王何宇 副教授

分类号：O241.5　　　　学校代码 10335

浙江大學

ZHEJIANG UNIVERSITY

博士学位论文

论文题目 两类非线性矩阵方程的数值算法研究

姓 吴乙荣

指导教师 黄正达 教授

学科(业) 计算数学

所 系 数学系

提交时间 2014 年 4

王何宇 Heyu Wang

Numerical Methods for Two Classes of Nonlinear Matrix Equations

by

Yirong Wu

Supervisor: Zhengda Huang

A Dissertation Submitted to Zhejiang University in Partial Fulfilment for the Degree of Doctor of Philosophy in Computational Mathematics

Department of Mathematics, Zhejiang University
Hangzhou, Zhejiang, 310027
P. R. China

March, 2016

摘　　要

关键词： Newton 法；Halley 法；

Abstract

Keywords: Newton's Method; Halley's Method;

目　　录

表　　格

插　图

第 1 章　绪论

1.1　研究背景

1.2　迭代法介绍

令 $\mathbb{X}$ 和 $\mathbb{Y}$ 是欧氏空间或一般的 Banach 空间, $\mathbb{D}$ 是 $\mathbb{X}$ 的一个开凸子集, 设 $F:\mathbb{D}\subset\mathbb{X}\to\mathbb{Y}$ 是一个 Fréchet 可导的非线性算子, 考虑如下一般的非线性算子方程：

$$F(x)=0. \tag{1.1}$$

求解非线性算子方程 (1.1) 的近似解是一个重要的数学问题. 有别于线性方程组的情形, 求解非线性算子方程 (1.1) 一般应用迭代方法. 目前, Newton 法是求解非线性算子方程 (1.1) 的最有效方法, 其迭代格式定义为 (初始点 x_0 给定)：

$$x_{k+1}=x_k-F'(x_k)^{-1}F(x_k),\quad k=0,1,2,\ldots. \tag{1.2}$$

用迭代法求解算子方程 (1.1) , 基本的途径是构造一个有效的迭代格式(如 Newton 法 (1.2)), 使得由给定的初始点出发, 逐步逼近到方程 (1.1) 的一个解. 于是, 迭代法的收敛性成为研究的一个核心问题. 一般情况下, 收敛性分析有以下三种类型：

1) 局部收敛性：该类型首先假定方程 (1.1) 存在解 x^* ，再根据 F 在 x^* 的局部条件 (例如，F 在 x^* 是连续可微的) 来研究有关迭代法的收敛性质，其中包括收敛速度、解的唯一性球及 (最重要的) 收敛球半径的最优性. 例如，关于 Newton 法 (1.2)，文献 [? ? ? ? ? ? ?] 分别研究了该迭代法在不同条件下的最优半径及相应局部收敛结果[1].

2) 半局部收敛性：这种类型在不知方程解 x^* 存在的情形下，根据 F 在某一近似初始点 x_0 的局部条件来研究有关迭代法的收敛性质，一般包括收敛

[1] 其中文献 [?] 拓展了文献 [?] 的结果；文献 [?] 的结果是在解析条件下得到的；文献 [?] 的结果是在 Hölder 条件下得到的；而文献 [? ? ?] 分别在不同的更一般的条件下统一了文献 [? ? ?] 的结果.

判据、收敛速度以及以 x_0 为中心的收敛球和解的唯一性球[2].

3) 全局收敛性：有别于前面两种，这种类型研究当 F 满足某些适当的 (全局) 条件时，可以保证取定义域内任意一点作为初始点时都可收敛到方程的某个解 (有解存在时). 同伦延拓法，线性搜索法和信赖域法是常用的全局化方法, 详见文献 [? ? ?].

事实上, 式 (1.2) 只是一种形式记号, 当应用 Newton 法对具体的非线性方程组进行求解时, 实际是求解如下线性方程组：

$$F'(x_k)\Delta x_k = -F(x_k), \quad k = 0, 1, 2, \ldots. \tag{1.3}$$

求得上述方程组的精确解 Δx_k 后, 由 $x_{k+1} = x_k + \Delta x_k$ 进行迭代修正. 从数值计算的角度看, Newton 法 (1.2) 具有收敛快的优点, 在实际计算时每一步运算只与前一步有关, 误差不传播且是自校正的. 因此, 在理论和实际应用上都是一种重要的方法. 但是, Newton 法亦有其不足, 例如, 设 $\mathbb{X} = \mathbb{Y} = \mathbb{C}^n$, 那么在每步计算中都要计算 n^2 个分量偏导数值和 n 个分量函数值, 并求一次矩阵的逆, 运算量较大. 为此, Newton 法有不少改进的算法.

这些修正的 Newton 法统称为 Newton 型迭代法, 针对不同的问题背景, 应用相应的 Newton 型迭代法. 常见的有以下几种类型 (详细论述可见文献 [? ? ?]):

- 一般的 Newton 法 (初始点 x_0 给定)：

$$F'(x_k)\Delta x_k = -F(x_k),\ x_{k+1} = x_k + \Delta x_k, \quad k = 0, 1, \ldots.$$

- 简化 Newton 法：这种变形的 Newton 法将每步计算 $F'(x_k)$ 改为固定的 $F'(x_0)$ (初始点 x_0 给定)：

$$F'(x_0)\Delta x_k = -F(x_k),\ x_{k+1} = x_k + \Delta x_k, \quad k = 0, 1, \ldots.$$

这样, 每步只需要计算 n 个分量函数, 但这种迭代法只有线性收敛.

[2] 这里的收敛球和解的唯一性球与局部收敛分析中的相应概念不同，文献 [?] 给出了局部收敛分析的有关定义.

- Newton 类法：当所要处理的方程组维数较大时, 很难求解 Newton 方程 (1.3) 的精确解 Δx_k, 进而无法得到精确的 Jacobi 矩阵 $F'(x_{k+1})$. 为了能够应用 Newton 法较好的处理这种情况, 得到了如下的一种变形 Newton 法 (初始点 x_0 给定):

  $$M(x_k)\Delta x_k = -F(x_k),\ x_{k+1} = x_k + \Delta x_k, \quad k = 0, 1, \ldots, \tag{1.4}$$

  其中 $M(x_k)$ 是近似于 $F'(x_k)$ 的矩阵.

- 非精确 Newton 法：同样考虑大规模方程组情形, 相比于 Newton 类法用一个近似的 Jacobi 矩阵来代替 $F'(x_k)$, 若考虑在求解 Newton 方程 (1.3) 时, 不去求其精确解而只需要求满足某种条件的近似解来作为迭代修正, 这样便得到如下的变形 Newton 法 (初始点 x_0 给定):

  $$F'(x_k)\Delta x_k = -F(x_k) + r_k,\ x_{k+1} = x_k + \Delta x_k, \quad k = 0, 1, \ldots, \tag{1.5}$$

  其中 $r_k \in \mathbb{Y}$ 一般应满足 $\|r_k\|/\|F(x_k)\| \leqslant \eta_k,\ k = 0, 1, \ldots,\ \{\eta_k\}$ 满足 $0 \leqslant \eta_k < 1$, 可能与 x_k 有关, 为控制序列, 用来控制求方程 (1.3) 的解的精确程度. 显然令 $\eta_k \equiv 0$ 时得到一般的 Newton 法.

- 拟 Newton 法：Newton 法的主要缺点之一是每步都要计算导数 $F'(x)$ 的值, 当分量函数较复杂时计算很不方便, 拟 Newton 法是针对这一缺点提出的算法, 其核心是用通过计算函数值来代替导数以避免求导:

  $$J_k\Delta x_k = -F(x_k),\ J_{k+1} = J_k + \Delta J_k, \quad k = 0, 1, \ldots,$$

  其中 J_k 是近似于 $F'(x_k)$ 的矩阵. 不同的 ΔJ_k 选择，可以得到不同的迭代法. 例如，当取$\Delta J_k = F(x_{k+1})\Delta x_k{}^{\mathrm{T}}/(\Delta x_k{}^{\mathrm{T}}\Delta x_k)$，则得到 Broyden 法。

- Gauss-Newton 法：这种变形的 Newton 法主要应用于求解非线性最小二乘 (约束/无约束) 问题, 迭代格式为:

  $$\|F'(x_k)\Delta x_k + F(x_k)\| = \min,\ x_{k+1} = x_k + \Delta x_k, \quad k = 0, 1, \ldots. \tag{1.6}$$

除上述几类重要的 Newton 型迭代法外, 还有几类高阶的 Newton 变形法. 如 Halley 法, 迭代格式为:

$$x_{k+1} = x_k - [\mathbf{I} - L_F(x_k)]^{-1}F'(x_k)^{-1}F(x_k), \quad k = 0, 1, \ldots, \tag{1.7}$$

及 Euler/Chebyshev 法, 迭代格式为:

$$x_{k+1} = x_k - [\mathbf{I} + L_F(x_k)]F'(x_k)^{-1}F(x_k), \quad k = 0, 1, \ldots, \tag{1.8}$$

其中 $L_F(x) = \frac{1}{2}F'(x)^{-1}F''(x)F'(x)^{-1}F(x)$.

为统一研究这两种迭代法的收敛性, Gutiérrez 和 Hernández 在文献 [?] 中提出了如下的 Halley-Euler 迭代族法:

$$x_{\alpha,k+1} = x_{\alpha,k} - \left\{\mathbf{I} + \frac{1}{2}L_F(x_{\alpha,k})[\mathbf{I} - \alpha L_F(x_{\alpha,k})]^{-1}\right\} F'(x_{\alpha,k})^{-1}F(x_{\alpha,k}), \quad k = 0, 1, \ldots, \tag{1.9}$$

其中 $\alpha \in [0, 1]$ 及 $L_F(x) = F'(x)^{-1}F''(x)F'(x)^{-1}F(x)$. 显然, $\alpha = 0$ 时 Halley-Euler 迭代族法 (1.9) 变为 Euler 法 (1.8); 当 $\alpha = \frac{1}{2}$ 时 Halley-Euler 迭代族法 (1.9) 变为 Halley 法 (1.7); 当 $\alpha = 1$ 时, Halley-Euler 迭代族法 (1.9) 变为快速 Halley 法.

关于 Newton 法 (1.2) 收敛的性质研究主要有以下两个方向:

1) Kantorovich 型收敛理论：理论上，Newton 法收敛性的一个最重要收敛结果是被称为 Newton-Kantorovich 半局部收敛定理 [?]. 该定理在理论和应用上都是相当重要的，它是解方程算法现代研究的起点. 大量的收敛结果都是基于所谓的 Kantorovich 型条件而得到的, 例如，[? ? ? ? ? ? ? ? ? ? ?].

2) Smale 点估计理论：该理论是由 Smale 于 1986 年提出，由 $\alpha-$理论和 $\gamma-$理论组成。在 $\alpha-$理论中，假设 F 在初始点 x_0 是解析的，给出了基于如下三个不变量的收敛判据 [?]:

$$\begin{cases} \alpha(F, x_0) = \beta(F, x_0)\gamma(F, x_0), \\ \beta(F, x_0) = \|F'(x_0)^{-1}F(x_0)\|, \\ \gamma(F, x_0) = \sup\limits_{k\geqslant 2} \left\|\dfrac{1}{k!}F'(x_0)^{-1}F^{(k)}(x_0)^{-1}\right\|^{\frac{1}{k-1}}. \end{cases}$$

而 $\gamma-$理论则研究了算子 F 在解析条件下的局部收敛性。定理 1.1 称为 $\gamma-$定理。王兴华等人改进并完善了 Smale 点估计理论 (见文献 [?] 及其所列文献). 值得指出的是，王兴华引入了 γ 条件并在此基础上系统建立了 Smale 原先在解析条件下的全部结果(见文献 [?] 及其所列文献).

定理 1.1 ([? , $\gamma-$定理]). *设 $F:\mathbb{C}^n\to\mathbb{C}^n$ 是解析的。设 ζ 为 F 的一个零点且 $F'(\zeta)^{-1}$ 存在。若 $z\in\mathbb{C}^n$ 满足*

$$\|z-\zeta\|<\frac{3-\sqrt{7}}{2\gamma(f,\zeta)}, \tag{1.10}$$

则 z 为 F 关于 ζ 的一个近似零点，即

$$\|\zeta-z_k\|\leq\left(\frac{1}{2}\right)^{2^k-1}\|\zeta-z\|,\quad k=0,1,2,\ldots, \tag{1.11}$$

其中 $\{z_k\}$ 为 Newton 法 (1.2) 以初始点$z_0=z$ 进行迭代所产生的序列。

对于其他的 Newton 型迭代法亦有很多研究结果是建立上述两种条件下而得到的，例如，关于非精确 Newton 法 (1.5) 研究有 [? ? ? ? ? ?]，关于 Gauss-Newton 法 (1.6) 的研究有 [? ? ? ? ? ?]。

Halley 法 (1.7) 和 Euler 法 (1.8) 是求解非线性方程 (1.1) 的 Newton 法的两种重要高阶的迭代法, 而 Newton-Kantorovich 型条件是研究 Newton 法收敛性的重要方向, 因而, 对于 Halley 法和 Euler 法的收敛性, 亦有很多收敛结果是在 Newton-Kantorovich 型条件下得到的, 如 [? ? ? ? ? ?].

对于一个迭代法，研究其收敛速度对实际计算是重要的. 为刻画收敛速度，本文引入 Q 收敛阶和 R 收敛阶. 对于二者关系的详细论述可见文献 [? ? ? ?].

定义 1.1 (Q 收敛阶). 设序列 $\{x_k\}$ 收敛到 x^* . 如果存在 $q\geqslant 1$ 及常数 $c\geqslant 0$ 和 $N\geqslant 0$ 使得当 $k\geqslant N$ 时有 $\|x^*-x_{k+1}\|\leqslant c\|x^*-x_k\|^q$，则称序列 $\{x_n\}$ 具有 Q 收敛阶至少为 q. 特别地，当 $q=2$ 时称为 (至少) Q 平方收敛，$q=3$ 时称为 (至少) Q 立方收敛.

定义 1.2 (R 收敛阶). 设序列 $\{x_k\}$ 收敛到 x^* . 如果存在 $\tau>1$ 及常数 $c\in(0,\infty)$ 和 $\theta\in(0,1)$ 使得对所有 $n\in\mathbb{N}$ 有 $\|x_k-x^*\|\leqslant c\theta^{\tau^k}$，则称 $\{x_k\}$ 具有 R 收敛阶至少为 τ.

1.3　矩阵函数

矩阵函数 $f:\mathbb{C}^{n\times n}\to\mathbb{C}^{n\times n}$ 有多种等价定义, 下面所给出的定义是基于 Jordan 典范形而得到的.

对于任意的矩阵 $A \in \mathbb{C}^{n\times n}$, 设其 Jordan 典范形为

$$Z^{-1}AZ = J = \operatorname{diag}(J_1, J_2, \ldots, J_s), \tag{1.12}$$

其中 $Z \in \mathbb{C}^{n\times n}$ 是非奇异的,

$$J_k = J_k(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix} \in \mathbb{C}^{m_k\times m_k}, \quad m_1 + m_2 + \cdots + m_s = n.$$

设 $\lambda_1, \ldots, \lambda_r$ 为 A 的所有不同的特征值, n_ℓ为属于特征值 λ_ℓ 的 Jordan 块的阶数, 称为 λ_ℓ 的次数.

定义 1.3. 对于任意函数 f, 如果

$$f^{(j)}(\lambda_\ell), \quad j = 0, \ldots, n_\ell - 1, \ \ \ell = 1, \ldots, r$$

的值都存在, 则称 f 在 A 的谱上有定义.

定义 1.4. 设函数 f 在矩阵 $A \in \mathbb{C}^{n\times n}$ 的谱上有定义, 且有 Jordan 典范形 (1.12), 则有

$$f(A) := Zf(J)Z^{-1} = Z\operatorname{diag}(f(J_k))Z^{-1}, \tag{1.13}$$

其中

$$f(J_k) = \begin{bmatrix} f(\lambda_k) & f'(\lambda_k) & \cdots & \dfrac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ & f(\lambda_k) & \ddots & \vdots \\ & & \ddots & f'(\lambda_k) \\ & & & f(\lambda_k) \end{bmatrix}.$$

下面的定理给出了矩阵函数的若干基本性质, 更多的其他性质参见 [**?**] 或 [**?**].

定理 1.2 ([**?** , 定理 1.13])**.** *设函数 f 在矩阵 $A \in \mathbb{C}^{n\times n}$ 的谱上有定义, 则有如下性质:*

(i) *$f(A)$ 与 A 可交换, 即 $f(A)A = Af(A)$;*

(ii) $f(A^{\mathrm{T}}) = f(A)^{\mathrm{T}}$;

(iii) $f(XAX^{-1}) = Xf(A)X^{-1}$;

(iv) $f(A)$ 的全部特征值分别为 $f(\lambda_\ell)$, 其中 $\lambda_\ell, \ell = 1, \ldots, n$ 为 A 的特征值;

(v) 如果 X 与 A 可交换, 那么 X 与 $f(A)$ 可交换.

定理 1.3 ([? , 定理 1.15]). 设函数 f, g 在矩阵 $A \in \mathbb{C}^{n\times n}$ 的谱上有定义.

(i) 若 $h(t) = f(t) + g(t)$, 则 $h(A) = f(A) + g(A)$;

(ii) 若 $h(t) = f(t)g(t)$, 则 $h(A) = f(A)g(A)$.

1.3.1　矩阵平方根

对于任意给定的矩阵 $A \in \mathbb{C}^{n\times n}$, 若存在矩阵 $X \in \mathbb{C}^{n\times n}$ 使得 $X^2 = A$ 成立, 则称 X 为 A 的一个平方根. 下面的定理给出了矩阵平方根存在性的充要条件.

定理 1.4 ([?]). 矩阵 $A \in \mathbb{C}^{n\times n}$ 存在一个平方根的充要条件是如下定义的递增的整数数列 $d_1, d_2, \ldots$ 没有两项都是相同的奇数:

$$d_\ell = \dim(\mathrm{null}(A^\ell)) - \dim(\mathrm{null}(A^{\ell-1})), \quad \ell = 1, 2, \ldots.$$

定理 1.5 (矩阵平方根的分类, [?]). 设非奇异矩阵 $A \in \mathbb{C}^{n\times n}$ 的 *Jordan* 典范形由 (1.12) 给出, 其所有不同的特征值数为 r. 如果 $r \le s$, 那么 A 有 2^r 准平方根, 由如下式子给出:

$$X_j = Z\mathrm{diag}(L_1^{(j_1)}, L_2^{(j_2)}, \ldots, L_s^{j_s})Z^{-1}, \quad j = 1, 2, \ldots, 2^r,$$

其中 $j_k = 1$ 或 2, 当 $\lambda_\ell = \lambda_k$ 时 $j_\ell = j_k$, $k = 1, 2, \ldots, s$. 特别地, 如果 $r < s$, 那么 A 存在非准平方根, 其形式为

$$X_j(U) = ZU\mathrm{diag}(L_1^{(j_1)}, L_2^{(j_2)}, \ldots, L_s^{j_s})U^{-1}Z^{-1}, \quad j = 2^r + 1, \ldots, 2^s,$$

其中 $j_k = 1$ 或 2, U 为任意的与 J 可交换的非奇异矩阵, 且对于每一个 j, 存在 ℓ 和 k, 使得当 $j_\ell \ne j_k$ 时 $\lambda_\ell = \lambda_k$.

定理 1.6 ([?]). 设矩阵 $A \in \mathbb{C}^{n\times n}$ 的所有特征值都不属于 $\mathbb{R}^- := (-\infty, 0]$. 则存在唯一的 A 的准平方根 X, 其所有特征值都属于复平面右半部分 $\{z \in \mathbb{C} : \operatorname{Re} z > 0\}$. 此时称 X 为矩阵 A 的主平方根, 记为 $X := A^{1/2}$. 若 A 是实矩阵, 则 $A^{1/2}$ 也是实矩阵.

下面考虑对非奇异矩阵 $A \in \mathbb{C}^{n\times n}$ 的主平方根 $A^{1/2}$ 的计算问题. 记 $f(A)$ 为 A 的任一准平方根, 并设 A 的 Schur 分解为 $A = QTQ^*$, 其中 Q 为酉矩阵而 T 为上三角矩阵. 由定理 1.2 知 $f(A) = Qf(T)Q^*$, 故为了计算矩阵 A 的主平方根, 只要计算上三角矩阵 T 的主平方根 $U = f(T)$ 即可. 设 $U = [u_{ij}]_{n\times n}, T = [t_{ij}]_{n\times n}$, 由 $U^2 = T$ 可得

$$\begin{aligned} u_{ii}^2 &= t_{ii}, \quad i = 1, 2, \ldots, n, \\ (u_{ii} + u_{jj})u_{ij} &= t_{ij} - \sum_{k=i+1}^{j-1} u_{ik}u_{kj}, \quad j > i. \end{aligned}$$

于是, 有如下计算非奇异矩阵平方根的算法, 该算法由 Björck & Hammarling 于 [?] 得到.

算法 1.1 计算矩阵平方根的 Schur 法 [?]

给定非奇异矩阵 $A \in \mathbb{C}^{n\times n}$, 本算法通过 Schur 分解来计算 A 的主平方根 $A^{1/2}$.

1. 计算矩阵 A 的 Schur 分解 $A = QRQ^*$;
2. 计算矩阵 U 各对角元素的主平方根 $u_{ii} = t_{ii}^{1/2}$, $i = 1, \ldots, n$;
3. 依次计算矩阵 U 的非对角元:
$$u_{ij} = \frac{t_{ij} - \displaystyle\sum_{k=i+1}^{j-1} u_{ik}u_{kj}}{u_{ii} + u_{jj}}, \quad j = 2, 3, \ldots, n,\ i = j-1, j-2, \ldots, 1;$$
4. 计算 $X = QUQ^*$.

算法 1.1 的总计算量为 $28\frac{1}{3}n^3$ flops, 其中 Schur 分解的计算量为 $25n^3$ flops, U 的计算量为 $\frac{1}{3}n^3$ flops, X 的计算量为 $3n^3$ flops.

当 A 是实矩阵时, Higham [?] 推广了算法 1.1 而得到了如下算法.

算法 1.2 计算矩阵平方根的实 Schur 法 [**?**]

给定矩阵 $A \in \mathbb{R}^{n\times n}$, 其所有特征值都不属于 $\mathbb{R}^- := (-\infty, 0]$. 本算法通过实 Schur 分解来计算 A 的主平方根 $A^{1/2}$.

1. 计算矩阵 A 的实 Schur 分解 $A = QRQ^{\mathrm{T}}$, 其中 R 是 $m \times m$ 的块矩阵.

2. 计算矩阵 U 各对角块的主平方根：当 $R_{ii} = [r_{ij}]_{1\times 1}$ 时, $U_{ii} = R_{ii}^{1/2}$；当 $R_{ii} = [r_{ij}]_{2\times 2}$ 时,

$$U_{ii} = \begin{bmatrix} \alpha + \frac{1}{4\alpha}(r_{11} - r_{22}) & \frac{1}{2\alpha} r_{12} \\ \frac{1}{2\alpha} r_{21} & \alpha - \frac{1}{4\alpha}(r_{11} - r_{22}) \end{bmatrix},$$

其中

$$\alpha = \begin{cases} \left(\dfrac{|\theta| + (\theta^2 + \mu^2)^{1/2}}{2} \right)^{1/2}, & \theta \geq 0, \\ \dfrac{\mu}{2\left(\dfrac{|\theta| + (\theta^2 + \mu^2)^{1/2}}{2} \right)^{1/2}}, & \theta < 0, \end{cases}$$

$$\theta = \frac{r_{11} + r_{12}}{2}, \quad \mu = \frac{(-(r_{11} - r_{22})^2 - 4r_{21}r_{22})^{1/2}}{2}.$$

3. 依次通过计算如下的方程而得到矩阵 U 的非对角块 U_{ij}：

$$U_{ii}U_{ij} + U_{ij}U_{jj} = R_{ij} - \sum_{k=i+1}^{j-1} U_{ik}U_{kj}, \quad j = 2, 3, \ldots, m, \ i = j-1, j-2, \ldots, 1.$$

4. 计算 $X = QUQ^{\mathrm{T}}$.

下面考虑应用 Newton 法来计算 $X^2 = A$. 设 Y 为该矩阵方程的一个近似解, 并记 $X = Y + E$, 则

$$A = (Y + E)^2 = Y^2 + YE + EY + E^2.$$

去掉上式中的 E^2 项后即可得如下的 Newton 法 (X_0 给定)：

$$X_{k+1} = X_k + E_k, \quad k = 0, 1, 2, \ldots, \tag{1.14}$$

其中 E_k 为如下 Sylvesterh 方程的解：

$$X_kE_k + E_kX_k = A - X_k^2.$$

一般地, 通过上述 Newton 法来计算矩阵的平方根所需要的计算成本比算法 1.1 或 1.2 要高很多. 但是, 下面的结果可以改进这个不足.

引理 1.1 ([? , 引理 6.8]). *假设 Newton 法 (1.14) 的初始点 X_0 与矩阵 A 可交换, 且所产生的序列 $\{X_k\}$ 是有定义的. 则对于所有的 $k \geq 1$, X_k 与 A 都是可交换的, 且此时 Newton 法的迭代格式为：*

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}A). \tag{1.15}$$

Higham 在文献 [?] 中得到了 Newton 法 (1.15) 在初始点取为 $X_0 = A$ 时的二阶收敛结果. 但需要指出的是, Newton 法 (1.15) 的数值稳定性较差 (详细分析见 [? , 6.4 节]). 为此, Denman & Beavers [?] 给出了Newton 法 (1.15) 的一个对偶形式：

$$\begin{cases} X_{k+1} = \frac{1}{2}(X_k + Y_k^{-1}), & X_0 = A, \\ Y_{k+1} = \frac{1}{2}(Y_k + X_k^{-1}), & Y_0 = \mathrm{I}. \end{cases} \tag{1.16}$$

注意到, (1.16) 是数值稳定的 (详见 [? , 6.4 节]), 且保持了 Newton 法的二阶收敛性. 特别地, 这种处理方法将在计算矩阵 $p\,(> 2)$ 次根中起到重要作用.

Note that Newton's method (2.12) and Halley's method (3.84) are special cases as (dual) Padé family of iterations which have recently received particular interest for computing both the principal pth root of a complex number and a matrix, see for example [? ? ? ?].

1.3.2　矩阵 p 次根

给定矩阵$A \in \mathbb{C}^{n\times n}$及任意的整数$p \geq 2$，如果存在矩阵$X \in \mathbb{C}^{n\times n}$使得$X^p = A$，那么称$X$为$A$的一个$p$次根。

Given a square matrix , a matrix is called a th root of A if for any integer . If A has no eigenvalues on $\mathbb{R}^-$, the closed negative real axis, there exists a unique principal pth root of A, denoted by $A^{1/p}$, which in turn has eigenvalues in the segment $\{z : -\pi/p < \arg(z) < \pi/p\}$ [? , Theorem 7.2].

定理 1.7 (矩阵 p 次根的存在性, [?]). 定义如下的整数列 $\{d_k\}$:

$$d_k = \dim(\mathrm{null}(A^k)) - \dim(\mathrm{null}(A^{k-1})), \quad k = 1, 2, \ldots.$$

矩阵 $A \in \mathbb{C}^{n\times n}$ 存在 p 次根的充要条件是对于任一整数 $\nu \geq 0$, 数列 $\{d_k\}$ 中至多只有一个元素处于 $p\nu$ 与 $p(\nu+1)$ 之间.

定理 1.8 (矩阵 p 次根的分类 [?]). 设非奇异矩阵 $A \in \mathbb{C}^{n\times n}$ 的 *Jordan* 典范形由 (1.12) 给出, 其所有不同的特征值数为 r. 如果 $r \leq s$, 那么 A 存在 p^r 个 p 次根, 由下式给出:

$$X_j = Z\mathrm{diag}(f_{j_1}(J_1), f_{j_2}(J_2), \ldots, f_{j_s}(J_s))Z^{-1}, \quad j = 1, 2, \ldots, p^r,$$

其中 $j_k \in \{1, 2, \ldots, p\}, k = 1, 2, \ldots, r$, 当 $\lambda_i = \lambda_k$ 时有 $j_i = j_k$. 特别地, 若$r < s$, 则 A 存在非 A 的函数的 p 次根, 由下式给出:

$$X_j(U) = ZU\mathrm{diag}(f_{j_1}(J_1), f_{j_2}(J_2), \ldots, f_{j_s}(J_s)), \quad j = p^r + 1, \ldots, p^s,$$

其中 $j_k \in \{1, 2, \ldots, p\}, k = 1, 2, \ldots, r$, U为任意与 J 可交换的非奇异矩阵, 对于每一个 j, 存在 i 和 k, 使得当 $\lambda_i = \lambda_k$ 时仍有 $j_i \neq j_k$.

定理 1.9 ([? , 定理 7.2]). 设矩阵 $A \in \mathbb{C}^{n\times n}$ 没有属于 $\mathbb{R}^- := (-\infty, 0]$ 的特征值, 则 A 存在唯一的 p 次根 X, 其所有的特征值均属于集合

$$\left\{ z \in \mathbb{C} : -\frac{\pi}{p} < \arg(z) < \frac{\pi}{p} \right\}.$$

此时, 称 X 为矩阵 A 的主 p 次根, 并记为 $X = A^{1/p}$. 若 A 是实矩阵, 则 $A^{1/p}$ 亦是实矩阵.

类似于矩阵平方根的情形, 计算一个给定矩阵的主 p 次根通常有直接法和迭代法两种途径. 关于直接法, Smith [?] 将算法 1.2 推广至主 p 次根的情形.

关于迭代法, 首先考虑 Newton 法. 类似于矩阵平方根的情形, 给定矩阵 $A \in \mathbb{C}^{n\times n}$, 应用 Newton 法对矩阵方程 $X^p - A = 0$, $(p > 2)$ 进行求解时, 可通过迭代格式 (1.14) 实现, 此时 E_k 通过如下的广义 Sylvester 方程得到:

$$\sum_{\ell=1}^{p} X_k^{p-\ell} E_k X_k^{\ell-1} = A - X_k^p.$$

特别地, 同矩阵平方根的情形一样, 当初始点 X_0 与 A 可交换时, 可得知对任意的 $k \geq 1$, X_k 均与 A 可交换. 于是得到如下计算矩阵主 p 次根的简化 Newton 法：

$$X_{k+1} = \frac{1}{p}[(p-1)X_k + X_k^{1-p}A], \quad X_0A = AX_0. \tag{1.17}$$

显然, 当初始点取 $X_0 = A$ 且 A 为正定矩阵时, Hoskins [?] 证明了 Newton 法 (1.17) 是二阶收敛的. 进一步, 当初始点取 $X_0 = A$ 且 A 为一般的矩阵时, Smith [?] 证明了 Newton 法 (1.17) 仍然是二阶收敛的. 当初始点取 $X_0 = \mathrm{I}$ 时，Iannazzo [?] 得到了如下的收敛性结果：

定理 1.10 ([?]). *给定矩阵 $A \in \mathbb{C}^{n\times n}$，设其谱为 $\sigma(A)$。若 $\sigma(A) \subset \mathcal{E}_1$, 其中*

$$\mathcal{E}_1 := \{z \in \mathbb{C} : \mathrm{Re}\, z > 0, |z| \leq 1\}, \tag{1.18}$$

则以 $X_0 = \mathrm{I}$ 为初始点的 Newton 法 (1.17) 所产生的矩阵序列 $\{X_k\}$ 收敛于矩阵 A 的主 p 次根 $A^{1/p}$.

之后，Iannazzo 在 [?] 得到了一个新的收敛域：

定理 1.11 ([?]). *给定矩阵 $A \in \mathbb{C}^{n\times n}$，设其谱为 $\sigma(A)$。若 $\sigma(A) \subset \mathcal{E}_2$, 其中*

$$\mathcal{E}_2 := \left\{z \in \mathbb{C} : |z| \leq 2, |\arg(z)| < \frac{\pi}{4}\right\}, \tag{1.19}$$

则以 $X_0 = \mathrm{I}$ 为初始点的 Newton 法 (1.17) 所产生的矩阵序列 $\{X_k\}$ 收敛于矩阵 A 的主 p 次根 $A^{1/p}$.

最近，Guo [?] 进一步得到了一个更好的收敛域：

定理 1.12 ([?]). *给定矩阵 $A \in \mathbb{C}^{n\times n}$，设其谱为 $\sigma(A)$。若 $\sigma(A) \subset \mathcal{E}_3$ 且零特征值 (若存在) 是半单的，其中*

$$\mathcal{E}_3 := \{z \in \mathbb{C} : |z-1| \leq 1\}, \tag{1.20}$$

则以 $X_0 = \mathrm{I}$ 为初始点的 Newton 法 (1.17) 所产生的矩阵序列 $\{X_k\}$ 收敛于矩阵 A 的主 p 次根 $A^{1/p}$, 且是二阶收敛的.

然而, 在实际计算中 Newton 法 (1.17) 的数值稳定性并不好, 详细的稳定性分析可见 [? , 3.2 节]. 于是，基于 Denman & Beavers 在计算矩阵平方根时所给出的具有数值稳定性的 Newton 法 (即对偶 Newton 法 (1.16)), Iannazzo [?]提出了如下的 Newton 迭代格式来计算矩阵主 p 次根：

$$\begin{cases} X_{k+1} = X_k \left(\dfrac{(p-1)\mathrm{I} + N_k}{p} \right), & X_0 = \mathrm{I}, \\ N_{k+1} = \left(\dfrac{(p-1)\mathrm{I} + N_k}{p} \right)^{-p} N_k, & N_0 = A. \end{cases} \tag{1.21}$$

称 (1.21) 为对偶 Newton 法，该迭代格式的一个优点是具有很好的数值稳定性。显然，当 $N_k \to \mathrm{I}$ 时 $X_k \to A^{1/p}$。在 [?] 中，Iannazzo 给出了计算矩阵主 p 次根的算法 1.3：

算法 1.3 计算矩阵主 p 次根的 Schur-Newton 法 [? , 算法 3]

给定矩阵 $A \in \mathbb{R}^{n\times n}$, 其所有特征值都不属于 $\mathbb{R}^- := (-\infty, 0]$. 给定整数 $p \geq 2$, 则存在整数 $k_0 \geq 0$ 及奇数 q 使得 $p = 2^{k_0} q$. 本算法通过实 Schur 分解和对偶 Newton 法 (1.21) 来计算 A 的主 p 次根 $A^{1/p}$.

1. 计算矩阵 A 的实 Schur 分解 $A = QRQ^{\mathrm{T}}$.
2. 若 $q = 1$，令 $k_1 = k_0$; 若 $q \neq 1$，则选取 $k_1 \geq k_0$ 使得存在正数 s 使任意 A 的特征值 λ 满足
$$s\lambda^{1/2^{k_1}} \in \left\{ z \in \mathbb{C} : \left| z - \frac{6}{5} \right| \leq \frac{3}{4} \right\}.$$
3. 通过算法 1.2 计算 $B = R^{1/2^{k_1}}$.
4. 若 $q = 1$, 则令 $X = QBQ^{\mathrm{T}}$；若 $q \neq 1$, 则通过对偶 Newton 法 (1.21) 来计算 $C = (B/s)^{1/q}$ 并令 $X = Q(Cs^{1/q})^{2^{k_1-k_0}}Q^{\mathrm{T}}$.

Guo & Higham [?] 给出了一种含参数的对偶 Newton 法：

$$\begin{cases} X_{k+1} = \left(\dfrac{(p+1)\mathrm{I} - N_k}{p} \right)^{-1} X_k, & X_0 = c\mathrm{I}, \\ N_{k+1} = \left(\dfrac{(p+1)\mathrm{I} - N_k}{p} \right)^{p} N_k, & N_0 = \dfrac{1}{c^p} A. \end{cases} \tag{1.22}$$

显然，当 $N_k \to \mathrm{I}$ 时 $X_k \to A^{1/p}$。此外，也给出了计算矩阵主 p 次根的算法 1.4：

算法 1.4 计算矩阵主 p 次根的含参数 Schur-Newton 法 [? , 算法 3.3]

给定矩阵 $A \in \mathbb{R}^{n\times n}$, 其所有特征值都不属于 $\mathbb{R}^- := (-\infty, 0]$. 给定整数 $p \geq 2$, 则存在整数 $k_0 \geq 0$ 及奇数 q 使得 $p = 2^{k_0}q$. 本算法通过实 Schur 分解和含参数 Newton 法 (1.22) 来计算 A 的主 p 次根 $A^{1/p}$.

1. 计算矩阵 A 的实 Schur 分解 $A = QRQ^{\mathrm{T}}$.
2. 若 $q = 1$，令 $k_1 = k_0$; 若 $q \neq 1$，则选取 $k_1 \geq k_0$ 使得$|\lambda_1/\lambda_n|^{1/2^{k_1}} \leq 2$, 其中 $\lambda_1, \ldots, \lambda_n$ 为 A 的特征值且满足 $|\lambda_n| \leq \cdots \leq |\lambda_1|$, 当 λ_ℓ 不全是实数时，重新选取 k_1 使得对任意的 $\ell \in \{1, 2, \ldots, n\}$ 都有
$$\arg(\lambda_\ell^{1/2^{k_1}}) \in \left(-\frac{\pi}{8}, \frac{\pi}{8}\right).$$
3. 通过算法 1.2 计算 $B = R^{1/2^{k_1}}$.
4. 若 $q = 1$, 则令 $X = QBQ^{\mathrm{T}}$；若 $q \neq 1$, 则先选取参数 c, 再通过含参数对偶 Newton 法 (1.22) 来计算 $C = B^{1/q}$ 并令 $X = QC^{2^{k_1-k_0}}Q^{\mathrm{T}}$.

Halley 法是计算矩阵主 p 次根的另一种重要的迭代法。类似于 Newton 法，若初始点 X_0 与矩阵 A 可交换，则可得如下的计算矩阵主 p 次根的简化 Halley 法：

$$X_{k+1} = X_k\left((p+1)X_k^p + (p-1)A\right)^{-1}\left((p-1)X_k^p + (p+1)A\right), \quad AX_0 = X_0A. \tag{1.23}$$

特别地，初始点取为 $X_0 = \mathrm{I}$ 是研究时主要考虑的情形，如 [? ? ?]。关于 Halley (1.23) 法的收敛性，Iannazzo [?] 得到了如下的结果：

定理 1.13 ([?]). *给定矩阵 $A \in \mathbb{C}^{n\times n}$，设其谱为 $\sigma(A)$。若 $\sigma(A) \subset \{z \in \mathbb{C} : \mathrm{Re}\, z > 0\}$, 则以 $X_0 = \mathrm{I}$ 为初始点的 Halley 法 (1.23) 所产生的矩阵序列 $\{X_k\}$ 收敛于矩阵 A 的主 p 次根 $A^{1/p}$.*

Guo 在 [?] 中进一步证明了当 $\sigma(A) \subset \{z \in \mathbb{C} : \mathrm{Re}\, z > 0\}$ 时 Halley 法 (1.23) 是三阶收敛的。应用在[?]中对Newton法稳定性的分析方法可知，Halley 法 (1.23) 同样在数值计算中是不稳定的。故类似于 Newton 法，可引入如下的

具有数值稳定的对偶形式的 Halley 法：

$$\begin{cases} X_{k+1} = X_k\left((p+1)\mathrm{I} + (p-1)N_k\right)^{-1}\left((p-1)\mathrm{I} + (p+1)N_k\right), & X_0 = \mathrm{I}, \\ N_{k+1} = N_k\left((p+1)\mathrm{I} + (p-1)N_k\right)^{-1}\left((p-1)\mathrm{I} + (p+1)N_k\right), & N_0 = A. \end{cases} \tag{1.24}$$

显然，当 $N_k \to \mathrm{I}$ 时 $X_k \to A^{1/p}$。算法 1.5 给出了应用对偶 Halley 法 (1.24) 来计算 A 的主 p 次根 $A^{1/p}$。

算法 1.5 计算矩阵主 p 次根的 Schur-Halley 法 [? , 算法 4]

给定矩阵 $A \in \mathbb{R}^{n\times n}$, 其所有特征值都不属于 $\mathbb{R}^- := (-\infty, 0]$. 给定整数 $p \geq 2$, 则存在整数 $k_0 \geq 0$ 及奇数 q 使得 $p = 2^{k_0}q$. 本算法通过实 Schur 分解和对偶 Halley 法 (1.24) 来计算 A 的主 p 次根 $A^{1/p}$.

1. 计算矩阵 A 的实 Schur 分解 $A = QRQ^{\mathrm{T}}$.

2. 若 $q = 1$，令 $k_1 = k_0$; 若 $q \neq 1$，则选取 $k_1 \geq k_0$ 使得存在正数 s 使任意 A 的特征值 λ 满足
$$s\lambda^{1/2^{k_1}} \in \left\{ z \in \mathbb{C} : \left| z - \frac{8}{5} \right| \leq 1 \right\}.$$

3. 通过算法 1.2 计算 $B = R^{1/2^{k_1}}$.

4. 若 $q = 1$, 则令 $X = QBQ^{\mathrm{T}}$；若 $q \neq 1$, 则通过对偶 Halley 法 (1.24) 来计算 $C = (B/s)^{1/q}$ 并令 $X = Q(Cs^{1/q})^{2^{k_1-k_0}}Q^{\mathrm{T}}$.

The results concerning convergence of these two methods have recently been studied under the assumption that the eigenvalues of A are all lie in some region, see for example [? ? ? ? ? ?]. In particular, Guo in [?] shown that the matrix sequence $\{X_k\}$ generated by Newton's method (2.12) starting from the identity matrix converges to the principal pth root of A if all of whose eigenvalues lie in the set $\mathcal{E}_1 := \{z \in \mathbb{C} : |z-1| \leq 1\}$. Iannazzo obtained in [?] that the matrix sequence $\{X_k\}$ generated by Halley's method (3.84) starting also from the identity matrix converges to $A^{1/p}$ for each A having eigenvalues in the set $\mathcal{E}_2 := \{z \in \mathbb{C} : \mathrm{Re}\, z > 0\}$. I

1.3.3　代数 Riccati 方程

1.4　论文的组织

第 2 章　预备知识

我们把后文中用到的混合有限元方法和构建混合元所用的几何遗传树结构，以及移动网格方法，在本章中做简要描述和历史回顾。

2.1　混合有限元

混合有限元方法的专著可以参考F.Brezzi和M.Fortin [?]。二阶椭圆性问题的混合有限元方法就是基于Hellinger-Reissner变分原理的有限元方法。在流体中混合元方法是把速度和压力耦合在一起求解，因而精度得到提高，体现出相对于分开求解速度和压力有限元方法的优势。另外，对于不可压流体中的Galenkin逼近，只能采取混合有限元。在这一节中，我们以Stokes方程为例，来介绍混合元方法。Stokes方程可以看做是Navier-stokes的简化，去掉了非线性项。Stokes方程有两个需要求解的函数速度$\boldsymbol{u}$和压力p是耦合的，并且速度要满足质量守恒的条件($\nabla \cdot \boldsymbol{u} = 0$)。因此，需要我们用混合有限元方法来求解方程。为确保方程的解存在唯一，速度和压力所在的有限元空间必须满足inf-sup条件。如果有限元不满足inf-sup(LBB) 条件，例如在四边形单元中的$Q_1 - Q_1$,$Q_1 - P_0$,三角形单元中的$P_1 - P_1$,$P_1 - P_0$元, 还需要在压力空间上做稳定化。

下面我们以Stokes方程为例，给出它的混合变分形式。Stokes方程可以表示为：

$$\begin{aligned} -\triangle \boldsymbol{u} + \nabla p = f \quad & \text{在}\Omega\text{内}, \\ \nabla \cdot \boldsymbol{u} = 0 \quad & \text{在}\Omega\text{内}, \\ \boldsymbol{u} = 0 \qquad & \text{在}\partial\Omega\text{上}, \end{aligned} \tag{2.1}$$

令

$$a(\boldsymbol{u}, \boldsymbol{v}) = \int_\Omega \nabla \boldsymbol{u} \cdot \nabla \boldsymbol{v}, \tag{2.2}$$

$$b(\boldsymbol{u}, q) = -\int_\Omega \nabla \cdot \boldsymbol{u} q. \tag{2.3}$$

为简便起见，我们这里只考虑二维情形。对(2.1)利用格林公式，可以得到混合变分形式：寻找$(\boldsymbol{u},p)\in(H_0^1(\Omega))^2\times L_0^2(\Omega)$ 使得

$$\begin{aligned} a(\boldsymbol{u},\boldsymbol{v})+b(\boldsymbol{v},p)= &\ \ (f,v),\\ b(\boldsymbol{u},q)= &\quad\ \ 0. \end{aligned}\tag{2.4}$$

对$\forall(\boldsymbol{v},q)\in X\times P, X\in(H_0^1(\Omega))^2, P\in(L_0^2)$, $L_0^2=\{q\in L^2(\Omega)|\int_\omega qdx=0$。其中有限元空间$X,P$要满足inf-sup条件或者LBB条件，即

$$\beta||q||_P\leq\sup_{v\in X}\frac{b(v,q)}{||\boldsymbol{v}||_X},\quad\forall q\in P.\tag{2.5}$$

这个条件的验证由很多专著中给出，例如([?],[?], [?]等)。我们不加证明地给出如下定理：

定理 2.1. *问题(2.4)的解存在唯一。*

在研究用混合元去离散Stokes问题时，我们构造的有限维子空间$(X_h,P_h)\subset(X,P)$要满足离散的inf-sup条件:

$$\beta_0||q_h||_{P_h}\leq\sup_{\boldsymbol{v}_h\in X_h}\frac{b(\boldsymbol{v}_h,q_h)}{||\boldsymbol{v}_h||_{X_h}},\quad\forall q_h\in P_h.\tag{2.6}$$

其中β_0是一个跟网格尺度无关的常量。Mini元，RT元以及Taylor-Hood元都是满足离散的LBB条件的。我们得到(2.4)对应的离散形式：寻找$(\boldsymbol{u}_h,p_h)\in X_h\times P_h$ 使得

$$\begin{aligned} a(\boldsymbol{u}_h,\boldsymbol{v}_h)+b(\boldsymbol{v}_h,p_h)= &\ \ (f_h,v_h),\\ b(\boldsymbol{u}_h,q_h)= &\qquad\ 0. \end{aligned}\tag{2.7}$$

求解(2.7)，得到的离散数值解$(\boldsymbol{u}_h,p_h)$, 如果$(\boldsymbol{u},p)$ 是弱形式(2.4)的解，则我们可以得到对应的误差估计：

$$||\boldsymbol{u}-\boldsymbol{u}_h||_1+||p-p_h||_0\leq C\{\inf_{\boldsymbol{v}_h\in X_h}||\boldsymbol{u}-\boldsymbol{v}_h||_1+\inf_{q_h\in P_h}||p-q_h||_0\}.\tag{2.8}$$

我们倾向于使用满足LBB条件的混合元Taylor-Hood元，如P_2-P_1,Q_2-Q_1元，即速度是二次元，压力是线性元。在实际的工程计算中，因为线性元简单，所以比较受欢迎。因此我们设想，如果把稳定的Taylor-Hood元中的速度二次元换成线性元，同时又要保证速度和压力满足inf-sup条件,这就

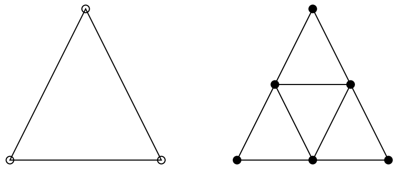


图 2.1: 左: 压力p 单元, ∘ 为p单元上的自由度; 右: 4个速度v 单元, ● 表示v单元上的自由度.

是$P_1isoP_2P_1$的想法。如图2.1所示，因为压力是线性元，因此在一个三角形单元上有三个自由度，分布在三角形的三个顶点上。同样地，速度是二次元，在这个三角形单元个有6个自由度，分布在三角形的三个顶点和三边的中点上。如果这个三角形单元加密一次，即连接三边中点，因此得到四个小三角形。这样在每个小三角形上有分布着三个自由度。在[?]中证明了$P_1isoP_2P_1$是满足inf-sup条件的。令$X_h = X^1_{h/2}, P_h = P^1_h$,则

定理 2.2. *如果Ω是一个多边形，并且对所有h都有$\Omega_h = \Omega$,如果Ω_h中所有的三角形都至少有一个顶点不在Ω的边界Γ上，X_h, P_h是满足LBB条件的，那么问题(2.7) 是有唯一解$(\boldsymbol{u}_h, p_h) \in X_h \times (P_h/R)$($p_h$除了一个常数外是确定的)。*

定理的过程我们不再给出，[?]中还给出了$P_1isoP_2P_1$元的误差估计:

$$|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)|_0 \le hK(||\boldsymbol{u}||_{(H^2(\Omega))^2} + ||p||_{H^1(\Omega)/R}), \tag{2.9}$$

$$|\nabla(p - p_h)|_0 \le K(||\boldsymbol{u}||_{(H^2(\Omega))^2} + ||p||_{H^1(\Omega)/R}). \tag{2.10}$$

但是需要注意到图2.1中，每个小三角形上的三个自由度,其实是大的三角形6个自由度中的3个。因此，在做速度单元和压力单元拼装的时候，需要先找到速度单元基函数所在的大单元，然后用大单元的基函数中正好位于当前速度单元上基函数与压力单元上的基函数进行拼装。这会带来一些技术上的困难。如果图2.1右中每个小三角形上的自由度都是局部在一个单元上，这时跟$P_1isoP_2P_1$拥有同样的网格结构，也满足inf-sup条件，但速度单元全部是线性元，因此称作$4P1 - P1$元。在这种情况下，只要建立速度单元和压力单元之间的索引，速度单元与压力单元间的拼装变得容易很多。这时速度和压力分别在两套网格上，其中速度网格由压力网格加密一次得到。

2.2　几何遗传树结构

在上一节中，我们提到速度单元和压力单元之间的索引，而这个索引的建立利用了[?] 中的几何遗传树结构。这种树结构是用来做h-自适应，下面我们就介绍一下这种树结构。

在h-自适应中，两个耦合在一起的物理量(如上一节提到的速度u和压力p)，有可能在不同的地方出现奇异，因此需要分别在两套网格上离散数值解。在拼装$\int_\Omega \nabla \cdot \boldsymbol{u} p$这一项时，因为$\boldsymbol{u}$ 和p在不同的网格上，因此需要构建两套网格之间的联系。如果函数在相对于计算区域来说，很小的区域内有奇异，好的网格需要在这个奇异出现的地方网格非常集中。特别地，如果最小的网格单元和最大的网格单元面积相差上百倍，这时候建立两个网格之间对应的代价是非常昂贵的。为了创建高效的多网格自适应策略，[?]中选取了几何遗传树，具有遗传性的网格结构可以描述为点、线、面、体。并且针对无结构网格，加密也是通过按层次来加密的。这种分层的数据结构可以提供简单的编程方法和快速的网格加密算法，以及多重网格求解器。

我们以二维区域Ω上的一个三角剖分$\mathcal{T}_0$为例，来介绍一下几何遗传树结构。为与有限元空间的单元区分开，称三角剖分中的单元为几何单元。τ为$\mathcal{T}_0$的几何单元。三角剖分中所有几何性质都有属于的关系，即如果一条边是一个三角形三边中的一条，则称这条边属于这个三角形。同样的，如果一个点是一条边两个顶点中的一个，则称这个点属于这条边。三角剖分中的一个三角形可以被加密成四个小的三角形。在三角形加密的过程中，三条边分别被分成了两部分。当三角剖分中所有的三角形都加密一次，我们可以得到一个更密的三角剖分$\mathcal{T}_1$, 这就是$\mathcal{T}_0$的一次全局加密。经过逐次加密，我们可以得到一系列的三角剖分$\{\mathcal{T}_n\}$。如果$\mathcal{T}_n$中的一个三角形τ_1位于$\mathcal{T}_{n-1}$中的三角形τ_0中，则τ_1被称作是τ_0的一个孩子。我们给$\mathcal{T}_0$中的所有三角形一个虚拟的父亲，与所有从$\mathcal{T}_0$的三角形中派生出的所有四叉树组成了一个树的数据结构，如图2.4所示。这颗树也被称为几何遗传树。 用几何遗传树结构来构建h-自适应的内容就不再详细介绍了。

2.3　移动网格方法

移动网格方法常常用在数值计算中，寻找好的网格(跟具体问题有关)，并在网格上进行数值求解，使得在不增加自由度的前提下，尽可能的提高计算精度。移动网格方法包括网格的移动策略，以及在方程在网格上的数值求解算

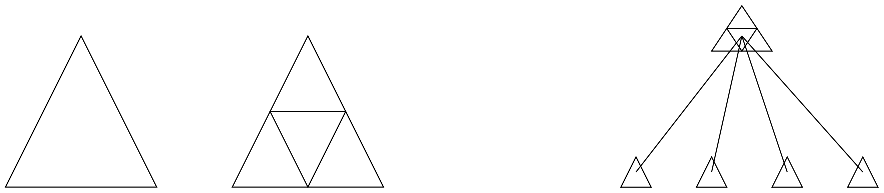


图 2.2: 全局加密一次的网格　　图 2.3: 它的几何遗传树

图 2.4: 几何遗传树结构

法。在[?]、[?]、[?] 中系统地给出了移动网格方法的介绍。移动网格的策略是基于等分布原则(Equi-distribute Principle)，这是de Boor在1974年[?] 提出的。等分布原则是根据一个能反映局部网格计算质量的指标，比如误差，移动网格节点使得这个指标在网格上均匀分布。如果在计算结果误差比较大的地方网格集中，尽量使得误差分布的比较均匀，从而提高计算精度。

一些符号的定义如下：令$\Omega \subset \mathbb{R}^N, N = 1, 2, 3$代表物理区域，$\boldsymbol{x}$代表物理区域上的坐标。同样地，$\Omega_c \subset \mathbb{R}^N$表示逻辑区域(或称计算区域)，$\boldsymbol{\xi}$为其坐标。

de Boor 给出了一维情形的证明，如下：

假设$\Omega = [a, b]$,$\Omega_c = [0, 1]$分别为物理区域和计算区域，在Ω_c上有一个均匀网格$\{\xi_j = \frac{j}{N} | j = 0, 1, \cdots, N\}$,

2.4　预处理方法

应用迭代法计算矩阵的 p 次根, 实际上是求解如下的计算矩阵方程

$$f(X) = X^p - A = 0, \quad p \geq 2. \tag{2.11}$$

若假设初始迭代点 X_0 与矩阵 A 是可交换的, 且对于任意的 $k \geq 0$, X_k 都是非奇异的, 则计算上述矩阵方程的 Newton 法可写成:

$$X_{k+1} = N(X_k), \quad k = 0, 1, 2, \ldots, \tag{2.12}$$

其中

$$N(X) := \frac{1}{p} X \left[(p-1)\mathrm{I} + AX^{-p} \right]. \tag{2.13}$$

本章将研究通过 Newton 法来计算矩阵 p 次根的一个新的收敛性结果。新的结果比其他现有的结果在计算矩阵 p 次根时有更好的计算效果。本章中总假定整数 $p \geq 2$.

2.5　收敛性结果

在给出本章的主要收敛结果前, 先引入一些记号。令

$$\phi_1(z) := 1 - u^p(z)(1-z), \quad z \in \mathcal{D}_{0,1}, \tag{2.14}$$

其中

$$u(z) := \frac{1}{1-\frac{1}{p}z}, \quad z \in \mathcal{D}_{0,1},$$

及

$$\mathcal{D}_{0,1} := \{z \in \mathbb{C} : |z| < p\}. \tag{2.15}$$

定义

$$\mathcal{R}_1 := \left\{ z \in \mathbb{C} : 1 - z \in \overline{\mathcal{D}}_1 \bigcup \left(\mathcal{D}_{0,1} \bigcap \mathcal{D}_{2,1} \right) \right\} \tag{2.16}$$

及

$$\widehat{\mathcal{R}}_1 := \left\{ z \in \mathbb{C} : 1 - z \in \mathcal{D}_1 \bigcup \left(\mathcal{D}_{0,1} \bigcap \mathcal{D}_{2,1} \right) \right\}, \tag{2.17}$$

其中 $\mathcal{D}_{0,1}$ 由 (2.15) 给出, $\mathcal{D}_1$ 定义为

$$\mathcal{D}_1 := \{z \in \mathbb{C} : |z| < 1\}, \tag{2.18}$$

$\overline{\mathcal{D}}_1$ 为 $\mathcal{D}_1$ 的闭包, 而

$$\mathcal{D}_{2,1} := \left\{ z \in \mathbb{C} : \sup_{m \geq 2} \left\{ \frac{|S_{1,m}(z)|}{|z|} \right\} \cdot \frac{|z| + |\phi_1(z)|}{\big||z| - |\phi_1(z)|\big|} < 1 \right\}, \tag{2.19}$$

其中 ϕ_ν 由 (2.14) 给出,

$$S_{1,m}(z) = \sum_{j=2}^{m} c_{1,j} z^j, \quad z \in \mathcal{D}_{0,1}, \tag{2.20}$$

其中

$$c_{1,j} = \frac{p(p+1) \cdot (p+j-2)}{(j-1)!p^{j-1}} - \frac{p(p+1) \cdot (p+j-1)}{j!p^j} > 0, \quad j = 2, 3, \ldots,$$

并满足

$$\sum_{j=2}^{\infty} c_{1,j} = 1,$$

故 $\phi_1(z)$ 有如下的 Maclaurin 级数形式

$$\phi_1(z)=\sum_{j=2}^{\infty}c_{1,j}z^j,\quad z\in\mathcal{D}_{0,1},\tag{2.21}$$

其中 $\mathcal{D}_{0,1}$ 由 (2.15) 给出.

下面给出应用 Newton 法 (2.12) 来计算矩阵主 p 次根时的收敛性结果:

定理 2.3. 设 $\mathcal{R}_1$ 由 (2.16) 定义，而 $\widehat{\mathcal{R}}_1$ 由 (2.17) 定义。如果矩阵 $A\in\mathbb{C}^{n\times n}$ 的所有特征值都属于 $\mathcal{R}_1$ 且所有零特征值都是半单的，那么以 $X_0=\mathrm{I}$ 为初始点的 *Newton* 法 (2.12) 迭代所产生的矩阵序列 $\{X_k\}$ 收敛于矩阵 A 的主 p 次根 $A^{1/p}$。特别地，如果 A 的所有特征值都属于 $\widehat{\mathcal{R}}_1\backslash\{0\}$，那么 *Newton* 法 (2.12) 是二次收敛的。

注 2.1. 图 2.5 所示的是，由 (2.16) 所定义的 $\mathcal{R}_1$ 的近似收敛域的九种情形，其中红色表示的是收敛域 $\{z\in\mathbb{C}:1-z\in\overline{\mathcal{D}}_1\}$，而蓝色表示的是收敛域 $\{z\in\mathbb{C}:1-z\in\mathcal{D}_{0,1}\bigcap\mathcal{D}_{2,1}\}$。由该图可知，对于某一固定的 p，当 m 分别取 $m=20,100,500$ 时，所得到的近似收敛域 (见 (a)-(c),或 (d)-(f)，或 (g)-(i)) 几乎一样。为此，在实际计算中只需要取 $m=20$ 即可。

2.6　预备引理

在证明定理 2.3 前，需要一些有用的预备引理.

给定任意一个复数 $\lambda\in\mathbb{C}$，令

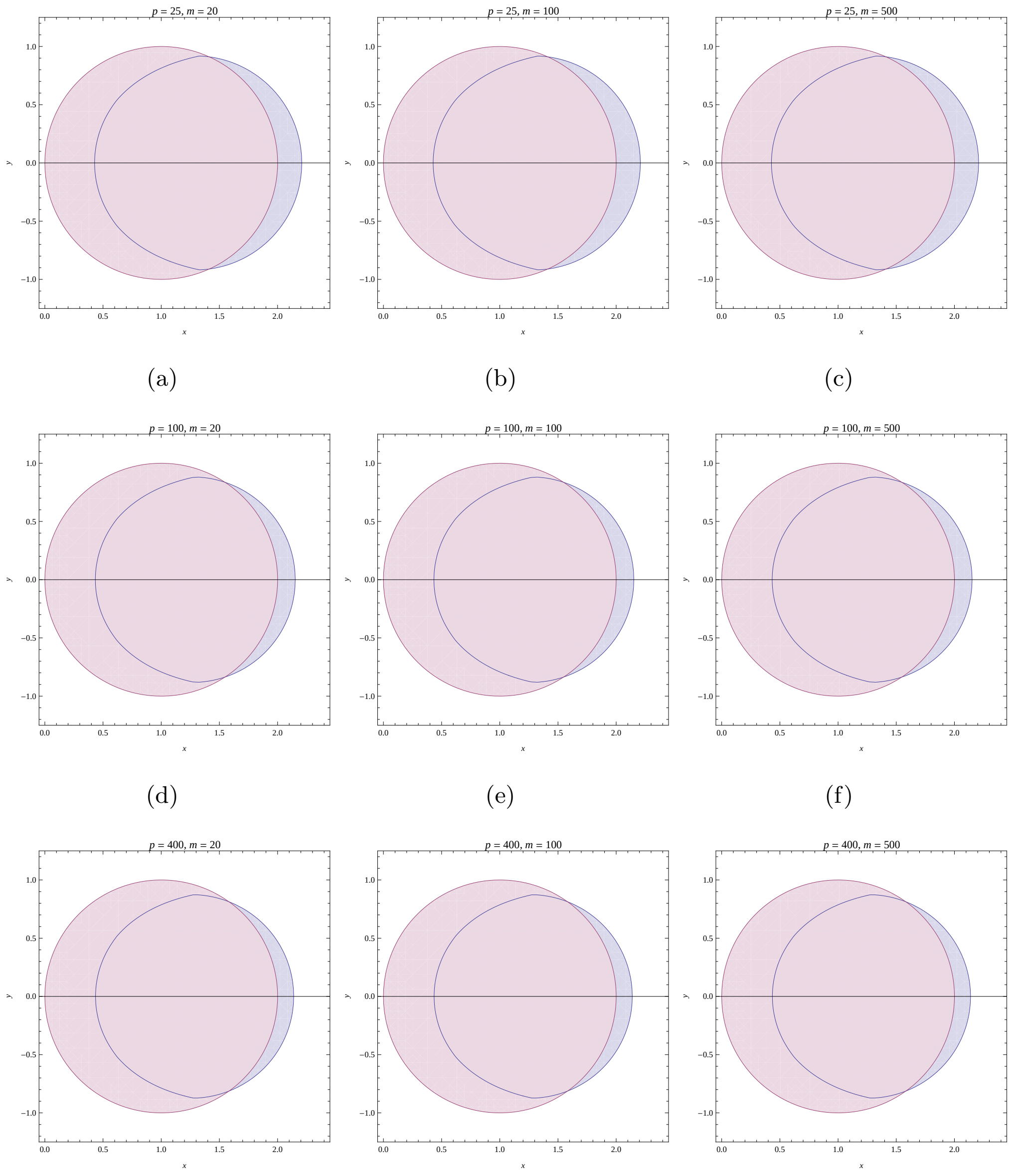
$$f(z):=z^p-\lambda,\quad z\in\mathbb{C}\tag{2.22}$$

及

$$r(z,\lambda):=1-\lambda z^{-p},\quad z\in\mathbb{C}\backslash\{0\}.\tag{2.23}$$

引理 2.1. 设 $r(z,\lambda)$ 由 (3.49) 定义. 对于某个复数 $z\in\mathbb{C}\backslash\{0\}$，若 $r(z,\lambda)\in\mathcal{D}_{0,1}\backslash\{1\}$，其中 $\mathcal{D}_{0,1}$ 由 (2.15) 给出，则由 (2.13) 的标量形式所得到的 $N(z)$ 有定义且

$$r(N(z),\lambda)=\phi_1(r(z,\lambda)),\tag{2.24}$$



(a)　(b)　(c)

(d)　(e)　(f)

(g)　(h)　(i)

图 2.5: 由 (2.16) 给出的 $\mathcal{R}_1$ 在九种不同情形 (p 分别取 25, 100, 400 及 m 分别取 20, 100, 500) 下的近似域, 其中红色域表示集合 $\{z \in \mathbb{C} : 1 - z \in \overline{\mathcal{D}}_1\}$, 而蓝色域表示集合 $\{z \in \mathbb{C} : 1 - z \in \mathcal{D}_{0,1} \bigcap \mathcal{D}_{2,1}\}$.

其中 $\phi_1(z)$ 由 (2.14) 定义. 此外, 有如下的估计: 当 $r(z,\lambda) \in \overline{\mathcal{D}}_1 \backslash \{0,1\}$ 时,

$$|r(N(z),\lambda)| < |r(z,\lambda)|^2. \tag{2.25}$$

当 $r(z,\lambda) \in \mathcal{D}_{0,1}\backslash\{1\}$ 时，

$$|r(N(z),\lambda)| \leq \sup_{m\geq 2}\left\{\left|\frac{S_{1,m}(u)}{u^2}\right|\right\} \cdot \frac{|u|+|r(z,\lambda)|}{|u|-|r(z,\lambda)|} \cdot |r(z,\lambda)|^2, \tag{2.26}$$

其中 $\overline{\mathcal{D}}_1$ 为 $\mathcal{D}_1$ 的闭包，而 $\mathcal{D}_1$ 由 (2.18) 定义，$S_{1,m}(u)$ 由 (2.20) 定义且对于每一个 $u \in \mathcal{D}_{0,1}$ 均满足 $|u| > |r(z,\lambda)|$, $m \geq 2$.

证明. 对于任意的复数 $z \in \mathbb{C}\backslash\{0\}$, 由 (2.13) 知 $N(z)$ 是存在的. 由于

$$N(z) = \frac{1}{p}z\left[(p-1)+\lambda z^{-p}\right] = z\left[1+\frac{1}{p}r(z,\lambda)\right] \neq 0. \tag{2.27}$$

故

$$r(N(z),\lambda) = 1-\left[1+\frac{1}{p}r(z,\lambda)\right]^{-p}(1-r(z,\lambda)) = \phi_1(r(z,\lambda)), \tag{2.28}$$

即得 (2.24) 是成立的.

若 $r(z,\lambda) \in \overline{\mathcal{D}}_1\backslash\{0,1\} \subset \mathcal{D}_{0,1}$, 则由 (2.21) 及 (2.28), 并注意到对于任意的 $w \in \overline{\mathcal{D}}_1\backslash\{1\}$, 不等式 $|c_3+c_4w| < c_3+c_4$ 都是成立的，于是可得

$$\begin{aligned}
|r(N(z),\lambda)| &= |\phi(r(z,\lambda))| \\
&= |r(z,\lambda)|^2\left|\sum_{j=2}^{\infty} c_{1,j}r^{j-2}(z,\lambda)\right| \\
&\leq |r(z,\lambda)|^2\left[|c_{1,3}+c_{1,4}r(z,\lambda)| + \sum_{j=4}^{\infty} c_{1,j}|r(z,\lambda)|^{j-2}\right] \\
&< |r(z,\lambda)|^2\sum_{j=2}^{\infty} c_{1,j} = |r(z,\lambda)|^2 \\
&\leq |r(z,\lambda)|.
\end{aligned} \tag{2.29}$$

即 (2.25) 是成立的.

若 $r(z,\lambda) \in \mathcal{D}_0\backslash\{1\}$, 则由 (2.21) 及 (2.28) 可知，对于任意的 $u \in \mathbb{C}\backslash\{0\}$ 有

$$\begin{aligned}
r(N(z),\lambda) &= \phi_1(r(z,\lambda)) \\
&= \sum_{j=2}^{\infty} c_{1,j}r^j(z,\lambda)
\end{aligned}$$

$$= \left[\sum_{j=2}^{\infty} c_{1,j} u^{j-2} \left(\frac{r(z,\lambda)}{u}\right)^{j-2}\right] \cdot r^2(z,\lambda).$$

对于任意的 $m \geq 2$, 应用 Abel 变换可得

$$\sum_{j=2}^{m} c_{1,j} u^{j-2} \left(\frac{r(z,\lambda)}{u}\right)^{j-2} = \sum_{j=2}^{m-1} \left(\sum_{\ell=2}^{j} c_{1,\ell} u^{\ell-2}\right) \left(1 - \frac{r(z,\lambda)}{u}\right) \left(\frac{r(z,\lambda)}{u}\right)^{j-2} + \left(\sum_{\ell=2}^{m} c_{1,\ell} u^{\ell-2}\right) \cdot \left(\frac{r(z,\lambda)}{u}\right)^{m-2}.$$

于是

$$\left|\sum_{j=2}^{m} c_{1,j} u^{j-2} \left(\frac{r(z,\lambda)}{u}\right)^{j-2}\right| \leq \sup_{2\leq j\leq m-1} \left\{\left|\sum_{\ell=2}^{j} c_{1,\ell} u^{\ell-2}\right|\right\} \cdot \left|1 - \frac{r(z,\lambda)}{u}\right| \cdot \sum_{j=2}^{m-1} \left|\frac{r(z,\lambda)}{u}\right|^{j-2} + \left|\frac{S_{1,m}(u)}{u^2}\right| \cdot \left|\frac{r(z,\lambda)}{u}\right|^{m-2}, \quad m > 2.$$

在上述不等式中令 $m \to \infty$, 则对于任意的 $r(z,\lambda) \in \mathcal{D}_0 \backslash \{1\}$ 及满足关系 $|u| > |r(z,\lambda)|$ 的任意复数 $u \in \mathcal{D}_0$, 有

$$|r(E(z),\lambda)| \leq \sup_{m\geq 2} \left\{\left|\sum_{j=2}^{m} c_{1,j} u^{j-2}\right|\right\} \cdot \left|1 - \frac{r(z,\lambda)}{u}\right| \cdot \sum_{m=2}^{\infty} \left|\frac{r(z,\lambda)}{u}\right|^{m-2} \cdot |r(z,\lambda)|^2$$
$$= \sup_{m\geq 2} \left\{\left|\frac{S_{1,m}(u)}{u^2}\right|\right\} \cdot \frac{\left|1 - \frac{r(z,\lambda)}{u}\right|}{1 - \left|\frac{r(z,\lambda)}{u}\right|} \cdot |r(z,\lambda)|^2$$
$$= \sup_{m\geq 2} \left\{\left|\frac{S_{1,m}(u)}{u^2}\right|\right\} \cdot \frac{|u| + |r(z,\lambda)|}{|u| - |r(z,\lambda)|} \cdot |r(z,\lambda)|^2.$$

从而 (2.26) 得证. 证完. □

引理 2.2. *设 $r(z,\lambda)$ 由 (3.49) 定义. 对于某个复数 $z_0 \in \mathbb{C}\backslash\{0\}$, 若 $r(z_0,\lambda) \in \overline{\mathcal{D}}_1\backslash\{0,1\}$, 则由 (2.12) 的标量形式 (以 z_0 为初始点) 迭代产生的序列 $\{z_k\}$ 是有定义的, 且有估计式:*

$$|r(z_k,\lambda)| \leq q_1^{2^{k-1}}(z_0), \quad k = 1,2,\ldots, \tag{2.30}$$

其中

$$q_1(z_0) = q_1(z_0,\lambda) := \left|\sum_{j=2}^{\infty} c_{1,j} r^{j-1}(z_0,\lambda)\right| < 1. \tag{2.31}$$

由此可知, 当 $k \to \infty$ 时 $|r(z_k,\lambda)|$ 收敛于 0, 且收敛速度是二阶的.

证明. 对于给定的 z_0, 类似于 (2.29) 的处理方式可知 $q_1(z_0) < 1$. 根据引理 2.1 中的 (2.25) 知 $z_1 = E(z_0)$ 是存在的并且有

$$|r(z_1,\lambda)| = q_1(z_0)|r(z_0,\lambda)| \leq q_1(z_0) < 1.$$

对于某一 $k \geq 1$, 假设 z_k 是存在的且 (2.30) 是成立的, 则由引理 2.1 可知 $z_{k+1} = E(z_k)$ 是存在的并且有

$$|r(z_{k+1},\lambda)| < |r(z_k,\lambda)|^2 \leq \left[q_1^{2^{k-1}}(z_0)\right]^2 = q_1^{2^k}(z_0).$$

由此知, (2.30) 对于 $k+1$ 的情形仍然成立. 于是由归纳法知, 序列 $\{z_k\}$ 是存在的且 (2.30) 恒成立. 证完. □

对于某个复数 $z_0 \in \mathbb{C}\backslash\{0\}$, 引理 2.2 说明当 $r(z_0,\lambda) \in \overline{\mathcal{D}}_1\backslash\{0,1\}$ 时可保证 $r(z_k,\lambda)$ 是二阶收敛于 0, 下面的引理表明, 除此之外, 当 $r(z_0,\lambda) \in \mathcal{D}_{2,1}\bigcap\mathcal{D}_{0,1}\backslash\{1\}$ 时, 仍然可以保证 $r(z_k,\lambda)$ 是二阶收敛于 0.

引理 2.3. *设 $r(z,\lambda)$ 由 (3.49) 定义. 对于某一复数 $z_0 \in \mathbb{C}\backslash\{0\}$, 若 $r(z_0,\lambda) \in \mathcal{D}_{2,1}\bigcap\mathcal{D}_{0,1}\backslash\{1\}$ 且*

$$q_2(z_0) = q_2(z_0,\lambda) := \sup_{m\geq 2}\left\{\frac{|S_{1,m}(r(z_0,\lambda))|}{|r(z_0,\lambda)|}\right\} \cdot \frac{|r(z_0,\lambda)| + |\phi_1(r(z_0,\lambda))|}{|r(z_0,\lambda)| - |\phi_1(r(z_0,\lambda))|} < 1, \tag{2.32}$$

其中 $\mathcal{D}_{2,1}$ 由 (2.19) 定义, 则由 (2.12) 的标量形式 (以 z_0 为初始点) 迭代产生的序列 $\{z_k\}$ 是有意义的, 且有如下估计:

$$|r(z_k,\lambda)| \leq q_2^{2^k-1}(z_0) \cdot |r(z_0,\lambda)|, \quad k = 0,1,\ldots. \tag{2.33}$$

由此可知, 当 $k \to \infty$ 时 $|r(z_k,\lambda)|$ 收敛于 0, 且收敛速度是二阶的.

证明. 显然, 对于给定的 z_0 有 $r(z_0,\lambda) \in \mathcal{D}_{0,1}\backslash\{1\}$. 故 $z_1 = N(z_0)$ 存在且由 (2.27) 知 $z_1 \neq 0$. 注意到, 当 $k \to \infty$ 时 $S_{1,m}(r(z_0,\lambda)) \to \phi_1(r(z_0,\lambda))$, 故由 (2.32) 可得

$$\left|\frac{\phi_1(r(z_0,\lambda))}{r(z_0,\lambda)}\right| \leq \sup_{m\geq 2}\left\{\left|\frac{S_{1,m}(r(z_0,\lambda))}{r(z_0,\lambda)}\right|\right\} < q_2^2(z_0) < 1,$$

于是有

$$|r(z_1,\lambda)| = |\phi_1(r(z_0,\lambda))| = \left|\frac{\phi_1(r(z_0,\lambda))}{r(z_0,\lambda)} r(z_0,\lambda)\right| < q_2^2(z_0) \cdot |r(z_0,\lambda)|,$$

即当 $k=1$ 时 (2.33) 是成立的. 现假设 $z_0, z_1, \ldots, z_k$ 都是存在的且满足 (2.33), 则

$$|r(z_k,\lambda)| \leq q_2^2(z_0)|r(z_0,\lambda)| < |r(z_0,\lambda)| < p.$$

由引理 2.1 (取 $u = r(z_0,\lambda)$) 可知 $z_{k+1} = N(z_k)$ 是存在的且 $z_{k+1} \neq 0$. 进一步有

$$\begin{aligned}
|r(z_{k+1},\lambda)| &= |\phi(r(z_k,\lambda))| \\
&\leq \sup_{m\geq 2}\left\{\frac{|S_{1,m}(r(z_0,\lambda))|}{|r(z_0,\lambda)|^2}\right\} \cdot \frac{|r(z_0,\lambda)| + |\phi_1(r(z_0,\lambda))|}{\big||r(z_0,\lambda)| - |\phi(r(z_0,\lambda))|\big|} \cdot |r(z_k,\lambda)|^2 \\
&\leq \sup_{m\geq 2}\left\{\frac{|S_{1,m}(r(z_0,\lambda))|}{|r(z_0,\lambda)|^2}\right\} \cdot \frac{|r(z_0,\lambda)| + |\phi_1(r(z_0,\lambda))|}{\big||r(z_0,\lambda)| - |\phi_1(r(z_0,\lambda))|\big|} \left[q_2^{2^k-1}(z_0)\right]^2 \cdot |r(z_0,\lambda)|^2 \\
&= \sup_{m\geq 2}\left\{\frac{|S_{1,m}(r(z_0,\lambda))|}{|r(z_0,\lambda)|}\right\} \cdot \frac{|r(z_0,\lambda)| + |\phi_1(r(z_0,\lambda))|}{\big||r(z_0,\lambda)| - |\phi_1(r(z_0,\lambda))|\big|} \cdot [q_2(z_0)]^{2^{k+1}-2} \cdot |r(z_0,\lambda)| \\
&= [q_2(z_0)]^{2^{k+1}-1} \cdot |r(z_0)|,
\end{aligned}$$

由此, 根据归纳法知 (2.33) 得证. 证完. □

基于上述几个引理, 可以得到如下的关于 (标量形式的) Newton 法 (2.12) 的收敛性结果. 为此, 定义

$$\mathcal{R}_{1,1} := \{\lambda \in \mathbb{C} : r(z_0,\lambda) \in \overline{\mathcal{D}}_1 \text{ for some } z_0 \in \mathbb{C}\backslash\{0\}\}, \tag{2.34}$$

及

$$\mathcal{R}_{1,2} := \left\{\lambda \in \mathbb{C} : r(z_0,\lambda) \in \mathcal{D}_{0,1} \bigcap \mathcal{D}_{2,1} \text{ for some } z_0 \in \mathbb{C}\backslash\{0\}\right\}, \tag{2.35}$$

其中 $\overline{\mathcal{D}}_1$ 为 $\mathcal{D}_1$ 的闭包, 而 $\mathcal{D}_1$ 由 (2.18) 定义, $\mathcal{D}_{0,1}$ 和 $\mathcal{D}_{2,1}$ 分别由 (2.15) 和 (2.19) 定义.

引理 2.4. 对于任意的 $\lambda \in \mathcal{R}_{1,1} \bigcup \mathcal{R}_{1,2}$, 其中 $\mathcal{R}_{1,1}$ 和 $\mathcal{R}_{1,2}$ 分别由 (2.34) 和 (2.35) 定义, 以 $z_0 \in \mathbb{C}\backslash\{0\}$ 为初始点的标量 *Newton* 法 (2.12) 迭代所产生的序列 $\{z_k(\lambda)\}$ 收敛于 λ 的主 p 次根 $\lambda^{1/p}$. 此外, 若 $\lambda \neq 0$, 则收敛速度是二阶的.

证明. 下面分四步来证明该引理.

第一步. 假设 $\mathcal{R}_c$ 为属于 $\mathcal{R}_{1,1}$ 或 $\mathcal{R}_{1,2}$ 的任一闭区域且 $0 \notin \mathcal{R}_c$. 首先证明, 对于任意的 $\lambda \in \mathcal{R}_c$, 序列 $\{z_k(\lambda)\}$ 一致收敛于 λ 的一个 p 次根 $z(\lambda)$.

记

$$r(z_k,\lambda)\triangleq r(z_k(\lambda),\lambda).$$

对于任意的 $\lambda\in\mathcal{R}_{1,1}\bigcup\mathcal{R}_{1,2}$, 由于级数

$$\sum_{j=2}^{\infty}c_{1,j}r^j(z_0,\lambda)$$

是解析的, 且 $\mathcal{R}_c\subset\mathcal{R}_{1,1}\bigcup\mathcal{R}_{1,2}$ 是有界的, 故根据解析函数的最大模定理知, 存在 $\widehat{\lambda}\in\partial\mathcal{R}_c$ 使得

$$\left|\sum_{j=2}^{\infty}c_{1,j}r^j(z_0,\widehat{\lambda})\right|=\max_{\lambda\in\mathcal{R}_c}\left|\sum_{j=2}^{\infty}c_{1,j}r^j(z_0,\lambda)\right|. \tag{2.36}$$

令

$$q(z_0):=\begin{cases}\max\limits_{\lambda\in\mathcal{R}_c}q_1(z_0,\lambda), & \text{if } \mathcal{R}_c\subset\mathcal{R}_{1,1},\\ \max\limits_{\lambda\in\mathcal{R}_c}q_2(z_0,\lambda), & \text{if } \mathcal{R}_c\subset\mathcal{R}_{1,2},\end{cases}$$

其中 $q_1(z_0,\lambda)$ 和 $q_2(z_0,\lambda)$ 分别由 (2.31) 和 (2.32) 定义. 则由 (3.61), 引理 2.2 和 2.3 可得

$$q(z_0)=\begin{cases}q_1(z_0,\widehat{\lambda})<1, & \text{if } \mathcal{R}_c\subset\mathcal{R}_{1,1},\\ q_2(z_0,\widehat{\lambda})<1, & \text{if } \mathcal{R}_c\subset\mathcal{R}_{1,2},\end{cases}$$

及

$$|r(z_k,\lambda)|\le\begin{cases}q^{2^{k-1}}(z_0), & \text{if } \mathcal{R}_c\subset\mathcal{R}_{1,1},\\ q^{2^k-1}(z_0)\cdot r_*, & \text{if } \mathcal{R}_c\subset\mathcal{R}_{1,2},\end{cases}\quad k=1,2,\dots,\lambda\in\mathcal{R}_c \tag{2.37}$$

其中 $r_*:=\max\limits_{\lambda\in\mathcal{R}_c}|r(z_0,\lambda)|$ 为不依赖于 $\lambda\in\mathcal{R}_c$ 的正实数. 于是, 对于任意的 $\lambda\in\mathcal{R}_c$, 当 $k\to\infty$ 时序列 $\{r(z_k,\lambda)\}$ 一致收敛于 0, 且收敛速度是二阶的. 此外, 由关系

$$z_k^p=\frac{\lambda}{1-r(z_k,\lambda)},\quad \lambda\in\mathcal{R}_c,\ k=0,1,\dots \tag{2.38}$$

可知, 对任意的 $\lambda\in\mathcal{R}_c$, 序列 $\{z_k(\lambda)\}$ 是一致有界的. 因而, 存在一个不依赖于 k 的常数 $M>0$ 及 $\lambda\in\mathcal{R}_c$ 使得

$$\frac{1}{p}|z_k(\lambda)|\le M,\quad k\ge 0,\lambda\in\mathcal{R}_c. \tag{2.39}$$

由 (2.13) 及 (3.64) 可得

$$|z_{k+1}(\lambda)-z_k(\lambda)|=\frac{1}{p}|z_k(\lambda)||r(z_k,\lambda)|\le M|r(z_k,\lambda)|,\quad \lambda\in\mathcal{R}_c,\ k=0,1,\dots,$$

结合 (3.62) 可得知序列 $\{z_{k+1}(\lambda)-z_k(\lambda)\}$ 收敛于 0. 故对于任意的 $\lambda\in\mathcal{R}_c$, $\{z_k(\lambda)\}$ 是 Cauchy 列. 于是存在定义于 $\mathcal{R}_c$ 的 $z(\lambda)$ 使得对于任意的 $\lambda\in\mathcal{R}_c$, 序列 $z_k(\lambda)$ 一致收敛于 $z(\lambda)$. 在 (3.63) 中令 $k\to\infty$ 即可得

$$z^p(\lambda)=\lambda,\quad \lambda\in\mathcal{R}_c.$$

因此, $z(\lambda)$ 是 $\lambda\in\mathcal{R}_c$ 的一个 p 次根.

第二步. 由于对任意的 $\lambda\in\mathcal{R}_{1,1}\bigcup\mathcal{R}_{1,2}$, 存在 $\mathcal{R}_{1,1}$ 或 $\mathcal{R}_{1,2}$ 的一个闭区域使得 λ 属于它. 由第一步知, 对于任意的 $\lambda\in\mathcal{R}_{1,1}\bigcup\mathcal{R}_{1,2}$, $z(\lambda)$ 都是存在的. 在这一步中, 将进一步证明, $z(\lambda)$ 在 $\mathcal{R}_{1,1}$ 的内部 $\mathrm{Int}(\mathcal{R}_{1,1})$ 及 $\mathcal{R}_{1,2}$ 上是解析的. 由此得到 $z(\lambda)$ 落在 $\mathrm{Int}(\mathcal{R}_{1,1})$ 和 $\mathcal{R}_{1,2}$ 上的根函数的单值分枝上.

事实上, 由 $z_k(\lambda),\lambda\in\mathcal{R}_{1,1}\bigcup\mathcal{R}_{1,2}$ 的定义知, $z_k(\lambda)$ 在 $\mathrm{Int}(\mathcal{R}_{1,1})$ 或 $\mathcal{R}_{1,2}$ 都是解析的. 又由第一步知 $\{z_k(\lambda)\}$ 一致收敛于 $z(\lambda)$, 故由 Weierstrass 定理得 $z_k(\lambda)$ 在 $\mathrm{Int}(\mathcal{R}_{1,1})$ 和 $\mathcal{R}_{1,2}$ 都是解析的. 而 $z(\lambda)$ 是 $\lambda\in\mathrm{Int}(\mathcal{R}_{1,1})$ 或 $\mathcal{R}_{1,2}$ 的 p 次根, 因此,$z(\lambda)$ 落在 $\mathrm{Int}(\mathcal{R}_{1,1})$ 和 $\mathcal{R}_{1,2}$ 上的根函数的单值分枝上.

第三步. 对于 $\lambda_0\in\partial\mathcal{R}_{1,1}$ 且 $\lambda_0\ne0$, 将证明当 $\lambda\to\lambda_0$ (从 $\mathcal{R}_{1,1}$ 的内部逼近) 时 $z(\lambda)\to z(\lambda_0)$.

显然, 对于任意的 $\lambda_0\in\partial\mathcal{R}_{1,1}$ 且 $\lambda_0\ne0$, 有 $|r(z_0,\lambda_0)|<1$. 故存在 $\delta_0>0$ 使得闭区域 $\overline{O}(\lambda_0,\delta_0)\bigcap\mathcal{R}_{1,1}$ 不包含 0 和 1, 其中

$$\overline{O}(\lambda_0,\delta_0):=\{\lambda\in\mathbb{C}:|\lambda-\lambda_0|\le\delta_0\}.$$

由第一步知, $\{z_k(\lambda)\}$ 一致收敛于 $z(\lambda)$, $\lambda\in\overline{O}(\lambda_0,\delta_0)\bigcap\mathcal{R}_{1,1}$. 故对任意的 $\varepsilon>0$, 存在整数 $K>0$ 使得对任意的 $k\ge K$ 及 $\lambda\in\overline{O}(\lambda_0,\delta_0)\bigcap\mathcal{R}_{1,1}$ 都有

$$|z_k(\lambda)-z(\lambda)|<\frac{\varepsilon}{3}. \tag{2.40}$$

因 $z_k(\lambda)$ 在 $\mathcal{R}_{1,1}$ 上是解析的, 故 $z_k(\lambda)$ 在 $\overline{O}(\lambda_0,\delta_0)\bigcap\mathcal{R}_{1,1}$ 是连续的. 于是, 存在 $0<\delta_1<\delta_0$ 使得 $\overline{O}(\lambda_0,\delta_1)\subset\overline{O}(\lambda_0,\delta_0)$ 且

$$|z_k(\lambda)-z_k(\lambda_0)|<\frac{\varepsilon}{3},\quad \forall\ \lambda\in\overline{O}(\lambda_0,\delta_1)\bigcap\mathcal{R}_{1,1}. \tag{2.41}$$

因而, 对任意的 $\lambda \in \overline{O}(\lambda_0, \delta_1) \bigcap \mathcal{R}_{1,1}$, 由 (2.40) 及 (2.41) 可得

$$\begin{aligned} |z(\lambda) - z(\lambda_0)| &\leq |z(\lambda) - z_k(\lambda)| + |z_k(\lambda) - z_k(\lambda_0)| + |z_k(\lambda_0) - z(\lambda_0)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \end{aligned}$$

即 $z(\lambda)$ 在 $\lambda = \lambda_0$ 是连续的. 由 λ_0 的任意性即得证.

第四步. 因 $z(\lambda)$ 在 $\mathrm{Int}(\mathcal{R}_{1,1})$ 是解析的且落在 p 次根的单值分支上, 又因对任意的 $k \geq 0$ 都有 $z_k(1) \equiv 1$, 故 $z(1) = 1$ 且 $z(\lambda)$ 落在包含 1 的单值分支中. 即对任意 $\lambda \in \mathrm{Int}(\mathcal{R}_{1,1})$, $z(\lambda)$ 都是其主 p 次根. 注意到, 对任意的 $\lambda_0 \in \partial\mathcal{R}_{1,1}$, 当 $\lambda \to \lambda_0$ $(\lambda \in \mathrm{Int}(\mathcal{R}_{1,1}))$ 时 $z(\lambda) \to z(\lambda_0)$, 因而可知 $z(\lambda_0)$ 亦是 $\lambda_0 \in \partial\mathcal{R}_{1,1}$ 的主 p 次根.

此时, 我们已经证明当 $\lambda \in \mathcal{R}_{1,1}$ 且 $\lambda \neq 0$ 时 $z(\lambda)$ 是其主 p 次根. 现在考虑当 $\lambda = 0$ 时的情形. 由 (2.12) 可得

$$z_k(0) = \frac{p-1}{p} z_{k-1}(0) = \left(\frac{p-1}{p}\right)^k z_0, \quad k = 0, 1, 2, \ldots.$$

故 $\{z_k(0)\}$ 线性收敛到 0 (亦即其自身的主 p 次根). 因此, 对任意的 $\lambda \in \mathcal{R}_{1,1}$, $z(\lambda)$ 都是其主 p 次根.

同样由第二步知, 对任意的 $\lambda \in \mathcal{R}_{1,2}$, $z(\lambda)$ 是解析的且落在 p 次根函数的单值分支中. 于是, 若我们能证明 $\mathcal{R}_{1,2}$ 包含 $\partial\mathcal{R}_{1,1}$ 某一部分, 那么可知, 当 $\lambda \in \mathcal{R}_{1,2}$ 时的 $z(\lambda)$ 的单值分支与当 $\lambda \in \mathcal{R}_{1,1}$ 时的 $z(\lambda)$ 的单值分支是一样的. 从而可得, 对任意 $\lambda \in \mathcal{R}_{1,1} \bigcup \mathcal{R}_{1,2}$, 有 $z(\lambda)$ 为其主 p 次根. 因此, 我们只须证明 $\mathcal{R}_{1,2}$ 包含了 $\partial\mathcal{R}_{1,1}$ 的某一部分即可.

记 $\lambda_0 = z_0^p$. 定义

$$\begin{aligned} S_{1,m}(z_0, \lambda) &:= \sum_{j=2}^{m} c_{1,j} r^{j-1}(z_0, \lambda), \\ S_\infty(z_0, \lambda) &:= \sum_{j=2}^{\infty} c_{1,j} r^{j-1}(z_0, \lambda), \quad m \geq 2, \lambda \in \partial\mathcal{R}_{1,1}. \end{aligned}$$

显然, $S_\infty(z_0, \lambda)$ 关于 λ 是连续的且

$$0 = |S_\infty(z_0, \lambda_0)| = \min_{\lambda \in \partial\mathcal{R}_{1,1}} |S_\infty(z_0, \lambda)| \leq |S_\infty(z_0, \lambda)| \leq 1, \quad \forall\, \lambda \in \partial\mathcal{R}_{1,1}. \tag{2.42}$$

由于对任意的 $\lambda \in \mathcal{R}_{1,1} \bigcup \mathcal{R}_{1,2}$ 都有 $|S_{1,m}(z_0,\lambda)| < 1$, 故可取实数 $M < 1$ 满足不等式

$$\sup_{m\geq 2} |S_{1,m}(z_0,\lambda)| < M, \quad \lambda \in \partial\mathcal{R}_{1,1}.$$

由 (3.67) 知, 存在 $\delta > 0$ 使得一旦 $|\lambda - \lambda_0| < \delta$ 便有

$$|S_\infty(z_0,\lambda)| < \frac{\frac{1}{M} - 1}{\frac{1}{M} + 1},$$

或其等价情形

$$q_2(z_0,\lambda) = \sup_{m\geq 2} |S_{1,m}(z_0,\lambda)| \cdot \frac{1 + |S_\infty(z_0,\lambda)|}{1 - |S_\infty(z_0,\lambda)|} < M \cdot \frac{1}{M} = 1.$$

因此, $\mathcal{R}_{1,2}$ 包含了 $\partial\mathcal{R}_{1,1}$ 的某一部分. 证完. □

对于分别由 (2.34) 和 (2.35) 定义的 $\mathcal{R}_{1,1}$ 和 $\mathcal{R}_{1,2}$, 若取 $z_0 \equiv 1$, 则有 $\mathcal{R} = \mathcal{R}_{1,1} \bigcup \mathcal{R}_{1,2}$. 于是, 由引理 2.4 可立即得到如下推论:

推论 2.1. *对于任意的 $\lambda \in \mathcal{R}_1$, 其中 $\mathcal{R}_1$ 由 (2.16) 定义, 由标量 Newton 法 (2.12) 以 $z_0 = 1$ 为初始点进行迭代产生的序列 $\{z_k(\lambda)\}$ 收敛于 λ 的主 p 次根 $\lambda^{1/p}$. 进一步, 若 $\lambda \neq 0$, 则其收敛速度是二阶的.*

下面我们考虑矩阵的情形. 对于由 (3.37) 给定的矩阵 $A \in \mathbb{C}^{n\times n}$, 令

$$R(X) := \mathrm{I} - AX^{-p}, \quad p \geq 2, \tag{2.43}$$

其中 $X \in \mathbb{C}^{n\times n}$ 是非奇异的.

引理 2.5. *设矩阵 $X \in \mathbb{C}^{n\times n}$ 是非奇异的且与 A 是可交换的, $R(X)$ 由 (2.43) 定义. 如果 $R(X)$ 的谱满足 $\sigma(R(X)) \subset \mathcal{D}_{0,1}$, 其中 $\mathcal{D}_{0,1}$ 由 (2.15) 定义, 那么由 (2.13) 定义 $N(X)$ 是非奇异的, 与 A 是可交换的且满足*

$$R(N(X)) = \mathrm{I} - \left[\mathrm{I} - \frac{1}{p}R(X)\right]^{-p} \cdot (\mathrm{I} - R(X)) = \phi_1(R(X)), \tag{2.44}$$

其中 ϕ_1 由 (2.14) 定义.

证明. 显然, 若 $X \in \mathbb{C}^{n\times n}$ 是非奇异矩阵, 则由 (2.13) 定义的 $N(X)$ 是存在的. 由于 $\sigma(R(X)) \subset \mathcal{D}_{0,1}\backslash\{0\}$, 根据 Neumann 引理可知矩阵

$$\mathrm{I} - \frac{1}{p}R(X)$$

是非奇异的. 于是由等式

$$N(X) = \frac{1}{p}X\left[(p-1)\mathrm{I} + AX^{-p}\right] = X\left[\mathrm{I} - \frac{1}{p}R(X)\right]$$

知矩阵 $N(X)$ 是非奇异的且与 A 是可交换的. 由此可进一步得到

$$R(N(X)) = \mathrm{I} - \left[\mathrm{I} - \frac{1}{p}R(X)\right]^{-p} \cdot (\mathrm{I} - R(X)) = \phi_1(R(X))$$

上述第一个等式是根据 X 与 A 是可交换的假设得到的. 证完. □

根据引理 2.5 可直接得到如下推论:

推论 2.2. *如果矩阵 $X_0 \in \mathbb{C}^{n\times n}$ 与 A 是可交换的且 $R(X_0)$ 的谱满足 $\sigma(R(X_0)) \subset \mathcal{D}_{0,1}$, 其中 $\mathcal{D}_{0,1}$ 由 (2.15) 定义, 那么由 Newton 法 (2.12) 以 X_0 为初始点进行迭代产生的矩阵序列 $\{X_k\}$ 是有定义的.*

引理 2.6. *设 $R(X)$ 由 (2.43) 定义, $X \in \mathbb{C}^{n\times n}$ 是非奇异的. 如果 $R(X)$ 的谱满足 $\sigma(R(X)) \subset \mathcal{D}_1\backslash\{0\}$, 其中 $\mathcal{D}_1$ 由 (2.18) 定义, 那么存在一个次可加的矩阵范数 $\|\cdot\|$ 使得 $\|R(X)\| \le 1$ 且*

$$\|R(E(X))\| \le \frac{\phi_1(\|R(X)\|)}{\|R(X)\|^2} \cdot \|R(X)\|^2 < \|R(X)\|^2, \tag{2.45}$$

其中 ϕ_1 由 (2.14) 定义.

证明. 由引理 2.5 易知 $R(N(X))$ 是存在的. 因 $\sigma(R(X)) \subset \mathcal{D}_1$, 故 $R(X)$ 的谱半径小于 1. 从而存在一个次可加的矩阵范数 $\|\cdot\|$ 使得 $\|R(X)\| < 1$. 注意到, 对于任意的 $u \in (0,1)$ 都有

$$\frac{\phi(u)}{u^2} = \sum_{j=2}^{\infty} c_{1,j}u^{j-2} < \sum_{j=2}^{\infty} c_{1,j} = 1.$$

于是, 结合 (2.44) 可得

$$\|R(N(X))\| = \|\phi_1(R(X))\| \le \phi_1(\|R(X)\|)$$

$$
\begin{aligned}
&= \frac{\phi_1(\|R(X)\|)}{\|R(X)\|^2}\|R(X)\|^2 \\
&< \|R(X)\|^2,
\end{aligned}
$$

因此 (2.45) 是成立的. 证完. □

引理 2.7. 设 $R(X)$ 由 (2.43) 定义, 矩阵 $X_0 \in \mathbb{C}^{n\times n}$ 是非奇异的. 设 $R(X_0)$ 的谱满足 $\sigma(R(X_0)) \subset \mathcal{D}_1 \backslash \{0\}$ 且存在一个次可加性的矩阵范数 $\|\cdot\|$ 使得 $\|R(X_0)\| < 1$, 其中 $\mathcal{D}_1$ 由 (2.18) 定义. 令 $\{X_k\}$ 是由 *Newton* 法 (2.12) 以 X_0 为初始点进行迭代产生的矩阵序列. 则有

$$
\|R(X_k)\| \le q^{2^k-1}(X_0) \cdot \|R(X_0)\|, \quad k = 1, 2, \ldots, \tag{2.46}
$$

其中

$$
q(X_0) := \frac{\phi(\|R(X_0)\|)}{\|R(X_0)\|} < 1, \tag{2.47}
$$

而 ϕ_1 由 (2.14) 定义.

证明. 对于取定的矩阵 X_0, 由引理 2.6 中的 (2.45) 可得

$$
\|R(X_1)\| \le \frac{\phi_1(\|R(X_0)\|)}{\|R(X_0)\|^2}\|R(X_0)\|^2 = q(X_0) \cdot \|R(X_0)\|.
$$

如果对于某个整数 $k \ge 1$, (2.46) 是成立的, 那么由引理 2.6 有

$$
\begin{aligned}
\|R(X_{k+1})\| &\le \frac{\phi_1(\|R(X_k)\|)}{\|R(X_k)\|^2}\|R(X_k)\|^2 \\
&\le \frac{\phi_1(\|R(X_0)\|)}{\|R(X_0)\|^2}\left[q^{2^k-1}(X_0)\right]^2 \cdot \|R(X_0)\|^2 \\
&= [q(X_0)]^{2^{k+1}-1} \cdot \|R(X_0)\|.
\end{aligned}
$$

因而, 由归纳法知, 对于任意的 $k \ge 1$, (2.46) 都是成立的. 证完. □

引理 2.8. 设 $R(X)$ 由 (2.43) 定义, 矩阵 $X_0 \in \mathbb{C}^{n\times n}$ 是非奇异的. 如果 $R(X_0)$ 的谱满足 $\sigma(R(X_0)) \subset \mathcal{D}_{2,1} \bigcap \mathcal{D}_{0,1} \backslash \{0\}$, 其中 $\mathcal{D}_{0,1}$ 和 $\mathcal{D}_{2,1}$ 分别由 (2.15) 和 (2.19) 定义. 设 $\{X_k\}$ 是由 *Newton* 法 (2.12) 以 X_0 为初始点进行迭代产生的矩阵序列. 则存在整数 $\widehat{N} > 0$ 使得

$$
\|R(X_k)\| \le [q(X_{\widehat{N}})]^{2^{k-\widehat{N}}-1} \cdot \|R(X_{\widehat{N}})\|, \quad \forall\ k > \widehat{N}, \tag{2.48}
$$

其中

$$q(X_{\widehat{N}}) := \frac{\phi(\|R(X_{\widehat{N}})\|)}{\|R(X_{\widehat{N}})\|} < 1,$$

而 ϕ_1 由 (2.14) 定义.

证明. 对于任意的 $r(z_0) \in \sigma(R(X_0))$, 设 $\{z_k\}$ 是由标量 Newton 法 (2.12) 以 z_0 为初始点进行迭代产生的复序列. 则由引理 2.3 知, 存在整数 $N > 0$ 使得 $|r(z_N)| < 1$. 于是, 由引理 2.2 可得

$$|r(z_k)| < |r(z_N)|^{2^{k-N}}, \quad \forall\ k > N.$$

定义

$$\widehat{N} := \max_{r(z_0)\in\sigma(R(X_0))} \{N : \text{choose a } N > 0 \text{ such that } |r(z_N)| < 1\}.$$

则存在一个次可加性的矩阵范数 $\|\cdot\|$ 使得 $\|X_{\widehat{N}}\| < 1$. 于是, 由引理 2.7 可知 (2.48) 是成立的. 证完. □

2.7　定理 2.3 的证明

下面的引理来自 [? , 定理 4.15]. 根据该引理并结合上节中的引理即可证明我们的收敛定理.

引理 2.9 ([? , 定理 4.15]). *设 $g(x,t)$ 是一个双变量的有理函数, $x^* = f(\lambda)$ 是如下迭代格式的一个吸引固定点:*

$$x_{k+1} = g(x_k, \lambda), \quad x_0 = \phi_0(\lambda),$$

其中 ϕ_0 是一个有理函数而 $\lambda \in \mathbb{C}$. 则由迭代格式

$$X_{k+1} = g(X_k, J(\lambda)), \quad X_0 = \phi_0(J(\lambda))$$

迭代产生的矩阵序列 $\{X_k\}$ 收敛于满足如下关系的矩阵 X^*:

$$(X^*)_{ii} \equiv f(\lambda), \quad i = 1, 2, \ldots, m,$$

其中 $J(\lambda) \in \mathbb{C}^{m\times m}$ 为 Jordan 块.

下面应用引理 3.10 来证明收敛性定理 2.3.

定理 2.3 的证明. 首先由推论 2.2 知, 当矩阵 A 的所有特征值都属于 $\mathcal{R}_1 \subset \mathcal{D}_{0,1}$, 则由 Newton 法 (2.12) 以 $X_0 = \mathrm{I}$ 为初始点进行迭代产生的矩阵序列 $\{X_k\}$ 是有定义的. 再由引理 3.10 及推论 2.1 知, 矩阵序列 $\{X_k\}$ 收敛于矩阵 A 的主 p 次根 $A^{1/p}$.

对于 $k \geq 0$, 令 $X_* = A^{1/p}$ 及 $E_k = X_k - X_*$. 由于 X_k 与 X_* 是可交换的, 则有

$$\begin{aligned} R(X_k) &= \mathrm{I} - AX_k^{-p} = (X_k^p - X_*^p)X_k^{-p} \\ &= (X_k - X_*)\left(X_k^{p-1} + X_k^{p-2}X_* + \cdots + X_kX_*^{p-2} + X_*^{p-1}\right)X_k^{-p} \\ &= E_k\left(X_k^{p-1} + X_k^{p-2}X_* + \cdots + X_kX_*^{p-2} + X_*^{p-1}\right)X_k^{-p}, \quad \forall\, k \geq 0 \qquad (2.49) \end{aligned}$$

记

$$Y_k := \sum_{i=1}^{p} X_k^{p-i}X_*^{i-1} = X_k^{p-1} + X_k^{p-2}X_* + \cdots + X_kX_*^{p-2} + X_*^{p-1}.$$

因 X_k 收敛于 $A^{1/p}$ 且 $A^{1/p}$ 的所有特征值均不属于 $\mathbb{R}^-$, 故存在非负整数 $N > 0$ 使得对于任意的 $k \geq N$, X_k 的特征值均不属于 $\mathbb{R}^-$. 进而知 Y_k 的特征值亦不在 $\mathbb{R}^-$, 从而对任意的 $k \geq N$ 矩阵 Y_k 都是非奇异的. 于是由 (3.74) 可得

$$E_{k+1} = R(X_{k+1})X_{k+1}^pY_{k+1}^{-1}, \quad k \geq N. \qquad (2.50)$$

由 (2.46) 和 (2.48) 可知, 存在 $K_0 > 0$ 使得对于所有的 $k \geq K_0$ 都有 $\|R(X_{k+1})\| < \|R(X_k)\|^2$. 所以, 应用 (3.74) 和 (3.75) 可得

$$\begin{aligned} \|E_{k+1}\| &\leq \|R(X_{k+1})\|\|X_{k+1}\|^p\|Y_{k+1}^{-1}\| \\ &< \|R(X_k)\|^2\|X_{k+1}\|^p\|Y_{k+1}^{-1}\| \\ &\leq \left(\|X_k^{-1}\|^p\|X_{k+1}\|^p\|Y_k\|\|Y_{k+1}^{-1}\|\right)\|E_k\|^2, \quad \forall\, k \geq K_0. \qquad (2.51) \end{aligned}$$

注意到因为 $\{X_k\}$ 是收敛的, 所以对于所有的 $k \geq K_0$, $\|X_k^{-1}\|^p\|X_{k+1}\|^p\|Y_k\|\|Y_{k+1}^{-1}\|$ 是有界的. 因此, 由 (3.76) 知收敛速度是二阶的. 证完. □

2.8　Numerical examples

Newton's method (2.12) and Halley's method (3.84) are usually not stable in a neighborhood of the principal pth root of A, see [?] or [?] for the stability analysis on Newton's method. Thus, these two iterative methods cannot be used directly to computing $A^{1/p}$. A stable version of Newton's method by introducing the auxiliary matrix N_k has been given in [?] as follows:

$$\begin{cases} X_{k+1} = X_k \left(\dfrac{(p-1)\mathrm{I} + N_k}{p} \right), & X_0 = \mathrm{I}, \\ N_{k+1} = \left(\dfrac{(p-1)\mathrm{I} + N_k}{p} \right)^{-p} N_k, & N_0 = A. \end{cases} \tag{2.52}$$

Clearly, $N_k = AX_k^{-p}$ and $\{X_k\}$ generated by (2.52) is same as the sequence of Newton's method (2.12). We call it coupled Newton iteration. When the sequence $\{X_k\}$ generated by (2.52) converges to $A^{1/p}$, N_k converges to I.

For Halley's method, a stable version has been given in [?] as follows:

$$\begin{cases} X_{k+1} = X_k \big((p+1)\mathrm{I} + (p-1)N_k\big)^{-1} \big((p-1)\mathrm{I} + (p+1)N_k\big), & X_0 = \mathrm{I}, \\ N_{k+1} = N_k \left(\big((p+1)\mathrm{I} + (p-1)N_k\big)^{-1} \big((p-1)\mathrm{I} + (p+1)N_k\big) \right)^{-p}, & N_0 = A. \end{cases} \tag{2.53}$$

Also, $\{X_k\}$ generated by (2.53) is same as the sequence of Halley's method (3.84). We call it coupled Halley iteration. When the sequence $\{X_k\}$ generated by (2.53) converges to $A^{1/p}$, N_k converges to I.

算法 2.1 Preprocessing iterative framework for computing $A^{1/p}$

Given $A \in \mathbb{C}^{n\times n}$ with no nonpositive real eigenvalues, an integer $p = 2^{k_0}q$ with $k_0 \geq 0$ and q odd. This algorithm computes the principal pth root of matrix A via a Schur decomposition and some iterative method.

1. Compute the Schur decomposition of $A = QRQ^*$;

2. If $q = 1$, then $k_1 = k_0$; else choose the smallest $k_1 \geq k_0$ such that for each eigenvalue λ of matrix A, $\lambda^{1/2^{k_1}}$ belongs to some region $\mathcal{R}$;

3. Compute $B = R^{1/2^{k_1}}$ by taking the square root k_1 times; if $q = 1$, then set $X = QBQ^{\mathrm{T}}$; else continue;

4. Compute $C = B^{1/q}$ by using some iterative method and set $X = QC^{2^{k_1-k_0}}Q^{\mathrm{T}}$.

To improve the convergence of iterative methods for computing $A^{1/p}$, an effective way is of using a preprocessing as in [?] so that the eigenvalues of the new matrix are in a smaller convergence region. Based on this way, a framework that computes the principal pth root of matrix $A \in \mathbb{C}^{n\times n}$ by using Schur decomposition and some iterative method is summarized as Algorithm A number of modifications to this framework are possible. For example, we can obtain an algorithm called Schur-Newton algorithm by choosing $\mathcal{R}_1^{\mathrm{N}}$ defined in (2.54) as the region $\mathcal{R}$ in step 2 and the coupled Newton iteration (2.52) as the iterative method in step 4.

注 2.2. In practice, it is not feasible to check whether a eigenvalue λ belongs to $\mathcal{R}_1$. This is due to the computational cost of (2.19) may be large even if we only choose $m = 20$. Thus, based on the observation from Figure 2.5, we now give a new convergence region which allows us to determine easily whether a eigenvalue belongs to it. Define

$$\mathcal{R}_1^{\mathrm{N}} = \mathcal{D}_3 \bigcup \mathcal{D}_4, \tag{2.54}$$

where

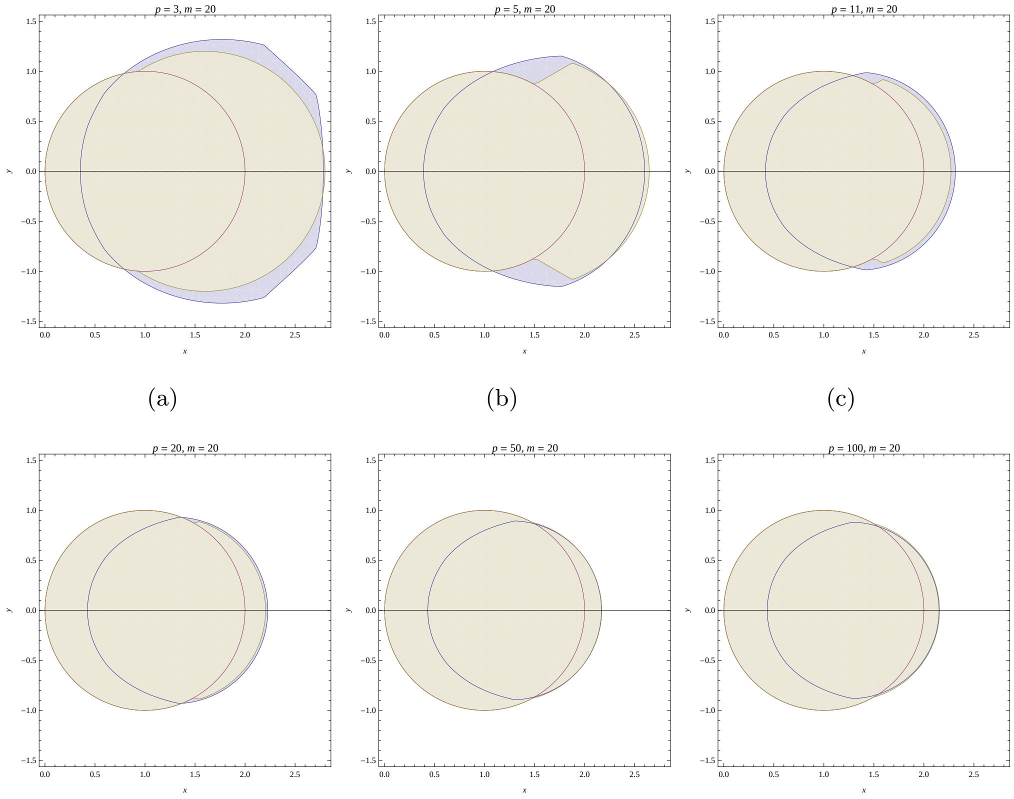
$$\mathcal{D}_3 := \{z \in \mathbb{C} : 1 - z \in \overline{\mathcal{D}}_1\}, \tag{2.55}$$

$$\mathcal{D}_4 := \begin{cases} \left\{ z \in \mathbb{C} : \left| \frac{8}{5} - z \right| < \frac{6}{5} \right\}, & p = 2, 3, 4, \\ \left\{ z \in \mathbb{C} : |\arg(z)| < \frac{\pi}{6}, \left| \frac{5p-13}{4(p-3)} - z \right| < \frac{43p-105}{48(p-3)} \right\}, & p \geq 5. \end{cases}$$

In Figure 2.6, we present the regions $\mathcal{R}_1$ and $\mathcal{R}_1^{\mathrm{N}}$ defined in (2.16) and (2.54), respectively, for Newton's method. We can observe that the new region $\mathcal{R}_1^{\mathrm{N}}$ is acceptable approximation to $\mathcal{R}_1$. So instead of using $\mathcal{R}_1$, in Section 4, we will use $\mathcal{R}_1^{\mathrm{N}}$ in our algorithm and numerical experiments.

Next, we give some numerical examples to illustrate that the convergence results obtained in Theorems 2.3 and 3.2 are better than the existing ones. For simplicity, in the following tests, we denote

1. SN: Schur-Newton algorithm by choosing $\mathcal{R}_1^{\mathrm{N}}$ defined in (2.54) as the region $\mathcal{R}$ in step 2 and the coupled Newton iteration (2.52) as the iterative method in step 4;



(a) (b) (c)

(d) (e) (f)

图 2.6: For fixed $m = 20$ and $p = 2, 5, 11, 20, 50, 100$, the actual convergence regions $\mathcal{R}_1$ defined in (2.16) (the union of the red and blue parts) and the approximate convergence regions $\mathcal{R}_1^{\mathrm{N}}$ defined in (2.54) (the yellow parts).

2. SN-old: Schur-Newton algorithm by choosing $\mathcal{D}_3$ defined in (2.55) as the region $\mathcal{R}$ in step 2 and the coupled Newton iteration (2.52) as the iterative method in step 4;

3. SH: Schur-Halley algorithm by choosing $\mathcal{R}_2^{\mathrm{H}}$ defined in (3.92) as the region $\mathcal{R}$ in step 2 and the coupled Halley iteration (2.53) as the iterative method in step 4;

4. SH-old: Schur-Halley algorithm by choosing the disk $\{z \in \mathbb{C} : |8/5 - z| \le 1\}$ as the region $\mathcal{R}$ in step 2 and the coupled Halley iteration (2.53) as the iterative method in step 4.

Recall that the convergence regions of SN-old and SH-old are presented by Guo in [**?**] and Iannazzo in [**?**], respectively. The iterations in the above four algorithms are stopped when $\|N_k - \mathrm{I}\| < \sqrt{n}u/2$, where n is the size of A and $u = 2^{-52} \approx 2.2204\mathrm{e} - 16$.

Our numerical experiments were carried out in MATLAB 7.0 running on a PC Intel Pentium P6200 of 2.13 GHz CPU. To measure of the quality of a computed solution X, we use the relative residual $\rho_A(X)$ and relative error $\mathrm{err}(X)$ as follows:

$$\rho_A(X) = \frac{\|A - X^p\|}{\|X\| \|\sum_{j=0}^{p-1} (X^{p-1-j})^{\mathrm{T}} \otimes X^j\|}, \quad \mathrm{err}(X) = \frac{\|A - X^p\|}{\|A\|}, \qquad (2.56)$$

where $\otimes$ denotes the Kronecker product and $\|\cdot\|$ denotes the Frobenius norm. Note that the relative residual $\rho_A(X)$ (given in [**?**]) is more practically useful definition of relative residual (e.g., for testing purposes) than relative error and that the averaged CPU time computed by the standard MATLAB function cputime. The averaged time was computed by repeating 100 times for each test matrix. Moreover, we use ‘iter’ to stand for the number of the iterations.

第 3 章　基于4P1-P1元求解Navier-Stokes方程的移动网格方法

根据上一章, 我们知道混合元是求解不可压Navier-Stokes方程的一种重要方法。为了保证解的存在唯一性，需要使速度和压力两个解所在的空间满足LBB条件。其中一种办法是使得速度空间相对压力空间来说，自由度足够多，例如mini元，Taylor-Hood元等等。另外一中方法是对压力空间施加一些约束，比如说稳定化的P_1-P_1元，P_1-P_0元。为了减少计算量，提高计算效率，通常会运用自适应网格方法。在文献[?],[?] [?]中，运用了h-自适应的P2P1元。在实际的工程计算中，我们通常倾向于使用线性元，而非高次元。[?]提出自适应网格和稳定化P1P1元、P1P0元相结合的策略. [?]将移动网格方法应用到不可压Navier-Stokes 方程的求解。我们移动网格部分的策略是基于文献[?]中的工作。然而，不满足LBB条件的有限元对应用到自适应网格方法上有一定的技术困难。综合以上考虑，我们选取稳定的$P_1isoP_2P_1$元，它自然满足LBB条件，详细见[?]. 在P1ISOP2P1元，速度单元所在的网格可以有压单元的网格加密一次得到，如Figure (2.1)注意到在速度单元$\boldsymbol{u}$ 和压力单元p 上均是线性元。但是，速度单元网格上的6个基函数不在同一个速度单元上。因此在拼装散度块矩阵，即拼装压力单元和速度单元，这个过程并不显然。

在本文的工作中，我们选取$4P1-P1$元，它与$P_1isoP_2P_1$拥有相同的网格结构，自然也满足inf-sup条件，但需要指出的是$P_1isoP_2P_1$元的速度单元上的基函数都局部的位于相同的速度单元内。这给我们拼装散度块矩阵提供了便利: 只要我们建立四个速度单元和大的压力单元之间的索引。索引的建立又依赖于两层网格的数据结构，这种两层网格用我们上一章提到的几何遗传树结构，可以建立两层网格间的对应关系。

基于这种网格遗传树的结构，我们可以很容易的建立四个小的速度单元与压力宏单元间的索引。同时在[?]中h-自适应方法也是基于这种树结构。自然的，我们可以将h-自适应方法应用到4P1-P1元上。自适应网格可以减少计算量，同时可以研究流体局部小尺度上的现象，比如涡。同时，移动网格网格方法也可以应用在4P1-P1元上，网格移动我们只移动压力网格，当压力网格移动

完，速度网格由压力网格加密一次即可得到。但是需要注意的是，求解不可压的Navier-Stokes方程，在将旧网格的解插值到移动后的网格的过程，要满足散度为0的条件，这个工作由[?]中给出。

在第一节中，我们介绍Navier-Stokes方程的知识。在第二节中我们展示4P1-P1的数据结构，以及单元间索引的建立过程，接下来我们用4P1-P1元近似Navier-Stokes方程。移动网格的策略将在第四节中给出。最后，我们给出数值例子。

3.1　数据结构

4P1-P1是基于两套不同的网格和两种有限元空间。速度网格可以由压力网格全局加密一次得到。网格的数据结构是基于[?]中的几何遗传树结构，如Figure(2.4)所示。一个宏压力单元对应这四个速度单元，通过遍历一次所有的速度单元，利用几何遗传树结构，就可以建立速度单元和压力单元的$1-1$对应。

在AFEPack中，速度网格和压力网格分别存储在两张非正则网格irregularMeshV和irregularMeshP上，这两张非正则网格都是建立在同一颗树上。非正则网格上，可以进行全局加密，局部加密以及疏化的操作，能进行这种操作得益于它的四叉树结构。所有的网格节点从根节点到叶子节点，都存储在非正则网格中。这里irregularMeshV 是由irregularMeshP进行全局一次加密得到的，根据非正则网格的树结构，我们可以通过遍历irregularMeshV的全部活动单元(即叶子节点)，通过活动单元来找到它的父亲单元(即叶子节点的父亲节点)。这样我们建立了从速度网格单元到压力网格单元之间的单向索引。还是根据活动单元的父亲节点也是具有树结构的，因此，我们可以根据父子节点单元(压力单元)来找到它对应的四个儿子单元(即四个速度单元)，这样我们也建立了从压力单元到速度单元的单向索引。到这里，速度和压力间的单元已经建立了$1-1$索引，注意到，只要不产生新的网格单元，我们建立的索引不需要重新改。应用到移动网格上，我们只需要一次构建索引就可以，不管网格如何移动，网格间的索引不会改变。建立索引的过程参见算法(3.1)。

Algorithm 3.1 构建速度单元和压力单元间的索引

```
achieveiterator $\leftarrow$ irregualerMeshV.beginActiveElement()
enditerator $\leftarrow$ irregualerMeshV.endActiveElement()
while achieveiterator $\neq$ enditerator do
    int index-v-element $\leftarrow$ achieveiterator$-\rangle$ index
    HElement $\langle$ DIM, DIM$\rangle *$ parent $\leftarrow$ activeiterator$-\rangle$ parent
    int index-p-element $\leftarrow$ parent$-\rangle$ index
    int n-child $\leftarrow$ parent$-\rangle$ n-child
    index-p2v[index-p-element].resize(n-child)
    while $i \geq 0$ and $i <$ n-child do
        HElement$\langle$DIM,DIM$\rangle$ $*$chi $\leftarrow$ parent$-\rangle$child[i]
        int index-v-element $\leftarrow$ child$-\rangle$index
        index-p2v[index-p-element][i] $\leftarrow$ index-v-element
        index-v2p[index-v-element] $\leftarrow$ index-p-element
    end while
end while
```

3.2　混合元近似

3.2.1　流体方程

在Stokes方程的基础上加一个对流项，我们可以得到一个稳态的Navier-Stokes方程

$$\begin{aligned} -\nu\nabla^2\boldsymbol{u}+\boldsymbol{u}\cdot\nabla\boldsymbol{u}+\nabla p = &\ \boldsymbol{f}, \\ \nabla\cdot\boldsymbol{u}=0. \end{aligned} \tag{3.1}$$

其中$\nu > 0$ 是一个常数，称作动力学粘性系数。跟Stokes 方程中类似，$\boldsymbol{u}$表示流体速度，p表示压力。对流项$\boldsymbol{u}\cdot\nabla\boldsymbol{u} := (\boldsymbol{u}\cdot\nabla)\boldsymbol{u}$，是非线性项，这也使得Navier-Stokes方程的解不唯一，这也给我们的数值计算带来了一定的挑战。系统(3.1)的计算区域是Ω，可以是二维的或三维的。在边界$\partial\Omega = \partial\Omega_D \cup \Omega_N$上的边界条件如下：

$$\begin{aligned} \boldsymbol{u}=\boldsymbol{w}, &\qquad \text{on } \partial\Omega_D, \\ \nu\frac{\partial\boldsymbol{u}}{\partial n}-p=\boldsymbol{0}, &\qquad \text{on } \partial\Omega_N. \end{aligned} \tag{3.2}$$

其中边界$\partial\Omega = \partial\Omega_D \bigcup \partial\Omega_N$, $\boldsymbol{n}$表示边界的外法向。Dirichlet边界根据边界上的速度和外法向的乘积，可以细分为：

$$\begin{aligned} \partial\Omega_+ &= x\text{在}\partial\Omega\text{上}|\boldsymbol{\omega}\cdot\boldsymbol{n} > 0, \quad \text{出流边界} \\ \partial\Omega_0 &= x\text{在}\partial\Omega\text{上}|\boldsymbol{\omega}\cdot\boldsymbol{n} = 0, \quad \text{特征边界} \\ \partial\Omega_- &= x\text{在}\partial\Omega\text{上}|\boldsymbol{\omega}\cdot\boldsymbol{n} < 0, \quad \text{入流边界} \end{aligned} \tag{3.3}$$

如果边界全部是Direchlet边界条件，即$\partial\Omega = \partial\Omega_D$，那么Navier-Stokes问题(3.1)和(3.2)的压力解除去一个常数外是唯一的。我们对(3.1)中的不可压约束应用散度定理，

$$0 = \int_\Omega \nabla\cdot\boldsymbol{u} = \int_{\partial\Omega} \boldsymbol{u}\cdot\boldsymbol{n} = \int_\Omega \boldsymbol{\omega}\cdot\boldsymbol{n}. \tag{3.4}$$

即边界值要满足相容性条件

$$\int_{\partial\Omega_+} \boldsymbol{\omega}\cdot\boldsymbol{n} + \int_{\partial\Omega_-} \boldsymbol{\omega}\cdot\boldsymbol{n} = 0. \tag{3.5}$$

简单的讲，就是说流入Ω的流体的体积，要与流出的体积相等。这也是压力不唯一的原因。在处理入流/出流的问题时，要注意保证这个相容性条件，否则Navier-Stokes问题的解有可能不存在。一般情况下，我们在出流边界设置自然条件，相容性条件会自然满足，因此这时候问题(3.1)和(3.2)的压力解是唯一的。

(3.1)是Navier-Stokes的原始变量形式我们定义解和检验空间如下：

$$\mathbf{H}^1_E := \left\{\boldsymbol{u}\in\mathcal{H}^1(\Omega)^d \middle| \boldsymbol{u} = \boldsymbol{w} \text{ on } \partial\Omega_D\right\}, \tag{3.6}$$

$$\mathbf{H}^1_{E_0} := \left\{\boldsymbol{v}\in\mathcal{H}^1(\Omega)^d \middle| \boldsymbol{u} = \mathbf{0} \text{ on } \partial\Omega_D\right\}, \tag{3.7}$$

那么变分形式为：寻找$(\boldsymbol{u}, p) \in (\mathbf{H}^1_E, L_2(\Omega))$ 使得

$$\int_\Omega \frac{\partial\boldsymbol{u}}{\partial t}\cdot\boldsymbol{v} + \nu\int_\Omega \nabla\boldsymbol{u} : \nabla\boldsymbol{v} + \int_\Omega (\boldsymbol{u}\cdot\nabla\boldsymbol{u})\cdot\boldsymbol{v} - \int p\,(\nabla\cdot\boldsymbol{v}) = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}, \quad \forall\boldsymbol{v}\in\mathbf{H}^1_{E_0}, \tag{3.8}$$

$$\int_\Omega q\,(\nabla\cdot\boldsymbol{u}) = 0, \quad \forall q\in L_2(\Omega). \tag{3.9}$$

其中$\nabla\boldsymbol{u} : \nabla\boldsymbol{v}$ 表示纯量的乘积，在二维中为$\nabla u_x\cdot\nabla v_x + \nabla u_y\cdot\nabla v_y$.

假设τ_h是Ω 上对压力网格的三角剖分，网格尺度$h = max_{T\in\tau_h} diam(T)$, T 为三角剖分τ_h 的单元。对应的，$\tau_{\frac{h}{2}}$ 是对速度网格的三角剖分。基于$\tau_{\frac{h}{2}}$ 和τ_h 上的有限元空间X_E^h 和P_h 满足

$$X_E^h \subset \mathcal{H}_E, \quad P_h \subset L_2(\Omega)$$

那么(3.8) 和(3.9) 可以写成如下形式: 寻找$(\boldsymbol{u}_h, p_h) \in X_E^h \times P_h$ 使得

$$\begin{aligned}
&\int_\Omega \frac{\partial \boldsymbol{u}_h}{\partial t}\cdot \boldsymbol{v}_h + \nu \int_\Omega \nabla \boldsymbol{u}_h : \nabla \boldsymbol{v}_h \\
+&\int_\Omega (\boldsymbol{u}_h\cdot\nabla\boldsymbol{u}_h)\cdot\boldsymbol{v}_h - \int_\Omega p_h\left(\nabla\cdot\boldsymbol{v}_h\right) &&= \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}_h, \qquad \forall \boldsymbol{v}_h \in \mathbf{X}_0^h; \\
&\int_\Omega q_h\left(\nabla\cdot\boldsymbol{u}_h\right) &&=0, \qquad \forall q_h \in P^h.
\end{aligned} \tag{3.10}$$

时间方向至少三阶Runge-Kutta方法才能保证数值稳定性。为了简便，我们这里只用显示Euler格式：$\forall(\boldsymbol{v}_h, q_h) \in \mathbf{X}_0^h \times P^h$

$$\begin{aligned}
\int_\Omega \frac{\boldsymbol{u}_h^{(n+1)} - \boldsymbol{u}_h^{(n)}}{\delta t} + \nu\int_\Omega \nabla\boldsymbol{u}_h^{(n+1)} : \nabla\boldsymbol{v}_h - \int_\Omega p_h^{(n+1)}\left(\nabla\cdot\boldsymbol{v}_h\right) &= \int_\Omega \left(\boldsymbol{u}_h^{(n)}\cdot\nabla\boldsymbol{u}_h^{(n)}\right) \\
\int_\Omega q_h \nabla\cdot\boldsymbol{u}_h^{(n+1)} &= 0.
\end{aligned} \tag{3.11}$$

令$\{\phi_j\}_{j=1}^n$ 和$\{\psi_k\}_{k=1}^m$ 分别为速度和压力的线性元基函数。则数值解$\boldsymbol{u}_h^{(n+1)} = (u_{xh}^{(n+1)}, u_{yh}^{(n+1)}), p_h$ 可以写成如下形式：

$$u_{xh}^{(n+1)} = \sum_{j=1}^n u_j\phi_j, \quad u_{yh}^{(n+1)} = \sum_{j=1}^n v_j\phi_j, \quad p_h^{(n+1)} = \sum_{k=1}^m p_k\psi_k$$

将$u_{xh}^{(n+1)}, u_{yh}^{(n+1)}, p_h^{(n+1)}$带入离散弱形式(3.11) 中，可以得到线性方程组

$$\begin{bmatrix} \frac{1}{dt}M + \nu A & 0 & B_x^T \\ 0 & \frac{1}{dt}M + \nu A & B_y^T \\ B_x & B_y & 0 \end{bmatrix}\begin{bmatrix} u_x \\ u_y \\ p \end{bmatrix} = \begin{bmatrix} f_x \\ f_y \\ g \end{bmatrix}, \tag{3.12}$$

其中M 是$n\times n$ 的质量矩阵，A是拉普拉斯矩阵，由以下形式：

$$A \;=\; [a_{ij}], \quad a_{ij} = \int_\Omega \nabla\phi_i\cdot\nabla\phi_j$$

$$
\begin{aligned}
M &= [m_{ij}], \quad m_{ij} = \int_\Omega \phi_i \phi_j \\
B_x^T &= [bx_{ik}^T], \quad bx_{ik}^T = \int_\Omega \psi_k \frac{\partial \phi_i}{\partial x} \\
B_y^T &= [by_{ik}^T], \quad by_{ik}^T = \int_\Omega \psi_k \frac{\partial \phi_i}{\partial y} \\
f_x &= [f_i], \quad f_i = \int_\Omega (\frac{u_{xh}^{(n)}}{dt} - (u_{xh}^{(n)} \frac{\partial u_{xh}^{(n)}}{\partial x} + u_{yh}^{(n)} \frac{\partial u_{xh}^{(n)}}{\partial y}) \phi_i \\
f_y &= [f_i], \quad f_i = \int_\Omega (\frac{u_{yh}^{(n)}}{dt} - (u_{xh}^{(n)} \frac{\partial u_{yh}^{(n)}}{\partial x} + u_{yh}^{(n)} \frac{\partial u_{yh}^{(n)}}{\partial y}) \phi_i \\
g &= [g_k], \quad g_i = 0.
\end{aligned}
$$

接下来我们着重解释一下散度块矩阵$B_x B_y, B_x^T, B_y^T$的拼装。以B_x^T为例，令$\triangle_{v_i}$ 为一个速度单元，我们可以通过上一节建立的单元索引来找到对应的压力单元$\triangle_{p_k}$(如Figure (3.1)). $\phi_{i_1}, \phi_{i_2}, \phi_{i_3}$ 是定义在速度单元$\triangle_{v_i}$ 上的线性元基函数，下标i_1, i_2, i_3 表示该顶点上的自由度在速度全部自由度的全局编号。同时$\psi_{k_1}, \psi_{k_2}, \psi_{k_3}$ 是定义在压力单元$\triangle_{p_k}$ 上的线性元基函数，k_1, k_2, k_3 代表单元$\triangle_{p_k}$ 上的自由度在压力全部自由度中的全局编号。因此单元$\triangle_{p_k}$ 和$\triangle_{v_i}$ 对B_x^T的贡献为

$$
\begin{bmatrix}
\int_{\triangle_{v_i}} \psi_{k_1} \frac{\partial \phi_{i_1}}{\partial x} & \int_{\triangle_{v_i}} \psi_{k_2} \frac{\partial \phi_{i_1}}{\partial x} & \int_{\triangle_{v_i}} \psi_{k_3} \frac{\partial \phi_{i_1}}{\partial x} \\
\int_{\triangle_{v_i}} \psi_{k_1} \frac{\partial \phi_{i_2}}{\partial x} & \int_{\triangle_{v_i}} \psi_{k_2} \frac{\partial \phi_{i_2}}{\partial x} & \int_{\triangle_{v_i}} \psi_{k_3} \frac{\partial \phi_{i_2}}{\partial x} \\
\int_{\triangle_{v_i}} \psi_{k_1} \frac{\partial \phi_{i_3}}{\partial x} & \int_{\triangle_{v_i}} \psi_{k_2} \frac{\partial \phi_{i_3}}{\partial x} & \int_{\triangle_{v_i}} \psi_{k_3} \frac{\partial \phi_{i_3}}{\partial x}
\end{bmatrix} \tag{3.13}
$$

将贡献矩阵(3.13) 加到B_x^T 中的相应位置(i_1, k_1), (i_1, k_2), (i_1, k_3), (i_2, k_1), (i_2, k_2), (i_2, k_3), (i_3, k_1), (i_3, k_2), (i_3, k_3) to B_x^T. 上。通过遍历所有的速度单元，重复上面的操作，B_x^T 拼装完成。B_y^T 也可以同样拼装。

B_x 和B_y 的拼装是遍历所有的压力单元，通过单元索引，找到对应的四个速度单元，然后分别用压力单元和四个速度单元进行拼装。过程相似，便不再赘述。

我们主要的想法是有两套网格，速度网格和压力网格，速度网格可以由压力网格加密一次得到。因此我们对压力网格进行移动，然后对移动完的压力网格全局加密一次，即可实现对速度网格的移动。为了更好的说明我们的数值方法，我们将算法的流程图如Algorithm(3.2)所示.

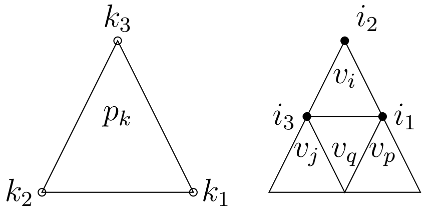


图 3.1: 左: p 单元; 右: 与压力单元对应的4个速度v单元

3.3　移动网格策略

在$t = t_{n+1}$ 时刻，用上一章的方法可以得到有限元解$(\boldsymbol{u}_h^{(n+1)}, p_h^{(n+1)})$. 下面的问题是如何用新的数值解和旧的网格$\mathcal{T}_h^{(n)}$ 来获得新的网格$\mathcal{T}_h^{(n+1)}$. 我们采取文献[?] 中的方法，注意到我们区域的边界均是Dirichlet边界，分为以下四步：

3.3.1　Step 1 获取Monitor

选择一个合适的控制函数，对于移动网格的结果是非常重要的。应用在不可压Navier-Stokes方程上的控制函数主要有如下几种. 令$m = \frac{1}{G}$ 其中m是(3.19)中的一个纯量函数。关于G有集中不同的选择.一种是基于涡量

$$G_0 = \sqrt{1 + \alpha|\omega|^{\beta}} \tag{3.14}$$

其中$\omega = \nabla \times \boldsymbol{u}$, α和β是两个正的常数。

另外一个选择，G是基于数值解的梯度

$$G_1 = \sqrt{1 + \alpha|\nabla \boldsymbol{u}|^{\beta}} \tag{3.15}$$

对于线性元v_h逼近真实解v，下面的后验误差估计公式可以用来近似计算误差

$$|v - v_h|_{1,\Omega} \sim \eta(v_h) := \sqrt{\sum_{l:\text{内部边界}} \int_l [\nabla v_h \cdot \boldsymbol{n}_l]^2 \, dl} \tag{3.16}$$

其中$[\cdot]_l$意味着边l上的跳跃，即$[v]_l = v|_{l^+} - v|_{l^+}$. 很自然的在每个单元上等分布数值误差$\eta(v_h)$, 控制函数为一下形式

$$G_2 = \sqrt{1 + \alpha\eta^2(v_h)} \tag{3.17}$$

Algorithm 3.2 移动网格方法来求解Navier Stokes方程

```
1: while $t_n < T$ do
2:     在速度网格$\triangle_v^{(n)}$和压力网格$\triangle_p^{(n)}$上，求解$t = t_n$时刻的Navier-Stokes方程(3.12), 获得数值解$\boldsymbol{u}_h^{(n)}, p_h^{(n)}$.
3:     在压力网格$\triangle_p^{(n)}$上，用$\boldsymbol{u}_h^{(n)}, p_h^{(n)}$来计算控制函数, 并通过求解(3.19), 获得$\boldsymbol{\xi}^*$.
4:     判断$\| \boldsymbol{\xi}^* - \boldsymbol{\xi}^{(0)} \|_{L^2}$ 是否小于容忍量$\epsilon$, 如果是迭代结束，否则，继续做5 - 8.
5:     用$\boldsymbol{\xi}^* - \boldsymbol{\xi}^{(0)}$两者之差来计算网格$\triangle_p^{(n)}$的移动量$\delta \boldsymbol{x}$.
6:     利用5中$\delta \boldsymbol{x}$, 在速度网格$\triangle_v^{(n)}$上求解更新数值解的方程(3.32), 得到新网格上的中间量$\boldsymbol{u}_{h,*}^{(n)}, p_{h,*}^{(n)}$.
7:     更新$\triangle_p^{(n)}$, 通过几何遗传树结构，来同步$\triangle_v^{(n)}$, 得到新的$\triangle_p^{(n+1)}$和$\triangle_v^{(n+1)}$
8:     回到3.
9:     在$\triangle_v^{(n+1)}$和$\triangle_p^{(n+1)}$ 上求解Navier-Stokes方程(3.12).从而真正获得$t = t_{n+1}$时刻的数值解$\boldsymbol{u}_h^{(n+1)}, p_h^{(n+1)}$.
10:    $n = n + 1$
11: end while
```

[?]中对(3.17)进行了改进

$$G_3 = \sqrt{1 + \alpha \left[\eta(v_h) / \mathrm{max} \eta(v_h) \right]^{\beta}} \tag{3.18}$$

其中$\beta > 2$时有更好的效果。

3.3.2　Step 2 获取新的逻辑网格

求解椭圆形方程

$$\begin{aligned} \nabla_{\boldsymbol{x}} \left(m \nabla_{\boldsymbol{x}} \boldsymbol{\xi} \right) &= 0 \\ \boldsymbol{\xi} &= \boldsymbol{\xi}_b \end{aligned} \tag{3.19}$$

其中m 是上一节中的纯量函数，通常依赖于$(\boldsymbol{u}_h^{(n+1)}, p_h^{(n+1)})$。我们定义初始的逻辑网格$\mathcal{T}_c$($\mathcal{A}^0$为它的节点)。一旦初始的逻辑网格给定，在整个的求解过程中，将一直保持不变。通过求解(3.19) 我们可以得到新的逻辑网格$\mathcal{T}_c^*$($\mathcal{A}^*$为它的节点)。

3.3.3 Step 3 物理网格的移动方向

我们先引入一些定义。$\mathcal{T}_h$ 为物理区域上的三角剖分。第i个点定义为X_i，以X_i 为顶点的单元的集合称之为T_i。相应的计算区域上的标记为$\mathcal{T}_c$, $\mathcal{A}_i$ 和$T_{i,c}$。$\mathcal{A}_i$ 点在计算区域上的坐标定义为$(\mathcal{A}_i^1, \mathcal{A}_i^2)^T$。在Step 1 结束后，我们得到了新的逻辑网格$\mathcal{T}_c^*$ 和它的顶点$\mathcal{A}_i^*$. 从而我们得到新旧逻辑网格的差:

$$\delta\mathcal{A}_i = \mathcal{A}^{(0)} - \mathcal{A}_i^* \tag{3.20}$$

对于一个给定的单元$E \in \mathcal{T}_h$, $X_{E_k}, 0 \le k \le 2$, 作为它的三个顶点。从$V_{\mathcal{T}_c^*}(\Omega)$ 到$V_{\mathcal{T}}(\Omega)$ 的分片线性映射在单元E 上的梯度是常数，并且满足下面的方程组：

$$\begin{aligned} &\begin{pmatrix} \mathcal{A}_{E_1}^{*,1} - \mathcal{A}_{E_0}^{*,1} & \mathcal{A}_{E_2}^{*,1} - \mathcal{A}_{E_0}^{*,1} \\ \mathcal{A}_{E_1}^{*,2} - \mathcal{A}_{E_0}^{*,2} & \mathcal{A}_{E_2}^{*,2} - \mathcal{A}_{E_0}^{*,2} \end{pmatrix} \begin{pmatrix} \frac{\partial x^1}{\partial \xi^1} & \frac{\partial x^1}{\partial \xi^2} \\ \frac{\partial x^2}{\partial \xi^1} & \frac{\partial x^2}{\partial \xi^2} \end{pmatrix} \\ = &\begin{pmatrix} X_{E_1}^1 - X_{E_0}^1 & X_{E_2}^1 - X_{E_0}^1 \\ X_{E_1}^2 - X_{E_0}^2 & X_{E_2}^2 - X_{E_0}^2 \end{pmatrix} \end{aligned}$$

求解上面的方程组，可以获得单元E 上的$\partial \boldsymbol{x}/\partial\xi$. 如果以单元的面积作为权重，则第i个点的加权平均的位移定义如下:

$$\delta X_i = \frac{\sum\limits_{E \in T_i} |E| \frac{\partial \boldsymbol{x}}{\partial \xi}|_{\text{in}E} \delta\mathcal{A}_i}{\sum\limits_{E \in T_i} |E|}. \tag{3.21}$$

其中$|E|$代表单元E的面积. 为了避免网格发生缠结，在网格移动向量前乘上一个常量μ, 即物理区域上新网格$\mathcal{T}^*$的节点表示为：

$$X_i^* = X_i + \mu\delta X_i. \tag{3.22}$$

文献[**?**]中提出μ按以下方式给出:

$$\begin{vmatrix} 1 & 1 & 1 \\ x_0^1 + \mu\delta x_0^1 & x_1^1 + \mu\delta x_1^1 & x_2^1 + \mu\delta x_2^1 \\ x_0^2 + \mu\delta x_0^2 & x_1^2 + \mu\delta x_1^2 & x_2^2 + \mu\delta x_2^2 \end{vmatrix} = 0 \tag{3.23}$$

其中$\boldsymbol{x}_i = (x_i^1, x_i^2), 0 \le i \le 2$表示第i个点的坐标。令$\mu_i^*$ 为方程(3.23)的最小正根，则令

$$\mu = \min(1, \frac{\mu_i^*}{2}). \tag{3.24}$$

3.3.4 Step 4 散度为0的插值

用移动网格方法求解不可压流体时，要保证插值的过程散度是为0的。通过求解一个对流方程，对流的速度是网格的移动速度，从而实现旧的物理网格上的数值解到新的物理网格上数值解的插值。令$u_h = \sum u_i\phi_i, u_h \in \mathcal{X}_E^h$，$\phi_i$是有限元空间$\mathcal{X}_h$的基函数。引入一个虚拟的时间$\tau$, 假设基函数$\phi_i$和$u_i$均是关于$\tau$的函数,即$\phi_i = \phi_i(\tau), u_i = u_i(\tau)$. 我们引入一个从旧网格$x^{旧}$到新网格$x^{新}$网格点的连续变换：

$$x_i(\tau) = X_i + \tau(X_i^* - X_i), \qquad \tau \in [0,1] \tag{3.25}$$

其中$X_i^* = x_i^{新}, X_i = x_i^{旧}$ 基于(3.25)的连续形式$x(\tau) = x_{旧} + \tau(x^{新} - x^{旧})$，基函数可以定义为$\phi_i(\tau) = \phi_i(x(\tau))$ 并且$u_i = u_i(x(\tau))$.

在插值的过程中，我们要保持解曲线$u_h = \sum u_i\phi_i$关于τ 在弱形式下是不变的.即对$\forall\psi \in \mathcal{X}_h$, $(\partial_\tau u_h, \psi) = 0$。通过直接计算可得

$$\frac{\partial\phi}{\partial\tau} = -\nabla_{\boldsymbol{x}}\phi_i \cdot \delta\boldsymbol{x} \tag{3.26}$$

其中$\delta\boldsymbol{x} = x^{旧} - x^{新}$。紧接着

$$\begin{aligned} 0 &= (\partial_\tau u_h, \psi) \\ &= (\partial_\tau \sum u_i(x(\tau))\phi_i, \psi) \\ &= (\sum \phi_i\partial u_i(x(\tau)) + \sum u_i\partial_\tau\phi_i) \\ &= (\sum \phi_i\partial_\tau u_i(x(\tau)) - \sum u_i\nabla_{\boldsymbol{x}}\phi_i \cdot \delta\boldsymbol{x}, \psi) \\ &= (\sum \phi_i\partial_\tau u_i(x(\tau)) - \nabla_{\boldsymbol{x}}u_h \cdot \delta\boldsymbol{x}, \psi) \end{aligned} \tag{3.27}$$

我们将(3.27)应用到不可压流上，即速度场要满足散度为0的条件。令$\mathcal{X}_h$ 为散度为0的空间：

$$\mathcal{X}_E^h = X_E^h \cap \{\boldsymbol{u}_h | \nabla \cdot \boldsymbol{u}_h = 0\} \tag{3.28}$$

那么(3.27)将变成：寻找$w_h \in \mathcal{X}_h$ 使得

$$\left(\sum \phi_i\partial_\tau u_i - \sum u_i\nabla_{\boldsymbol{x}}\phi_i \cdot \delta\boldsymbol{x}, z_h\right) = 0 \quad \forall z_h \in \mathcal{X}_h. \tag{3.29}$$

上面的结果意味着

$$\sum \phi_i\partial_\tau u_i - \sum u_i\nabla_{\boldsymbol{x}}\phi_i \cdot \delta\boldsymbol{x} \in \mathcal{X}_h^\perp \tag{3.30}$$

其中$\mathcal{X}_h^\perp + \mathcal{X}_h = L^2$. 根据文献[?]中的定理2.7, 如果区域$\Omega$ 是单连通的,那么

$$\mathcal{X}_h^\perp = \{\nabla q | q \in H^1(\Omega)\} \tag{3.31}$$

则存在$\nabla p \in \mathcal{X}_h^\perp$使得

$$\begin{aligned} \sum \phi_i \partial_\tau u_i - \sum u_i \nabla_{\boldsymbol{x}} \phi_i \cdot \delta \boldsymbol{x} = -\nabla p \\ \nabla_{\boldsymbol{x}} \cdot u_h = 0. \end{aligned} \tag{3.32}$$

注 3.1. 这里的p跟外部Navier-Stokes方程的解p不一致，只是一个辅助量。

(3.32)的弱形式：寻找$(\boldsymbol{u}_h, p_h) \in X_E^h \times P_h$ 使得

$$\begin{aligned} \left(\sum \phi_i \partial_\tau u_i - \sum u_i \nabla_{\boldsymbol{x}} \phi_i \cdot \delta \boldsymbol{x}, v_h\right) = (p_h, \nabla v_h), \qquad & \forall v_h \in X_E^h. \\ (\nabla_{\boldsymbol{x}} \cdot u_h, q_h) = 0, \qquad & \forall q_h \in P_h \end{aligned} \tag{3.33}$$

(3.32)和(3.33)的初值为在$t = t_n$时刻的网格上，$t = t_{n+1}$时刻外部Navier-Stokes方程的解。

时间方向的离散我们暂时先用线性Euler方法:

$$\begin{aligned} \left(\frac{\sum \phi_i u_{i,*}^{(n)} - \sum \phi_i u_i^{(n)}}{\Delta \tau}, v_h\right) - \left(\sum u_i^{(n)} \nabla \phi_i \cdot \delta \boldsymbol{x}, v_h\right) = \left(p_h^{(n)}, \nabla v_h\right). \\ \left(\nabla \cdot u_{h,*}^{(n)}, q_h\right) = 0 \end{aligned} \tag{3.34}$$

注意到在做数值解插值的过程中，物理网格还没有发生移动，这时候基函数ϕ仍然是$t = t_n$时刻网格上的基函数。所以$u_{h,*}^{(n)} = \sum u_{i,*}^{(n)} \phi_i^{(n)}$，$u_h^{(n)} = \sum u_i^{(n)} \phi_i^{(n)}$简单整理一下：

$$\begin{aligned} \left(\frac{u_{h,*}^{(n)} - u_h^{(n)}}{\Delta \tau}, v_h\right) - \left(\nabla u_h^{(n)} \cdot \delta \boldsymbol{x}, v_h\right) = \left(p_h^{(n)}, \nabla v_h\right). \\ \left(\nabla \cdot u_{h,*}^{(n)}, q_h\right) = 0 \end{aligned} \tag{3.35}$$

其中$u_h^{(n)}$ 和p_h 是在$t = t_n$时刻的网格上，$t = t_{n+1}$ 时刻，Navier-Stokes 方程的数值解。而$u_{h,*}^{(n)}$ 和$p_{h,*}^{(n)}$是在新的网格上$t = t_{n+1}$时刻更新的解。但这组解不能当作外部Navier-Stokes方程的解。需要在新网格上重新求解Navier-Stokes 方程，得到的解才是我们想要的解。

3.4　数值算例

3.4.1　Colliding Flow

这个例子为稳态Stokes方程的精确解，粘性系数$\nu = 1.0$:

$$u_x = 20xy^3; \quad u_y = 5x^4 - 5y^4; \quad p = 60x^2y - 20y^3 + \text{constant}. \tag{3.36}$$

其中计算区域$\Omega = [-1, -1] \times [1, 1]$, 边界条件全部是Dirichlet 条件。这个例子是用来检验移动网格方法的收敛阶，此时解比较光滑。从文献[?] 可知，我们期望移动网格的收敛阶：速度有二阶收敛，压力一阶收敛。我们先给出均匀网格时，误差的收敛阶，如Table(3.1) 和Table(3.2)所示。

网格	$\|\|\boldsymbol{u} - \boldsymbol{u}_h\|\|_{L^2}$	误差阶	$\|\|\boldsymbol{u} - \boldsymbol{u}_h\|\|_{H^1}$	$\|\|p - p_h\|\|_{L^0}$	误差阶	$\|\|p - p_h\|\|_{H^1}$
10×10	1.42×10^{-1}		3.65×10^0	1.26×10^0		2.06×10^1
20×20	3.54×10^{-2}	2.01	1.81×10^0	3.90×10^{-1}	1.62	1.22×10^1
40×40	8.82×10^{-3}	2.01	9.03×10^{-1}	1.14×10^{-1}	1.71	6.72×10^0
80×80	2.20×10^{-3}	2.00	4.51×10^{-1}	3.39×10^{-2}	1.68	3.90×10^0

表 3.1: 用均匀网格计算碰撞流的误差, $\nu = 1.0$.

网格	涡量L^2误差	误差阶	散度L^2误差	误差阶
10×10	2.62×10^0		2.54×10^0	
20×20	1.31×10^0	1.00	1.25×10^0	1.02
40×40	6.54×10^{-1}	1.00	6.22×10^{-1}	1.00
80×80	3.27×10^{-1}	1.00	3.10×10^{-1}	1.00

表 3.2: 用均匀网格计算碰撞流的散度和涡量误差, $\nu = 1.0$.

我们采用上面一章中的移动网格方法来求解这个算例。选取(3.15)为控制函数，其中$\boldsymbol{u} = (u_x, u_y)^T$. α和β分别取为0.002和2. 从Table(3.3)中可以看出速度L^2 误差有二阶，压力收敛阶是一阶。从Table(3.4)可以看出速度散度和涡量均有一阶收敛。

首先我们判断网格是不是往正确的方向移动。从Figure(3.2)中看出控制函数G_1最大的地方分布在区域的四个顶角上，中间区域控制函数的值是比较小

网格	$\|\|\boldsymbol{u}-\boldsymbol{u}_h\|\|_{L^2}$	误差阶	$\|\|\boldsymbol{u}-\boldsymbol{u}_h\|\|_{H^1}$	$\|\|p-p_h\|\|_{L^0}$	误差阶	$\|\|p-p_h\|\|_{H^1}$
10×10	1.18×10^{-1}		3.37×10^{0}	8.49×10^{-1}		1.84×10^{1}
20×20	2.94×10^{-2}	2.01	1.67×10^{0}	2.330×10^{-1}	1.82	1.06×10^{1}
40×40	7.37×10^{-3}	1.99	8.35×10^{-1}	6.54×10^{-2}	1.78	5.89×10^{0}
80×80						

表 3.3: 用移动网格计算碰撞流的误差, $\nu=1.0$.

网格	涡量L^2误差	误差阶	散度L^2误差	误差阶
10×10	2.35×10^{0}		2.41×10^{0}	
20×20	1.16×10^{0}	1.01	1.20×10^{0}	1.00
40×40	5.79×10^{-1}	1.00	6.01×10^{-1}	1.00
80×80				

表 3.4: 用移动网格计算碰撞流的散度和涡量误差, $\nu=1.0$.

的。网格应该从控制函数小的地方移动到控制函数值大的地方。而网格的移动方向也确实是往四个顶角上移动。因此我们可以确定网格移动方向是对的。再者，判断选取G_1为控制函数是否合适。从Figure (3.3)中看到：在均匀网格下，速度L^2误差最大的地方分布在四个顶角上，这与控制函数的分布是一致的。所以选取G_1为控制函数是合理的。再来对比Table(3.1)和Table(3.3)中速度和压力的误差。可以发现，移动网格方法下速度和压力的误差均有所下降，但是下降的很小。

注意到在Figure(3.4)中，移动网格相对于均匀网格移动的并不是很明显。在Table(3.5)中我们选取不同的α的值，查看速度和压力误差的变化。网格如Figure(3.5) 所示。随着α的值变大, 虽然网格移动的效果越来越明显, 但此时的速度和压力误差却变大了。

3.5 Introduction

Let integer $p\geq 2$. A matrix $X\in\mathbb{C}^{n\times n}$ is called a pth root of a matrix $A\in\mathbb{C}^{n\times n}$ if $X^p=A$. If A has no eigenvalues on $\mathbb{R}^-$, the closed negative real axis, and all zero eigenvalues of A are semisimple, there exists a unique principal pth root of A, denoted by $A^{1/p}$ [?]. Eigenvalues of $A^{1/p}$ have argument less

图 3.2: 左：$m = \frac{1}{G_1}$的等高线和网格移动方向；右：压力等高线和速度的流速线$\alpha = 0.002, \beta = 2.0$.

α	$\|\|\boldsymbol{u} - \boldsymbol{u}_h\|\|_{L^2}$	$\|\|\boldsymbol{u} - \boldsymbol{u}_h\|\|_{H^1}$	$\|\|p - p_h\|\|_{L^0}$	$\|\|p - p_h\|\|_{H^1}$
0.0(不移动)	3.54×10^{-2}	1.81×10^{0}	3.90×10^{-1}	1.22×10^{1}
0.002	2.94×10^{-2}	1.67×10^{0}	2.330×10^{-1}	1.06×10^{1}
0.005	2.98×10^{-2}	1.64×10^{0}	2.08×10^{-1}	1.06×10^{1}
0.01	3.22×10^{-2}	1.65×10^{0}	2.02×10^{-1}	1.08×10^{1}
0.1	5.28×10^{-2}	2.09×10^{0}	3.39×10^{-1}	1.40×10^{1}

表 3.5: 不同α的值，用移动网格计算碰撞流的误差, 网格20×20.

in modulus than π/p. An application of pth root is in the computation of the matrix logarithm through the relation $\log A = p \log A^{1/p}$, where p is chosen so that $A^{1/p}$ can be well approximated by a polynomial or rational function. One can see [?] or the recent survey paper [?] for more applications.

There are two various ways to deal with the matrix pth root. The first way is to use the direct Schur decomposition of A, say $Q^*AQ = R$, to solve the equation $Y^p = R$, instead of the equation $X^p = A$, by appropriate recurrences on the elements of Y, see for example [? ? ?]. The second way is to use the matrix iterations which converge to the principal pth root of A, see for example [? ? ? ? ? ? ? ? ? ?].

图 3.3: 速度L^2误差分布，左：均匀网格；右：移动网格$\alpha = 0.002, \beta = 2.0$.

In this paper, we are interested in the method of matrix iterations. They are mainly deduced from iterative method for solving nonlinear equation. When Newton's method is applied on the matrix equation

$$f(X) = X^p - A = 0, \tag{3.37}$$

Iannazzo shown that the matrix sequence $\{X_k\}$ generated by Newton's method starting from the identity matrix converges to the principal pth root $A^{1/p}$ if all the eigenvalues of A are in the set $\mathcal{D} = \{z \in \mathbb{C} : |z| \leq 1, \operatorname{Re} z > 0\}$ or in the set $\mathcal{E} = \{z \in \mathbb{C} : 0 < |z| \leq 2, |\arg(z)| < \pi/4\}$ in [?] and [?], respectively. Subsequently, Guo obtained in [?] that the matrix sequence converges to the principal pth root of A also when the eigenvalues of A lie in the set $\mathcal{F} = \{z \in \mathbb{C} : |z - 1| \leq 1\}$.

The convergence study for Halley's method applied to (3.37) is studied in [?] where Iannazzo proved that the matrix sequence $\{X_k\}$ generated by Halley's method starting from the identity matrix converges to the principal pth root $A^{1/p}$ for each A having eigenvalues in the set $\mathcal{G} = \{z \in \mathbb{C} : \operatorname{Re} z > 0\}$. Guo further proved the order of convergence is quadratic and cubic for Newton iteration and Halley iteration in [?], respectively.

Methods based on Padé approximation (Padé family of iterations and dual

图 3.4: 网格对比，左：均匀网格20×20；右：移动网格$\alpha = 0.002, \beta = 2.0$.

Padé family of iterations) for computing the principal pth root are also investigated by several authors, see for example [**? ? ?**]. Note that Newton's method and Halley's method are special cases as dual Padé family of iterations. Meanwhile, Halley's method is also a special case of Padé family of iterations, see [**?**] for more details.

As you known, Euler's method is also a famous iteration for solving nonlinear operator equation. Euler's method for (3.37) in classical form

$$X_{k+1} = X_k - [\mathbf{I} + L_f(X_k)]f'(X_k)^{-1}f(X_k), \quad k = 0, 1, 2, \ldots,$$

where $\mathbf{I}$ denotes the identity operator and operator $L_f(X) = \frac{1}{2}f'(X)^{-1}f''(X)f'(X)^{-1}f(X)$, can be rewritten as follows:

$$X_{k+1} = E(X_k), \quad k = 0, 1, 2, \ldots, \tag{3.38}$$

provided X_0 commutes with A and X_k is nonsingular for any $k \geq 0$, where

$$E(X) = \frac{1}{2p^2}X\left[(2p^2 - 3p + 1)\mathrm{I} + 2(2p - 1)AX^{-p} - (p - 1)\left(AX^{-p}\right)^2\right] \tag{3.39}$$

for nonsingular matrix $X \in \mathbb{C}^{n\times n}$.

A natural question is, how it should be when we apply this method to solve the principal pth root of A? It becomes the main goal of this paper. It is worth

图 3.5: 网格对比，左；移动网格$\alpha = 0.01, \beta = 2.0$ 右：移动网格$\alpha = 0.1, \beta = 2.0$.

noting that, Newton's method and Euler's method (3.38) are special cases as the family of Schröder iteration which has recently been studied for the pth root of complex numbers, see for example [? ?].

In the case of scalar Newton iteration the only existing fixed points are the pth roots of λ. For scalar Halley iteration, it has not only the pth roots of λ as the fixed points, but has the extra fixed point $z = 0$. While for Euler's method (3.38), the extra fixed points are $z = 0$ as well as the pth roots of $(p-1)\lambda/(3p-1)$. Based on the above intrinsic properties, we cannot get the convergence region of Halley's method obtained in [?] for Euler's method (3.38). However, we determine a certain convergence region for Euler's method (3.38) including that one for Newton's method given in [?]. Furthermore, we also analyze the robust of Euler's method (3.38) and give a modification based on Schur decomposition which has been applied to Newton and Halley's methods in [?] and [?], respectively. Numerical experiments illustrate that the modified algorithm has good numerical properties as the existing algorithms and confirm that the Euler method is a good choice for solving the matrix pth root.

This paper is organized as follows. In Section 2, we state our main result on the convergence of Euler's method (3.38) for computing the matrix pth root. And then we prove this result in Section 3. In Section 4, we describe our general algorithm and discuss some related computational issues. Finally in Section 5 we present some numerical experiments and compare our method with others existing methods. These results confirm the numerical stability the overall good performance of the new algorithm.

3.6 Convergence results for Euler's method

Throughout the whole paper, we always suppose that the integer $p \geq 2$. To state the convergence results for Euler's method (3.38), we introduce some notations. Let

$$u(z) := \frac{1}{1 - \frac{1}{p}z - \frac{1}{2}\left(\frac{1}{p} - \frac{1}{p^2}\right) z^2}, \quad z \in \mathcal{D}_0$$

and

$$\phi(z) := 1 - u(z)^p(1 - z), \quad z \in \mathcal{D}_0, \tag{3.40}$$

where

$$\mathcal{D}_0 := \left\{ z \in \mathbb{C} : |z| < \frac{p}{p-1}(\sqrt{2p-1} - 1) \right\}. \tag{3.41}$$

Define

$$\mathcal{R} := 1 - \overline{\mathcal{D}}_1 \bigcup \left(\mathcal{D}_0 \bigcap \mathcal{D}_2 \right) := \left\{ 1 - z : z \in \overline{\mathcal{D}}_1 \bigcup \left(\mathcal{D}_0 \bigcap \mathcal{D}_2 \right) \right\} \tag{3.42}$$

and

$$\widehat{\mathcal{R}} := 1 - \mathcal{D}_1 \bigcup \left(\mathcal{D}_0 \bigcap \mathcal{D}_2 \right) := \left\{ 1 - z : z \in \mathcal{D}_1 \bigcup \left(\mathcal{D}_0 \bigcap \mathcal{D}_2 \right) \right\}, \tag{3.43}$$

where $\mathcal{D}_0$ is defined in (3.86), $\overline{\mathcal{D}}_1$ is the closure of $\mathcal{D}_1$ defined by

$$\mathcal{D}_1 := \{ z \in \mathbb{C} : |z| < 1 \}, \tag{3.44}$$

and

$$\mathcal{D}_2 := \left\{ z \in \mathbb{C} : \sup_{m \geq 3} \left\{ \frac{|S_m(z)|}{|z|} \right\} \cdot \frac{|z| + |\phi(z)|}{|z| - |\phi(z)|} < 1 \right\}, \tag{3.45}$$

where $S_m(z) = \sum_{j=3}^{m} c_j z^j$, $c_j = \phi^{(j)}(0)/j!, j = 3, 4, \ldots$, and ϕ is defined in (3.85). That is,

$$\mathcal{R} = \{ z \in \mathbb{C} : 1 - z \in \overline{\mathcal{D}}_1 \} \bigcup \{ z \in \mathbb{C} : 1 - z \in \mathcal{D}_0 \bigcap \mathcal{D}_2 \}.$$

We have

定理 3.1. *If all eigenvalues of $A \in \mathbb{C}^{n\times n}$ are in $\mathcal{R}$ defined in (3.87) and all zero eigenvalues of A are semisimple, then the matrix sequence $\{X_k\}$ generated by Euler's method (3.38) starting from $X_0 = \mathrm{I}$ converges to the principal pth root $A^{1/p}$. Moreover, if all eigenvalues are in $\widehat{\mathcal{R}}\backslash\{0\}$ defined in (3.88), then the convergence is cubic.*

注 3.2. Figure 3.6 plots the approximating region of $\mathcal{R}$ with nine cases in the complex plane, where the red and blue regions denote the sets $\{z \in \mathbb{C} : 1 - z \in \overline{\mathcal{D}}_1\}$ and $\{z \in \mathbb{C} : 1 - z \in \mathcal{D}_0 \bigcap \mathcal{D}_2\}$, respectively. Since

$$\sum_{j=1}^{m} c_j^1 (1-z)^{j-1} \to \frac{\phi_1(1-z)}{1-z}, \quad n \to \infty,$$

for each $z \in \{z \in \mathbb{C} : 1 - z \in \mathcal{D}_0\}$ by Lemma 3.1 in Section 3.1, we can see from Figure 3.6 that approximating regions ((a)-(c), (d)-(f), (g)-(i)) are almost the same for a fixed p when $m = 20, 100, 500$, respectively. To this end, it suffices to choose $m = 20$ in practical numerical computation.

注 3.3. The set $\mathcal{D}_0$ defined in (3.86) can be enlarged as

$$\left\{ z \in \mathbb{C} : \frac{1}{p}|z| \left| 1 + \frac{1}{2}(1 - \frac{1}{p})z \right| < 1 \right\}.$$

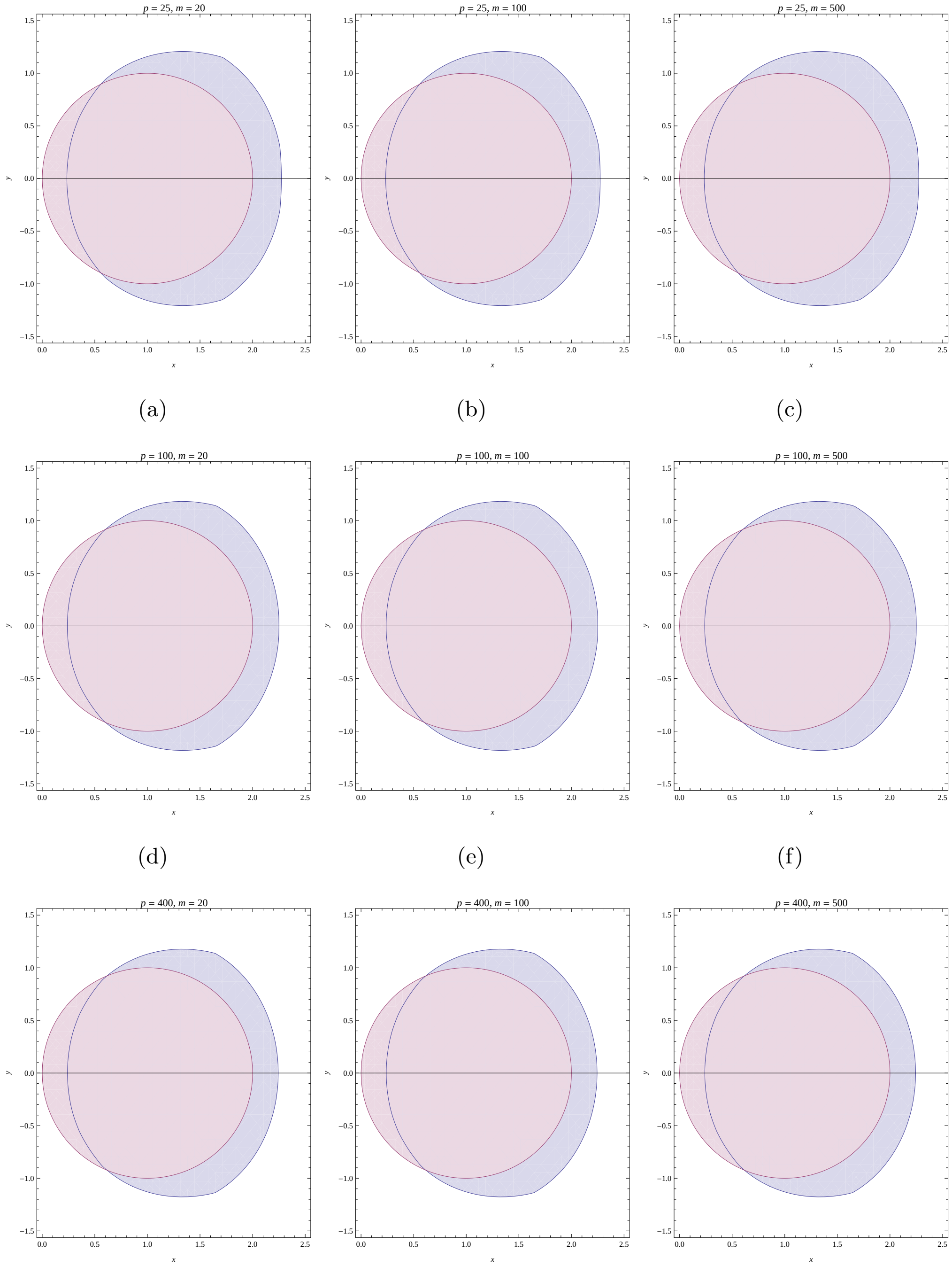
However, this improvement does not affect the convergence region $\mathcal{R}$ defined in (3.87). So, for convenience, we still use $\mathcal{D}_0$ in our convergence analysis in Section 3.

注 3.4. In practice, it is not feasible to check whether a eigenvalue λ belongs to $\mathcal{R}$. This is due to the computational cost of (3.90) may be large even if we only choose $m = 20$. Thus, based on the observation from Figure 3.6, we now give a new convergence region which allows us to determine easily whether a eigenvalue belongs to it. Define

$$\mathcal{R}_{\mathrm{E}} := \mathcal{D}_3 \bigcup \mathcal{D}_4, \tag{3.46}$$

where

$$\mathcal{D}_3 := \{z \in \mathbb{C} : 1 - z \in \overline{\mathcal{D}}_1\},$$



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

图 3.6: The approximating regions of $\mathcal{R}$ defined in (3.87) for $p = 25, 100, 400$ and $m = 20, 100, 500$, where the red and blue regions denote the set $\{z \in \mathbb{C} : 1 - z \in \overline{\mathcal{D}}_1\}$ and $\{z \in \mathbb{C} : 1 - z \in \mathcal{D}_0 \bigcap \mathcal{D}_2\}$, respectively.

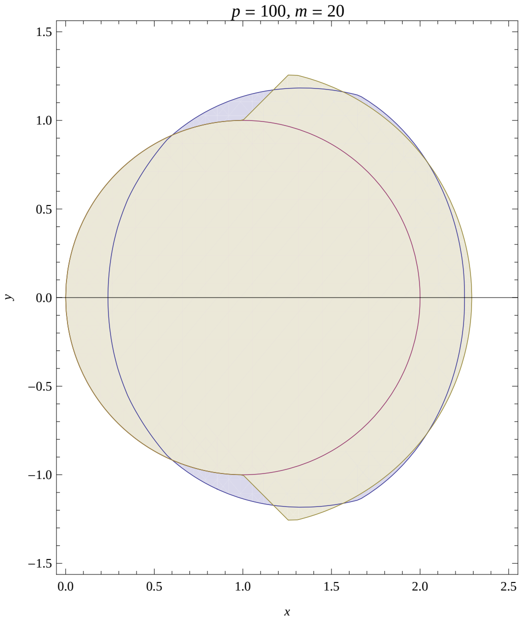


图 3.7: For $p = 100, m = 20$, the actual convergence regions $\mathcal{R}$ defined in (3.87) (the union of the red and blue parts) and the approximate convergence regions $\mathcal{R}_{\mathrm{E}}$ defined in (3.46) of Euler's method (the yellow parts).

$$\mathcal{D}_4 := \left\{ z \in \mathbb{C} : |\arg(z)| < \frac{\pi}{4}, |1 - z| < \frac{31}{24} \right\}.$$

In view of $p(\sqrt{2p-1} - 1)/(p-1) > 31/24$ for any $p \geq 2$, one has that $\mathcal{D}_4 \subset \mathcal{D}_0$. In Figure 3.7, for $p = 100, m = 20$, we present the the regions $\mathcal{R}$ and $\mathcal{R}_{\mathrm{E}}$ defined in (3.87) and (3.46), respectively, of Euler's method (3.38) for computing the pth root of a matrix. We can observe that the new region $\mathcal{R}_{\mathrm{E}}$ is acceptable approximation to $\mathcal{R}$. So instead of using $\mathcal{R}$, in Sections 4 and 5, we will use $\mathcal{R}_{\mathrm{E}}$ in our algorithm and numerical experiments.

3.7 Proof of Theorem 3.1

In this section, we will prove Theorem 3.1, the results on convergence and convergence order for Euler's method (3.38).

3.7.1 Technical lemmas

The following lemma is taken from [? , Theorem 3.2].

引理 3.1 ([?]). *The Maclaurin series of the function $\phi(z)$ defined by (3.85) has the form*

$$\phi(z) = \sum_{j=0}^{\infty} c_j z^j, \quad z \in \mathcal{D}_0, \tag{3.47}$$

where $\mathcal{D}_0$ is defined in (3.86), and the coefficients $c_j = \phi^{(j)}(0)/j!$ satisfying $c_j > 0$ for all $j \geq 3$ and $\sum\limits_{j=3}^{\infty} c_j = 1$.

For a complex number $\lambda \in \mathbb{C}$, let

$$f(z) := z^p - \lambda, \quad z \in \mathbb{C} \tag{3.48}$$

and

$$r(z, \lambda) := 1 - \lambda z^{-p}, \quad z \in \mathbb{C}\backslash\{0\}. \tag{3.49}$$

引理 3.2. *Let $r(z,\lambda)$ be defined in (3.49). If $r(z,\lambda) \in \mathcal{D}_0\backslash\{1\}$ for some $z \in \mathbb{C}\backslash\{0\}$, where $\mathcal{D}_0$ is defined in (3.86), then $E(z)$ generated by the scalar case of (3.39) for $f(z)$ defined in (3.48) exists and*

$$r(E(z), \lambda) = \phi(r(z, \lambda)), \tag{3.50}$$

where ϕ is defined by (3.85). Moreover, we have

$$\begin{cases} |r(E(z),\lambda)| < |r(z,\lambda)|^3, & \text{if } \ r(z,\lambda) \in \overline{\mathcal{D}}_1\backslash\{0,1\}, \quad (3.51) \\ |r(E(z),\lambda)| \leq \sup\limits_{m \geq 3} \left\{ \left| \dfrac{S_m(u)}{u^3} \right| \right\} \cdot \dfrac{|u| + |r(z,\lambda)|}{|u| - |r(z,\lambda)|} \cdot |r(z,\lambda)|^3, & \text{if } \ r(z,\lambda) \in \mathcal{D}_0\backslash\{1\}, \quad (3.52) \end{cases}$$

where $\overline{\mathcal{D}}_1$ is the closure of $\mathcal{D}_1$ defined in (3.89), and $S_m(u)$ is the mth partial sum of the series (3.47) for each $u \in \mathcal{D}_0$ satisfying $|u| > |r(z,\lambda)|$, $m \geq 3$.

证明. For any $z \in \mathbb{C}\backslash\{0\}$, $E(z)$ exists by (3.39). Furthermore, when $r(z,\lambda) \in \mathcal{D}_0\backslash\{0,1\}$, since

$$\frac{1}{2p^2} \left| (2p^2 - 3p + 1) + 2(2p-1)\lambda z^{-p} - (p-1)(\lambda z^{-p})^2 \right|$$

$$
\begin{aligned}
&= \left|\frac{2p^2-3p+1}{2p^2}+\frac{2p-1}{p^2}(1-r(z,\lambda))-\frac{p-1}{2p^2}(1-r(z,\lambda))^2\right| \\
&= \left|1-\frac{1}{p}r(z,\lambda)-\frac{1}{2}\left(\frac{1}{p}-\frac{1}{p^2}\right)r^2(z,\lambda)\right| \\
&> 1-\left[\frac{1}{p}|r(z,\lambda)|+\frac{1}{2}\left(\frac{1}{p}-\frac{1}{p^2}\right)|r(z,\lambda)|^2\right] \\
&> 0,
\end{aligned}
$$

we know

$$E(z)=\frac{1}{2p^2}z\left[(2p^2-3p+1)+2(2p-1)\lambda z^{-p}-(p-1)(\lambda z^{-p})^2\right]\neq 0$$

and

$$r(E(z),\lambda)=1-\left[1-\frac{1}{p}r(z,\lambda)-\frac{1}{2}\left(\frac{1}{p}-\frac{1}{p^2}\right)r^2(z,\lambda)\right]^{-p}\cdot(1-r(z,\lambda))=\phi(r(z,\lambda)), \tag{3.53}$$

i.e., (3.50) holds. (3.53) shows that $r(E(z),\lambda)\neq 1$ when $r(z,\lambda)\neq 1$.

If $r(z,\lambda)\in\overline{\mathcal{D}}_1\backslash\{0,1\}\subset\mathcal{D}_0$, then, by (3.47) in Lemma 3.1 and (3.53), one has that

$$
\begin{aligned}
|r(E(z),\lambda)|=|\phi(r(z,\lambda))| &= |r(z,\lambda)|^3\cdot\left|\sum_{j=3}^{\infty}c_j r^{j-3}(z,\lambda)\right| \\
&\leq |r(z,\lambda)|^3\left[|c_3+c_4 r(z,\lambda)|+\sum_{j=5}^{\infty}c_j|r(z,\lambda)|^{j-3}\right] \\
&< |r(z,\lambda)|^3\sum_{j=3}^{\infty}c_j=|r(z,\lambda)|^3 \\
&\leq |r(z,\lambda)| \qquad (3.54)
\end{aligned}
$$

from Lemma 3.1 and the fact that $|c_3+c_4w|<c_3+c_4$ for all $w\in\overline{\mathcal{D}}_1\backslash\{1\}$. Thus, (3.51) is proved.

If $r(z,\lambda)\in\mathcal{D}_0\backslash\{1\}$, based on (3.47) in Lemma 3.1 and (3.53) again, we have

$$r(E(z),\lambda)=\phi(r(z,\lambda))=\sum_{j=3}^{\infty}c_j r^j(z,\lambda)=\left[\sum_{j=3}^{\infty}c_j u^{j-3}\left(\frac{r(z,\lambda)}{u}\right)^{j-3}\right]\cdot r^3(z,\lambda)$$

holds for any $u \in \mathbb{C}\backslash\{0\}$. Since, for any $m \geq 3$,

$$\begin{aligned}\sum_{j=3}^{m} c_j u^{j-3}\left(\frac{r(z,\lambda)}{u}\right)^{j-3} &= \sum_{j=3}^{m-1}\left(\sum_{\ell=3}^{j} c_\ell u^{\ell-3}\right)\left(1-\frac{r(z,\lambda)}{u}\right)\left(\frac{r(z,\lambda)}{u}\right)^{j-3} \\ &\quad + \left(\sum_{\ell=3}^{m} c_\ell u^{\ell-3}\right)\cdot\left(\frac{r(z,\lambda)}{u}\right)^{m-3}\end{aligned}$$

by Abel transformation, we have

$$\begin{aligned}\left|\sum_{j=3}^{m} c_j u^{j-3}\left(\frac{r(z,\lambda)}{u}\right)^{j-3}\right| &\leq \sup_{3\leq j\leq m-1}\left\{\left|\sum_{\ell=3}^{j} c_\ell u^{\ell-3}\right|\right\}\cdot\left|1-\frac{r(z,\lambda)}{u}\right|\cdot\sum_{j=3}^{m-1}\left|\frac{r(z,\lambda)}{u}\right|^{j-3} \\ &\quad + \left|\frac{S_m(u)}{u^3}\right|\cdot\left|\frac{r(z,\lambda)}{u}\right|^{m-3}, \quad m > 3.\end{aligned}$$

Then, letting $m \to \infty$ in the above inequality, it follows that

$$\begin{aligned}|r(E(z),\lambda)| &\leq \sup_{m\geq 3}\left\{\left|\sum_{j=3}^{m} c_j u^{j-3}\right|\right\}\cdot\left|1-\frac{r(z,\lambda)}{u}\right|\cdot\sum_{m=3}^{\infty}\left|\frac{r(z,\lambda)}{u}\right|^{m-3}\cdot|r(z,\lambda)|^3 \\ &= \sup_{m\geq 3}\left\{\left|\frac{S_m(u)}{u^3}\right|\right\}\cdot\frac{\left|1-\frac{r(z,\lambda)}{u}\right|}{1-\left|\frac{r(z,\lambda)}{u}\right|}\cdot|r(z,\lambda)|^3 \\ &= \sup_{m\geq 3}\left\{\left|\frac{S_m(u)}{u^3}\right|\right\}\cdot\frac{|u|+|r(z,\lambda)|}{|u|-|r(z,\lambda)|}\cdot|r(z,\lambda)|^3\end{aligned}$$

for any $r(z,\lambda) \in \mathcal{D}_0\backslash\{1\}$ and any $u \in \mathcal{D}_0$ subject to $|u| > |r(z,\lambda)|$, which verifies (3.52). The proof is completed. □

Based on Lemma 3.2, we have the following lemma which is more refined than the one obtained in [**?** , Corollary 3.1].

引理 3.3. *Let* $r(z,\lambda)$ *be defined in* (3.49). *If* $r(z_0,\lambda) \in \overline{\mathcal{D}}_1\backslash\{0,1\}$ *for some* $z_0 \in \mathbb{C}\backslash\{0\}$, *then the sequence* $\{z_k\}$ *starting from* z_0 *generated by the scalar case of* (3.38) *for solving* (3.48) *exists,*

$$|r(z_k,\lambda)| \leq q_1^{3^{k-1}}(z_0), \quad k = 1, 2, \ldots, \tag{3.55}$$

where

$$q_1(z_0) = q_1(z_0,\lambda) := \left|\sum_{j=3}^{\infty} c_j r^{j-1}(z_0,\lambda)\right| < 1, \tag{3.56}$$

and so $|r(z_k,\lambda)| \to 0$ *with order* 3 *as* $k \to \infty$.

证明. For z_0 chosen, $q_1(z_0) < 1$ follows from the same arguments used in (3.54). By (3.51) in Lemma 3.2, $z_1 = E(z_0)$ exists and

$$|r(z_1, \lambda)| = q_1(z_0)|r(z_0, \lambda)| \leq q_1(z_0) < 1.$$

Suppose that z_k exists and (3.55) holds for some $k \geq 1$, then by Lemma 3.2 again, $z_{k+1} = E(z_k)$ exists and

$$|r(z_{k+1}, \lambda)| < |r(z_k, \lambda)|^3 \leq \left[q_1^{3^{k-1}}(z_0)\right]^3 = q_1^{3^k}(z_0).$$

Thus, (3.55) holds for $k+1$. By induction, $\{z_k\}$ exists and (3.55) holds. Furthermore, $r(z_k, \lambda) \to 0$ with order 3 as $k \to \infty$, which completes the proof. □

The following lemma says that, besides in $\overline{\mathcal{D}}_1 \backslash \{0, 1\}$, it also guarantees $r(z_k, \lambda)$ converges cubically to 0 as $k \to \infty$ when $r(z_0, \lambda) \in \mathcal{D}_2 \bigcap \mathcal{D}_0 \backslash \{1\}$ for some $z_0 \in \mathbb{C} \backslash \{0\}$.

引理 3.4. 我们*Let $r(z, \lambda)$ be defined in (3.49). For any $z_0 \in \mathbb{C}\backslash\{0\}$ satisfying $r(z_0, \lambda) \in \mathcal{D}_2 \bigcap \mathcal{D}_0 \backslash \{1\}$ and*

$$q_2(z_0) = q_2(z_0, \lambda) := \sqrt{\sup_{m \geq 3} \left\{ \frac{|S_m(r(z_0, \lambda))|}{|r(z_0, \lambda)|} \right\} \cdot \frac{|r(z_0, \lambda)| + |\phi(r(z_0, \lambda))|}{||r(z_0, \lambda)| - |\phi(r(z_0, \lambda))||}} < 1, \tag{3.57}$$

where $\mathcal{D}_2$ is defined in (3.90), the sequence $\{z_k\}$ generated by the scalar form of (3.38) starting from z_0 for solving (3.48) exists,

$$|r(z_k, \lambda)| \leq q_2^{3^k - 1}(z_0) \cdot |r(z_0, \lambda)|, \quad k = 0, 1, \ldots, \tag{3.58}$$

and so $|r(z_k, \lambda)| \to 0$ with order 3 as $k \to \infty$.

证明. For z_0 chosen, we have $r(z_0) \in \mathcal{D}_0 \backslash \{1\}$. So, $z_1 = E(z_0)$ exists and $z_1 \neq 0$ by Lemma 3.2. Recall that $S_m(r(z_0)) \to \phi(r(z_0))$ as $k \to \infty$ implies

$$\left| \frac{\phi(r(z_0))}{r(z_0)} \right| \leq \sup_{m \geq 3} \left\{ \left| \frac{S_m(r(z_0))}{r(z_0)} \right| \right\} < q_2^2(z_0) < 1,$$

by (3.57), we have

$$|r(z_1)| = |\phi(r(z_0))| = \left| \frac{\phi(r(z_0))}{r(z_0)} r(z_0) \right| < q_2^2(z_0) \cdot |r(z_0)|,$$

and (3.58) holds for $k=1$. Assume $z_0, z_1, \ldots, z_k$ exist and satisfy (3.58). Then

$$|r(z_k)| \leq q_2^2(z_0)|r(z_0)| < |r(z_0)| < \frac{p}{p-1}(\sqrt{2p-1}-1).$$

So, by Lemma 3.2 with $u = r(z_0)$, $z_{k+1} = E(z_k)$ exists, $z_{k+1} \neq 0$ and

$$\begin{aligned}
|r(z_{k+1})| &= |\phi(r(z_k))| \\
&\leq \sup_{m\geq 3}\left\{\frac{|S_m(r(z_0))|}{|r(z_0)|^3}\right\} \cdot \frac{|r(z_0)|+|\phi(r(z_0))|}{||r(z_0)|-|\phi(r(z_0))||} \cdot |r(z_k)|^3 \\
&\leq \sup_{m\geq 3}\left\{\frac{|S_m(r(z_0))|}{|r(z_0)|^3}\right\} \cdot \frac{|r(z_0)|+|\phi(r(z_0))|}{||r(z_0)|-|\phi(r(z_0))||} \left[q_2^{3^k-1}(z_0)\right]^3 \cdot |r(z_0)|^3 \\
&= \sup_{m\geq 3}\left\{\frac{|S_m(r(z_0))|}{|r(z_0)|}\right\} \cdot \frac{|r(z_0)|+|\phi(r(z_0))|}{||r(z_0)|-|\phi(r(z_0))||} \cdot [q_2(z_0)]^{3^{k+1}-3} \cdot |r(z_0)| \\
&= [q_2(z_0)]^{3^{k+1}-1} \cdot |r(z_0)|,
\end{aligned}$$

which shows (3.58) by induction. The proof is completed. □

Now, based on the above lemmas, we can obtain the following convergence results for scalar Euler's method (3.38). Define

$$\mathcal{R}_1 := \{\lambda \in \mathbb{C} : r(z_0,\lambda) \in \overline{\mathcal{D}}_1 \text{ for some } z_0 \in \mathbb{C}\backslash\{0\}\}, \tag{3.59}$$

and

$$\mathcal{R}_2 := \left\{\lambda \in \mathbb{C} : r(z_0,\lambda) \in \mathcal{D}_0 \bigcap \mathcal{D}_2 \text{ for some } z_0 \in \mathbb{C}\backslash\{0\}\right\}, \tag{3.60}$$

where $\overline{\mathcal{D}}_1$ is the closure of $\mathcal{D}_1$ defined in (3.89), $\mathcal{D}_0$ and $\mathcal{D}_2$ are defined in (3.86) and (3.90), respectively.

引理 3.5. *For any $\lambda \in \mathcal{R}_1 \bigcup \mathcal{R}_2$, where $\mathcal{R}_1$ and $\mathcal{R}_2$ are defined by (3.59) and (3.60), respectively, the sequence $\{z_k(\lambda)\}$ generated by scalar Euler iteration (3.38) with some $z_0 \in \mathbb{C}\backslash\{0\}$ for solving (3.48) converges to the principal pth root $\lambda^{1/p}$. Moreover, if $\lambda \neq 0$, then the convergence order is 3.*

证明. We prove this lemma by four steps as follows.

Step 1. Suppose $\mathcal{R}_c$ is any closed domain in $\mathcal{R}_1$ or $\mathcal{R}_2$ and that $0 \notin \mathcal{R}_c$. We will prove in this step $\{z_k(\lambda)\}$ converges uniformly to $z(\lambda)$, a pth root of each $\lambda \in \mathcal{R}_c$, and that $z(\lambda)$ exists for each $\lambda \in \mathcal{R}_1 \bigcup \mathcal{R}_2$.

We write $r(z_k, \lambda) \triangleq r(z_k(\lambda), \lambda)$ for short below. Since $\sum\limits_{j=3}^{\infty} c_j r^j(z_0, \lambda)$ is analytic for each $\lambda \in \mathcal{R}_1 \bigcup \mathcal{R}_2$ and $\mathcal{R}_c \subset \mathcal{R}_1 \bigcup \mathcal{R}_2$ is bounded, there is a $\widehat{\lambda} \in \partial \mathcal{R}_c$ such that

$$\left| \sum_{j=3}^{\infty} c_j r^j(z_0, \widehat{\lambda}) \right| = \max_{\lambda \in \mathcal{R}_c} \left| \sum_{j=3}^{\infty} c_j r^j(z_0, \lambda) \right| \tag{3.61}$$

by the theorem of maximum modulus of analytic function. Let

$$q(z_0) := \begin{cases} \max\limits_{\lambda \in \mathcal{R}_c} q_1(z_0, \lambda), & \text{if } \mathcal{R}_c \subset \mathcal{R}_1, \\ \max\limits_{\lambda \in \mathcal{R}_c} q_2(z_0, \lambda), & \text{if } \mathcal{R}_c \subset \mathcal{R}_2, \end{cases}$$

where $q_1(z_0, \lambda)$ and $q_2(z_0, \lambda)$ are defined in (3.56) and (3.57), respectively. Then

$$q(z_0) = \begin{cases} q_1(z_0, \widehat{\lambda}) < 1, & \text{if } \mathcal{R}_c \subset \mathcal{R}_1, \\ q_2(z_0, \widehat{\lambda}) < 1, & \text{if } \mathcal{R}_c \subset \mathcal{R}_2, \end{cases}$$

and

$$|r(z_k, \lambda)| \leq \begin{cases} q^{3^{k-1}}(z_0), & \text{if } \mathcal{R}_c \subset \mathcal{R}_1, \\ q^{3^k - 1}(z_0) \cdot r_*, & \text{if } \mathcal{R}_c \subset \mathcal{R}_2, \end{cases} \quad k = 1, 2, \ldots, \lambda \in \mathcal{R}_c \tag{3.62}$$

by (3.61), Lemmas 3.3 and 3.4, where $r_* := \max\limits_{\lambda \in \mathcal{R}_c} |r(z_0, \lambda)|$ is a positive real independent on $\lambda \in \mathcal{R}_c$. It follows $\{r(z_k, \lambda)\}$ converges uniformly to 0 with order 3 as $k \to \infty$ for all $\lambda \in \mathcal{R}_c$ and that identities

$$z_k^p = \frac{\lambda}{1 - r(z_k, \lambda)}, \quad \lambda \in \mathcal{R}_c, \ k = 0, 1, \ldots \tag{3.63}$$

give the sequence $\{z_k(\lambda)\}$ is bounded uniformly for all $\lambda \in \mathcal{R}_c$. Thus, there is a $M > 0$, independent on k and $\lambda \in \mathcal{R}_c$, such that

$$\frac{1}{2p^2} |z_k(\lambda)| |2p + (p-1) r(z_k, \lambda)| \leq M, \quad k \geq 0, \lambda \in \mathcal{R}_c. \tag{3.64}$$

By (3.39) and (3.64),

$$\begin{aligned} |z_{k+1}(\lambda) - z_k(\lambda)| &= \left| \frac{1}{2p^2} z_k(\lambda) \left[(3p-1) - 2(2p-1)\lambda z_k^{-p}(\lambda) + (p-1)\lambda^2 z_k^{-2p}(\lambda) \right] \right| \\ &= \left| \frac{1}{2p^2} z_k(\lambda) \left[(1-p) + 2(2p-1) r(z_k, \lambda) + (p-1)(1 - r(z_k, \lambda))^2 \right] \right| \end{aligned}$$

$$
\begin{aligned}
&= \left| \frac{1}{2p^2} z_k(\lambda) \left[2pr(z_k, \lambda) + (p-1) r(z_k, \lambda)^2 \right] \right| \\
&= \frac{1}{2p^2} |z_k(\lambda)| |r(z_k, \lambda)| |2p + (p-1) r(z_k, \lambda)| \\
&\leq M |r(z_k, \lambda)|, \quad \lambda \in \mathcal{R}_c, \ k = 0, 1, \ldots,
\end{aligned}
$$

which and (3.62) conclude that $\{z_{k+1}(\lambda) - z_k(\lambda)\}$ is majoriant by a geometrical sequence which converges to zero and independent on $\lambda \in \mathcal{R}_c$. So, $\{z_k(\lambda)\}$ is a Cauchy sequence for each $\lambda \in \mathcal{R}_c$ and there is $z(\lambda)$ defined on $\mathcal{R}_c$ such that $z_k(\lambda)$ converges uniformly to $z(\lambda)$ for all $\lambda \in \mathcal{R}_c$. Let $k \to \infty$ in (3.63), we have

$$
z^p(\lambda) = \lambda, \quad \lambda \in \mathcal{R}_c.
$$

That is, $z(\lambda)$ is a pth root of $\lambda \in \mathcal{R}_c$.

Step 2. Since for any $\lambda \in \mathcal{R}_1 \bigcup \mathcal{R}_2$, there is a closed domain of $\mathcal{R}_1$ or $\mathcal{R}_2$ such that λ belongs to it. By Step 1, $z(\lambda)$ exists for each $\lambda \in \mathcal{R}_1 \bigcup \mathcal{R}_2$. In this step, we will show $z(\lambda)$ obtained in Step 1 is analytic in $\mathrm{Int}(\mathcal{R}_1)$, the interior of $\mathcal{R}_1$, and $\mathcal{R}_2$. Therefore, $z(\lambda)$ located in a single-valued branch of root function in $\mathrm{Int}(\mathcal{R}_1)$ and $\mathcal{R}_2$.

In fact, by the definition of $z_k(\lambda), \lambda \in \mathcal{R}_1 \bigcup \mathcal{R}_2$, we have $z_k(\lambda)$ is analytic in $\mathrm{Int}(\mathcal{R}_1)$ or $\mathcal{R}_2$. Since $\{z_k(\lambda)\}$ converges uniformly to $z(\lambda)$ by Step 1, $z_k(\lambda)$ is analytic in $\mathrm{Int}(\mathcal{R}_1)$ and $\mathcal{R}_2$ by Weierstrass theorem, seperately. Recall that $z(\lambda)$ is a pth root of $\lambda \in \mathrm{Int}(\mathcal{R}_1)$ or $\mathcal{R}_2$, we get $z(\lambda)$ located in a single-valued branch of root function for $\lambda \in \mathrm{Int}(\mathcal{R}_1)$ or $\lambda \in \mathcal{R}_2$, seperately.

Step 3. In this step, we will show that $z(\lambda) \to z(\lambda_0)$ as $\lambda \to \lambda_0$ from the inner side of $\mathcal{R}_1$, where $\lambda_0 \in \partial \mathcal{R}_1$ and $\lambda_0 \neq 0$.

It is clear that $|r(z_0, \lambda_0)| < 1$ for any $\lambda_0 \in \partial \mathcal{R}_1$ and $\lambda_0 \neq 0$. Then, there is $\delta_0 > 0$ such that closed domain $\overline{O}(\lambda_0, \delta_0) \bigcap \mathcal{R}_1$ does not contains 0 and 1, where $\overline{O}(\lambda_0, \delta_0) := \{\lambda \in \mathbb{C} : |\lambda - \lambda_0| \leq \delta_0\}$. By Step 1, $\{z_k(\lambda)\}$ converges uniformly to $z(\lambda)$ for any λ in $\overline{O}(\lambda_0, \delta_0) \bigcap \mathcal{R}_1$. So, for any $\varepsilon > 0$, there exists $K > 0$ such that for all $k \geq K$ and $\lambda \in \overline{O}(\lambda_0, \delta_0) \bigcap \mathcal{R}_1$,

$$
|z_k(\lambda) - z(\lambda)| < \frac{\varepsilon}{3}. \tag{3.65}
$$

Since $z_k(\lambda)$ is analytic in $\mathcal{R}_1$ derives $z_k(\lambda)$ is continuous in $\overline{O}(\lambda_0, \delta_0) \bigcap \mathcal{R}_1$, there exists $0 < \delta_1 < \delta_0$ such that $\overline{O}(\lambda_0, \delta_1) \subset \overline{O}(\lambda_0, \delta_0)$ and

$$|z_k(\lambda) - z_k(\lambda_0)| < \frac{\varepsilon}{3}, \quad \forall\ \lambda \in \overline{O}(\lambda_0, \delta_1) \bigcap \mathcal{R}_1. \tag{3.66}$$

Thus, for all λ in $\overline{O}(\lambda_0, \delta_1) \bigcap \mathcal{R}_1$, we have

$$|z(\lambda) - z(\lambda_0)| \leq |z(\lambda) - z_k(\lambda)| + |z_k(\lambda) - z_k(\lambda_0)| + |z_k(\lambda_0) - z(\lambda_0)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$$

by (3.65) and (3.66). That is, $z(\lambda)$ is continuous at $\lambda = \lambda_0$. The arbitrariness of λ_0 completes the proof of Step 3.

Step 4. This is the last step of the proof.

By Step 2, $z(\lambda)$ is analytic in $\mathrm{Int}(\mathcal{R}_1)$ and it is located in a single-valued branch of pth root function. Since $z_k(1) \equiv 1$ for all k implies $z(1) = 1$, we have that $z(\lambda)$ located in the single-valued branch containing 1. That is, $z(\lambda)$ is the principal pth root of each λ in $\mathrm{Int}(\mathcal{R}_1)$. Since $z(\lambda) \to z(\lambda_0)$ as $\lambda \to \lambda_0$ ($\lambda \in \mathrm{Int}(\mathcal{R}_1)$) for each $\lambda_0 \in \partial\mathcal{R}_1$ by Step 3, we get $z(\lambda_0)$ is also the principal pth root of $\lambda_0 \in \partial\mathcal{R}_1$.

By now, we have proved that $z(\lambda)$ is the principal pth root of each $\lambda \in \mathcal{R}_1$ except $\lambda = 0$. When $\lambda = 0$, we have by (3.38) that

$$z_k(0) = \frac{2p^2 - 3p + 1}{2p^2} z_{k-1}(0) = \left(\frac{2p^2 - 3p + 1}{2p^2}\right)^k z_0, \quad k = 0, 1, 2, \ldots.$$

So, $\{z_k(0)\}$ converges to 0, the principal pth root of 0, linearly. Therefore, $z(\lambda)$ is the principal pth root of each $\lambda \in \mathcal{R}_1$.

By Step 2 again, $z(\lambda)$ is analytic and locates in a single-valued branch of pth root function for each $\lambda \in \mathcal{R}_2$. Now, we can see that, if we can show $\mathcal{R}_2$ contains some part of $\partial\mathcal{R}_1$, then the single-valued branch of $z(\lambda)$ for $\lambda \in \mathcal{R}_2$ is the same branch of $z(\lambda)$ for $\lambda \in \mathcal{R}_1$, which deduces that $z(\lambda)$ is the principle pth root of $\lambda \in \mathcal{R}_1 \bigcup \mathcal{R}_2$. So, what we need to do is to prove $\mathcal{R}_2$ contains some part of $\partial\mathcal{R}_1$.

Set $\lambda_0 = z_0^p$ and define

$$S_m(z_0, \lambda) := \sum_{j=3}^{m} c_j r^{j-1}(z_0, \lambda), \quad S_\infty(z_0, \lambda) := \sum_{j=3}^{\infty} c_j r^{j-1}(z_0, \lambda), \quad m \geq 3, \lambda \in \partial\mathcal{R}_1.$$

Clearly, $S_\infty(z_0,\lambda)$ is continuous with respect to λ on $\partial\mathcal{R}_1$ and

$$0 = |S_\infty(z_0,\lambda_0)| = \min_{\lambda\in\partial\mathcal{R}_1} |S_\infty(z_0,\lambda)| \le |S_\infty(z_0,\lambda)| \le 1, \quad \forall\ \lambda \in \partial\mathcal{R}_1. \tag{3.67}$$

Since $|S_m(z_0,\lambda)| < 1$ holds for all $\lambda \in \mathcal{R}_1 \bigcup \mathcal{R}_2$, we can choose $M < 1$ be a real satisfying

$$\sup_{m\ge 3} |S_m(z_0,\lambda)| < M, \quad \lambda \in \partial\mathcal{R}_1.$$

By (3.67), there is $\delta > 0$ such that once $|\lambda - \lambda_0| < \delta$, then

$$|S_\infty(z_0,\lambda)| < \frac{\frac{1}{M}-1}{\frac{1}{M}+1},$$

or equivalently,

$$q_2(z_0,\lambda) = \sup_{m\ge 3} |S_m(z_0,\lambda)| \cdot \frac{1+|S_\infty(z_0,\lambda)|}{1-|S_\infty(z_0,\lambda)|} < M \cdot \frac{1}{M} = 1.$$

Therefore, $\mathcal{R}_2$ contains some part of $\partial\mathcal{R}_1$. The proof is completed. □

Choosing $z_0 \equiv 1$ in $\mathcal{R}_1$ and $\mathcal{R}_2$ given by (3.59) and (3.60), respectively, one has that $\mathcal{R} = \mathcal{R}_1 \bigcup \mathcal{R}_2$. So, we obtain from Lemma 3.5 the following corollary:

推论 3.1. *For any $\lambda \in \mathcal{R}$ defined in (3.87), the sequence $\{z_k(\lambda)\}$ generated by scalar Euler iteration (3.38) with $z_0 = 1$ for solving (3.48) converges to the principal pth root $\lambda^{1/p}$. Moreover, if $\lambda \neq 0$, then the convergence order is 3.*

Next, we consider the case of matrix form. For the matrix $A \in \mathbb{C}^{n\times n}$ defined in (3.37), let

$$R(X) := \mathrm{I} - AX^{-p}, \quad p \ge 2 \tag{3.68}$$

for any nonsingular matrix $X \in \mathbb{C}^{n\times n}$.

引理 3.6. *Let $X \in \mathbb{C}^{n\times n}$ be a nonsingular matrix commuting with A and $R(X)$ be defined in (3.68). If the spectrum $\sigma(R(X)) \subset \mathcal{D}_0$ for some nonsingular matrix $X \in \mathbb{C}^{n\times n}$, where $\mathcal{D}_0$ is defined in (3.86), then $E(X)$ generated by (3.39) is nonsingular, commutes with A and*

$$R(E(X)) = \mathrm{I} - \left[\mathrm{I} - \frac{1}{p}R(X) - \frac{1}{2}\left(\frac{1}{p} - \frac{1}{p^2}\right)R^2(X)\right]^{-p} \cdot (\mathrm{I} - R(X)) = \phi(R(X)). \tag{3.69}$$

证明. Clearly, $E(X)$ given by (3.39) exists for any nonsingular matrix $X \in \mathbb{C}^{n\times n}$. Since $\sigma(R(X)) \subset \mathcal{D}_0\backslash\{0\}$, we get

$$\begin{aligned}\rho\left(\frac{1}{p}R(X)+\frac{1}{2}\left(\frac{1}{p}-\frac{1}{p^2}\right)R^2(X)\right) &\le \frac{1}{p}\cdot\rho(R(X))+\frac{1}{2}\left(\frac{1}{p}-\frac{1}{p^2}\right)\rho^2(R(X))\\ &= \frac{1}{p}\cdot\rho(R(X))\cdot\left(1+\frac{p-1}{2p}\rho(R(X))\right)\\ &< \frac{1}{p}\cdot\frac{p}{p-1}\left(\sqrt{2p-1}-1\right)\cdot\left(1+\frac{\sqrt{2p-1}-1}{2}\right)\\ &= 1.\end{aligned}$$

Here, the first inequality follows from $\frac{1}{p}R(X)+\frac{1}{2}(\frac{1}{p}-\frac{1}{p^2})R^2(X)$ is a polynomial of the matrix $R(X)$. So

$$\mathrm{I}-\frac{1}{p}R(X)-\frac{1}{2}\left(\frac{1}{p}-\frac{1}{p^2}\right)R^2(X)$$

is nonsingular by Neumann Lemma. It follows that

$$\begin{aligned}E(X) &= \frac{1}{2p^2}X\left[(2p^2-3p+1)\mathrm{I}+2(2p-1)AX^{-p}-(p-1)(AX)^2\right]\\ &= X\left[\frac{2p^2-3p+1}{2p^2}\mathrm{I}+\frac{2p-1}{p^2}(\mathrm{I}-R(X))-\frac{p-1}{2p^2}(1-R(X))^2\right]\\ &= X\left[\mathrm{I}-\frac{1}{p}R(X)-\frac{1}{2}\left(\frac{1}{p}-\frac{1}{p^2}\right)R^2(X)\right]\end{aligned}$$

is also nonsingular and commutes with A, and

$$R(E(X)) = \mathrm{I}-\left[\mathrm{I}-\frac{1}{p}R(X)-\frac{1}{2}\left(\frac{1}{p}-\frac{1}{p^2}\right)R^2(X)\right]^{-p}\cdot(\mathrm{I}-R(X)) = \phi(R(X))$$

by the assumption of X commutes with A. This completes the proof. □

Thanks to Lemma 3.6, we have

推论 3.2. *If $X_0 \in \mathbb{C}^{n\times n}$ commutes with A and the spectrum $\sigma(R(X_0)) \subset \mathcal{D}_0$, where $\mathcal{D}_0$ is defined in (3.86), then the sequence $\{X_k\}$ starting from X_0 generated by Euler's method (3.38) for solving (3.37) exists.*

引理 3.7. *If the spectrum $\sigma(R(X)) \subset \mathcal{D}_1\backslash\{0\}$ for some nonsingular matrix $X \in \mathbb{C}^{n\times n}$, where $R(X)$ be defined in (3.68) and $\mathcal{D}_1$ is defined in (3.89), then, there is a sub-multiplicative matrix norm $\|\cdot\|$ such that $\|R(X)\| \le 1$ and*

$$\|R(E(X))\| \le \frac{\phi(\|R(X)\|)}{\|R(X)\|^3} \cdot \|R(X)\|^3 < \|R(X)\|^3, \tag{3.70}$$

where ϕ is defined by (3.85).

证明. It follows from Lemma 3.6 that $R(E(X))$ exists. Since the spectrum $\sigma(R(X)) \subset \mathcal{D}_1$, the spectral radius of $R(X)$ is less than 1. So, there is a sub-multiplicative matrix norm $\|\cdot\|$ such that $\|R(X)\| < 1$. Note that, for any $u \in (0,1)$, It follows from Lemma 3.1 that

$$\frac{\phi(u)}{u^3} = \sum_{j=3}^{\infty} c_j u^{j-3} < \sum_{j=3}^{\infty} c_j = 1.$$

Thus, this together with (3.69) gives that

$$\|R(E(X))\| = \|\phi(R(X))\| \le \phi(\|R(X)\|) = \frac{\phi(\|R(X)\|)}{\|R(X)\|^3}\|R(X)\|^3 < \|R(X)\|^3,$$

which shows (3.70). The proof is completed. □

引理 3.8. *Let $R(X)$ be defined in (3.68). Suppose the spectrum $\sigma(R(X_0)) \subset \mathcal{D}_1\backslash\{0\}$ for some nonsingular matrix $X_0 \in \mathbb{C}^{n\times n}$ and that $\|R(X_0)\| < 1$ for a sub-multiplicative matrix norm $\|\cdot\|$, where $\mathcal{D}_1$ is defined in (3.89). Let $\{X_k\}$ be the sequence starting from X_0 generated by Euler's method (3.38) for solving (3.37). Then we have*

$$\|R(X_k)\| \le q^{3^k-1}(X_0) \cdot \|R(X_0)\|, \quad k = 1, 2, \ldots, \tag{3.71}$$

where

$$q(X_0) := \sqrt{\frac{\phi(\|R(X_0)\|)}{\|R(X_0)\|}} < 1 \tag{3.72}$$

and ϕ is defined by (3.85).

证明. For X_0 chosen, by (3.70) in Lemma 3.7, we have

$$\|R(X_1)\| \le \frac{\phi(\|R(X_0)\|)}{\|R(X_0)\|^3}\|R(X_0)\|^3 = q^2(X_0)\cdot\|R(X_0)\|.$$

If (3.71) holds for some $k \ge 1$, then by Lemma 3.7 again, one has that

$$\begin{aligned}\|R(X_{k+1})\| &\le \frac{\phi(\|R(X_k)\|)}{\|R(X_k)\|^3}\|R(X_k)\|^3\\ &\le \frac{\phi(\|R(X_0)\|)}{\|R(X_0)\|^3}\left[q^{3^k-1}(X_0)\right]^3\cdot\|R(X_0)\|^3\\ &= [q(X_0)]^{3^{k+1}-1}\cdot\|R(X_0)\|.\end{aligned}$$

Thus, by induction, (3.71) holds for all $k \ge 1$. This completes the proof. □

引理 3.9. *Let $R(X)$ be defined in (3.68). If the spectrum $\sigma(R(X_0)) \subset \mathcal{D}_2 \bigcap \mathcal{D}_0 \backslash \{0\}$ for some nonsingular matrix $X_0 \in \mathbb{C}^{n\times n}$, where $\mathcal{D}_0$ and $\mathcal{D}_2$ are defined in (3.86) and (3.90), respectively. Let $\{X_k\}$ be the sequence starting from X_0 generated by Euler's method (3.38) for solving (3.37). Then, there exists $\widehat{N} > 0$ such that*

$$\|R(X_k)\| \le [q(X_{\widehat{N}})]^{3^{k-\widehat{N}}-1}\cdot\|R(X_{\widehat{N}})\|, \quad \forall\ k > \widehat{N}, \tag{3.73}$$

where $q(X_{\widehat{N}})$ is defined as $q(X_0)$ in (3.72) by substituting $X_{\widehat{N}}$ for X_0.

证明. For any $r(z_0) \in \sigma(R(X_0))$, let $\{z_k\}$ be the sequence generated by the scalar form of (3.38) starting from z_0. It follows from Lemma 3.4 that there exists $N > 0$ such that $|r(z_N)| < 1$. Then, we can get from Lemma 3.3 that

$$|r(z_k)| < |r(z_N)|^{3^{k-N}}, \quad \forall\ k > N.$$

Define

$$\widehat{N} := \max_{r(z_0)\in\sigma(R(X_0))}\{N : \text{choose a } N > 0 \text{ such that } |r(z_N)| < 1\}.$$

Then, there exists a sub-multiplicative matrix norm $\|\cdot\|$ such that $\|X_{\widehat{N}}\| < 1$. Thus, (3.73) follows from Lemma 3.8. This completes the proof. □

3.7.2 Proof of Theorem 3.1

The following lemma is taken from [? , Theorem 4.15], which allows us to deduce convergence of the matrix iteration sequence generated by Euler's method (3.38).

引理 3.10 ([?]). *Suppose that $g(x,t)$ is a rational function with respect to its two variables and that $x^* = f(\lambda)$ is an attracting fixed point of the iteration $x_{k+1} = g(x_k, \lambda), x_0 = \phi_0(\lambda)$, where ϕ_0 is a rational function and $\lambda \in \mathbb{C}$. Then, the matrix sequence generated by $X_{k+1} = g(X_k, J(\lambda)), X_0 = \phi_0(J(\lambda))$, converges to a matrix X^* with $(X^*)_{ii} \equiv f(\lambda)$, $i = 1, 2, \ldots, m$, where $J(\lambda) \in \mathbb{C}^{m\times m}$ is a Jordan block.*

Now, we can prove Theorem 3.1 by applying the above lemma together with the lemmas proposed in Section 3.1.

The proof of Theorem 3.1. By Corollary 3.2, the matrix sequence $\{X_k\}$ generated by Euler's method (3.38) starting from $X_0 = \mathrm{I}$ is well defined when the eigenvalues of A are in $\mathcal{R} \subset \mathcal{D}_0$. Thanks to Lemma 3.10, $\{X_k\}$ converges to the principal pth root of A follows from Corollary 3.1. The first part of theorem is completed.

For the second part, let $X_* = A^{1/p}$ and $E_k = X_k - X_*$ for $k \geq 0$. Due to the commutativity of X_k and X_*, we have

$$\begin{aligned} R(X_k) &= \mathrm{I} - AX_k^{-p} = (X_k^p - X_*^p)X_k^{-p} \\ &= (X_k - X_*)\left(X_k^{p-1} + X_k^{p-2}X_* + \cdots + X_kX_*^{p-2} + X_*^{p-1}\right)X_k^{-p} \\ &= E_k\left(X_k^{p-1} + X_k^{p-2}X_* + \cdots + X_kX_*^{p-2} + X_*^{p-1}\right)X_k^{-p}, \quad \forall\, k \geq 0 \end{aligned} \tag{3.74}$$

Set

$$Y_k := \sum_{i=1}^{p} X_k^{p-i}X_*^{i-1} = X_k^{p-1} + X_k^{p-2}X_* + \cdots + X_kX_*^{p-2} + X_*^{p-1}.$$

Since X_k converges to $A^{1/p}$ and all the eigenvalues of $A^{1/p}$ are not in $\mathbb{R}^-$, there exists nonnegative integer $N > 0$ such that the eigenvalues X_k are not in $\mathbb{R}^-$ for all $k \geq N$. Thus, the eigenvalues of Y_k are also not in $\mathbb{R}^-$ and so Y_k is nonsingular for $k \geq N$. Then, it follows from (3.74) that

$$E_{k+1} = R(X_{k+1})X_{k+1}^p Y_{k+1}^{-1}, \quad k \geq N. \tag{3.75}$$

Thanks to (3.71) and (3.73) in Lemmas 3.8 and 3.9, respectively, there exists $K_0 > 0$ such that $\|R(X_{k+1})\| < \|R(X_k)\|^3$ for any $k \geq K_0$. It follows from (3.74) and (3.75) that

$$\begin{aligned}\|E_{k+1}\| &\leq \|R(X_{k+1})\|\|X_{k+1}\|^p\|Y_{k+1}^{-1}\| \\ &< \|R(X_k)\|^3\|X_{k+1}\|^p\|Y_{k+1}^{-1}\| \\ &\leq \left(\|X_k^{-1}\|^p\|X_{k+1}\|^p\|Y_k\|\|Y_{k+1}^{-1}\|\right)\|E_k\|^3, \quad \forall\, k \geq K_0. \end{aligned} \tag{3.76}$$

Thus $\{X_k\}$ is convergent guarantees that $\|X_k^{-1}\|^p\|X_{k+1}\|^p\|Y_k\|\|Y_{k+1}^{-1}\|$ is bounded for all $k \geq K_0$. (3.76) concludes that the local convergence order is 3. This completes the proof. □

3.8 A Preconditioned Schur Modification

Let's begin with considering the stability of Euler iteration (3.38). The analysis approach here follows the line of stability analysis for Newton iteration in [?].

Let A be nonsingular and diagonalizable with n different eigenvalues $\lambda_1, \ldots, \lambda_n$. Then, there exists a nonsingular matrix Z such that

$$Z^{-1}AZ = \Lambda := \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Suppose that the sequence $\{X_k\}$ generated by Euler iteration (3.38) starting from $X_0 = \mathrm{I}$ converges to the principle pth root $A^{1/p}$. Set $D_k := Z^{-1}X_kZ$, $k \geq 0$. It follows from (3.38) that

$$D_{k+1} = \frac{1}{2p^2}D_k\left[(2p^2 - 3p + 1)\mathrm{I} + 2(2p-1)\Lambda D_k^{-p} - (p-1)\left(\Lambda D_k^{-p}\right)^2\right], \quad k \geq 0.$$

Since $D_0 = Z^{-1}X_0Z = \mathrm{I}$, D_k is diagonal for any $k \geq 1$. Thus, if we set $D_k := \mathrm{diag}(d_1^{(k)}, \ldots, d_n^{(k)})$, then $d_j^{(k)}$ converges to $\lambda_j^{1/p}$ as $k \to \infty$. Let $\{\widetilde{X}_k\}$ be the sequence of computed iterates, then

$$\begin{aligned}\widetilde{X}_{k+1} &= \frac{1}{2p^2}\left((2p^2-3p+1)\widetilde{X}_k + 2(2p-1)A\widetilde{X}_k^{1-p} - (p-1)A^2\widetilde{X}_k^{1-2p}\right) \\ &= \frac{2p^2-3p+1}{2p^2}(X_k+\Delta_k) + \frac{2p-1}{p^2}A(X_k+\Delta_k)^{1-p} - \frac{p-1}{2p^2}A^2(X_k+\Delta_k)^{1-2p}.\end{aligned} \tag{3.77}$$

Set $\Delta_k = \widetilde{X}_k - X_k$ and $\widetilde{\Delta}_k := (\widetilde{\delta}_{ij}^{(k)}) = Z^{-1}\Delta_k Z$. Applying the following expansion [?]

$$(X+\Delta_X)^{1-p} = X^{1-p} - \sum_{\ell=1}^{p-1} X^{\ell-p}\Delta_X X^{-\ell} + O(\|\Delta_X\|^2)$$

to (3.77), where $\|\cdot\|$ denotes the 2-norm, one has that

$$\begin{aligned}\widetilde{X}_{k+1} = \frac{2p^2-3p+1}{2p^2}(X_k+\Delta_k) &+ \frac{2p-1}{p^2}A\left[X_k^{1-p} - \sum_{\ell=1}^{p-1} X_k^{\ell-p}\Delta_k X_k^{-\ell}\right] \\ &- \frac{p-1}{2p^2}A^2\left[X_k^{1-2p} - \sum_{\ell=1}^{2p-1} X_k^{\ell-2p}\Delta_k X_k^{-\ell}\right] + O(\|\Delta_k\|^2), \quad k \ge 0.\end{aligned}$$

We have for each $k \ge 0$,

$$\begin{aligned}\Delta_{k+1} &= \widetilde{X}_{k+1} - X_{k+1} \\ &= \frac{2p^2-3p+1}{2p^2}\Delta_k - \frac{2p-1}{p^2}A\sum_{\ell=1}^{p-1} X_k^{\ell-p}\Delta_k X_k^{-\ell} + \frac{p-1}{2p^2}A^2\sum_{\ell=1}^{2p-1} X_k^{\ell-2p}\Delta_k X_k^{-\ell},\end{aligned}$$

and

$$\begin{aligned}\widetilde{\Delta}_{k+1} = \frac{2p^2-3p+1}{2p^2}\widetilde{\Delta}_k &- \frac{2p-1}{p^2}\Lambda\sum_{\ell=1}^{p-1} D_k^{\ell-p}\widetilde{\Delta}_k D_k^{-\ell} \\ &+ \frac{p-1}{2p^2}\Lambda^2\sum_{\ell=1}^{2p-1} D_k^{\ell-2p}\widetilde{\Delta}_k D_k^{-\ell} + O(\|\Delta_k\|^2).\end{aligned}$$

This gives

$$\widetilde{\delta}_{ij}^{(k+1)} = \widetilde{\delta}_{ij}^{(k)}\pi_{ij}^{(k)} + O(\|\Delta_k\|^2), \quad i,j = 1,2,\ldots,n, \ k \ge 0, \tag{3.78}$$

where

$$\pi_{ij}^{(k)} = \frac{2p^2-3p+1}{2p^2} - \frac{2p-1}{p^2}\lambda_i \sum_{\ell=1}^{p-1} \frac{1}{(d_i^{(k)})^{p-\ell}(d_j^{(k)})^{\ell}} + \frac{p-1}{2p^2}\lambda_i^2 \sum_{\ell=1}^{2p-1} \frac{1}{(d_i^{(k)})^{2p-\ell}(d_j^{(k)})^{\ell}}.$$

Since $d_j^{(k)}$ converges to $\lambda_j^{1/p}$ as $k \to \infty$ for each $j = 1, \ldots, n$, we may write $d_j^{(k)} = \lambda_j^{1/p} + \varepsilon_j^{(k)}, j = 1, 2, \ldots, n,\ k \geq 0$, where $\varepsilon_j^{(k)} \to 0$ as $k \to \infty$. Hence, for each $i, j = 1, 2, \ldots, n,\ k = 0, 1, \ldots,$

$$\pi_{ij}^{(k)} = \frac{2p^2-3p+1}{2p^2} - \frac{2p-1}{p^2} \sum_{j=1}^{p-1} \left(\frac{\lambda_i}{\lambda_j}\right)^{j/p} + \frac{p-1}{2p^2} \sum_{j=1}^{2p-1} \left(\frac{\lambda_i}{\lambda_j}\right)^{j/p} + O(\varepsilon^{(k)}). \quad (3.79)$$

where $\varepsilon^{(k)} := \max_i |\varepsilon_i^{(k)}|$ and the constant interfere with “O” is independent on $i, j = 1, 2, \ldots, n$.

To guarantee the numerical stability of Euler iteration (3.38), we should require from (3.78) and (3.79) that

$$\left| \frac{2p^2-3p+1}{2p^2} - \frac{2p-1}{p^2} \sum_{j=1}^{p-1} \left(\frac{\lambda_i}{\lambda_j}\right)^{j/p} + \frac{p-1}{2p^2} \sum_{j=1}^{2p-1} \left(\frac{\lambda_i}{\lambda_j}\right)^{j/p} \right| \leq 1, \quad i, j = 1, \ldots, n. \quad (3.80)$$

Obviously, this is a very restrictive condition on A, even for that A is Hermitian positive definite. For example, in the case $p = 2$, (3.80) becomes

$$-3 \leq -5\kappa_2(A)^{1/2} + \kappa_2(A) + \kappa_2(A)^{3/2} \leq 5,$$

which is equivalent to $\kappa_2(A) \leq 5$, where $\kappa_2(A) := \|A\|_2 \|A^{-1}\|_2$.

The above defects of Euler iteration (3.38) should be weakened. One way often used is to modify Euler iteration (3.38) into the following coupled version by introducing the auxiliary matrix $N_k (k \geq 0)$:

$$\begin{cases} X_{k+1} = X_k \left(\dfrac{(2p^2-3p+1)\mathrm{I} + 2(2p-1)N_k - (p-1)N_k^2}{2p^2} \right), & X_0 = \mathrm{I}, \\ N_{k+1} = \left(\dfrac{(2p^2-3p+1)\mathrm{I} + 2(2p-1)N_k - (p-1)N_k^2}{2p^2} \right)^{-p} N_k, & N_0 = A. \end{cases} \quad (3.81)$$

Clearly, $N_k = AX_k^{-p}$ and $\{X_k\}$ generated by (3.81) is same as the sequence of Euler method (3.38). We call it coupled Euler's iteration. When the sequence

$\{X_k\}$ generated by (3.81) converges to $A^{1/p}$, N_k converges to I. The computational cost of iteration (3.81) is $2(4+\vartheta\log_2 p)n^3$ flops per step for some $\vartheta \in [1,2]$ by means of the binary powering technique [**?** , Algorithm 11.2.2].

Note that, while we purely use coupled Euler iteration (3.81) to compute $A^{1/p}$, bad numerical results will still appear. A simple example (see TEST 1 in Section 4) illustrates this observation.

In order to avoid poor numerical results, we make Schur decomposition before we begin to use coupled Euler iteration (3.81), similar to Algorithms 3 and 4 in [**?**], based on the idea of Algorithm 3.3 in [**?**]. New algorithm is given as follows.

算法 3.3 Schur-Euler algorithm using (3.81) for computing $A^{1/p}$

Given $A \in \mathbb{C}^{n\times n}$ with no nonpositive real eigenvalues, an integer $p = 2^{k_0}q$ with $k_0 \geq 0$ and q odd. This algorithm computes $A^{1/p}$ via a Schur decomposition and Euler iteration.

1. Compute the Schur decomposition of $A = QRQ^*$;

2. If $q = 1$, then $k_1 = k_0$; else choose the smallest $k_1 \geq k_0$ such that for each eigenvalue λ of A, $\lambda^{1/2^{k_1}} \in \mathcal{R}_{\mathrm{E}}$ defined in (3.46);

3. Compute $B = R^{1/2^{k_1}}$ by taking the square root k_1 times; if $q = 1$, then set $X = QBQ^{\mathrm{T}}$; else continue;

4. Compute $C = B^{1/q}$ by using the coupled Euler iteration (3.81) and set $X = QC^{2^{k_1-k_0}}Q^{\mathrm{T}}$.

Recall that, one square root costs $n^3/3$ flops and the evaluation of (3.81) costs $2(4+\vartheta\log_2 p)n^3$ flops, where $\vartheta \in [1,2]$. Thus, $25n^3$ flops for the real Schur decomposition plus k_1 times square root, $k_1 - k_0$ times matrix multiplications and $4n^3$ flops to form X, the total computational cost of Algorithm 3.3 is about

$$\left(29 + \frac{k_1}{3} + 2(k_1 - k_0) + 2k_2(4 + \vartheta\log_2 p)\right) n^3 \text{ flops}, \quad \vartheta \in [1,2],$$

where we assume that k_2 iterations of (3.81) are done exactly. Numerical experiments in Section 4 show that, Algorithm 3.3 has a well numerical behavior. Moreover, in most cases, Algorithm 3.3 has less computational time and also can save some square roots in preprocessing.

3.9 Numerical Experiments

The goal of numerical experiments in this section is to illustrate that Euler's method is as good as Newton method and Halley method for computing the principal pth root of a matrix. To support this claim, we now test Algorithm 3.3 and compare its computational performance with existing three algorithms presented in [? ?] on computational error, computational time and the number of iterations.

Upto 6 Tests are performed here. Test 1 is to show that it may suffers from bad numerical results when we purely use coupled Euler iteration (3.81) to compute $A^{1/p}$ even though the matrix A is simple. Tests 2-6 used by many authors [? ? ? ? ?] are to compare numerical behavior with other algorithms. The results show that, in most cases, Algorithm 3.3 requires less computational time and less steps of iterations. Meanwhile, it has more numerical accuracy of the computed solution than others algorithms.

Our numerical experiments were carried out in MATLAB 7.0 running on a PC Intel Pentium P6200 of 2.13 GHz CPU. To measure of the quality of a computed solution X, we use the relative residual $\rho_A(X)$ and relative error $\mathrm{err}(X)$ as follows:

$$\rho_A(X) = \frac{\|A - X^p\|}{\|X\| \|\sum_{j=0}^{p-1} (X^{p-1-j})^{\mathrm{T}} \otimes X^j\|}, \quad \mathrm{err}(X) = \frac{\|A - X^p\|}{\|A\|}, \tag{3.82}$$

where $\otimes$ denotes the Kronecker product and $\|\cdot\|$ denotes the Frobenius norm. Note that the relative residual $\rho_A(X)$ (given in [?]) is more practically useful definition of relative residual (e.g., for testing purposes) than relative error and that the averaged CPU time computed by the standard MATLAB function cputime. The averaged time was computed by repeating 100 times for each test matrix. Moreover, we use 'iter' to stand for the number of the iterations.

For simplicity, in the following tests, we denote

1. SE: Schur-Euler algorithm, Algorithm 3.3;
2. PSN: parameter Schur-Newton algorithm from [? , Algorithm 3.3];
3. SN: Schur-Newton algorithm from [? , Algorithm 3];
4. SH: Schur-Halley algorithm from [? , Algorithm 4].

The iterations in the above four algorithms are stopped when $\|N_k - \mathrm{I}\| < \sqrt{n}u/2$, where n is the size of A and $u = 2^{-52} \approx 2.2204\mathrm{e}-16$.

TEST 1. We first give a simple example to illustrate that it usually suffers from bad numerical results when we purely use coupled Euler iteration (3.81) to compute $A^{1/p}$. Consider the following simple 3×3 matrix

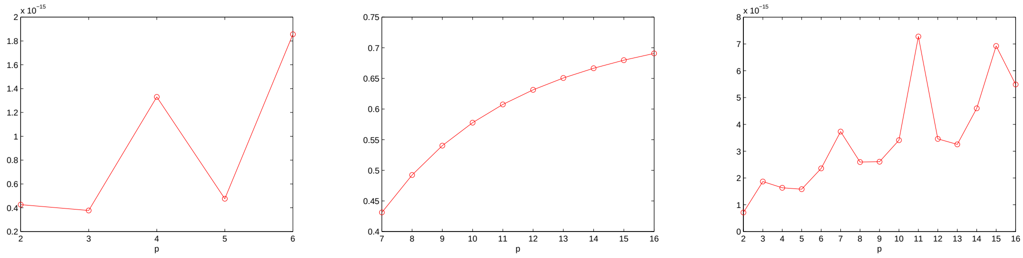
$$A = \begin{bmatrix} 1 & 1/2 & 0 \\ 1/2 & 2 & 1/2 \\ 0 & 1/2 & 3 \end{bmatrix},$$

and compute the pth root $A^{1/p}$ for $p = 2:15$. In Figure 3.8, we give the relative error $\mathrm{err}(X)$ for each p. As one can see, for $p = 2:6$ the coupled Euler iteration (3.81) gives good relative error (Figure 3.8 (a)), but the errors deteriorate as $p = 7:15$ (Figure 3.8 (b)). Algorithm 3.3 (taking some preprocessing before using oupled Euler iteration (3.81)) shows good numerical stability (Figure 3.8 (c)).

TEST 2. In this test, we compare the relative error on computing the principal pth root of the matrices [?] (depend on variable ε)

$$A(\varepsilon) = \begin{bmatrix} 1 & 1 \\ 0 & 1+\varepsilon \end{bmatrix} \tag{3.83}$$

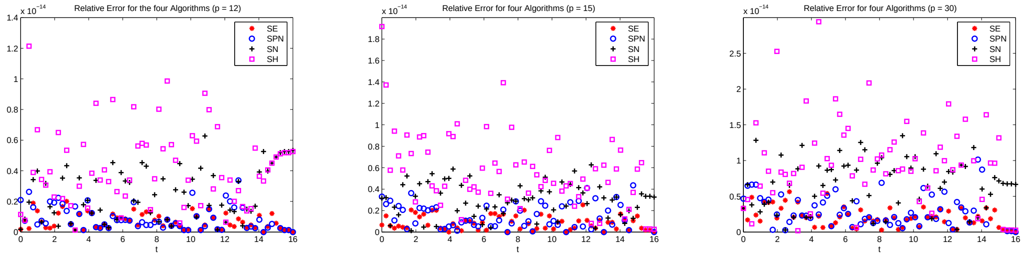
for $p = 12, 15, 30$. We choose $\varepsilon = 10^{-t}$ with 65 equally spaced values of $t \in [0, 16]$. As is pointed out in [?], $A(\varepsilon)$ approaches a defective matrix as t increases. The relative errors for the four algorithms are shown in Figure 3.9. The numerical results of Algorithm 3.3 are very good. Note that, there exists 9 values of t



(a)　(b)　(c)

图 3.8: (a) and (b) show the relative error for the pth root $A^{1/p}$ by using coupled Euler iteration (3.81); (c) is the result of Algorithm 3.3.

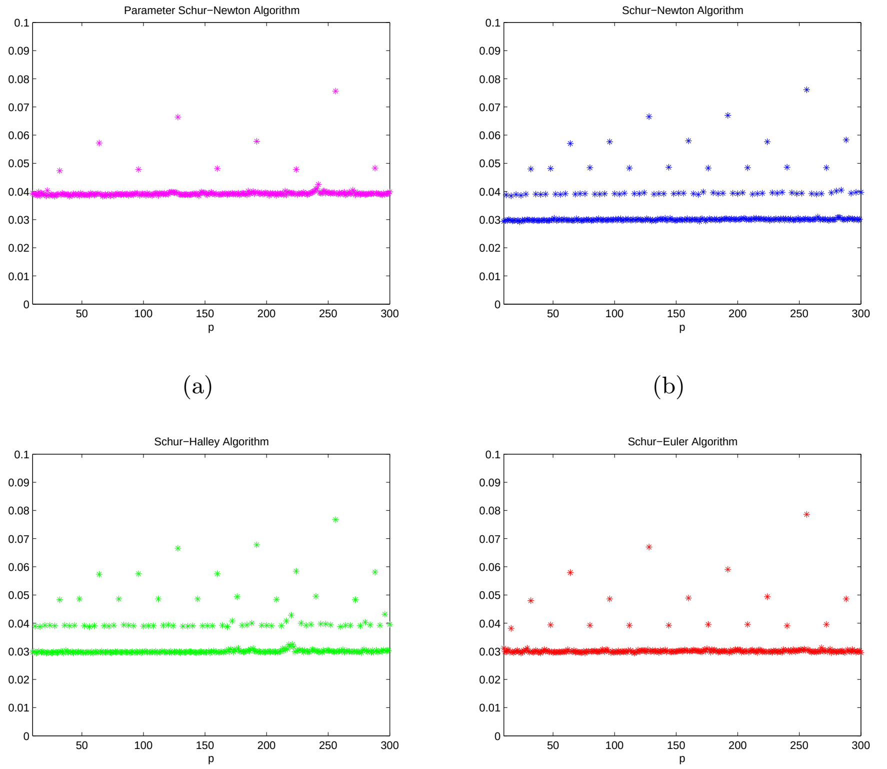
cannot be computed by using PSN while t approaches to 16. We also computed the relative residuals of these four algorithms on matrix $A(\varepsilon)$, the results were broadly similar to those shown in Figure 3.9.



(a) $p = 12$　(b) $p = 15$　(c) $p = 30$

图 3.9: The relative errors for the four algorithms on matrix (3.83).

TEST 3. To compare the computational time of the existing three algorithms with Algorithm 3.3, we select the classical test matrix (Frank matrix) from the MATLAB gallery function with 15 × 15 matrix which no nonpositive real eigenvalues and compute the principal pth root for $10 \leq p \leq 300$. Note that, the Frank matrix is an upper Hessenbery with determinant 1. This test matrix was also used in [? ? ?]. The eigenvalues of a Frank matrix are positive and occur in reciprocal pairs, half of which are ill-conditioned. The averaged CPU time for any value of p is shown in Fingure 3.10. As one can see in Figure 3.10, the required CPU time of Algorithm 3.3 is generally less than other algorithms, especially, PSN.



(a)

(b)

(c)

(d)

图 3.10: The CPU time (in seconds) required by using the four Algorithms to compute the principal pth root of 15×15 Frank matrix for $p = 10 : 300$.

TEST 4. Consider the following two matrices which take from [**?**] and [**?**], respectively.

$$S_1 = \begin{bmatrix} 0.44 & -0.88 & -0.38 & -0.50 \\ 0.68 & 2.15 & 0.48 & 0.11 \\ 0.61 & 0.77 & 2.14 & 1.04 \\ -0.16 & -0.30 & -0.67 & 1.33 \end{bmatrix}, \quad S_2 = \begin{bmatrix} -1 & -2 & 2 \\ -4 & -6 & 6 \\ -4 & -16 & 13 \end{bmatrix}.$$

Let $A_1 = S_1^5$ and $A_2 = S_2^{15}$. The eigenvalues of A_1 are $15.2477, 0.2724 \pm 16.0066\,\mathrm{i}, 1.1030$ and of A_2 are $1, 2, 3$. We now compute $A_1^{1/5}$ and $A_2^{1/15}$ using the four algorithms. The computational results are given in Table 3.6. The relative error is computed by $\|X - S\|/\|S\|$ and the relative residuals is computed by using (3.82). As is shown in Table 3.6, Algorithm 3.3 has a less pronounced advantage for computing $A_1^{1/5}$ and $A_2^{1/15}$.

	Algorithm	CPU time (s)	iter	k_1	$\rho(X)$	err(X)
$A_1^{1/5}$						
	PSN	3.91e-03	5	2	1.75e-15	2.05e-15
	SN	3.28e-03	6	2	4.67e-16	1.05e-15
	SH	3.13e-03	3	2	6.15e-16	9.63e-16
	SE	3.59e-03	3	3	8.77e-16	1.55e-15
$A_2^{1/15}$						
	PSN	6.09e-03	5	5	1.48e-15	2.67e-08
	SN	5.17e-03	6	4	5.74e-15	2.67e-08
	SH	5.00e-03	3	4	7.82e-15	2.67e-08
	SE	4.53e-03	5	5	2.25e-14	2.67e-08

表 3.6: Results for computing $A_1^{1/5}$ and $A_2^{1/15}$ by using the four Algorithms.

TEST 5. Consider a random nonnormal 8 × 8 matrix constructed as $A = QRQ^{\mathrm{T}}$, where Q is a random orthogonal matrix and R is in the Schur form with eigenvalues $\alpha_j \pm \mathrm{i}\beta_j, \alpha_j = -j^2/10, \beta_j = -j, j = 1 : n/2$ and elements $(2j, 2j+1)$ equal to -450. This example was used in [?] and [?] to compare the behavior of Algorithms PSN, SN and SH. We use the four algorithms to compute the pth of this random matrix and list in Table 3.7 the results in terms of CPU time, number of iterations, relative residual and relative error. Algorithm 3.3 shows a good accuracy and requires less computational time.

TEST 6. We select the classical test matrices (prolate matrix and Frank matrix) with no nonpositive real eigenvalues from the MATLAB gallery function together with Hilbert matrix, and then compute their principal pth root for $p = 18, 33, 81$ by using the four algorithms. These test matrices also were used in [? ? ?]. Note that, the prolate matrix is a symmetric ill-conditioned Toeplitz matrix whose elements are $A_{ii} = 1/2, A_{ij} = \sin(\pi(j-i)/2)/(\pi(j-i))$. The Hilbert matrix is a notable example of an ill-conditioned matrix. The elements

	Algorithm	CPU time (s)	iter	k_1	$\rho(X)$	err(X)
$p = 5$						
	PSN	2.34e-03	5	3	8.68e-15	3.11e-12
	SN	2.97e-03	5	2	8.62e-15	3.08e-12
	SH	3.13e-03	3	2	8.54e-15	3.06e-12
	SE	1.41e-03	3	2	8.52e-15	3.05e-12
$p = 7$						
	PSN	5.00e-03	5	3	1.45e-14	3.83e-12
	SN	4.22e-03	4	2	1.50e-14	3.96e-12
	SH	4.38e-03	3	2	1.55e-14	4.07e-12
	SE	4.06e-03	4	2	1.47e-14	3.87e-12

表 3.7: Results for a random nonnormal matrix by using the four Algorithms.

of the Hilbert matrix are $H_{ij} = 1/(i + j - 1)$. The results of the comparison are summarized in Table 3.8. As we can see that, in most cases, Algorithm 3.3 has less computational time and smaller errors than others algorithms. It also confirms that Euler's method is a very good choice for computing pth roots of a matrix.

	$p = 18$					$p = 33$					$p = 81$				
	CPU time	iter	k_1	$\rho(X)$	err(X)	CPU time	iter	k_1	$\rho(X)$	err(X)	CPU time	iter	k_1	$\rho(X)$	err(X
Hilbert matrix (7×7)															
PSN	1.06e-02	5	5	1.31e-14	6.47e-14	1.06e-02	5	5	3.77e-14	2.12e-13	1.05e-02	5	5	6.77e-14	4.30e-
SN	8.44e-03	6	4	6.09e-15	3.01e-14	8.59e-03	6	4	1.65e-14	9.31e-14	8.59e-02	6	4	2.63e-14	1.67e-
SH	8.59e-03	4	4	1.34e-15	6.60e-15	8.59e-03	4	4	2.43e-14	1.37e-13	8.59e-03	4	4	2.51e-14	1.60e-
SE	7.03e-03	5	3	3.19e-15	1.57e-14	8.44e-03	4	4	4.53e-15	2.55e-14	8.44e-03	4	4	2.45e-14	1.56e-
Prolate matrix (10×10)															
PSN	2.16e-02	5	5	3.96e-15	3.45e-14	2.03e-02	5	5	1.33e-14	1.03e-13	2.31e-02	4	5	3.79e-14	3.64e-
SN	1.81e-02	6	4	1.29e-15	1.12e-14	1.81e-02	6	4	5.74e-15	5.26e-14	1.77e-02	6	4	9.05e-15	8.70e-
SH	1.72e-02	4	4	1.84e-15	1.61e-14	1.59e-02	4	4	1.03e-14	9.47e-14	1.81e-02	4	4	2.55e-14	2.45e-
SE	1.23e-02	4	3	1.29e-15	1.13e-14	1.33e-02	4	3	3.41e-15	3.13e-14	1.25e-02	4	3	1.08e-14	1.04e-
Frank matrix (12×12)															
PSN	2.31e-02	5	4	1.46e-15	1.54e-08	2.39e-02	5	4	7.65e-15	2.51e-08	2.44e-02	5	4	1.17e-13	6.58e-
SN	1.69e-02	6	3	1.46e-15	1.55e-08	1.67e-02	6	3	5.06e-15	1.66e-08	1.70e-02	6	3	1.15e-13	6.45e-
SH	1.77e-02	4	3	1.52e-15	1.61e-08	1.67e-02	4	3	8.09e-15	2.65e-08	1.80e-02	4	3	8.24e-14	4.73e-
SE	1.63e-02	4	3	1.17e-15	1.24e-08	1.80e-02	4	3	7.45e-15	2.44e-08	1.78e-02	4	3	1.07e-13	6.00e-

表 3.8: Results for computing principal pth root of 7×7 Hilbert matrix, 10×10 Prolate matrix and 12×12 Frank matrix by using the four Algorithms.

3.10　关于 Halley 法的注记

While the iterative form of Halley's method for computing $A^{1/p}$ is given by:

$$X_{k+1} = X_k\left[(p+1)X_k^p + (p-1)A\right]^{-1}\left[(p-1)X_k^p + (p+1)A\right], \quad k = 0, 1, 2, \dots, \tag{3.84}$$

provided X_0 commutes with A and that $(p+1)X_k^p + (p-1)A$ is nonsingular for any $k \geq 0$.

令

$$\phi_\nu(z) := 1 - u^p(z)(1-z), \quad z \in \mathcal{D}_{0,\nu},\ \nu = 1, 2, \tag{3.85}$$

其中

$$u_2(z) := \frac{1 - \frac{p-1}{2p}z}{1 - \frac{p+1}{2p}z}, z \in \mathcal{D}_{0,2},$$

$$\mathcal{D}_{0,2} := \left\{z \in \mathbb{C} : |z| < \frac{2p}{p+1}\right\}. \tag{3.86}$$

Define

$$\mathcal{R}_\nu := \left\{z \in \mathbb{C} : 1 - z \in \overline{\mathcal{D}}_1 \bigcup \left(\mathcal{D}_{0,\nu} \bigcap \mathcal{D}_{2,\nu}\right)\right\} \tag{3.87}$$

and

$$\widehat{\mathcal{R}}_\nu := \left\{z \in \mathbb{C} : 1 - z \in \mathcal{D}_1 \bigcup \left(\mathcal{D}_{0,\nu} \bigcap \mathcal{D}_{2,\nu}\right)\right\}, \quad \nu = 1, 2, \tag{3.88}$$

where $\mathcal{D}_{0,\nu}$ are defined in (3.86), $\overline{\mathcal{D}}_1$ is the closure of $\mathcal{D}_1$ defined by

$$\mathcal{D}_1 := \{z \in \mathbb{C} : |z| < 1\}, \tag{3.89}$$

and

$$\begin{cases} \mathcal{D}_{2,1} := \left\{z \in \mathbb{C} : \sup\limits_{m \geq 2}\left\{\dfrac{|S_{1,m}(z)|}{|z|}\right\} \cdot \dfrac{|z| + |\phi_1(z)|}{||z| - |\phi_1(z)||} < 1\right\}, \\ \mathcal{D}_{2,2} := \left\{z \in \mathbb{C} : \sup\limits_{m \geq 3}\left\{\dfrac{|S_{2,m}(z)|}{|z|}\right\} \cdot \dfrac{|z| + |\phi_2(z)|}{||z| - |\phi_2(z)||} < 1\right\}, \end{cases} \tag{3.90}$$

where

$$\begin{cases} S_{1,m}(z) = \sum\limits_{j=2}^{m} c_{1,j} z^j, \quad z \in \mathcal{D}_{0,1}, \\ S_{2,m}(z) = \sum\limits_{j=3}^{m} c_{2,j} z^j, \quad z \in \mathcal{D}_{0,2}, \end{cases} \tag{3.91}$$

$c_{\nu,j} = \phi_\nu^{(j)}(0)/j!$ and ϕ_ν are defined in (3.85), $\nu = 1, 2$. As for $\phi_2(z)$, the similar result has been proved in [**?**].

Similar to Newton's method (2.12), we can obtain the following theorem on convergence and convergence order for Halley's method (3.84).

定理 3.2. *If all eigenvalues of $A \in \mathbb{C}^{n\times n}$ are in $\mathcal{R}_2$ defined by (3.87) and all zero eigenvalues of A are semisimple, then the matrix sequence $\{X_k\}$ generated by Halley's method (3.84) starting from $X_0 = \mathrm{I}$ converges to the principal pth root $A^{1/p}$. Moreover, if all eigenvalues are in $\widehat{\mathcal{R}}_2\backslash\{0\}$ defined by (3.88), then the convergence is cubic.*

注 3.5. Similar to Newton's method, in Figure 3.11, we plot the approximating region $\mathcal{R}_2$ defined in (3.87) with nine cases in the complex plane for Halley's method, where the red and blue regions denote the sets $\{z \in \mathbb{C} : 1 - z \in \overline{\mathcal{D}}_1\}$ and $\{z \in \mathbb{C} : 1 - z \in \mathcal{D}_{0,2} \bigcap \mathcal{D}_{2,2}\}$, respectively. We can see from Figure 3.11 that, for a fixed p, approximating regions (cf. (a)-(c) for $p = 25$, (d)-(f) for $p = 100$ and (g)-(i) for $p = 400$) are almost the same when $m = 20, 100, 500$, respectively. To this end, it suffices to choose $m = 20$ in practical numerical computation. Furthermore, we find that, for a fixed $m(\geq 20)$, the approximating regions $\mathcal{R}_2$ with various p are also almost the same.

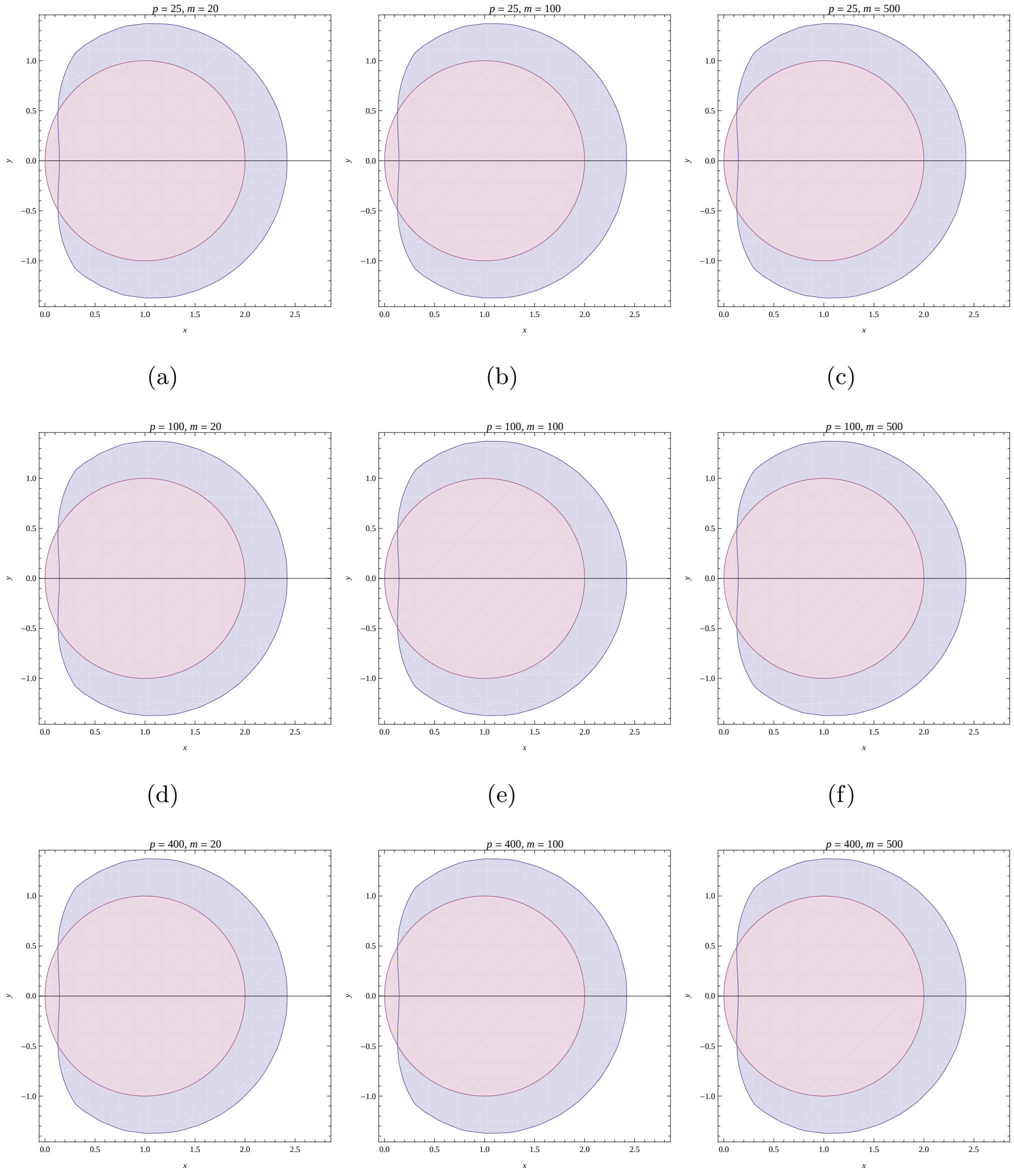
注 3.6. It is also inconvenient, similar to the case of $\mathcal{R}_1$ defined in (3.87) for Newton's method, to check whether a eigenvalue λ belongs to $\mathcal{R}_2$ defined in (3.87) in practice. We can also define a feasible region which is acceptable approximation to $\mathcal{R}_2$ and allows us to determine easily whether a eigenvalue belongs to it. Define

$$\mathcal{R}_2^{\mathrm{H}} = \mathcal{D}_3 \bigcup \mathcal{D}_5, \tag{3.92}$$

where $\mathcal{D}_3$ is defined in (2.55) and

$$\mathcal{D}_5 := \left\{ z \in \mathbb{C} : |\arg(z)| < \frac{\pi}{3}, |1 - z| < \frac{7}{5} \right\}.$$

The actual convergence region $\mathcal{R}_2$ and the new feasible region $\mathcal{R}_2^{\mathrm{H}}$ are depicted in Figure 3.12 with $p = 7, 20, 100$ and fixed $m = 20$, where the yellow parts denote the region $\mathcal{R}_2^{\mathrm{H}}$.



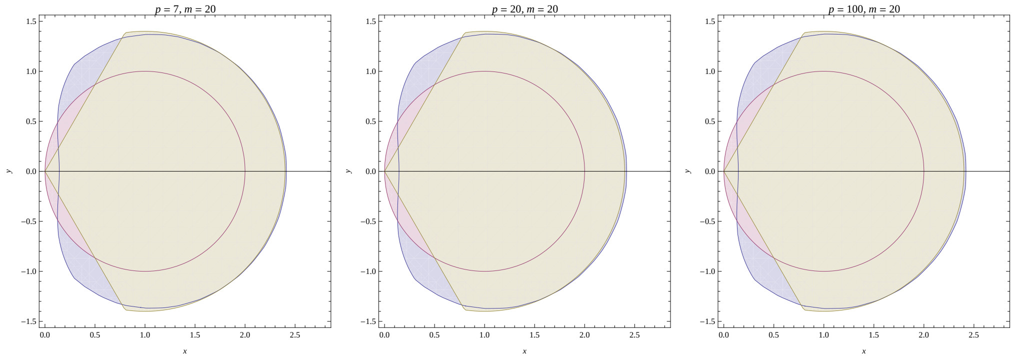
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

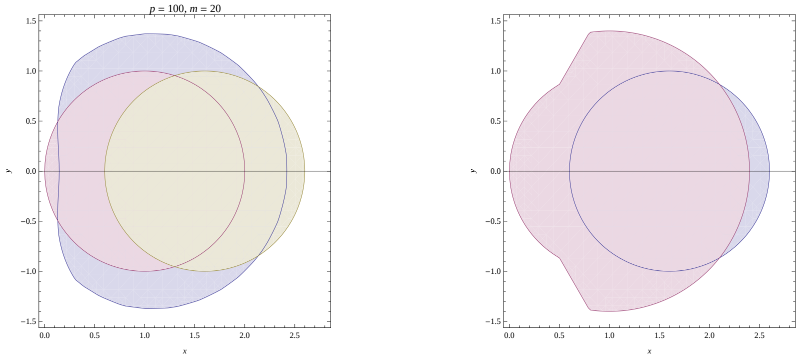
图 3.11: The approximating regions of $\mathcal{R}_2$ defined in (3.87) for $p = 25, 100, 400$ and $m = 20, 100, 500$, where the red and blue regions denote the sets $\{z \in \mathbb{C} : 1 - z \in \overline{\mathcal{D}}_1\}$ and $\{z \in \mathbb{C} : 1 - z \in \mathcal{D}_{0,2} \bigcap \mathcal{D}_{2,2}\}$, respectively.

注 3.7. It is worth noting that, the convergence regions $\mathcal{R}_2$ defined in (3.87) for Halley's method (3.84) is more useful than the one given by Iannazzo in [**?**



(a) (b) (c)

图 3.12: For fixed $m = 20$ and $p = 7, 20, 100$, the actual convergence regions $\mathcal{R}_2$ defined in (3.87) (the union of the red and blue parts) and the approximate convergence regions $\mathcal{R}_2^{\mathrm{H}}$ defined in (3.92) (the yellow parts).



(a) (b)

图 3.13: (a) The convergence region $\mathcal{R}_2$ defined in (3.87) with $p = 100$ and $m = 20$ (the union of the red and blue parts) and the the disk of center 8/5 and radius 1 given by Iannazzo (the yellow parts); (b) the feasible region $\mathcal{R}_2^{\mathrm{H}}$ defined in (3.92) (the red part) and the disk of center 8/5 and radius 1 (the blue part).

, Algorithm 4] in which the choice of the region (the disk of center 8/5 and radius 1) is heuristic and is based on the observation the experimental regions of convergence. The comparison of the regions $\mathcal{R}_2$, $\mathcal{R}_2^{\mathrm{H}}$ and the disk of center 8/5 and radius 1 is shown in Figure 3.13.

第 4 章　计算代数 Riccati 方程的 ULM 方法

第 5 章　计算代数 Riccati 方程的修正 Newton 方法

攻读学位期间取得的研究成果

[1] An analysis on efficiency of Euler's method for computing the matrix pth root, 2013, in preparation.

[2] A note on Newton's method and Halley's method for computing the matrix pth root, 2013, in preparation.

[3] On the semilocal convergence behavior for Halley's method, Comput. Optim. Appl., 2014.

[4] A globally convergent inexact Newton-like Cayley transform method for inverse eigenvalue problems, J. Appl. Math., 2013.

[5] Convergence behavior for Newton-Steffensen's method under gamma-condition of second derivative, Abstr. Appl. Anal., 2013.

致　　谢

本文是在黄正达教授的悉心指导下完成的. 非常感谢黄老师给予我自由选择研究课题的余地, 使我能够有机会在一个非常有意思的领域内做一些研究工作. 黄老师渊博的知识、严谨的治学态度使我受益匪浅, 同时在思想、生活上对我也很关心和照顾, 值此毕业之际, 谨向他致以崇高的敬意和诚挚的谢意.

借此机会, 我还要感谢我的同门师兄姐弟妹们, 三年来在讨论班及实验室里的交流使我受益匪浅, 他们是：孔祥银、高芹、孔维镇、周小燕、王会迪、郑璇、陈敏红、黄敏、郭明、王湘美、王玉芳、郑聪、赵晓芃、潜陈印、张燕、张勇.

感谢所有关心我、支持我、帮助我的亲人和朋友们!

浙江大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果．论文中除了特别加以标注和致谢的地方外，不包含其他人或其他机构已经发表或撰写过的研究成果．其他同志对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了谢意．本人完全意识到本声明的法律结果由本人承担.

作者签名：　　　　　　　　　　　　日期：　　　　年　　月　　日

学位论文使用授权声明

本人完全了解浙江大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关机关或机构送交论文的复印件和电子文档，允许论文被查阅和借阅，可以采用影印、缩印或扫描等手段保存、汇编学位论文．同意浙江大学可以用不同方式在不同媒体上发表、传播论文的全部或部分内容.

保密的学位论文在解密后遵守此协议.

作者签名：　　　　　　　导师签名：　　　　　　　日期：　　　　年　　月　　日