max_{U \in \mathbb{S}_i^n} \min_{x \in U, x \neq 0} \frac{x^T B x}{x^T x} \\ &= \max_{U \in \mathbb{S}_i^n} \min_{x \in U, x \neq 0} \left( \frac{x^T A x}{x^T x} + \frac{x^T E x}{x^T x} \right) \\ &\geq \max_{U \in \mathbb{S}_i^n} \min_{x \in U, x \neq 0} \left( \frac{x^T A x}{x^T x} + \mu_n \right) \end{aligned}$$





$$\begin{aligned}
 &= \max_{U \in \mathbb{S}_i^n} \min_{x \in U, x \neq 0} \frac{x^T A x}{x^T x} + \mu_n \\
 &= \lambda_i + \mu_n.
 \end{aligned}$$

所以定理结论成立. □

根据这个定理, 我们立即可以得到下面的 Weyl 定理.

**定理 5.25 (Weyl)** 设  $A \in \mathbb{R}^{n \times n}$  和  $B = A + E \in \mathbb{R}^{n \times n}$  都是对称矩阵, 其特征值分别为  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$  和  $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_n$ , 则

$$|\tilde{\lambda}_j - \lambda_j| \leq \|E\|_2, \quad j = 1, 2, \dots, n.$$

该定理的结论可以推广到奇异值情形. 我们首先给出下面的引理.

**引理 5.26** 设  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ) 的奇异值分解为  $A = U \Sigma V$ , 其中  $U = [u_1, \dots, u_n] \in \mathbb{R}^{m \times n}$  为列正交矩阵,  $V = [v_1, \dots, v_n] \in \mathbb{R}^{n \times n}$  为正交矩阵,  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ . 将  $U$  扩展成  $n \times n$  的正交矩阵  $[U, \tilde{U}] = [u_1, \dots, u_n, \tilde{u}_1, \dots, \tilde{u}_{m-n}]$ , 令

$$H = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix} \in \mathbb{R}^{(m+n) \times (m+n)},$$

则  $H$  对称, 且特征值为  $\pm \sigma_i$  和 0 (其中 0 至少为  $m - n$  重特征值), 对应的特征向量分别为  $\frac{\sqrt{2}}{2} \begin{bmatrix} v_i \\ \pm u_i \end{bmatrix}$ ,  $i = 1, 2, \dots, n$ , 和  $\begin{bmatrix} 0 \\ \tilde{u}_j \end{bmatrix}$ ,  $j = 1, 2, \dots, m - n$ .

(留作课外自习, 直接代入验证即可)

由上面的引理和 Weyl 定理 5.25 立即可得

**定理 5.27** 设  $A, B \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ), 它们的奇异值分别为  $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$  和  $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_n$ . 则

$$|\tilde{\sigma}_j - \sigma_j| \leq \|B - A\|_2, \quad j = 1, 2, \dots, n.$$

最后给出一个基于 F-范数的扰动性质 [71].

**定理 5.28** 设  $A, E \in \mathbb{C}^{n \times n}$  且  $A$  是 Hermite 的,  $A + E$  是正规矩阵. 并设  $A$  的特征值满足

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n,$$

$A + E$  的特征值  $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_n$  满足

$$\text{Re}(\tilde{\lambda}_1) \geq \text{Re}(\tilde{\lambda}_2) \geq \cdots \geq \text{Re}(\tilde{\lambda}_n).$$

则

$$\sum_{i=1}^n |\tilde{\lambda}_i - \lambda_i|^2 \leq \|E\|_F^2.$$


### 5.7.3 对称矩阵特征向量的扰动



**定义 5.3** 设  $A \in \mathbb{R}^{n \times n}$  的特征值为  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ , 则  $\lambda_i$  与其余特征值之间的 **间隙 (gap)** 定义为

$$\text{gap}(\lambda_i, A) = \min_{j \neq i} |\lambda_j - \lambda_i|.$$

有时简记为  $\text{gap}(\lambda_i)$ .

 特征向量的敏感性依赖于其对应的特征值的 gap, 一般来说, gap 越小, 特征向量越敏感.

**例 5.2** 设

$$A = \begin{bmatrix} 1+g & \\ & 1 \end{bmatrix}, \quad E = \begin{bmatrix} 0 & \varepsilon \\ \varepsilon & 0 \end{bmatrix}, \quad (0 < \varepsilon < g)$$

则  $A$  的特征值为  $\lambda_1 = 1+g, \lambda_2 = 1$ , 对应的单位特征向量为  $q_1 = e_1, q_2 = e_2$ . 当  $\varepsilon$  充分小时,  $A+E$  的特征值为  $\hat{\lambda}_{1,2} = 1 + (g \pm \sqrt{g^2 + 4\varepsilon^2})/2$ , 对应的单位特征向量为

$$\begin{aligned} \hat{q}_1 &= \beta_1 \cdot \begin{bmatrix} 1 \\ \frac{\sqrt{1+4\varepsilon^2/g^2}-1}{2\varepsilon/g} \end{bmatrix} = \beta_1 \cdot \begin{bmatrix} 1 \\ \frac{\sqrt{(1+2\varepsilon^2/g^2)^2-4(\varepsilon/g)^4}-1}{2\varepsilon/g} \end{bmatrix} \\ &\approx \beta_1 \cdot \begin{bmatrix} 1 \\ \frac{(1+2\varepsilon^2/g^2)-1}{2\varepsilon/g} \end{bmatrix} \\ &= \frac{1}{\sqrt{1+\varepsilon^2/g^2}} \begin{bmatrix} 1 \\ \varepsilon/g \end{bmatrix}, \\ \hat{q}_2 &= \beta_2 \cdot \begin{bmatrix} 1 \\ \frac{-\sqrt{1+4\varepsilon^2/g^2}-1}{2\varepsilon/g} \end{bmatrix} \approx \frac{1}{\sqrt{1+\varepsilon^2/g^2}} \begin{bmatrix} -\varepsilon/g \\ 1 \end{bmatrix}, \end{aligned}$$

其中  $\beta_1, \beta_2$  为规范化因子. 故特征向量的扰动约为  $\varepsilon/g$ , 与特征值的间隙  $\text{gap}(\lambda_i, A) = g$  成反比.

**定理 5.29** 设  $A = Q\Lambda Q^T$  和  $A+E = \tilde{Q}\tilde{\Lambda}\tilde{Q}^T$  分别为对称矩阵  $A \in \mathbb{R}^{n \times n}$  和  $A+E \in \mathbb{R}^{n \times n}$  的特征值分解, 其中  $Q = [q_1, q_2, \dots, q_n]$  和  $\tilde{Q} = [\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_n]$  均为正交矩阵, 且  $\tilde{q}_i$  为  $q_i$  对应的扰动特征向量. 用  $\theta_i$  表示  $q_i$  和  $\tilde{q}_i$  之间的锐角, 则当  $\text{gap}(\lambda_i, A) > 0$  时

$$\frac{1}{2} \sin 2\theta_i \leq \frac{\|E\|_2}{\text{gap}(\lambda_i, A)}.$$

类似地, 当  $\text{gap}(\tilde{\lambda}_i, A+E) > 0$  时

$$\frac{1}{2} \sin 2\theta_i \leq \frac{\|E\|_2}{\text{gap}(\tilde{\lambda}_i, A+E)}.$$

(留作课外自习)

**证明.** 如右图所示, 令  $d = \tilde{q}_i / \cos \theta_i - q_i$ , 即  $\tilde{q}_i = (q_i + d) \cos \theta_i$ . 则



$$d^\top q_i = 0, \quad \tan \theta_i = \|d\|_2, \quad \sec \theta_i = \|q_i + d\|_2.$$

令  $\eta = \tilde{\lambda}_i - \lambda_i$ , 由  $(A + E)\tilde{q}_i = \tilde{\lambda}_i \tilde{q}_i$  可得

$$(A + E)(q_i + d) = (\eta + \lambda_i)(q_i + d),$$

将  $Aq_i = \lambda_i q_i$  代入后整理可得

$$(\eta I - E)(q_i + d) = (A - \lambda_i I)d.$$

又  $q_i^\top (A - \lambda_i I) = ((A - \lambda_i I)q_i)^\top = 0$ , 故用  $q_i^\top$  左乘上式两边可得

$$q_i^\top (\eta I - E)(q_i + d) = q_i^\top (A - \lambda_i I)d = 0, \quad (5.17)$$

即  $(\eta I - E)(q_i + d) \in \text{span}\{q_i\}^\perp = \text{span}\{q_1, \dots, q_{i-1}, q_{i+1}, \dots, q_n\}$ . 所以可设  $(\eta I - E)(q_i + d) = \sum_{j \neq i} \alpha_j q_j$ . 又  $q_i^\top d = 0$ , 故可设  $d = \sum_{j \neq i} \delta_j q_j$ . 所以

$$\begin{aligned} \sum_{j \neq i} \alpha_j q_j &= (\eta I - E)(q_i + d) = (A - \lambda_i I)d \\ &= (A - \lambda_i I) \sum_{j \neq i} \delta_j q_j = \sum_{j \neq i} \delta_j (A - \lambda_i I)q_j = \sum_{j \neq i} \delta_j (\lambda_j - \lambda_i) q_j. \end{aligned}$$

由于  $q_1, q_2, \dots, q_n$  线性无关, 故可得  $\delta_j (\lambda_j - \lambda_i) = \alpha_j$ . 又  $\text{gap}(\lambda_i, A) > 0$ , 即  $j \neq i$  时  $\lambda_j \neq \lambda_i$ , 所以  $\delta_i = \frac{\alpha_i}{\lambda_j - \lambda_i}$ , 因此

$$d = \sum_{j \neq i} \frac{\alpha_j}{\lambda_j - \lambda_i} q_j.$$

注意到  $q_i^\top d = 0$  且  $q_i^\top q_i = 1$ , 所以由 (5.17) 可得

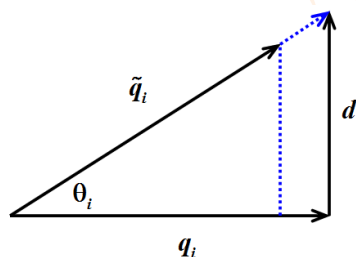
$$\eta = q_i^\top E(q_i + d).$$

故

$$\begin{aligned} (\eta I - E)(q_i + d) &= (q_i + d)\eta - E(q_i + d) \\ &= (q_i + d)q_i^\top E(q_i + d) - E(q_i + d) \\ &= ((q_i + d)q_i^\top - I)E(q_i + d). \end{aligned}$$

由习题 5.4 可知  $\|(q_i + d)q_i^\top - I\|_2 = \|q_i + d\|_2$ , 故

$$\begin{aligned} \tan \theta_i &= \|d\|_2 = \left( \sum_{j \neq i} \left| \frac{\alpha_j}{\lambda_j - \lambda_i} \right|^2 \right)^{1/2} \\ &\leq \left( \sum_{j \neq i} \left| \frac{\alpha_j}{\text{gap}(\lambda_i, A)} \right|^2 \right)^{1/2} \\ &= \frac{1}{\text{gap}(\lambda_i, A)} \left( \sum_{j \neq i} \alpha_j^2 \right)^{1/2} \\ &= \frac{1}{\text{gap}(\lambda_i, A)} \left\| \sum_{j \neq i} \alpha_j q_j \right\|_2 \\ &= \frac{1}{\text{gap}(\lambda_i, A)} \|(\eta I - E)(q_i + d)\|_2 \end{aligned}$$







$$\begin{aligned}
&\leq \frac{1}{\text{gap}(\lambda_i, A)} \|(q_i + d)q_i^\top - I\|_2 \cdot \|E\|_2 \cdot \|(q_i + d)\|_2 \\
&= \frac{1}{\text{gap}(\lambda_i, A)} \|(q_i + d)\|_2^2 \cdot \|E\|_2 \\
&= \frac{1}{\text{gap}(\lambda_i, A)} \cdot \frac{1}{\cos^2 \theta_i} \|E\|_2.
\end{aligned}$$

即

$$\frac{1}{2} \sin 2\theta_i = \sin \theta_i \cos \theta_i = \tan \theta_i \cos^2 \theta_i \leq \frac{1}{\text{gap}(\lambda_i, A)} \|E\|_2.$$

将  $A + E$  看作原矩阵,  $(A + E) - E$  看作是扰动矩阵, 则可证明第二个结论.  $\square$

-  当  $\theta_i \ll 1$  时,  $\frac{1}{2} \sin 2\theta_i \approx \theta_i \approx \sin \theta_i$ ;
-  若  $\|E\|_2 \geq \frac{1}{2} \text{gap}(\lambda_i, A)$ , 则定理中给出的上界就失去了实际意义;
-  在该定理中, 没有对特征值进行排序;
-  在实际计算中, 我们通常所知道的是  $\text{gap}(\tilde{\lambda}_i, A + E)$ .

#### 5.7.4 Rayleigh 商逼近

**定理 5.30** 设对称矩阵  $A \in \mathbb{R}^{n \times n}$  的特征值为  $\lambda_1, \lambda_2, \dots, \lambda_n$ ,

(1) 若  $x \in \mathbb{R}^n$  是单位向量,  $\beta \in \mathbb{R}$ , 则

$$\min_{1 \leq i \leq n} |\lambda_i - \beta| \leq \|Ax - \beta x\|_2; \quad (5.18)$$

(2) 对于给定的非零向量  $x \in \mathbb{R}^n$ , 当  $\beta = \rho(x)$  时,  $\|Ax - \beta x\|_2$  达到最小, 即

$$\min_{\beta \in \mathbb{R}} \|Ax - \beta x\|_2 = \|Ax - \rho(x)x\|_2; \quad (5.19)$$

(3) 令  $r = Ax - \rho(x)x$ , 设  $\lambda_i$  是距离  $\rho(x)$  最近的特征值,  $\text{gap}' = \min_{j \neq i} |\lambda_j - \rho(x)|$ ,  $\theta$  是  $x$  和  $q_i$  之间的锐角, 其中  $q_i$  是  $\lambda_i$  对应的单位特征向量, 则

$$\sin \theta \leq \frac{\|r\|_2}{\text{gap}'} \quad \text{且} \quad |\lambda_i - \rho(x)| \leq \frac{\|r\|_2^2}{\text{gap}'}. \quad (5.20)$$

(留作课外自习)

**证明.** (1) 若  $\beta$  是  $A$  的特征值, 则结论显然成立.

若  $\beta$  不是  $A$  的特征值, 则  $A - \beta I$  非奇异, 故

$$1 = \|x\|_2 = \|(A - \beta I)^{-1}(A - \beta I)x\|_2 \leq \|(A - \beta I)^{-1}\|_2 \cdot \|(A - \beta I)x\|_2. \quad (5.21)$$

由于  $A - \beta I$  对称, 且特征值为  $\lambda_i - \beta$ , 故

$$\|(A - \beta I)^{-1}\|_2 = \frac{1}{\min_{1 \leq i \leq n} |\lambda_i - \beta|}.$$

代入 (5.21) 即可知结论成立.

(2) 由于

$$x^\top (Ax - \rho(x)x) = x^\top Ax - \frac{x^\top Ax}{x^\top x} x^\top x = 0,$$




即  $x \perp (Ax - \rho(x)x)$ . 所以


$$\begin{aligned}\|Ax - \beta x\|_2^2 &= \|(A - \rho(x))x + (\rho(x) - \beta)x\|_2^2 \\ &= \|Ax - \rho(x)x\|_2^2 + \|(\rho(x) - \beta)x\|_2^2 \\ &\geq \|Ax - \rho(x)x\|_2^2,\end{aligned}$$

所以当  $\beta = \rho(x)$  时,  $\|Ax - \beta x\|_2$  达到最小.

(3) 略

□

 由 (5.18) 可知, 在幂迭代和反迭代中可以使用残量  $\|Ax - \tilde{\lambda}x\|_2 < tol$  作为停机准则, 这里  $\tilde{\lambda}$  是迭代过程中计算得到的近似特征值. 等式 (5.19) 则解释了为什么用 Rayleigh 商来近似特征值.

 不等式 (5.20) 表明  $|\lambda_i - \rho(x)|$  的值与残量范数  $\|r\|_2$  的平方成正比, 这个结论是 Rayleigh 商迭代局部三次收敛的基础.

### 5.7.5 相对扰动分析

这里主要讨论  $A$  和  $X^TAX$  的特征值和特征向量之间的扰动关系, 其中  $X$  非奇异且满足  $\|X^TX - I\|_2 = \varepsilon$ . 这是因为在计算特征向量时, 由于舍入误差的原因, 最后得到的正交矩阵  $Q$  会带有误差, 从而失去正交性.

**定理 5.31 (相对 Weyl 定理)** 设对称矩阵  $A$  和  $X^TAX$  的特征值分别为  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$  和  $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_n$ , 令  $\varepsilon = \|X^TX - I\|_2$ , 则

$$|\tilde{\lambda}_i - \lambda_i| \leq \varepsilon |\lambda_i| \quad \text{或} \quad \frac{|\tilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \leq \varepsilon \quad (\text{if } \lambda_i \neq 0).$$

(留作课外自习)

**证明.** 因为  $A - \lambda_i I$  的第  $i$  个特征值为 0, 故由 Sylvester 惯性定理 5.11 可知


$$X^T(A - \lambda_i I)X = (X^TAX - \lambda_i I) + \lambda_i(I - X^TX)$$

的第  $i$  个特征值也为 0. 由 Weyl 定理 5.25 可知

$$|(\tilde{\lambda}_i - \lambda_i) - 0| \leq \|\lambda_i(I - X^TX)\|_2 = \varepsilon |\lambda_i|,$$

即定理结论成立

□

 当  $X$  正交时,  $\varepsilon = 0$ , 故  $X^TAX$  与  $A$  有相同的特征值. 当  $X$  几乎正交时,  $\varepsilon$  很小, 此时  $X^TAX$  与  $A$  的特征值几乎相同.

**推论 5.32** 设  $G$  和  $Y^TGY$  的奇异值分别为  $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$  和  $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_n$ , 令

$\varepsilon = \max \{ \|X^T X - I\|_2, \|Y^T Y - I\|_2 \}$ , 则

$$|\tilde{\sigma}_i - \sigma_i| \leq \varepsilon |\sigma_i| \quad \text{或} \quad \frac{|\tilde{\sigma}_i - \sigma_i|}{|\sigma_i|} \leq \varepsilon \quad (\text{if } \sigma_i \neq 0).$$

下面给出特征向量的相对扰动性质.

**定义 5.4** 设  $A \in \mathbb{R}^{n \times n}$  的特征值为  $\lambda_1, \lambda_2, \dots, \lambda_n$ , 若  $\lambda_i \neq 0$ , 则  $\lambda_i$  与其余特征值之间的**相对间隙 (relative gap)** 定义为

$$\text{relgap}(\lambda_i, A) = \min_{j \neq i} \frac{|\lambda_j - \lambda_i|}{|\lambda_i|}.$$

**定理 5.33** 设  $A \in \mathbb{R}^{n \times n}$  和  $X^T A X \in \mathbb{R}^{n \times n}$  的特征值分解分别为  $A = Q \Lambda Q^T$  和  $X^T A X = \tilde{Q} \tilde{\Lambda} \tilde{Q}^T$ , 其中  $Q = [q_1, q_2, \dots, q_n]$  和  $\tilde{Q} = [\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_n]$  均为正交矩阵,  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ ,  $\tilde{\Lambda} = \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_n)$  且  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ ,  $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_n$ . 设  $\theta_i$  表示  $q_i$  和  $\tilde{q}_i$  之间的锐角, 令  $\varepsilon_1 = \|I - X^{-T} X^{-1}\|_2$ ,  $\varepsilon_2 = \|X - I\|_2$ , 若  $\varepsilon_1 < 1$  且  $\text{relgap}(\tilde{\lambda}_i, X^T A X) > 0$ , 则

$$\frac{1}{2} \sin 2\theta_i \leq \frac{\varepsilon_1}{1 - \varepsilon_1} \cdot \frac{1}{\text{relgap}(\tilde{\lambda}_i, X^T A X)} + \varepsilon_2.$$

(留作课外自习)

**证明.** 设  $\eta = \tilde{\lambda}_i - \lambda_i$ ,  $H = A - \tilde{\lambda}_i I$ ,  $F = \tilde{\lambda}_i(I - X^{-T} X^{-1})$ , 则  $H q_i = A q_i - \tilde{\lambda}_i q_i = -\eta q_i$ ,

$$H + F = A - \tilde{\lambda}_i X^{-T} X^{-1} = X^{-T} (X^T A X - \tilde{\lambda}_i I) X^{-1}.$$

故  $(H + F)(X \tilde{q}_i) = 0$ . 即  $X \tilde{q}_i$  是  $H + F$  的第  $i$  个特征值  $\tilde{\lambda}_i = 0$  的一个特征向量. 设  $\theta_1$  是  $q_i$  与  $X \tilde{q}_i$  之间的锐角, 由定理 5.29 可知

$$\frac{1}{2} \sin 2\theta_1 \leq \frac{\|F\|_2}{\text{gap}(\lambda_i, H + F)} = \frac{\varepsilon_1 |\tilde{\lambda}_i|}{\text{gap}(\lambda_i, H + F)}. \quad (5.22)$$

由于  $\tilde{\lambda}_i = 0$ , 故  $\text{gap}(\lambda_i, H + F)$  即为  $H + F$  的最小非零特征值的绝对值. 又  $X^T (H + F) X = X^T A X - \tilde{\lambda}_i I$  的特征值为  $\tilde{\lambda}_j - \tilde{\lambda}_i, j = 1, 2, \dots, n$ , 且

$$\tilde{\lambda}_1 - \tilde{\lambda}_i \geq \tilde{\lambda}_2 - \tilde{\lambda}_i \geq \dots \geq \tilde{\lambda}_n - \tilde{\lambda}_i,$$

所以由相对 Weyl 定理 5.31 可知

$$|\hat{\lambda}_j - (\tilde{\lambda}_j - \tilde{\lambda}_i)| \leq \varepsilon_1 |\tilde{\lambda}_j - \tilde{\lambda}_i|.$$

这里  $\hat{\lambda}_j$  表示  $H + F$  的第  $j$  个特征值 (按降序排列). 因此  $|\hat{\lambda}_j| \geq (1 - \varepsilon_1) |\tilde{\lambda}_j - \tilde{\lambda}_i|$ , 故

$$\text{gap}(\hat{\lambda}_i, H + F) \geq (1 - \varepsilon_1) \text{gap}(\tilde{\lambda}_i, X^T A X).$$

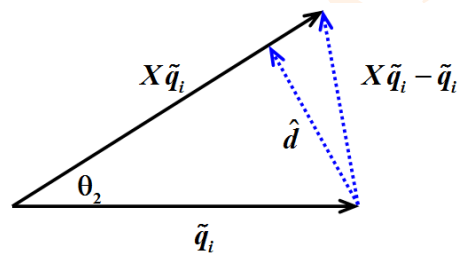
代入 (5.22) 可得

$$\frac{1}{2} \sin 2\theta_1 \leq \frac{\varepsilon_1 |\tilde{\lambda}_i|}{(1 - \varepsilon_1) \text{gap}(\tilde{\lambda}_i, X^T A X)} = \frac{\varepsilon_1}{1 - \varepsilon_1} \cdot \frac{1}{\text{relgap}(\tilde{\lambda}_i, X^T A X)}.$$

设  $\theta_2$  是  $X \tilde{q}_i$  与  $\tilde{q}_i$  之间的锐角, 则由右图可知



$$\begin{aligned}
 \sin \theta_2 &= \|\tilde{q}_i\|_2 \sin \theta_2 \leq \|X\tilde{q}_i - \tilde{q}_i\|_2 \\
 &\leq \|X - I\|_2 \cdot \|\tilde{q}_i\|_2 \\
 &= \varepsilon_2.
 \end{aligned}$$



又  $\theta_i \leq \theta_1 + \theta_2$ , 故

$$\begin{aligned}
 \frac{1}{2} \sin 2\theta_i &\leq \frac{1}{2} \sin 2\theta_1 + \frac{1}{2} \sin 2\theta_2 \\
 &\leq \frac{1}{2} \sin 2\theta_1 + \sin \theta_2 \\
 &\leq \frac{\varepsilon_1}{1 - \varepsilon_1} \cdot \frac{1}{\text{relgap}(\tilde{\lambda}_i, X^\top AX)} + \varepsilon_2,
 \end{aligned}$$

即定理结论成立. □

## 5.8 应用

### 5.8.1 SVD 与图像压缩

关于 SVD 的几何意义, 可以参考 AMS 的 Feature Column 的一篇文献: [We Recommend a Singular Value Decomposition](#), 2009.

用矩阵  $A$  表示图像, 然后求它的 SVD, 最后保留前  $k$  个奇异值, 即用  $A$  的截断 SVD 来近似  $A$ , 从而达到图像压缩的目的.

**例 5.3** 举例: [LS\\_SVD\\_ImageCompress\\_01.m](#), [LS\\_SVD\\_ImageCompress\\_02.m](#)





## 5.9 课后习题

**练习 5.1** 设  $x, y \in \mathbb{R}^n$ , 试证明:  $\det(I + xy^\top) = 1 + y^\top x$ . (注: 在复数域也成立)

**练习 5.2** 设  $A = D + uu^\top$ , 其中  $D = \text{diag}(d_1, d_2, \dots, d_n)$  满足  $d_1 \geq d_2 \geq \dots \geq d_n$ ,  $u = [u_1, u_2, \dots, u_n]^\top \in \mathbb{R}^n$ .

(1) 证明:  $d_i$  是  $A$  的特征值的充要条件是  $d_i = d_{i+1}$  或  $d_i = d_{i-1}$  或  $u_i = 0$ ;

(2) 若  $u_i = 0$ , 则  $e_i$  是与  $d_i$  对应的特征向量;

(3) 若  $d_{i-1} > d_i = d_{i+1} > d_{i+2}$  且  $u_i \neq 0$ , 证明: 对应于  $\lambda = d_i$  的特征向量  $x$  除  $x_i$  和  $x_{i+1}$  外, 其余分量全部为 0, 且  $x_i u_i + x_{i+1} u_{i+1} = 0$ .

**思考: 如果  $d_{i-1} > d_i = d_{i+1} = d_{i+2} > d_{i+3}$ , 则结论如何?**

**练习 5.3\*** 设  $D = \text{diag}(d_1, d_2, \dots, d_n)$ . 若矩阵  $D + \alpha uu^\top$  和  $D + \hat{u}\hat{u}^\top$  具有相同的特征值, 记为  $\lambda_1, \lambda_2, \dots, \lambda_n$ , 且满足交错性质  $\lambda_1 > d_1 > \lambda_2 > d_2 > \dots > \lambda_n > d_n$ , 则它们具有相同的特征向量.

**思考: 如果没有交错性质, 则结论如何?**

**练习 5.4\*** 设  $q \in \mathbb{R}^n$  满足  $\|q\|_2 = 1$ . 对任意与  $q$  正交的向量  $d \in \mathbb{R}^n$ , 试证明:

$$\|(q+d)q^\top - I\|_2 = \|q+d\|_2.$$

**练习 5.5** 设  $S \in \mathbb{C}^{n \times n}$  是 skew-Hermite 矩阵, 即  $S^* = -S$ . 证明:

(1)  $S$  的非零特征值是纯虚数;

(2)  $I + S$  非奇异;

(3) 矩阵  $(I + S)^{-1}(I - S)$  是酉矩阵. (该矩阵称为  $S$  的 Cayley 变换)

**练习 5.6** 设  $B \in \mathbb{R}^{m \times n}$ ,  $m \geq n$  且  $\|B\|_2 < 1$ . 若  $A = \begin{bmatrix} I & B \\ B^\top & I \end{bmatrix}$ , 证明:

$$\kappa_2(A) = \frac{1 + \|B\|_2}{1 - \|B\|_2}.$$

**练习 5.7\*** 设  $B$  是二对角矩阵

$$B = \begin{bmatrix} a_1 & b_1 & & \\ & \ddots & \ddots & \\ & & \ddots & b_{n-1} \\ & & & a_n \end{bmatrix}.$$

证明: 存在正交矩阵  $Q_1$  和  $Q_2$ , 使得  $Q_1^\top B Q_2$  仍然是二对角矩阵且所有元素都非负.

**练习 5.8** 设  $x, y \in \mathbb{R}^n$ , 若  $y^\top x$  只有零特征值, 证明:  $xy^\top$  也只有零特征值.

设  $X, Y \in \mathbb{R}^{n \times 2}$ , 若  $Y^\top X$  只有零特征值, 则  $XY^\top$  是否也只有零特征值?

**练习 5.9\***(极分解) 设  $A \in \mathbb{C}^{n \times n}$ . 证明:

(1) 存在酉矩阵  $U$  和唯一的 Hermite 半正定矩阵  $P$ , 使得  $A = PU$ .

(2) 进一步, 若  $A$  非奇异, 则  $U$  也唯一.

**练习 5.10\*** 设  $A \in \mathbb{C}^{n \times n}$ . 证明:  $A$  可对角化当且仅当存在 Hermite 正定矩阵  $P$  使得  $P^{-1}AP$  是正



规矩阵.

(提示: 利用极分解, 但不是对  $A$  进行极分解)

### ..... 以下为可选题 .....

**练习 5.11** 设  $\lambda \in \mathbb{R}$  是对称矩阵  $A \in \mathbb{R}^{n \times n}$  的一个特征值, 对应的特征向量为  $x \in \mathbb{R}^n$ . 若  $\tilde{x} \in \mathbb{R}^n$  是  $x$  的一个  $\mathcal{O}(\varepsilon)$  近似, 即  $\tilde{x} = x + \mathcal{O}(\varepsilon)$ , 证明:

$$\frac{\tilde{x}^\top A \tilde{x}}{\tilde{x}^\top \tilde{x}} = \lambda + \mathcal{O}(\varepsilon^2),$$

即  $\tilde{x}$  对应的 Rayleigh 商是  $\lambda$  的  $\mathcal{O}(\varepsilon^2)$  逼近.

**练习 5.12** 设  $A, E \in \mathbb{R}^{n \times n}$  都是对称矩阵, 它们的特征值分别为  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$  和  $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_n$ . 设  $A + E$  的特征值为  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n$ , 试证明

$$\lambda_i + \theta_n \leq \hat{\lambda}_i \leq \lambda_i + \theta_1, \quad i = 1, 2, \dots, n.$$

并由此可知, 若  $E$  对称正定, 则  $\hat{\lambda}_i \geq \lambda_i$ .

**练习 5.13** 设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵,  $A_{n-1}$  是  $A$  的  $n-1$  阶顺序主子矩阵, 它们的特征值分别为  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$  和  $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_{n-1}$ . 试证明

$$\lambda_1 \geq \theta_1 \geq \lambda_2 \geq \theta_2 \geq \cdots \geq \lambda_{n-1} \geq \theta_{n-1} \geq \lambda_n.$$

更一般地, 记  $\lambda_i(B)$  为对称矩阵  $B$  的第  $i$  个特征值 (按降序排列), 设  $A_k$  是  $A$  任意一个  $k$  阶主子矩阵 ( $1 \leq k \leq n-1$ ), 则有

$$\lambda_i(A) \geq \lambda_i(A_k) \geq \lambda_{n-k+i}(A), \quad i = 1, 2, \dots, k.$$

**练习 5.14\*** 证明以下结论:

(1) 设  $x \in \mathbb{R}^n$  是一个正向量, 即  $x_i > 0$ . 证明: 由  $x$  定义的 Cauchy 矩阵  $A$

$$a_{ij} = \frac{1}{x_i + x_j}$$

是对称半正定的. 进一步, 若  $x_i$  互不相等, 则  $A$  对称正定. (参见 [145])

(2) 证明: Hilbert 矩阵是对称正定的.

### ..... 以下为实践题 .....

**练习 5.15** 编写程序, 实现对称矩阵  $A \in \mathbb{R}^{n \times n}$  的三对角化.

**练习 5.16** 编写程序, 实现矩阵  $A \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ) 的二对角化.

**练习 5.17** 编写程序, 实现计算对称三对角矩阵特征值的带 Wilkinson 位移的 QR 算法.

**练习 5.18** 编写程序, 实现计算二对角矩阵奇异值的 dqds 算法.



## 第六讲 线性方程组定常迭代法

考虑线性方程组

$$Ax = b, \quad A \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n.$$

目前, 求解线性方程组的方法有:

- 直接法: PLU 分解, LDL<sup>T</sup> 分解, Cholesky 分解等
- 迭代法:
  - ▷ 定常迭代法: Jacobi, Gauss-Seidel, SOR, SSOR, AOR 等
  - ▷ Krylov 子空间迭代法: CG, MINRES, GMRES, BiCGStab 等
- 快速方法:
  - ▷ 基于各种快速变换, 如 FFT, DCT, DST 等
  - ▷ 代数多重网格法 (Algebraic multigrid)
  - ▷ 快速多极子算法 (Fast multipole)

有些方法可能只适用于某类特殊方程, 如快速方法. 在实际应用中, 这些方法常常结合使用, 如混合 (hybrid) 方法, 预处理 (preconditioning) 方法等.

直接法的优点是稳定可靠, 能在有限步内得到近似解 (如果不考虑误差, 则得到精确解), 而且所需存储量和运算量都是可知的. 缺点是所需运算量约为  $\mathcal{O}(n^3)$ , 这对于大规模线性方程组来说是非常巨大的. 而且在实际应用中, 很多问题中需要求解的大规模线性方程组都是稀疏的, 如偏微分方程的有限差分/有限元离散, 但直接法很难有效地利用问题的稀疏性来降低总运算量. 而迭代法则可以很好地利用问题的稀疏性, 大大降低运算量.

从历史上看, 最早的迭代法可以追溯到十九世纪 Gauss, Jacobi, Seidel 和 Nekrasov 等工作 [13, 73]. 但是针对迭代法的系统研究主要还是在计算机出现以后, 大约是从二十世纪五十年代开始. 在开始阶段主要研究的是定常迭代法, 典型代表有 Jacobi, GS, SOR, SSOR, ADI, Chebyshev 迭代, 等等. 在这期间, 有两本非常有名的经典著作, 一本是 Varga 的 “Matrix Iterative Analysis” (1962) [132], 另一本是 Yong 的 “Iterative Solution of Large Linear Systems” (1971) [134]. 定常迭代法的收敛性有着非常完美的理论分析, 但在实际使用中却存在着许多不足, 比如收敛速度较慢, 最优参数估计困难, 等等.

从二十世纪七十年代中期开始, 研究重点慢慢转向 Krylov 子空间迭代法. 事实上, 早在 1952 年, Lanczos [82] 和 Hestenes & Stiefel [66] 就同时独立地提出了求解对称正定线性方程组的共轭梯度法 (CG). 对于一个  $n$  阶的线性方程组, 如果不考虑舍入误差的影响, 则共轭梯度法在  $n$  步后就一定会得到精确解. 因此共轭梯度法一开始被认作是直接法. 但在实际使用中发现, 由于舍入误差的影响, 迭代步数可能会超过  $n$ , 特别是对于坏条件问题. 而对于条件数较小的线性方程组, 迭代步数则可能远远小于  $n$ , 这使得共轭梯度法具有一定的吸引力. 但由于种种原因 [54], 共轭梯度法提出后并没有受到重视, 在其出现后的近二十年里, 主流方法仍然是 Gauss 消去法, SOR 方法和 Chebyshev 迭代法.

1971 年, Reid 在其论文 [102] 中指出, 对于好条件的大规模稀疏线性方程组, 共轭梯度法能在很少的迭代步数内得到一个很好的近似解. (事实上, Engeli 等 [40] 在 1959 年就发现了该现象, 但并没有引起关注). 这极大地促发了大家对共轭梯度法的研究兴趣, 包括各种改进和推广, 如求解对称不定线性方程组的 MINRES 方法和 SYMMLQ 方法 [95], 求解一般线性方程组的 GMRES 方法 [107], QMR 方法 [47], BiCGSTAB 方法 [122], 等等. 目前, Krylov 子空间迭代法已成为求解大规模稀疏线性方程组的主流方法.

本讲介绍常用的定常迭代法, 关于 Krylov 子空间迭代法, 我们将在下一讲介绍.

#### 关于线性方程组定常迭代法的相关参考资料

- G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2013. [56]
- R. S. Varga, *Matrix Iterative Analysis*, 2nd edition, 2000. [132]
- D. M. Young, *Iterative Solution of Large Linear Systems*, 1971. [134]
- O. Axelsson, *Iterative Solution Methods*, 1994. [6]
- R. Barrett, et.al, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, 1994. [11]
- 徐树方, 矩阵计算的理论与方法, 1995. [143]
- ▷ Y. Saad and H. A. van der Vorst, *Iterative solution of linear systems in the 20th century*, 2000. [110]

更多迭代法可参见 [11].



## 6.1 定常迭代法

随着矩阵规模的增大, 直接法的运算量也随之快速增长. 对于大规模的线性方程组, 由于运算量太大, 直接法一般不再被采用, 取而代之的是迭代法.

当直接求解  $Ax = b$  比较困难时, 我们可以求解一个比较容易求解的近似等价方程组  $Mx = b$ , 其中  $M$  可以看作是  $A$  在某种意义下的近似. 设  $Mx = b$  的解为  $x^{(1)}$ . 易知它与原方程组的解  $x_* = A^{-1}b$  之间的差距满足

$$A(x_* - x^{(1)}) = b - Ax^{(1)}.$$

如果  $x^{(1)}$  已经满足精度要求, 即非常接近真解  $x_*$ , 则可以停止计算, 否则需要修正. 记  $\Delta x \triangleq x_* - x^{(1)}$ , 则  $\Delta x$  满足方程  $A\Delta x = b - Ax^{(1)}$ . 但由于直接求解该方程组比较困难 (与求解原方程组一样困难), 因此我们还是通过求解近似方程组

$$M\Delta x^{(1)} = b - Ax^{(1)},$$

得到一个修正量  $\Delta x^{(1)}$ . 于是修正后的近似解为

$$x^{(2)} = x^{(1)} + \Delta x^{(1)} = x^{(1)} + M^{-1}(b - Ax^{(1)}).$$

如果  $x^{(2)}$  已经满足精度要求, 则停止计算, 否则继续按以上的方式进行修正, 即求解  $M\Delta x^{(2)} = b - Ax^{(2)}$  得到修正量  $\Delta x^{(2)}$ , 然后加到  $x^{(2)}$  上得到  $x^{(3)}$ :

$$x^{(3)} = x^{(2)} + \Delta x^{(2)} = x^{(2)} + M^{-1}(b - Ax^{(2)}).$$

不断重复以上步骤, 于是, 我们就得到一个序列

$$x^{(1)}, x^{(2)}, \dots, x^{(k)}, \dots$$

满足以下递推关系

$$x^{(k+1)} = x^{(k)} + M^{-1}(b - Ax^{(k)}), \quad k = 1, 2, \dots$$

这就构成了一个迭代法. 由于每次迭代的格式是一样的, 因此称为 **定常迭代法**.

通常, 构造一个好的定常迭代, 需要考虑以下两点:

- (1) 以  $M$  为系数矩阵的线性方程组要比原线性方程组更容易求解;
- (2)  $M$  应该是  $A$  的一个很好的近似, 而且迭代序列  $\{x^{(k)}\}$  要收敛.

常用的定常迭代法是基于**矩阵分裂**的迭代法, 主要包括:

- Richardson 方法
- Jacobi 方法
- Gauss-Seidel (G-S) 方法
- 超松弛 (SOR, Successive Over-Relaxation) 方法
- 对称超松弛 (SSOR, Symmetric SOR) 方法
- 加速超松弛 (AOR, Accelerated Over-Relaxation) 方法



## 6.2 矩阵分裂迭代法

首先给出 **矩阵分裂** 的定义.

**定义 6.1 (矩阵分裂 Matrix Splitting)** 设  $A \in \mathbb{R}^{n \times n}$  非奇异, 称

$$A = M - N \quad (6.1)$$

为  $A$  的一个矩阵分裂, 其中  $M$  非奇异.

考虑线性方程组

$$Ax = b, \quad (6.2)$$

其中  $A \in \mathbb{R}^{n \times n}$  非奇异. 迭代法的基本思想: 给定一个迭代初始值  $x^{(0)}$ , 通过一定的迭代格式生成一个迭代序列  $\{x^{(k)}\}_{k=0}^{\infty}$ , 使得

$$\lim_{k \rightarrow \infty} x^{(k)} = x_* \triangleq A^{-1}b.$$

给定一个矩阵分裂 (6.1), 则原方程组 (6.2) 就等价于  $Mx = Nx + b$ . 于是我们就可以构造出以下的迭代格式

$$Mx^{(k+1)} = Nx^{(k)} + b, \quad k = 0, 1, 2, \dots,$$

或

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b \triangleq Gx^{(k)} + g, \quad k = 0, 1, 2, \dots, \quad (6.3)$$

其中  $G \triangleq M^{-1}N$  称为 **迭代矩阵**. 这就是基于矩阵分裂 (6.1) 的迭代法. 易知, 选取不同的  $M$ , 就可以构造出不同的迭代法.

### 6.2.1 Jacobi 迭代法

将矩阵  $A$  写成

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

其中  $D$  为  $A$  的对角线部分,  $-L$  和  $-U$  分别为  $A$  的严格下三角和严格上三角部分.

在矩阵分裂  $A = M - N$  中取  $M = D$ ,  $N = L + U$ , 则可得 **Jacobi 迭代** 方法:


$$x^{(k+1)} = D^{-1}(L + U)x^{(k)} + D^{-1}b, \quad k = 0, 1, 2, \dots \quad (6.4)$$

对应的迭代矩阵为

$$G_J = D^{-1}(L + U).$$

写成分量形式即为

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1, j \neq i}^n a_{ij}x_j^{(k)} \right), \quad i = 1, 2, \dots, n.$$

 由于 Jacobi 迭代中  $x_i^{(k+1)}$  的更新顺序与  $i$  无关, 即可以按顺序  $i = 1, 2, \dots, n$  计算, 也可以按顺序  $i = n, n-1, \dots, 2, 1$  计算, 或者乱序计算. 因此 Jacobi 迭代非常适合并行计算.



**算法 6.1.** Jacobi 迭代法

```

1: Given an initial guess $x^{(0)}$
2: while not converge do
3: for $i = 1$ to n do
4: $x_i^{(k+1)} = \left(b_i - \sum_{j=1, j \neq i}^n a_{ij} x_j^{(k)} \right) / a_{ii}$
5: end for
6: end while

```

我们有时也将 Jacobi 迭代格式写为

$$x^{(k+1)} = x^{(k)} + D^{-1}(b - Ax^{(k)}) = x^{(k)} + D^{-1}r_k, \quad k = 0, 1, 2, \dots,$$

其中  $r_k \triangleq b - Ax^{(k)}$  是  $k$  次迭代后的残量. 这表明,  $x^{(k+1)}$  是通过对  $x^{(k)}$  做一个修正得到的.

**6.2.2 Gauss-Seidel 迭代法**

在分裂  $A = M - N$  中取  $M = D - L$ ,  $N = U$ , 即可得 **Gauss-Seidel (G-S) 迭代** 方法:

$$x^{(k+1)} = (D - L)^{-1}Ux^{(k)} + (D - L)^{-1}b. \quad (6.5)$$

对应的迭代矩阵为

$$G_{GS} = (D - L)^{-1}U.$$

将 G-S 迭代改写为

$$Dx^{(k+1)} = Lx^{(k+1)} + Ux^{(k)} + b,$$

即可得分量形式

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right), \quad i = 1, 2, \dots, n.$$

**算法 6.2.** Gauss-Seidel 迭代法

```

1: Given an initial guess $x^{(0)}$
2: while not converge do
3: for $i = 1$ to n do
4: $x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right)$
5: end for
6: end while

```

 G-S 方法的主要优点是充分利用了已经获得的最新数据.

 但在 G-S 方法中, 未知量的更新必须按自然顺序进行, 因此不适合并行计算.





### 6.2.3 SOR 迭代法

在 G-S 方法的基础上, 我们可以通过引入一个松弛参数  $\omega$  来加快收敛速度. 这就是 SOR (Successive Overrelaxation) 方法 [133]. 该方法的基本思想是将 G-S 方法中的第  $k+1$  步近似解与第  $k$  步近似解做一个加权平均, 从而给出一个新的近似解, 即

$$x^{(k+1)} = (1 - \omega)x^{(k)} + \omega D^{-1} (Lx^{(k+1)} + Ux^{(k)} + b). \quad (6.6)$$

整理后即得

$$x^{(k+1)} = (D - \omega L)^{-1} ((1 - \omega)D + \omega U)x^{(k)} + \omega(D - \omega L)^{-1}b,$$

其中  $\omega$  称为**松弛参数** (relaxation parameter).

- 当  $\omega = 1$  时, SOR 即为 G-S 方法,
- 当  $\omega < 1$  时, 称为**低松弛** (under relaxation) 方法,
- 当  $\omega > 1$  时, 称为**超松弛** (over relaxation) 方法.

在大多数情况下, 当  $\omega > 1$  时会取得比较好的收敛效果.

 SOR 方法曾经在很长一段时间内是科学计算中求解线性方程组的首选方法.

SOR 的迭代矩阵为

$$G_{\text{SOR}} = (D - \omega L)^{-1} ((1 - \omega)D + \omega U),$$

对应的矩阵分裂为


$$M = \frac{1}{\omega}D - L, \quad N = \frac{1 - \omega}{\omega}D + U.$$

由 (6.6) 可知 SOR 迭代的分量形式为

$$\begin{aligned} x_i^{(k+1)} &= (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right) \\ &= x_i^{(k)} + \frac{\omega}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i}^n a_{ij}x_j^{(k)} \right) \end{aligned}$$

#### 算法 6.3. 求解线性方程组的 SOR 迭代法

- 1: Given an initial guess  $x^{(0)}$  and parameter  $\omega$
- 2: **while** not converge **do**
- 3:     **for**  $i = 1$  to  $n$  **do**
- 4:          $x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right)$
- 5:     **end for**
- 6: **end while**

 SOR 方法最大的优点是引入了松弛参数  $\omega$ : 通过选取适当的  $\omega$  就可以大大提高方法的收敛速度. 但是 SOR 方法最大的难点就是如何选取最优的参数.





## 6.2.4 SSOR 迭代法

将 SOR 方法中的  $L$  和  $U$  相互交换位置, 则可得迭代格式

$$x^{(k+1)} = (D - \omega U)^{-1} ((1 - \omega)D + \omega L) x^{(k)} + \omega(D - \omega U)^{-1} b.$$

将这个迭代格式与 SOR 相结合, 就可以得到下面的两步迭代法

$$\begin{cases} x^{(k+\frac{1}{2})} = (D - \omega L)^{-1} [(1 - \omega)D + \omega U] x^{(k)} + \omega(D - \omega L)^{-1} b, \\ x^{(k+1)} = (D - \omega U)^{-1} [(1 - \omega)D + \omega L] x^{(k+\frac{1}{2})} + \omega(D - \omega U)^{-1} b. \end{cases}$$

这就是 **对称超松弛 (SSOR)** 迭代法, 相当于将  $L$  与  $U$  同等看待, 交替做两次 SOR 迭代.

消去中间迭代向量  $x^{(k+\frac{1}{2})}$ , 可得

$$x^{(k+1)} = G_{\text{SSOR}} x^{(k)} + g,$$


其中迭代矩阵

$$G_{\text{SSOR}} = (D - \omega U)^{-1} [(1 - \omega)D + \omega L] (D - \omega L)^{-1} [(1 - \omega)D + \omega U].$$

对应的矩阵分裂为

$$\begin{aligned} M &= \frac{1}{\omega(2 - \omega)} [D - \omega(L + U) + \omega^2 L D^{-1} U] \\ &= \frac{1}{\omega(2 - \omega)} (D - \omega L) D^{-1} (D - \omega U), \\ N &= \frac{1}{\omega(2 - \omega)} [(1 - \omega)D + \omega L] D^{-1} [(1 - \omega)D + \omega U]. \end{aligned}$$

 对于某些特殊问题, SOR 方法不收敛, 但仍然可能构造出收敛的 SSOR 方法.

 一般来说, SOR 方法的渐进收敛速度对参数  $\omega$  比较敏感, 但 SSOR 对参数  $\omega$  不是很敏感.

## 算法 6.4. SSOR 迭代法

- 1: Given an initial guess  $v^{(0)}$  and parameter  $\omega$
- 2: **while** not converge **do**
- 3:     **for**  $i = 1$  to  $n$  **do**
- 4:          $x_i^{(k+\frac{1}{2})} = (1 - \omega)x_i^{(k)} + \frac{\omega}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+\frac{1}{2})} - \sum_{j=i+1}^n a_{ij}x_j^{(k)} \right)$
- 5:     **end for**
- 6:     **for**  $i = n$  to  $1$  **do**
- 7:          $x_i^{(k+1)} = (1 - \omega)x_i^{(k+\frac{1}{2})} + \frac{\omega}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+\frac{1}{2})} - \sum_{j=i+1}^n a_{ij}x_j^{(k+1)} \right)$
- 8:     **end for**
- 9: **end while**



### 6.2.5 AOR 迭代法

Hadjidimos 于 1978 年提出了 **加速超松弛 (AOR, Accelerated Over-Relaxation)** 迭代方法, 迭代矩阵为


$$G_{\text{AOR}} = (D - \gamma L)^{-1} [(1 - \omega)D + (\omega - \gamma)L + \omega U],$$

其中  $\gamma$  和  $\omega$  为松弛参数. 对应的矩阵分解为

$$M = \frac{1}{\omega}(D - \gamma L), \quad N = \frac{1}{\omega}[(1 - \omega)D + (\omega - \gamma)L + \omega U].$$

易知:

- (1) 当  $\gamma = \omega$  时, AOR 方法即为 SOR 方法;
- (2) 当  $\gamma = \omega = 1$  时, AOR 方法即为 G-S 方法;
- (3) 当  $\gamma = 0, \omega = 1$  时, AOR 方法即为 Jacobi 方法.

 AOR 迭代法中含有两个参数. 因此在理论上, 通过选取合适的参数, AOR 迭代法会收敛得更快. 但也是因为含有两个参数, 使得参数的选取变得更加困难, 因此较少使用.

 与 SSOR 类似, 我们也可以定义 SAOR 迭代法.

### 6.2.6 Richardson 迭代法

Richardson 方法是一类形式非常简单的算法, 其迭代格式为

$$x^{(k+1)} = x^{(k)} + \omega(b - Ax^{(k)}), \quad k = 0, 1, 2, \dots$$

它可以看作是基于一矩阵分裂的迭代法:

$$M = \frac{1}{\omega}I, \quad N = \frac{1}{\omega}I - A.$$

对应的迭代矩阵为

$$G_{\text{R}} = I - \omega A.$$

**定理 6.1** 设  $A \in \mathbb{R}^{n \times n}$  是对称正定矩阵,  $\lambda_1$  和  $\lambda_n$  分别是  $A$  的最大和最小特征值, 则 Richardson 方法收敛当且仅当

$$0 < \omega < \frac{2}{\lambda_1}.$$

另外, Richardson 方法的最优参数为

$$\omega_* = \arg \min_{\omega} \rho(G_{\text{R}}) = \frac{2}{\lambda_1 + \lambda_n},$$

即当  $\omega = \omega_*$  时, 迭代矩阵的谱半径达到最小, 且有

$$\rho(G_{\text{R}}) = \begin{cases} 1 - \omega\lambda_n & \text{if } \omega \leq \omega_* \\ \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = \frac{\kappa(A) - 1}{\kappa(A) + 1} & \text{if } \omega = \omega_* \\ \omega\lambda_1 - 1 & \text{if } \omega \geq \omega_*. \end{cases}$$



如果在每次迭代时取不同的参数, 即

$$x^{(k+1)} = x^{(k)} + \omega_k(b - Ax^{(k)}), \quad k = 0, 1, 2, \dots,$$

则每次迭代的格式就不一样了, 因此不再是定常迭代, 而是 **非定常 (Nonstationary)** 迭代. 此时称为 **非定常 Richardson 方法**.

### 6.2.7 分块迭代法

如果  $A$  的对角线中出现零, 则 Jacobi, G-S, SOR 等方法就不再有定义. 此时我们可以采用分块迭代格式. 另外, 分块迭代法也能提升算法的计算效率.

将  $A$  写成如下的分块形式 (如右图所示):

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & A_{22} & \cdots & A_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pp} \end{bmatrix},$$

|          |          |  |  |  |          |
|----------|----------|--|--|--|----------|
| $A_{11}$ |          |  |  |  |          |
|          | $A_{22}$ |  |  |  |          |
|          |          |  |  |  |          |
|          |          |  |  |  |          |
|          |          |  |  |  |          |
|          |          |  |  |  | $A_{pp}$ |

则相应的分块 Jacobi, 分块 G-S 和分块 SOR 迭代法分别定义为:

- 分块 Jacobi 迭代:

$$A_{ii}x_i^{(k+1)} = b_i - \sum_{j=1, j \neq i}^p A_{ij}x_j^{(k)}, \quad i = 1, 2, \dots, p.$$

- 分块 Gauss-seidel 迭代:

$$A_{ii}x_i^{(k+1)} = b_i - \sum_{j=1}^{i-1} A_{ij}x_j^{(k+1)} - \sum_{j=i+1}^p A_{ij}x_j^{(k)}, \quad i = 1, 2, \dots, p.$$

- 分块 SOR 迭代:

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \omega A_{ii}^{-1} \left( b_i - \sum_{j=1}^{i-1} A_{ij}x_j^{(k+1)} - \sum_{j=i+1}^p A_{ij}x_j^{(k)} \right),$$

$$i = 1, 2, \dots, p.$$

### 6.3 应用: Poisson 方程求解

Poisson 方程是一类典型的偏微分方程, 经过五点差分离散后, 转化为求解一个线性方程组. 我们以这个线性方程组为例, 对 Jacobi, G-S 和 SOR 方法进行测试.

#### 6.3.1 一维 Poisson 方程

首先介绍一维 Poisson 方程的离散. 考虑如下带 Dirichlet 边界条件的一维 Poisson 方程

$$\begin{cases} -\frac{d^2 u(x)}{dx^2} = f(x), & 0 < x < 1, \\ u(0) = a, u(1) = b, \end{cases} \quad (6.7)$$

其中  $f(x)$  是给定的函数,  $u(x)$  是需要计算的未知函数.

#### 差分离散

取步长  $h = \frac{1}{n+1}$ , 节点设为  $x_i = ih, i = 0, 1, 2, \dots, n+1$ . 我们采用二阶中心差分来近似二阶导数, 可得

$$-\frac{d^2 u(x)}{dx^2} \Big|_{x_i} = \frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1}))}{h^2} + O\left(h^2 \cdot \left\| \frac{d^4 u}{dx^4} \right\|_{\infty}\right), \quad i = 1, 2, \dots, n.$$

将其代入 (6.7), 舍去高阶项后就可得到 Poisson 方程在  $x_i$  点的近似离散方程

$$-u_{i-1} + 2u_i - u_{i+1} = h^2 f_i,$$

其中  $f_i = f(x_i)$ ,  $u_i$  为  $u(x_i)$  的近似. 令  $i = 1, 2, \dots, n$ , 则可得  $n$  个线性方程. 写成矩阵形式为

$$T_n u = f, \quad (6.8)$$

其中

$$T_n = \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} \triangleq \text{tridiag}(-1, 2, -1), \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \\ u_n \end{bmatrix}, \quad f = \begin{bmatrix} f_1 + u_0 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n + u_{n+1} \end{bmatrix}. \quad (6.9)$$

#### 系数矩阵 $T_n$ 的性质

易知,  $T_n$  是不可约弱对角占优的. 另外,  $T_n$  是对称矩阵, 因此其特征值都是实数. 事实上, 我们有下面的结论.

**引理 6.2**  $T_n$  的特征值和对应的特征向量分别为

$$\lambda_k = 2 - 2 \cos \frac{k\pi}{n+1},$$

$$z_k = \sqrt{\frac{2}{n+1}} \cdot \left[ \sin \frac{k\pi}{n+1}, \sin \frac{2k\pi}{n+1}, \dots, \sin \frac{nk\pi}{n+1} \right]^T, \quad k = 1, 2, \dots, n,$$

即  $T_n = Z \Lambda Z^T$ , 其中  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  是对角矩阵,  $Z = [z_1, z_2, \dots, z_n]$  是正交矩阵.



(留作课外自习, 直接代入验证即可)

由此可知,  $T_n$  是对称正定的. 更一般地, 我们有下面的结论.

**推论 6.3** 设  $T = \text{tridiag}(a, b, c) \in \mathbb{R}^{n \times n}$ , 则  $T$  的特征值为

$$\lambda_k = b - 2\sqrt{ac} \cos \frac{k\pi}{n+1}, \quad k = 1, 2, \dots, n,$$

对应的特征向量为  $z_k$ , 其第  $j$  个分量为

$$z_k(j) = \left(\frac{a}{c}\right)^{\frac{j}{2}} \sin \frac{jk\pi}{n+1}.$$

特别地, 若  $a = c = 1$ , 则对应的单位特征向量为

$$z_k = \sqrt{\frac{2}{n+1}} \cdot \left[ \sin \frac{k\pi}{n+1}, \sin \frac{2k\pi}{n+1}, \dots, \sin \frac{nk\pi}{n+1} \right]^T.$$

(留作练习)

由引理 6.2 可知,  $T_n$  的最大特征值为


$$2 \left( 1 - \cos \frac{n\pi}{n+1} \right) = 4 \sin^2 \frac{n\pi}{2(n+1)} \approx 4,$$

最小特征值为

$$2 \left( 1 - \cos \frac{\pi}{n+1} \right) = 4 \sin^2 \frac{\pi}{2(n+1)} \approx \left( \frac{\pi}{n+1} \right)^2.$$

因此, 当  $n$  很大时,  $T_n$  的谱条件数约为

$$\kappa_2(T_n) \approx \frac{4(n+1)^2}{\pi^2}.$$

 矩阵  $T_n$  有时也称为 **二阶差分矩阵**, 可以写成一阶差分矩阵的乘积, 即  $T_n = DD^T$ , 其中

$$D = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n \times (n+1)}.$$

其中  $D$  是 **一阶差分矩阵**. 需要注意的是,  $D$  不是方阵, 所以从表面上看, 无法像 LU 分解那样用来求解线性方程组  $T_n u = f$ , 但事实上是可以的, 对应的方法称为 **变系数追赶法**, 感兴趣的读者可参见 [147].

### 6.3.2 二维 Poisson 方程

下面考虑二维 Poisson 方程的离散. 同样是带 Dirichlet 边界条件, 即

$$\begin{cases} -\Delta u(x, y) = -\frac{\partial^2 u(x, y)}{\partial x^2} - \frac{\partial^2 u(x, y)}{\partial y^2} = f(x, y), & (x, y) \in \Omega, \\ u(x, y) = u_0(x, y), & (x, y) \in \partial\Omega, \end{cases} \quad (6.10)$$

其中  $\Omega = [0, 1] \times [0, 1]$  为求解区域,  $\partial\Omega$  表示  $\Omega$  的边界.



### 五点差分离散

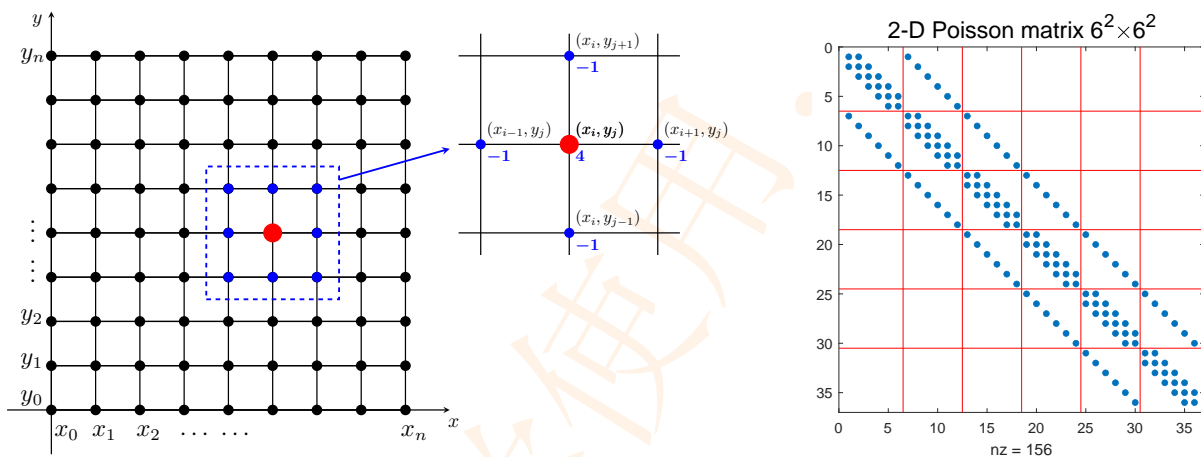
为了简单起见, 我们在  $x$ -方向和  $y$ -方向取相同的步长  $h = \frac{1}{n+1}$ , 节点设为  $x_i = ih, y_j = jh$ ,  $i, j = 0, 1, 2, \dots, n+1$ . 在  $x$ -方向和  $y$ -方向同时采用二阶中心差分近似, 可得

$$\begin{aligned} -\frac{\partial^2 u(x, y)}{\partial x^2} \Big|_{(x_i, y_j)} &\approx \frac{2u(x_i, y_j) - u(x_{i-1}, y_j) - u(x_{i+1}, y_j)}{h^2}, \\ -\frac{\partial^2 u(x, y)}{\partial y^2} \Big|_{(x_i, y_j)} &\approx \frac{2u(x_i, y_j) - u(x_i, y_{j-1}) - u(x_i, y_{j+1})}{h^2}. \end{aligned}$$

代入 (6.10) 后, 就可以得到二维 Poisson 方程在  $(x_i, y_j)$  点的近似离散方程

$$4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} = h^2 f_{i,j},$$

其中  $f_{ij} = f(x_i, y_j)$ ,  $u_{i,j}$  为  $u(x_i, y_j)$  的近似. 这个离散格式可以用下面的图 (左图) 来描述.



写成矩阵形式即为

$$T_N u = h^2 f, \quad (6.11)$$

其中

$$T_N \triangleq I \otimes T_n + T_n \otimes I, \quad u = [u_{1,1}, \dots, u_{n,1}, u_{1,2}, \dots, u_{n,2}, \dots, u_{1,n}, \dots, u_{n,n}].$$

这里  $\otimes$  表示 Kronecker 乘积,  $T_n$  为一维 Poisson 方程离散后的系数矩阵, 即 (6.9). 上图 (右图) 为  $N = 6^2$  时的矩阵示例图.

需要注意的是, 系数矩阵与网格点的排序有关, 不同的排序方式对应不同的系数矩阵. 这里是按自然顺序排列的.

类似地, 如果对三维 poisson 方程进行中心差分离散, 则对应的系数矩阵为

$$T_n \otimes I \otimes I + I \otimes T_n \otimes I + I \otimes I \otimes T_n.$$

### 系数矩阵 $T_N$ 的性质



**定理 6.4** 设  $T_n = Z\Lambda Z^T$ , 其中  $Z = [z_1, z_2, \dots, z_n]$  为正交阵,  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  为对角阵, 则  $T_N$  的特征值分解为

$$T_N = (Z \otimes Z)(I \otimes \Lambda + \Lambda \otimes I)(Z \otimes Z)^T,$$

即  $T_N$  的特征值为  $\lambda_i + \lambda_j$ , 对应的特征向量为  $z_i \otimes z_j$ ,  $i, j = 1, 2, \dots, n$ .

由于  $T_N$  对称正定, 其条件数为

$$\kappa(T_N) = \frac{\lambda_{\max}(T_N)}{\lambda_{\min}(T_N)} = \frac{1 - \cos \frac{n\pi}{n+1}}{1 - \cos \frac{\pi}{n+1}} = \frac{\sin^2 \frac{n\pi}{2(n+1)}}{\sin^2 \frac{\pi}{2(n+1)}} \approx \frac{4(n+1)^2}{\pi^2}.$$

故当  $n$  越来越大时,  $\kappa(T_N) \rightarrow \infty$ , 即  $T_N$  越来越病态.

### 求解二维离散 Poisson 方程的 Jacobi 方法

**算法 6.5.** 求解二维离散 Poisson 方程的 Jacobi 迭代法

```

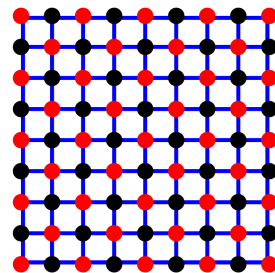
1: Given an initial guess $u^{(0)} \in \mathbb{R}^N$
2: while not converge do
3: for $i = 1$ to n do
4: for $j = 1$ to n do
5: $u_{i,j}^{(k+1)} = \frac{1}{4} \left(h^2 f_{i,j} + u_{i+1,j}^{(k)} + u_{i-1,j}^{(k)} + u_{i,j+1}^{(k)} + u_{i,j-1}^{(k)} \right)$
6: end for
7: end for
8: end while

```

### 求解二维离散 Poisson 方程的 G-S 方法

基于自然顺序的 G-S 方法不适合并行计算. 下面我们介绍一种新的适合并行计算的未知量排序方法: **红黑排序**, 即将二维网格点依次做红黑记号, 如右图所示.

在计算过程中, 对未知量的值进行更新时, 我们可以先更新红色节点, 此时所使用的只是黑色节点的数据, 然后再更新黑色节点, 这时使用的是红色节点的数据.



由于在更新红点时, 各个红点之间是相互独立的, 因此可以并行计算. 同样, 在更新黑点时, 各个黑点之间也是相互独立的, 因此也可以并行计算.

**算法 6.6.** 求解二维离散 Poisson 方程的红黑排序 G-S 迭代法

```

1: Given an initial guess $u^{(0)} \in \mathbb{R}^N$
2: while not converge do

```

```

3: for (i, j) 为红色节点 do
4: $u_{i,j}^{(k+1)} = \frac{1}{4} \left(h^2 f_{i,j} + u_{i+1,j}^{(k)} + u_{i-1,j}^{(k)} + u_{i,j+1}^{(k)} + u_{i,j-1}^{(k)} \right)$
5: end for
6: for (i, j) 为黑色节点 do
7: $u_{i,j}^{(k+1)} = \frac{1}{4} \left(h^2 f_{i,j} + u_{i+1,j}^{(k+1)} + u_{i-1,j}^{(k+1)} + u_{i,j+1}^{(k+1)} + u_{i,j-1}^{(k+1)} \right)$
8: end for
9: end while

```


### 求解二维离散 Poisson 方程的 SOR 方法

#### 算法 6.7. 求解二维离散 Poisson 方程的 SOR 迭代法

```

1: Given an initial guess $u^{(0)} \in \mathbb{R}^N$ and a parameter ω
2: while not converge do
3: for $i = 1$ to n do
4: for $j = 1$ to n do
5: $u_{i,j}^{(k+1)} = (1 - \omega)u_{i,j}^{(k)} + \frac{\omega}{4} \left(h^2 f_{i,j} + u_{i+1,j}^{(k)} + u_{i-1,j}^{(k)} + u_{i,j+1}^{(k)} + u_{i,j-1}^{(k)} \right)$
6: end for
7: end for
8: end while

```

 上面的 SOR 算法是基于自然排序, 基于红黑排序的 SOR 算法请读者自己写.

#### 例 6.1 已知二维 Poisson 方程

$$\begin{cases} -\Delta u(x, y) = -1, & (x, y) \in \Omega \\ u(x, y) = \frac{x^2 + y^2}{4}, & (x, y) \in \partial\Omega \end{cases}$$

其中  $\Omega = (0, 1) \times (0, 1)$ . 该方程的解析解是  $u(x, y) = \frac{x^2 + y^2}{4}$ . 用五点差分格式离散后得到一个线性方程组.

(1) 分别用 Jacobi, G-S 和 SOR 方法计算这个方程组的解.

(2) 分别用 SOR 和 SSOR 方法求解方程, 观察参数  $\omega$  对方法收敛的影响.

**解.** (1) MATLAB 程序参见 [Poisson\\_Jacobi\\_GS\\_SOR.m](#). 定义近似解的相对误差:

$$\text{relerr}_k \triangleq \frac{\|u^{(k)} - u_*\|_2}{u_*},$$

其中  $u_*$  表示精确解. 下图画出了  $n = 16$  (即  $N = 256$ ) 时三种方法的近似解相对误差的下降过程 (前 100 个迭代步).





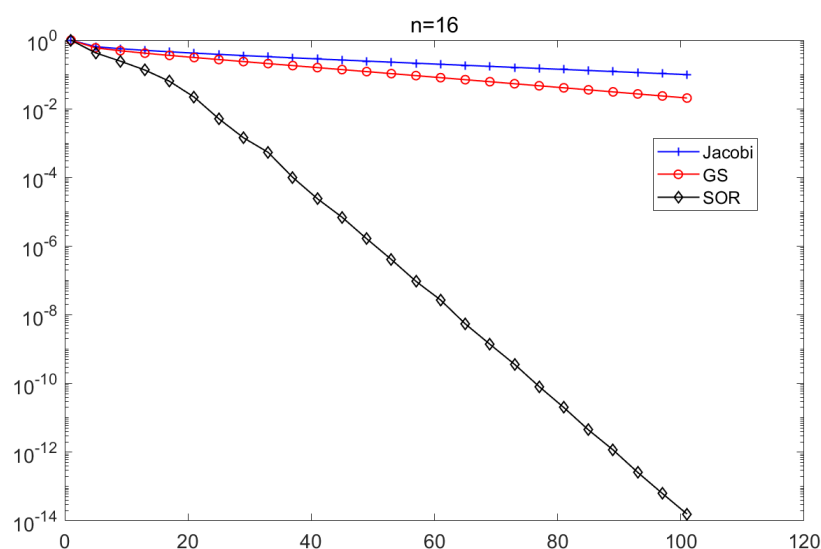


图 6.1. Jacobi, G-S 和 SOR 的近似解相对误差的下降曲线.

(2) MATLAB 程序参见 [Poisson\\_SOR\\_omega.m](#) 和 [Poisson\\_SSOR\\_omega.m](#). 下图中画出了  $n = 8$  (即  $N = 64$ ) 时, SOR 和 SSOR 收敛结果与参数  $\omega$  取值之间的关系.

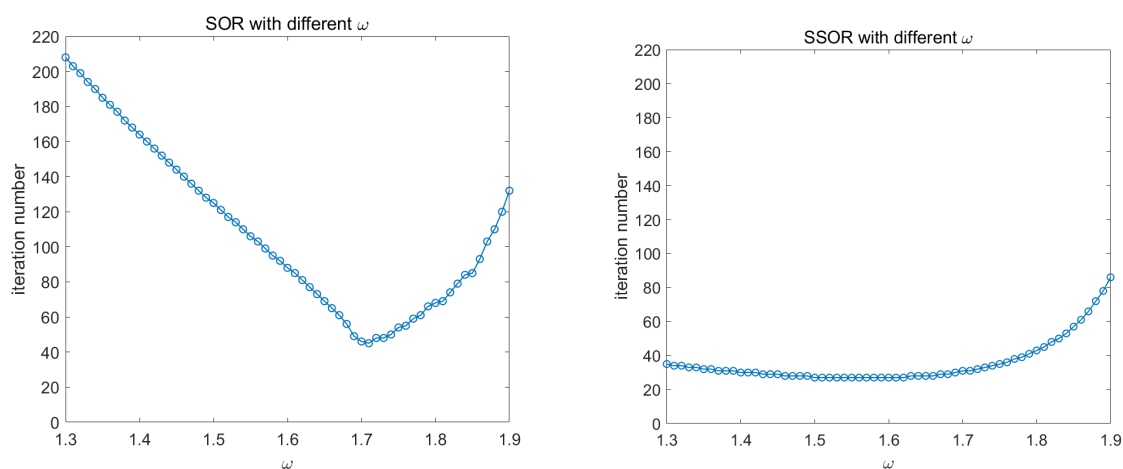


图 6.2. SOR 和 SSOR 的收敛结果与  $\omega$  取值的关系.

在编写程序时, 我们不需要生成和存放系数矩阵. 此外, 我们用矩阵形式来存放未知量  $u_{ij}$ , 是为了与算法描述相一致, 便于计算矩阵向量乘积,

### 6.3.3 求解方法小结

由于 Poisson 方程的特殊结构和性质, 除了常规求解方法以外, 人们还设计出了一些特殊的快速算法.



假定网格剖分为  $n \times n$ , 并记  $N = n^2$ .

|              | 方法                   | 串行时间                    | 存储空间                    |
|--------------|----------------------|-------------------------|-------------------------|
| 直接法          | 稠密 Cholesky 分解       | $\mathcal{O}(N^3)$      | $\mathcal{O}(N^2)$      |
|              | 显式求逆                 | $\mathcal{O}(N^2)$      | $\mathcal{O}(N^2)$      |
|              | 带状 Cholesky 分解       | $\mathcal{O}(N^2)$      | $\mathcal{O}(N^{3/2})$  |
|              | 稀疏 Cholesky 分解       | $\mathcal{O}(N^{3/2})$  | $\mathcal{O}(N \log N)$ |
| 基本迭代法        | Jacobi               | $\mathcal{O}(N^2)$      | $\mathcal{O}(N)$        |
|              | Gauss-Seidel         | $\mathcal{O}(N^2)$      | $\mathcal{O}(N)$        |
|              | SOR                  | $\mathcal{O}(N^{3/2})$  | $\mathcal{O}(N)$        |
|              | 带 Chebyshev 加速的 SSOR | $\mathcal{O}(N^{5/4})$  | $\mathcal{O}(N)$        |
| Krylov 子空间迭代 | CG (共轭梯度法)           | $\mathcal{O}(N^{3/2})$  | $\mathcal{O}(N)$        |
|              | CG (带修正 IC 预处理)      | $\mathcal{O}(N^{5/4})$  | $\mathcal{O}(N)$        |
| 快速方法         | DST (快速 Sine 变换)     | $\mathcal{O}(N \log N)$ | $\mathcal{O}(N)$        |
|              | 块循环约化                | $\mathcal{O}(N \log N)$ | $\mathcal{O}(N)$        |
|              | Multigrid            | $\mathcal{O}(N)$        | $\mathcal{O}(N)$        |



## 6.4 收敛性分析

### 6.4.1 向量与矩阵序列收敛基本概念

首先给出向量序列收敛的定义.

**定义 6.2 (向量序列的收敛)** 设  $\{x^{(k)}\}_{k=1}^{\infty}$  是  $\mathbb{R}^n$  (或  $\mathbb{C}^n$ ) 中的一个向量序列. 如果存在向量  $x = [x_1, x_2, \dots, x_n]^T \in \mathbb{R}^n$  (或  $\mathbb{C}^n$ ) 使得

$$\lim_{k \rightarrow \infty} x_i^{(k)} = x_i, \quad i = 1, 2, \dots, n,$$

其中  $x_i^{(k)}$  表示  $x^{(k)}$  的第  $i$  个分量, 则称  $\{x^{(k)}\}$  (按分量) 收敛到  $x$ , 即  $x$  是  $x^{(k)}$  的极限, 记为

$$\lim_{k \rightarrow \infty} x^{(k)} = x.$$

类似地, 我们可以给出矩阵序列收敛的定义.

**定义 6.3 (矩阵序列的收敛)** 设  $\{A^{(k)} = [a_{ij}^{(k)}]\}_{k=0}^{\infty}$  是  $\mathbb{R}^{m \times n}$  (或  $\mathbb{C}^{m \times n}$ ) 中的一个矩阵序列. 如果存在矩阵  $A = [a_{ij}] \in \mathbb{R}^{m \times n}$  (或  $\mathbb{C}^{m \times n}$ ) 使得

$$\lim_{k \rightarrow \infty} a_{ij}^{(k)} = a_{ij}, \quad i = 1, 2, \dots, m, j = 1, 2, \dots, n,$$

则称  $A^{(k)}$  收敛到  $A$ , 即  $A$  是  $A^{(k)}$  的极限, 记为

$$\lim_{k \rightarrow \infty} A^{(k)} = A.$$

关于向量序列和矩阵序列的收敛性, 我们有下面的基本判别方法 [137].

**定理 6.5** 设向量序列  $\{x^{(k)}\}_{k=0}^{\infty} \subset \mathbb{R}^n$  (或  $\mathbb{C}^n$ ), 矩阵序列  $\{A^{(k)} = [a_{ij}^{(k)}]\}_{k=0}^{\infty} \subset \mathbb{R}^{m \times n}$  (或  $\mathbb{C}^{m \times n}$ ), 则

- (1)  $\lim_{k \rightarrow \infty} x^{(k)} = x \iff \lim_{k \rightarrow \infty} \|x^{(k)} - x\| = 0$ , 其中  $\|\cdot\|$  为任一向量范数;
- (2)  $\lim_{k \rightarrow \infty} A^{(k)} = A \iff \lim_{k \rightarrow \infty} \|A^{(k)} - A\| = 0$ , 其中  $\|\cdot\|$  为任一矩阵范数;
- (3)  $\lim_{k \rightarrow \infty} A^{(k)} = 0 \iff \lim_{k \rightarrow \infty} A^{(k)}x = 0, \forall x \in \mathbb{R}^n$  (或  $\mathbb{C}^n$ ).

由定理 6.5 和引理 1.40, 我们可以立即得到下面的结论.

**定理 6.6** 设矩阵  $A \in \mathbb{R}^{n \times n}$  (或  $\mathbb{C}^{n \times n}$ ), 则  $\lim_{k \rightarrow \infty} A^k = 0$  当且仅当  $\rho(A) < 1$ .

(板书)

**证明.** 充分性: 设  $\rho(A) < 1$ , 则由引理 1.40 可知, 存在某个矩阵范数  $\|\cdot\|$  使得  $\|A\| < 1$ . 因此由  $0 \leq \|A^k\| \leq \|A\|^k \rightarrow 0$  ( $k \rightarrow \infty$ ) 可得

$$\lim_{k \rightarrow \infty} \|A^k\| = 0.$$

根据定理 6.5, 我们有  $\lim_{k \rightarrow \infty} A^k = 0$ .



必要性: 设  $\lim_{k \rightarrow \infty} A^k = 0$ , 则由定理 6.5 可知

$$\lim_{k \rightarrow \infty} \|A^k\| = 0.$$

所以由  $0 \leq \rho(A)^k = \rho(A^k) \leq \|A^k\|$  可得

$$\lim_{k \rightarrow \infty} \rho(A)^k = 0,$$

即  $\rho(A) < 1$ . □

下面给出谱半径与范数之间的一个非常重要的性质, 有时也用来定义矩阵的谱半径.

**定理 6.7** 设  $A \in \mathbb{R}^{n \times n}$  (或  $\mathbb{C}^{n \times n}$ ), 则对任意矩阵范数  $\|\cdot\|$ , 有

$$\rho(A) = \lim_{k \rightarrow \infty} \|A^k\|^{\frac{1}{k}}.$$

(板书)

**证明.** 由谱半径与范数之间的关系可知

$$\rho(A)^k = \rho(A^k) \leq \|A^k\|.$$

另一方面, 对任意  $\varepsilon > 0$ , 构造矩阵

$$A_\varepsilon = \frac{1}{\rho(A) + \varepsilon} A.$$

则  $\rho(A_\varepsilon) < 1$ , 故  $\lim_{k \rightarrow \infty} A_\varepsilon^k = 0$ . 因此存在正整数  $N$ , 使得当  $k > N$  时, 有  $\|A_\varepsilon^k\| < 1$ , 即

$$\left\| \frac{A^k}{(\rho(A) + \varepsilon)^k} \right\| = \frac{\|A^k\|}{(\rho(A) + \varepsilon)^k} < 1.$$

所以

$$\rho(A) \leq \|A^k\|^{\frac{1}{k}} \leq \rho(A) + \varepsilon,$$

即对任意  $\varepsilon > 0$ , 存在正整数  $N$ , 使得当  $k > N$  时, 有

$$\left| \|A^k\|^{\frac{1}{k}} - \rho(A) \right| < \varepsilon.$$

根据极限定义可知,  $\|A^k\|^{\frac{1}{k}} \rightarrow \rho(A)$ . □

下面是关于 **收敛速度** 的定义.

**定义 6.4** 设点列  $\{\varepsilon_k\}_{k=1}^\infty$  收敛, 且  $\lim_{k \rightarrow \infty} \varepsilon_k = 0$ . 若存在一个有界常数  $0 < c < \infty$ , 使得

$$\lim_{k \rightarrow \infty} \frac{|\varepsilon_{k+1}|}{|\varepsilon_k|^p} = c,$$

则称点列  $\{\varepsilon_k\}$  是  **$p$  次 (渐进) 收敛** 的. 若  $1 < p < 2$  或  $p = 1$  且  $c = 0$ , 则称点列是 **超线性收敛** 的.

## 6.4.2 定常迭代法的收敛性

首先给出定常迭代法收敛的定义.



**定义 6.5** 考虑定常迭代法 (6.3), 如果对任意的初始向量  $x^{(0)}$ , 都有

$$\lim_{k \rightarrow \infty} x^{(k)} \rightarrow x_*,$$

则称定常迭代法 (6.3) 是**收敛**的, 否则就称其为**发散**的.

 基于矩阵分裂的迭代法, 其收敛性取决于迭代矩阵的谱半径.

下面考虑定常迭代法 6.3 的收敛性. 首先给出一个迭代法收敛的充分条件.

**引理 6.8** 若存在算子范数  $\|\cdot\|$ , 使得  $\|G\| < 1$ , 则迭代法 6.3 收敛.

(板书)

**证明.** 由  $x^{(k+1)} = Gx^{(k)} + g$  和  $x_* = Gx_* + g$  可得

$$x^{(k+1)} - x_* = G(x^{(k)} - x_*).$$

故


$$\|x^{(k+1)} - x_*\| \leq \|G\| \cdot \|x^{(k)} - x_*\|.$$

依此类推, 可得

$$\|x^{(k+1)} - x_*\| \leq \|G\|^{k+1} \cdot \|x^{(0)} - x_*\|.$$

由于  $\|G\| < 1$ , 当  $k \rightarrow \infty$  时, 上式右端  $\rightarrow 0$ , 即  $\|x^{(k+1)} - x_*\| \rightarrow 0$ . 因此方法收敛.  $\square$

 事实上, 由习题 ?? 可知, 引理 6.8 条件中的算子范数可以改为任意矩阵范数.

 我们记  $e^{(k)} \triangleq x^{(k)} - x_*$  为第  $k$  步迭代解  $x^{(k)}$  的**误差向量**.

**定理 6.9 (收敛性定理)** 对任意迭代初始向量  $x^{(0)}$ , 迭代法 6.3 收敛的充要条件是  $\rho(G) < 1$ .

(板书)

**证明.** 必要性: 用反证法, 假设  $\rho(G) \geq 1$ . 设  $\lambda$  为  $G$  的模最大的特征值, 即  $|\lambda| = \rho(G) \geq 1$ . 令  $x \neq 0$  为其对应的特征向量. 取迭代初始向量  $x^{(0)} = x_* + x$ , 则

$$x^{(k)} - x_* = G(x^{(k-1)} - x_*) = \cdots = G^k(x^{(0)} - x_*) = G^k x = \lambda^k x,$$

不可能收敛到 0, 即方法不收敛. 故假设不成立, 因此  $\rho(G) < 1$ .

充分性: 若  $\rho(G) < 1$ , 则由引理 1.40 可知, 存在一个算子范数  $\|\cdot\|_\varepsilon$ , 使得  $\|G\|_\varepsilon < 1$ . 再由引理 6.8 可知, 方法收敛.  $\square$

**定义 6.6** 设  $G$  是迭代矩阵, 则迭代法 6.3 的**平均收敛速度**定义为

$$R_k(G) \triangleq -\ln \|G^k\|^{\frac{1}{k}},$$

渐进收敛速度定义为

$$R(G) \triangleq \lim_{k \rightarrow \infty} R_k(G) = -\ln \rho(G).$$

平均收敛速度与迭代步数和所用的范数有关,但渐进收敛速度只依赖于迭代矩阵的谱半径.

**定理 6.10** 考虑迭代法 6.3. 如果存在某个算子范数  $\|\cdot\|$  使得  $\|G\| = q < 1$ , 则

- (1)  $\|x^{(k)} - x_*\| \leq q^k \|x^{(0)} - x_*\|$ ;
- (2)  $\|x^{(k)} - x_*\| \leq \frac{q}{1-q} \|x^{(k)} - x^{(k-1)}\|$ ;
- (3)  $\|x^{(k)} - x_*\| \leq \frac{q^k}{1-q} \|x^{(1)} - x^{(0)}\|$ .

(板书)

证明.

- (1) 由  $x^{(k)} = Gx^{(k-1)} + g$  和  $x_* = Gx_* + g$  可得

$$x^{(k)} - x_* = G(x^{(k-1)} - x_*).$$

因此有

$$\|x^{(k)} - x_*\| \leq \|G\| \cdot \|x^{(k-1)} - x_*\| = q \|x^{(k-1)} - x_*\|. \quad (6.12)$$

反复利用结论 (6.12) 即可得  $\|x^{(k)} - x_*\| \leq q^k \|x^{(0)} - x_*\|$ .

- (2) 由  $x^{(k+1)} = Gx^{(k)} + g$  和  $x_* = Gx_* + g$  可得

$$x^{(k+1)} - x^{(k)} = G(x^{(k)} - x^{(k-1)})$$

因此有

$$\|x^{(k+1)} - x^{(k)}\| \leq q \|x^{(k)} - x^{(k-1)}\|, \quad (6.13)$$

根据 (6.12), 我们有  $\|x^{(k+1)} - x_*\| \leq q \|x^{(k)} - x_*\|$ . 所以

$$\begin{aligned} \|x^{(k+1)} - x^{(k)}\| &= \|x_* - x^{(k)} + x^{(k+1)} - x_*\| \\ &\geq \|x_* - x^{(k)}\| - \|x^{(k+1)} - x_*\| \\ &\geq (1-q) \|x^{(k)} - x_*\|. \end{aligned}$$

结合 (6.13) 可得

$$\|x_* - x^{(k)}\| \leq \frac{1}{1-q} \|x^{(k+1)} - x^{(k)}\| \leq \frac{q}{1-q} \|x^{(k)} - x^{(k-1)}\|.$$

- (3) 反复利用 (6.13) 即可得

$$\|x^{(k+1)} - x^{(k)}\| \leq q^k \|x^{(1)} - x^{(0)}\|.$$

所以

$$\|x_* - x^{(k)}\| \leq \frac{1}{1-q} \|x^{(k+1)} - x^{(k)}\| \leq \frac{q^k}{1-q} \|x^{(1)} - x^{(0)}\|.$$

□



一般来说, 好的迭代法应该满足:

- (1)  $\rho(G)$  很小;
- (2) 以  $M$  为系数矩阵的线性方程组比较容易求解.

### 6.4.3 二维离散 Poisson 方程情形

对于定常迭代法, 其收敛的充要条件是迭代矩阵的谱半径小于 1. 当谱半径不可求时, 我们可以根据迭代矩阵的范数来判断, 即如果迭代矩阵的某个算子范数小于 1, 则方法也收敛.

本小节考虑求解二维离散 Poisson 方程的 Jacobi, G-S 和 SOR 方法的收敛性. 考虑这些方法的收敛性, 只需研究相应的迭代矩阵的谱半径即可. 对于二维离散 Poisson 方程, 系数矩阵为

$$A = T = I \otimes T_n + T_n \otimes I.$$

故 Jacobi 方法的迭代矩阵为

$$G_J = D^{-1}(L + U) = (4I)^{-1}(4I - T) = I - T/4. \quad (6.14)$$

由于  $T$  的特征值为

$$\lambda_i + \lambda_j = 2 \left( 1 - \cos \frac{\pi i}{n+1} \right) + 2 \left( 1 - \cos \frac{\pi j}{n+1} \right) = 4 - 2 \left( \cos \frac{\pi i}{n+1} + \cos \frac{\pi j}{n+1} \right),$$

所以  $G_J$  的特征值为

$$1 - (\lambda_i + \lambda_j)/4 = \frac{1}{2} \left( \cos \frac{\pi i}{n+1} + \cos \frac{\pi j}{n+1} \right).$$

故

$$\rho(G_J) = \frac{1}{2} \max_{i,j} \left\{ \left| \cos \frac{\pi i}{n+1} + \cos \frac{\pi j}{n+1} \right| \right\} = \cos \frac{\pi}{n+1} < 1,$$

即 Jacobi 方法是收敛的.

注意当  $n$  越来越大时,  $\kappa(T) \rightarrow \infty$ , 即  $T$  越来越病态, 此时  $\rho(G_J) \rightarrow 1$ , 即 Jacobi 方法收敛越来越慢.

通常, 问题越病态就越难求解.

关于 G-S 方法和 SOR 方法, 我们有下面的结论.

**定理 6.11** 设  $G_{GS}$  和  $G_{SOR}$  分别表示求解二维 Poisson 方程的红黑排序的 G-S 方法和 SOR 方法的迭代矩阵, 则有

$$\rho(G_{GS}) = \rho(G_J)^2 = \cos^2 \frac{\pi}{n+1} < 1 \quad (6.15)$$

$$\rho(G_{SOR}) = \frac{\cos^2 \frac{\pi}{n+1}}{\left( 1 + \sin \frac{\pi}{n+1} \right)^2} < 1, \quad \omega = \frac{2}{1 + \sin \frac{\pi}{n+1}}. \quad (6.16)$$

在上述结论中, SOR 方法中的  $\omega$  是最优参数, 即此时的  $\rho(G_{SOR})$  最小. 由 Taylor 公式可知, 当



$n$  很大时, 有

$$\rho(G_J) = \cos \frac{\pi}{n+1} \approx 1 - \frac{\pi^2}{2(n+1)^2} = 1 - O\left(\frac{1}{n^2}\right),$$

$$\rho(G_{\text{SOR}}) = \frac{\cos^2 \frac{\pi}{n+1}}{\left(1 + \sin \frac{\pi}{n+1}\right)^2} \approx 1 - \frac{2\pi}{n+1} = 1 - O\left(\frac{1}{n}\right).$$

由于当  $n$  很大时有

$$\left(1 - \frac{1}{n}\right)^k \approx 1 - \frac{k}{n} = 1 - \frac{kn}{n^2} \approx \left(1 - \frac{1}{n^2}\right)^{kn},$$

即 SOR 方法迭代  $k$  步后误差的减小量与 Jacobi 方法迭代  $kn$  步后误差减小量差不多. 因此, 对于二维离散 Poisson 方程, 当 SOR 方法取最优参数时, 收敛速度大约是 Jacobi 方法的  $n$  倍.

这里需要指出的是, 对于一般线性方程组, 上述结论不一定成立.

由于  $\rho(G_{\text{GS}}) = \rho(G_J)^2$ , 因此, 对于二维离散 Poisson 方程, G-S 方法的收敛速度大约是 Jacobi 方法的 2 倍.

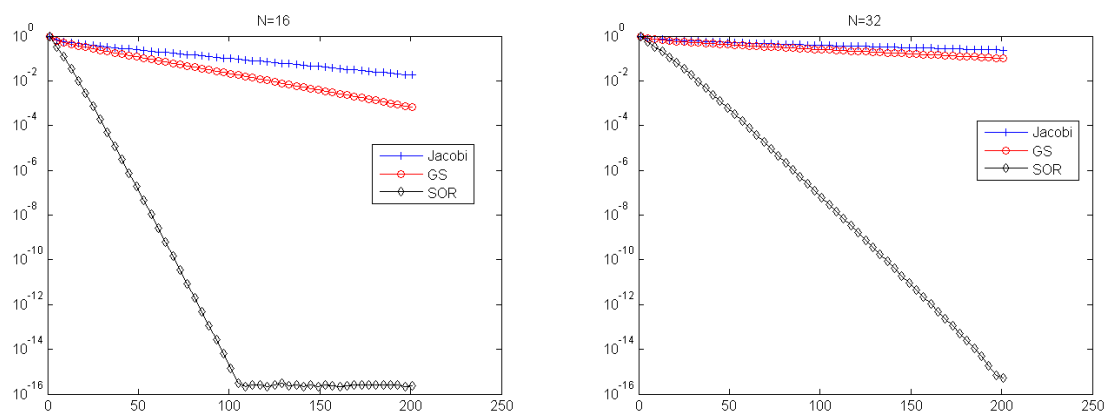
事实上, 当  $n$  很大时, 这三个方法的收敛速度都很慢.

**例 6.2** 已知二维 Poisson 方程

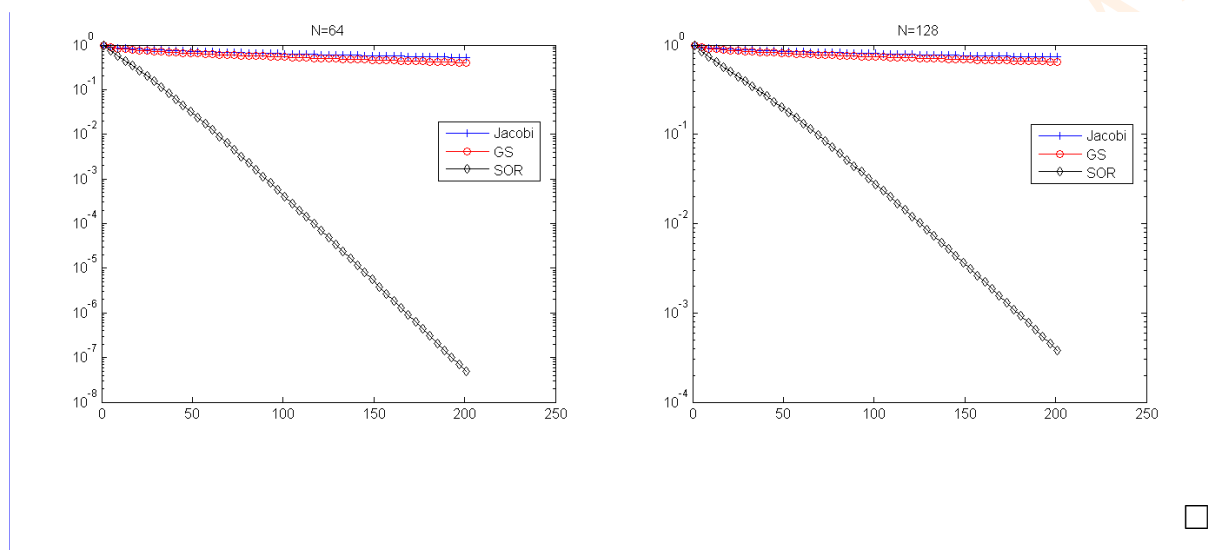
$$\begin{cases} -\Delta u(x, y) = -1, & (x, y) \in \Omega \\ u(x, y) = \frac{x^2 + y^2}{4}, & (x, y) \in \partial\Omega \end{cases}$$

其中  $\Omega = (0, 1) \times (0, 1)$ . 该方程的解析解是  $u(x, y) = \frac{x^2 + y^2}{4}$ . 用五点差分格式离散后得到一个线性方程组, 分别用 Jacobi, G-S 和 SOR 方法计算这个方程组的解, 并比较收敛效果.

**解.** 参见 MATLAB 程序 `Poisson_Jacobi_GS_SOR_convergence.m`. 下图中画出了  $N = 16, 32, 64, 128$  时, 这三种方法的相对误差下降情况.







#### 6.4.4 不可约对角占优矩阵

这里我们考虑  $A$  是严格对角占优或不可约弱对角占优情形. 由前面的结论可知,  $A$  是非奇异的.

**定理 6.12** 设  $A \in \mathbb{R}^{n \times n}$ , 若  $A$  严格对角占优, 则 Jacobi 方法和 G-S 方法都收敛, 且

$$\|G_{GS}\|_{\infty} \leq \|G_J\|_{\infty} < 1.$$

(板书)

**证明.** 首先证明  $\|G_J\|_{\infty} < 1$ . 由于  $A$  严格行对角占优, 故  $\sum_{j \neq i} |a_{ij}|/|a_{ii}| < 1$ . 所以

$$\|G_J\|_{\infty} = \|D^{-1}(L+U)\|_{\infty} = \max_{1 \leq i \leq n} \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

下面证明  $\|G_{GS}\|_{\infty} \leq \|G_J\|_{\infty}$ , 只需证明  $|G_{GS}|e \leq |G_J|e$  即可, 其中  $e = [1, 1, \dots, 1]^T$ .

令  $\tilde{L} = D^{-1}L$  和  $\tilde{U} = D^{-1}U$ . 则  $\tilde{L}$  是严格下三角矩阵, 故  $\tilde{L}^n = 0$ , 所以

$$(I - \tilde{L})^{-1} = I + \tilde{L} + \tilde{L}^2 + \dots + \tilde{L}^{n-1}.$$

于是

$$\begin{aligned} |G_{GS}|e &= |(D-L)^{-1}U|e = |(I - \tilde{L})^{-1}\tilde{U}|e \\ &= |(I + \tilde{L} + \tilde{L}^2 + \dots + \tilde{L}^{n-1})\tilde{U}|e \\ &\leq (I + |\tilde{L}| + |\tilde{L}|^2 + \dots + |\tilde{L}|^{n-1})|\tilde{U}|e \\ &= (I - |\tilde{L}|)^{-1}|\tilde{U}|e. \end{aligned} \quad (6.17)$$

由  $A$  的严格行对角占优性可知  $1 - \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} > 0$ , 即  $(I - |\tilde{L}| - |\tilde{U}|)e > 0$ . 两边同乘  $|\tilde{L}| \geq 0$  可得


$$\begin{aligned} 0 &\leq |\tilde{L}|(I - |\tilde{L}| - |\tilde{U}|)e = (|\tilde{L}| - |\tilde{L}|^2 + |\tilde{U}| - |\tilde{L}| \cdot |\tilde{U}| - |\tilde{U}|)e \\ &= ((I - |\tilde{L}|)(|\tilde{L}| + |\tilde{U}|) - |\tilde{U}|)e, \end{aligned}$$



即  $|\tilde{U}|e \leq (I - |\tilde{L}|)(|\tilde{L}| + |\tilde{U}|)e$ . 两边同乘  $(I - |\tilde{L}|)^{-1} \geq 0$  可得

$$(I - |\tilde{L}|)^{-1}|\tilde{U}|e \leq (|\tilde{L}| + |\tilde{U}|)e.$$


又  $|\tilde{L}| + |\tilde{U}| = |D^{-1}(L + U)| = |G_J|$ , 由 (6.17) 可知  $|G_{GS}|e \leq |G_J|e$ , 即定理结论成立.  $\square$

 当  $A$  是严格列对角占优时, 该结论也成立.


**定理 6.13** 设  $A \in \mathbb{R}^{n \times n}$ , 若  $A$  是弱对角占优且不可约, 则 Jacobi 方法和 G-S 方法都收敛. 进一步, 若  $A$  是非负矩阵, 则

$$\rho(G_{GS}) < \rho(G_J) < 1.$$

(留作练习, 可参见 [132])

 若  $A \in \mathbb{R}^{n \times n}$  严格对角占优且非负, 则是否有结论  $\rho(G_{GS}) \leq \rho(G_J)$ ?

 二维离散 Poisson 方程是弱行对角占优且不可约, 故对 Jacobi 方法和 G-S 方法都收敛.

 上述定理中的结论对一般矩阵并不成立: 对某些矩阵, Jacobi 方法收敛, 但 G-S 方法却不一定收敛, 见思考题 6.26.

关于 SOR 方法, 我们有下面的结论.

**定理 6.14** 设  $A \in \mathbb{R}^{n \times n}$ , 若  $A$  严格对角占优且  $0 < \omega \leq 1$ , 则 SOR 方法收敛.

(板书)

**证明.** 反证法. 假设 SOR 方法不收敛, 即  $\rho(G_{SOR}) \geq 1$ . 因此  $G_{SOR}$  存在特征值  $\lambda$ , 满足  $|\lambda| \geq 1$ . 由  $\det(\lambda I - G_{SOR}) = 0$  可知

$$\begin{aligned} \det(\lambda I - (D - \omega L)^{-1}((1 - \omega)D + \omega U)) &= \det((D - \omega L)^{-1}) \cdot \det(\lambda(D - \omega L) - (1 - \omega)D - \omega U) \\ &= \det((D - \omega L)^{-1}) \cdot \det((\lambda + \omega - 1)D - \lambda\omega L - \omega U) \\ &= 0. \end{aligned}$$

又  $\det((D - \omega L)^{-1}) \neq 0$ , 所以

$$\det((\lambda + \omega - 1)D - \lambda\omega L - \omega U) = 0.$$

由  $0 < \omega \leq 1$  和  $|\lambda| \geq 1$  可知,  $\lambda + \omega - 1 \neq 0$ . 所以  $\det(\tilde{G}) = 0$ , 其中

$$\tilde{G} = D - \frac{\lambda\omega}{\lambda + \omega - 1}L - \frac{\omega}{\lambda + \omega - 1}U. \quad (6.18)$$

令  $\lambda = a + \mathbf{i}b$ , 其中  $a, b \in \mathbb{R}$ . 则根据  $0 < \omega \leq 1$  和  $|\lambda| \geq 1$ , 可得

$$\begin{aligned} |\lambda + \omega - 1|^2 - |\lambda\omega|^2 &= (a + \omega - 1)^2 + b^2 - \omega^2(a^2 + b^2) \\ &= (1 - \omega)((a - 1)^2 + \omega(a^2 + b^2 - 1) + b^2) \\ &\geq 0. \end{aligned} \quad (6.19)$$



所以

$$\frac{|\omega|}{|\lambda + \omega - 1|} \leq \frac{|\lambda\omega|}{|\lambda + \omega - 1|} \leq 1.$$

又  $A$  是严格对角占优的, 所以由 (6.18) 和 (6.19) 可知,  $\tilde{G}$  也严格对角占优. 这意味着  $\det(\tilde{G}) \neq 0$ , 矛盾.  $\square$

**定理 6.15** 设  $A \in \mathbb{R}^{n \times n}$ , 若  $A$  是弱对角占优且不可约, 且  $0 < \omega \leq 1$ , 则 SOR 方法收敛.

### 6.4.5 对称正定矩阵

在给出收敛性结论之前, 也介绍两个需要用到的引理.

**引理 6.16** 设  $A \in \mathbb{C}^{n \times n}$  是 Hermite 的,  $A = M - N$ , 其中  $M$  非奇异, 则  $M^* + N$  也是 Hermite 的, 且对任意  $x \in \mathbb{C}^n$  有

$$x^*Ax - \tilde{x}^*A\tilde{x} = u^*(M^* + N)u,$$

其中  $\tilde{x} = M^{-1}Nx$ ,  $u = x - \tilde{x}$ .

(板书)

**证明.** 由于  $A$  是 Hermite 的, 所以  $M^* + N = M^* + M - A$  也是 Hermite 的.

由于  $\tilde{x} = M^{-1}Nx$ , 所以  $M\tilde{x} = Nx$ , 因此

$$Mu = Mx - M\tilde{x} = Mx - Nx = Ax,$$

$$Nu = Nx - N\tilde{x} = M\tilde{x} - N\tilde{x} = A\tilde{x}.$$

由  $M^* + N$  的对称性可知  $M = M^* + N - N^*$ . 又  $(Nx)^* = (M\tilde{x})^*$ , 所以

$$\begin{aligned} x^*Ax - \tilde{x}^*A\tilde{x} &= x^*Mu - \tilde{x}^*Nu \\ &= x^*(M^* + N - N^*)u - \tilde{x}^*Nu \\ &= x^*M^*u - x^*N^*u + x^*Nu - \tilde{x}^*Nu \\ &= x^*M^*u - \tilde{x}^*M^*u + u^*Nu \\ &= u^*(M^* + N)u. \end{aligned}$$

$\square$

**引理 6.17** 设  $A \in \mathbb{R}^{n \times n}$  对称,  $A = M - N$ , 其中  $M$  非奇异.

- (1) 如果  $A$  和  $M^T + N$  都是正定的, 则  $\rho(M^{-1}N) < 1$ ;
- (2) 如果  $\rho(M^{-1}N) < 1$  且  $M^T + N$  正定, 则  $A$  是正定的.

(板书)

**证明.** (1) 设  $\lambda \in \mathbb{C}$  是  $M^{-1}N$  的一个特征值, 对应的特征向量为  $x \neq 0$ , 即

$$M^{-1}Nx = \lambda x.$$

我们首先说明  $\lambda \neq 1$ . 假设  $\lambda = 1$ , 则可得  $Mx = Nx$ , 即  $Ax = 0$ . 由于  $A$  非奇异, 所以  $x = 0$ , 矛盾.



令  $\tilde{x} = M^{-1}Nx = \lambda x$ ,  $u = x - \tilde{x} = (1 - \lambda)x$ . 则由引理 6.16 可得

$$(1 - |\lambda|^2)x^*Ax = |1 - \lambda|^2x^*(M^T + N)x.$$

由于  $A$  和  $M^T + N$  都是正定矩阵, 所以上式右端为正, 故  $|\lambda| < 1$ . 因此  $\rho(M^{-1}N) < 1$ .

(2) 反证法. 假设  $A$  不是正定的. 由于  $\rho(M^{-1}N) < 1$ , 所以  $A = M(I - M^{-1}N)$  非奇异, 因此存在  $x^{(0)} \in \mathbb{R}^n$ , 使得

$$\eta \triangleq (x^{(0)})^T Ax^{(0)} < 0.$$

以  $x^{(0)}$  为初始点, 构造迭代序列

$$x^{(k)} = M^{-1}Nx^{(k-1)}, \quad k = 1, 2, \dots$$

由  $\rho(M^{-1}N) < 1$  可知

$$\lim_{k \rightarrow \infty} x^{(k)} = \lim_{k \rightarrow \infty} (M^{-1}N)^k x^{(0)} = 0. \quad (6.20)$$

令  $u^{(k)} = x^{(k-1)} - x^{(k)}$ , 则由引理 6.16 可得

$$(x^{(k-1)})^T Ax^{(k-1)} - (x^{(k)})^T Ax^{(k)} = (u^{(k)})^T (M^T + N)u^{(k)}.$$

由于  $M^T + N$  对称正定, 上式右端非负, 所以

$$(x^{(k)})^T Ax^{(k)} \leq (x^{(k-1)})^T Ax^{(k-1)}.$$

依此类推, 可得

$$(x^{(k)})^T Ax^{(k)} \leq (x^{(0)})^T Ax^{(0)} = \eta < 0.$$

这与 (6.20) 矛盾. 因此  $A$  一定是正定的. □

我们首先给出 SOR 迭代收敛的一个必要条件.

**定理 6.18** 对于 SOR 方法, 有  $\rho(G_{\text{SOR}}) \geq |1 - \omega|$ , 故 SOR 方法收敛的必要条件是  $0 < \omega < 2$ .

(板书)

**证明.** SOR 方法的迭代矩阵为

$$G_{\text{SOR}} = (D - \omega L)^{-1}((1 - \omega)D + \omega U) = (I - \omega \tilde{L})^{-1}((1 - \omega)I + \omega \tilde{U}).$$

所以  $G_{\text{SOR}}$  的行列式为

$$\det(G_{\text{SOR}}) = \det((I - \omega \tilde{L})^{-1}) \cdot \det((1 - \omega)I + \omega \tilde{U}) = (1 - \omega)^n.$$

设  $G_{\text{SOR}}$  的特征为  $\lambda_1, \lambda_2, \dots, \lambda_n$ , 则

$$\lambda_1 \lambda_2 \cdots \lambda_n = \det(G_{\text{SOR}}) = (1 - \omega)^n,$$

故至少有一个特征值的模不小于  $|1 - \omega|$ , 即  $\rho(G_{\text{SOR}}) \geq |1 - \omega|$ .

若 SOR 收敛, 则  $\rho(G_{\text{SOR}}) < 1$ , 因此  $|1 - \omega| < 1$ , 即  $0 < \omega < 2$ . □

**定理 6.19** 设  $A \in \mathbb{R}^{n \times n}$  对称正定.

- (1) 若  $2D - A$  正定, 则 Jacobi 迭代收敛.
- (2) 若  $0 < \omega < 2$ , 则 SOR 和 SSOR 收敛.



(3) G-S 迭代收敛.

(留作课外自习, 可利用引理 6.17)

 若系数矩阵对称正定, 则 SOR 方法收敛的充要条件是  $0 < \omega < 2$ .

 对于二维离散 Poisson 方程, 其系数矩阵是对称正定的, 故当  $0 < \omega < 2$  时, SOR 方法收敛.

**定理 6.20** 设  $A \in \mathbb{R}^{n \times n}$  对称, 且  $D$  正定.

- (1) 若 Jacobi 迭代收敛, 则  $A$  和  $2D - A$  都正定;
- (2) 若存在  $\omega \in (0, 2)$  使得 SOR (或 SSOR) 收敛, 则  $A$  正定;
- (3) 若 G-S 迭代收敛, 则  $A$  正定.

(板书, 只证明 (1), (2) (3) 利用引理 6.17 即可)

**证明.** 设 Jacobi 迭代收敛. 由于  $A$  对称,  $D$  对称正定, 所以  $D^{-1}A = D^{-\frac{1}{2}} \left( D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \right) D^{\frac{1}{2}}$  的特征值都是实数. 又

$$G_J = D^{-1}(D - A) = I - D^{-1}A,$$

由  $\rho(G_J) < 1$  可知  $\lambda(D^{-1}A) > 0$ , 即  $\lambda \left( D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \right) > 0$ , 所以  $\lambda(A) > 0$ , 故  $A$  正定.

下面证明  $2D - A$  的正定性. 根据  $\rho(G_J) < 1$  可知

$$0 < \lambda(D^{-1}A) < 2,$$

所以

$$0 < \lambda(2I - D^{-1}A) < 2.$$

又  $2D - A$  与  $D^{-\frac{1}{2}}(2D - A)D^{-\frac{1}{2}}$  合同, 而

$$D^{-\frac{1}{2}}(2D - A)D^{-\frac{1}{2}} = 2I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = D^{\frac{1}{2}}(2I - D^{-1}A)D^{-\frac{1}{2}},$$

即  $D^{-\frac{1}{2}}(2D - A)D^{-\frac{1}{2}}$  与  $2I - D^{-1}A$  有相同的特征值, 所以  $2D - A$  的特征值也都是正数, 即  $2D - A$  正定.  $\square$

**推论 6.21** 设  $A \in \mathbb{R}^{n \times n}$  对称且对角线均为正, 则 Jacobi 迭代收敛的充要条件是  $A$  和  $2D - A$  都正定.

#### 6.4.6 相容次序矩阵

针对一类特殊的矩阵, Jacobi 迭代矩阵和 SOR 迭代矩阵的特征值之间存在一种特殊的关系式, 根据这个关系式, 在某些场合下就可以确定 SOR 的最优参数的取值.

**定义 6.7** 设  $A \in \mathbb{R}^{n \times n}$  的对角线元素全不为零, 记  $\tilde{L} = D^{-1}L$ ,  $\tilde{U} = D^{-1}U$ , 即  $A = D(I - \tilde{L} - \tilde{U})$ . 若矩阵  $G(\alpha) \triangleq \alpha\tilde{L} + \frac{1}{\alpha}\tilde{U}$  的特征值与  $\alpha$  无关 ( $\alpha \neq 0$ ), 则称  $A$  是**相容次序矩阵**.

下面我们讨论一类具有相容次序的特殊矩阵. 首先给出一个引理.



**引理 6.22** 设  $B \in \mathbb{R}^{n \times n}$  具有下面的结构

$$B = \begin{bmatrix} 0 & B_{12} \\ B_{21} & 0 \end{bmatrix}, \quad \text{其中 } B_{12} \in \mathbb{R}^{k \times (n-k)}, B_{21} \in \mathbb{R}^{(n-k) \times k}, \quad 0 \leq k \leq n.$$

令  $B_L$  和  $B_U$  分别表示  $B$  的下三角和上三角部分, 则

- (1) 若  $\mu$  是  $B$  的特征值, 则  $-\mu$  也是  $B$  的特征值;
- (2)  $B(\alpha)$  的特征值与  $\alpha$  无关, 其中

$$B(\alpha) = \alpha B_L + \frac{1}{\alpha} B_U, \quad \alpha \neq 0.$$


(板书)

**证明.** (1) 若  $[x, y]$  是  $B$  的对应于  $\mu$  的特征向量, 则  $[x, -y]$  是  $B$  对应于  $-\mu$  的特征向量.

(2) 由于  $\alpha \neq 0$ , 我们有

$$\begin{bmatrix} I & 0 \\ 0 & \alpha I \end{bmatrix}^{-1} B(\alpha) \begin{bmatrix} I & 0 \\ 0 & \alpha I \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \frac{1}{\alpha} I \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{\alpha} B_{12} \\ \alpha B_{21} & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \alpha I \end{bmatrix} = \begin{bmatrix} 0 & B_{12} \\ B_{21} & 0 \end{bmatrix},$$

故引理结论成立. □

 该结论对复矩阵也成立.

 由引理 6.22 中结论 (2) 可知,  $B(\alpha) + \beta I$  的特征值也与  $\alpha$  无关, 其中  $\beta$  为任意常数.

**定义 6.8** 设  $A \in \mathbb{R}^{n \times n}$ , 如果存在一个置换矩阵  $P$ , 使得


$$PAP^T = \begin{bmatrix} D_1 & F \\ E & D_2 \end{bmatrix}, \quad (6.21)$$


其中  $D_1, D_2$  为对角矩阵, 则称  $A$  具有性质 A.

**例 6.3** 对于二维离散 Poisson 方程, 系数矩阵  $T_{N^2}$  具有性质 A. 事实上, 如果  $\tilde{T}_{N^2}$  是采用红黑排序后得到的系数矩阵, 则  $\tilde{T}_{N^2}$  具有 (6.21) 的结构.

根据引理 6.22, 我们可以直接得到下面的结论.

**定理 6.23** 设  $A$  的对角线元素全不为零, 若  $A$  具有性质 A, 则存在置换矩阵  $P$ , 使得  $PAP^T$  是相容次序矩阵.

 如果矩阵  $A$  本身就具有 (6.21) 的形式, 则  $A$  就是相容次序矩阵.

 事实上, (6.21) 可推广到分块三对角矩阵情形, 见下面的例子.

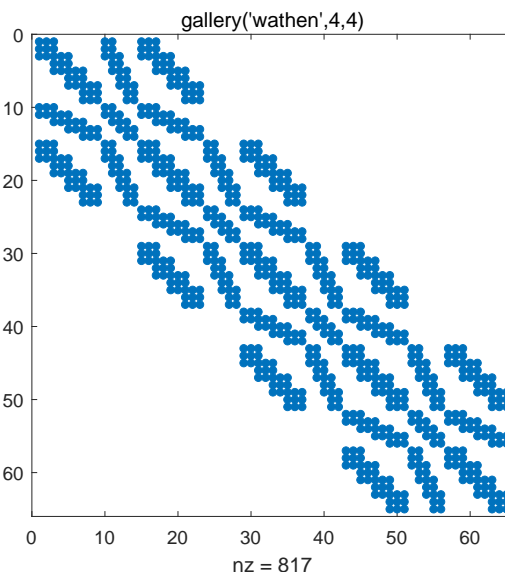
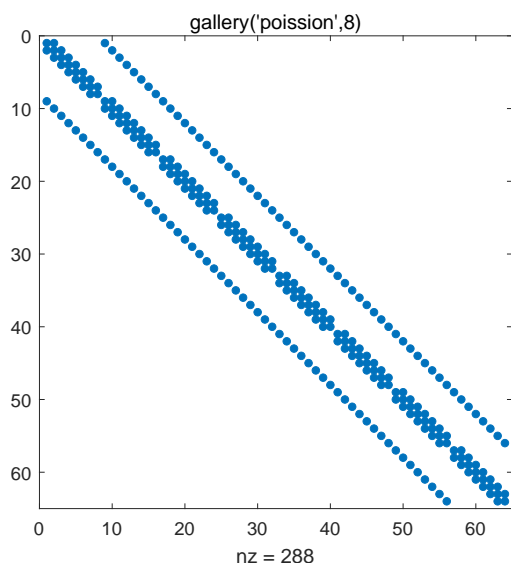


**例 6.4** 设  $D_i$  是非奇异的对角矩阵, 则任意块三对角矩阵

$$\begin{bmatrix} D_1 & A_1 & & \\ B_1 & \ddots & \ddots & \\ & \ddots & \ddots & A_{N-1} \\ & & B_{N-1} & D_N \end{bmatrix}$$

都是相容次序矩阵.

(留作练习)



分别用有限差分和有限元方法离散的二维偏微分方程

**定理 6.24** 设  $A$  是相容次序矩阵且  $\omega \neq 0$ , 则下列命题成立

- (1) Jacobi 迭代矩阵  $G_J$  的特征值正负成对出现;
- (2) 若  $\mu$  是  $G_J$  的一个特征值且  $\lambda$  满足

$$(\lambda + \omega - 1)^2 = \lambda \omega^2 \mu^2, \quad (6.22)$$

则  $\lambda$  是 SOR 迭代法矩阵  $G_{\text{SOR}}$  的特征值;

- (3) 反之, 若  $\lambda \neq 0$  是  $G_{\text{SOR}}$  的一个特征值且  $\mu$  满足 (6.22), 则  $\mu$  是  $G_J$  的特征值.

(板书)

**证明.** (1) 易知  $G_J = G(1)$ . 设  $\mu$  是  $G(1)$  的特征值, 因为  $G(-1) = -G(1)$ , 所以  $-\mu$  是  $G(-1)$  的特征值. 又  $G(\alpha)$  的特征值与  $\alpha$  无关, 故  $-\mu$  也是  $G(1)$  的特征值, 所以命题结论成立.

(2) 若  $\lambda = 0$ , 则由 (6.22) 可知  $\omega = 1$ , 此时  $G_{\text{SOR}} = (I - \tilde{L})^{-1}\tilde{U}$  为奇异矩阵, 故  $\lambda = 0$  是其特征值, 即命题结论成立.

若  $\lambda \neq 0$ , 则  $G_{\text{SOR}}$  的特征多项式为

$$\begin{aligned} \det(\lambda I - G_{\text{SOR}}) &= \det\left(\lambda I - (I - \omega \tilde{L})^{-1}((1 - \omega)I + \omega \tilde{U})\right) \\ &= \det\left((I - \omega \tilde{L})^{-1}\right) \cdot \det(\lambda(I - \omega \tilde{L}) - (1 - \omega)I - \omega \tilde{U}) \\ &= \det((\lambda + \omega - 1)I - \lambda \omega \tilde{L} - \omega \tilde{U}) \end{aligned}$$



$$\begin{aligned}
 &= \det \left( \sqrt{\lambda\omega^2} \left( \left( \frac{\lambda + \omega - 1}{\sqrt{\lambda\omega^2}} \right) I - \sqrt{\lambda} \tilde{L} - \frac{1}{\sqrt{\lambda}} \tilde{U} \right) \right) \\
 &= (\lambda\omega^2)^{n/2} \det \left( \left( \frac{\lambda + \omega - 1}{\sqrt{\lambda\omega^2}} \right) I - \tilde{L} - \tilde{U} \right), \quad (6.23)
 \end{aligned}$$

其中最后一个等式由引理 6.22 推得. 令

$$\mu = \frac{\lambda + \omega - 1}{\sqrt{\lambda\omega^2}} \quad \text{即} \quad (\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2,$$

则由 (6.23) 和性质 (1) 可知  $\lambda$  是  $G_{\text{SOR}}$  的特征值当且仅当  $\mu$  是  $G_J = \tilde{L} + \tilde{U}$  的特征值, 即命题结论成立.

(3) 已经由 (2) 证明. □

**推论 6.25** 设  $A$  是相容次序矩阵. 若  $G_J$  的特征值全部为实数, 则 SOR 迭代法收敛的充要条件是  $0 < \omega < 2$  且  $\rho(G_J) < 1$ . (留作练习)

**推论 6.26** 若  $A$  是相容次序矩阵, 则  $\rho(G_{\text{GS}}) = \rho(G_J)^2$ , 即当 Jacobi 迭代法收敛时, Gauss-Seidel 迭代法比 Jacobi 迭代法快一倍.

下面是关于 SOR 迭代法的最优参数选取.

**定理 6.27** 设  $A$  是相容次序矩阵. 若  $G_J$  的特征值全部为实数, 且  $\rho_J \triangleq \rho(G_J) < 1$ , 则 SOR 迭代法的最优参数为

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \rho_J^2}},$$

此时

$$\rho(G_{\text{SOR}}) = \omega_{\text{opt}} - 1 = \frac{\rho_J^2}{\left(1 + \sqrt{1 - \rho_J^2}\right)^2}.$$

进一步, 有

$$\rho(G_{\text{SOR}}) = \begin{cases} \omega - 1, & \omega_{\text{opt}} \leq \omega \leq 2 \\ 1 - \omega + \frac{1}{2}\omega^2\rho_J^2 + \omega\rho_J\sqrt{1 - \omega + \frac{1}{4}\omega^2\rho_J^2}, & 0 < \omega \leq \omega_{\text{opt}} \end{cases}.$$

(留作课外自习, 直接求解等式 (6.22), 分情况讨论即可, 可参见 [134, page 173])

**例 6.5** 采用红黑排序的二维离散 Poisson 问题的系数矩阵  $\tilde{T}_{N^2}$  是相容次序矩阵, 且  $G_J$  是对称的, 即  $G_J$  的特征值都是实的. 又由系数矩阵的弱对角占优和不可约性质可知  $\rho(G_J) < 1$ , 故上述定理的条件均满足.





## 6.5 加速方法

当迭代解  $x^{(0)}, x^{(1)}, x^{(2)}, \dots, x^{(k)}$  已经计算出来后, 我们可以对其进行组合, 得到一个新的近似解, 这样就可以对原方法进行加速.

### 6.5.1 外推技术


设原迭代格式为

$$x^{(k+1)} = Gx^{(k)} + g. \quad (6.24)$$

由  $x^{(k)}$  和  $x^{(k+1)}$  加权组合后可得新的近似解

$$x^{(k+1)} = (1 - \omega)x^{(k)} + \omega(Gx^{(k)} + g), \quad (6.25)$$

其中  $\omega$  是参数. 这种加速方法就称为 **外推方法**.

 如果原迭代格式是 Gauss-Seidel 迭代, 虽然 SOR 与外推方法 (6.25) 都是通过加权进行加速, 但格式是不一样的: 前者是在单个分量计算出来后就进行加权, 而后者则是当所有分量都计算出来后才加权.

为了使得迭代格式 (6.25) 尽可能快地收敛, 需要选择  $\omega$  使得其迭代矩阵  $G_\omega \triangleq (1 - \omega)I + \omega G$  的谱半径尽可能地小. 假设  $G$  的特征值都是实数, 且最大特征值和最小特征值分别为  $\lambda_1$  和  $\lambda_n$ . 于是

$$\rho(G_\omega) = \max_{\lambda \in \sigma(G)} |(1 - \omega) + \omega\lambda| = \max\{|1 - \omega + \omega\lambda_1|, |1 - \omega + \omega\lambda_n|\}.$$

**定理 6.28** 设  $G$  的特征值都是实数, 其最大和最小特征值分别为  $\lambda_1$  和  $\lambda_n$ , 且  $1 \notin [\lambda_n, \lambda_1]$ , 则

$$\omega_* = \arg \min_{\omega} \rho(G_\omega) = \frac{2}{2 - (\lambda_1 + \lambda_n)},$$

此时

$$\rho(G_{\omega_*}) = 1 - |\omega_*|d,$$

其中  $d$  是 1 到  $[\lambda_n, \lambda_1]$  的距离, 即当  $\lambda_n \leq \lambda_1 < 1$  时,  $d = 1 - \lambda_1$ , 当  $\lambda_1 \geq \lambda_n > 1$  时,  $d = \lambda_n - 1$ .

(板书)

**证明.** 易知  $|1 - \omega + \omega\lambda_1| = |1 - \omega + \omega\lambda_n|$  当且仅当  $\lambda_1 = \lambda_n$  或  $\omega = 0$  或  $\omega = \omega_* = \frac{2}{2 - (\lambda_1 + \lambda_n)}$ .

若  $\lambda_1 = \lambda_n$ , 则当  $\omega = \frac{1}{1 - \lambda_1} = \omega_*$  时  $\rho(G_{\omega_*})$  达到最小值 0.

下面假定  $\lambda_1 \neq \lambda_n$ . 先考虑  $\lambda_1 < 1$  的情形. 此时有

$$\max_{\lambda_n \leq \lambda \leq \lambda_1} |(1 - \omega) + \omega\lambda| = \begin{cases} 1 - \omega + \lambda_n\omega, & \omega \leq 0; \\ 1 - \omega + \lambda_1\omega, & 0 < \omega \leq \omega_*; \\ -(1 - \omega + \lambda_n\omega), & \omega > \omega_*. \end{cases}$$


通过函数图像可知, 当  $\omega = \omega_*$  时取最小值. 此时

$$\rho(G_{\omega_*}) = 1 - (1 - \lambda_1)\omega_* = 1 - \frac{2(1 - \lambda_1)}{2 - (\lambda_1 + \lambda_n)} = \frac{\lambda_1 - \lambda_n}{2 - (\lambda_1 + \lambda_n)}.$$



对于  $\lambda_n > 1$  的情形, 可以进行类似的讨论.  $\square$

由定理 6.28 可知,  $\rho(G_{\omega_*}) = 1 - |\omega_*|d$ , 且当  $\omega_* \neq 1$  时, 外推迭代 (6.25) 比原迭代法收敛要更快一些.

 最优参数依赖于原迭代矩阵  $G$  的特征值, 因此实用性不强. 在实际应用时可以估计特征值所在的区间  $[\alpha, \beta]$ , 然后用  $\alpha, \beta$  来代替  $\lambda_n$  和  $\lambda_1$ .

## JOR 方法

对 Jacobi 迭代进行外推加速, 则可得 JOR (Jacobi over-relaxation) 方法:

$$\begin{aligned} x^{(k+1)} &= (1 - \omega)x^{(k)} + \omega(D^{-1}(L + U)x^{(k)} + D^{-1}b) \\ &= x^{(k)} + \omega D^{-1}(b - Ax^{(k)}), \quad k = 0, 1, 2, \dots \end{aligned}$$

**定理 6.29** 设  $A$  对称正定. 若

$$0 < \omega < \frac{2}{\rho(D^{-1}A)},$$

则 JOR 方法收敛.

## 6.5.2 Chebyshev 多项式加速

本节对外推技巧进行推广.

假定通过迭代格式 (6.24) 已经计算出  $x^{(0)}, x^{(1)}, \dots, x^{(k)}$ , 下面考虑如何将这些近似解进行组合, 以便得到更精确的近似解.

记  $\varepsilon_k = x^{(k)} - x_*$  为第  $k$  步迭代解的误差, 则有

$$\varepsilon_k = G\varepsilon_{k-1} = G^2\varepsilon_{k-2} = \dots = G^k\varepsilon_0.$$

设  $\tilde{x}^{(k)}$  为  $x^{(0)}, x^{(1)}, \dots, x^{(k)}$  的一个线性组合, 即

$$\tilde{x}^{(k)} = \alpha_0 x^{(0)} + \alpha_1 x^{(1)} + \dots + \alpha_k x^{(k)}, \quad (6.26)$$

其中  $\alpha_i$  为待定系数, 且满足  $\sum_{i=0}^k \alpha_i = 1$ . 于是

$$\tilde{x}^{(k)} - x_* = \alpha_0 \varepsilon_0 + \alpha_1 G\varepsilon_0 + \dots + \alpha_k G^k \varepsilon_0 \triangleq p_k(G)\varepsilon_0, \quad (6.27)$$

其中  $p_k(t) = \sum_{i=0}^k \alpha_i t^i$  为  $k$  次多项式, 且满足  $p_k(1) = 1$ .

我们希望通过适当选取参数  $\alpha_i$ , 使得  $\tilde{x}^{(k)} - x_*$  尽可能地小, 即使得  $\tilde{x}^{(k)}$  收敛到  $x_*$  速度远远快于  $x^{(k)}$  收敛到  $x_*$  速度. 这种加速方法就称为 **多项式加速** 或 **半迭代法 (semi-iterative method)**.

**例 6.6** 设  $p_n(t)$  为  $G$  的特征多项式, 令  $\tilde{p}_n(t) \triangleq p_n(t)/p_n(1)$ , 则  $\tilde{p}_n(1) = 1$  且  $\tilde{p}_n(G) = 0$ , 所以选取  $\alpha_i$  为  $\tilde{p}_n$  的系数, 则  $\tilde{x}^{(n)} - x_* = 0$ . 但这种选取方法不实用, 原因是:



- (1)  $\tilde{p}_n(t)$  的系数并不知道;
- (2) 我们通常希望收敛所需的迭代步数  $\ll n$ .

下面讨论参数  $\alpha_i$  的较实用的选取方法. 由 (6.27) 可知

$$\|\tilde{x}^{(k)} - x_*\|_2 = \|p_k(G)\varepsilon_0\|_2 \leq \|p_k(G)\|_2 \cdot \|\varepsilon_0\|_2.$$

因此我们需要求解下面的极小化问题

$$\min_{p \in \mathbb{P}_k, p(1)=1} \|p(G)\|_2, \quad (6.28)$$

其中  $\mathbb{P}_k$  表示所有次数不超过  $k$  的多项式组成的集合. 一般来说, 这个问题是非常困难的. 但在一些特殊情况下, 我们可以给出其 (近似) 最优解.


假设迭代矩阵  $G$  是对称矩阵, 即  $G$  存在特征值分解

$$G = Q\Lambda Q^T,$$

其中  $\Lambda$  是对角矩阵, 且对角线元素都是实的,  $Q$  是正交矩阵. 于是有

$$\begin{aligned} \min_{p \in \mathbb{P}_k, p(1)=1} \|p(G)\|_2 &= \min_{p \in \mathbb{P}_k, p(1)=1} \|p(\Lambda)\|_2 \\ &= \min_{p \in \mathbb{P}_k, p(1)=1} \max_{1 \leq i \leq n} \{ |p(\lambda_i)| \} \\ &\leq \min_{p \in \mathbb{P}_k, p(1)=1} \max_{\lambda \in [\lambda_n, \lambda_1]} \{ |p(\lambda)| \}, \end{aligned} \quad (6.29)$$

其中  $\lambda_1, \lambda_n$  分别表示  $G$  的最大和最小特征值. 这是带归一化条件的多项式最佳一致逼近问题 (与零的偏差最小). 该问题的解与著名的 Chebyshev 多项式有关.

 由于所有算子范数  $\|p_k(G)\|$  的下确界是  $\rho(p_k(G))$ , 因此, 一种较实用的选取方法是使得  $p_k(G)$  的谱半径尽可能地小.

### Chebyshev 多项式

Chebyshev 多项式是一类很重要的正交多项式, 在函数逼近, 函数插值, 数值积分等方面都有着重要的应用.

Chebyshev 多项式  $T_k(t)$  可以通过下面的递归方式来定义:

$$\begin{aligned} T_0(t) &= 1, \quad T_1(t) = t, \\ T_k(t) &= 2tT_{k-1}(t) - T_{k-2}(t), \quad k = 2, 3, \dots, \end{aligned} \quad (6.30)$$

也可以直接由下面的式子定义

$$T_k(t) = \begin{cases} \cos(k \arccos t), & |t| \leq 1 \\ \cosh(k \operatorname{arccosh} t), & |t| > 1 \end{cases},$$

其中  $\cosh$  为双曲余弦, 即  $\cosh(t) = \frac{e^t + e^{-t}}{2}$ .



**引理 6.30** Chebyshev 多项式具有下面的性质:

- (1)  $T_k(1) = 1$ ;
- (2)  $T_k(t)$  的首项系数为  $2^{k-1}$ ;
- (3)  $T_k(-t) = (-1)^k T_k(t)$ , 即  $T_{2k}(t)$  只含偶次项,  $T_{2k+1}(t)$  只含奇次项;
- (4) 当  $|t| \leq 1$  时  $|T_k(t)| \leq 1$ ; 当  $|t| > 1$  时  $|T_k(t)| > 1$ ;
- (5)  $T_k(t) = 0$  的解为  $t_i = \cos \frac{(2i-1)\pi}{2k}$ ,  $i = 1, \dots, k$ ;
- (6)  $T_k(t)$  有  $n-1$  个极值点:  $t_i = \cos \frac{k\pi}{n}$ ,  $i = 1, 2, \dots, k-1$ ;

下面的结论表明, 在所有首项系数为 1 的  $k$  次多项式中,  $p_k(t) = \frac{1}{2^{k-1}} T_k(t)$  在  $[-1, 1]$  上与零的偏差是最小的 (在无穷范数意义下).

**定理 6.31** 设  $p_k(t) = \frac{1}{2^{k-1}} T_k(t)$ , 则

$$\max_{-1 \leq t \leq 1} |p_k(t)| \leq \max_{-1 \leq t \leq 1} |p(t)|, \quad \forall p(t) \in \tilde{\mathbb{P}}_k,$$

其中  $\tilde{\mathbb{P}}_k$  表示所有首项系数为 1 的  $k$  次多项式组成的集合, 即

$$\|p_k(t)\|_\infty = \min_{p(t) \in \tilde{\mathbb{P}}_k} \|p(t)\|_\infty.$$

这里的范数  $\|\cdot\|_\infty$  是  $C[-1, 1]$  上的无穷范数, 即  $\|f\|_\infty = \max_{x \in [-1, 1]} |f(x)|$ ,  $f(x) \in C[-1, 1]$ .

该性质可用于计算首项系数非零的  $n$  次多项式在  $[-1, 1]$  上的  $n-1$  次最佳一致逼近多项式. 利用这个性质, 我们可以采用 Chebyshev 多项式的零点作为节点进行多项式插值, 以使得插值的总体误差达到最小化.

Chebyshev 另外一个重要性质是下面的最小最大性质.

**定理 6.32** 设  $\eta \in \mathbb{R}$  满足  $|\eta| > 1$ , 则下面的最小最大问题

$$\min_{p(t) \in \mathbb{P}_k, p(\eta)=1} \max_{-1 \leq t \leq 1} |p(t)|$$

的唯一解为

$$\tilde{T}_k(t) \triangleq \frac{T_k(t)}{T_k(\eta)}.$$

通过简单的仿射变换, 该定理的结论可以推广到一般区间.

**定理 6.33** 设  $\alpha, \beta, \eta \in \mathbb{R}$  满足  $\alpha < \beta$  且  $|\eta| \notin [\alpha, \beta]$ . 则下面的最小最大问题

$$\min_{p(t) \in \mathbb{P}_k, p(\eta)=1} \max_{\alpha \leq x \leq \beta} |p(x)|$$



的唯一解为

$$\hat{T}_k(t) \triangleq \frac{T_k\left(\frac{2t - (\beta + \alpha)}{\beta - \alpha}\right)}{T_k\left(\frac{2\eta - (\beta + \alpha)}{\beta - \alpha}\right)},$$

其中  $\mathbb{P}_k$  表示所有次数不超过  $k$  的实系数多项式组成的集合.

### Chebyshev 加速方法

考虑迭代格式 (6.24), 我们假定:

- (1) 迭代矩阵  $G$  的特征值都是实数;
- (2) 迭代矩阵谱半径  $\rho = \rho(G) < 1$ , 故  $\lambda(G) \in [-\rho, \rho] \subset (-1, 1)$ .

于是最小最大问题 (6.29) 就转化为

$$\min_{p \in \mathbb{P}_k, p(1)=1} \max_{\lambda \in [-\rho, \rho]} \{|p(\lambda)|\}.$$

由于  $1 \notin [-\rho, \rho]$ , 根据定理 6.33, 上述问题的解为

$$p_k(t) = \frac{T_k(t/\rho)}{T_k(1/\rho)}.$$

下面考虑  $\tilde{x}^{(k)}$  的计算. 我们无需先计算出  $x^{(0)}, x^{(1)}, \dots, x^{(k)}$ , 然后再通过线性组合 (6.26) 来计算  $\tilde{x}^{(k)}$ . 事实上, 我们可以通过 Chebyshev 多项式的三项递推公式 (6.30), 由  $\tilde{x}^{(k-1)}$  和  $\tilde{x}^{(k-2)}$  直接计算出  $\tilde{x}^{(k)}$ . 这样做的另一个好处是无需存储所有的  $\tilde{x}^{(i)}$ . 下面给出具体的推导公式.

$$\text{令 } \mu_k = \frac{1}{T_k(1/\rho)}, \text{ 即 } T_k(1/\rho) = \frac{1}{\mu_k}. \text{ 由三项递推公式 (6.30) 可得}$$

$$\frac{1}{\mu_k} = \frac{2}{\rho} \cdot \frac{1}{\mu_{k-1}} - \frac{1}{\mu_{k-2}}.$$

所以

$$\begin{aligned} \tilde{x}^{(k)} - x_* &= p_k(G) \varepsilon_0 = \mu_k T_k(G/\rho) \varepsilon_0 \\ &= \mu_k \left[ \frac{2G}{\rho} \cdot T_{k-1}(G/\rho) - T_{k-2}(G/\rho) \right] \varepsilon_0 \\ &= \mu_k \left[ \frac{2G}{\rho} \cdot \frac{1}{\mu_{k-1}} p_{k-1}(G) \varepsilon_0 - \frac{1}{\mu_{k-2}} p_{k-2}(G) \varepsilon_0 \right] \\ &= \mu_k \left[ \frac{2G}{\rho} \cdot \frac{1}{\mu_{k-1}} (\tilde{x}^{(k-1)} - x_*) - \frac{1}{\mu_{k-2}} (\tilde{x}^{(k-2)} - x_*) \right]. \end{aligned}$$

整理后可得

$$\tilde{x}^{(k)} = \frac{2\mu_k}{\mu_{k-1}} \cdot \frac{G}{\rho} \tilde{x}^{(k-1)} - \frac{\mu_k}{\mu_{k-2}} \tilde{x}^{(k-2)} + d_k,$$

其中

$$\begin{aligned} d_k &= x_* - \frac{2\mu_k}{\mu_{k-1}} \cdot \frac{G}{\rho} x_* + \frac{\mu_k}{\mu_{k-2}} x_* \\ &= x_* - \frac{2\mu_k}{\mu_{k-1}} \cdot \frac{x_* - g}{\rho} + \frac{\mu_k}{\mu_{k-2}} x_* \\ &= \mu_k \left( \frac{1}{\mu_k} - \frac{2}{\rho \mu_{k-1}} + \frac{1}{\mu_{k-2}} \right) x_* + \frac{2\mu_k g}{\mu_{k-1} \rho} \end{aligned}$$



$$= \frac{2\mu_k g}{\mu_{k-1}\rho}.$$

由此, 我们可以得到迭代格式 (6.24) 的 Chebyshev 加速方法.

#### 算法 6.8. Chebyshev 加速方法

- 1: Set  $\mu_0 = 1, \mu_1 = \rho = \rho(G), \tilde{x}^{(0)} = x^{(0)}, k = 1$
- 2: compute  $\tilde{x}^{(1)} = Gx^{(0)} + g$
- 3: **while** not converge **do**
- 4:      $k = k + 1$
- 5:      $\mu_k = \left( \frac{2}{\rho} \cdot \frac{1}{\mu_{k-1}} - \frac{1}{\mu_{k-2}} \right)^{-1}$
- 6:      $\tilde{x}^{(k)} = \frac{2\mu_k}{\mu_{k-1}} \cdot \frac{G}{\rho} \tilde{x}^{(k-1)} - \frac{\mu_k}{\mu_{k-2}} \tilde{x}^{(k-2)} + \frac{2\mu_k}{\mu_{k-1}\rho} \cdot g$
- 7: **end while**

该方法的每步迭代中只有一次矩阵向量乘积, 故方法每个迭代步的整体运算量与原迭代格式的每个迭代步的运算量基本相当.

设  $\lambda(G) \in [\alpha, \beta]$ , 且  $-1 < \alpha \leq \beta < 1$ , 则我们也可以构造出相应的 Chebyshev 加速方法.

**例 6.7** 针对二维离散 Poisson 方程, 比较 Jacobi, JOR 和 Chebyshev 加速方法的收敛性质. (绘制收敛曲线)

#### SSOR 方法的 Chebyshev 加速

SSOR 迭代矩阵为

$$G_{\text{SSOR}} = (D - \omega U)^{-1} [(1 - \omega)D + \omega L] (D - \omega L)^{-1} [(1 - \omega)D + \omega U].$$

当  $A$  对称时, 有  $L = U^T$ , 故


$$\begin{aligned} & (D - \omega U)G_{\text{SSOR}}(D - \omega U)^{-1} \\ &= [(1 - \omega)D + \omega L](D - \omega L)^{-1} [(1 - \omega)D + \omega L^T](D - \omega L^T)^{-1} \\ &= [(2 - \omega)D(D - \omega L)^{-1} - I][(2 - \omega)D(D - \omega L^T)^{-1} - I] \\ &= I - (2 - \omega)D[(D - \omega L)^{-1} + (D - \omega L^T)^{-1}] + (2 - \omega)^2 D(D - \omega L)^{-1} D(I - \omega L^T)^{-1}. \end{aligned}$$

假定  $D$  的对角线元素全是正的, 则

$$\begin{aligned} & D^{-1/2}(D - \omega U)G_{\text{SSOR}}(D - \omega U)^{-1}D^{1/2} \\ &= I - (2 - \omega)D^{-1/2}[(D - \omega L)^{-1} + (D - \omega L^T)^{-1}]D^{1/2} \\ &\quad + (2 - \omega)^2 D^{-1/2}(D - \omega L)^{-1} D(I - \omega L^T)^{-1} D^{1/2}. \end{aligned}$$

这是一个对称矩阵, 故  $G_{\text{SSOR}}$  具有实特征值. 所以我们可以对其实行 Chebyshev 加速. 但我们需要估计  $G_{\text{SSOR}}$  的谱半径.



 若存在矩阵  $W$  使得  $W^{-1}AW$  是对称矩阵, 则称  $A$  是对称化的, 即  $A$  相似于一个对称矩阵.

## 6.6 交替方向法与 HSS 迭代法

### 6.6.1 多步迭代法

设  $A = M_1 - N_1 = M_2 - N_2$  是  $A$  的两个矩阵分裂, 则可以构造迭代格式

$$\begin{cases} M_1 x^{(k+\frac{1}{2})} = N_1 x^{(k)} + b, \\ M_2 x^{(k+1)} = N_2 x^{(k+\frac{1}{2})} + b, \end{cases} \quad k = 0, 1, 2, \dots \quad (6.31)$$

这就是两步迭代法, 对应的分裂称为二重分裂. 易知, 两步迭代格式 (6.31) 的迭代矩阵为

$$G = M_2^{-1} N_2 M_1^{-1} N_1.$$

因此, 其收敛的充要条件是  $\rho(M_2^{-1} N_2 M_1^{-1} N_1) < 1$ .

类似地, 我们可以推广到多步迭代法. 设  $l$  是一个正整数, 则  $A$  的  $l$  重分裂为

$$A = M_1 - N_1 = M_2 - N_2 = \dots = M_l - N_l,$$

相应的多步迭代法为

$$\begin{cases} M_1 x^{(k+\frac{1}{l})} = N_1 x^{(k)} + b, \\ M_2 x^{(k+\frac{2}{l})} = N_2 x^{(k+\frac{1}{l})} + b, \\ \dots \\ M_l x^{(k+1)} = N_l x^{(k+\frac{l-1}{l})} + b, \end{cases} \quad k = 0, 1, 2, \dots$$

### 6.6.2 交替方向法

**交替方向法** (alternating direction implicit, **ADI**) 是由 Peaceman 和 Rachford [99] 于 1955 年提出, 用于计算偏微分方程的数值解, 因此也称为 PR 方法. 其本质上也可以看成是一个两步迭代法.

设  $A = A_1 + A_2$ , 则 ADI 迭代格式为

$$\begin{cases} (\alpha I + A_1) x^{(k+\frac{1}{2})} = (\alpha I - A_2) x^{(k)} + b, \\ (\alpha I + A_2) x^{(k+1)} = (\alpha I - A_1) x^{(k+\frac{1}{2})} + b, \end{cases} \quad k = 0, 1, 2, \dots, \quad (6.32)$$

其中  $\alpha \in \mathbb{R}$  是迭代参数. 易知 ADI 方法的迭代矩阵为

$$G_{\text{ADI}} = (\alpha I + A_2)^{-1} (\alpha I - A_1) (\alpha I + A_1)^{-1} (\alpha I - A_2).$$

它相似于

$$\tilde{G} \triangleq (\alpha I - A_1) (\alpha I + A_1)^{-1} (\alpha I - A_2) (\alpha I + A_2)^{-1}.$$

所以 ADI 迭代 (6.32) 收敛的充要条件是

$$\rho(\tilde{G}) < 1.$$

若  $A$  对称正定, 且  $A_1$  和  $A_2$  中有一个是对称正定, 另一个是对称半正定, 则有下面的收敛定理.





**定理 6.34** 设  $A \in \mathbb{R}^{n \times n}$  对称正定,  $A = A_1 + A_2$ , 其中  $A_1$  和  $A_2$  中有一个是对称正定, 另一个是对称半正定, 则对任意正数  $\alpha > 0$ , 有  $\rho(\tilde{G}) < 1$ , 即 ADI 迭代法 (6.32) 收敛.

(板书)

**证明.** 不妨假设  $A_1$  对称正定,  $A_2$  对称半正定, 则  $(\alpha I - A_1)(\alpha I + A_1)^{-1}$  和  $(\alpha I - A_2)(\alpha I + A_2)^{-1}$  都对称, 故

$$\begin{aligned}\|(\alpha I - A_1)(\alpha I + A_1)^{-1}\|_2 &= \max_{\lambda \in \sigma(A_1)} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right| < 1, \\ \|(\alpha I - A_2)(\alpha I + A_2)^{-1}\|_2 &= \max_{\lambda \in \sigma(A_2)} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right| \leq 1.\end{aligned}$$

所以,

$$\rho(\tilde{G}) \leq \|\tilde{G}\|_2 \leq \|(\alpha I - A_1)(\alpha I + A_1)^{-1}\|_2 \cdot \|(\alpha I - A_2)(\alpha I + A_2)^{-1}\|_2 < 1.$$

□


### 6.6.3 HSS 方法

**HSS 方法** 全称为 **H**ermitian and **S**kew-Hermit **S**plitting method, 是由 Bai, Golub 和 Ng [8] 于 2003 年提出.

设  $A = H + S$ , 其中  $H$  和  $S$  分别是  $A$  的对称与反对称 (斜对称, Skew-Hermit) 部分, 即

$$H = \frac{A + A^T}{2}, \quad S = \frac{A - A^T}{2}.$$

该分裂就称为 HS 分裂, 即 HSS.

 任何一个矩阵都具有 HS 分裂. 如果  $A \in \mathbb{C}^{n \times n}$ , 则取共轭转置.

类似于 ADI 方法, 我们可得下面的 HSS 方法

$$\begin{cases} (\alpha I + H)x^{(k+\frac{1}{2})} = (\alpha I - S)x^{(k)} + b, \\ (\alpha I + S)x^{(k+1)} = (\alpha I - H)x^{(k+\frac{1}{2})} + b, \end{cases} \quad k = 0, 1, 2, \dots \quad (6.33)$$

易知, HSS 方法的迭代矩阵为

$$G_{\text{HSS}} = (\alpha I + S)^{-1}(\alpha I - H)(\alpha I + H)^{-1}(\alpha I - S).$$

同样, 它相似于

$$\tilde{G} \triangleq (\alpha I - H)(\alpha I + H)^{-1}(\alpha I - S)(\alpha I + S)^{-1}.$$

我们首先考察矩阵  $(\alpha I - S)(\alpha I + S)^{-1}$ . 事实上, 它是一个酉矩阵 (见习题 (5.5)). 因此

$$\|(\alpha I - S)(\alpha I + S)^{-1}\|_2 = 1.$$

由于  $H$  是对称矩阵, 因此

$$\|(\alpha I - H)(\alpha I + H)^{-1}\|_2 = \max_{\lambda \in \sigma(H)} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right|.$$



**定理 6.35** 设  $A \in \mathbb{R}^{n \times n}$  正定, 则对任意正数  $\alpha > 0$ , 有  $\rho(\tilde{G}) < 1$ , 即 HSS 迭代法 (6.33) 收敛.

(板书)

**证明.** 由于  $A$  正定, 即  $H$  对称正定, 故

$$\|(\alpha I - H)(\alpha I + H)^{-1}\|_2 = \max_{\lambda \in \sigma(H)} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right| < 1.$$

所以, 结论成立. □

### 参数 $\alpha$ 的选取

为了达到最快收敛效果, 我们希望迭代矩阵的谱半径越小越好. 但是在一般情况下, 谱半径很难计算或估计. 因此要极小化谱半径是非常困难的, 或者说是不可能的. 此时, 我们能做的往往是极小化它的一个上界.

由前面的分析可知

$$\rho(G_{\text{HSS}}) = \rho(\tilde{G}) \leq \max_{\lambda \in \sigma(H)} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right| \triangleq \sigma(\alpha).$$

下面考虑  $\sigma(\alpha)$  的极小值点. 设  $H$  的最大和最小特征值分别为  $\lambda_{\max}(H)$  和  $\lambda_{\min}(H)$ , 则我们有下面的结论.

**定理 6.36** [8] 设  $A \in \mathbb{R}^{n \times n}$  正定, 则极小极大问题


$$\min_{\alpha > 0} \max_{\lambda_{\min}(H) \leq \lambda \leq \lambda_{\max}(H)} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right|$$

的解为

$$\alpha_* = \sqrt{\lambda_{\max}(H)\lambda_{\min}(H)}.$$

此时

$$\sigma(\alpha_*) = \frac{\sqrt{\lambda_{\max}(H)} - \sqrt{\lambda_{\min}(H)}}{\sqrt{\lambda_{\max}(H)} + \sqrt{\lambda_{\min}(H)}} = \frac{\sqrt{\kappa(H)} - 1}{\sqrt{\kappa(H)} + 1}.$$

 HSS 自从被提出来以后, 很多学者将其进行了推广, 如 PSS, NSS, AHSS 等, 感兴趣的读者可以参考相关文献.




## 6.7 Poisson 方程快速求解方法

如果已经知道矩阵  $A$  的特征值分解  $A = X\Lambda X^{-1}$ , 则  $Ax = b$  的解可表示为

$$x = A^{-1}b = X\Lambda^{-1}X^{-1}b.$$

如果  $A$  是正规矩阵, 即  $X$  是酉矩阵, 则

$$x = A^{-1}b = X\Lambda^{-1}X^*b.$$

 一般来说, 我们不会采用这种特征值分解的方法来解线性方程组, 因为计算特征值分解通常比解线性方程组更困难. 但在某些特殊情况下, 我们可以由此得到快速方法.

考虑二维离散 Poisson 方程

$$Tu = h^2 f, \quad (6.34)$$

其中

$$T = I \otimes T_n + T_n \otimes I, \quad T_n = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{n \times n}.$$

由定理 6.4 可知

$$T = (Z \otimes Z)(I \otimes \Lambda + \Lambda \otimes I)(Z \otimes Z)^T,$$

其中  $Z = [z_1, z_2, \dots, z_n]$  是正交矩阵. 这里

$$z_k = \sqrt{\frac{2}{n+1}} \cdot \left[ \sin \frac{k\pi}{n+1}, \sin \frac{2k\pi}{n+1}, \dots, \sin \frac{nk\pi}{n+1} \right]^T, \quad k = 1, 2, \dots, n.$$

所以, 方程 (6.34) 的解为

$$u = T^{-1}h^2 f = [(Z \otimes Z)(I \otimes \Lambda + \Lambda \otimes I)^{-1}(Z \otimes Z)^T] h^2 f.$$

因此, 主要的运算是  $Z \otimes Z$  与向量的乘积, 以及  $(Z \otimes Z)^T$  与向量的乘积. 而这些乘积可以通过快速 Sine 变换来实现.

### 6.7.1 快速 Fourier 变换

**快速 Fourier 变换 (FFT)** 用来计算 **离散 Fourier 变换 (DFT)** 矩阵与向量乘积的一种快速方法.


设  $x = [x_0, x_1, \dots, x_{n-1}]^T \in \mathbb{C}^n$ , 其 DFT 定义为  $y = \text{DFT}(x) = [y_0, y_1, \dots, y_{n-1}]^T \in \mathbb{C}^n$ , 其中

$$y_k = \sum_{j=0}^{n-1} \omega_n^{kj} x_j, \quad k = 0, 1, \dots, n-1.$$

这里

$$\omega_n = e^{\frac{-2\pi i}{n}} = \cos(2\pi/n) - i \sin(2\pi/n)$$

是 1 的一个  $n$  次本原根 (primitive  $n$ -th root of unity),  $i$  是虚部单位.

 这里说  $\omega_n$  是 primitive  $n$ -th root of unity 是指  $\omega_n^n = 1$  且

$$\omega_n^k \neq 1, \quad k = 1, 2, \dots, n-1.$$

构造矩阵  $F_n = [f_{kj}] \in \mathbb{C}^{n \times n}$ , 其中  $f_{kj} = \omega_n^{kj} = e^{\frac{-2kj\pi i}{n}} = \cos(2kj\pi/n) - i \sin(2kj\pi/n)$ , 即

$$F_n = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \omega_n^2 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \cdots & \omega_n^{2(n-1)} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \cdots & \omega_n^{(n-1)^2} \end{bmatrix},$$

则有

$$y = \text{DFT}(x) = F_n x.$$

我们称矩阵  $F_n$  为 DFT 矩阵. 易知 DFT 矩阵具有以下性质:

(1)  $F_n$  是对称矩阵 (但不是 Hermite 对称);

(2)  $F_n^* F_n = nI$ , 所以  $\frac{1}{\sqrt{n}} F_n$  是酉矩阵.

相应的离散 Fourier 反变换定义为  $x = \text{IDFT}(y)$ , 其中

$$x_j = \frac{1}{n} \sum_{k=0}^{n-1} \omega_n^{-jk} y_k, \quad j = 0, 1, \dots, n-1.$$


写成矩阵形式为

$$x = \frac{1}{n} F_n^* y.$$

DFT 和 IDFT 满足下面的性质:

$$\text{IDFT}(\text{DFT}(x)) = x,$$

$$\text{DFT}(\text{IDFT}(y)) = y.$$

 在 MATLAB 中, 计算 DFT 和 IDFT 的函数分别为 `fft` 和 `ifft`, 即:  $y = \text{fft}(x)$ ,  $x = \text{ifft}(y)$ .  
(测试代码见 `FFT_test.m`)

## 6.7.2 离散 Sine 变换

**离散 Sine 变换 (DST)** 有多种定义, 我们这里只介绍与求解 Poisson 方程有关的一种定义, 其它定义可参见 [17].

设  $x = [x_1, x_2, \dots, x_n]^T \in \mathbb{R}^n$ , 其离散 Sine 变换定义为  $y = \text{DST}(x) = [y_1, y_2, \dots, y_n]^T \in \mathbb{R}^n$ , 其中

$$y_k = \sum_{j=1}^n x_j \sin\left(\frac{kj\pi}{n+1}\right), \quad k = 1, 2, \dots, n.$$

对应的离散 Sine 反变换记为 IDST, 即  $x = \text{IDST}(y)$ , 其中

$$x_j = \frac{2}{n+1} \sum_{k=1}^n y_k \sin\left(\frac{jk\pi}{n+1}\right), \quad j = 1, 2, \dots, n.$$

DST 和 IDST 满足下面的性质:

$$\text{IDST}(\text{DST}(x)) = x,$$



$$\text{DST}(\text{IDST}(y)) = y.$$

在 MATLAB 中, 计算 DST 和 IDST 的函数分别为 `dst` 和 `idst`, 即: `y=dst(x)`, `x=idst(y)`. (测试代码见 `DST_test.m`)

可以通过公式  $T_n = Z\Lambda Z^T$  来计算一维离散 Poisson 矩阵的特征值: 记  $[\alpha_1, \alpha_2, \dots, \alpha_n]^T$  和  $[\beta_1, \beta_2, \dots, \beta_n]^T$  分别为  $Z^T T_n$  和  $Z^T$  的第一列, 则  $T_n$  的特征值为

$$\lambda_i = \frac{\alpha_i}{\beta_i}.$$

对应的 MATLAB 代码为: `Lam=idst([2,-1,zeros(1,n-2)]') ./ idst(eye(n,1))`

### 6.7.3 Poisson 方程与 DST

我们首先考虑矩阵  $Z$  与一个任意给定向量  $b$  的乘积. 设  $y = Zb$ , 则

$$y_k = \sum_{j=1}^n Z(k, j)b_j = \sqrt{\frac{2}{n+1}} \sum_{j=1}^n b_j \sin\left(\frac{kj\pi}{n+1}\right) = \sqrt{\frac{2}{n+1}} \cdot \text{DST}(b).$$

因此, 乘积  $y = Zb$  可以通过 DST 来实现. 类似地, 乘积  $y = Z^T b = Z^{-1}b$  可以通过离散 Sine 反变换 IDST 实现, 即

$$y = Z^T b = Z^{-1}b = \left(\sqrt{\frac{2}{n+1}}\right)^{-1} \text{IDST}(b).$$

所以对于一维离散 Poisson 方程, 其解为

$$u = T_n^{-1}(h^2 f) = (Z\Lambda^{-1}Z^T)(h^2 f) = h^2 Z\Lambda^{-1}Z^T f = h^2 \cdot \text{DST}(\Lambda^{-1} \text{IDST}(b)).$$

而对于二维离散 Poisson 方程, 我们需要计算  $(Z \otimes Z)b$  和  $(Z^T \otimes Z^T)b$ . 它们对应的是二维离散 Sine 变换和二维离散 Sine 反变换.

设  $b = [b_1^T, b_2^T, \dots, b_n^T]^T \in \mathbb{R}^{n^2}$ , 其中  $b_k \in \mathbb{R}^{n \times n}$ . 令  $B = [b_1, b_2, \dots, b_n] \in \mathbb{R}^{n \times n}$ , 则由 Kronecker 乘积的性质可知

$$(Z \otimes Z)b = (Z \otimes Z)\text{vec}(B) = \text{vec}(ZBZ^T) = \text{vec}((Z(ZB)^T)^T).$$

因此, 我们仍然可以使用 DST 来计算  $(Z \otimes Z)b$ . 类似地, 我们可以使用 IDST 来计算  $(Z^T \otimes Z^T)b$ .

#### 算法 6.9. 二维离散 Poisson 方程的快速方法

- 1: 计算  $b = h^2 f$
- 2:  $B = \text{reshape}(b, n, n)$
- 3:  $B_1 = (Z^T B)^T = (\text{IDST}(B))^T$
- 4:  $B_2 = (Z^T B_1)^T = (\text{IDST}(B_1))^T$
- 5:  $b_1 = (I \otimes \Lambda + \Lambda \otimes I)^{-1} \text{vec}(B_2)$
- 6:  $B_3 = \text{reshape}(b_1, n, n)$
- 7:  $B_4 = (ZB_3)^T = (\text{DST}(B_3))^T$



$$8: B_5 = (ZB_4)^T = (\text{DST}(B_4))^T$$

$$9: u = \text{reshape}(B_5, n^2, 1)$$

MATLAB 程序见 [Poisson\\_DST.m](#)



## 6.8 课后习题

练习 6.1 已知  $A = \begin{bmatrix} a & 1 & 3 \\ 1 & a & 2 \\ -3 & 2 & a \end{bmatrix} \in \mathbb{R}^{3 \times 3}$ , 且 Jacobi 方法收敛, 求  $a$  的取值范围.

练习 6.2 写出求解线性方程组  $Ax = b$  的 Jacobi, GS 和 SOR 的分量形式, 并讨论它们的收敛性, 其

$$\text{中 } A = \begin{bmatrix} 5 & -3 & 2 \\ -3 & 6 & -1 \\ 2 & -1 & 3 \end{bmatrix}, b = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}.$$

练习 6.3 设矩阵  $A \in \mathbb{C}^{n \times n}$  是块三对角矩阵, 且主对角块为 0, 即

$$A = \begin{bmatrix} 0 & B_1 & & & \\ A_1 & 0 & B_2 & & \\ & A_2 & \ddots & \ddots & \\ & & \ddots & \ddots & B_{N-1} \\ & & & A_{N-1} & 0 \end{bmatrix}.$$

试证明: 矩阵  $A(\alpha) \triangleq \alpha L + \frac{1}{\alpha} U$  的特征值与  $\alpha$  无关 ( $\alpha \in \mathbb{C}$  且  $\alpha \neq 0$ ), 其中  $L$  和  $U$  分别是  $A$  的下三角和上三角部分.

练习 6.4 若存在非奇异矩阵  $W$ , 使得  $W(I - G)W^{-1}$  是对称正定矩阵, 则称迭代法

$$x^{(k+1)} = Gx^{(k)} + g, \quad k = 0, 1, 2, \dots$$

是**可对称化**的. 现假定上述迭代法是**可对称化**的, 试证

- (1)  $G$  的特征值全部为实数, 且都小于 1;
- (2) 存在  $\gamma$ , 使得下面的**外推迭代法**收敛

$$x^{(k+1)} = (1 - \gamma)x^{(k)} + \gamma(Gx^{(k)} + g), \quad k = 0, 1, 2, \dots$$

练习 6.5 设  $A = \begin{bmatrix} 2 & 0 \\ 3 & 1 \end{bmatrix}$ , 构造迭代法

$$x^{(k+1)} = x^{(k)} + \alpha(b - Ax^{(k)}), \quad k = 0, 1, 2, \dots$$

问: 当  $\alpha \in \mathbb{R}$  取何值时, 该迭代法收敛, 什么时候收敛最快?

练习 6.6 设  $A \in \mathbb{R}^{n \times n}$ . 证明:  $\rho(A) < 1$  的充要条件是  $I - A$  非奇异, 且  $(I - A)^{-1}(I + A)$  的特征值具有正实部.



练习 6.7 设  $A \in \mathbb{R}^{n \times n}$  是对称三对角矩阵

$$A = \begin{bmatrix} a & b & & \\ b & \ddots & \ddots & \\ & \ddots & \ddots & b \\ & & b & a \end{bmatrix}.$$

计算矩阵  $A$  的特征值和特征向量.

练习 6.8 设  $A \in \mathbb{R}^{n \times n}$  对称正定,  $B \in \mathbb{R}^{n \times k}$  列满秩, 试证明  $B^T A B$  对称正定.

练习 6.9\* 设  $A = D - E - E^* \in \mathbb{C}^{n \times n}$ , 其中  $D$  是 Hermite 正定的. 令

$$G_\omega = (D - \omega E)^{-1}[(1 - \omega)D + \omega E^*],$$

其中  $\omega$  是实数. 证明  $\rho(G_\omega) < 1$  当且仅当

(1)  $A$  是正定的, 且  $0 < \omega < 2$ , 或者

(2)  $A$  是负定的, 且  $\omega \notin [0, 2]$ .

(提示: 利用和仿照引理 6.17 中的证明思路)

练习 6.10 设迭代格式  $x^{(k+1)} = Gx^{(k)} + g$ , 其中  $G$  对称且  $\lambda(G) \in [\alpha, \beta]$ ,  $-1 < \alpha \leq \beta < 1$ . 试构造出相应的 Chebyshev 加速方法.

练习 6.11 (定理 6.36) 设  $\lambda_{\max} \geq \lambda_{\min} > 0$ , 则最小最大问题

$$\min_{\alpha > 0} \max_{\lambda_{\min} \leq \lambda \leq \lambda_{\max}} \left| \frac{\alpha - \lambda}{\alpha + \lambda} \right|$$

的解为

$$\alpha_* = \sqrt{\lambda_{\min} \lambda_{\max}}.$$

练习 6.12\* 设  $A \in \mathbb{R}^{n \times n}$  是对称正定的. 证明: 对任意正实数  $\alpha$ , 都有

$$\rho((\alpha I + A)^{-1}(\alpha I - A)) < 1.$$

如果  $A$  只是正定呢?

练习 6.13\* (定理 6.13) 设  $A \in \mathbb{R}^{n \times n}$ , 若  $A$  是不可约对角占优, 证明: Jacobi 迭代法和 G-S 迭代法都收敛.

练习 6.14\* 证明: 对任意矩阵  $A \in \mathbb{C}^{n \times n}$  和任意正数  $\varepsilon > 0$ , 存在非奇异矩阵  $L$ , 使得

$$\|A\|_L \leq \rho(A) + \varepsilon,$$

其中  $\|A\|_L \triangleq \|LAL^{-1}\|_2$ .

练习 6.15\* 设  $A \in \mathbb{R}^{n \times n}$ . 证明:  $\rho(A) < 1$  的充要条件是存在对称正定矩阵  $P \in \mathbb{R}^{n \times n}$  使得  $P - APA^T$  正定.

..... 以下为可选题 .....

练习 6.16 试给出  $T_n$  (见 (6.9)) 的 Cholesky 分解.





练习 6.17 证明推论 6.3 (三对角 Toeplitz 矩阵的特征值分解).

练习 6.18\* 设  $A \in \mathbb{R}^{n \times n}$ , 如果矩阵分裂  $A = M - N$  满足  $M^{-1} \geq 0, N \geq 0$ , 则称这个分裂为**正则分裂**. 假定  $A^{-1} \geq 0$ , 且  $A = M - N$  是正则分裂, 试证明

$$\rho(M^{-1}N) = \frac{\rho(A^{-1}N)}{1 + \rho(A^{-1}N)} < 1.$$

练习 6.19 (定理 6.19, 对称正定矩阵与定常迭代的收敛性)

设  $A \in \mathbb{R}^{n \times n}$  对称正定. 证明:

- (1) 若  $2D - A$  正定, 则 Jacobi 迭代收敛;
- (2) 若  $0 < \omega < 2$ , 则 SOR 和 SSOR 收敛;
- (3) G-S 迭代收敛.

练习 6.20 (定理 6.20, 定常迭代的收敛性与矩阵的对称正定性)

设  $A \in \mathbb{R}^{n \times n}$  对称. 证明:

- (1) 若  $2D - A$  正定且 Jacobi 迭代收敛, 则  $A$  正定;
- (2) 若  $D$  正定, 且存在  $\omega \in (0, 2)$  使得 SOR (或 SSOR) 收敛, 则  $A$  正定;
- (3) 若  $D$  正定, 且 G-S 迭代收敛, 则  $A$  正定.

练习 6.21\* 设  $A \in \mathbb{R}^{n \times n}$  严格对角占优且非负, 则结论  $\rho(G_{GS}) \leq \rho(G_J)$  是否成立? 若成立则给出证明, 若不成立则给出反例.

练习 6.22 初值的选取对迭代法的收敛速度也有着很大的影响, 请以 Jacobi 迭代法为例, 初值取何时收敛最快 (真解除外)? 什么初值收敛最慢?

练习 6.23\* 设  $A \in \mathbb{R}^{n \times n}$  对称正定,  $B \in \mathbb{R}^{n \times m}$  满秩, 其中  $n \geq m$ . 记  $A$  的最大和最小特征值分别为  $\mu_1$  和  $\mu_n$ , 记  $B$  的最大和最小奇异值分别为  $\sigma_1$  和  $\sigma_m$ . 证明:  $\mathcal{A} = \begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix}$  的特征值满足

$$\lambda(\mathcal{A}) \in I^- \cup I^+,$$

其中

$$I^- = \left[ \frac{1}{2} \left( \mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2} \right), \frac{1}{2} \left( \mu_1 - \sqrt{\mu_1^2 + 4\sigma_m^2} \right) \right],$$

$$I^+ = \left[ \mu_n, \frac{1}{2} \left( \mu_1 + \sqrt{\mu_1^2 + 4\sigma_1^2} \right) \right].$$

练习 6.24\* 设  $A \in \mathbb{R}^{n \times n}$  对称正定,  $B \in \mathbb{R}^{n \times m}$  满秩,  $C \in \mathbb{R}^{m \times m}$  对称半正定, 其中  $n \geq m$ , 记  $A$  的最大和最小特征值分别为  $\mu_1$  和  $\mu_n$ ,  $C$  的最大特征值为  $\nu_1$ ,  $B$  的最大和最小奇异值分别为  $\sigma_1$  和  $\sigma_m$ . 证明:  $\mathcal{A} = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$  的特征值满足

$$\lambda(\mathcal{A}) \in I^- \cup I^+,$$

其中

$$I^- = \left[ \frac{1}{2} \left( \mu_n - \nu_1 - \sqrt{(\mu_n + \nu_1)^2 + 4\sigma_1^2} \right), \frac{1}{2} \left( \mu_1 - \sqrt{\mu_1^2 + 4\sigma_m^2} \right) \right],$$



$$I^+ = \left[ \mu_n, \frac{1}{2} \left( \mu_1 + \sqrt{\mu_1^2 + 4\sigma_1^2} \right) \right].$$

**练习 6.25\*** 设  $A \in \mathbb{R}^{n \times n}$  对称正定,  $B \in \mathbb{R}^{n \times m}$  满秩, 其中  $n \geq m$ . 记  $A$  的最小特征值为  $\mu_n$ . 证明:

如果  $\mu_n \geq 4\|S\|_2$ , 其中  $S = B^T A^{-1} B$ , 则  $\tilde{A} = \begin{bmatrix} A & B \\ -B^T & 0 \end{bmatrix}$  的特征值都是正实数.

**练习 6.26\*** 试举例说明, 存在线性方程组  $Ax = b$ , 使得 Jacobi 迭代法收敛, 但 G-S 方法不收敛.

..... **以下为实践题** .....

**练习 6.27** 写出 Poisson 方程红黑排序的 SOR 和 SSOR 方法的迭代格式, 并编程实现. (以例 6.2 中的 Poisson 方程为例)



## 第七讲 Krylov 子空间迭代法

### 子空间迭代法的基本思想

在一个维数较低的子空间中寻找解析解的一个“最佳”近似. 子空间迭代法的主要过程可以分解为下面三步:

- (1) 寻找合适的子空间;
- (2) 在该子空间中求“最佳近似解”;
- (3) 若这个近似解满足精度要求, 则停止计算; 否则, 重新构造一个新的子空间, 并返回第(2)步.

### 主要涉及到的两个关键问题

- (1) 如果选择和更新子空间;
- (2) 如何在给定的子空间中寻找“最佳近似解”.

关于第一个问题, 目前较成功的解决方案就是使用 **Krylov 子空间**.

### 关于 Krylov 子空间迭代法的相关资料

- [1] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd edition, 2003 [109]
- [2] J. Demmel, *Applied Numerical Linear Algebra*, 1997 [30]
- [3] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, 1997 [57]
- [4] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th, 2013 [56]
- [5] L.N. Trefethen and D. Bau, *Numerical Linear Algebra*, 1997 [119]

## 7.1 Krylov 子空间

### 7.1.1 Arnoldi 过程与 Lanczos 过程

设  $A \in \mathbb{R}^{n \times n}$ ,  $r \in \mathbb{R}^n$ , 则由  $A$  和  $r$  生成的 Krylov 子空间定义为

$$\mathcal{K}_m(A, r) = \text{span}\{r, Ar, A^2r, \dots, A^{m-1}r\}, \quad m \leq n,$$

通常简记为  $\mathcal{K}_m$ .

设解析解在  $\mathcal{K}_m$  中的“最佳近似解”为  $x^{(m)}$ . 我们假定  $r, Ar, A^2r, \dots, A^{m-1}r$  线性无关, 则  $\dim \mathcal{K}_m = m$ . 令  $v_1, v_2, \dots, v_m$  是  $\mathcal{K}_m$  的一组基, 则  $\mathcal{K}_m$  中的任意向量  $x$  均可表示为

$$x = y_1 v_1 + y_2 v_2 + \dots + y_m v_m = V_m y,$$

其中  $y = [y_1, y_2, \dots, y_m]^T$  为线性表出系数,  $V_m = [v_1, v_2, \dots, v_m]$ . 于是, 寻找“最佳近似解”  $x^{(m)}$  就转化为下面两个子问题:

- (1) 寻找一组合适的基  $v_1, v_2, \dots, v_m$ ;
- (2) 求出  $x^{(m)}$  在这组基下面的线性表出系数  $y^{(m)} = [y_1^{(m)}, y_2^{(m)}, \dots, y_m^{(m)}]^T$ .

### Arnoldi 过程

首先考虑基的选取. 假定  $r, Ar, A^2r, \dots, A^{m-1}r$  线性无关, 因此它们就自然地组成  $\mathcal{K}_m$  的一组基. 但为了确保算法的稳定性, 一般来说, 我们通常希望选取一组标准正交基. 这并不困难, 只需对向量组  $\{r, Ar, A^2r, \dots, A^{m-1}r\}$  进行单位正交化即可. 对这个过程略加修改, 就得到下面的 **Arnoldi 过程**.

#### 算法 7.1. Arnoldi 过程 (MGS)

```

1: $v_1 = r / \|r\|_2$
2: for $j = 1$ to $m - 1$ do
3: $z = Av_j$
4: for $i = 1$ to j do % MGS
5: $h_{i,j} = (v_i, z)$
6: $z = z - h_{i,j}v_i$
7: end for
8: $h_{j+1,j} = \|z\|_2$
9: if $h_{j+1,j} = 0$ then
10: break
11: end if
12: $v_{j+1} = z / h_{j+1,j}$
13: end for
```

可以证明, Arnoldi 过程生成的向量  $v_1, v_2, \dots, v_m$  构成  $\mathcal{K}_m$  的一组标准正交基.



**引理 7.1** 如果 Arnoldi 过程不中断, 则

$$\mathcal{K}_m = \text{span}\{v_1, v_2, \dots, v_m\}.$$

(留作练习, 归纳法)

记  $V_m = [v_1, v_2, \dots, v_m]$ ,

$$H_{m+1,m} = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & \cdots & h_{1,m} \\ h_{2,1} & h_{2,2} & h_{2,3} & \cdots & h_{2,m} \\ & h_{3,2} & h_{3,3} & \cdots & h_{3,m} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & h_{m,m} \\ & & & & h_{m+1,m} \end{bmatrix} \in \mathbb{R}^{(m+1) \times m},$$

则由 Arnoldi 过程可知

$$h_{j+1,j}v_{j+1} = Av_j - h_{1,j}v_1 - h_{2,j}v_2 - \cdots - h_{j,j}v_j,$$

即

$$Av_j = \sum_{i=1}^{j+1} h_{i,j}v_i = V_{m+1} \begin{bmatrix} h_{1,j} \\ \vdots \\ h_{j+1,j} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = V_{m+1} H_{m+1,m}(:, j).$$

所以有

$$AV_m = V_{m+1} H_{m+1,m} = V_m H_m + h_{m+1,m} v_{m+1} e_m^T, \quad (7.1)$$

其中  $H_m$  是由  $H_{m+1,m}$  的前  $m$  行组成的矩阵, 即  $H_m = H_{m+1,m}(1:m, 1:m)$ ,  $e_m = [0, \dots, 0, 1]^T \in \mathbb{R}^m$ . 由于  $V_m$  是列正交矩阵, 上式两边同乘  $V_m^T$  可得

$$V_m^T AV_m = H_m. \quad (7.2)$$

等式 (7.1) 和 (7.2) 是 Arnoldi 过程的两个重要性质. 这两个性质也可以通过下图来表示.

需要指出的是, 如果  $r, Ar, A^2r, \dots, A^{m-1}r$  线性相关, 则 Arnoldi 过程就会提前中断. 此时, 我们会得到一个不变子空间.

**定理 7.2** 如果 Arnoldi 过程在第  $k$  步时中断, 即  $h_{k+1,k} = 0$ , 其中  $k < m$ . 则有  $AV_k = V_k H_k$ , 即  $\mathcal{K}_k$  是  $A$  的一个不变子空间.



(留作课外自习, 直接利用等式 (7.1))

**Lanczos 过程**

如果  $A$  是对称矩阵, 则  $H_m$  为对称三对角矩阵, 此时将其记为  $T_m$ , 即

$$T_m = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{m-1} \\ & & \beta_{m-1} & \alpha_m \end{bmatrix}. \quad (7.3)$$

与 Arnoldi 过程类似, 我们有下面的性质

$$AV_m = V_m T_m + \beta_m v_{m+1} e_m^T, \quad (7.4)$$

$$V_m^T A V_m = T_m. \quad (7.5)$$

考察关系式 (7.4) 两边的第  $j$  列可知

$$\beta_j v_{j+1} = A v_j - \alpha_j v_j - \beta_{j-1} v_{j-1}, \quad j = 1, 2, \dots,$$

这里我们令  $v_0 = 0$  和  $\beta_0 = 0$ . 根据这个三项递推公式, Arnoldi 过程可简化为下面的 **Lanczos 过程**.

**算法 7.2. Lanczos 过程**

```

1: Set $v_0 = 0$ and $\beta_0 = 0$
2: $v_1 = r / \|r\|_2$
3: for $j = 1$ to $m - 1$ do
4: $z = A v_j$
5: $\alpha_j = (v_j, z)$
6: $z = z - \alpha_j v_j - \beta_{j-1} v_{j-1}$
7: $\beta_j = \|z\|_2$
8: if $\beta_j = 0$ then
9: break
10: end if
11: $v_{j+1} = z / \beta_j$
12: end for

```

可以证明, 由 Lanczos 过程得到的向量组  $\{v_1, v_2, \dots, v_m\}$  是单位正交的.

**定理 7.3** 设  $\{v_1, v_2, \dots, v_m\}$  是由 Lanczos 过程得到的向量组, 则

$$(v_i, v_j) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \quad i, j = 1, 2, \dots, m.$$

(留作练习)



## 7.1.2 Krylov 子空间方法一般格式

## Krylov 子空间迭代法的一般过程

- (1) 令  $m = 1$
- (2) 定义 Krylov 子空间  $\mathcal{K}_m$ ;
- (3) 找出仿射空间  $x^{(0)} + \mathcal{K}_m$  中的“最佳近似”解;
- (4) 如果这个近似解满足精度要求, 则迭代结束; 否则令  $m \leftarrow m + 1$ , 即将 Krylov 子空间的维数增加一维, 并返回第 (3) 步.

在实际计算中, 为了充分利用迭代初值中所包含的有用信息, 我们通常是在仿射空间  $x^{(0)} + \mathcal{K}_m$  中寻找“最佳近似解”.

## 算法 7.3. Krylov 子空间迭代算法

- 1: 选取初始向量  $x^{(0)}$
- 2: 计算  $r_0 = b - Ax^{(0)}$ ,  $v_1 = r_0 / \|r_0\|_2$
- 3: 寻找“最佳近似解”:  $x^{(1)} \in x^{(0)} + \mathcal{K}_1 = x^{(0)} + \text{span}\{v_1\}$
- 4: **if**  $x^{(1)}$  满足精度要求 **then**
- 5:     终止迭代
- 6: **end if**
- 7: **for**  $m = 2$  to  $n$  **do**
- 8:     调用 Arnoldi 或 Lanczos 过程计算向量  $v_m$
- 9:     寻找“最佳近似解”:  $x^{(m)} \in x^{(0)} + \mathcal{K}_m = x^{(0)} + \text{span}\{v_1, \dots, v_m\}$
- 10:     **if**  $x^{(m)}$  满足精度要求 **then**
- 11:         终止迭代
- 12:     **end if**
- 13: **end for**

子空间的选取和更新问题可以通过 Krylov 子空间来解决. 下面需要考虑如何寻找方程组在仿射空间  $x^{(0)} + \mathcal{K}_m$  中的“最佳近似”解  $x^{(m)}$ . 首先, 我们必须给出“最佳”的定义, 即  $x^{(m)}$  满足什么条件时才是“最佳”的, 不同的定义会导出不同的算法.

我们很自然地想到用近似解与真解之间的距离来衡量“最佳”, 即使得  $x^{(m)} - x_*$  在某种意义上最小, 如  $\|x^{(m)} - x_*\|_2$  达到最小. 但是由于  $x_*$  不知道, 因此这种方式往往是不实用.

实用的“最佳”判别方式有:

- (1) 使得  $\|r_m\|_2 = \|b - Ax^{(m)}\|_2$  最小, 即极小化残量  $r_m = b - Ax^{(m)}$ . 这种方式是可行的, 当  $A$  对称时, 相应的算法即为 **MINRES 方法**, 当  $A$  不对称时, 相应的算法即为 **GMRES 方法**;
- (2) 若  $A$  对称正定, 我们也可以极小化残量的能量范数  $\|r_m\|_{A^{-1}}$ , 相应的算法即为 **CG 方法**. 事

实上, 我们有

$$\begin{aligned}
 \|r_m\|_{A^{-1}} &\triangleq (r_m^T A^{-1} r_m)^{\frac{1}{2}} = \left( (b - Ax^{(m)})^T A^{-1} (b - Ax^{(m)}) \right)^{\frac{1}{2}} \\
 &= \left( (A^{-1}b - x^{(m)})^T A (A^{-1}b - x^{(m)}) \right)^{\frac{1}{2}} \\
 &= \left( (x_* - x^{(m)})^T A (x_* - x^{(m)}) \right)^{\frac{1}{2}} \\
 &= \|x_* - x^{(m)}\|_A.
 \end{aligned}$$

因此 CG 方法也可以看作是极小化能量范数  $\|x_* - x^{(m)}\|_A$  的算法.

### 7.1.3 常用的 Krylov 子空间迭代算法

Krylov 子空间算法一览表

|      |                 |                              |
|------|-----------------|------------------------------|
| 对称   | CG (1952)       | 对称正定, 正交投影法 (Galerkin)       |
|      | MINRES (1975)   | 对称不定, 斜投影法 (Petrov-Galerkin) |
|      | SYMMLQ (1975)   | 对称不定, 正交投影法 (Galerkin)       |
|      | SQMR (1994)     | 对称不定                         |
| 非对称  | FOM (1981)      | 正交投影法 (Galerkin)             |
|      | GMRES (1984)    | 斜投影法 (Petrov-Galerkin)       |
|      | BiCG (1976)     | 双正交 (biorthogonalization)    |
|      | QMR (1991)      | 双正交 (biorthogonalization)    |
|      | CGS (1989)      | 免转置                          |
|      | BiCGStab (1992) | 免转置                          |
|      | TFQMR (1993)    | 免转置                          |
| 正规方程 | FGMRES (1993)   | Flexible GMARES              |
|      | CGLS (1982)     | 法方程 (最小二乘)                   |
|      | LSQR (1982)     | 鞍点问题                         |

我们主要介绍 GMRES 和 CG 方法.





## 7.2 GMRES 方法

### 7.2.1 算法描述

GMRES 方法是目前求解非对称线性方程组的最常用算法之一. 在该算法中, “最佳近似解”的判别方法为“使得  $\|r_m\|_2 = \|b - Ax^{(m)}\|_2$  最小”, 即

$$x^{(m)} = \arg \min_{x \in x^{(0)} + \mathcal{K}_m} \|b - Ax\|_2. \quad (7.6)$$

下面我们就根据这个最优性条件来推导出 GMRES 方法.

设迭代初始向量为  $x^{(0)}$ , 则对任意向量  $x \in x^{(0)} + \mathcal{K}_m$ , 可设  $x = x^{(0)} + V_m y$ , 其中  $y \in \mathbb{R}^m$ . 于是有

$$\begin{aligned} r &= b - Ax \\ &= b - A(x^{(0)} + V_m y) \\ &= r_0 - AV_m y \\ &= \beta v_1 - V_{m+1} H_{m+1,m} y \\ &= V_{m+1} (\beta e_1 - H_{m+1,m} y), \end{aligned}$$

这里  $\beta = \|r_0\|_2$ . 由于  $V_{m+1}$  列正交, 所以

$$\|r\|_2 = \|V_{m+1}(\beta e_1 - H_{m+1,m} y)\|_2 = \|\beta e_1 - H_{m+1,m} y\|_2.$$

于是最优性条件 (7.6) 就等价于

$$y^{(m)} = \arg \min_{y \in \mathbb{R}^m} \|\beta e_1 - H_{m+1,m} y\|_2. \quad (7.7)$$

这是一个最小二乘问题. 由于  $H_{m+1,m}$  是一个上 Hessenberg 矩阵, 且通常  $m$  不是很大, 所以我们可以用基于 Givens 变换的 QR 分解来求解. 下面就是 GMRES 方法的基本框架.

#### 算法 7.4. GMRES 方法基本框架

- 1: 选取初值  $x^{(0)}$ , 停机标准  $\varepsilon > 0$ , 以及最大迭代步数 IterMax
- 2:  $r_0 = b - Ax^{(0)}$ ,  $\beta = \|r_0\|_2$
- 3:  $v_1 = r_0/\beta$
- 4: **for**  $j = 1$  to IterMax **do**
- 5:      $w = Av_j$
- 6:     **for**  $i = 1$  to  $j$  **do**     % Arnoldi 过程
- 7:          $h_{i,j} = (v_i, w)$
- 8:          $w = w - h_{i,j}v_i$
- 9:     **end for**
- 10:      $h_{j+1,j} = \|w\|_2$
- 11:     **if**  $h_{j+1,j} = 0$  **then**
- 12:          $m = j$ , break
- 13:     **end if**



```

14: $v_{j+1} = w/h_{j+1,j}$
15: $\text{relres} = \|r_j\|_2/\beta$ % 相对残量
16: if $\text{relres} < \varepsilon$ then % 检测是否收敛
17: $m = j$, break
18: end if
19: end for
20: 解最小二乘问题 (7.7), 得到 $y^{(m)}$
21: $x^{(m)} = x^{(0)} + V_m y^{(m)}$

```

### 7.2.2 具体实施细节

需要解决的问题有:

- (1) 如何计算残量  $r_j \triangleq b - Ax^{(j)}$  的范数?
- (2) 如何求解最小二乘问题 (7.7)?

这两个问题可以同时处理. 首先采用 QR 分解来求解最小二乘问题. 设  $H_{m+1,m}$  的 QR 分解为

$$H_{m+1,m} = Q_{m+1}^T R_{m+1,m},$$

其中  $Q_{m+1} \in \mathbb{R}^{(m+1) \times (m+1)}$  是正交矩阵,  $R_{m+1,m} \in \mathbb{R}^{(m+1) \times m}$  是上三角矩阵. 则

$$\|\beta e_1 - H_{m+1,m} y\|_2 = \|\beta Q_{m+1} e_1 - R_{m+1,m} y\|_2 = \left\| \beta q_1 - \begin{bmatrix} R_m \\ 0 \end{bmatrix} y \right\|_2, \quad (7.8)$$

其中  $R_m \in \mathbb{R}^{m \times m}$  是非奇异上三角矩阵 (这里假定  $H_{m+1,m}$  不可约). 所以问题 (7.7) 的解为

$$y^{(m)} = \beta R_m^{-1} q_1(1:m),$$

且

$$\|r_m\|_2 = \|b - Ax^{(m)}\|_2 = \|\beta e_1 - H_{m+1,m} y^{(m)}\|_2 = \beta \cdot |q_1(m+1)|,$$

其中  $q_1(m+1)$  表示  $q_1$  的第  $m+1$  个分量.

### $H_{m+1,m}$ 的 QR 分解的递推计算方法

由于  $H_{m+1,m}$  是上 Hessenberg 矩阵, 因此我们采用 Givens 变换.

- (1) 当  $m = 1$  时,  $H_{21} = \begin{bmatrix} h_{11} \\ h_{21} \end{bmatrix}$ , 构造 Givens 变换  $G_1$  使得

$$G_1 H_{21} = \begin{bmatrix} * \\ 0 \end{bmatrix} = R_{21}, \quad \text{即} \quad H_{21} = G_1^T R_{21}.$$

- (2) 假定存在  $G_1, G_2, \dots, G_{m-1}$ , 使得

$$(G_{m-1} \cdots G_2 G_1) H_{m,m-1} = R_{m,m-1},$$

即

$$H_{m,m-1} = (G_{m-1} \cdots G_2 G_1)^T R_{m,m-1} \triangleq Q_m^T R_{m,m-1}.$$



 为了书写方便, 这里假定  $G_i$  的维数自动扩张, 以满足矩阵乘积的需要.

(3) 考虑  $H_{m+1,m}$  的 QR 分解. 易知

$$H_{m+1,m} = \begin{bmatrix} H_{m,m-1} & h_m \\ 0 & h_{m+1,m} \end{bmatrix}, \quad \text{其中 } h_m = [h_{1m}, h_{2m}, \dots, h_{mm}]^\top.$$

所以有

$$\begin{bmatrix} Q_m & 0 \\ 0 & 1 \end{bmatrix} H_{m+1,m} = \begin{bmatrix} R_{m,m-1} & Q_m h_m \\ 0 & h_{m+1,m} \end{bmatrix} = \begin{bmatrix} R_{m-1} & \tilde{h}_{m-1} \\ 0 & \hat{h}_{mm} \\ 0 & h_{m+1,m} \end{bmatrix},$$

其中  $\tilde{h}_{m-1}$  是  $Q_m h_m$  的前  $m-1$  个元素组成的向量,  $\hat{h}_{mm}$  是  $Q_m h_m$  的最后一个元素. 构造 Givens 变换  $G_m$ :

$$G_m = \begin{bmatrix} I_{m-1} & 0 & 0 \\ 0 & c_m & s_m \\ 0 & -s_m & c_m \end{bmatrix} \in \mathbb{R}^{(m+1) \times (m+1)},$$

其中

$$c_m = \frac{\hat{h}_{m,m}}{\tilde{h}_{m,m}}, \quad s_m = \frac{h_{m+1,m}}{\tilde{h}_{m,m}}, \quad \tilde{h}_{m,m} = \sqrt{\hat{h}_{m,m}^2 + h_{m+1,m}^2}.$$


令

$$Q_{m+1} = G_m \begin{bmatrix} Q_m & 0 \\ 0 & 1 \end{bmatrix},$$

则

$$Q_{m+1} H_{m+1,m} = G_m \begin{bmatrix} R_{m-1} & \tilde{h}_{m-1} \\ 0 & \hat{h}_{m,m} \\ 0 & h_{m+1,m} \end{bmatrix} = \begin{bmatrix} R_{m-1} & \tilde{h}_{m-1} \\ 0 & \tilde{h}_{m,m} \\ 0 & 0 \end{bmatrix} \triangleq R_{m+1,m}.$$

所以可得  $H_{m+1,m}$  的 QR 分解  $H_{m+1,m} = Q_{m+1}^\top R_{m+1,m}$ .

 由  $H_{m,m-1}$  的 QR 分解到  $H_{m+1,m}$  的 QR 分解, 我们需要

(1) 计算  $Q_m h_m$ , 即将之前的  $m-1$  个 Givens 变换作用到  $H_{m+1,m}$  的最后一列的前  $m$  个元素上, 所以我们需要保留所有的 Givens 变换;

(2) 残量计算:  $\|r_m\|_2 = |\beta q_1(m+1)| = |\beta Q_{m+1}(m+1, 1)|$ , 即

$$G_m G_{m-1} \cdots G_2 G_1(\beta e_1)$$

的最后一个分量的绝对值. 由于在计算  $r_{m-1}$  时就已经计算出  $G_{m-1} \cdots G_2 G_1(\beta e_1)$ , 因此这里只需做一次 Givens 变换即可;

(3)  $y^{(m)}$  的计算: 当相对残量满足精度要求时, 需要计算  $y^{(m)} = R_m^{-1} q_1(1:m)$ , 而  $q_1$  即为

$$G_m G_{m-1} \cdots G_2 G_1 (\beta e_1).$$

### 算法 7.5. 实用 GMRES 方法

```

1: 选取初值 $x^{(0)}$, 停机标准 $\varepsilon > 0$, 以及最大迭代步数 IterMax
2: $r_0 = b - Ax^{(0)}$, $\beta = \|r_0\|_2$
3: if $\beta/\|b\|_2 < \varepsilon$ then
4: 停止计算, 输出近似解 $x^{(0)}$
5: end if
6: $v_1 = r_0/\beta$
7: $\xi = \beta e_1$ % 记录 q_1
8: for $j = 1$ to IterMax do
9: $w = Av_j$
10: for $i = 1$ to j do % Arnoldi 过程
11: $h_{i,j} = (v_i, w)$
12: $w = w - h_{i,j}v_i$
13: end for
14: $h_{j+1,j} = \|w\|_2$
15: if $h_{j+1,j} = 0$ then % 迭代中断
16: $m = j$, break
17: end if
18: $v_{j+1} = w/h_{j+1,j}$
19: for $i = 1$ to $j - 1$ do % 计算 $G_{j-1} \cdots G_2 G_1 H_{j+1,j}(1:j, j)$
20:
$$\begin{bmatrix} h_{ij} \\ h_{i+1,j} \end{bmatrix} = \begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix} \begin{bmatrix} h_{ij} \\ h_{i+1,j} \end{bmatrix}$$

21: end for
22: if $|h_{jj}| > |h_{j+1,j}|$ then % 构造 Givens 变换 G_j
23: $\tau = h_{j+1,j}/h_{jj}$, $c_j = 1/\sqrt{1+\tau^2}$, $s_j = c_j\tau$
24: else
25: $\tau = h_{jj}/h_{j+1,j}$, $s_j = 1/\sqrt{1+\tau^2}$, $c_j = s_j\tau$
26: end if
27: $h_{jj} = c_j h_{jj} + s_j h_{j+1,j}$ % 计算 $G_j H_{j+1,j}(1:j, j)$
28: $h_{j+1,j} = 0$
29:
$$\begin{bmatrix} \xi_j \\ \xi_{j+1} \end{bmatrix} = \begin{bmatrix} c_j & s_j \\ -s_j & c_j \end{bmatrix} \begin{bmatrix} \xi_j \\ 0 \end{bmatrix}$$
 % 计算 $G_j(\beta G_{j-1} \cdots G_2 G_1 e_1)$
30: relres = $|\xi_{j+1}|/\beta$ % 相对残量
31: if relres $< \varepsilon$ then
32: $m = j$, break

```



```

33: end if
34: end for
35: m = j
36: $y^{(m)} = H(1:m, 1:m) \backslash \xi(1:m)$ % 求最小二乘问题的解, 回代求解
37: $x^{(m)} = x^{(0)} + V_m y^{(m)}$
38: if relres < ε then
39: 输出近似解 x 及相关信息
40: else
41: 输出算法失败信息
42: end if

```

### 7.2.3 GMRES 方法的中断

在上面的 GMRES 方法中, 当执行到某一步时有  $h_{k+1,k} = 0$ , 则算法会中断 (breakdown). 如果出现这种中断, 则我们就找到了精确解.

**定理 7.4** 设  $A \in \mathbb{R}^{n \times n}$  非奇异且  $r_0 \neq 0$ . 若  $h_{i+1,i} \neq 0, i = 1, 2, \dots, k-1$ , 则  $h_{k+1,k} = 0$  当且仅当  $x^{(k)}$  是方程组的精确解. (不考虑舍入误差) (板书)

**证明.** 设  $h_{k+1,k} = 0$ , 则有

$$AV_k = V_k H_k, \quad y^{(k)} = H_k^{-1}(\beta e_1).$$

所以

$$\begin{aligned} \|r_k\|_2 &= \|b - Ax^{(k)}\|_2 = \|b - A(x^{(0)} + V_k y^{(k)})\|_2 \\ &= \|r_0 - V_k H_k y^{(k)}\|_2 = \|\beta v_1 - V_k(\beta e_1)\|_2 = 0. \end{aligned}$$

反之, 设  $x^{(k)}$  是精确解, 则

$$0 = b - Ax^{(k)} = r_0 - V_{k+1} H_{k+1,k} y^{(k)} = V_{k+1}(\beta e_1 - H_{k+1,k} y^{(k)}).$$

反证法, 假设  $h_{k+1,k} \neq 0$ , 则  $v_{k+1} \neq 0$ . 因此  $V_{k+1}$  单位列正交, 故列满秩, 所以由上式可知

$$\beta e_1 - H_{k+1,k} y^{(k)} = 0.$$

由于  $H_{k+1,k}$  是上 Hessenberg 矩阵, 且  $h_{i+1,i} \neq 0, i = 1, 2, \dots, k$ . 通过向后回代求解可得  $y^{(k)} = 0$ , 于是  $\beta = 0$ . 这与  $r_0 \neq 0$  矛盾. 所以  $h_{k+1,k} = 0$   $\square$

### 7.2.4 带重启的 GMRES 方法

由于随着迭代步数的增加, GMRES 方法的每一步所需的运算量和存储量都会越来越大. 因此当迭代步数很大时, GMRES 方法就不太实用. 通常的解决方法就是重启, 即事先设定一个重启迭代步数  $k$ , 如  $k = 20$  或  $50$  等等, 当 GMRES 达到这个迭代步数时仍不收敛, 则计算出方程组在  $x^{(0)} + \mathcal{K}_k$  中的最佳近似解  $x^{(k)}$ , 然后令  $x^{(0)} = x^{(k)}$ , 并重新开始新的 GMRES 迭代. 不断重复该过程, 直到收敛为止.



**算法 7.6.** GMRES( $k$ ): 带重启的 GMRES 方法

```

1: 设定重启步数 k ($k \ll n$)
2: 选取初值 $x^{(0)}$, 停机标准 $\varepsilon > 0$, 以及最大迭代步数 IterMax
3: $r_0 = b - Ax^{(0)}$, $\beta = \|r_0\|_2$
4: if $\beta/\|b\|_2 < \varepsilon$ then
5: 停止计算, 输出近似解 $x = x^{(0)}$
6: end if
7: for iter=1 to $\text{ceil}(\text{IterMax}/k)$ do % 外循环
8: $v_1 = r_0/\beta$
9: $\xi = \beta e_1$
10: for $j = 1$ to k do
11: 调用 GMRES 循环
12: end for
13: $m = j$
14: $y^{(m)} = H(1:m, 1:m) \backslash \xi(1:m)$
15: $x^{(m)} = x^{(0)} + V_m y^{(m)}$
16: if relres $< \varepsilon$ then
17: break
18: end if
19: $x^{(0)} = x^{(m)}$ % 重启 GMRES
20: $r_0 = b - Ax^{(0)}$, $\beta = \|r_0\|_2$
21: end for
22: if relres $< \varepsilon$ then
23: 输出近似解 $x^{(m)}$ 及相关信息
24: else
25: 输出算法失败信息
26: end if

```

带重启的 GMRES 方法需要注意的问题:

- (1) 如何选取合适的重启步数  $k$ ? 一般只能依靠经验来选取;
- (2) 不带重启的 GMRES 方法能保证算法的收敛性, 但带重启的 GMRES 方法却无法保证, 有时可能出现停滞现象 (*stagnation*).



### 7.3 共轭梯度法

**共轭梯度法** (Conjugate Gradient, CG) 是当前求解对称正定线性方程组的首选方法. CG 算法可以从优化角度得出, 也可以从代数角度推导. 我们这里就从代数角度来介绍 CG 算法.

#### 7.3.1 算法基本过程

我们首先给出 CG 算法的一个性质, 然后根据这个性质来推导 CG 算法.

**定理 7.5** 设  $A$  对称正定, 则

$$x^{(m)} = \arg \min_{x \in x^{(0)} + \mathcal{K}_m} \|x - x_*\|_A \quad (7.9)$$

的充要条件是

$$x^{(m)} \in x^{(0)} + \mathcal{K}_m \quad \text{且} \quad b - Ax^{(m)} \perp \mathcal{K}_m. \quad (7.10)$$

(板书)

**证明.** 首先证明充分性. 设  $x^{(m)}$  满足 (7.10). 记  $\tilde{x} = x^{(m)} - x^{(0)}$ , 则  $\tilde{x} \in \mathcal{K}_m$  且

$$r_0 - A\tilde{x} \perp \mathcal{K}_m.$$

由正交投影的性质可知,  $\tilde{x}$  是最佳逼近问题

$$\min_{x \in \mathcal{K}_m} \|x - A^{-1}r_0\|_A$$

的解, 即

$$\begin{aligned} \tilde{x} &= \arg \min_{x \in \mathcal{K}_m} \|x - A^{-1}r^{(0)}\|_A \\ &= \arg \min_{x \in \mathcal{K}_m} \|x - A^{-1}(b - Ax^{(0)})\|_A \\ &= \arg \min_{x \in \mathcal{K}_m} \|(x^{(0)} + x) - x_*\|_A. \end{aligned}$$

所以,  $x^{(m)} = x^{(0)} + \tilde{x}$  是最佳逼近问题

$$\min_{x \in x^{(0)} + \mathcal{K}_m} \|x - x_*\|_A$$

的解, 即结论 (7.9) 成立.

必要性只需利用正交投影的性质即可, 见作业 7.3. □

当  $A$  对称正定时, Arnoldi 过程就转化为 Lanczos 过程, 且

$$\begin{aligned} AV_m &= V_{m+1}T_{m+1,m} = V_mT_m + \beta_m v_{m+1}e_m^T, \\ V_m^T AV_m &= T_m, \end{aligned}$$

其中  $T_m = \text{tridiag}(\beta_i, \alpha_{i+1}, \beta_{i+1})$ , 见 (7.3). 由定理 7.5 可知, 此时我们需要在  $x^{(0)} + \mathcal{K}_m$  寻找最优近似解  $x^{(m)}$ , 满足

$$b - Ax^{(m)} \perp \mathcal{K}_m. \quad (7.11)$$

根据这个性质, 我们就可以给出 CG 算法.



设  $x^{(m)} = x^{(0)} + V_m z^{(m)}$ , 其中  $z^{(m)} \in \mathbb{R}^m$ . 由 (7.11) 可知

$$0 = V_m^T(b - Ax^{(m)}) = V_m^T(r_0 - AV_m z^{(m)}) = V_m^T(\beta v_1) - V_m^T AV_m z^{(m)} = \beta e_1 - T_m z^{(m)}.$$

因此,

$$z^{(m)} = T_m^{-1}(\beta e_1).$$

于是可得

$$x^{(m)} = x^{(0)} + V_m z^{(m)} = x^{(0)} + V_m T_m^{-1}(\beta e_1).$$

如果  $x^{(m)}$  满足精度要求, 则计算结束. 否则令  $m \leftarrow m+1$ , 继续下一步迭代. 这就是 CG 方法的基本过程.

### 7.3.2 实用迭代格式

下面我们推导 CG 方法的具体实施过程. 由于  $A$  是对称正定的, 因此我们可以借助三项递推公式来简化运算.

首先, 根据  $T_m$  对称正定性, 我们知道  $T_m$  存在  $LDL^T$  分解, 即  $T_m = L_m D_m L_m^T$ . 于是

$$x^{(m)} = x^{(0)} + V_m z^{(m)} = x^{(0)} + V_m T_m^{-1}(\beta e_1^{(m)}) = x^{(0)} + (V_m L_m^{-T})(\beta D_m^{-1} L_m^{-1} e_1^{(m)}),$$

其中  $e_1^{(m)}$  表示  $m$  阶单位矩阵的第一列. 如果  $x^{(m)}$  满足精度要求, 则计算结束. 否则我们需要计算  $x^{(m+1)}$ , 即

$$x^{(m+1)} = x^{(0)} + V_{m+1} T_{m+1}^{-1}(\beta e_1^{(m+1)}) = x^{(0)} + (V_{m+1} L_{m+1}^{-T})(\beta D_{m+1}^{-1} L_{m+1}^{-1} e_1^{(m+1)}).$$

这里假定  $T_{m+1}$  的  $LDL^T$  分解为  $T_{m+1} = L_{m+1} D_{m+1} L_{m+1}^T$ . 下面就考虑如何利用递推方式来计算  $x^{(m+1)}$ . 记

$$\tilde{P}_m \triangleq V_m L_m^{-T} = [\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_m] \in \mathbb{R}^{n \times m},$$

$$y_m \triangleq \beta D_m^{-1} L_m^{-1} e_1^{(m)} = [\eta_1, \dots, \eta_m]^T \in \mathbb{R}^m.$$

我们首先证明下面的结论.

**引理 7.6** 设  $\tilde{P}_m$  和  $y_m$  由上面的式子所定义, 则下面的递推公式成立:

$$\tilde{P}_{m+1} \triangleq V_{m+1} L_{m+1}^{-T} = [\tilde{P}_m, \tilde{p}_{m+1}],$$

$$y_{m+1} \triangleq \beta D_{m+1}^{-1} L_{m+1}^{-1} e_1^{(m+1)} = [y_m^T, \eta_{m+1}]^T, \quad m = 1, 2, \dots$$

(板书)

**证明.** 设  $T_m$  的  $LDL^T$  分解为

$$T_m = L_m D_m L_m^T = \begin{bmatrix} 1 & & & \\ l_1 & 1 & & \\ & \ddots & \ddots & \\ & & l_{m-1} & 1 \end{bmatrix} \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_m \end{bmatrix} \begin{bmatrix} 1 & & & \\ l_1 & 1 & & \\ & \ddots & \ddots & \\ & & l_{m-1} & 1 \end{bmatrix}^T.$$

由待定系数法可知  $d_1 = \alpha_1$ ,

$$l_i = \beta_i / d_i, \quad d_{i+1} = \alpha_{i+1} - l_i \beta_i, \quad i = 1, 2, \dots, m-1.$$





记  $\gamma = [0, \dots, 0, \beta_m]^\top \in \mathbb{R}^m$ , 则  $T_{m+1}$  的  $\text{LDL}^\top$  分解为

$$T_{m+1} = \begin{bmatrix} T_m & \gamma_m \\ \gamma_m^\top & \alpha_{m+1} \end{bmatrix} = L_{m+1} D_{m+1} L_{m+1}^\top$$

$$= \begin{bmatrix} 1 & & & & \\ l_1 & 1 & & & \\ & \ddots & \ddots & & \\ & & l_{m-1} & 1 & \\ & & & l_m & 1 \end{bmatrix} \begin{bmatrix} d_1 & & & & \\ & d_2 & & & \\ & & \ddots & & \\ & & & d_m & \\ & & & & d_{m+1} \end{bmatrix} \begin{bmatrix} 1 & & & & \\ l_1 & 1 & & & \\ & \ddots & \ddots & & \\ & & l_{m-1} & 1 & \\ & & & l_m & 1 \end{bmatrix}^\top,$$

其中  $l_m = \beta_m/d_m$ ,  $d_{m+1} = \alpha_{m+1} - l_m \beta_m$ . 记  $\tilde{\gamma} = [0, \dots, 0, l_m]^\top \in \mathbb{R}^m$ , 则

$$L_{m+1} = \begin{bmatrix} L_m & 0 \\ \tilde{\gamma}^\top & 1 \end{bmatrix} \quad \text{and} \quad L_{m+1}^{-1} = \begin{bmatrix} L_m^{-1} & 0 \\ -\tilde{\gamma}^\top L_m^{-1} & 1 \end{bmatrix}.$$

所以有

$$\tilde{P}_{m+1} = V_{m+1} L_{m+1}^{-T} = [V_m, v_{m+1}] \begin{bmatrix} L_m^{-T} & -L_m^{-T} \tilde{\gamma} \\ 0 & 1 \end{bmatrix} = [V_m L_m^{-T}, -V_m L_m^{-T} \tilde{\gamma} + v_{m+1}].$$

又  $V_m L_m^{-T} \tilde{\gamma} = \tilde{P}_m [0, \dots, 0, l_m]^\top = l_m \tilde{p}_m$ , 所以  $\tilde{P}_{m+1} = [\tilde{P}_m, \tilde{p}_{m+1}]$ , 其中

$$\tilde{p}_{m+1} = v_{m+1} - l_m \tilde{p}_m. \quad (7.12)$$

另外,

$$\begin{aligned} y_{m+1} &= \beta D_{m+1}^{-1} L_{m+1}^{-1} e_1^{(m+1)} = \beta \begin{bmatrix} D_m^{-1} & 0 \\ 0 & d_{m+1}^{-1} \end{bmatrix} \begin{bmatrix} L_m^{-1} & 0 \\ -\tilde{\gamma}^\top L_m^{-1} & 1 \end{bmatrix} \begin{bmatrix} e_1^{(m)} \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \beta D_m^{-1} L_m^{-1} e_1^{(m)} \\ -\beta d_{m+1}^{-1} \tilde{\gamma}^\top L_m^{-1} e_1^{(m)} \end{bmatrix} \\ &= \begin{bmatrix} y_m \\ \eta_{m+1} \end{bmatrix}, \end{aligned}$$

其中  $\eta_{m+1} \triangleq -\beta d_{m+1}^{-1} \tilde{\gamma}^\top L_m^{-1} e_1^{(m)}$ . 所以结论成立.  $\square$

有了上面的性质, 我们就可以得到从  $x^{(m)}$  到  $x^{(m+1)}$  的递推公式

$$x^{(m+1)} = \tilde{P}_{m+1} y_{m+1} = [\tilde{P}_m, \tilde{p}_{m+1}] \begin{bmatrix} y_m \\ \eta_{m+1} \end{bmatrix} = x^{(m)} + \eta_{m+1} \tilde{p}_{m+1}.$$

为了判断近似解是否满足要求, 我们还需要计算残量. 直接计算可知, 残量满足下面的递推公式:

$$r_{m+1} = b - Ax^{(m+1)} = b - A(x^{(m)} + \eta_{m+1} \tilde{p}_{m+1}) = r_m - \eta_{m+1} A \tilde{p}_{m+1}.$$



另一方面, 我们有

$$\begin{aligned}
 r_m &= b - Ax^{(m)} = b - A(x^{(0)} + V_m z^{(m)}) \\
 &= r_0 - AV_m z^{(m)} \\
 &= \beta v_1 - V_m T_m z^{(m)} - \beta_m v_{m+1} e_m^T z^{(m)} \\
 &= -\beta_m (e_m^T z^{(m)}) v_{m+1},
 \end{aligned}$$

即  $r_m$  与  $v_{m+1}$  平行. 记

$$r_m = \tau_m v_{m+1}, \quad m = 0, 1, 2, \dots, \quad (7.13)$$

其中

$$\tau_0 = \beta = \|r_0\|_2, \quad \tau_m = -\beta_m (e_m^T z^{(m)}), \quad m = 1, 2, \dots$$

引入变量

$$p_m \triangleq \tau_{m-1} \tilde{p}_m, \quad m = 1, 2, \dots$$

由 (7.12) 和 (7.13) 可知,  $\{p_m\}$  满足下面的递推公式:

$$p_{m+1} = \tau_m \tilde{p}_{m+1} = \tau_m (v_{m+1} - l_m \tilde{p}_m) = r_m + \mu_m p_m, \quad (7.14)$$

其中  $\mu_m = -\frac{l_m \tau_m}{\tau_{m-1}}, m = 1, 2, \dots$

于是我们就得到递推公式 ( $m = 1, 2, \dots$ )

$$x^{(m+1)} = x^{(m)} + \eta_{m+1} \tilde{p}_{m+1} = x^{(m)} + \xi_{m+1} p_{m+1}, \quad (7.15)$$

$$r_{m+1} = r_m - \eta_{m+1} A \tilde{p}_{m+1} = r_m - \xi_{m+1} A p_{m+1}, \quad (7.16)$$

其中  $\xi_{m+1} = \frac{\eta_{m+1}}{\tau_m}, m = 1, 2, \dots$

下面需要考虑系数  $\xi_{m+1}$  和  $\mu_m$  的计算方法. 首先给出下面的性质.

**引理 7.7** 对于共轭梯度法, 下面的结论成立:

- (1)  $r_1, r_2, \dots, r_m$  相互正交;
- (2)  $p_1, p_2, \dots, p_m$  相互  $A$ -共轭 (或  $A$ -正交), 即当  $i \neq j$  时有  $p_i^T A p_j = 0$ .

(板书)

**证明.** (1) 由于  $r_k$  与  $v_{k+1}$  平行, 所以结论成立.

(2) 只需证明  $P_m^T A P_m$  是对角矩阵即可, 即证  $\tilde{P}_m^T A \tilde{P}_m$  是对角矩阵. 通过直接计算可得

$$\begin{aligned}
 \tilde{P}_m^T A \tilde{P}_m &= (V_m L_m^{-T})^T A V_m L_m^{-T} \\
 &= L_m^{-1} V_m^T A V_m L_m^{-T} \\
 &= L_m^{-1} T_m L_m^{-T} \\
 &= L_m^{-1} (L_m D_m L_m^T) L_m^{-T} \\
 &= D_m.
 \end{aligned}$$

□



下面给出  $\xi_{m+1}$  和  $\mu_m$  的计算公式. 在等式 (7.14) 两边同时左乘  $p_{m+1}^\top A$  可得

$$p_{m+1}^\top A p_{m+1} = p_{m+1}^\top A r_m + \mu_m p_{m+1}^\top A p_m = r_m^\top A p_{m+1}.$$

再用  $r_m^\top$  左乘方程 (7.16) 可得

$$0 = r_m^\top r_{m+1} = r_m^\top r_m - \xi_{m+1} r_m^\top A p_{m+1}.$$

于是

$$\xi_{m+1} = \frac{r_m^\top r_m}{r_m^\top A p_{m+1}} = \frac{r_m^\top r_m}{p_{m+1}^\top A p_{m+1}}. \quad (7.17)$$

在等式 (7.14) 两边同时左乘  $p_m^\top A$  可得

$$0 = p_m^\top A p_{m+1} = p_m^\top A r_m + \mu_m p_m^\top A p_m.$$

于是

$$\mu_m = -\frac{p_m^\top A r_m}{p_m^\top A p_m} = -\frac{r_m^\top A p_m}{p_m^\top A p_m}. \quad (7.18)$$

为了进一步减少运算量, 我们还可以将上式进行简化. 用  $r_{m+1}^\top$  左乘方程 (7.16) 可得

$$r_{m+1}^\top r_{m+1} = r_{m+1}^\top r_m - \xi_{m+1} r_{m+1}^\top A p_{m+1} = -\frac{r_m^\top r_m}{p_{m+1}^\top A p_{m+1}} \cdot r_{m+1}^\top A p_{m+1}.$$

于是

$$r_{m+1}^\top A p_{m+1} = -\frac{r_{m+1}^\top r_{m+1}}{r_m^\top r_m} \cdot p_{m+1}^\top A p_{m+1}.$$

所以  $r_m^\top A p_m = -\frac{r_m^\top r_m}{r_{m-1}^\top r_{m-1}} \cdot p_m^\top A p_m$ , 代入 (7.18) 可得

$$\mu_m = -\frac{r_m^\top A p_m}{p_m^\top A p_m} = \frac{r_m^\top r_m}{r_{m-1}^\top r_{m-1}}. \quad (7.19)$$

综上, 由 (7.19), (7.14), (7.17) 和 (7.15), (7.16) 就构成了 CG 算法的迭代过程, 即

$$\begin{aligned} p_{m+1} &= r_m + \mu_m p_m, \quad \text{其中} \quad \mu_m = \frac{r_m^\top r_m}{r_{m-1}^\top r_{m-1}}, \\ x^{(m+1)} &= x^{(m)} + \xi_{m+1} p_{m+1}, \\ r_{m+1} &= r_m - \xi_{m+1} A p_{m+1}, \quad \text{其中} \quad \xi_{m+1} = \frac{r_m^\top r_m}{p_{m+1}^\top A p_{m+1}}. \end{aligned}$$

注意, 以上递推公式是从  $m=1$  开始的. 因此  $m=0$  时的计算公式需要另外推导.

首先, 由  $\tilde{p}_1$  的定义可知

$$\tilde{p}_1 = \tilde{P}_1 = V_1 L_1^{-T} = v_1.$$

因此

$$p_1 = \tau_0 \tilde{p}_1 = \beta v_1 = r_0.$$



其次, 由 Lanczos 过程可知  $T_1 = \alpha_1 = v_1^T A v_1$ . 注意到  $\beta = r_0^T r_0$ , 于是

$$x^{(1)} = x^{(0)} + V_1 T_1^{-1} (\beta e_1) = x^{(0)} + \frac{\beta}{v_1^T A v_1} v_1 = x^{(0)} + \frac{r_0^T r_0}{p_1^T A p_1} p_1.$$

令  $\xi_1 = \frac{r_0^T r_0}{p_1^T A p_1}$  (注: 之前的  $\xi_{m+1}$  计算公式 (7.17) 只对  $m \geq 1$  有定义), 则当  $m = 0$  时关于  $x^{(m+1)}$  的递推公式仍然成立.

最后考虑残量. 易知

$$r_1 = b - A x^{(1)} = b - A x^{(0)} - \frac{r_0^T r_0}{p_1^T A p_1} A p_1 = r_0 - \xi_1 A p_1,$$

即当  $m = 0$  时关于  $r_{m+1}$  的递推公式也成立.

于是, 我们就可以给出下面的 CG 算法.

#### 算法 7.7. 共轭梯度法 (CG)


```

1: 选取初值 $x^{(0)}$, 停机准则 $\varepsilon > 0$, 最大迭代步数 IterMax
2: $r_0 = b - A x^{(0)}$
3: $\beta = \|r_0\|_2$
4: if $\beta < \varepsilon$ then
5: 停止迭代, 输出近似解 $x^{(0)}$
6: end if
7: for $m = 1$ to IterMax do
8: $\rho = r_{m-1}^T r_{m-1}$
9: if $m > 1$ then
10: $\mu_{m-1} = \rho / \rho_0$
11: $p_m = r_{m-1} + \mu_{m-1} p_{m-1}$
12: else
13: $p_m = r_0$
14: end if
15: $q_m = A p_m$
16: $\xi_m = \rho / (p_m^T q_m)$
17: $x^{(m)} = x^{(m-1)} + \xi_m p_m$
18: $r_m = r_{m-1} - \xi_m q_m$
19: $\text{relres} = \|r_m\|_2 / \beta$
20: if $\text{relres} < \varepsilon$ then
21: 停止迭代, 输出近似解 $x^{(m)}$
22: end if
23: $\rho_0 = \rho$
24: end for

```



```
25: if relres $< \varepsilon$ then
26: 输出近似解 $x^{(m)}$ 及相关信息
27: else
28: 输出算法失败信息
29: end if
```

 CG 算法的每个迭代步的主要运算为一个矩阵向量乘积和两个向量内积;

## 7.4 收敛性分析

### 7.4.1 CG 的收敛性

设  $x_*$  是解析解,  $x^{(m)} \in x^{(0)} + \mathcal{K}_m$  是 CG 算法在  $x^{(0)} + \mathcal{K}_m$  中找到的近似解, 即

$$x^{(m)} = \arg \min_{x \in x^{(0)} + \mathcal{K}_m} \|x - x_*\|_A.$$

记  $\mathbb{P}_k$  为所有次数不超过  $k$  的多项式的集合. 对任意  $x \in x^{(0)} + \mathcal{K}_m$ , 存在  $p(t) \in \mathbb{P}_{m-1}$ , 使得

$$x = x^{(0)} + p(A)r_0.$$

于是有

$$x - x_* = \varepsilon_0 + p(A)(b - Ax^{(0)}) = \varepsilon_0 + p(A)(Ax_* - Ax^{(0)}) = (I - Ap(A))\varepsilon_0 \triangleq q(A)\varepsilon_0,$$

其中  $\varepsilon_0 = x^{(0)} - x_*$ , 多项式  $q(t) = 1 - tp(t) \in \mathbb{P}_m$  且  $q(0) = 1$ . 所以

$$\|x - x_*\|_A^2 = \varepsilon_0^T q(A)^T A q(A) \varepsilon_0.$$

设  $A = Q\Lambda Q^T$  是  $A$  的特征值分解, 其中  $Q$  是正交矩阵,  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  是对角阵, 且  $\lambda_i > 0$ . 记  $y = [y_1, y_2, \dots, y_n]^T \triangleq Q^T \varepsilon_0$ , 则

$$\begin{aligned} \|x^{(m)} - x_*\|_A^2 &= \min_{x \in x^{(0)} + \mathcal{K}_m} \|x - x_*\|_A^2 \\ &= \min_{q \in \mathbb{P}_m, q(0)=1} \varepsilon_0^T q(A)^T A q(A) \varepsilon_0 \\ &= \min_{q \in \mathbb{P}_m, q(0)=1} \varepsilon_0^T Q q(\Lambda)^T \Lambda q(\Lambda) Q^T \varepsilon_0 \\ &= \min_{q \in \mathbb{P}_m, q(0)=1} y^T q(\Lambda)^T \Lambda q(\Lambda) y \\ &= \min_{q \in \mathbb{P}_m, q(0)=1} \sum_{i=1}^n y_i^2 \lambda_i q(\lambda_i)^2 \\ &\leq \min_{q \in \mathbb{P}_m, q(0)=1} \max_{1 \leq i \leq n} \{q(\lambda_i)^2\} \sum_{i=1}^n y_i^2 \lambda_i \\ &= \min_{q \in \mathbb{P}_m, q(0)=1} \max_{1 \leq i \leq n} \{q(\lambda_i)^2\} y^T \Lambda y \\ &= \min_{q \in \mathbb{P}_m, q(0)=1} \max_{1 \leq i \leq n} \{q(\lambda_i)^2\} \varepsilon_0^T A \varepsilon_0 \\ &= \min_{q \in \mathbb{P}_m, q(0)=1} \max_{1 \leq i \leq n} \{q(\lambda_i)^2\} \|\varepsilon_0\|_A^2, \end{aligned}$$

即

$$\frac{\|x^{(m)} - x_*\|_A^2}{\|x^{(0)} - x_*\|_A^2} \leq \min_{q \in \mathbb{P}_m, q(0)=1} \max_{1 \leq i \leq n} \{q(\lambda_i)^2\}.$$

因此, 我们有下面的结论.

**引理 7.8** 设  $A \in \mathbb{R}^{n \times n}$  对称正定,  $x^{(m)}$  是 CG 算法迭代  $m$  步后得到的近似解. 则

$$\frac{\|x^{(m)} - x_*\|_A}{\|x^{(0)} - x_*\|_A} \leq \min_{q \in \mathbb{P}_m, q(0)=1} \max_{1 \leq i \leq n} |q(\lambda_i)|. \quad (7.20)$$



由于  $A$  的特征值通常是不知道的, 因此不等式 (7.20) 右端通常难以计算. 但在许多情况下, 我们可以通过一些近似方法估计出  $A$  的最大特征值  $\lambda_1$  和最小特征值  $\lambda_n$ . 假设  $\lambda_1$  和  $\lambda_n$  已知, 则不等式 (7.20) 右端可以放缩为

$$\min_{q \in \mathbb{P}_m, q(0)=1} \max_{\lambda_n \leq \lambda \leq \lambda_1} |q(\lambda)|. \quad (7.21)$$

由 Chebyshev 多项式的最佳逼近性质 (定理 6.33) 可知, 最小最大问题 (7.21) 的解为

$$q_*(t) = \frac{T_m\left(\frac{2t - (\lambda_1 + \lambda_n)}{\lambda_1 - \lambda_n}\right)}{T_m\left(-\frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n}\right)},$$

其中  $T_m$  为  $m$  次 Chebyshev 多项式. 由于  $T_m(t) = (-1)^m T_m(t)$ , 且当  $|t| \leq 1$  时有  $|T_m(t)| \leq 1$ , 所以

$$|q_*(t)| \leq \frac{1}{T_m\left(\frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n}\right)}.$$

又当  $|t| > 1$  时

$$T_m(t) = \frac{1}{2} \left[ \left(t + \sqrt{t^2 - 1}\right)^m + \left(t + \sqrt{t^2 - 1}\right)^{-m} \right] \geq \frac{1}{2} \left(t + \sqrt{t^2 - 1}\right)^m,$$

所以

$$\begin{aligned} T_m\left(\frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n}\right) &\geq \frac{1}{2} \left( \frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n} + \sqrt{\frac{(\lambda_1 + \lambda_n)^2}{(\lambda_1 - \lambda_n)^2} - 1} \right)^m \\ &= \frac{1}{2} \left( \frac{\lambda_1 + \lambda_n}{\lambda_1 - \lambda_n} + \frac{2\sqrt{\lambda_1 \lambda_n}}{\lambda_1 - \lambda_n} \right)^m \\ &= \frac{1}{2} \left( \frac{\sqrt{\lambda_1} + \sqrt{\lambda_n}}{\sqrt{\lambda_1} - \sqrt{\lambda_n}} \right)^m \\ &= \frac{1}{2} \left( \frac{\sqrt{\kappa(A)} + 1}{\sqrt{\kappa(A)} - 1} \right)^m. \end{aligned}$$

其中  $\kappa(A) = \frac{\lambda_1}{\lambda_n}$  为  $A$  的条件数. 因此我们可以得到下面的收敛性定理.

**定理 7.9** 设  $A \in \mathbb{R}^{n \times n}$  对称正定,  $x^{(m)}$  是 CG 算法迭代  $m$  步后得到的近似解. 则

$$\frac{\|x^{(m)} - x_*\|_A}{\|x^{(0)} - x_*\|_A} \leq 2 \left( \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^m. \quad (7.22)$$

#### 7.4.2 CG 的超收敛性

如果我们能够获得  $A$  的更多的特征值信息, 则能得到更好的误差限 [6].

**定理 7.10** 设  $A$  对称正定, 特征值为

$$0 < \lambda_n \leq \cdots \leq \lambda_{n+1-i} \leq b_1 \leq \lambda_{n-i} \leq \cdots \leq \lambda_{j+1} \leq b_2 \leq \lambda_j \leq \cdots \leq \lambda_1.$$



则当  $m \geq i + j$  时有

$$\frac{\|x^{(m)} - x_*\|_A}{\|x^{(0)} - x_*\|_A} \leq 2 \left( \frac{b-1}{b+1} \right)^{m-i-j} \max_{\lambda \in [b_1, b_2]} \left\{ \prod_{k=n+1-i}^n \left( \frac{\lambda - \lambda_k}{\lambda_k} \right) \prod_{k=1}^j \left( \frac{\lambda_k - \lambda}{\lambda_k} \right) \right\}, \quad (7.23)$$

其中

$$b = \left( \frac{b_2}{b_1} \right)^{\frac{1}{2}} \geq 1.$$

由此可知, 当  $b_1$  与  $b_2$  非常接近时, 迭代  $i + j$  步后, CG 算法收敛会非常快.

**推论 7.11** 设  $A$  对称正定, 特征值为

$$0 < \delta \leq \lambda_n \leq \cdots \leq \lambda_{n+1-i} \leq 1 - \varepsilon \leq \lambda_{n-i} \leq \cdots \leq \lambda_{j+1} \leq 1 + \varepsilon \leq \lambda_j \leq \cdots \leq \lambda_1.$$

则当  $m \geq i + j$  时有

$$\frac{\|x^{(m)} - x_*\|_A}{\|x^{(0)} - x_*\|_A} \leq 2 \left( \frac{1 + \varepsilon}{\delta} \right)^i \varepsilon^{m-i-j}. \quad (7.24)$$

### 7.4.3 GMRES 的收敛性

#### 正规矩阵情形

设  $A$  是正规矩阵, 即

$$A = U\Lambda U^*,$$

其中  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  的对角线元素  $\lambda_i \in \mathbb{C}$  为  $A$  的特征值.

设  $x \in x^{(0)} + \mathcal{K}_k(A, r_0)$ , 则存在多项式  $p(t) \in \mathbb{P}_{k-1}$  使得  $x = x^{(0)} + p(A)r_0$ . 于是

$$b - Ax = b - Ax^{(0)} - Ap(A)r_0 = (I - Ap(A))r_0 \triangleq q(A)r_0, \quad (7.25)$$

其中  $q(t) = 1 - tp(t) \in \mathbb{P}_k$  满足  $q(0) = 1$ . 直接计算可知

$$\|b - Ax\|_2 = \|q(A)r_0\|_2 = \|Uq(\Lambda)U^*r_0\|_2 \leq \|U\|_2\|U^*\|_2\|q(\Lambda)\|_2\|r_0\|_2 = \|q(\Lambda)\|_2\|r_0\|_2.$$

又  $\Lambda$  是对角矩阵, 所以

$$\|q(\Lambda)\|_2 = \max_{1 \leq i \leq n} |q(\lambda_i)|.$$

设  $x^{(k)}$  是近似解, 则由 GMRES 方法的最优性可知,  $x^{(k)}$  极小化残量的 2-范数. 因此,

$$\begin{aligned} \|b - Ax^{(k)}\|_2 &= \min_{x \in x^{(0)} + \mathcal{K}_k(A, r_0)} \|b - Ax\|_2 \\ &= \min_{q \in \mathbb{P}_k, q(0)=1} \|q(A)r_0\|_2 \\ &\leq \min_{q \in \mathbb{P}_k, q(0)=1} \|q(\Lambda)\|_2\|r_0\|_2 \\ &= \|r_0\|_2 \min_{q \in \mathbb{P}_k, q(0)=1} \max_{1 \leq i \leq n} |q(\lambda_i)|. \end{aligned}$$

于是, 我们有下面的结论.





**定理 7.12** 设  $A \in \mathbb{R}^{n \times n}$  是正规矩阵,  $x^{(k)}$  是 GMRES 方法得到的近似解, 则

$$\frac{\|b - Ax^{(k)}\|_2}{\|r_0\|_2} \leq \min_{q \in \mathbb{P}_k, q(0)=1} \max_{1 \leq i \leq n} |q(\lambda_i)|. \quad (7.26)$$

需要指出的是, 上界 (7.29) 是紧凑的 [58, 76]. 与 CG 算法类似, 我们也可以选取一个包含  $A$  的所有特征值的区域  $\Omega \subset \mathbb{C}$ , 然后将定理 7.12 中的上界缩放为

$$\min_{q \in \mathbb{P}_k, q(0)=1} \max_{\lambda \in \Omega} |q(\lambda)|. \quad (7.27)$$


通常  $\Omega$  必须是连通的, 否则 (7.27) 的求解非常困难. 即使是两个区间的并都没法求解 [44]. 另外,  $\Omega$  不能包含原点.

我们首先考虑一种最简单的情形, 即  $\Omega$  是复平面上一个圆盘, 圆心为  $C(c, 0)$ , 半径  $r > 0$ . 这里假定  $r < c$ , 因此原点不在  $\Omega$  内.

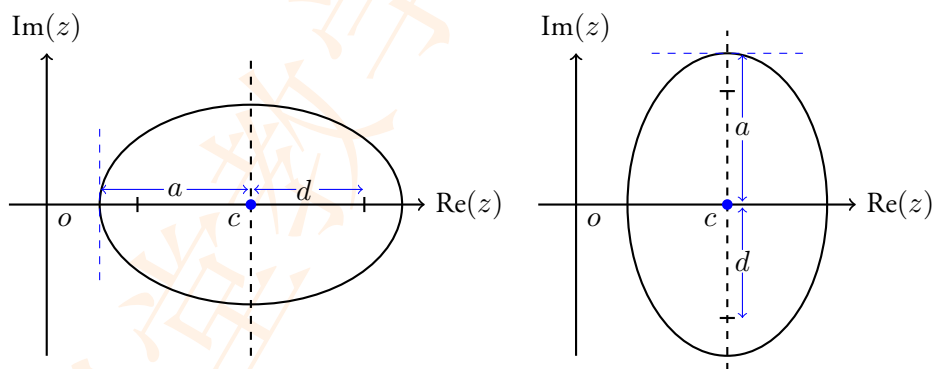
定义复系数多项式  $q_k(z) = ((c - z)/c)^k$ , 则  $q_k(z) \in \mathbb{P}_k$  且  $q_k(0) = 1$ . 所以

$$\min_{q \in \mathbb{P}_k, q(0)=1} \max_{1 \leq i \leq n} |q(\lambda_i)| \leq \max_{1 \leq i \leq n} |q_k(\lambda_i)| = \max_{1 \leq i \leq n} \left| \frac{(c - \lambda_i)^k}{c^k} \right| \leq \frac{r^k}{|c|^k}.$$

由此可见, 当  $r$  越小或  $c$  越大时, 右式趋向于 0 的速度就越快.

 设  $A$  是正规矩阵, 其特征值包含在一个圆盘内, 则圆盘的半径越小或者圆心离原点越远, 则 GMRES 收敛越快.

如果用椭圆来代替圆盘, 则可得出更紧凑的上界. 设  $A$  的特征值全部包含在椭圆  $E(c, d, a)$  内, 其中  $d > 0$  为焦距,  $a > 0$  为半长轴长,  $C(c, 0)$  为椭圆中心. 并且假定原点不在椭圆内. 如下图所示.



令  $T_k$  为  $k$  次复 Chebyshev 多项式, 则多项式

$$\tilde{T}_k(z) = \frac{T_k(\frac{c-z}{d})}{T_k(\frac{c}{d})}$$

就是极小极大问题

$$\min_{q \in \mathbb{P}_k, q(0)=1} \max_{\lambda \in E(c, d, a)} |q(\lambda)| \quad (7.28)$$

的渐进最优解. 于是

$$\min_{q \in \mathbb{P}_k, q(0)=1} \max_{1 \leq i \leq n} |q(\lambda_i)| \leq \min_{q \in \mathbb{P}_k, q(0)=1} \max_{\lambda \in E(c, d, a)} |q(\lambda)| \leq \max_{\lambda \in E(c, d, a)} |\tilde{T}_k(\lambda)|.$$

由复 Chebyshev 多项式的性质可知

$$\max_{\lambda \in E(c,d,a)} |\tilde{T}_k(\lambda)| = \frac{T_k(\frac{a}{d})}{|T_k(\frac{c}{d})|}.$$

因此

$$\min_{q \in \mathbb{P}_k, q(0)=1} \max_{1 \leq i \leq n} |q(\lambda_i)| \leq \frac{T_k(\frac{a}{d})}{|T_k(\frac{c}{d})|}.$$

所以我们可得下面的结论.

**推论 7.13** 设  $A \in \mathbb{R}^{n \times n}$  是正规矩阵,  $x^{(k)}$  是由 GMRES 得到的近似解, 则

$$\frac{\|b - Ax^{(k)}\|_2}{\|r_0\|_2} \leq \frac{T_k(\frac{a}{d})}{|T_k(\frac{c}{d})|}. \quad (7.29)$$

虽然  $\tilde{T}_k(z)$  通常不是极小极大问题 (7.28) 的解, 但上界 (7.29) 却往往能给出 GMRES 方法收敛速度的一个很好的估计 [88].

### 非正规情形

当  $A$  不是正规矩阵时, 在一般情况下, GMRES 方法的收敛性很难分析.

如果  $A \in \mathbb{R}^{n \times n}$  是可对角化的, 即

$$A = X\Lambda X^{-1},$$

其中  $X \in \mathbb{C}^{n \times n}$  非奇异,  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  的对角线元素  $\lambda_i \in \mathbb{C}$  为  $A$  的特征值, 则

$$\|b - Ax^{(k)}\|_2 = \min_{x \in x^{(0)} + \mathcal{K}_k(A, r_0)} \|b - Ax\|_2 = \min_{q \in \mathbb{P}_k, q(0)=1} \|q(A)r_0\|_2. \quad (7.30)$$

相类似地, 我们可以得到下面的结论 [107].

**定理 7.14** 设  $A = X\Lambda X^{-1}$ , 其中  $X \in \mathbb{C}^{n \times n}$  非奇异,  $\Lambda$  是对角矩阵,  $x^{(k)}$  是 GMRES 方法得到的近似解, 则

$$\begin{aligned} \frac{\|b - Ax^{(k)}\|_2}{\|r_0\|_2} &\leq \|X\|_2 \|X^{-1}\|_2 \min_{q \in \mathbb{P}_k, q(0)=1} \max_{1 \leq i \leq n} |q(\lambda_i)| \\ &= \kappa(X) \min_{q \in \mathbb{P}_k, q(0)=1} \max_{1 \leq i \leq n} |q(\lambda_i)|, \end{aligned} \quad (7.31)$$

其中  $\kappa(X)$  是  $X$  的谱条件数.

如果  $A$  接近正规, 则  $\kappa(X) \approx 1$ . 此时上界 (7.31) 在一定程度上能描述 GMRES 的收敛速度. 当如果  $X$  远非正交, 则  $\kappa(X)$  会很大, 此时该上界就失去实际意义了.

需要指出的是, 上面的分析并不意味着非正规矩阵就一定比正规矩阵收敛慢. 事实上, 对任意一个非正规矩阵, 总存在一个相应的正规矩阵, 使得 GMRES 方法的收敛速度是一样的 (假定初始残量相同) [3, 59, 60, 87].

虽然 GMRES 方法的收敛性与系数矩阵的特征值有关, 但显然并不仅仅取决于特征值的分布. 事实上, 我们有下面的结论.



**定理 7.15** 对于任意给定的特征值分布和一条不增的收敛曲线, 则总存在一个矩阵  $A$  和一个右端项  $b$ , 使得  $A$  具有指定的特征值分布, 且 GMRES 方法的收敛曲线与给定的收敛曲线相同.

**例 7.1** 考虑线性方程组  $Ax = b$  其中

$$A = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ a_0 & a_1 & a_2 & \cdots & a_{n-1} \end{bmatrix}, \quad b = e_1.$$

当  $a_0 \neq 0$  时,  $A$  非奇异. 易知,  $A$  的特征值多项式为

$$p(x) = \lambda^n - a_{n-1}\lambda^{n-1} - a_{n-2}\lambda^{n-2} - \cdots - a_1\lambda - a_0.$$

方程组的精确解为

$$x = [-a_1/a_0, 1, 0, \dots, 0]^T.$$

以零向量为迭代初值, 则 GMRES 迭代到第  $n$  步时才收敛. (前  $n-1$  步残量范数不变).

(GMRES\_example01.m)

如果  $A$  不可以对角化, 我们在分析 GMRES 方法的收敛性时, 通常会想办法用一个新的极小化问题来近似原来的极小化问题 (7.30). 当然, 这个新的极小化问题应该是比较容易求解的.

事实上, 我们有

$$\frac{\|b - Ax^{(k)}\|_2}{\|r_0\|_2} = \frac{\min_{q \in \mathbb{P}_k, q(0)=1} \|q(A)r_0\|_2}{\|r_0\|_2} \leq \max_{\|v\|_2=1} \min_{q \in \mathbb{P}_k, q(0)=1} \|q(A)v\|_2 \quad (7.32)$$

$$\leq \min_{q \in \mathbb{P}_k, q(0)=1} \|q(A)\|_2. \quad (7.33)$$

不等式 (7.32) 右端代表的是在最坏情况下的 GMRES 收敛性, 而且是紧凑的, 即它是所能找到的不依赖于  $r_0$  的最好上界. 但我们仍然不清楚, 到底是  $A$  的那些性质决定着这个上界 [42].

可以证明, 当  $A$  是正规矩阵时, 上界 (7.32) 和 (7.33) 是相等的 [58, 76]. 但是, 对于大多数非正规矩阵而言, 这两者是否相等或者非常接近, 迄今仍不太清楚.

最后需要指出的是, 算法的收敛性也依赖于迭代初值和右端项. 所以定理 7.9, 7.12 和 7.14 中的上界描述的都是最坏情况下的收敛速度. 也就是说, 在实际计算中, 算法的收敛速度可能会比预想的要快得多.

## 7.5 预处理方法

Nothing will be more central to computational science in the next century than the art of transforming a problem that appears intractable into another whose solution can be approximated rapidly. For Krylov subspace matrix iterations, this is **preconditioning**.

— Trefethen & Bau III [119, page 319], 1997.

为了改善 krylov 迭代方法的收敛性与可靠性, 我们通常需要运用预处理技术 (preconditioning, 也称为预条件). 通俗地说, 预处理就是将原来难以求解的问题转化成一个等价的但比较容易求解的新问题. 预处理技术的研究是目前科学计算领域中的重要研究课题之一.

早在 1948 年, 著名数学家 Turing (也被称为计算机科学之父和人工智能之父) 就提出了预处理这个概念, 用于改善线性方程组的性态.

The experiences of Fox, Huskey, and Wilkinson prompted Turing to write a remarkable paper “Rounding-off errors in matrix processes” [121, 1948]. In this paper, Turing made several important contributions. He formulated the LU (actually, the LDU) factorization of a matrix, .... He used the word “preconditioning” to mean improving the condition of a system of linear equations (a term that did not come into popular use until the 1970s).

— Nicholas J. Higham [68, page 184], 2002.

随后, Evans 在 1968 年提出了类似的想法, 用于降低线性方程组的条件数.

To obtain improved convergence rates for the methods of successive displacement we require the coefficient matrix to have a  $P$ -condition number as small as possible. If this criterion is not satisfied, then it is advisable to prepare the system or “precondition” it beforehand.

— D. J. Evans [41], 1968.

关于预处理的主要参考文献有 [13, 124].

### 7.5.1 预处理方法介绍

对于线性方程组而言, 预处理就是对系数矩阵进行适当的线性转换, 转换为一个新矩阵.

考虑线性方程组

$$Ax = b, \quad A \in \mathbb{R}^{n \times n} \text{ 非奇异}, \quad b \in \mathbb{R}^n. \quad (7.34)$$

如果在方程组的两边同时左乘一个非奇异矩阵  $P \in \mathbb{R}^{n \times n}$  的逆, 则可得

$$P^{-1}Ax = P^{-1}b. \quad (7.35)$$



这个新方程组就是预处理后的方程组,  $P$  就称为**预处理子** (preconditioner). 当 Krylov 子空间方法被用于求解 (7.35) 时, 就称为预处理 Krylov 子空间方法.

理论上讲, 任何一个非奇异矩阵都可以作为预处理子. 但一个好的预处理子  $P$  通常需满足下面两个要求:

- (1) 预处理后的线性方程组更容易求解;
  - (2) 预处理子  $P$  的构造和使用所增加的额外计算成本较低.
- (1)  $P^{-1}A$  具有更好的特征值分布及/或更小的条件数;
  - (2)  $Pz = r$  容易求解.

其中第一个要求是为了确保预处理后的线性方程组更容易求解, 也就是说, 选取的预处理方法是有效的, 通常要求  $P^{-1}A$  具有更小的条件数和/或更好的特征值分布. 对于 Krylov 子空间方法而言, 第二个要求主要是指以  $P$  为系数矩阵的线性方程组很容易求解.

预处理方式 (7.35) 称为**左预处理**. 相应地, 我们可以在方程组两边同时右乘  $P^{-1}$ , 这就是**右预处理**, 即

$$AP^{-1}u = b, \quad x = P^{-1}u. \quad (7.36)$$

另外, 我们也可以将  $P$  分解成两个矩阵的乘积, 即  $P = LR$ . 于是我们可以用下面的方式对原方程组 (7.34) 进行预处理

$$L^{-1}AR^{-1}u = b, \quad x = R^{-1}u. \quad (7.37)$$

这就是**两边预处理**.

以上是三种常用的预处理方式. 这三种方式预处理后的系数矩阵分别为  $P^{-1}A$ ,  $AP^{-1}$  和  $L^{-1}AR^{-1}$ . 由于它们是相似的, 所以具有相同的特征值分布. 如果  $A$  是对称正定的, 则使用共轭梯度法求解时, 这三种方式的预处理效果基本上是一样的. 但对于非对称 (特别是非正规) 情形, 效果可能会相差很大.

在实际使用中, 该选取哪种预处理方式, 需要根据问题本身和所用的方法来确定. 如对于对称正定线性方程组的 CG 方法, 三种方式都可以, 而对于 GMRES 方法, 则选取右预处理比较合适. 一方面是实际使用时, 得到的残量范数与原方程组的残量范数是一样的, 另一方面是, 右预处理极小化的是原始残量范数, 而左预处理极小化的是预处理后的残量.

这里需要指出的是, 在实际求解预处理后的方程组时, 我们并不会显式地计算  $P^{-1}$  (除非  $P^{-1}$  非常容易计算), 更不会显式地计算  $P^{-1}A$ .

### 7.5.2 预处理 CG 方法

设  $A \in \mathbb{R}^{n \times n}$  对称正定, 并假定预处理子  $P$  也是对称正定的. 为了保证预处理后的系数矩阵仍然是对称正定的, 我们考虑使用两边预处理方式. 设  $P$  的 Cholesky 分解为

$$P = LL^T.$$

于是我们得到下面的预处理方程组

$$L^{-1}A(L^T)^{-1}u = L^{-1}b, \quad x = (L^T)^{-1}u. \quad (7.38)$$



用 CG 方法求解上述方程组, 迭代  $k$  步后, 得到的近似解记为  $u^{(k)}$ , 预处理残量记为  $\tilde{r}_k \triangleq L^{-1}b - L^{-1}A(L^T)^{-1}u^{(k)}$ . 于是, 求解预处理方程组 (7.38) 的 CG 方法可描述如下:

#### 算法 7.8. 两边预处理 CG 方法

```

1: 给定初值 $x^{(0)}$
2: 计算 $r_0 = b - Ax^{(0)}$
3: 令 $\tilde{r}_0 = L^{-1}r_0, \tilde{p}_1 = \tilde{r}_0$
4: for $k = 1, 2, \dots$ do
5: $\xi_k = \frac{(\tilde{r}_{k-1}, \tilde{r}_{k-1})}{(L^{-1}A(L^T)^{-1}\tilde{p}_k, \tilde{p}_k)}$
6: $u^{(k)} = u^{(k-1)} + \xi_k \tilde{p}_k$
7: $\tilde{r}_k = \tilde{r}_{k-1} - \xi_k L^{-1}A(L^T)^{-1}\tilde{p}_k$
8: $\mu_k = \frac{(\tilde{r}_k, \tilde{r}_k)}{(\tilde{r}_{k-1}, \tilde{r}_{k-1})}$
9: $\tilde{p}_{k+1} = \tilde{r}_k + \mu_k \tilde{p}_k$
10: end for

```

由于算法 7.8 中得到的是预处理后的近似解  $u^{(k)}$  和预处理残量  $\tilde{r}_k$ , 因此我们还需要考虑原方程的近似解和残量的计算.

我们对上述算法中的迭代公式进行适当的改写. 首先引入一个辅助变量  $p_k$ :

$$p_k \triangleq (L^T)^{-1}\tilde{p}_k, \quad k = 1, 2, \dots$$

于是有

$$p_{k+1} = (L^T)^{-1}p_{k+1} = (L^T)^{-1}\tilde{r}_k + \mu_k(L^T)^{-1}\tilde{p}_k = (L^T)^{-1}\tilde{r}_k + \mu_k p_k.$$

又由  $\tilde{r}_k$  的表达式可知, 原方程组的残量  $r_k = L\tilde{r}_k$ . 因此可得到下面的递推公式:

$$\begin{aligned}
r_k &= L\tilde{r}_k = L(\tilde{r}_{k-1} - \xi_k L^{-1}A(L^T)^{-1}\tilde{p}_k) = r_{k-1} - \xi_k A p_k, \\
p_{k+1} &= (L^T)^{-1}\tilde{r}_k + \mu_k p_k = (L^T)^{-1}L^{-1}r_k + \mu_k p_k = P^{-1}r_k + \mu_k p_k, \\
x^{(k)} &= (L^T)^{-1}u^{(k)} = (L^T)^{-1}(u^{(k-1)} + \xi_k \tilde{p}_k) = x^{(k-1)} + \xi_k p_k,
\end{aligned}$$

其中

$$\begin{aligned}
\xi_k &= \frac{(L^{-1}r_{k-1}, L^{-1}r_{k-1})}{(L^{-1}A(L^T)^{-1}\tilde{p}_k, \tilde{p}_k)} = \frac{(r_{k-1}, (L^T)^{-1}L^{-1}r_{k-1})}{(A(L^T)^{-1}\tilde{p}_k, (L^T)^{-1}\tilde{p}_k)} = \frac{(r_{k-1}, P^{-1}r_{k-1})}{(Ap_k, p_k)}, \\
\mu_k &= \frac{(L^{-1}r_k, L^{-1}r_k)}{(L^{-1}r_{k-1}, L^{-1}r_{k-1})} = \frac{(r_k, P^{-1}r_k)}{(r_{k-1}, P^{-1}r_{k-1})}.
\end{aligned}$$

记  $z_k \triangleq P^{-1}r_k$ , 则可得下面的预处理 CG 方法.

#### 算法 7.9. PCG: 预处理 CG 方法

```

1: 给定初值 $x^{(0)}$, (相对) 精度要求 $\varepsilon > 0$ 和最大迭代步数 IterMax
2: 计算 $r_0 = b - Ax^{(0)}$ 和 $\beta = \|r_0\|_2$

```



```

3: 令 $z_0 = P^{-1}r_0, p_1 = z_0$
4: 计算 $\rho = r_0^\top z_0$
5: for $k = 1$ to IterMax do
6: $q_k = Ap_k$
7: $\xi_k = \rho / (p_k^\top q_k)$
8: $x^{(k)} = x^{(k-1)} + \xi_k p_k$
9: $r_k = r_{k-1} - \xi_k q_k$
10: $\text{relres} = \|r_k\|_2 / \beta$
11: if $\text{relres} < \varepsilon$ then
12: break
13: end if
14: $\rho_0 = \rho$
15: $z_k = P^{-1}r_k$
16: $\rho = r_k^\top z_k$
17: $\mu_k = \rho / \rho_0$
18: $p_{k+1} = z_k + \mu_k p_k$
19: end for
20: if $\text{relres} < \varepsilon$ then
21: 输出近似解 $x^{(k)}$ 及相关信息
22: else
23: 输出算法失败信息
24: end if

```

我们注意到, 矩阵  $L$  并没有出现在算法中, 这意味着我们并不需要对  $P$  做 Cholesky 分解.

在算法 7.9 中, 每步迭代都需要计算  $z_k = P^{-1}r_k$ . 一般情况下, 都是通过求解方程组  $Pz_k = r_k$  来实现, 除非  $P^{-1}$  非常容易得到, 或者是直接给出的.

事实上, PCG 方法 7.9 也可以从左预处理方式导出. 考虑左预处理方程组

$$P^{-1}Ax = P^{-1}b. \quad (7.39)$$

易知  $P^{-1}A$  是正定的, 但通常不是对称的, 因此我们不能直接对 (7.39) 实施 CG 方法. 此时我们需要定义一个新的内积:  $P$ -内积, 即

$$(x, y)_P \triangleq (Px, y) = (x, Py). \quad (7.40)$$

由于  $P$  是对称正定的, 所以这个定义是有效的. 在该内积下, 有

$$(P^{-1}Ax, y)_P = (Ax, y) = (x, Ay) = (x, P(P^{-1}Ay)) = (x, P^{-1}Ay)_P,$$

即  $P^{-1}A$  关于  $P$ -内积是自伴随的. 也就是说, 在  $P$ -内积意义下,  $P^{-1}A$  是“对称”的. 这时我们就可以用 CG 方法来求解预处理方程组 (7.39), 但需要将欧拉内积改为  $P$ -内积.

引入辅助变量  $z_k \triangleq P^{-1}r_k$ , 则求解预处理方程组 (7.39) 的 CG 方法可描述如下:





**算法 7.10.** 基于  $P$ -内积的 CG 方法

```

1: 给定初值 $x^{(0)}$, (相对) 精度要求 $\varepsilon > 0$ 和最大迭代步数 IterMax
2: 计算 $r_0 = b - Ax^{(0)}$ 和 $\beta = \|r_0\|_2$
3: 令 $z_0 = P^{-1}r_0, p_1 = z_0$
4: for $k = 1$ to IterMax do
5: $\xi_k = \frac{(z_{k-1}, z_{k-1})_P}{(P^{-1}Ap_k, p_k)_P} = \frac{(r_{k-1}, z_{k-1})}{(Ap_k, p_k)}$
6: $x^{(k)} = x^{(k-1)} + \xi_k p_k$
7: $r_k = r_{k-1} - \xi_k Ap_k$
8: relres = $\|r_k\|_2 / \beta$
9: if relres $< \varepsilon$ then
10: break
11: end if
12: $z_k = z_{k-1} - \alpha_k P^{-1}Ap_k = P^{-1}r_k$
13: $\mu_k = \frac{(z_k, z_k)_P}{(z_{k-1}, z_{k-1})_P} = \frac{(r_k, z_k)}{(r_{k-1}, z_{k-1})}$
14: $p_{k+1} = z_k + \mu_k p_k$
15: end for

```

不难看出, 算法 7.10 和算法 7.9 是完全一样的.

类似地, 我们也可以从右预处理方式来推导 PCG 方法, 具体过程留作练习.

**7.5.3 预处理 GMRES 方法**

对于非对称 Krylov 子空间方法, 也有三种预处理方式. 但与对称正定情形不同的是, 这三种方式并不等价, 而且有时效果会相差很大.

如果采用左预处理方式, 则原方程组转化为

$$P^{-1}Ax = P^{-1}b. \quad (7.41)$$

记  $r_k$  为原线性方程组的残量, 即  $r_k = b - Ax^{(k)}$ , 则预处理方程组 (7.41) 的残量为  $\tilde{r}_k = P^{-1}r_k$ . 因此在对应的算法中, 停机准则是针对  $\tilde{r}_k$ , 而不是真实的残量  $r_k$ . 所以我们通常采用右预处理方式.

设  $P$  是预处理子, 则右预处理后的方程组为

$$AP^{-1}u = b, \quad x = P^{-1}u. \quad (7.42)$$

给定迭代初值  $x^{(0)}$ , 则  $u^{(0)} = Px^{(0)}$ . 相应的预处理初始残量为

$$\tilde{r}_0 = b - AP^{-1}u^{(0)} = b - Ax^{(0)} = r^{(0)},$$

即预处理残量与原方程组残量是一样的. 因此无需计算  $u^{(0)}$ . 设  $u^{(m)} = u^{(0)} + V_m y_m$  是迭代  $m$  步





后的近似解, 则对应的原方程组近似解  $x^{(m)}$  为

$$x^{(m)} = P^{-1}u^{(m)} = P^{-1}(u^{(0)} + V_m y_m) = x^{(0)} + P^{-1}V_m y_m,$$

原方程组的残量为

$$r_m = b - Ax^{(m)} = r_0 - AP^{-1}V_m y_m = \beta v_1 - V_{m+1}H_{m+1,m}y_m = V_{m+1}(\beta e_1 - H_{m+1,m}y_m).$$

由于  $y_m$  是最小二乘问题

$$\min_{y \in \mathbb{R}^m} \|\beta e_1 - H_{m+1,m}y\|_2$$

的解, 因此

$$\|r_m\|_2 = \|\beta e_1 - H_{m+1,m}y_m\|_2 = \beta \cdot |q_1(m+1)|.$$

这与不带预处理的 GMRES 方法是一样的. 因此在实际求解过程中我们可以直接得到原方程组残量的模, 而无需计算  $u^{(m)}$  和  $x^{(m)}$ . 这是右预处理方式与左预处理方式的主要区别.

右预处理 GMRES 方法具体描述如下:

#### 算法 7.11. 右预处理 GMRES 方法

```

1: 给定初值 $x^{(0)}$ 和 (相对) 精度要求 $\varepsilon > 0$
2: 计算 $r_0 = b - Ax^{(0)}$ 和 $\beta = \|r_0\|_2$
3: 令 $v_1 = \tilde{r}_0/\beta, \xi = \beta e_1$
4: for $j = 1, 2, \dots$ do
5: $\tilde{w}_j = P^{-1}v_j$ % Apply preconditioning
6: $w_j = A\tilde{w}_j$
7: for $i = 1, 2, \dots, j$ do
8: $h_{ij} = (w_j, v_i)$
9: $w_j = w_j - h_{ij}v_i$
10: end for
11: $h_{j+1,j} = \|w_j\|_2$
12: for $i = 1, 2, \dots, j-1$ do % Apply G_{j-1}, \dots, G_1 to the last column of $H_{j+1,j}$
13:
$$\begin{bmatrix} h_{i,j} \\ h_{i+1,j} \end{bmatrix} = \begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix} \begin{bmatrix} h_{i,j} \\ h_{i+1,j} \end{bmatrix}$$

14: end for
15: if $h_{j+1,j} = 0$ then
16: set $m = j$ and break
17: end if
18: $v_{j+1} = w_j/h_{j+1,j}$
19: if $|h_{j,j}| > |h_{j+1,j}|$ then % Form the Givens rotation G_j
20: $c_j = \frac{1}{\sqrt{1+\tau^2}}, s_j = c_j\tau$ where $\tau = \frac{h_{j+1,j}}{h_{j,j}}$
21: else
```



```

22: $s_j = \frac{1}{\sqrt{1+\tau^2}}, c_j = s_j \tau$ where $\tau = \frac{h_{j,j}}{h_{j+1,j}}$
23: end if
24: $h_{j,j} = c_j h_{j,j} + s_j h_{j+1,j}, h_{j+1,j} = 0$ % Apply G_j to last column of $H_{j+1,j}$
25: $\begin{bmatrix} \xi_j \\ \xi_{j+1} \end{bmatrix} = \begin{bmatrix} c_j & s_j \\ -s_j & c_j \end{bmatrix} \begin{bmatrix} \xi_j \\ 0 \end{bmatrix}$ % Apply G_j to the right-hand side
26: if $|\xi_{j+1}|/\beta < \varepsilon$ then % Check convergence: $\|r_j\|_2 = |\xi_{j+1}|$
27: set $m = j$ and break
28: end if
29: end for
30: 计算 $y^{(m)} = R_m^{-1} \xi^{(m)}$, 其中 $R_m = H(1:m, 1:m), \xi^{(m)} = \xi(1:m)$
31: 计算近似解 $x^{(m)} = x^{(0)} + P^{-1} V_m y^{(m)}$

```

关于左、右预处理 GMRES, 我们有下面的最优性质.

**定理 7.16** 设  $x_L^{(k)}$  和  $x_R^{(k)}$  分别是左预处理 GMRES 方法和右预处理 GMRES 方法迭代  $k$  步后得到的近似解, 且迭代初始值均为  $x^{(0)}$ . 则  $x_L^{(k)}$  是  $\|P^{-1}(b - Ax)\|_2$  在  $x^{(0)} + \mathcal{K}_k(P^{-1}A, P^{-1}r_0)$  的极小值点, 而  $x_R^{(k)}$  则是  $\|b - Ax\|_2$  在同一个子空间中的极小值点.

#### 7.5.4 预处理器构造

This process of “preconditioning” is essential to most successful applications of iterative methods.

— Trefethen & Bau III [119, page 313], 1997.

Finding a good preconditioner to solve a given sparse linear system is often viewed as a combination of art and science. Theoretical results are rare and some methods work surprisingly well, often despite expectations.


— Saad [109], 2003.

预处理能否取得成功的关键就是能否找到一个好的预处理器. 一个公认的事实是, 好的预处理器通常是与问题本身的特征密切相关的.

🔴 既要有很好的通用性, 又要有很好的加速效果的预处理器往往是可遇而不可求的.

关于预处理技术的理论分析很少, 大多数情况下只能根据经验来构造. 尽管如此, 在实际应用中, 这些根据经验构造出来的预处理器往往能取得很好的数值效果, 有时甚至会大大出乎人们的意料.



 预处理方法主要是用来提升 Krylov 子空间迭代法的收敛速度, 因此, 对 Krylov 子空间迭代法的收敛性质的深刻理解对构造好的预处理方法有很大的帮助.

A preconditioner  $M$  is good if  $M^{-1}A$  is not too far from normal and its eigenvalues are clustered.

— Trefethen & Bau III [119, page 314], 1997.

一般来说, 预处理子可以分为两大类

(a) 代数预处理子 (Algebraic Preconditioner), 即仅仅根据所给的矩阵构造出来的预处理子.

(b) 专用预处理子 (Problem-Specific Preconditioner), 即根据问题的物理背景所构造的预处理子.

显然, 由于专用预处理子充分利用了问题的物理背景知识, 所以它们往往具有很好的数值表现, 如多重网格, 区域分解, 快速变换等等. 但它们严重依赖于原问题的物理背景, 因此通用性较差.

我们这里只考虑代数预处理子, 即仅仅根据所给的系数矩阵来构造预处理方法. 这种预处理方法具有较好的通用性. 这类预处理子的一般来源有:

- 定常迭代法: 设有矩阵分裂  $A = M - N$ , 则  $M$  可作为一个预处理子, 如 Jacobi, G-S, SOR, HSS 等;
- 直接法: 如不完全 LU 分解, 不完全 Cholesky 分解等;
- 近似逆: 即选取矩阵  $P$ , 使得  $P^{-1} \approx A^{-1}$ ;
- 块状结构: 如基于 Schur 补的块对角矩阵, 块三角矩阵等.

### 基于矩阵分裂的预处理子

考虑线性方程组

$$Ax = b, \quad A \in \mathbb{R}^{n \times n},$$

对  $A$  做如下的矩阵分裂:

$$A = M - N \tag{7.43}$$

其中  $M$  非奇异, 则可以得到下面的迭代方法

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b = x^{(k)} + \left(M^{-1}b - M^{-1}Ax^{(k)}\right), \quad k = 0, 1, 2, \dots$$

这等价于求解下面的方程组

$$M^{-1}Ax = M^{-1}b. \tag{7.44}$$

这就是与矩阵分裂 (7.43) 相对应的 **预处理线性方程组**. 将 Krylov 子空间方法用于求解方程组 (7.44), 就得到预处理 Krylov 子空间方法. 矩阵  $M$  就是由矩阵分裂 (7.43) 所定义的预处理子.

理论上讲, 任何一个矩阵分裂都可以定义一个预处理子. 但为了使得预处理子能有很好的预处理效果, 往往需要其在一定意义下与  $A$  充分接近.

设  $A = D - L - U$ , 其中  $D$ ,  $-L$ ,  $-U$  分别是  $A$  的对角部分, 严格下三角部分和严格上三角部分, 并假定  $D$  非奇异. 则由我们之前讨论的定常迭代法, 可以立即得到下面的预处理子:



- Jacobi 预处理子, 即取  $A$  的对角部分作为预处理子:

$$P_J = D.$$

- G-S 预处理子, 即取  $A$  的下三角部分作为预处理子:

$$P_{GS} = D - L.$$

- SOR 预处理子, 即

$$P_{SOR} = \frac{1}{\omega} (\omega D - L).$$

- SSOR 预处理子, 即

$$P_{SSOR} = \frac{1}{\omega(2-\omega)} [(1-\omega)D + \omega L] D^{-1} [(1-\omega)D + \omega U].$$

由于 SSOR 对参数  $\omega$  的取值不是很敏感, 因此我们通常令  $\omega = 1$ , 对应的预处理子就是对称 Gauss-Seidel (SGS) 预处理子: 即

$$P_{SGS} = (D - L)D^{-1}(D - U).$$

- HSS 预处理子, 即

$$P_{HSS} = \frac{1}{2\alpha} (\alpha I + H)(\alpha I + S),$$

其中  $H = \frac{1}{2}(A + A^T)$ ,  $S = \frac{1}{2}(A - A^T)$ .

## 不完全 LU 分解

对于大规模稀疏线性方程组, 有一类比较常用的预处理方法是**不完全分解**. 不完全分解首先由 Buleev [18, 19] 于上个世纪五十年代提出, 当时主要用来求解五点差分方程 (2D) 和七点差分方程 (3D) [74], 取得了非常好的效果, 但没有任何理论结果. 之后有更多学者对不完全分解进行了研究, 如 Varga (1960) [131], Oliphant (1962) [93], 等等 (参见 [74]). 但真正的突破是 1977 年 Meijerink 和 van der Vorst [90] 将不完全 Cholesky 分解用作共轭梯度法的预处理子, 提出了不完全 Cholesky 分解共轭梯度法 (ICCG). 它的出现, 才使得共轭梯度法成为求解大规模对称正定线性方程组的首选方法. 这时离共轭梯度法的提出已经过去了整整 25 年. 同时, 这也标志着 Krylov 子空间方法逐渐替代定常迭代方法, 成为求解大规模稀疏线性方程组的主流方法. 从那以后, 关于不完全矩阵分解的研究工作越来越多, 直至今天, 仍然是科学与工程计算领域的一个核心研究课题.

不完全 LU 分解的基本思想就是在对系数矩阵  $A$  做 LU 分解时, 要求  $L$  和  $U$  满足一定的稀疏性, 比如与  $A$  具有相同的稀疏模式, 即  $A$  在  $(i, j)$  ( $i \neq j$ ) 位置上的元素为 0, 则  $L(i, j)$  和  $U(i, j)$  也必须为 0. 于是得到的不完全 LU 分解为

$$A = LU + R,$$

其中  $R$  是误差矩阵. 按这种方式定义的不完全 LU 分解记为  $ILU(0)$ .

为了增加  $LU$  与  $A$  的近似度, 可以适当放宽对  $L$  和  $U$  的稀疏性要求, 即允许  $L$  和  $U$  中出现额外的非零元, 这些非零元称为**填充 (fill-in)**. 但为了控制预处理子  $LU$  的使用成本, 需要控制填充的数量. 于是学者们提出了各种限制填充出现的规则, 比如设定填充的大小或填充的个数, 即只有大于给定阈值的填充才保留, 或者每行填充的个数不能超过一个给定的阈值, 等等. 更多关于这方面的内容可以参见 [65, 109, 127].



## 7.6 课后习题

练习 7.1 证明引理 7.1, 即

$$\mathcal{K}_m = \text{span}\{v_1, v_2, \dots, v_m\}.$$

(提示: 用归纳法证明  $v_k \in \mathcal{K}_k$ ,  $k = 1, 2, \dots, m$  即可)

练习 7.2 证明定理 7.3, 即由 Lanczos 过程生成的向量是相互正交的.

练习 7.3 证明定理 7.5 中的必要性.

练习 7.4 设  $H_{m+1,m}$  不可约, 试证明 (7.8) 中的  $R_m$  非奇异.

练习 7.5 设  $A \in \mathbb{R}^{n \times n}$  对称正定, 非零向量  $p_1, p_2, \dots, p_m$  是  $A$ -共轭的, 即当  $i \neq j$  时有  $p_i^T A p_j = 0$ .  
试证明:  $p_1, p_2, \dots, p_m$  线性无关.

练习 7.6 用右预处理方式推导 PCG 算法.

练习 7.7 设  $A \in \mathbb{R}^{n \times n}$  可对角化且特征值都是实数, 试证明: 存在某内积  $(\cdot, \cdot)$ , 使得在此内积下  $A$  是自伴随的.

..... 以下为实践题 .....

练习 7.8 编程实现 GMRES 方法.

练习 7.9 编程实现 CG 算法.



## 第八讲 特征值问题的迭代解法

当矩阵规模很大时, 计算其所有的特征值和特征向量是非常困难的. 而在实际应用中, 我们通常也只对其中的某些特征值和 (或) 特征向量感兴趣, 因此也没有必要计算所有的特征值和特征向量.

本讲主要介绍计算部分特征值和 (或) 特征值向量的迭代解法. 这些算法的存储量一般要远小于  $\mathcal{O}(n^2)$ , 运算量也远小于  $\mathcal{O}(n^3)$ .

目前主要的几类方法:

- 梯度下降法 (Gradient-type methods): 如 SD, CG and LOPCG. 主要思想是将特征值问题看作一个非线性问题, 然后用梯度下降法来求解.
- 子空间投影法 (Krylov subspace methods): 如: Lanczos, Arnoldi and its variants.
- Davidson-type methods: Davidson and Jacobi-Davidson

关于大规模矩阵特征值问题的参考资料有:

- B.N. Parlett, The Symmetric Eigenvalue Problem, Prentice Hall, Englewood Cliffs, NJ, 1980. (Republished by SIAM, Philadelphia, 1998.)
- Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide, SIAM, Philadelphia, 2000.
- Y. Saad, Numerical Methods for Large Eigenvalue Problems, 2nd revised edition. SIAM, Philadelphia, 2011.
- G. H. Golub and Ch. van Loan, Matrix Computations, 4th edition, Johns Hopkins University Press, Baltimore, 2013.
- Z. Bai, R. C. Li, and Y.F. Su, Lecture notes on Matrix Eigenvalue Computations, 2009.
- 课程: Prof. Bai, Large Scale Scientific Computing, 2013.  
<http://www.cs.ucdavis.edu/~bai/ECS231/>
- 课程: Prof. Dr. Peter Arbenz, Numerical Methods for Solving Large Scale Eigenvalue Problems, 2014.  
<http://people.inf.ethz.ch/arbenz/ewp/>
- 软件: ARPACK – collection of Fortran subroutines designed to compute a few eigenvalues and eigenvectors of large scale sparse matrices and pencils  
<http://www.caam.rice.edu/software/ARPACK/>

### 8.1 投影算法

最简单的特征值问题就是仅仅计算一个特征值, 如计算模最大的特征值. 这时我们可以使用幂迭代方法, 这在前面章节已经介绍. 为了后面讨论方便, 我们仍将该算法描述出来.

**算法 8.1.** 幂迭代: 计算最大特征值

```

1: Given $x^{(0)}$
2: for $i = 1, 2, \dots$, until converge do
3: $y^{(i)} = Ax^{(i-1)}$
4: $x^{(i)} = y^{(i)} / \|y^{(i)}\|_2$
5: end for

```

幂迭代所产生的迭代向量  $x^{(0)}, x^{(1)}, \dots, x^{(m-1)}$  生成一个 Krylov 子空间

$$\mathcal{K}_m(A, x^{(0)}) = \text{span} \{x^{(0)}, Ax^{(0)}, \dots, A^{m-1}x^{(0)}\} = \text{span} \{x^{(0)}, x^{(1)}, \dots, x^{(m-1)}\}.$$

在幂迭代中, 我们取  $x^{(m-1)}$  为近似特征向量. 显然, 如果我们在  $\mathcal{K}_m(A, x^{(0)})$  中找出“最佳”的近似特征向量, 则收敛速度就可能会大大加快.

下面我们讨论如何在  $\mathcal{K}_m = \mathcal{K}_m(A, x^{(0)})$  中寻找“最佳”的近似特征向量. 设  $A \in \mathbb{R}^{n \times n}$ , 并设  $\mathcal{K}_m$  和  $\mathcal{L}_m$  是  $\mathbb{R}^n$  的两个  $m$  维子空间. 投影算法就是寻找  $A$  的近似特征对  $(\tilde{\lambda}, \tilde{x})$ , 满足下面的 **Petrov-Galerkin 条件**

$$\text{find } \tilde{\lambda} \in \mathbb{C} \text{ and } \tilde{x} \in \mathcal{K}_m \text{ such that } A\tilde{x} - \tilde{\lambda}\tilde{x} \perp \mathcal{L}_m. \quad (8.1)$$

这样的算法我们称为**斜投影算法**. 如果我们取  $\mathcal{L}_m = \mathcal{K}_m$ , 则上面的算法就是一个**正交投影算法**, 此时条件 (8.1) 称为 **Galerkin 条件**.

设  $v_0, v_1, \dots, v_{m-1}$  和  $w_0, w_1, \dots, w_{m-1}$  分别是  $\mathcal{K}_m$  和  $\mathcal{L}_m$  的一组标准正交基. 令  $V_m = [v_0, v_1, \dots, v_{m-1}]$ ,  $W_m = [w_0, w_1, \dots, w_{m-1}]$ . 则对任意  $\tilde{x} \in \mathcal{K}_m$ , 存在向量  $y \in \mathbb{R}^m$  使得  $\tilde{x} = V_m y$ , 即  $\tilde{x}$  可以由  $v_0, v_1, \dots, v_{m-1}$  线性表出. 根据条件 (8.1), 我们可得

$$(AV_m y - \tilde{\lambda} V_m y, w_i) = 0, \quad i = 1, 2, \dots, m,$$

即

$$W_m^T AV_m y = \tilde{\lambda} W_m^T V_m y. \quad (8.2)$$

这是一个广义特征值问题. 如果我们取  $\mathcal{L}_m = \mathcal{K}_m$ , 并令  $W_m = V_m$ , 则 (8.2) 就转化为

$$T_m y = \tilde{\lambda} y, \quad (8.3)$$

其中  $T_m = V_m^T AV_m \in \mathbb{R}^{m \times m}$ . 这意味着  $(\tilde{\lambda}, y)$  是矩阵  $T_m$  的一个特征对. 由于  $m$  通常比较小, 因此我们可以使用先前讨论的方法 (如 QR 迭代) 来计算  $(\tilde{\lambda}, y)$ . 这样我们就可以计算出  $A$  的一个近似特征对  $(\tilde{\lambda}, \tilde{x})$ , 其中  $\tilde{x} = V_m y$ .

## 8.2 Rayleigh-Ritz 方法

事实上, 我们可以在  $\mathcal{K}_m(A, x^{(0)})$  中找出  $m$  个最佳近似特征向量及相应的最佳近似特征值. 这些近似特征值和近似特征向量就称为 **Ritz 值** 和 **Ritz 向量**.





**定义 8.1** 设  $\mathcal{K}_m$  是  $\mathbb{R}^{n \times n}$  的一个  $m$  维子空间, 它的一组标准正交基为  $v_0, v_1, \dots, v_{m-1}$ , 并令  $V_m = [v_0, v_1, \dots, v_{m-1}]$ . 记  $T_m = V_m^T A V_m$ , 设  $(\tilde{\lambda}, y)$  是  $T_m$  的一组特征对, 即  $T_m y = \tilde{\lambda} y$  且  $\|y\|_2 = 1$ . 则称  $\tilde{\lambda}$  是  $A$  的一个 **Ritz 值**,  $\tilde{x} = V_m y$  是  $A$  的一个 **Ritz 向量**.

**Rayleigh-Ritz 方法** 就是用 Ritz 值和 Ritz 向量来近似  $A$  的特征值与特征向量.

### 算法 8.2. Rayleigh Ritz 算法

- 1: 计算  $\mathcal{K}_m$  的一组单位正交基:  $V_m = [v_0, v_1, \dots, v_{m-1}]$ .
- 2: 计算矩阵  $T_m = V_m^T A V_m$  的特征值 (即  $A$  的 Ritz 值)
- 3: 选取其中的  $k$  个想要的特征值:  $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_k$ , 其中  $k \leq m$ .
- 4: 计算对应的特征向量:  $y_1, y_2, \dots, y_k$ , 即  $T_m y_i = \tilde{\lambda}_i y_i$
- 5: 计算  $A$  的 Ritz 向量:  $\tilde{x}_i = V_m y_i, i = 1, 2, \dots, k$ .

下面我们讨论  $A$  是对称矩阵时, Rayleigh Ritz 方法的性质.

设  $V_u \in \mathbb{R}^{n \times n-m}$  是一个列正交矩阵, 且使得  $V = [V_m, V_u] \in \mathbb{R}^{n \times n}$  是一个正交矩阵. 于是我们有

$$V^T A V = [V_m, V_u]^T A [V_m, V_u] = \begin{bmatrix} V_m^T A V_m & V_m^T A V_u \\ V_u^T A V_m & V_u^T A V_u \end{bmatrix}.$$

由于  $A$  对称的, 故  $T_m = V_m^T A V_m \in \mathbb{R}^{m \times m}$  也是对称的. 此时, Ritz 值和 Ritz 向量具有下面的最优性质.

**定理 8.1** 设  $A \in \mathbb{R}^{n \times n}$  对称, 则对任意的对称矩阵  $R \in \mathbb{R}^{m \times m}$ , 有

$$\|AV_m - V_m R\|_2 \geq \|AV_m - V_m T_m\|_2,$$

即  $\|AV_m - V_m R\|_2$  在  $R = T_m$  处取最小值, 此时  $\|AV_m - V_m R\|_2 = \|V_u^T A V_m\|_2$ .

**证明.** 令  $R = T_m + Z$ , 其中  $Z \in \mathbb{R}^{m \times m}$  是任意一个对称矩阵. 由于  $A$  和  $T_m$  都是对称矩阵, 且  $T_m = V_m^T A V_m$ , 因此由

$$\begin{aligned} \|AV_m - V_m R\|_2^2 &= (AV_m - V_m R)^T (AV_m - V_m R) \\ &= (AV_m - V_m(T_m + Z))^T (AV_m - V_m(T_m + Z)) \\ &= \|AV_m - V_m T_m\|_2^2 - V_m^T A V_m Z + T_m V_m^T V_m Z \\ &\quad - Z V_m^T A V_m + Z V_m^T V_m T_m + Z^T V_m^T V_m Z \\ &= \|AV_m - V_m T_m\|_2^2 + \|Z\|_2^2. \end{aligned}$$

所以当  $Z = 0$  时  $\|AV_m - V_m R\|_2^2$  达到最小, 且

$$\begin{aligned} \|AV_m - V_m T_m\|_2 &= \|V V^T A V_m - V_m T_m\|_2 \\ &= \left\| (V_m, V_u) \begin{bmatrix} V_m^T A V_m \\ V_u^T A V_m \end{bmatrix} - V_m T_m \right\|_2 \end{aligned}$$





$$\begin{aligned}
 &= \|V_u(V_u^T AV_m)\|_2 \\
 &= \|V_u^T AV_m\|_2.
 \end{aligned}$$

□

定理 8.1 中的 2-范数可以改成任意的酉不变范数, 如  $F$ -范数.

**定理 8.2** 设  $A \in \mathbb{R}^{n \times n}$  对称, 并设  $T_m = U\Lambda U^T$  是  $T_m = V_m^T AV_m$  的特征值分解. 设  $Q \in \mathbb{R}^{n \times m}$  是满足  $\text{span}(Q) = \mathcal{K}$  的任意单位列正交矩阵,  $D \in \mathbb{R}^{m \times m}$  是任意对角矩阵. 我们有

$$\|AQ - QD\|_2 \geq \|AV_m - V_m T_m\|_2,$$

且当  $Q = V_m U$ ,  $D = \Lambda$  时等式成立.

**证明.** 由于  $\text{span}(Q) = \mathcal{K} = \text{span}(V_m)$ , 所以存在矩阵  $W \in \mathbb{R}^{m \times m}$  使得  $Q = V_m W$ . 又  $Q$  是单位列正交的, 因此

$$I = Q^T Q = (V_m W)^T V_m W = W^T V_m^T V_m W = W^T W.$$

这表明  $W$  是正交矩阵. 设  $WDW^T = T_m + Z$ , 则

$$\begin{aligned}
 \|AQ - QD\|_2^2 &= \|AV_m W - V_m W D\|_2^2 \\
 &= \|AV_m - V_m W D W^T\|_2^2 \\
 &= \|AV_m - V_m T_m\|_2^2 + \|Z\|_2^2 \\
 &\geq \|AV_m - V_m T_m\|_2^2.
 \end{aligned}$$

如果取  $W = U$  和  $D = \Lambda$ , 则  $Z = W D W^T - T_m = U \Lambda U^T - T_m = 0$ . 此时上式中的等号成立. □

定理 8.2 表明, 在所有满足  $\text{span}(Q) = \mathcal{K}$  单位列正交矩阵  $Q \in \mathbb{R}^{n \times m}$  和任意的对角矩阵  $D \in \mathbb{R}^m$  中, 当  $Q = V_m U$  和  $D = \Lambda$  时,  $\|AQ - QD\|_2$  取到极小值.

定理 8.1 和定理 8.2 表明, 在  $\|AQ - QD\|_2$  极小的意义下, Ritz 值是特征值的“最佳”近似. 所以我们用 Ritz 值作为特征值的近似是有道理的.

### 8.3 Lanczos 方法

设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵. Lanczos 方法就是利用 Lanczos 过程来计算  $\mathcal{K}_m$  的基和  $T_m = V_m^T AV_m$ , 然后计算  $A$  的 Ritz 值和 Ritz 向量.

#### 算法 8.3. Lanczos 算法

- 1: Choose a vector  $v_0$  such that  $\|v_0\| = 1$ , and set  $\beta_0 = 0$
- 2: **for**  $j = 0, 1, \dots$  **do**
- 3:     Compute  $w = Av_j - \beta_j v_{j-1}$
- 4:      $\alpha_{j+1} = (w, v_j)$
- 5:      $w = w - \alpha_{j+1} v_j$



```

6: $\beta_{j+1} = \|w\|_2$
7: if $\beta_{j+1} = 0$ then
8: stop
9: end if
10: $v_{j+1} = w/\beta_{j+1}$
11: Compute the eigenvalues and eigenvectors of T_j
12: Check the convergence
13: end for

```

在 Lanczos 算法 8.3 中, 迭代  $m$  步后, 向量  $v_0, v_1, \dots, v_{m-1}$  构成子空间

$$\mathcal{K}_m(A, v_0) = \text{span} \{v_0, Av_0, \dots, A^{m-1}v_0\},$$

的一组基, 并且有

$$AV_m = V_m T_m + \beta_m v_m e_m^T,$$

其中  $e_m = [0, 0, \dots, 0, 1]^T \in \mathbb{R}^m$ ,

$$T_m = V_m^T A V_m = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{m-1} \\ & & & \beta_{m-1} & \alpha_m \end{bmatrix}.$$

设  $(\tilde{\lambda}, y)$  是  $T_m$  的一个特征对, 则有

$$A(V_m y) = V_m T_m y + \beta_m v_m e_m^T y = \tilde{\lambda} V_m y + \beta_m (e_m^T y) v_m.$$

于是

$$Ax - \tilde{\lambda}x = \beta_m (e_m^T y) v_m,$$

即

$$\|Ax - \tilde{\lambda}x\|_2 = |\beta_m (e_m^T y)|,$$

其中  $x = V_m y$ . 如果  $|\beta_m (e_m^T y)|$  很小, 则我们就认为  $\tilde{\lambda}$  是  $A$  的某个特征值的很好的近似. 事实上, 关于 Ritz 值  $\tilde{\lambda}$ , 我们有下面的性质.

**引理 8.3** 设  $A \in \mathbb{R}^{n \times n}$  是对称矩阵, 设  $r = Ax - \tilde{\lambda}x$ , 其中  $x \neq 0$ . 则

$$\min_{\lambda \in \sigma(A)} |\lambda - \tilde{\lambda}| \leq \frac{\|r\|_2}{\|x\|_2},$$

其中  $\sigma(A)$  表示  $A$  的谱, 即所有特征值组成的集合.

**证明.** 设  $A = U\Lambda U^T$  是矩阵  $A$  的特征值分解. 则

$$r = (A - \tilde{\lambda}I)x = U(\Lambda - \tilde{\lambda}I)U^T x.$$



记  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ . 对任意向量  $z = [z_1, z_2, \dots, z_n]^T \in \mathbb{R}^n$ , 有

$$\begin{aligned}\|(\Lambda - \tilde{\lambda}I)z\|_2^2 &= \sum_{i=1}^n (\lambda_i - \tilde{\lambda})^2 z_i^2 \\ &\geq \sum_{i=1}^n \min_{\lambda \in \sigma(A)} |\lambda - \tilde{\lambda}|^2 z_i^2 \\ &= \|z\|_2^2 \min_{\lambda \in \sigma(A)} |\lambda - \tilde{\lambda}|^2.\end{aligned}$$

所以

$$\|r\|_2 = \|U^T r\|_2 = \|(\Lambda - \tilde{\lambda}I)U^T x\|_2 \geq \|U^T x\|_2 \min_{\lambda \in \sigma(A)} |\lambda - \tilde{\lambda}| = \|x\|_2 \cdot \min_{\lambda \in \sigma(A)} |\lambda - \tilde{\lambda}|.$$

□

由引理 8.3 可知, 存在  $A$  的某个特征值  $\lambda$ , 使得

$$|\lambda - \tilde{\lambda}| \leq \frac{\|\beta_m(e_m^T y)v_m\|_2}{\|V_m y\|_2} = \frac{|\beta_m| \cdot |e_m^T y|}{\|y\|_2} = |\beta_m| \cdot |e_m^T y|. \quad (8.4)$$

在前面的讨论中, 我们并没有考虑实际计算时可能的舍入误差. 在实际计算中, 由于浮点运算的舍入误差, 即使  $m$  很小 (如  $m = 10$  或  $m = 20$ ), 也可能会导致向量  $\{v_i\}$  失去正交性. 这时我们必须采取一些补救措施, 最简单的方法就是对它们重新来一次正交化, 即在算法 8.3 的第 5 步后加上一条语句

$$w = w - \sum_{i=1}^j (w, v_i) v_i.$$

这个过程就称为带全正交过程的 Lanczos 算法. 显然, 这个过程是非常费时的. 另外一个可行的方法就是选择性正交.

## 8.4 Arnoldi 方法

这里考虑非对称情形, 即计算非对称矩阵  $A$  的特征值. 与 Lanczos 算法相类似, 我们可以通过 Arnoldi 过程来计算  $K_m$  的标准正交基  $v_0, v_1, \dots, v_{m-1}$  和上 Hessenberg 矩阵  $H_m = V_m^T A V_m$ , 使得

$$A V_m = V_m H_m + h_{m+1,m} v_m e_m^T \quad \text{and} \quad V_m^T A V_m = H_m.$$

但此时  $H_m$  只是上 Hessenberg, 而不是对称三对角. 但我们同样可以通过计算  $H_m$  的特征值和特征向量来得到  $A$  的 Ritz 值的 Ritz 向量, 并用它们来近似  $A$  的特征值和特征向量.

设  $(\tilde{\lambda}, y)$  是  $H_m$  的一个特征对, 其中  $\|y\|_2 = 1$ , 则

$$A(V_m y) = V_m H_m y + h_{m+1,m} v_m e_m^T y = \tilde{\lambda} V_m y + h_{m+1,m} (e_m^T y) v_m,$$

所以

$$\|Ax - \tilde{\lambda}x\|_2 = \|h_{m+1,m}(e_m^T y)v_m\|_2 = |h_{m+1,m}| \cdot |e_m^T y|,$$

其中  $x = V_m y$ . 若  $|h_{m+1,m}| \cdot |e_m^T y|$  足够小, 我就认为  $(\tilde{\lambda}, x)$  是  $A$  的某个特征对的近似.



**算法 8.4.** Arnoldi 算法

```

1: Choose a vector v_0 such that $\|v_0\|_2 = 1$
2: for $j = 0, 1, \dots$ do
3: Compute $w_{j+1} = Av_j$
4: for $i = 0, 1, \dots, j$ do
5: $h_{ij} = (w_{j+1}, v_i)$
6: $w_{j+1} = w_{j+1} - h_{ij}v_i$
7: end for
8: $h_{j+1,j} = \|w_{j+1}\|_2$
9: if $h_{j+1,j} = 0$ then
10: stop
11: end if
12: $v_{j+1} = w_{j+1}/h_{j+1,j}$
13: Compute the eigenvalues and eigenvectors of T_j
14: Check the convergence
15: end for

```

由于  $A$  是非对称的, 其特征值可能是复的, 或者是坏条件的, 此时 Lanczos 方法的一些最优性质就不再成立. 尽管如此, 目前还是存在一些有效 Arnoldi 方法的实现方式, 可参见 [84, 108, 112].

**8.5 非对称 Lanczos 方法**

非对称 Lanczos 方法就是 Lanczos 方法在非对称矩阵上的推广, 它是基于 Lanczos 双正交化过程.

设  $v_0$  和  $w_0$  是任意的非零向量, 并设

$$\mathcal{K}_m(A, v_0) = \text{span}\{v_0, Av_0, \dots, A^{m-1}v_0\}$$

和

$$\mathcal{K}_m(A^T, w_0) = \text{span}\{w_0, A^T w_0, \dots, (A^T)^{m-1}w_0\}.$$

Lanczos 双正交化过程就是计算  $\mathcal{K}_m(A, v_0)$  和  $\mathcal{K}_m(A, w_0)$  的基  $\{v_i\}$  和  $\{w_i\}$ , 满足  $\{v_i\}$  和  $\{w_i\}$  相互正交, 即

$$(v_i, w_j) = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases}.$$

所以

$$W_m^T V_m = I,$$

其中  $V_m = [v_0, v_1, \dots, v_{m-1}]$ ,  $W_m = [w_0, w_1, \dots, w_{m-1}]$ .

注意, 通常  $\{v_i\}_{i=0}^m$  或  $\{w_j\}_{j=0}^m$  本身并不一定正交, 故  $V_m$  和  $W_m$  通常并不列正交.



根据 Lanczos 双正交化过程, 我们可得

$$\begin{aligned} AV_m &= V_m T_m + \gamma_m v_m e_m^T, \\ A^T W_m &= W_m T_m^T + \beta_m w_m e_m^T, \end{aligned}$$

其中  $e_m = [0, \dots, 0, 1]^T \in \mathbb{R}^m$ ,

$$T_m = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ \gamma_1 & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{m-1} \\ & & \gamma_{m-1} & \alpha_m \end{bmatrix}.$$

所以

$$W_m^T AV_m = W_m^T V_m T_m + \gamma_m (W_m^T v_m) e_m^T = T_m.$$

设  $\tilde{\lambda}$  是  $T_m$  的特征值, 其对应的右特征向量和左特征向量分别为  $y$  和  $z$ , 且  $\|y\|_2 = \|z\|_2 = 1$ , 即

$$T_m y = \tilde{\lambda} y \quad \text{and} \quad z^T T_m = \tilde{\lambda} z^T.$$

于是有

$$\begin{aligned} AV_m y &= V_m T_m y + \gamma_m v_m e_m^T y = \tilde{\lambda} V_m y + \gamma_m e_m^T y v_m, \\ (W_m z)^T A &= (A^T W_m z)^T = (W_m T_m^T z)^T + \beta_m e_m^T z w_m^T = \tilde{\lambda} (W_m z)^T + \beta_m e_m^T z w_m^T. \end{aligned}$$

若  $|\gamma_m(e_m^T y)|$  和  $|\beta_m(e_m^T z)|$  足够小, 我们就认为  $\tilde{\lambda}$  是  $A$  的某个特征值的近似, 而  $V_m y$  和  $W_m z$  就是相应的右特征向量和左特征向量的近似.

#### 算法 8.5. 非对称 Lanczos 算法

- 1: Choose two vectors  $v_0$  and  $w_0$  such that  $(v_0, w_0) = 1$
- 2: Set  $\beta_0 = 0$  and  $\gamma_0 = 0$
- 3: **for**  $j = 0, 1, \dots$  **do**
- 4:   Compute  $\alpha_{j+1} = (Av_j, w_j)$
- 5:    $\tilde{v}_{j+1} = Av_j - \alpha_{j+1}v_j - \beta_j v_{j-1}$
- 6:    $\tilde{w}_{j+1} = A^T w_j - \alpha_{j+1}w_j - \gamma_j w_{j-1}$
- 7:    $\gamma_{j+1} = |(\tilde{v}_{j+1}, \tilde{w}_{j+1})|^{1/2}$
- 8:   **if**  $\gamma_{j+1} = 0$  **then**
- 9:     stop
- 10:   **end if**
- 11:    $\beta_{j+1} = (\tilde{v}_{j+1}, \tilde{w}_{j+1})/\gamma_{j+1}$
- 12:    $v_{j+1} = \tilde{v}_{j+1}/\gamma_{j+1}$
- 13:    $w_{j+1} = \tilde{w}_{j+1}/\beta_{j+1}$
- 14:   Compute the eigenvalues and eigenvectors of  $T_j$
- 15:   Check the convergence



16: **end for**

非对称 Lanczos 算法的显著优点就是节省运算量, 缺点是更容易被中断.



## 附录 A IEEE 浮点运算标准

在数值计算中, 小数在内存中是以浮点数格式表示和参与运算的. 浮点数是数字 (或者说数值) 在内存中的一种存储格式, 它和定点数是相对的.

### A.1 浮点数与定点数

浮点数和定点数中的“点”指的是小数点.

所谓定点数, 就是指小数点的位置是固定的, 不会向前或者向后移动. 比如我们用 4 个字节 (32 位字长) 来存储无符号的定点数  $x$  (通常用 4 个字节存储单精度数), 并且约定: 前 16 位表示整数部分, 后 16 位表示小数部分, 如下图所示:



这种表示方法的优点是: 整数部分和小数部分一目了然, 非常直观.

✍ 对于整数, 所有位都用来存储整数部分, 所以一般采用定点数格式存储.

但定点数格式有个很大的缺点, 就是所能表示的数的取值范围比较小. 比如在前面的例子中 (4 个字节), 所能表示的最大值和最小值 (非零) 是

$$x_{\max} = 1111111111111111.1111111111111111_2 \approx 2^{16} = 65536,$$

$$x_{\min} = 0000000000000000.0000000000000001_2 = 2^{-16} \approx 1.5 \times 10^{-5}.$$

这在科学计算中显然是远远不够的. 比如电子的质量大约是  $9 \times 10^{-28}$  克, 用 4 位定点数格式就无法表示. 即使是采用 8 个字节 (通常用 8 个字节存储双精度数), 假定前 32 位表示整数, 后 32 位表示小数, 则所能表示的数的范围为 (0 除外)


$$x_{\max} \approx 2^{32} \approx 4.3 \times 10^9, \quad x_{\min} = 2^{-32} \approx 2.3 \times 10^{-10}.$$

为了克服这个缺点, 人们发明了一种更加科学的存储格式, 即浮点数格式, 也就是通常所说的科学计数法. 该格式以指数形式存储数字, 不但节省内存, 也非常直观, 而且所能表示的数的范围也大大增加.

### A.2 IEEE 中的浮点数的表示方法

自计算机发明以来, 曾出现许多不同的浮点数表示方式, 但目前最通用的是 IEEE 二进制浮点数算术标准 (IEEE Standard for Binary Floating-Point Arithmetic, 简称 IEEE 754 标准).

IEEE 754 标准的主要起草者是加州大学伯克利分校数学系的 William Kahan 教授, 他帮助 Intel 公司设计了 8087 浮点处理器, 并以此为基础形成了 IEEE 754 标准, Kahan 教授也因此获得了 1987 年的图灵奖.

 William Kahan 由于在浮点运算标准的制定上的杰出贡献, 于 1990 年 1 月获得了图灵奖.

通常一个浮点数由符号、尾数、基和指数组成, 如:

$$-0.31415926_{10} \times 10^2, \quad 0.10101_2 \times 2^3.$$

这里要求小数点前面为零, 小数点后面的数称为**尾数**. 若尾数的首位数字不为 0 时, 我们称其为**正规数** (或**规范化数**), 否则称为**次正规数** (或**非规范化数**). 如  $0.314_{10} \times 10^2$  是正规数, 而  $0.00314_{10} \times 10^4$  是次正规数. 正规化表示方法可以使得每个浮点数的表示方式唯一, 而且可以空出一个位置, 使得表示精度更高.

- IEEE 754 标准中定义了表示浮点数的四种格式:
  - 两种基本的浮点数: **单精度** (32 位字长) 和 **双精度** (64 位字长).  
其中单精度格式具有 24 位有效数字 (二进制), 而双精度格式具有 53 位有效数字 (二进制), 相对于十进制来说, 分别是 7 位 ( $2^{24} \approx 10^7$ ) 和 16 位 ( $2^{53} \approx 10^{16}$ ) 有效数字.
  - 两种扩展的浮点数: **单精度扩展** 和 **双精度扩展**.  
IEEE 754 标准中并未规定扩展格式的精度和大小, 但它指定了最小精度和字长: 单精度扩展需 43 位字长以上, 双精度扩展需 79 位字长以上 (64 位有效数字). 单精度扩展很少使用, 而对于双精度扩展, 不同的机器架构中有着不同的规定, 有的为 80 位字长 (如 X86), 有的为 128 位字长 (如 SPARC).
- 一般来说, 描述一个浮点数的三个基本要素为:
  - 基**: 计算机一般都以 2 为基;
  - 尾数**的位数: 确定有效数字的位数, 即精度;
  - 指数**的位数: 确定所能表示的数的范围.
- 在 IEEE 754 标准中, 浮点数是用二进制表示的, 由三部分组成: 符号 (sign, 其值用  $s$  表示), 指数 (exponent, 其值用  $e$  表示) 和尾数 (fraction, 其值用  $f$  表示), 见图 A.1. 单精度数占 32 位字长 (4 个字节), 第 1 位是符号位, 第 2 至 9 位 (8 位字长) 是指数位, 最后 23 位是尾数. 双精度数占 64 位字长 (8 个字节), 第 1 位是符号位, 第 2 至 12 位 (11 位字长) 是指数位, 最后 52 位是尾数.

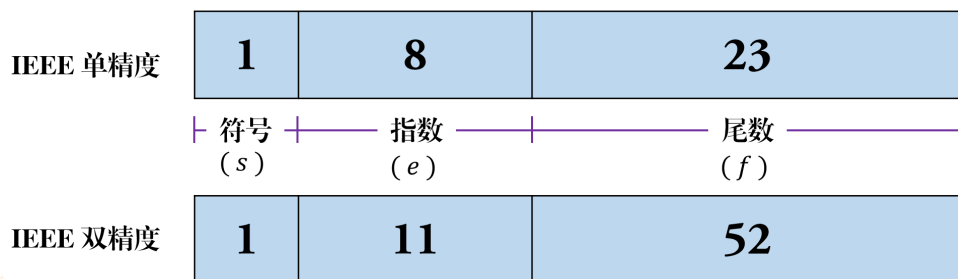


图 A.1. IEEE 754 中单精度格式与双精度格式的位模式

- 单精度格式**: 用 8 位字长的二进制数来表示指数, 因此  $e$  的取值范围为  $[0, 255]$ . 当  $0 \leq e < 255$  时, 按单精度格式存储的数, 其对应的值是使用以下方法得到的:






将二进制基数点 (小数点) 插入到尾数  $f$  最高有效位的左侧, 并将一个**隐含位**插入到二进制基数点的左侧, 从而得到的是一个二进制带分数 (整数加小数).

由此构成的带分数就是**单精度格式有效数字**. 隐含位的值并不是显式指定的 (即不存储), 而是通过指数  $e$  的值来隐式指定:

- 当  $0 < e < 255$  时, 表示该数为**二进制正规数**, 此时隐含位设为 1.
- 当  $e = 0$  时, 表示该数为**二进制次正规数**, 隐含位设为 0.

 由于引入了隐含位 (为了尽可能地增加所能表示的数的精度), 这里的正规数概念与前面的定义有点区别, 因此我们加上“二进制”三个字.

单精度格式位模式中的尾数只有 23 位, 但由于使用了隐含位, 所以能提供 24 位有效数字 (二进制). 因此, 在 IEEE 中, 单精度数的表示方法为

$$\begin{aligned} &(-1)^s \times 1.f \times 2^{e-127} \quad (\text{二进制正规数}) \\ &(-1)^s \times 0.f \times 2^{-126} \quad (\text{二进制次正规数}) \end{aligned}$$

完整的对应关系是

| 单精度格式位模式                | 值                                             |
|-------------------------|-----------------------------------------------|
| $0 < e < 255$           | $(-1)^s \times 1.f \times 2^{e-127}$ (二进制正规数) |
| $e = 0, f \neq 0$       | $(-1)^s \times 0.f \times 2^{-126}$ (二进制次正规数) |
| $e = 0, f = 0$          | $(-1)^s \times 0.0$ (有符号的零)                   |
| $e = 255, f = 0, s = 0$ | $+\text{inf}$ (正无穷大)                          |
| $e = 255, f = 0, s = 1$ | $-\text{inf}$ (负无穷大)                          |
| $e = 255, f \neq 0$     | NaN (非数、非确定值)                                 |

其中 127 是单精度格式的**指数偏移值** (exponent bias), 在 IEEE 标准中, 这个值定义为  $2^{(\text{指数位长}-1)-1}$ .

1. 所以对于单精度格式, 指数偏移值就是  $2^{8-1} - 1 = 127$ , 而对于双精度格式, 这个值为  $2^{11-1} - 1 = 1023$ .

- **双精度格式**: 与单精度格式类似, 完整的对应关系是

| 双精度格式位模式                 | 值                                              |
|--------------------------|------------------------------------------------|
| $0 < e < 2047$           | $(-1)^s \times 1.f \times 2^{e-1023}$ (二进制正规数) |
| $e = 0, f \neq 0$        | $(-1)^s \times 0.f \times 2^{-1022}$ (二进制次正规数) |
| $e = 0, f = 0$           | $(-1)^s \times 0.0$ (有符号的零)                    |
| $e = 2047, f = 0, s = 0$ | $+\text{inf}$ (正无穷大)                           |
| $e = 2047, f = 0, s = 1$ | $-\text{inf}$ (负无穷大)                           |
| $e = 2047, f \neq 0$     | NaN (非数、非确定值)                                  |


**例 A.1** 单精度格式所能表示的十进制数范围.

- 最大二进制正规数为  $7\text{FFFFFF}_{16} = 3.40282347 \times 10^{38}$
- 最小 (正的) 二进制正规数为  $00800000_{16} = 1.17549435 \times 10^{-38}$
- 最大二进制次正规数为  $007\text{FFFFFF}_{16} = 1.17549421 \times 10^{-38}$
- 最小 (正的) 二进制次正规数为  $00000001_{16} = 1.40129846 \times 10^{-45}$

由此可见, 浮点数所能表示的数的范围比定点数要大很多.

**例 A.2** 双精度格式所能表示的十进制数范围.

- 最大二进制正规数为  $7\text{FEFFFFFF FFFFFFFF}_{16} = 1.7976931348623157 \times 10^{308}$
- 最小 (正的) 二进制正规数为  $00100000 00000000_{16} = 2.2250738585072014 \times 10^{-308}$
- 最大二进制次正规数为  $000\text{FFFFFF FFFFFFFF}_{16} = 2.2250738585072009 \times 10^{-308}$
- 最小 (正的) 二进制次正规数为  $00000000 00000001_{16} = 4.9406564584124654 \times 10^{-324}$
- $3\text{FF00000 00000000}_{16} = 1$  (补码)

 在 MATLAB 中, `hex2num` 可以将一个由 16 个 16 进制数组成的字符串转化为其所对应的浮点数 (根据 IEEE 标准), 类似的命令有 `hex2dec`, `bin2dec`, `base2dec`, `num2hex`, `dec2bin`, ...

**例 A.3** 把二进制数  $(1001.0101)_2$  转换成十进制数.

$$\begin{aligned}
 (1001.0101)_2 &= 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 \\
 &\quad + 0 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-2} \\
 &= 9.3125_{10}
 \end{aligned}$$

**例 A.4** 把十进制数  $13.125_{10}$  转换成二进制数.

整数部分:  $13_{10} = 1101_2$  (辗转相除法)

小数部分:

- $0.125 \times 2 = 0.25$ , 整数位是  $0 \rightarrow .0$ ;
- $0.25 \times 2 = 0.5$ , 整数位是  $0 \rightarrow .00$ ;
- $0.5 \times 2 = 1$ , 整数位是  $1 \rightarrow .001$ ; (小数部分已变为 0, 运算结束)

所以  $13.125_{10} = 1101.001_2$

一个十进制数能否用二进制浮点数精确表示, 关键在于小数部分.

**例 A.5** 十进制数  $0.1_{10}$  能否用二进制数精确表示?

- $0.1 \times 2 = 0.2$ , 整数位是  $0 \rightarrow .0$ ;
- $0.2 \times 2 = 0.4$ , 整数位是  $0 \rightarrow .00$ ;



- $0.4 \times 2 = 0.8$ , 整数位是 0  $\rightarrow .000$ ;
- $0.8 \times 2 = 1.6$ , 整数位是 1  $\rightarrow .0001$ ;
- $0.6 \times 2 = 1.2$ , 整数位是 1  $\rightarrow .00011$ ;
- $0.2 \times 2 = 0.4$ , 整数位是 0  $\rightarrow .000110$ ;
- ... ..

得到一个无限循环的二进制小数, 显然用有限位字长是无法表示的, 因此  $0.1_{10}$  无法用 IEEE 754 浮点数精确表示.

同理可知  $0.2, 0.4, 0.6, 0.8, 0.3, 0.7, 0.9$  也是无法精确表示的. 故  $0.1$  至  $0.9$  的 9 个小数中, 只有  $0.5$  可以精确表示.

**例 A.6** 能用二进制数精确表示的十进制小数. 易知

$$\begin{aligned} 0.1_2 &= 2_{10}^{-1} = 0.5 \\ 0.01_2 &= 2_{10}^{-2} = 0.25 \\ 0.001_2 &= 2_{10}^{-3} = 0.125 \\ 0.0001_2 &= 2_{10}^{-4} = 0.0625 \\ 0.00001_2 &= 2_{10}^{-5} = 0.03125 \\ 0.000001_2 &= 2_{10}^{-6} = 0.015625 \\ 0.0000001_2 &= 2_{10}^{-7} = 0.0078125 \\ 0.00000001_2 &= 2_{10}^{-8} = 0.00390625 \\ &\dots \dots \end{aligned}$$

由此可知, 一个十进制小数要能用浮点数精确表示, 最后一位必须是 5. 当然这是必要条件, 并非充分条件. 如  $0.35$  就无法精确表示.

**例 A.7**  $N$  位二进制小数能精确表示的非零十进制小数总共有多少个?

- 1 位二进制小数能精确表示的有  $2^0 = 1$  个 ( $0.1_2 = 0.5_{10}$ );
- 2 位二进制小数能精确表示的有  $2^1 = 2$  个 ( $0.01_2 = 0.25_{10}, 0.11_2 = 0.75_{10}$ );
- 3 位二进制小数能精确表示的有  $2^2 = 4$  个
- ... ..
- $N$  位二进制小数能精确表示的有  $2^{N-1}$  个

所以  $N$  位二进制小数能精确表示的十进制小数总共有  $2^{N-1}$  个.


### A.3 IEEE 中的浮点数运算

- IEEE 754 标准也定义了浮点数的运算规则:
  - 加、减、乘、除、平方根、余数、将浮点格式的数舍入为整数值、在不同浮点格式之间转换、在浮点和整数格式之间转换以及比较: IEEE 对以上浮点运算的准确度作了规定:



求余和比较运算必须精确无误. 其他运算必须向其目标提供精确的结果, 除非没有此类结果, 或者该结果不满足目标格式, 此时运算必须按照下面介绍的舍入模式对精确结果进行最低限度的修改, 并将经过修改的结果提供给运算的目标.

- **在十进制字符串和两种基本浮点格式之一的二进制浮点数之间进行转换的准确度、单一性和一致性要求:** 对于在指定范围内的操作数, 这些转换必须生成精确的结果 (如果可能的话), 或者按照规定的舍入模式, 对此类精确结果进行最低限度的修改. 对于不在指定范围内的操作数, 这些转换生成的结果与精确结果之间的差值不得超过取决于舍入模式的指定误差.
- 五种类型的 IEEE 浮点异常: **无效运算 (如  $0/0$ ,  $\infty/\infty$  等), 被零除, 上溢, 下溢和不精确**, 以及用于向用户指示发生这些类型异常的条件.
- 四种舍入模式: 设  $x$  是所要表示的数,
  - (1) **就近舍入**: 用最接近  $x$  的可表示的值来代替, 类似于整数的四舍五入. 如果  $x$  正好在两个相邻的可表示值的中间, 则首选二进制“偶数”(二进制最后一位为 0);
  - (2) **向下舍入**: 用不大于  $x$  的可表示的值来代替 (向负无穷大方向截断);
  - (3) **向上舍入**: 用不小于  $x$  的可表示的值来代替 (向正无穷大方向截断);
  - (4) **向 0 舍入**: 当  $x > 0$  时采用向下舍入, 当  $x < 0$  时采用向上舍入.

 我们将后面三种舍入模式统称为 **截断**.

 不同编译器对舍入可能有不同的处理方式.

#### • 下溢

当运算结果非常小时, 就会发生下溢. 下表是下溢阈值.

| 目标的精度 | 下溢阈值   |                                       |
|-------|--------|---------------------------------------|
| 单精度   | 最小正规数  | $1.17549435 \times 10^{-38}$          |
|       | 最大次正规数 | $1.17549421 \times 10^{-38}$          |
| 双精度   | 最小正规数  | $2.2250738585072014 \times 10^{-308}$ |
|       | 最大次正规数 | $2.2250738585072009 \times 10^{-308}$ |

IEEE 算法处理下溢的方式是**渐进下溢**: 当生成的正确结果的数量级低于最小正正规数时, 就会生成次正规数, 而不是返回零.

- **机器精度**: 将 1.0 与大于 1.0 的最小浮点数之间的距离记为  $\varepsilon_m$ . 它的一半称为 unit roundoff, 记为  $\varepsilon_u$ , 它是计算机表示一个浮点数时的相对误差界,

$$\text{fl}(x) = x(1 + \delta) \quad \text{或} \quad \text{fl}(x) = \frac{x}{1 + \delta}, \quad |\delta| \leq \varepsilon_u.$$

这里  $\text{fl}(x)$  表示  $x$  在计算机中实际存储的 IEEE 浮点数.

在 IEEE 标准下, 单精度和双精度浮点运算的最大相对误差  $\varepsilon_u$  分别为

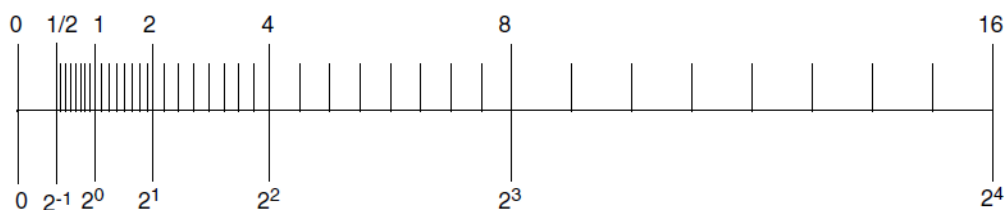


| 精度  | 最大相对误差                                     |
|-----|--------------------------------------------|
| 单精度 | $2^{-24} \approx 5.960464 \times 10^{-8}$  |
| 双精度 | $2^{-53} \approx 1.110223 \times 10^{-16}$ |

如果采用的不是就近舍入模式, 而是其他三种舍入模式 (即截断), 则最大相对误差为  $\varepsilon_m$ .

有的文献中称  $\varepsilon_u$  为机器精度 (**machine epsilon**, **machine precision**, or **macheps**), 如 Demmel [30], LAPACK, Scilab, Wikipedia. 也有的文献称  $\varepsilon_m$  为机器精度, 如 Higham [68], MATLAB, Mathematica. 我们采用前面一种方式, 即“机器精度”指的是  $\varepsilon_u$ .

**例 A.8** 假定要使用只有三个精度位的二进制算法. 那么, 最大相对误差为  $2^{-3}$ . 在任意两个 2 的幂之间, 只有  $2^3 - 1 = 7$  个可表示数字, 如下图所示.



数轴显示了数字之间的差距是随着指数增加而加倍增加的.

在 IEEE 单精度格式中, 两个最小正次正规数之间的差大约是  $10^{-45}$ , 而两个最大有限数之间的数量级差大约是  $10^{31}$ !

精确是偶然的, 误差是必然的. 做数值算法, 惟一能做的就是尽量使误差的传播和累积能够得到有效的控制.

## A.4 浮点运算舍入误差分析

由于计算机无法精确表示所有的浮点数, 在做浮点运算时, 如果计算结果无法精确表示, 此时就会产生的误差, 这就是浮点运算的**舍入误差**. 根据 IEEE 浮点运算标准, 如果  $a \odot b$  的结果无法精确表示, 则用一个最接近的浮点数来代替 (浮点运算时一般采用就近舍入模式), 记为  $\text{fl}(a \odot b)$ . 这里的  $\odot$  表示加、减、乘、除四种运算符.

在不考虑溢出的情况下, 我们有

$$\text{fl}(a \odot b) = (a \odot b)(1 + \delta) \quad \text{或} \quad \text{fl}(a \odot b) = \frac{a \odot b}{1 + \delta}, \quad (\text{A.1})$$

其中  $\delta$  表示浮点运算的相对误差, 满足  $|\delta| \leq \varepsilon_u$ . 公式 (A.1) 是分析浮点运算舍入误差的基础 (标准模型) [68, page 40]. IEEE 浮点运算标准同时也指出, 对于开根号运算, 产生的误差同样也满足

(A.1).

一个比较有趣的问题是何时两个浮点数能进行精确的相减运算.

**定理 A.1 (Ferguson, 1995)** [68] 设浮点数  $x$  和  $y$  满足

$$e(x - y) \leq \min\{e(x), e(y)\} \quad (\text{A.2})$$

其中  $e(\cdot)$  表示一个正规浮点数的指数. 则  $\text{fl}(x - y) = x - y$ . (假定  $x - y$  不会下溢)

这里只考虑浮点运算误差, 所以假定  $x, y$  都是计算机可以精确表示的. 事实上, 由 (A.2) 可知,  $x$  和  $y$  的指数至多差 1.

上面的定理对任何基都成立. 如果基是 2, 则有下列的结论.

**推论 A.2 (Sterbenz, 1974)** 设浮点数  $x$  和  $y$  满足

$$y/2 \leq x \leq 2y,$$

则  $\text{fl}(x - y) = x - y$ . (假定  $x - y$  不会下溢)

下面的结论对估计浮点运算的舍入误差非常有用 [68, page 63].

**引理 A.3** 设  $|\delta_i| \leq \varepsilon_u$  且  $n\varepsilon_u < 1$ , 则

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

其中  $\rho_i = \pm 1$ .

**例 A.9 (多项式运算的舍入误差分析)** : 已知多项式  $p(x) = \sum_{k=0}^n a_k x^k$ . 对于给定的  $x$ , 分析计算  $p(x)$  的值时的舍入误差.

**解.** 在对多项式  $p(x) = a_n x^n + a_{n-1} x^{n-1} \cdots + a_1 x + a_0$  求值时, 我们通常采用 Horner 法则.

**算法 A.1.** Horner 法则

```

1: $p = a_n$
2: for $i = n - 1 : -1 : 0$ do
3: $p = x * p + a_i$
4: end for
```

先看一个具体的例子. 设  $p(x) = (x - 2)^9$ , 观测  $x = 2$  附近的舍入误差. 我们通过画图来显示误差情况 (见图 A.2). 观察发现, 用 Horner 法则求值时会在  $x = 2$  附近会出现“噪声带”.

下面我们基于舍入误差分析模型 (A.1) 来分析多项式求值的误差情况. 带舍入误差的 Horner 算法可写为



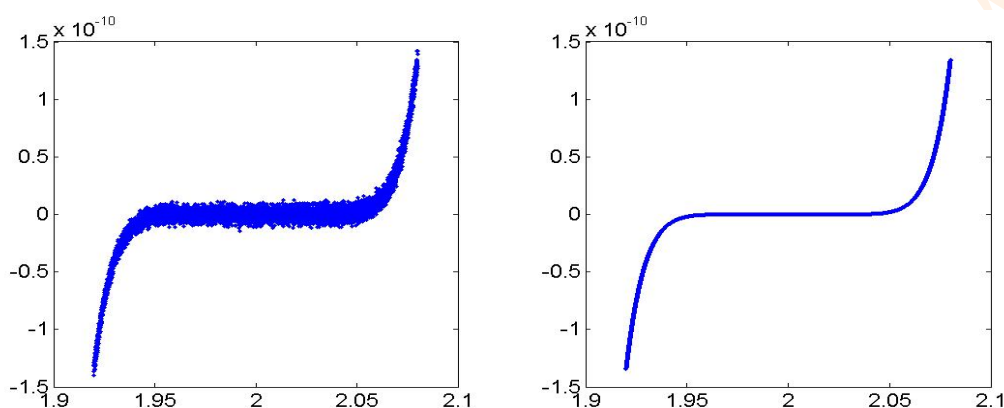


图 A.2. 分别利用 Horner 法则（左图）和  $y = (x - 2)^9$ （右图）在区间  $[1.92, 2.08]$  上的 8000 个等分点上求值的结果

#### 算法 A.2. 带舍入误差的 Horner 法则

```

1: $p = a_n$
2: for $i = n - 1 : -1 : 0$ do
3: $p = ((x * p)(1 + \delta_i) + a_i)(1 + \tilde{\delta}_i)$ % $|\delta_i| \leq \varepsilon_u, |\tilde{\delta}_i| \leq \varepsilon_u$
4: end for

```

所以得到的最终计算结果为

$$\text{fl}(p(x)) = \sum_{i=0}^{n-1} \left( (1 + \tilde{\delta}_i) \prod_{j=0}^{i-1} (1 + \delta_j)(1 + \tilde{\delta}_j) \right) a_i x^i + \left( \prod_{j=0}^{n-1} (1 + \delta_j)(1 + \tilde{\delta}_j) \right) a_n x^n. \quad (\text{A.3})$$

假定  $j \cdot \varepsilon_u \ll 1$ , 则有

$$\begin{aligned} (1 + \delta_1) \cdots (1 + \delta_j) &\leq (1 + \varepsilon_u)^j \lesssim 1 + j \cdot \varepsilon_u, \\ (1 + \delta_1) \cdots (1 + \delta_j) &\geq (1 - \varepsilon_u)^j \geq 1 - j \cdot \varepsilon_u. \end{aligned}$$

于是 (A.3) 可改写为

$$\text{fl}(p(x)) = \sum_{i=0}^n (1 + \hat{\delta}_i) a_i x^i \triangleq \sum_{i=0}^n \tilde{a}_i x^i, \quad \text{其中 } |\hat{\delta}_i| \leq 2n \cdot \varepsilon_u. \quad (\text{A.4})$$

所以  $\text{fl}(p(x))$  可看作是另一个多项式的精确值, 且这个多项式的系数是原多项式系数的一个小的扰动. 这说明多项式计算的 Horner 算法是向后稳定的, 且系数的向后相对误差不超过  $2n \cdot \varepsilon_u$ . 因此, 绝对误差为

$$\begin{aligned} |\text{fl}(p(x)) - p(x)| &= \left| \sum_{i=0}^n \tilde{a}_i x^i - \sum_{i=0}^n a_i x^i \right| \\ &= \left| \sum_{i=0}^n \hat{\delta}_i a_i x^i \right| \leq \sum_{i=0}^n |\hat{\delta}_i a_i x^i| \leq 2n \cdot \varepsilon_u \sum_{i=0}^n |a_i x^i|. \end{aligned}$$





相对误差为

$$\frac{|\text{fl}(p(x)) - p(x)|}{|p(x)|} \leq 2n \cdot \varepsilon_u \frac{\sum_{i=0}^n |a_i x^i|}{\left| \sum_{i=0}^n a_i x^i \right|}.$$

□

## A.5 课后习题

**练习 1.1** 考虑抽样数据  $x_1, x_2, \dots, x_n$ , 其均值为  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , 样本方差为 (无偏估计)

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

试证明:

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^2 - n\bar{x}^2).$$

对于数值计算, 上述两种计算方法哪一个更可靠, 为什么?

**练习 1.2** 根据浮点运算的舍入误差模型 (A.1), 在不考虑溢出的情况下, 证明 [68]

$$\text{fl} \left( \sum_{i=1}^n x_i y_i \right) = \sum_{i=1}^n x_i y_i (1 + \theta_i), \quad \text{其中 } |\theta_i| \leq \gamma_i \triangleq \frac{i\varepsilon_u}{1 - i\varepsilon_u}, \quad i = 1, 2, \dots, n.$$





## 附录 B 数值计算中的误差

数值方法的特点之一就是所求得解是近似解,总是存在一定的误差.因此,误差分析是数值分析中一个很重要的课题.

**误差**是人们用来描述数值计算中近似解的精确程度,是科学计算中的一个十分重要的概念.

误差大致可分为以下几种类型:

- **模型误差**: 数学模型是对实际问题的数学描述,它往往是抓住问题的主要因素而略去次要因素,因此,它是实际问题的一个近似.
- **观测误差**: 在数学模型中通常包含一些参量(数据),这些参量的值一般都是通过测量或实验的方法所得到的,因此也存在误差.
- **截断误差**: 也称**方法误差**,在对数学模型进行数值求解时,需要做一些近似,如对导数离散时可用差商代替.
- **舍入误差**: 由于机器字长有限,由于机器字长有限,计算机对浮点数的表示和算术运算都存在一定的误差.

在数值分析中,我们总是假定数学模型和所给的数据都是准确的,因而不考虑模型误差和观测误差,主要研究截断误差和舍入误差对计算结果的影响.

**例 B.1** 近似计算  $\int_0^1 e^{-x^2} dx$  的值.

**解.** 这里我们采用 Taylor 展开,即

$$\begin{aligned}\int_0^1 e^{-x^2} dx &= \int_0^1 \left( x - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \cdots \right) \\ &= 1 - \frac{1}{3} + \frac{1}{2!} \times \frac{1}{5} - \frac{1}{3!} \times \frac{1}{7} + \frac{1}{4!} \times \frac{1}{9} - \cdots \\ &\triangleq S_4 + R_4\end{aligned}$$

其中  $S_4$  为前四项的部分和,  $R_4$  为剩余部分. 如果我们以  $S_4$  作为定积分的近似值, 则  $R_4$  就是由此产生的误差, 这种误差就称为截断误差.

在计算  $S_4$  的值, 假定我们保留小数点后 4 位有效数字, 则

$$S_4 = 1 - \frac{1}{3} + \frac{1}{10} - \frac{1}{42} \approx 1 - 0.3333 + 0.1000 - 0.0238 = 0.7429$$

这就是我们最后得到的近似值. 这里, 在计算  $S_4$  时所产生的误差就是舍入误差. □

### B.1 误差与有效数字

**定义 B.1** 设  $x^*$  是精确值,  $x$  是其近似值, 则**绝对误差**  $e$  和**相对误差**  $e_r$  分别定义为

$$e = x - x^* \quad \text{和} \quad e_r = \frac{x - x^*}{x^*}.$$

若存在  $\varepsilon > 0$  满足  $|e| = |x - x^*| \leq \varepsilon$ , 则称  $\varepsilon$  为**绝对误差限**, 简称**误差限**. 类似地, 若存在  $\varepsilon_r > 0$  满足  $|e_r| \leq \varepsilon_r$ , 则称  $\varepsilon_r$  为**相对误差限**.

几点说明:

- 绝对误差可能为正, 也可能为负;
- 绝对误差越小越具有参考价, 但绝对误差却不能很好地表示近似值的精确程度;
- 近似值的精确程度取决于相对误差的大小;
- 实际计算中我们所能估计的通常是误差限或相对误差限;
- 由于真值难以求出, 通常也使用下面的定义作为相对误差限

$$e_r = \frac{x - x^*}{x}.$$

- 工程中通常用下面的表达式来刻画近似值的精度:

$$x^* = x \pm \varepsilon,$$

表示精确值在区间  $[x - \varepsilon, x + \varepsilon]$  中.

**定义 B.2** 若近似值  $x$  的误差限是某一位的半个单位, 且该位到  $x$  的第一位非零数字共有  $n$  位, 则称  $x$  有  $n$  位**有效数字**.

关于有效数字的判断, 我们可以使用下面的方法.

**定理 B.1** 设  $x$  是  $x^*$  的近似值, 若  $x$  可表示为

$$x = \pm 0.a_1a_2 \dots a_n \dots \times 10^m,$$

其中  $a_i$  是 0 到 9 中的数字, 且  $a_1 \neq 0$ . 若

$$|x - x^*| \leq 0.5 \times 10^{m-n},$$

则  $x$  至少有  $n$  位有效数字.

换言之, 若  $|x - x^*| \leq 0.5 \times 10^k$ , 则  $x$  至少有  $m - k$  位有效数字.

**例 B.2** 设  $x_1 = 3.1415$  和  $x_2 = 3.1416$  是  $\pi = 3.14159265\dots$  的近似值, 则  $x_1$  有 4 位有效数字, 而  $x_2$  有 5 位有效数字.

**例 B.3** 根据四舍五入原则, 写出下列各数的具有 5 位有效数字的近似值:

$$187.9325, \quad 0.03785551, \quad 8.000033.$$

**解.** 这三个数的具有 5 位有效数字的近似值分别为: 187.93, 0.037856, 8.0000. □

有两点需要注意的是:



- 按四舍五入原则得到的数字是有效数字;
- 一个数末尾的 0 不可以随意添加或省略.

**定理 B.2 (有效数字与相对误差限)** 设  $x$  是  $x^*$  的近似值, 若  $x$  可表示为

$$x = \pm 0.a_1a_2 \dots a_n \dots \times 10^m,$$

其中  $a_i$  是 0 到 9 中的数字, 且  $a_1 \neq 0$ . 若  $x$  具有  $n$  位有效数字, 则其相对误差限满足

$$\varepsilon_r \leq \frac{1}{2a_1} \times 10^{-n+1}.$$

反之, 若  $x$  的相对误差限满足

$$\varepsilon_r \leq \frac{1}{2(a_1+1)} \times 10^{-n+1}.$$

则  $x$  至少有  $n$  位有效数字.

**证明.** 由  $x$  的表达式可知

$$a_1 \times 10^{m-1} \leq |x| \leq (a_1 + 1) \times 10^{m-1}.$$

若  $x$  具有  $n$  位有效数字, 则

$$\varepsilon_r = \frac{|x - x^*|}{|x|} \leq \frac{0.5 \times 10^{m-n}}{a_1 \times 10^{m-1}} = \frac{1}{2a_1} \times 10^{-n+1}.$$

反之, 若  $\varepsilon_r \leq \frac{1}{2(a_1+1)} \times 10^{-n+1}$ , 则

$$|x - x^*| = |x| \cdot \varepsilon_r \leq 0.5 \times 10^{m-n}.$$

故  $x$  至少有  $n$  位有效数字. □

从这个定理可以看出, 有效数字与相对误差是紧密相关的: 有效数字越多, 相对误差就越小; 反之, 相对误差越小, 有效数字就越多.

### B.1.1 基本算术运算的误差估计

误差估计主要是指如何估计误差限或相对误差限. 我们用  $\varepsilon(x)$  表示  $x$  的误差限, 则有

$$\begin{aligned} \varepsilon(x_1 \pm x_2) &\leq \varepsilon(x_1) + \varepsilon(x_2), \\ \varepsilon(x_1 x_2) &\leq |x_2| \varepsilon(x_1) + |x_1| \varepsilon(x_2), \\ \varepsilon\left(\frac{x_1}{x_2}\right) &\leq \frac{|x_1| \varepsilon(x_1) + |x_1| \varepsilon(x_2)}{|x_2|^2}. \end{aligned}$$

### B.1.2 函数求值的误差估计

一般地, 设  $f(x)$  是可微函数,  $x$  为  $x^*$  的近似值, 则由 Taylor 公式可知, 存在  $\xi$  使得

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

所以有

$$\varepsilon(f(x)) \leq |f'(x^*)| \varepsilon(x) + \frac{|f''(\xi)|}{2} \varepsilon^2(x).$$



当  $|f''(\xi)|$  与  $|f'(x^*)|$  的比值不是很大时, 我们可以舍去二次项, 从而得到

$$\varepsilon(f(x)) \lesssim |f'(x^*)|\varepsilon(x).$$

由于  $x^*$  通常是不知道的, 所以我们也用  $f'(x)$  来近似  $f'(x^*)$ , 即

$$\varepsilon(f(x)) \lesssim |f'(x)|\varepsilon(x).$$

关于相对误差限, 我们有如下的估计:

$$\varepsilon_r(f(x)) = \left| \frac{f(x) - f(x^*)}{f(x^*)} \right| \approx \left| \frac{f'(x^*)(x - x^*)}{f(x^*)} \right| = \left| \frac{x^* f'(x^*)}{f(x^*)} \right| \cdot \left| \frac{x - x^*}{x^*} \right| = C_p \varepsilon_r(x),$$

其中  $C_p = \left| \frac{x^* f'(x^*)}{f(x^*)} \right|$  称为  $f(x)$  的 **条件数**.

对于多元可微函数  $f(x_1, x_2, \dots, x_n)$ , 设  $x = (x_1, x_2, \dots, x_n)$  是  $x = (x_1^*, x_2^*, \dots, x_n^*)$  的近似值, 则有

$$\varepsilon(f(x)) \approx \sum_{k=1}^n \left| \frac{\partial f(x)}{\partial x_k^*} \right| \varepsilon(x_k).$$

## B.2 误差分析

- 数值计算中的误差分析很重要, 但也很复杂;
- 在计算过程中, 误差会传播、积累、对消;
- 实际计算中的运算次数通常都在千万次以上, 因此对每一步运算都做误差分析比较不切实际.

误差分析一般可分为定量分析和定性分析.

### B.2.1 定量分析

- 主要方法有: 向前误差分析法, 向后误差分析法, 区间误差分析法, 概率分析法等.
- **向前误差分析**: 用输入数据的误差和数值方法本身的误差来分析计算结果的误差.
- **向后误差分析**: 用某个算法计算  $f(x)$ , 得到的近似解为  $\tilde{f}$ , 假定  $\tilde{f}$  是  $f(x)$  对应于某个数据  $\tilde{x}$  的精确解, 即  $\tilde{f} = f(\tilde{x})$ , 分析  $\tilde{x} - x$  的大小就是向后误差分析.

向后误差分析法 (backward error analysis) 由著名数值分析专家 J. H. Wilkinson 于 1960 年提出, 这是误差理论中最基本的误差分析方法之一.

向后误差分析是一种 **先验误差估计** 方法, 不仅可以用来讨论算法的稳定性, 还可以用于讨论算法的收敛性.


**后验误差估计** 则是利用得到的数值结果来估计近似解的误差, 如在解方程组时, 可以利用残量来估计解的误差.

### B.2.2 定性分析

- 目前在数值计算中更关注的是误差的定性分析;
- 定性分析包括研究数学问题的适定性, 数学问题与原问题的相容性, 数值算法的稳定性, 避免扩大误差的准则等;



- 定性分析的核心是原始数据的误差和计算过程中产生的误差对最终计算结果的影响.

 算法有“优劣”之分, 问题有“好坏”之别, 即使不能定量地估计出最终误差, 但是若能确保计算过程中误差不会被任意放大, 那就能放心地实施计算, 这就是研究定性分析的初衷.

### B.3 数值稳定性


数值计算中的稳定性包括数学问题的稳定性和数值算法的稳定性.

#### B.3.1 数学问题的稳定性

如果数学问题满足

- (1) 对任意满足一定条件的输入数据, 存在一个解,
- (2) 对任意满足一定条件的输入数据, 解是唯一的,
- (3) 问题的解关于输入数据是连续的,

则称该数学问题是**适定的** (well-posed), 否则就称为**不适定的** (ill-posed).

 如果输入数据的微小扰动会引起输出数据 (即计算结果) 的很大变化 (误差), 则称该数值问题是**病态的**.

**例 B.4** 解线性方程组 
$$\begin{cases} x + \alpha y = 1 \\ \alpha x + y = 0 \end{cases}$$

**解.** 易知当  $\alpha = 1$  时, 方程组无解. 当  $\alpha \neq 1$  时, 解为

$$x = \frac{1}{1 - \alpha^2}, \quad y = \frac{-\alpha}{1 - \alpha^2}.$$

当  $\alpha \approx 1$  时, 解的误差可能会很大. 比如当  $\alpha = 0.999$  时,  $x \approx 500.25$ . 假定输入数据  $\alpha$  带有 0.0001 的误差, 即输入数据为  $\alpha^* = 0.9991$ , 则此时有  $x^* \approx 555.81$ , 解的误差约为 55.56, 是输入数据误差的五十多万倍, 因此该问题的病态的. □

#### B.3.2 病态问题与条件数

设  $f(x)$  可导, 则其**条件数**定义为

$$C_p = \left| \frac{x f'(x)}{f(x)} \right|.$$

- 一般情况下, 条件数大于 10 时, 就认为问题是病态的;
- 条件数越大问题病态就越严重;
- 病态是问题本身固有的性质, 与数值算法无关;
- 对于病态问题, 选择数值算法时需要谨慎.



## B.3.3 算法的稳定性

## 例 B.5 近似计算

$$S_n = \int_0^1 \frac{x^n}{x+5} dx, \quad n = 1, 2, \dots, 8.$$

解. 通过观察可知

$$S_n + 5S_{n-1} = \int_0^1 \frac{x^n + 5x^{n-1}}{x+5} dx = \int_0^1 x^{n-1} = \frac{1}{n},$$

因此,

$$S_n = \frac{1}{n} - 5S_{n-1}. \quad (\text{B.1})$$

易知  $S_0 = \ln 6 - \ln 5 \approx 0.182$  (保留三位有效数字), 利用上面的递推公式可得 (保留三位有效数字)

$$S_1 = 0.0900, \quad S_2 = 0.0500, \quad S_3 = 0.0833, \quad S_4 = -0.166,$$

$$S_5 = 1.03, \quad S_6 = -4.98, \quad S_7 = 25.0, \quad S_8 = -125.$$

另一方面, 我们有

$$\frac{1}{6(n+1)} = \int_0^1 \frac{x^n}{6} dx \leq \int_0^1 \frac{x^n}{x+5} dx \leq \int_0^1 \frac{x^n}{5} dx = \frac{1}{5(n+1)}. \quad (\text{B.2})$$

因此, 上面计算的  $S_4, \dots, S_8$  显然是不对的. 原因是什么呢? 误差!

设  $S_n^*$  是  $S_n$  的近似值, 则

$$e(S_n^*) = S_n^* - S_n = \left( \frac{1}{n} - 5S_{n-1}^* \right) - \left( \frac{1}{n} - 5S_{n-1} \right) = -5(S_{n-1}^* - S_{n-1}) = -5e(S_{n-1}^*).$$

即误差是以 5 倍速度增长, 这说明计算过程是不稳定的, 因此我们不能使用该算法.

事实上, 递推公式 (B.1) 可以改写为

$$S_{n-1} = \frac{1}{5n} - \frac{1}{5}S_n.$$

因此, 我们可以先估计  $S_8$  的值, 然后通过反向递推, 得到其它值.

我们可以根据 (B.2) 对  $S_8$  做简单的估计, 即

$$S_8 \approx \frac{1}{2} \left( \int_0^1 \frac{x^8}{6} dx + \int_0^1 \frac{x^8}{5} dx \right) \approx 0.0204.$$

于是

$$S_7 = 0.0209, \quad S_6 = 0.0244, \quad S_5 = 0.0285, \quad S_4 = 0.0343,$$

$$S_3 = 0.0431, \quad S_2 = 0.0580, \quad S_1 = 0.0884, \quad S_0 = 0.182.$$

对比精确值  $S_n^*$  可知, 此时计算出来的  $S_n$  精度要好很多.

通过误差分析可知, 误差是以  $\frac{1}{5}$  的速度减小, 因此计算过程是稳定的.  $\square$

**算法的稳定性:** 通俗地讲, 对于某个给定的算法, 如果输入数据的误差在运算过程不断增长而得不到控制, 那么我们就说该算法是数值不稳定的, 否则就是数值稳定的.

假设输入数据的误差为  $e_0$ , 经  $n$  次运算后的计算结果的误差为  $e_n$ . 如果  $e_n \approx c_1 n e_0$ , 其中  $c_1$  是与  $n$  无关的常数, 则称误差是线性增长的. 如果  $e_n \approx c_2 k^n e_0$ , 其中  $c_2, k$  是与  $n$  无关的常数且





$k > 1$ , 则称误差是指数增长的. 如果算法的误差增长是线性的, 则该算法是数值稳定的, 如果算法的误差是指数增长的, 则该算法是数值不稳定的.

显然误差的线性增长是不可避免的, 而指数增长是必须避免的. 在数值计算中, 误差不可避免, 因此算法的稳定性是一个非常重要的性质.

在数值计算中, 不要采用不稳定的算法!

**定义 B.3 (算法的向后稳定性)** 设用某个算法来计算  $f(x)$ , 得到的近似值为  $\tilde{f}(x)$ . 若对任意的  $x$ , 都存在一个“小”的  $\delta x$ , 使得  $f(x + \delta x) = \tilde{f}(x)$ , 则称该算法是**向后稳定的**, 其中  $\delta x$  称为**向后误差**. 这种误差分析方法就是**向后误差分析**.

若一个算法是向后稳定的, 则有

$$|\tilde{f}(x) - f(x)| = |f(x + \delta x) - f(x)| \approx |f'(x)| \cdot |\delta x|.$$

由于  $\delta x$  很小, 所以当  $|f'(x)|$  不是很大时, 误差总是很小. 因此向后稳定是一个好的算法的基本性质.

向后误差分析将舍入误差归入到截断误差中, 使得误差分析相对简单化.

### B.3.4 数值计算注意事项

在用计算机进行数值计算时, 舍入误差不可避免, 但我们要尽可能地减小舍入误差对计算结果的影响. 在计算过程中, 我们应注意以下几点.

- (1) **避免相近的数相减**. 如果两个相近的数相减, 则会损失有效数字, 如  $0.12346 - 0.12345 = 0.00001$ , 操作数有 5 位有效数字, 但结果却只有 1 为有效数字. 下面给出几个避免相近的数相减的方法:

$$\sqrt{x + \varepsilon} - \sqrt{x} = \frac{\varepsilon}{\sqrt{x + \varepsilon} + \sqrt{x}}$$

$$\ln(x + \varepsilon) - \ln(x) = \ln\left(1 + \frac{\varepsilon}{x}\right)$$

$$1 - \cos(x) = 2 \sin^2 \frac{x}{2}, \quad |x| \ll 1$$

$$e^x - 1 = x \left(1 + \frac{1}{2}x + \frac{1}{6}x^2 + \cdots\right), \quad |x| \ll 1$$

- (2) **避免数量级相差很大的数相除**. 可能会产生溢出, 即超出计算机所能表示的数的范围.
- (3) **避免大数吃小数**. 如直接计算  $(10^9 + 10^{-9} - 10^9)/10^{-9}$  时, 结果可能为 0. 另外, 在对一组数求和时, 应按绝对值从小到大求和.
- (4) **简化计算**. 尽量减少运算次数, 避免误差积累.
- (5) **选用稳定的算法**. 尽可能避免由于算法本身导致的误差增大.



## B.4 课后习题

### 实践题

练习 2.1 数值计算中的误差. 已知  $\sin(x)$  的幂级数展开为

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$$

试利用该公式编程计算  $\sin(\pi/2)$  和  $\sin(33\pi/2)$  的值, 并与实际值做比较, 误差分别多大? 分析原因.





## 附录 C 高性能计算 – 科学计算软件介绍

### C.1 科学计算发展

本节内容摘自 [Predictions for scientific computing 50 years from now](#), by N. Trefethen in 2000.

- Before 1940
  - Newton's method
  - Gaussian elimination
  - Gauss quadrature
  - Least-squares fitting
  - Adams and Runge-Kutta formulas
  - Richardson extrapolation
- 1940 – 1970
  - Floating point arithmetic
  - Fortran
  - Finite differences
  - Finite elements
  - Simplex algorithm
  - Monte Carlo
  - Orthogonal linear algebra
  - Splines
  - FFT
- 1970 – 2000
  - Quasi-Newton iterations
  - Adaptivity
  - Stiff ODE solvers
  - Software libraries
  - MATLAB
  - Multigrid
  - Sparse and iterative linear algebra
  - Spectral methods
  - Interior point method
  - Wavelets
- 2000 –

- Multipole methods
- Breakthroughs in preconditioners, spectral methods, time stepping for PDE
- . . . . .

### C.1.1 数值分析经典论文

这里是 Trefethen 建议的 13 篇论文阅读列表:

(见 [Classic Papers in Numerical Analysis](#), by N. Trefethen, 1993)

- (1) [Cooley & Tukey \(1965\)](#) → [the Fast Fourier Transform](#)  
James W. Cooley and John W. Tukey, “An algorithm for the machine calculation of complex Fourier series,” *Mathematics of Computation*, 19 (1965), 297–301.
- (2) [Courant, Friedrichs & Lewy \(1928\)](#) → [finite difference methods for PDE](#)  
R. Courant, K. O. Friedrichs and H. Lewy, “Ueber die partiellen Differenzengleichungen der mathematischen Physik,” *Mathematische Annalen*, 100 (1928), 32–74. Translated as: “On the partial difference equations of mathematical physics,” *IBM Journal of Research and Development*, 11 (1967), 215–234.
- (3) [Householder \(1958\)](#) → [QR factorization of matrices](#)  
A. S. Householder, “Unitary triangularization of a nonsymmetric matrix,” *Journal of the Association of Computing Machinery*, 5 (1958), 339–342.
- (4) [Curtiss & Hirschfelder \(1952\)](#) → [stiffness of ODEs; BD formulas](#)  
C. F. Curtiss and J. O. Hirschfelder, “Integration of stiff equations,” *Proceedings of the National Academy of Sciences*, 38 (1952), 235–243.
- (5) [de Boor \(1972\)](#) → [calculations with B-splines](#)  
C. de Boor, “On calculating with B-splines,” *Journal of Approximation Theory*, 6 (1972), 50–62.
- (6) [Courant \(1943\)](#) → [finite element methods for PDE](#)  
R. Courant, “Variational methods for the solution of problems of equilibrium and vibrations,” *Bulletin of the American Mathematical Society*, 49 (1943), 1–23.
- (7) [Golub & Kahan \(1965\)](#) → [the singular value decomposition](#)  
G. Golub and W. Kahan, “Calculating the singular values and pseudo-inverse of a matrix,” *SIAM Journal on Numerical Analysis*, 2 (1965), 205–224.
- (8) [Brandt \(1977\)](#) → [multigrid algorithms](#)  
A. Brandt, “Multi-level adaptive solutions to boundary-value problems,” *Mathematics of Computation*, 31 (1977), 333–390.
- (9) [Hestenes & Stiefel \(1952\)](#) → [the conjugate gradient iteration](#)  
Magnus R. Hestenes and Eduard Stiefel, “Methods of conjugate gradients for solving linear systems,” *Journal of Research of the National Bureau of Standards*, 49 (1952), 409–436.
- (10) [Fletcher & Powell \(1963\)](#) → [optimization via quasi-Newton updates](#)  
R. Fletcher and M. J. D. Powell, “A rapidly convergent descent method for minimization,” *Computer Journal*, 6 (1963), 163–168.



(11) **Wanner, Hairer & Norsett (1978) → order stars and applications to ODE**

G. Wanner, E. Hairer and S. P. Norsett, "Order stars and stability theorems," BIT, 18 (1974), 475–489.

(12) **Karmarkar (1984) → interior point methods for linear programming**

N. Karmarkar, "A new polynomial-time algorithm for linear programming," Combinatorica, 4 (1984), 373–395.

(13) **Greengard & Rokhlin (1987) → multipole methods for particles**

L. Greengard and V. Rokhlin, "A fast algorithm for particle simulations," Journal of Computational Physics, 72 (1987), 325–348.

**C.1.2 Longer list of papers**

一个更长的阅读论文列表 (N. Trefethen, 1993)

• **LINEAR ALGEBRA – SYSTEMS OF EQUATIONS AND LEAST-SQUARES**

- Frankel (1950) optimal omega for SOR iteration
- Hestenes & Stiefel (1952) the conjugate gradient iteration
- Young (1954) theory of classical iterative methods
- Householder (1958) QR decomposition
- Wilkinson (1961) error analysis for systems of eqs.
- Golub (1965) least-squares problems
- Strassen (1969) Gaussian elimination is not optimal
- George (1973) nested dissection
- Gill, Golub, Murray & Saunders (1974) updating matrix factorizations
- Concus, Golub & O’Leary (1976) preconditioned conjugate gradients
- Meijerink & van der Vorst (1977) incomplete LU preconditioning
- Skeel (1980) iterative refinement and stability
- Saad & Schultz (1986) GMRES for nonsymmetric systems

• **LINEAR ALGEBRA - EIGENVALUES AND SVD**

- Jacobi (1846) Jacobi’s method for matrix eigenvalues
- Henrici (1958) convergence of the Jacobi method
- Rutishauser (1958) the LR algorithm
- Kublanovskaya (1961) the QR algorithm
- Francis (1961) the QR algorithm
- Golub & Kahan (1965) computation of the SVD
- Moler & Stewart (1973) QZ algorithm for gen’d eigenvalues
- Cuppen (1981) divide and conquer for eigenvalues

• **OPTIMIZATION**

- Dantzig (1951) simplex method for linear programming



- Davidon (1959) variable metric methods
- Fletcher & Powell (1963) DFP quasi-Newton update formula
- Broyden, Fletcher, Goldfarb & Shanno (1970) BFGS quasi-Newton update formula
- Karmarkar (1984) interior pt methods for linear prog.
- INTEGRATION
  - Golub & Welsch (1969) Gauss quadrature rules
  - de Boor (1971) adaptive quadrature algorithms
- APPROXIMATION
  - Remes (1934) Remes algorithm for Chebyshev approx.
  - Schoenberg (1946) splines
  - Powell (1967) near-optimality of Chebyshev interp.
  - Reinsch (1967) smoothing with splines
  - Cox (1972) calculation with B-splines
  - de Boor (1972) calculation with B-splines
- OTHER
  - Aitken (1932) Aitken extrapolation
  - Cooley & Tukey (1965) the fast Fourier transform
  - Greengard & Rokhlin (1987) fast multipole methods
- ODEs
  - Curtiss & Hirschfelder (1952) stiffness and BD formulas
  - Dahlquist (1956) stability and convergence
  - Dahlquist (1963) A-stability
  - Butcher (1965) Runge-Kutta methods
  - Gear (1969) stiff ODEs
  - Wanner, Hairer & Norsett (1978) order stars and stability theorems
- ELLIPTIC PDEs
  - Peaceman & Rachford (1955) ADI
  - Douglas (1955) ADI
  - Strang (1971 or 1973) finite elements and approx. theory
  - Buzbee, Golub & Nielsen (1970) fast Poisson via cyclic reduction
  - Hockney (1965) fast Poisson via FFT
  - Fedorenko (1961) multigrid methods
  - Brandt (1977) multigrid methods
- PARABOLIC AND HYPERBOLIC PDEs
  - Courant, Friedrichs & Lewy (1928) the CFL condition
  - Crank & Nicolson (1947) finite differences for parabolic PDE



- O'Brien, Hyman & Kaplan (1951) Von Neumann stability analysis
- Lax & Richtmyer (1956) general stability theory
- Lax & Wendroff (1960,1962,1964) methods for solving conservation laws
- Kreiss (1962) more general stability theory
- Orszag (1971) spectral methods
- Kreiss & Oliger (1972) spectral methods
- Gustafsson, Kreiss & Sundstrom (1972) stability of boundary conditions
- Chorin (1973) vortex methods for CFD
- Engquist & Majda (1977) absorbing boundary conditions


## C.2 矩阵运算的复杂度

算法的执行效率依赖于算法中所涉及的运算类型和运算次数. 一般来说, 数值算法所涉及的运算主要是加减乘除, 以及开根号运算. 开根号运算依赖于具体的实现算法, 一般最终也可以归结为加减乘除运算, 而且开根号运算次数往往比加减乘除运算具有更低的数量级. 因此, 评价算法的一个重要标准就是加减乘除运算的次数.

一般情况下, 加减运算次数与乘法运算次数具有相同的数量级, 而除法运算次数往往比乘法运算具有更低的数量级.


常见数值运算的计算量:

| $\mathcal{O}(n)$ 量级         | $\mathcal{O}(n^2)$ 量级                                   | $\mathcal{O}(n^3)$ 量级              |
|-----------------------------|---------------------------------------------------------|------------------------------------|
| $x \leftarrow \alpha x$     | $x \leftarrow Ax$                                       | $B \leftarrow \alpha AB$           |
| $y \leftarrow y + \alpha x$ | $y \leftarrow \alpha Ax + \beta y$                      | $C \leftarrow \alpha AB + \beta C$ |
| $s \leftarrow y^* x$ (内积)   | $x \leftarrow A^{-1}x$ ( $A$ 三角矩阵)                      |                                    |
| $s \leftarrow \ x\ _2$      | $A \leftarrow A + \alpha xy^*$ (秩 1 修正)                 |                                    |
| $s \leftarrow \ x\ _1$      | $A \leftarrow A + \alpha xy^* + \alpha yx^*$ (对称秩 2 修正) |                                    |
| 向量的交换, 复制, 比较               |                                                         |                                    |

 为了提高程序执行效率, 节约程序开发成本, 建议尽可能地使用现有的高性能优秀程序库, 如 BLAS, LAPACK 等.


## C.3 矩阵乘积的快速算法

通常, 两个  $n$  阶稠密矩阵相乘, 运算量为  $\mathcal{O}(n^3)$  (包括  $n^3$  次乘法和  $n^3 - n^2$  次加法). 1969 年, Strassen [118] 采用了分而治之技术, 将运算量降低至  $\mathcal{O}(n^{\log 7}) \approx \mathcal{O}(n^{2.807355})$ .

 Volker Strassen 是一位出生于 1936 年的德国数学家. 他因为在概率论上的工作而广为人知, 但在计算机科学和算法领域, 他却因为矩阵相乘算法而被大部分人认识, 这个算法目前仍



然是比通用矩阵相乘算法性能好的主要算法之一。

 Strassen 在 1969 年提出了这个算法, 并据此说明复杂度为  $\mathcal{O}(n^3)$  的算法并不是最优算法, 由此触发了矩阵乘积快速算法领域的更多研究。

我们首先考虑两个  $2 \times 2$  矩阵的乘积. 设

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

如果按照通常方法计算  $C = AB$ , 则需要 8 次乘法和 4 次加法. Strassen 提出了一个新的计算方法, 只需要 7 次乘法和 18 次加法. 记

$$p_1 = (a_{11} + a_{22})(b_{11} + b_{22})$$

$$p_2 = (a_{21} + a_{22})b_{11}$$

$$p_3 = a_{11}(b_{12} - b_{22})$$

$$p_4 = a_{22}(b_{21} - b_{11})$$

$$p_5 = (a_{11} + a_{12})b_{22}$$

$$p_6 = (a_{21} - a_{11})(b_{11} + b_{12})$$

$$p_7 = (a_{12} - a_{22})(b_{21} + b_{22})$$

则

$$c_{11} = p_1 + p_4 - p_5 + p_7,$$

$$c_{12} = p_3 + p_5,$$

$$c_{21} = p_2 + p_4,$$

$$c_{22} = p_1 - p_2 + p_3 + p_6.$$

下面考虑  $n$  阶矩阵的乘积. 假定  $n = 2m$ , 首先将  $A$  和  $B$  写成  $2 \times 2$  分块形式, 每个子块都是  $m \times m$  的矩阵. 然后则利用上面的技巧, 则需  $7m^3$  次乘积和  $7(m^3 - m^2) + 18m^2 = 7m^3 + 11m^2$  次加法. 与普通矩阵乘法的  $8m^3$  乘法和  $8m^3 - 4m^2$  相比, Strassen 算法的运算量大约是普通算法的  $7/8$ , 特别是  $m$  很大时.

事实上, Strassen 算法的运算量比这要少的多. 我们可以采用递归技术, 即计算子矩阵乘积时, 仍可以进行分块计算. 假定  $n = 2^k$ , 则可以不断分块, 直至子矩阵为  $1 \times 1$ . 设总的乘法运算次数为  $t(k)$ , 则

$$t(k) = 7t(k-1).$$

因此

$$t(k) = 7^k = n^{\log(7)}.$$



当  $n = 10^3$  时, Strassen 算法的运算量大约是普通算法的 0.27 倍, 而当  $n = 10^6$  时, Strassen 算法的运算量大约是普通算法的 0.07 倍.

在实际计算中, 当矩阵维数足够小时, 可以直接计算, 无需再分块.

Strassen 算法并不是当前矩阵乘积的最快算法. 2010 年之前, 矩阵乘积的最快算法是 Coppersmith-Winograd [27] 算法, 由 Coppersmith 和 Winograd 于 1990 年提出, 运算量大约为  $\mathcal{O}(n^{2.375477})$ . 2010 年, Stothers [117] 将算法运算量降低到  $\mathcal{O}(n^{2.374})$ . 2011 年, Williams [129] 再次优化算法, 将运算量降低到  $\mathcal{O}(n^{2.3728642})$ . 2014, Le Gall [83] 简化了 Williams 算法, 将运算量进一步降低到  $\mathcal{O}(n^{2.3728639})$ .

## C.4 数值线性代数程序库

BLAS, LAPACK, ARPACK 是当前数值计算中优秀程序库的典型代表, 许多科学计算软件都是基于这些程序库, 如著名的 MATLAB.

### C.4.1 BLAS

BLAS (Basic Linear Algebra Subprograms) 是一组高质量的子程序, 用于实现基本的向量和矩阵运算, 最初发布于 1979 年. 在 BLAS 中, 所有的子程序都经过精心的优化以确保其高效性, 同时, 程序的书写和语句的选择都很规范, 以便于移植.

长期以来, 专家们一直呼吁使用 BLAS 来建立和开发自己的线性代数软件包. 这样做, 不仅可以缩短开发周期, 而且便于形成统一的调用接口. 在高性能计算领域, BLAS 被广泛使用. 合理的调用 BLAS 子程序, 可以大大提高程序的性能.

目前 BLAS 库有多种不同的优化实现, 如 BLAS, Goto BLAS, ATLAS 等. 为提高性能, 各软硬件厂商都会对其产品的 BLAS 接口实现进行高度优化.

在 BLAS 中, 向量和矩阵运算被分为三个层次. 第一层次 (Level 1) BLAS 是指标量与向量, 向量与向量之间的运算. 第二层次 (Level 2) BLAS 是指矩阵与向量的运算. 而第三层次 (Level 3) BLAS 是指矩阵与矩阵之间的运算.

BLAS 主页为 <http://www.netlib.org/blas/>.

### C.4.2 LAPACK

LAPACK (Linear Algebra PACKage) 是由美国 Tennessee 大学, 加州大学 Berkeley 分校, 科罗拉多大学丹佛分校和 NAG (Numerical Algorithms Group) 公司联合开发的线性代数子程序库. 用于在不同高性能计算环境上高效求解数值线性代数问题, 包含了求解科学与工程计算中最常见的数值线性代数的计算问题, 如线性方程组, 线性最小二乘问题, 特征值问题, 奇异值问题等.

LAPACK 主页为 <http://www.netlib.org/lapack/>





### C.4.3 ARPACK

ARPACK (Arnoldi PACKage) 是由 Rice 大学开发, 用于求解大规模特征值问题. 该软件包基于 Implicitly Restarted Arnoldi 算法, 非常适合求解大型稀疏或结构化矩阵的特征值问题.

ARPACK 主页为 <http://www.caam.rice.edu/software/ARPACK/>

### C.4.4 其它

- C++ 符号计算库 GiNaC : <http://www.ginac.de/>
- C++ 线性代数库 Eigen : <http://eigen.tuxfamily.org/>
- C++ 线性代数库 Armadillo : <http://arma.sourceforge.net/>
- 多精度整数和有理数 MPIR (Multiple Precision Integers and Rationals): <http://www.mpir.org/>

## C.5 交互式数学软件

本小节介绍几个优秀的交互式数学软件:

MATLAB, Maple, Mathematica, Sage.

其中 MATLAB 是数值计算功能最强的, 而 Maple 和 Mathematica 则是符号计算软件中的佼佼者. 这三个都是商业软件, 而且价格不菲. Sage 则是 2005 年新开发出来的开源免费数学软件, 其目标就是创建一个有活力的自由开源软件以替代 MATLAB, Maple 和 Mathematica.

关于这四个软件的更多介绍, 可以参考其使用手册, 或其它相关资料, 如 [70].

### C.5.1 MATLAB

MATLAB 是由美国 Mathworks 公司开发的科学计算软件, 集数值计算, 数据可视化和交互式程序设计于一身, 为科学研究和工程设计的众多科学领域提供了一种全面的解决方案, 并在很大程度上摆脱了传统非交互式程序设计语言 (如 C、Fortran) 的编辑模式, 代表了当今国际科学计算软件的先进水平.

MATLAB 起源于 20 世纪七十年代末. 时任新墨西哥大学计算机科学系主任的 Cleve Moler 教授, 在给学生开设线性代数课程时, 为了给学生上机提供方便, 减轻学生编程负担, 编写了一组 LINPACK 和 EISPACK 程序库的“通俗易懂”的程序接口, 并取名为 MATLAB, 即 **MAT**rix **LAB**oratory.

几年后, 工程师 John Little 与 Moler, Steve Bangert 一起开发了第二代 MATLAB, 所有内核改用 C 语言编写, 除原有的数值计算功能外, 还新增了数据可视化功能, 并于 1984 年成立 Mathworks 公司, 推出第一个商业版 MATLAB.

MATLAB 以出现后的短短几年, 就以其良好的开放性和运行的可靠性, 非常受欢迎. 到了 20 世纪九十年代, 在国际上数学类科技应用软件中, MATLAB 在数值计算方面独占鳌头.





在欧美大学里, 诸如应用代数, 数理统计, 自动控制, 数字信号处理, 模拟与数字通信, 时间序列分析, 动态系统仿真等课程的教科书都把 MATLAB 作为内容. 这几乎成了九十年代教科书与旧版书籍的区别性标志. 在那里, MATLAB 是攻读学位的大学生, 硕士生, 博士生必须掌握的基本工具.

现在的 MATLAB 软件主要包括 MATLAB 和 Simulink 两大部分, 以及各种工具箱. 每年更新两次, 并以年份作为版本号. 最新版本中的矩阵计算主要是基于 LAPACK 程序库.

MATLAB 的主要优势: (1) 友好的工作平台和编程环境; (2) 简单易用的程序语言; (3) 强大的科学计算和数据处理能力; (4) 出色的绘图功能, 以及 (5) 实用的程序接口和发布平台. 因而深受广大科技工作者的欢迎.

MATLAB 官方主页为 <http://www.mathworks.com>

### C.5.2 Mathematica

Mathematica 是一款符号计算软件, 美国 Wolfram Research 公司开发, 很好地结合了符号计算, 数值计算, 图形系统, 编程语言, 文本系统等. 很多功能在相应领域内处于世界领先地位, 是使用最广泛的数学软件之一, 与 MATLAB 和 Maple 并称三大数学软件.



Mathematica 自 1988 发布以来, 已经对如何在科技和其它领域运用计算机产生了深刻的影响. 人们常说, Mathematica 的发布标志着现代科技计算的开始. Mathematica 的基本概念是用一个连贯和统一的方法创造一个能适用于科技计算各个领域的软件系统. 实现这一点的关键之处是发明了一种新的计算机符号语言, 这种语言能仅仅用很少量的基本元素制造出广泛的物体, 满足科技计算的广泛性. 这在人类历史上还是第一次. 当 Mathematica 刚发布时, 纽约时代报写道: “这个软件的重要性不可忽视”. 紧跟着, 商业周刊又将 Mathematica 评为当年十大最重要产品. 在科技界, Mathematica 被形容为智能和实践的革命.

Mathematica 分为两部分: 内核和前端. 内核负责对代码进行解释, 并且返回结果. 前端提供了一个 GUI, 使得用户可以创建并且编辑一个“笔记本文档”. 该笔记本文档可以包含程序代码和其它格式化的文本, 如公式, 图像, 表格, 声音等, 并且支持标准文字处理功能. 所有的内容和格式都可以通过算法生成或者通过交互式方法进行编辑.

文档可以使用层次式单元进行结构化处理, 这样便于对文档划分章节. 文档也可以表示为幻灯片形式, 便于进行演讲. 笔记本与其内容均以 Mathematica 表达式的形式存储, 并且可用使用 Mathematica 程序进行创建, 编辑和修改, 而且还可以转化为其它格式, 比如  $\text{\LaTeX}$  或者 XML.

前端还包括其它一些功能, 如程序调试, 输入自动补全, 自动语法着色等. 默认情况下, Mathematica 使用一个标准前端, 但也有其它前端可供选择. 此外, Mathematica 还包括一个命令行前端 (Mathematica Kernel).

Mathematica 的官方主页为 <http://www.wolfram.com>



### C.5.3 Maple

Maple 是目前世界上最为通用的数学和工程计算软件之一, 在数学和科学领域享有盛誉, 以符号运算为主, 有“数学家的软件”之称. Maple 被广泛地应用于科学, 工程和教育等领域.



Maple 系统包括强大的符号计算, 无限精度数值计算, 创新的互联网连接等. 其数学和分析功能几乎覆盖所有的数学分支, 如微积分, 微分方程, 特殊函数, 线性代数, 图像声音处理, 统计, 动力系统.

Maple 最初由加拿大滑铁卢大学的 Symbolic Computation Group 于 1980 年开发. 1988 年成立 Waterloo Maple 公司, 负责 Maple 的商业运营.

Maple 不仅仅是编程工具, 更能提供数学知识. 用户通过 Maple 可以在单一的环境中完成多领域建模和仿真, 符号计算, 数值计算, 程序设计, 报告演示, 算法开发, 外部程序连接等功能, 满足各个层次用户的需要.

Maple 的官方主页为 <http://www.maplesoft.com>

### C.5.4 SageMath

SageMath 是一个基于 GPL 协议的开源数学软件. 它使用 Python 作为通用接口, 将现有的许多开源软件包整合在一起, 构建一个统一的计算平台.



SageMath 的第一个版本发布于 2005 年, 由华盛顿大学 William Stein 教授领导开发. 其使命是 “Creating a viable free open source alternative to Magma, Maple, Mathematica and Matlab.”

SageMath 有 Linux 和 MacOS 版本, 在 Windows 下可以通过虚拟 Linux 环境来实现. 不过用户也可以通过浏览器直接在线使用.

SageMath 的官方主页为 [www.sagemath.org](http://www.sagemath.org), 有中文版的入门手册, <http://www.sagemath.org/zh/>.

## C.6 集群管理

HPC (High-Performance Computing, 高性能计算): 分成若干可以并行的子任务, 并行的子任务间联系很紧密, 并需要大量的数据交换. 作业管理软件: OpenPBS, Torque, Slurm

HTC (High-Throughput Computing, 高吞吐量计算): 分成若干可以并行的子任务, 但各个子任务彼此间没有什么关联. 作业管理软件: HTCCondor

集群管理软件比较: [https://en.m.wikipedia.org/wiki/Comparison\\_of\\_cluster\\_software](https://en.m.wikipedia.org/wiki/Comparison_of_cluster_software)



## 参考文献

- [1] J.O. Aasen, On the reduction of a symmetric matrix to tridiagonal form, *BIT*, 11 (1971), 233–242. 65
- [2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney and D. Sorensen, *LAPACK Users' Guide*, 3rd Edition, SIAM, Philadelphia, 1999. <https://netlib.org/lapack/lug/> 192
- [3] M. Arioli, V. Pták, and Z. Strakoš, Krylov sequences of maximal length and convergence of GMRES, *BIT*, 38 (1998), 636–643. 278
- [4] C. Ashcraft, R.G. Grimes and J.G. Lewis, Accurate symmetric indefinite linear equation solvers, *SIAM Journal on Matrix Analysis and Applications*, 20 (1998), 513–561. 65
- [5] J.L. Aurentz, T. Mach, R. Vandebril and D.S. Watkins, Fast and backward stable computation of roots of polynomials, *SIAM Journal on Matrix Analysis and Applications*, 36 (2015), 942–973. 158, 159
- [6] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1994. 208, 275
- [7] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, 2000. 133
- [8] Z.-Z. Bai, G.H. Golub and M.K. Ng, Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems, *SIAM Journal on Matrix Analysis and Applications*, 24 (2003), 603–626. 245, 246
- [9] M. Van Barel, G. Heinig and P. Kravanja, A stabilized superfast solver for nonsymmetric toeplitz systems, *SIAM Journal on Matrix Analysis and Applications*, 23 (2001), 494–510. FORTRAN code <http://people.cs.kuleuven.be/~marc.vanbarel/software/>
- [10] V. Bargmann, C. Montgomery and J. von Neumann, *Solution of Linear Systems of High Order*, Princeton, N.J.: Institute for Advanced Study, 1946.
- [11] R. Barrett, et.al, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, 1994. (<http://www.netlib.org/templates/index.html>) 208
- [12] V. Barwell and A. George, A comparison of algorithms for solving symmetric indefinite systems of linear equations, *ACM Transactions on Mathematical Software*, 2 (1976), 242–251. 65
- [13] M. Benzi, Preconditioning techniques for large linear systems: A survey, *Journal of Computational Physics*, 182 (2002), 418–477. 207, 280
- [14] D.A. Bini, P. Boito, Y. Eidelman, L. Gemignani and I. Gohberg, A fast implicit QR eigenvalue algorithm for companion matrices, *Linear Algebra and its Applications*, 432 (2010), 2006–2031. 158, 159
- [15] Åke Björck, Solving linear least square problems by Gram-Schmidt orthogonalization, *BIT*, 7 (1967), 1–21. 105
- [16] Åke Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, 1996. 86, 105
- [17] V. Britanak, P. Yip and K. Rao, *Discrete Cosine and Sine Transforms: General properties, Fast algorithms and Integer Approximations*, Academic Press, 2007. 248
- [18] N.I. Buleev, A numerical method for solving two-dimensional diffusion equations, *Atomic Energy*, 6 (1959), 338. (in Russian) 288
- [19] N.I. Buleev, A numerical method for solving two- and three-dimensional diffusion equations, *Mathematical Collection*, 51 (1960), 227–238. (in Russian) 288
- [20] J.R. Bunch, Analysis of the diagonal pivoting method, *SIAM Journal on Numerical Analysis*, 8 (1971), 656–680. 64
- [21] J.R. Bunch and L. Kaufman, Some stable methods for calculating inertia and solving symmetric linear systems, *Mathematics of Computation*, 31 (1977), 163–179. 65
- [22] J. R. Bunch and B. N. Parlett, Direct methods for solving symmetric indefinite systems of linear equa-



- tions, *SIAM Journal on Numerical Analysis*, 8 (1971), 639–655. 64
- [23] R. P. Brent, Stability of fast algorithms for structured linear systems, <http://arxiv.org/pdf/1005.0671v1.pdf>, 2010.
- [24] S. Chandrasekaran, M. Gu, J. Xia and J. Zhu, A fast QR algorithm for companion matrices, In: J. A. Ball, Y. Eidelman, J.W. Helton, V. Olshevsky, J. Rovnyak (eds) *Recent Advances in Matrix and Operator Theory*, 111–143, 2007. *Operator Theory: Advances and Applications* book series, vol 179, 158, 159
- [25] S. H. Cheng, *Symmetric Indefinite Matrices: Linear System Solvers and Modified Inertia Problems*, Ph.D. thesis, University of Manchester, 1998. 65
- [26] G. Codevico, G. Heinig and M. Van Barel, A superfast solver for real symmetric Toeplitz systems using real trigonometric transformations, *Numer. Linear Algebra Appl.*, 12 (2005), 699–713. MATLAB code <http://people.cs.kuleuven.be/~marc.vanbareel/software/>
- [27] D. Coppersmith and S. Winograd, Matrix multiplication via arithmetic progressions, *Journal of Symbolic Computation*, 9 (1990), 251–280. 323
- [28] J.J.M. Cuppen, A Divide and Conquer Method for the Symmetric Tridiagonal Eigenproblem, *Numerische Mathematik*, 36 (1981), 177–195. 175, 178
- [29] T. A. Davis, *Direct Methods for Sparse Linear Systems*, SIAM, 2006. 47
- [30] J.W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997. 47, 166, 184, 255, 305
- [31] J. J. Dongarra, I. S. Duff, D. C. Sorensen and H. A. van der Vorst, *Numerical Linear Algebra for High-Performance Computers*, SIAM, Philadelphia, PA, 1998. 65
- [32] J.J. Dongarra and V. Eijkhout, Numerical linear algebra algorithms and software, *Journal of Computational and Applied Mathematics*, 123 (2000), 489–514. 47
- [33] Z. Drmač and K. Veselić, New fast and accurate jacobi SVD algorithm. I *SIAM Journal on Matrix Analysis and Applications*, 29 (2008), 1322–1342. 188, 193
- [34] Z. Drmač and K. Veselić, New fast and accurate jacobi SVD algorithm. II *SIAM Journal on Matrix Analysis and Applications*, 29 (2008), 1343–1362. 188, 193
- [35] A. A. Dubrulle, Householder Transformations Revisited, *SIAM Journal on Matrix Analysis and Applications*, 22 (2000), 33–40. 94
- [36] J. Durbin, The fitting of time-series models, *Review of the International Statistical Institute*, 28 (1960), 233–243. <https://doi.org/10.2307/1401322> 70
- [37] I. S. Duff, A. M. Erisman and J. K. Reid, *Direct Methods for Sparse Matrices*, 2nd edition, Oxford, 2017. 47, 82
- [38] I. S. Duff, N. I. M. Gould, J. K. Reid, J. A. Scott and K. Turner, The factorization of sparse symmetric indefinite matrices, *IMA Journal of Numerical Analysis*, 11 (1991), 181–204. 65
- [39] I. S. Duff and J. K. Reid, The multifrontal solution of indefinite sparse symmetric linear equations, *ACM Transactions on Mathematical Software*, 9 (1983), 302–325. 65
- [40] M. Engeli, Th. Ginsburg, H. Rutishauser and E. Stiefel, *Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems*, Birkhäuser, Basel/Stuttgart, 1959. 208
- [41] D. J. Evans, The use of pre-conditioning in iterative methods for solving linear systems with symmetric positive definite matrices, *Journal of the Institute of Mathematics and Applications*, 4 (1968), 295–314. 280
- [42] V. Faber, W. Joubert, E. Knill and T. Manteuffel, Minimal residual method stronger than polynomial preconditioning, *SIAM Journal on Matrix Analysis and Applications*, 17 (1996), 707–729. 279
- [43] K. Fernando and B. Parlett, Accurate singular values and differential qd algorithms, *Numerische Mathematik*, 67 (1994), 191–229. 188, 191





- [44] B. Fischer, *Polynomial based iteration methods for symmetric linear systems*, Wiley-Teubner Series Advances in Numerical Mathematics, John Wiley & Sons Ltd., Chichester, 1996. 277
- [45] J. G. F. Francis, The QR transformation: A unitary analogue to the LR transformation — Part 1, *The Computer Journal*, 4 (1961) 265–271. <https://doi.org/10.1093/comjnl/4.3.265>
- [46] J. G. F. Francis, The QR transformation — Part II, *The Computer Journal*, 4 (1962) 332–345. <https://doi.org/10.1093/comjnl/4.4.332>
- [47] R. Freund and N. M. Nachtigal, QMR: A quasi-minimal residual method for non-Hermitian linear systems, *Numerische Mathematik*, 60 (1991), 315–339. 208
- [48] N. Gastinel, *Linear Numerical Analysis*, Kershaw Publishing, London, 1083. 77
- [49] J. F. Grcar, Mathematicians of Gaussian Elimination, *Notices of the AMS*, 58 (2011), 782–792. 47
- [50] A. George and J. W-H Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981. 63
- [51] G. H. Golub, History of numerical linear algebra: A personal view, Stanford, 2007. Available at <http://forum.stanford.edu/events/2007slides/plenary/history-revised-2007-03-19-golub.pdf> 3
- [52] G. H. Golub and W. Kahan, Calculating the singular values and pseudo-inverse of a matrix, *SIAM Journal on Numerical Analysis*, Series B, 2 (1965), 205–224. 188
- [53] G. H. Golub and C. Reinsch, Singular value decomposition and least squares solution, *Numerische Mathematik*, 14 (1970), 403–420. 188
- [54] G. H. Golub and D. P. O'Leary, Some history of the conjugate gradient and Lanczos methods: 1948–1976, *SIAM Review*, 31 (1989), 50–102. 207
- [55] G. H. Golub and H. A. van der Vorst, Eigenvalue computation in the 20th century, *Journal of Computational and Applied Mathematics*, 123 (2000), 35–65. 133
- [56] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The 4th Edition, The Johns Hopkins University Press, Baltimore, MD, 2013. 47, 49, 50, 90, 102, 107, 109, 132, 133, 172, 188, 189, 208, 255
- [57] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, 1997. 255
- [58] A. Greenbaum and L. Gurvits, Max-min properties of matrix factor norms, *SIAM Journal on Scientific Computing*, 15 (1994), 348–358. *SIAM J. Sci. Comput.*, 15 (1994), 348–358. 277, 279
- [59] A. Greenbaum, V. Pták and Z. Strakoš, Any nonincreasing convergence curve is possible for GMRES, *SIAM Journal on Matrix Analysis and Applications*, 17 (1996), 465–469. 278
- [60] A. Greenbaum and Z. Strakoš, Matrices that generate the same Krylov residual spaces, in *Recent Advances in Iterative Methods*, vol. 60 of IMA Vol. Math. Appl., Springer, New York, 1994, pp. 95–118. 278
- [61] J. F. Grcar, Mathematicians of Gaussian Elimination, *Notices of the AMS*, 58 (2011), Number 6, 782–792. 47
- [62] M. Gu and S. C. Eisenstat, A stable algorithm for the rank-1 modification of the symmetric eigenproblem, *SIAM Journal on Matrix Analysis and Applications*, 15 (1994), 1266–1276. 180
- [63] M. Gu and S. C. Eisenstat, A Divide-and-Conquer algorithm for the bidiagonal SVD, *SIAM Journal on Matrix Analysis and Applications*, 16 (1995), 79–92. 182
- [64] M. Gu and S. C. Eisenstat, A Divide-and-Conquer algorithm for the symmetric tridiagonal eigenproblem, *SIAM Journal on Matrix Analysis and Applications*, 16 (1995), 172–191. 175, 182
- [65] I. Gustafsson, A class of first order factorization methods, *BIT*, 18 (1978), 142–156. 288
- [66] M. R. Hestenes and E. L. Stiefel, Methods of conjugate gradients for solving linear systems, *Journal of research of the National Bureau of Standards*, 49 (1952), 2379. 207
- [67] N. J. Higham, Stability of the diagonal pivoting method with partial pivoting, *SIAM Journal on Matrix Analysis and Applications*, (18) 1997, 52–65. 65



- [68] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Second Edition, SIAM, Philadelphia, 2002. 54, 65, 79, 80, 81, 105, 125, 280, 305, 306, 308
- [69] N. J. Higham, Gaussian elimination, *WIREs: Computational Statistics*, 3 (2011), 230–238.
- [70] L. Hogben, *Handbook of Linear Algebra*, 2nd, CRC Press, 2014. 324
- [71] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985. 2nd edition, 2013. 6, 9, 22, 33, 97, 108, 194, 197
- [72] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991. 6, 113, 114, 115, 194
- [73] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964. 207
- [74] V. P. Il'in, *Iterative Incomplete Factorization Methods*, World Scientific, Singapore, 1992. 288
- [75] C. G. J. Jacobi, Concerning an easy process for solving equations occurring in the theory of secular disturbances, *Journal für die reine und angewandte mathematik*, 30 (1846), 51–94.
- [76] W. Joubert, A robust GMRES-based adaptive polynomial preconditioning algorithm for nonsymmetric linear systems, *SIAM Journal on Scientific Computing*, 15 (1994), 427–439. 277, 279
- [77] W. Kahan Numerical Linear Algebra, *Canadian mathematical bulletin*, 9 (1966), 757–801. 77
- [78] T. Kailath and A. H. Sayed, *Fast Reliable Algorithms for Matrices with Structure*, SIAM, Philadelphia, 1999. 70, 72
- [79] P. A. Knight, D. Ruiz and B. Uçar A symmetry preserving algorithm for matrix scaling, *SIAM Journal on Matrix Analysis and Applications*, 35 (2014), 931–955. 82
- [80] D. Kressner, *Numerical Methods for General and Structured Eigenvalue Problems*, Lecture Notes in Computational Sciences and Engineering 46, Springer-Verlag, 2005. 157
- [81] V. N. Kublanovskaya On some algorithms for the solution of the complete eigenvalue problem, *USSR Computational Mathematics and Mathematical Physics*, 3 (1961), 637–657.
- [82] C. Lanczos Solutions of systems of linear equations by minimized iterations, *Journal of research of the National Bureau of Standards*, 49 (1952), 2341. 207
- [83] F. Le Gall, Powers of tensors and fast matrix multiplication, *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation (ISSAC 2014)*, <http://arxiv.org/abs/1401.7714> 323
- [84] R. Lehoucq, *Analysis and Implementation of an Implicitly Restarted Arnoldi Iteration*, Ph.D. thesis, Rice University, Houston, TX, 1995. 296
- [85] N. Levinson, The Wiener RMS (root mean square) error criterion in filter design and prediction, *Journal of Mathematical Physics*, 25 (1946), 261–278. <https://doi.org/10.1002/sapm1946251261> 70, 72
- [86] S. Li, M. Gu and B. N. Parlett, An Improved DQDS Algorithm, *SIAM Journal on Scientific Computing*, 36 (2014), C290–C308. 192
- [87] J. Liesen, Computable convergence bounds for GMRES, *SIAM Journal on Matrix Analysis and Applications*, 21 (2000), 882–903. 278
- [88] J. Liesen and P. Tichý, Convergence analysis of Krylov subspace methods, *GAMM-Mitteilungen*, 27 (2004), 153–173. 278
- [89] Joseph W. H. Liu, A partial pivoting strategy for sparse symmetric matrix decomposition, *ACM Transactions on Mathematical Software*, 13 (1987), 173–182. 65
- [90] J. A. Meijerink and H. A. van der Vorst, An iterative solution method for linear systems of which the coefficient matrix is a symmetric  $M$ -matrix, *Mathematics of Computation*, 31 (1977), 148–162. 288
- [91] E. H. Moore, On the reciprocal of the general algebraic matrix, *Bulletin of the American Mathematical Society*, 26 (1920), 394–395. 121
- [92] L. Neal and G. Poole, A geometric analysis of Gaussian elimination. II, *Linear Algebra and its Applica-*



- tions, 173 (1992), 239–264. 65
- [93] T. A. Oliphant, An extrapolation procedure for solving linear systems, *Quarterly of Applied Mathematics*, 20 (1962), 257–265. 288
- [94] C. C. Paige, M. Rozložník and Z. Strakoš, Modified Gram–Schmidt (MGS), least squares, and backward stability of MGS-GMRES, *SIAM Journal on Matrix Analysis and Applications*, (28) 2006, 264–284. 100, 105
- [95] C. C. Paige and M. A. Saunders, Solution of sparse indefinite systems of linear equations, *SIAM Journal on Numerical Analysis*, 12 (1975), 617–629. 208
- [96] B. N. Parlett, *The Symmetric Eigenvalue Problem*, The 2nd Edition, SIAM, Philadelphia, PA, 1998. 133, 166, 170, 172
- [97] B. N. Parlett, The QR algorithm, *Computing in Science & Engineering*, 2 (2000), 38–42.
- [98] B. N. Parlett and O. Marques, An implementation of the dqds algorithm (positive case), *Linear Algebra and its Applications*, 309 (2000), 217–259. 192
- [99] D. W. Peaceman and H. H. Rachford, Jr., The numerical solution of parabolic and elliptic differential equations, *Journal of the Society for Industrial and Applied Mathematics*, 3 (1955), 28–41. 244
- [100] R. Penrose, A generalized inverse for matrices, *Mathematical Proceedings of the Cambridge Philosophical Society*, 51 (1955), 406–413. 121
- [101] A. D. Polyanin and A. V. Manzhirov, *Handbook of Mathematics for Engineers and Scientists*, Chapman & Hall/CRC, Taylor & Francis Group, 2007. 23
- [102] J. K. Reid, On the method of conjugate gradients for the solution of large sparse systems of linear equations, in *Large Sparse Sets of Linear Equations*, edited by J. K. Reid, p. 231, Academic Press, New York, 1971. 208
- [103] M. Rozložník, G. Shklarski, S. Toledo, Partitioned triangular tridiagonalization, *ACM Transactions on Mathematical Software*, 37(4) (2011), No. 38. 65
- [104] H. Rutishauser, Der Quotienten-Differenzen-Algorithmus, *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 5 (1954), 233–251. <https://link.springer.com/article/10.1007/BF01600331>
- [105] H. Rutishauser, Solution of eigenvalue problems with the LR-transformation, *NBS Applied Mathematics Series*, 49 (1958), 47–81.
- [106] J. Rutter, *A Serial Implementation of Cuppen's Divide and Conquer Algorithm for the Symmetric Eigenvalue Problem*, Master's Thesis, University of California, 1994. 178
- [107] Y. Saad and M. H. Schultz, GMRES: A generalized minimal residual method for solving nonsymmetric linear systems, *SIAM Journal on Scientific & Statistical Computing*, 7 (1986), 856–869. 208, 278
- [108] Y. Saad, *Numerical Methods for Large Eigenvalue Problems: Theory and Algorithms*, Manchester University Press, Manchester, UK, 1992. 296
- [109] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd edition, SIAM, 2003. 255, 286, 288
- [110] Y. Saad and H. A. van der Vorst, Iterative solution of linear systems in the 20th century, *Journal of Computational and Applied Mathematics*, 123 (2000), 1–33. 208
- [111] J. Scott, Scaling and pivoting in an out-of-core sparse direct solver, *ACM Transactions on Mathematical Software*, 37 (2010), Issue 2, Article No. 19. 82
- [112] D. Sorensen, Implicit application of polynomial filters in a  $k$ -step Arnoldi method, *SIAM Journal on Matrix Analysis and Applications*, 13 (1992), 357–385. 296
- [113] G. W. Stewart, *Matrix Algorithms, Vol I: Basic Decomposition*, SIAM, Philadelphia, PA, 1998. 112, 115
- [114] G. W. Stewart, *Matrix Algorithms, Vol II: Eigensystems*, SIAM, Philadelphia, PA, 2001. 133, 157
- [115] M. Stewart, Fast algorithms for structured matrix computations, in *Handbook of Linear Algebra*, 2nd,



- section 62, CRC Press, 2014.
- [116] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990. 79, 194
  - [117] A. Stothers, *On the complexity of matrix multiplication*, Ph.D. Thesis, U. Edinburgh, 2010. 323
  - [118] V. Strassen, Gaussian elimination is not optimal, *Numerische Mathematik*, 13 (1969), 354–356. 321
  - [119] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997. 3, 47, 109, 255, 280, 286, 287
  - [120] L. N. Trefethen, Numerical Analysis, in *Princeton Companion to Mathematics*, Edited by T. Gowers, J. Barrow-Green and I. Leader, Princeton University Press, 2008. 3
  - [121] A. M. Turing, Rounding-off errors in matrix processes, *The Quarterly Journal of Mechanics and Applied Mathematics*, 1 (1948), 287–308. Reprinted in *Collected Works of A. M. Turing: Pure Mathematics*, by J. L. Britton (Ed.), North-Holland, The Netherlands, 1992, with summary and notes (including corrections) 280
  - [122] H. A. van der Vorst, Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems, *SIAM Journal on Scientific and Statistical Computing*, 12 (1992), 631–644. 208
  - [123] John Von Neumann and H. H. Goldstine, Numerical inverting of matrices of high order, *Bulletin of the American Mathematical Society*, 53 (1947), 1021–1099. <http://www.ams.org/journals/bull/1947-53-11/S0002-9904-1947-08909-6/S0002-9904-1947-08909-6.pdf> 2
  - [124] A. J. Wathen, Preconditioning, *Acta Numerica*, (2015), 329–376. 280
  - [125] D. S. Watkins, *The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods*, SIAM, Philadelphia, 2007. 139, 154
  - [126] D. S. Watkins and L. Elsner, Convergence of algorithms of decomposition type for the eigenvalue problem, *Linear Algebra and its Applications*, 143 (1991), 19–47. 139
  - [127] J. W. Watts III, A conjugate gradient truncated direct method for the iterative solution of the reservoir simulation pressure equation, *Society of Petroleum Engineers Journal*, 21 (1981), 345–353. 288
  - [128] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford University, Oxford, 1965. 93, 105, 133, 166
  - [129] V. Williams, Breaking the Coppersmith-Winograd barrier, 2011. <http://theory.stanford.edu/~virgi/matrixmult-f.pdf> 323
  - [130] M. Van Barel, R. Vandebril and P. Van Dooren, Computing the eigenvalues of a companion matrix, 2008. <http://bezout.dm.unipi.it/cortona08/slides/VanBarel.pdf> 158, 159
  - [131] R. S. Varga, Factorizations and normalized iterative methods, in *Boundary Problems in Differential Equations*, edited by R.E. Langer, p. 121, Univ. Wisconsin Press, Madison, 1960. 288
  - [132] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962. 2nd edition, Springer-Verlag, Berlin, 2000. 207, 208, 230
  - [133] D. M. Young, *Iterative Methods for Solving Partial Difference Equations of Elliptic Type*, PhD thesis, Harvard University, 1950. 212
  - [134] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971. 207, 208, 236
  - [135] 北京大学数学系, 高等代数 (第三版), 高等教育出版社, 2003. 11
  - [136] 陈志明, 科学计算: 科技创新的第三种方法, 中国科学院院刊, 27 (2012), 161–166. 2
  - [137] 戴华, refbook 矩阵论, 学出版社, 2001. 6, 223
  - [138] 胡家赣, 线性代数方程组的迭代解法, 科学出版社, 1991. 37
  - [139] 蒋尔雄, 矩阵计算, 科学出版社, 2008. 180
  - [140] 石钟慈, 第三种科学方法 — 计算机时代的科学计算, 清华大学出版社, 暨南大学出版社, 2000. 2, 3





- [141] 孙继广, 矩阵扰动分析, 科学出版社, 北京, 2001. 79, 194
- [142] 魏木生, 李莹, 赵建立, 广义最小二乘问题的理论与计算, 第二版, 科学出版社, 北京, 2020. 86, 121, 123
- [143] 徐树方, 矩阵计算的理论与方法, 北京大学出版社, 北京, 1995. 21, 65, 208
- [144] 徐树方, 钱江, 矩阵计算六讲, 高等教育出版社, 北京, 2011. 157, 159, 188, 189
- [145] 詹兴致, 矩阵论, 高等教育出版社, 北京, 2008. 206
- [146] 张恭庆, 林源渠, 泛函分析讲义, 上册, 北京大学出版社, 北京, 1987. 23
- [147] 张凯院, 徐仲, 数值代数, 第二版, 科技出版社, 2006. 217

