数值代数— 绪论

邓建松

2018 年 9 月 11 日

数据处理问题

- 现代的科学技术发展十分迅速，它们有一个共同的特点，就是都有大量的数据问题

数据处理问题

- 现代的科学技术发展十分迅速，它们有一个共同的特点，就是都有大量的数据问题
  1. 发射一颗卫星，从卫星设计开始到发射、回收为止，科学家和工程技术人员、工人就要对卫星的总体部件进行全面的设计和生产，要对选用的火箭进行设计和生产

数据处理问题

- 现代的科学技术发展十分迅速，它们有一个共同的特点，就是都有大量的数据问题
  1. 发射一颗卫星，从卫星设计开始到发射、回收为止，科学家和工程技术人员、工人就要对卫星的总体部件进行全面的设计和生产，要对选用的火箭进行设计和生产
  2. 在高能加速器里进行高能物理试验，研究具有很高能量的基本粒子的性质、它们之间的相互作用和转化规律，这里面也有大量的数据计算问题

科学活动的新手段

- 1946年第一台电子计算机诞生

科学活动的新手段

- 1946年第一台电子计算机诞生
- 理论研究、科学试验、科学计算为当今世界科学活动的三种主要方式

科学活动的新手段

- 1946年第一台电子计算机诞生
- 理论研究、科学试验、科学计算为当今世界科学活动的三种主要方式
- 为科学与工程问题提供计算方法，提高计算的可靠性、有效性和精确性，是科学与工程计算领域的主要研究内容

科学活动的新手段

- 1946年第一台电子计算机诞生
- 理论研究、科学试验、科学计算为当今世界科学活动的三种主要方式
- 为科学与工程问题提供计算方法，提高计算的可靠性、有效性和精确性，是科学与工程计算领域的主要研究内容
- 研究计算问题的解决方法和有关数学理论问题的一门学科就叫做计算数学

科学活动的新手段

- 1946年第一台电子计算机诞生
- 理论研究、科学试验、科学计算为当今世界科学活动的三种主要方式
- 为科学与工程问题提供计算方法，提高计算的可靠性、有效性和精确性，是科学与工程计算领域的主要研究内容
- 研究计算问题的解决方法和有关数学理论问题的一门学科就叫做计算数学
- 计算数学属于应用数学的范畴

数值线性代数

- 简称为数值代数，也称为矩阵计算

数值线性代数

- 简称为数值代数，也称为矩阵计算
- 它是科学与工程计算的核心

数值线性代数

- 简称为数值代数，也称为矩阵计算
- 它是科学与工程计算的核心
  - 大部分科学与工程计算问题最终都归结为一个矩阵计算问题，特别是大规模矩阵计算问题

数值线性代数

- 简称为数值代数，也称为矩阵计算
- 它是科学与工程计算的核心
    - 大部分科学与工程计算问题最终都归结为一个矩阵计算问题，特别是大规模矩阵计算问题
- 数值代数的研究内容就是针对各类科学与工程问题的特点，设计出相应的快速可靠的算法

“计算数学”专业课程

- 本科阶段

“计算数学”专业课程

- 本科阶段
  - 数值代数、数值分析、偏微分方程数值解、有限元、计算机图形学

“计算数学”专业课程

- 本科阶段
  - 数值代数、数值分析、偏微分方程数值解、有限元、计算机图形学
- 研究生阶段

“计算数学”专业课程

- 本科阶段
  - 数值代数、数值分析、偏微分方程数值解、有限元、计算机图形学
- 研究生阶段
  - 计算流体力学、高级有限元、并行计算及算法、计算机辅助几何设计、样条函数与逼近论、多变量函数逼近论、计算代数几何、小波分析、高级几何建模与图形学

基本问题

- 求解线性方程组：给定n阶非奇异矩阵A和n维向量b, 求解方程$Ax = b$, 其中x是未知的n维向量

基本问题

- 求解线性方程组：给定n阶非奇异矩阵A和n维向量b, 求解方程$Ax = b$, 其中x是未知的n维向量
- 线性最小二乘问题：给出$m \times n$阶矩阵A和m维向量b, 求n维向量x使得

$$\|Ax - b\|_2 = \min\{\|Ay - b\|_2 : y \in \mathbb{R}^n\}$$

基本问题

- 求解线性方程组：给定n阶非奇异矩阵A和n维向量b, 求解方程$Ax = b$, 其中x是未知的n维向量
- 线性最小二乘问题：给出$m \times n$阶矩阵A和m维向量b, 求n维向量x使得

$$\|Ax - b\|_2 = \min\{\|Ay - b\|_2 : y \in \mathbb{R}^n\}$$

- 矩阵特征值问题：计算给定方阵A的部分或全部特征值与对应的特征向量

基本问题的变体

- 约束最小二乘问题、完全最小二乘问题、矩阵方程的求解、矩阵函数的计算、广义特征值问题、非线性特征值问题、特征值反问题、奇异值分解的计算

基本问题的变体

- 约束最小二乘问题、完全最小二乘问题、矩阵方程的求解、矩阵函数的计算、广义特征值问题、非线性特征值问题、特征值反问题、奇异值分解的计算
- 特别地，奇异值分解的计算有广泛的应用，也有人称其为数值代数的第四大问题

Cramer法则

上述问题的数学理论已相当完善，但是理论上漂亮的结果在实际计算时有可能不可行：先以Cramer法则为例

- Cramer法则：把线性方程组的求解归结为计算$n+1$个n阶行列式

Cramer法则

上述问题的数学理论已相当完善，但是理论上漂亮的结果在实际计算时有可能不可行：先以Cramer法则为例

- Cramer法则：把线性方程组的求解归结为计算$n+1$个n阶行列式
- Laplace展开定理：计算行列式的理论公式—n阶行列式的乘法次数$> n!$

Cramer法则

- 采用Cramer法则求解线性方程组的乘法运算次数大于$(n+1)!$. 例如，求解一个25阶线性方程组，采用此方法，需要约13亿年

Cramer法则

- 采用Cramer法则求解线性方程组的乘法运算次数大于$(n+1)!$. 例如，求解一个25阶线性方程组，采用此方法，需要约13亿年
- 而采用Gauss消元法，则可以不超过1秒钟完成求解

Jordan分解

- 基于Jordan标准型与分解定理，可以清楚地知道矩阵所有的特征值及有关信息：几何重数、代数重数、对应的特征向量等

Jordan分解

- 基于Jordan标准型与分解定理，可以清楚地知道矩阵所有的特征值及有关信息：几何重数、代数重数、对应的特征向量等
- 而Jordan分解是非常不稳定的，变换矩阵常常是非常病态的

Jordan分解

- 基于Jordan标准型与分解定理，可以清楚地知道矩阵所有的特征值及有关信息：几何重数、代数重数、对应的特征向量等
- 而Jordan分解是非常不稳定的，变换矩阵常常是非常病态的
- 实际计算采用的通常是具有良好数值性态的Schur分解

现状与问题

- 数值代数只有半个多世纪的历史

现状与问题

- 数值代数只有半个多世纪的历史
- 相关的方法和理论已发展得相对成熟

现状与问题

- 数值代数只有半个多世纪的历史
- 相关的方法和理论已发展得相对成熟
- 大规模矩阵计算问题仍是目前的研究核心问题之一

矩阵分解

- 如何根据矩阵计算问题的特点，设计出有效的计算方法，这是我们关注的首要问题

矩阵分解

- 如何根据矩阵计算问题的特点，设计出有效的计算方法，这是我们关注的首要问题
- 基本想法就是将一般的问题转化为一个或几个易于求解的特殊问题

矩阵分解

- 如何根据矩阵计算问题的特点，设计出有效的计算方法，这是我们关注的首要问题
- 基本想法就是将一般的问题转化为一个或几个易于求解的特殊问题
- 完成这一转化的基本技巧就是矩阵分解

例

- 对于线性方程组求解问题：当系数矩阵A为下三角或者上三角时，方程组的求解变得非常容易

例

- 对于线性方程组求解问题：当系数矩阵A为下三角或者上三角时，方程组的求解变得非常容易
- A为一般矩阵时，可以首先将A分解为$PA = LU$，其中P为排列方阵，U和L分别是上、下三角阵

例

- 对于线性方程组求解问题：当系数矩阵A为下三角或者上三角时，方程组的求解变得非常容易
- A为一般矩阵时，可以首先将A分解为$PA = LU$，其中P为排列方阵，U和L分别是上、下三角阵
- 通过求解$Ly = Pb$, $Ux = y$就可以得到原方程的解

误差

误差来源于

- 原始数据的误差

误差

误差来源于

- 原始数据的误差
- 计算过程产生的误差

误差

误差来源于

- 原始数据的误差
- 计算过程产生的误差

- 误差是不可避免的

误差

误差来源于

- 原始数据的误差
- 计算过程产生的误差

- 误差是不可避免的
- 计算解与真解的差是多少？

敏度分解

研究原始数据的微小变化会引起解的多大变化。假设考虑函数$f(x)$的计算问题

- 中值定理与Taylor展开：可行性不大

敏度分解

研究原始数据的微小变化会引起解的多大变化。假设考虑函数$f(x)$的计算问题

- 中值定理与Taylor展开：可行性不大
- 当$\delta x/|x|$很小时，确定一个尽可能小的正数$c(x)$, 满足

$$\frac{|f(x+\delta x)-f(x)|}{|f(x)|} \leqslant c(x)\frac{|\delta x|}{|x|}$$

条件数

$c(x)$的大小在一定程度上反映了自变量的微小变化对函数值的影响程度

- $c(x)$称为f在x点的条件数

条件数

$c(x)$的大小在一定程度上反映了自变量的微小变化对函数值的影响程度

- $c(x)$称为f在x点的条件数
- $c(x)$很大时，自变量的小变化有可能引起函数值的大变化，因此称f在x点是病态的；
  当$c(x)$很小时，称f在x点是良态的

条件数

$c(x)$的大小在一定程度上反映了自变量的微小变化对函数值的影响程度

- $c(x)$称为f在x点的条件数
- $c(x)$很大时，自变量的小变化有可能引起函数值的大变化，因此称f在x点是病态的；当$c(x)$很小时，称f在x点是良态的
- 计算问题是否病态是问题自身的固有属性，与计算方法无关

舍入误差

- 计算机是有效精度的。这引起的误差称为舍入误差。

舍入误差

- 计算机是有效精度的。这引起的误差称为舍入误差。
- 分析舍入误差对算法结果的影响，是衡量算法优劣的重要标志

舍入误差

- 计算机是有效精度的。这引起的误差称为舍入误差。
- 分析舍入误差对算法结果的影响，是衡量算法优劣的重要标志
- 仍以函数求值为例

向后误差分析法

- 设采用某算法后函数f在x点的值为$\hat{y}$

向后误差分析法

- 设采用某算法后函数f在x点的值为$\hat{y}$
- 通过对每步具体运算的误差分析可以证明：存在δx满足$\hat{y} = f(x + \delta x)$, $|\delta x| \leqslant |x|\varepsilon$, 其中$\varepsilon$是一个与计算机精度和算法有关的正数

向后误差分析法

- 设采用某算法后函数f在x点的值为$\hat{y}$
- 通过对每步具体运算的误差分析可以证明：存在δx满足$\hat{y} = f(x + \delta x)$, $|\delta x| \leqslant |x|\varepsilon$, 其中$\varepsilon$是一个与计算机精度和算法有关的正数
- 这种把计算结果归结为原始数据经扰动后的精确结果的误差分析方法称为向后误差分析法

数值稳定性

- ε越小，说明舍入误差对算法的影响越小，因此称算法为数值稳定的，否则称为数值不稳定的

数值稳定性

- ε越小，说明舍入误差对算法的影响越小，因此称算法为数值稳定的，否则称为数值不稳定的
- 算法的数值稳定性是算法本身的固有属性，与计算问题是否病态无关

精度估计

- 综合前述敏度分析和误差分析之后，计算结果的精度估计如下：

$$\frac{|\hat{y} - f(x)|}{|f(x)|} = \frac{|f(x+\delta x) - f(x)|}{|f(x)|} \leqslant c(x)\varepsilon$$

精度估计

- 综合前述敏度分析和误差分析之后，计算结果的精度估计如下：

$$\frac{|\hat{y}-f(x)|}{|f(x)|}=\frac{|f(x+\delta x)-f(x)|}{|f(x)|}\leqslant c(x)\varepsilon$$

- 计算结果是否可靠，依赖于计算问题是否病态以及所用的算法是否数值稳定

算法的种类

- 直接法：在没有误差的情况下可在有限步得到计算问题精确解

算法的种类

- 直接法：在没有误差的情况下可在有限步得到计算问题精确解
    - 运算量的大小可以作为其速度的一个主要标志

算法的种类

- 直接法：在没有误差的情况下可在有限步得到计算问题精确解
  - 运算量的大小可以作为其速度的一个主要标志
- 迭代法：采用逐次逼近的方法逼近问题的精确解，而在任意有限步都不能得到其精确解

算法的种类

- 直接法：在没有误差的情况下可在有限步得到计算问题精确解
  - 运算量的大小可以作为其速度的一个主要标志
- 迭代法：采用逐次逼近的方法逼近问题的精确解，而在任意有限步都不能得到其精确解
  - 除估计每步运算量的大小，还需要对收敛性进行分析

算法复杂性

计算或者估计算法的运算量

- 上世纪90年代之前：通常只计算乘除运算的次数

算法复杂性

计算或者估计算法的运算量

- 上世纪90年代之前：通常只计算乘除运算的次数
- 进入90年代之后：算法所有运算次数总和

算法复杂性

计算或者估计算法的运算量

- 上世纪90年代之前：通常只计算乘除运算的次数
- 进入90年代之后：算法所有运算次数总和
- 如果某一算法的运算量是关于n的多项式，通常就略去低阶项，用最高阶项称为算法的运算量

运算量与算法快慢

- 运算量只在一定程度上反映了算法的速度

运算量与算法快慢

- 运算量只在一定程度上反映了算法的速度
- 现代计算机中运算速度远远高于数据的传输速度，因此算法的速度在很大程度上依赖于算法实现后数据传输量的大小

收敛速度

针对迭代法

- 假设某一迭代法产生的序列$\{x_k\}$满足

$$\|x_k - x\| \leqslant c\|x_{k-1} - x\|^{\alpha}, \quad k = 1, 2, \ldots$$

收敛速度

针对迭代法

- 假设某一迭代法产生的序列$\{x_k\}$满足

$$\|x_k - x\| \leqslant c\|x_{k-1} - x\|^{\alpha}, \quad k = 1, 2, \ldots$$

- 若$0 < c < 1$, $\alpha = 1$, 则称算法线性收敛

收敛速度

针对迭代法

- 假设某一迭代法产生的序列$\{x_k\}$满足

$$\|x_k - x\| \leqslant c\|x_{k-1} - x\|^{\alpha}, \quad k = 1, 2, \ldots$$

  - 若$0 < c < 1$, $\alpha = 1$, 则称算法线性收敛
  - 若$c > 0$, $\alpha = 2$, 则称算法平方收敛（二次收敛）。此时c越小越好

教材与参考书

- 教材：数值线性代数（第二版），徐树方等编著，北京大学出版社

教材与参考书

- 教材：数值线性代数（第二版），徐树方等编著，北京大学出版社
- 参考书

教材与参考书

- 教材：数值线性代数（第二版），徐树方等编著，北京大学出版社
- 参考书
  - 矩阵计算（第四版），G. H. Golub著，人民邮电出版社影印，2014年

教材与参考书

- 教材：数值线性代数（第二版），徐树方等编著，北京大学出版社
- 参考书
  - 矩阵计算（第四版），G. H. Golub著，人民邮电出版社影印，2014年
  - 代数特征值问题，J. H. Wilkinson著，科学出版社，2018年再版

考核方式

- 课后习题：20% (注意网络上的解答有些是错的)

考核方式

- 课后习题：20% (注意网络上的解答有些是错的)
- 编程作业：35% (课本上机习题+SVD分解)

考核方式

- 课后习题：20% (注意网络上的解答有些是错的)
- 编程作业：35% (课本上机习题+SVD分解)
- 期终考试：50% (闭卷?，课后习题和书上定理证明占60–70%)

考核方式

- 课后习题：20% (注意网络上的解答有些是错的)
- 编程作业：35% (课本上机习题+SVD分解)
- 期终考试：50% (闭卷?，课后习题和书上定理证明占60–70%)
- 所有作业在布置后的周三上交。特殊情况提前向助教说明，一周内补交有效！

考核方式

- 课后习题：20% (注意网络上的解答有些是错的)
- 编程作业：35% (课本上机习题+SVD分解)
- 期终考试：50% (闭卷?，课后习题和书上定理证明占60–70%)
- 所有作业在布置后的周三上交。特殊情况提前向助教说明，一周内补交有效!
- 发现照抄，双方均不得分!

考核方式

- 课后习题：20% (注意网络上的解答有些是错的)
- 编程作业：35% (课本上机习题+SVD分解)
- 期终考试：50% (闭卷?，课后习题和书上定理证明占60–70%)
- 所有作业在布置后的周三上交。特殊情况提前向助教说明，一周内补交有效!
- 发现照抄，双方均不得分!
- 以上各项可讨论，下周三确定!

编程作业的要求

- C++

编程作业的要求

- C++
- Python 3.6

编程作业的要求

- C++
- Python 3.6
- 具体编译平台，请在大群中与助教讨论确认。各语言最好只接受一种平台

编程作业的要求

- C++
- Python 3.6
- 具体编译平台，请在大群中与助教讨论确认。各语言最好只接受一种平台
- 课堂演示：C++

线性方程组的直接解法

邓建松

2018 年 9 月 19 日

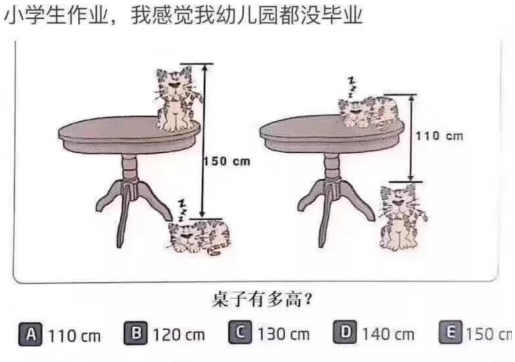
线性方程组

- 结构分析、网络分析、大地测量、数据分析、最优化及非线性方程组和微分方程组数值解等，都常常遇到线性方程的求解问题

线性方程组

- 结构分析、网络分析、大地测量、数据分析、最优化及非线性方程组和微分方程组数值解等，都常常遇到线性方程的求解问题
- 这也是一个历史悠久的问题：《九章算术》中就记载有消元法

  今有上禾三秉，中禾二秉，下禾一秉，实三十九斗；上禾二秉，中禾三秉，下禾一秉，实三十四斗；上禾一秉，中禾二秉，下禾三秉，实二十六斗。问上、中、下禾实一秉各几何？


小学生作业，我感觉我幼儿园都没毕业
150 cm
110 cm
桌子有多高？
A 110 cm
B 120 cm
C 130 cm
D 140 cm
E 150 cm

求解方法

- 求解大型线性方程组是计算机问世后才有可能的事情

求解方法

- 求解大型线性方程组是计算机问世后才有可能的事情
- 直接法（精确法）：在没有舍入误差的情况下经过有限次运算可求得方程组的精确解

求解方法

- 求解大型线性方程组是计算机问世后才有可能的事情
- 直接法（精确法）：在没有舍入误差的情况下经过有限次运算可求得方程组的精确解
- 迭代法：采取逐次逼近方法，从一个初始向量出发，按照一定的计算格式，得到一个向量的无穷序列，其极限是方程组的精确解。只经过有限次运算得不到精确解

Gauss消去法

- Gauss消去法是一类最基本的直接求解方法

Gauss消去法

- Gauss消去法是一类最基本的直接求解方法
- 它是目前求解中小规模（阶数一般不超过1000）线性方程组的最常用方法

Gauss消去法

- Gauss消去法是一类最基本的直接求解方法
- 它是目前求解中小规模（阶数一般不超过1000）线性方程组的最常用方法
- 用于系数矩阵没有任何特殊结构的方程组

下三角形方程组

考虑下三角形方程组$Ly=b$, 其中$b=(b_1,\dots,b_n)^T$已知，$y=(y_1,\dots,y_n)^T$未知，而系数阵L是已知的非奇异下三角阵，即

$$L=\begin{pmatrix} \ell_{11} & & & \\ \ell_{21} & \ell_{22} & & \\ \vdots & \vdots & \ddots & \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nn} \end{pmatrix}$$

求解方法

矩阵非奇异，等价于$\ell_{ii} \neq 0, i = 1, \ldots, n$

- $y_1 = b_1/\ell_{11}$

求解方法

矩阵非奇异，等价于$\ell_{ii} \neq 0$, $i = 1, \ldots, n$

- $y_1 = b_1/\ell_{11}$
- $y_2 = (b_2 - \ell_{21}y_1)/\ell_{22}$

求解方法

矩阵非奇异，等价于$\ell_{ii} \neq 0,\ i = 1, \ldots, n$

- $y_1 = b_1/\ell_{11}$
- $y_2 = (b_2 - \ell_{21}y_1)/\ell_{22}$
- ……

求解方法

矩阵非奇异，等价于$\ell_{ii} \neq 0$, $i = 1, \ldots, n$

- $y_1 = b_1/\ell_{11}$
- $y_2 = (b_2 - \ell_{21}y_1)/\ell_{22}$
- ……
- $y_i = \dfrac{b_i - \sum_{j=1}^{i-1} \ell_{ij}y_j}{\ell_{ii}}$

前代法

- 前述方法称为**前代法**

前代法

- 前述方法称为前代法
- 实现时为了节省存贮空间，可以把y_i放在b_i所用的存贮单元中

前代法

- 前述方法称为**前代法**
- 实现时为了节省存贮空间，可以把y_i放在b_i所用的存贮单元中
- 具体实现时可以在算出y_i时，马上从后面各项b_i中减去相应的量

前代法

- 前述方法称为前代法
- 实现时为了节省存贮空间，可以把y_i放在b_i所用的存贮单元中
- 具体实现时可以在算出y_i时，马上从后面各项b_i中减去相应的量
- 运算量：

$$\sum_{i=1}^{n}(2i-1)=n^2$$

上三角形方程组

考虑上三角形方程组$Ux=y$, 其中$y=(y_1,\ldots,y_n)^T$已知，$x=(x_1,\ldots,x_n)^T$未知，而系数阵U是已知的非奇异上三角阵，即

$$U=\begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ & & \ddots & \vdots \\ & & & u_{nn} \end{pmatrix}$$

解法：回代法

从方程组的最后一个方程出发依次求解

$$x_i = \frac{y_i - \sum_{j=i+1}^{n} u_{ij} x_j}{u_{ii}}, i = n, n-1, \ldots, 1$$

- 算法的运算量还是n^2

一般方程组

对于一般的线性方程组$Ax = b$, 如果能把A分解为$A = LU$, 那么

- 用前代法求解$Ly = b$

一般方程组

对于一般的线性方程组$Ax = b$, 如果能把A分解为$A = LU$, 那么

- 用前代法求解$Ly = b$
- 用回代法求解$Ux = y$

一般方程组

对于一般的线性方程组$Ax = b$, 如果能把A分解为$A = LU$, 那么

- 用前代法求解$Ly = b$
- 用回代法求解$Ux = y$
- 所以关键是如何进行分解

初等变换

期望通过一系列初等变换把A约化为上三角阵，而这些变换的乘积是一个下三角阵，即

- 对于一个任意给定的向量$x \in \mathbb{R}^n$, 找一个尽可能简单的下三角阵，使x经这一矩阵作用之后的第$k+1$至n个分量都是零

初等变换

期望通过一系列初等变换把A约化为上三角阵，而这些变换的乘积是一个下三角阵，即

- 对于一个任意给定的向量$x \in \mathbb{R}^n$，找一个尽可能简单的下三角阵，使x经这一矩阵作用之后的第$k+1$至n个分量都是零
- 下三角阵的乘积仍是下三角阵

Gauss变换

Gauss变换矩阵定义为

$$L_k = I - \ell_k e_k^T,$$

其中

$$\ell_k = (0, 0, \ldots, 0, \ell_{k+1,k}, \ldots, \ell_{nk})^T$$

称为Gauss向量

Gauss变换

$$L_k = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -\ell_{k+1,k} & 1 & & \\ & & \vdots & & \ddots & \\ & & -\ell_{nk} & & & 1 \end{pmatrix}$$

消去

由于

$$L_k x = (x_1, \ldots, x_k, x_{k+1} - x_k L_{k+1,k}, \ldots, x_n - x_k \ell_{nk})^T$$

所以取

$$\ell_{ik} = \frac{x_i}{x_k}, \quad i = k+1, \ldots, n$$

即有

$$L_k x = (x_1, \ldots, x_k, 0, \ldots, 0)^T$$

Gauss变换矩阵的性质

- 逆容易计算：由于$e_k^T\ell_k = 0$, 所以

$$(I - \ell_k e_k^T)(I + \ell_k e_k^T) = I,$$

从而

$$L_k^{-1} = I + \ell_k e_k^T$$

Gauss变换矩阵的性质

- 逆容易计算：由于$e_k^T \ell_k = 0$, 所以
$$(I - \ell_k e_k^T)(I + \ell_k e_k^T) = I,$$
从而
$$L_k^{-1} = I + \ell_k e_k^T$$
- Gauss变换作用于矩阵A相当于对矩阵进行秩1的修正
$$L_k A = A - \ell_k (e_k^T A)$$

三角分解的计算

`Mathematica 1.1.3.nb`

- 对一般n阶矩阵A, 在一定的条件下，可以得到$n-1$个Gauss变换$L_1, \dots, L_{n-1}$, 使得$L_{n-1} \cdots L_1 A$为上三角矩阵

三角分解的计算

`Mathematica 1.1.3.nb`

- 对一般n阶矩阵A, 在一定的条件下，可以得到$n-1$个Gauss变换$L_1, \ldots, L_{n-1}$, 使得$L_{n-1} \cdots L_1 A$为上三角矩阵
- 这里的条件就是第k步中对角元素不能是零

三角分解的计算

`Mathematica 1.1.3.nb`

- 对一般n阶矩阵A, 在一定的条件下，可以得到$n-1$个Gauss变换$L_1, \ldots, L_{n-1}$, 使得$L_{n-1} \cdots L_1 A$为上三角矩阵
- 这里的条件就是第k步中对角元素不能是零
- $L_{n-1} \cdots L_1$是一个对角元全为1的下三角矩阵

三角分解的计算

令

$$L = (L_{n-1} \cdots L_1)^{-1}, U = L_{n-1} \cdots L_1 A$$

则$A = LU$就是所期望的三角分解（LU分解），因为L也是一个单位下三角阵，而且

$$\begin{aligned} L &= L_1^{-1} \cdots L_{n-1}^{-1} \\ &= (I + \ell_1 e_1^T) \cdots (I + \ell_{n-1} e_{n-1}^T) \\ &= I + \ell_1 e_1^T + \cdots + \ell_{n-1} e_n^T \end{aligned}$$

Gauss消去法

- 上述方法称为Gauss消去法

Gauss消去法

- 上述方法称为Gauss消去法
- 具体实现时，可以在原有的矩阵上存贮中间矩阵和最终的矩阵U

Gauss消去法

- 上述方法称为Gauss消去法
- 具体实现时，可以在原有的矩阵上存贮中间矩阵和最终的矩阵U
- 同时，可以利用A的下三角部分存贮L

Gauss消去法

- 上述方法称为Gauss消去法
- 具体实现时，可以在原有的矩阵上存贮中间矩阵和最终的矩阵U
- 同时，可以利用A的下三角部分存贮L
- 运算量：$\frac{2}{3}n^3 + O(n^2)$

主元

- 称Gauss消去过程中的对角元$a_{kk}^{(k-1)}$为主元

主元

- 称Gauss消去过程中的对角元$a_{kk}^{(k-1)}$为主元
- 当且仅当所有的主元均不为零时，上述算法才能进行到底

主元

- 称Gauss消去过程中的对角元$a_{kk}^{(k-1)}$为主元
- 当且仅当所有的主元均不为零时，上述算法才能进行到底

定理

所有主元均不为零当且仅当A的各阶顺序主子式均不为零

三角分解存在定理

定理

若A的各阶顺序主子式均不为零，则存在唯一的单位下三角阵L和上三角阵U，使得

$$A = LU$$

三角分解存在定理

定理

若A的各阶顺序主子式均不为零，则存在唯一的单位下三角阵L和上三角阵U，使得

$$A = LU$$

- 唯一性证明需要注意

三角分解存在定理

定理

若A的各阶顺序主子式均不为零，则存在唯一的单位下三角阵L和上三角阵U，使得

$$A = LU$$

- 唯一性证明需要注意
- 若A的前$n-1$个顺序主子式非零，但A奇异，定理仍成立：分块证明

非奇异矩阵

- 只要A非奇异，那么方程组$Ax = b$就有唯一解

非奇异矩阵

- 只要A非奇异，那么方程组$Ax = b$就有唯一解
- A非奇异并不能保证其各阶顺序主子式均不为零

非奇异矩阵

- 只要A非奇异，那么方程组$Ax = b$就有唯一解
- A非奇异并不能保证其各阶顺序主子式均不为零
- 从而A非奇异并不能保证Gauss消去过程的完整进行

非奇异矩阵

- 只要A非奇异，那么方程组$Ax = b$就有唯一解
- A非奇异并不能保证其各阶顺序主子式均不为零
- 从而A非奇异并不能保证Gauss消去过程的完整进行
- 主元非零，但很小，也会导致一些问题。`Mathematica1.2.1.nb`

例

- 取$A = \begin{pmatrix} 1/1000 & 1 \\ 1 & 2 \end{pmatrix}$, $b = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$

例

- 取$A = \begin{pmatrix} 1/1000 & 1 \\ 1 & 2 \end{pmatrix}$, $b = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$
- 方程$Ax = b$精确解为

$$x = \begin{pmatrix} \frac{500}{499} \\ \frac{997}{998} \end{pmatrix}$$

例

设计算精度只有三位浮点数，即第一个有效数字开始共三位。

- 矩阵$A = \begin{pmatrix} 1.00 \times 10^{-3} & 1.00 \\ 1.00 & 2.00 \end{pmatrix}$，
  $b = \begin{pmatrix} 1.00 \\ 3.00 \end{pmatrix}$

例

设计算精度只有三位浮点数，即第一个有效数字开始共三位。

- 矩阵$A = \begin{pmatrix} 1.00 \times 10^{-3} & 1.00 \\ 1.00 & 2.00 \end{pmatrix}$,

  $b = \begin{pmatrix} 1.00 \\ 3.00 \end{pmatrix}$
- 方程$Ax = b$精确解为

$$x = \begin{pmatrix} 1.00 \\ 1.00 \end{pmatrix}$$

实际计算

- $L = \begin{pmatrix} 1.00 & 1.00 \\ 0.00 \times 10^3 & 1.00 \end{pmatrix}$

实际计算

- $L = \begin{pmatrix} 1.00 & 1.00 \\ 0.00 \times 10^3 & 1.00 \end{pmatrix}$
- $L^{-1}A = \begin{pmatrix} 1.00 \times 10^{-3} & 1.00 \\ 0 & -1.00 \times 10^3 \end{pmatrix}$

实际计算

- $L = \begin{pmatrix} 1.00 & 1.00 \\ 0.00 \times 10^3 & 1.00 \end{pmatrix}$
- $L^{-1}A = \begin{pmatrix} 1.00 \times 10^{-3} & 1.00 \\ 0 & -1.00 \times 10^3 \end{pmatrix}$
- 这里注意在三位精度下$1.00 \times 10^3 - 3.00 = 1.00 \times 10^3$

求解

- 由$Ly = b$得到$y = (1.00, -1.00 \times 10^3)^T$

求解

- 由$Ly = b$得到$y = (1.00, -1.00 \times 10^3)^T$
- 由$Ux = y$得到$x = (0.00, 1.00)^T$

求解

- 由$Ly=b$得到$y=(1.00,-1.00\times10^3)^T$
- 由$Ux=y$得到$x=(0.00,1.00)^T$
- 与精确解的近似值$(1.00,1.00)^T$相差甚远

解决方法

- 问题主要是由于小主元引起的，使得运算时发生精度丢失

解决方法

- 问题主要是由于小主元引起的，使得运算时发生精度丢失
- 交换矩阵的两行，即交换两个方程的顺序，重复上述过程，得到近似解为$(1.00, 1.00)^T$

解决方法

- 问题主要是由于小主元引起的，使得运算时发生精度丢失
- 交换矩阵的两行，即交换两个方程的顺序，重复上述过程，得到近似解为$(1.00, 1.00)^T$
- 交换矩阵的两列，这时相当于交换两个变量的顺序

初等置换矩阵

定义矩阵

$$I_{pq} = (e_1, \dots, e_{p-1}, e_q, e_{p+1}, \dots, e_{q-1}, e_p, e_{q+1}, \dots, e_n)$$

用这个矩阵左乘A交换第p行和第q行，右乘A交换第p列和第q列

行列交换与Gauss消去

- 假定消去过程已进行了$k-1$步，即

$$
\begin{aligned}
A^{(k-1)} &= L_{k-1}P_{k-1}\cdots L_1P_1AQ_1\cdots Q_{k-1} \\
&= \begin{pmatrix} A_{11}^{(k-1)} & A_{12}^{(k-1)} \\ 0 & A_{22}^{(k-1)} \end{pmatrix}
\end{aligned}
$$

其中L_i为Gauss变换，P_i, Q_i为初等置换矩阵，$A_{11}^{(k-1)}$为$k-1$阶上三角阵

行列交换与Gauss消去

- 那么在第k步我们先在$A_{22}^{(k-1)}$中选择元素$a_{pq}^{(k-1)}$，其模在所有元素中最大

行列交换与Gauss消去

- 那么在第k步我们先在$A_{22}^{(k-1)}$中选择元素$a_{pq}^{(k-1)}$，其模在所有元素中最大
  - 如果$a_{pq}^{(k-1)} = 0$, 则A为奇异阵

行列交换与Gauss消去

- 那么在第k步我们先在$A_{22}^{(k-1)}$中选择元素$a_{pq}^{(k-1)}$，其模在所有元素中最大
  - 如果$a_{pq}^{(k-1)} = 0$, 则A为奇异阵
- 交换$A^{(k-1)}$的k,p行与k,q列，相当于左、右乘I_{kp}和I_{kq}

行列交换与Gauss消去

- 那么在第k步我们先在$A_{22}^{(k-1)}$中选择元素$a_{pq}^{(k-1)}$，其模在所有元素中最大
  - 如果$a_{pq}^{(k-1)} = 0$, 则A为奇异阵
- 交换$A^{(k-1)}$的k, p行与k, q列，相当于左、右乘I_{kp}和I_{kq}
- 然后进行Gauss变换

全主元Gauss消去法

- 上述消去过程称为全主元Gauss消去法

全主元Gauss消去法

- 上述消去过程称为全主元Gauss消去法
- 得到$(L_r P_r \cdots L_1 P_1) A (Q_1 \cdots Q_r) = U$为上三角阵

全主元Gauss消去法

- 上述消去过程称为全主元Gauss消去法
- 得到$(L_rP_r\cdots L_1P_1)A(Q_1\cdots Q_r)=U$为上三角阵
- 记

$$Q=Q_1\cdots Q_r$$
$$P=P_r\cdots P_1$$
$$L=P(L_rP_r\cdots L_1P_1)^{-1}$$

则$PAQ=LU$

L是单位下三角阵

- $L = P_r \cdots P_2 L_1^{-1} P_2 L_2^{-1} \cdots P_r L_r^{-1}$

L是单位下三角阵

- $L = P_r \cdots P_2 L_1^{-1} P_2 L_2^{-1} \cdots P_r L_r^{-1}$
- 记$L^{(1)} = L_1^{-1}$, $L^{(k)} = P_k L^{(k-1)} P_k L_k^{-1}$

L是单位下三角阵

- $L = P_r \cdots P_2 L_1^{-1} P_2 L_2^{-1} \cdots P_r L_r^{-1}$
- 记$L^{(1)} = L_1^{-1}$, $L^{(k)} = P_k L^{(k-1)} P_k L_k^{-1}$
- 归纳证明$L^{(k)}$具有形式

$$L^{(k)} = \begin{pmatrix} L_{11}^{(k)} & 0 \\ L_{21}^{(k)} & I_{n-k} \end{pmatrix}$$

其中$L_{11}^{(k)}$是所有元素之模均不大于1的k阶单位下三角阵，$L_{21}^{(k)}$是所有元素模均不大于1的$(n-k) \times k$阶矩阵

归纳证明之关键

- $P_k L^{(k-1)} P_k$是对$L^{(k-1)}$进行第k, p行和k, p列$(k \leqslant p)$交换, 因此只有$L_{21}^{(k-1)}$交换了两行— 类似于魔方中的局部交换技术

归纳证明之关键

- $P_k L^{(k-1)} P_k$是对$L^{(k-1)}$进行第k, p行和k, p列($k \leqslant p$)交换，因此只有$L_{21}^{(k-1)}$交换了两行— 类似于魔方中的局部交换技术
- 再右乘L_k^{-1}则使得I_{n-k+1}的第一列发生变化

全主元三角分解

- 如前所得的$PAQ = LU$称为全主元三角分解

全主元三角分解

- 如前所得的$PAQ = LU$称为全主元三角分解

定理

对于n阶方阵，存在排列矩阵P, Q以及单位下三角阵L和上三角阵U使得$PAQ = LU$，其中L的所有元素模均不大于1，U的非零对角元的个数恰好等于A的秩

选主元的运算量

- 当A非奇异时，选主元需要进行比较的次数

$$\sum_{k=1}^{n-1}(n-k+1)^2 = \frac{1}{3}n^3 + O(n^2)$$

选主元的运算量

- 当A非奇异时，选主元需要进行比较的次数

$$\sum_{k=1}^{n-1}(n-k+1)^2 = \frac{1}{3}n^3 + O(n^2)$$

- 这个计算量几乎是进行Gauss消去的计算量的一半

方程组的求解

- 设经全主元Gauss消去后得到$PAQ = LU$, 那么$PA = LUQ^{-1}$

方程组的求解

- 设经全主元Gauss消去后得到$PAQ = LU$, 那么$PA = LUQ^{-1}$
- 求解$Ly = Pb$得到y

方程组的求解

- 设经全主元Gauss消去后得到$PAQ = LU$, 那么$PA = LUQ^{-1}$
- 求解$Ly = Pb$得到y
- 求解$Uz = y$得到z

方程组的求解

- 设经全主元Gauss消去后得到$PAQ = LU$, 那么$PA = LUQ^{-1}$
- 求解$Ly = Pb$得到y
- 求解$Uz = y$得到z
- 计算$x = Qz$得到解x，这里即根据记录的交换指标对z进行元素交换即得到x

列主元消去

- 在第k步中只在$A_{22}^{(k-1)}$的第k列的元素中寻找模最大的元素，如此得到$PA = LU$, 称为列主元三角分解

列主元消去

- 在第k步中只在$A_{22}^{(k-1)}$的第k列的元素中寻找模最大的元素，如此得到$PA = LU$, 称为列主元三角分解
- 这里P的元素模不一定全不大于1

列主元消去

- 在第k步中只在$A_{22}^{(k-1)}$的第k列的元素中寻找模最大的元素，如此得到$PA = LU$, 称为列主元三角分解
- 这里P的元素模不一定全不大于1
- 例子：C代码`example1_2_2()`

对称正定线性方程组

- 对一般的方阵，为了消除LU分解的局限性和误差的积累，我们采用选取主元的方法

对称正定线性方程组

- 对一般的方阵，为了消除LU分解的局限性和误差的积累，我们采用选取主元的方法
- 对于对称正定矩阵而言，不必要选取主元

Cholesky分解定理

定理

若A为对称正定的，那么存在唯一的对角元均为正数的下三角阵L满足

$$A = LL^T$$

上式称为Cholesky分解，

L称为A的Cholesky因子

证明

- 对称正定$\Longrightarrow$不需要进行主元选择的Gauss消去法可行$\Longrightarrow$ 存在单位下三角阵$\tilde{L}$和上三角阵U, 使得$A = \tilde{L}U$

证明

- 对称正定$\Longrightarrow$不需要进行主元选择的Gauss消去法可行$\Longrightarrow$ 存在单位下三角阵$\tilde{L}$和上三角阵U, 使得$A=\tilde{L}U$
- 用U的对角元构造矩阵D, $\tilde{U}=D^{-1}U$, 则

$$\tilde{U}^T D\tilde{L}^T = A^T = A = \tilde{L}D\tilde{U}$$

证明

- 对称正定$\Longrightarrow$不需要进行主元选择的Gauss消去法可行$\Longrightarrow$ 存在单位下三角阵$\tilde{L}$和上三角阵U, 使得$A=\tilde{L}U$
- 用U的对角元构造矩阵D, $\tilde{U}=D^{-1}U$, 则

$$\tilde{U}^T D\tilde{L}^T = A^T = A = \tilde{L}D\tilde{U}$$

- 我们可有

$$\tilde{L}^T\tilde{U}^{-1} = D^{-1}\tilde{U}^{-T}\tilde{L}D$$

证明

$$\tilde{L}^T\tilde{U}^{-1} = D^{-1}\tilde{U}^{-T}\tilde{L}D$$

- $\tilde{L}^T\tilde{U}^{-1}$是单位上三角阵

证明

$$\tilde{L}^T\tilde{U}^{-1} = D^{-1}\tilde{U}^{-T}\tilde{L}D$$

- $\tilde{L}^T\tilde{U}^{-1}$是单位上三角阵
- $D^{-1}\tilde{U}^{-T}\tilde{L}D$是单位下三角阵

证明

$$\tilde{L}^T\tilde{U}^{-1} = D^{-1}\tilde{U}^{-T}\tilde{L}D$$

- $\tilde{L}^T\tilde{U}^{-1}$是单位上三角阵
- $D^{-1}\tilde{U}^{-T}\tilde{L}D$是单位下三角阵
- 所以两者都是单位阵，即$\tilde{U} = \tilde{L}^T$

分解的构造

- 从而有$A = \tilde{L}D\tilde{L}^T$, 而且$D$的对角元全为正数

分解的构造

- 从而有$A = \tilde{L}D\tilde{L}^T$, 而且$D$的对角元全为正数
- 取$L = \tilde{L}\,\mathrm{diag}(\sqrt{u_{11}}, \ldots, \sqrt{u_{nn}})$, 则有

$$A = LL^T$$

分解的构造

- 从而有$A=\tilde{L}D\tilde{L}^T$, 而且D的对角元全为正数
- 取$L=\tilde{L}\,\mathrm{diag}(\sqrt{u_{11}},\ldots,\sqrt{u_{nn}})$, 则有

$$A=LL^T$$

- 类似可证分解的唯一性

方程组的求解

- 计算A的Cholesky分解$A = LL^T$

方程组的求解

- 计算A的Cholesky分解$A = LL^T$
- 求解$Ly = b$得y

方程组的求解

- 计算A的Cholesky分解$A = LL^T$
- 求解$Ly = b$得y
- 求解$L^Tx = y$得x

平方根法

- 在计算分解时，可不必按前方法进行，而是采用待定系数法

平方根法

- 在计算分解时，可不必按前方法进行，而是采用待定系数法
- 待定下三角阵$L = (\ell_{ij})$, 比较$A = LL^T$两边对应的元素，可得

$$a_{ij} = \sum_{p=1}^{j} \ell_{ip}\ell_{jp}, \quad 1 \leqslant j \leqslant i \leqslant n$$

平方根法

- 在计算分解时，可不必按前方法进行，而是采用待定系数法
- 待定下三角阵$L=(\ell_{ij})$, 比较$A=LL^T$两边对应的元素，可得

$$a_{ij}=\sum_{p=1}^{j}\ell_{ip}\ell_{jp},\quad 1\leqslant j\leqslant i\leqslant n$$

- 由$a_{11}=\ell_{11}^2$得$\ell_{11}=\sqrt{a_{11}}$

平方根法

- 在计算分解时，可不必按前方法进行，而是采用待定系数法
- 待定下三角阵$L = (\ell_{ij})$, 比较$A = LL^T$两边对应的元素，可得

$$a_{ij} = \sum_{p=1}^{j} \ell_{ip}\ell_{jp}, \quad 1 \leqslant j \leqslant i \leqslant n$$

- 由$a_{11} = \ell_{11}^2$得$\ell_{11} = \sqrt{a_{11}}$
- 由$a_{i1} = \ell_{11}\ell_{i1}$得$\ell_{i1} = a_{i1}/\ell_{11}$, $i > 1$

平方根法

- 设已算出L的前$k-1$列元素

平方根法

- 设已算出L的前$k-1$列元素
- 由$a_{kk}=\sum_{p=1}^{k}\ell_{kp}^2$得到

$$\ell_{kk}=\sqrt{a_{kk}-\sum_{p=1}^{k-1}\ell_{kp}^2}$$

平方根法

- 设已算出L的前$k-1$列元素
- 由$a_{kk}=\sum_{p=1}^{k}\ell_{kp}^2$得到

$$\ell_{kk}=\sqrt{a_{kk}-\sum_{p=1}^{k-1}\ell_{kp}^2}$$

- 再由$a_{ik}=\sum_{p=1}^{k-1}\ell_{ip}\ell_{kp}+\ell_{ik}\ell_{kk}$得到

$$\ell_{ij}=\left(a_{ik}-\sum_{p=1}^{k-1}\ell_{ip}\ell_{kp}\right)/\ell_{kk}$$

注解

- 也可以按行来进行计算

注解

- 也可以按行来进行计算
- 可以在A中存贮新计算出来的L

注解

- 也可以按行来进行计算
- 可以在A中存贮新计算出来的L
- 上述方法的运算量是$n^3/3$，是Gauss消去法的一半

LDL^T分解

- 在Cholesky分解中用到了开方运算

LDL^T分解

- 在Cholesky分解中用到了开方运算
- 为了避免开方，可以计算A的如下形式分解

$$A = LDL^T$$

其中L是单位下三角阵，D是对角元均为正数的对角阵

LDL^T分解

- 在Cholesky分解中用到了开方运算
- 为了避免开方，可以计算A的如下形式分解

$$A = LDL^T$$

其中L是单位下三角阵，D是对角元均为正数的对角阵

- 这是Cholesky分解的变形

改进的平方根方法

比较$A = LDL^T$的对应元素，我们有

$$a_{ij} = \sum_{k=1}^{j-1} \ell_{ik} d_k \ell_{jk} + \ell_{ij} d_j, \quad 1 \leqslant j \leqslant i \leqslant n$$

则对$j = 1, \ldots,$,

$$v_k = d_k \ell_{jk}, \quad d_j = a_{jj} - \sum_{k=1}^{j-1} \ell_{jk} v_k,$$

$$\ell_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} v_k \right) / d_j, i = j+1, \ldots, n$$

注解

- 实际计算时，可以把L的严格下三角元素和D的对角元存储在A的对应位置上

注解

- 实际计算时，可以把L的严格下三角元素和D的对角元存储在A的对应位置上
- 算法运算量也是$n^3/3$, 而且不需要开方运算

方程组求解

- 求得LDL^T分解后，再只需求解

$$Ly = b, \quad DL^T x = y$$

就可以得到方程组的解

方程组求解

- 求得LDL^T分解后，再只需求解
$$Ly = b, \quad DL^T x = y$$
就可以得到方程组的解
- 如此方法是Gauss消去法的一半，而且不需要选主元

方程组求解

- 求得LDL^T分解后，再只需求解

$$Ly = b, \quad DL^T x = y$$

  就可以得到方程组的解
- 如此方法是Gauss消去法的一半，而且不需要选主元
- 由构造过程可知$|\ell_{ij}| \leqslant \sqrt{a_{ii}}$, 因此分解中的量受控的，从而计算过程稳定

例

- C程序example1_3_1()

例

- C程序example1_3_1()
- C程序example1_3_2()

线性方程组的敏度分析与消去法的舍入误差分析

邓建松

2018 年 10 月 10 日

向量范数

- 采用某种方法求解了线性方程组后，我们还需要考虑因素

  对解的影响

向量范数

- 采用某种方法求解了线性方程组后，我们还需要考虑因素
  - 数据不准确

  对解的影响

向量范数

- 采用某种方法求解了线性方程组后，我们还需要考虑因素
  - 数据不准确
  - 机器的有限精度

  对解的影响

向量范数

- 采用某种方法求解了线性方程组后，我们还需要考虑因素
  - 数据不准确
  - 机器的有限精度

  对解的影响
- 为此，我们需要用范数来描述向量与矩阵的扰动大小

向量范数

- 一个从$\mathbb{R}^n$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^n$上的向量范数，是指它满足

向量范数

- 一个从$\mathbb{R}^n$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^n$上的向量范数，是指它满足
  - 正定性：$\forall x \in \mathbb{R}^n$, $\|x\| \geqslant 0$; 且$\|x\| = 0$当且仅当$x = 0$

向量范数

- 一个从$\mathbb{R}^n$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^n$上的向量范数，是指它满足
  - 正定性：$\forall x \in \mathbb{R}^n$, $\|x\| \geqslant 0$; 且$\|x\| = 0$当且仅当$x = 0$
  - 齐次性：$\forall x \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, $\|\alpha x\| = |\alpha| \, \|x\|$

向量范数

- 一个从$\mathbb{R}^n$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^n$上的向量范数，是指它满足
  - 正定性：$\forall x \in \mathbb{R}^n$, $\|x\| \geqslant 0$; 且$\|x\| = 0$当且仅当$x = 0$
  - 齐次性：$\forall x \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, $\|\alpha x\| = |\alpha| \, \|x\|$
  - 三角不等式：$\forall x, y \in \mathbb{R}^n$, $\|x + y\| \leqslant \|x\| + \|y\|$

向量范数的性质

由范数的定义，易得

- $\forall x, y \in \mathbb{R}^n$,

$$\big|\|x\| - \|y\|\big| \leqslant \|x - y\| \leqslant \max_{1 \leqslant i \leqslant n} \|e_i\| \sum_{i=1}^{n} |x_i - y_i|$$

向量范数的性质

由范数的定义，易得

- $\forall x, y \in \mathbb{R}^n$,

$$\big|\|x\| - \|y\|\big| \leqslant \|x-y\| \leqslant \max_{1\leqslant i\leqslant n} \|e_i\| \sum_{i=1}^{n} |x_i - y_i|$$

- 所以$\|\cdot\|$作为$\mathbb{R}^n$上的实函数是连续的

p范数

最典型的一类范数是p范数

- 也称为Hölder范数

p范数

最典型的一类范数是p范数

- 也称为Hölder范数
- $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p},\ p \geqslant 1$

p范数

最典型的一类范数是p范数

- 也称为Hölder范数
- $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$, $p \geqslant 1$
- $p = 1, 2, \infty$是最常见的，也是最重要的

常用的p范数

- 1范数：$\|x\|_1 = |x_1| + \cdots + |x_n|$

常用的p范数

- 1范数：$\|x\|_1 = |x_1| + \cdots + |x_n|$
- 2范数：

$$\|x\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2} = \sqrt{x^T x}$$

常用的p范数

- 1范数： $\|x\|_1 = |x_1| + \cdots + |x_n|$
- 2范数：
  $\|x\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2} = \sqrt{x^T x}$
- ∞范数： $\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$

范数的等价性

定理

设$\|\cdot\|_\alpha$和$\|\cdot\|_\beta$是$\mathbb{R}^n$上任意两个范数，则存在正常数c_1，c_2，使得对一切$x \in \mathbb{R}^n$，

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

例

- (数分练习8.1 第6题)

$$\|x\|_2 \leqslant \|x\|_1 \leqslant \sqrt{n}\|x\|_2$$

例

- (数分练习8.1 第6题)

$$\|x\|_2 \leqslant \|x\|_1 \leqslant \sqrt{n}\|x\|_2$$

- (数分练习8.1 第7题)

$$\|x\|_\infty \leqslant \|x\|_2 \leqslant \sqrt{n}\|x\|_\infty$$

例

- (数分练习8.1 第6题)

$$\|x\|_2 \leqslant \|x\|_1 \leqslant \sqrt{n}\|x\|_2$$

- (数分练习8.1 第7题)

$$\|x\|_\infty \leqslant \|x\|_2 \leqslant \sqrt{n}\|x\|_\infty$$

- 显然有

$$\|x\|_\infty \leqslant \|x\|_1 \leqslant n\|x\|_\infty$$

收敛定理

定理

设$x_k \in \mathbb{R}^n$，则$\lim_{k\to\infty} \|x_k - x\| = 0$当且仅当

$$\lim_{k\to\infty} \left|x_i^{(k)} - x_i\right| = 0, \quad i = 1, \ldots, n$$

即向量序列的范数收敛等价于其分量收敛

向量范数→矩阵范数

- 通过把矩阵中的元素“拉直”构成向量，可以把矩阵的范数定义为拉直后向量的范数

向量范数→矩阵范数

- 通过把矩阵中的元素“拉直”构成向量，可以把矩阵的范数定义为拉直后向量的范数
- 这种范数称为广义矩阵范数

向量范数→矩阵范数

- 通过把矩阵中的元素“拉直”构成向量，可以把矩阵的范数定义为拉直后向量的范数
- 这种范数称为广义矩阵范数
- 矩阵的尺寸是任意的，并不需要是方阵

向量范数→矩阵范数

- 通过把矩阵中的元素“拉直”构成向量，可以把矩阵的范数定义为拉直后向量的范数
- 这种范数称为广义矩阵范数
- 矩阵的尺寸是任意的，并不需要是方阵
- 如此定义没有考虑矩阵的乘法运算

方阵的矩阵范数

一个从$\mathbb{R}^{n\times n}$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^{n\times n}$上的矩阵范数，是指

- 正定性：$\forall A\in\mathbb{R}^{n\times n}$, $\|A\|\geqslant 0$; 且$\|A\|=0$当且仅当$A=0$

方阵的矩阵范数

一个从$\mathbb{R}^{n\times n}$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^{n\times n}$上的矩阵范数，是指

- 正定性：$\forall A \in \mathbb{R}^{n\times n}$, $\|A\| \geqslant 0$; 且$\|A\| = 0$当且仅当$A = 0$
- 齐次性：$\forall A \in \mathbb{R}^{n\times n}$, $\alpha \in \mathbb{R}$, $\|\alpha A\| = |\alpha| \, \|A\|$

方阵的矩阵范数

一个从$\mathbb{R}^{n\times n}$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^{n\times n}$上的矩阵范数，是指

- 正定性：$\forall A\in\mathbb{R}^{n\times n}$, $\|A\|\geqslant 0$; 且$\|A\|=0$当且仅当$A=0$
- 齐次性：$\forall A\in\mathbb{R}^{n\times n}$, $\alpha\in\mathbb{R}$, $\|\alpha A\|=|\alpha|\,\|A\|$
- 三角不等式：$\forall A,B\in\mathbb{R}^{n\times n}$, $\|A+B\|\leqslant\|A\|+\|B\|$

方阵的矩阵范数

一个从$\mathbb{R}^{n\times n}$到$\mathbb{R}$的非负函数$\|\cdot\|$称为$\mathbb{R}^{n\times n}$上的矩阵范数，是指

- 正定性：$\forall A\in\mathbb{R}^{n\times n}$, $\|A\|\geqslant 0$; 且$\|A\|=0$当且仅当$A=0$
- 齐次性：$\forall A\in\mathbb{R}^{n\times n}$, $\alpha\in\mathbb{R}$, $\|\alpha A\|=|\alpha|\,\|A\|$
- 三角不等式：$\forall A,B\in\mathbb{R}^{n\times n}$, $\|A+B\|\leqslant\|A\|+\|B\|$
- 相容性：$\forall A,B\in\mathbb{R}^{n\times n}$, $\|AB\|\leqslant\|A\|\,\|B\|$

矩阵范数的性质

- $\mathbb{R}^{n\times n}$上的矩阵范数可以看作$\mathbb{R}^{n^2}$上的向量范数

矩阵范数的性质

- $\mathbb{R}^{n\times n}$上的矩阵范数可以看作$\mathbb{R}^{n^2}$上的向量范数
- $\mathbb{R}^{n\times n}$上的任意两个矩阵范数是等价的

矩阵范数的性质

- $\mathbb{R}^{n\times n}$上的矩阵范数可以看作$\mathbb{R}^{n^2}$上的向量范数
- $\mathbb{R}^{n\times n}$上的任意两个矩阵范数是等价的
- 矩阵序列的范数收敛等价于逐元素收敛

矩阵范数与向量范数的协调性

- 矩阵与向量的乘积是矩阵计算中经常用到的运算

矩阵范数与向量范数的协调性

- 矩阵与向量的乘积是矩阵计算中经常用到的运算
- 自然期望矩阵范数与向量范数之间具有某种协调性：把向量看作特殊的矩阵，那么定义中的相容性就是所期望的性质

相容性

- 若矩阵范数$\|\cdot\|_M$和向量范数$\|\cdot\|_v$满足

  $$\|Ax\|_v \leqslant \|A\|_M\|x\|_v, \quad \forall A \in \mathbb{R}^{n\times n}, v \in \mathbb{R}^n$$

  则称矩阵范数$\|\cdot\|_M$与$\|\cdot\|_v$是相容的

相容性

- 若矩阵范数$\|\cdot\|_M$和向量范数$\|\cdot\|_v$满足

  $$\|Ax\|_v \leqslant \|A\|_M\|x\|_v, \quad \forall A \in \mathbb{R}^{n\times n}, v \in \mathbb{R}^n$$

  则称矩阵范数$\|\cdot\|_M$与$\|\cdot\|_v$是**相容的**
- 今后凡同时涉及到向量范数与矩阵范数，都假定它们是相容的

相容性

- 若矩阵范数$\|\cdot\|_M$和向量范数$\|\cdot\|_v$满足

  $$\|Ax\|_v \leqslant \|A\|_M\|x\|_v, \quad \forall A \in \mathbb{R}^{n\times n}, v \in \mathbb{R}^n$$

  则称矩阵范数$\|\cdot\|_M$与$\|\cdot\|_v$是相容的
- 今后凡同时涉及到向量范数与矩阵范数，都假定它们是相容的
- 对任意给定的向量范数，我们都可以定义一个矩阵范数，使之与向量范数是相容的

从属于向量范数的矩阵范数

定理

设$\|\cdot\|$是$\mathbb{R}^n$上的一个向量范数。若定义

$$\|A\| = \max_{\|x\|=1} \|Ax\|, \quad A \in \mathbb{R}^{n\times n}$$

则$\|A\|$是$\mathbb{R}^{n\times n}$上的一个矩阵范数

证明

- 最大值可达到：$\|x\| = 1$是有界闭集，$\|\cdot\|$是连续的

证明

- 最大值可达到：$\|x\| = 1$是有界闭集，$\|\cdot\|$是连续的
- 矩阵与向量乘法的相容性：

$$\frac{\|Ax\|}{\|x\|} = \left\|A\frac{x}{\|x\|}\right\| \leqslant \|A\|$$

$$\Longrightarrow \|Ax\| \leqslant \|A\|\|x\|, \forall x \in \mathbb{R}^n$$

证明

- 正定性：取$A \neq 0$, 不妨设A的第i列是非零向量，即$Ae_i \neq 0$. 则

$$0 < \|Ae_i\| \leqslant \|A\|\|e_i\| \Longrightarrow \|A\| > 0$$

证明

- 正定性：取$A \neq 0$, 不妨设A的第i列是非零向量，即$Ae_i \neq 0$. 则

$$0 < \|Ae_i\| \leqslant \|A\| \|e_i\| \Longrightarrow \|A\| > 0$$

- 齐次性：显然

证明

- 三角不等式：取x, $\|x\| = 1$使得$\|(A+B)x\| = \|A+B\|$, 则

$$\begin{aligned}\|A+B\| = \|(A+B)x\| &\leqslant \|Ax\| + \|Bx\| \\ &\leqslant \|A\|\|x\| + \|B\|\|x\| = \|A\| + \|B\|\end{aligned}$$

证明

- 三角不等式：取x, $\|x\| = 1$使得$\|(A+B)x\| = \|A+B\|$, 则

$$\begin{aligned}\|A+B\| &= \|(A+B)x\| \leqslant \|Ax\| + \|Bx\| \\ &\leqslant \|A\|\|x\| + \|B\|\|x\| = \|A\| + \|B\|\end{aligned}$$

- 相容性：取x, $\|x\| = 1$使得$\|ABx\| = \|AB\|$,

$$\begin{aligned}\|AB\| &= \|ABx\| \leqslant \|A\|\|Bx\| \\ &\leqslant \|A\|\|B\|\|x\| = \|A\|\|B\|\end{aligned}$$

算子范数

- 称由

$$\|A\| = \max_{\|x\|=1} \|Ax\|, \quad A \in \mathbb{R}^{n\times n}$$

所定义的范数$\|\cdot\|$为从属于向量范数$\|\cdot\|$的矩阵范数，也称为由向量范数$\|\cdot\|$诱导出的算子范数

算子范数

- 称由

$$\|A\| = \max_{\|x\|=1} \|Ax\|, \quad A \in \mathbb{R}^{n\times n}$$

所定义的范数$\|\cdot\|$为从属于向量范数$\|\cdot\|$的矩阵范数，也称为由向量范数$\|\cdot\|$诱导出的算子范数

- 今后仍把$\|\cdot\|$记作$\|\cdot\|$

单位阵的范数

- 对广义矩阵范数，$\|I\| > 0$

单位阵的范数

- 对广义矩阵范数，$\|I\| > 0$
- 对矩阵范数，$\|I\| \geqslant 1$

单位阵的范数

- 对广义矩阵范数，$\|I\| > 0$
- 对矩阵范数，$\|I\| \geqslant 1$
- 对算子范数，$\|I\| = 1$

算子范数$\|\cdot\|_p$

- 前面定义了向量的p范数

算子范数$\|\cdot\|_p$

- 前面定义了向量的p范数
- 由其诱导出矩阵的p范数如下：

$$\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p$$

算子范数$\|\cdot\|_p$

- 前面定义了向量的p范数
- 由其诱导出矩阵的p范数如下：

$$\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p$$

- $p = 1, 2, \infty$时对应的算子范数具有简捷的表示

“列和”范数

对于矩阵$A=(a_{ij})\in\mathbb{R}^{n\times n}$, 我们有

$$\|A\|_1=\max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}|a_{ij}|$$

- 记A的第j列为向量a_j, 不妨设$\delta=\|a_1\|=\max_{1\leqslant j\leqslant n}\|a_j\|_1$

“列和”范数

对于矩阵$A = (a_{ij}) \in \mathbb{R}^{n\times n}$, 我们有

$$\|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|$$

- 记A的第j列为向量a_j, 不妨
  设$\delta = \|a_1\| = \max\limits_{1\leqslant j\leqslant n} \|a_j\|_1$
- 取x, $\|x\|_1 = 1$, 则

$$\|Ax\|_1 = \left\| \sum_{j=1}^{n} x_j a_j \right\|_1 \leqslant \sum_{j=1}^{n} |x_j| \|a_j\|_1 \leqslant \delta$$

- 注意到$\|e_1\|_1 = 1$

- 注意到$\|e_1\|_1 = 1$
- $\|Ae_1\|_1 = \|a_1\|_1 = \delta$

- 注意到$\|e_1\|_1 = 1$
- $\|Ae_1\|_1 = \|a_1\|_1 = \delta$
- 所以$\|A\|_1 = \delta = \max\limits_{1\leqslant j\leqslant n} \sum\limits_{i=1}^{n} |a_{ij}|$

“行和”范数

对于矩阵$A=(a_{ij})\in\mathbb{R}^{n\times n}$, 我们有

$$\|A\|_\infty=\max_{1\leqslant i\leqslant n}\sum_{j=1}^n|a_{ij}|$$

- 对于任意x, $\|x\|_\infty=1$, 我们有

$$\|Ax\|_\infty=\max_{1\leqslant i\leqslant n}\left|\sum_{j=1}^n a_{ij}x_j\right|$$

$$\leqslant\max_{1\leqslant i\leqslant n}\sum_{j=1}^n|a_{ij}||x_j|\leqslant\max_{1\leqslant i\leqslant n}\sum_{j=1}^n|a_{ij}|$$

- 不妨设A中第1行对应向量的1范数最大，取

$$\bar{x} = (\operatorname{sgn} a_{11}, \ldots, \operatorname{sgn} a_{1n})^T$$

- 不妨设A中第1行对应向量的1范数最大，取

$$\bar{x} = (\operatorname{sgn} a_{11}, \ldots, \operatorname{sgn} a_{1n})^T$$

- $A \neq 0 \Longrightarrow \|\bar{x}\|_\infty = 1$，而且

$$\|A\bar{x}\|_\infty = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} |a_{ij}|$$

谱范数

对于矩阵$A = (a_{ij}) \in \mathbb{R}^{n\times n}$, 我们有

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$$

这里$\lambda_{\max}$表示参数矩阵的最大特征值

- 根据定义

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=1} \sqrt{x^TA^TAx}$$

- 设对称半正定矩阵A^TA的特征值为

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_n \geqslant 0$$

对应规范正交特征向量为$v_1, \cdots, v_n$

- 设对称半正定矩阵A^TA的特征值为

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \lambda_n \geqslant 0$$

  对应规范正交特征向量为$v_1, \cdots, v_n$
- 取x, $\|x\|_2 = 1$, 则

$$x = \sum_{i=1}^{n} \alpha_i v_i, \quad \sum_{i=1}^{n} \alpha_i^2 = 1$$

- $$x^TA^TAx = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leqslant \lambda_1$$

- $$x^T A^T A x = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leqslant \lambda_1$$
- 取$x = v_1$, 则
$$x^T A^T A x = v_1^T A^T A v_1 = v_1^T \lambda_1 v_1 = \lambda_1$$

- $$x^TA^TAx = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leqslant \lambda_1$$

- 取$x = v_1$，则
$$x^TA^TAx = v_1^TA^TAv_1 = v_1^T\lambda_1 v_1 = \lambda_1$$

- 从而
$$\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$$

- $$x^TA^TAx = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leqslant \lambda_1$$
- 取$x = v_1$, 则
$$x^TA^TAx = v_1^TA^TAv_1 = v_1^T\lambda_1 v_1 = \lambda_1$$
- 从而
$$\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$$
- A^TA的特征值的开方称为A的奇异值

算子范数的性质

- $\|A\|_2 = \max\{|y^T Ax| : x, y \in \mathbb{C}^n, \|x\|_2 = \|y\|_2 = 1\}$

算子范数的性质

- $\|A\|_2 = \max\{|y^T Ax| : x, y \in \mathbb{C}^n, \|x\|_2 = \|y\|_2 = 1\}$
- $\|A^T\|_2 = \|A\|_2 = \sqrt{\|A^T A\|_2}$

算子范数的性质

- $\|A\|_2 = \max\{|y^TAx| : x, y \in \mathbb{C}^n, \|x\|_2 = \|y\|_2 = 1\}$
- $\|A^T\|_2 = \|A\|_2 = \sqrt{\|A^TA\|_2}$
- 对任意n阶正交阵U, V, 我们有$\|UA\|_2 = \|AV\|_2 = \|A\|_2$

算子范数的性质

- $\|A\|_2 = \max\{|y^T Ax| : x, y \in \mathbb{C}^n, \|x\|_2 = \|y\|_2 = 1\}$
- $\|A^T\|_2 = \|A\|_2 = \sqrt{\|A^T A\|_2}$
- 对任意n阶正交阵U, V, 我们有$\|UA\|_2 = \|AV\|_2 = \|A\|_2$
- $\|A\|_\infty = \|A^T\|_1$

Frobenius范数

- 常用且易于计算的矩阵范数：

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2\right)^{1/2}$$

Frobenius范数

- 常用且易于计算的矩阵范数：

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

- 这是向量2范数的自然推广，也称为Hilbert-Schmidt范数

Frobenius范数

- 常用且易于计算的矩阵范数：

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

- 这是向量2范数的自然推广，也称为Hilbert-Schmidt范数
- 该范数满足对矩阵乘法的相容性

Frobenius范数

- 常用且易于计算的矩阵范数：

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

- 这是向量2范数的自然推广，也称为Hilbert-Schmidt范数
- 该范数满足对矩阵乘法的相容性
- 该范数与向量的2范数是相容的

Frobenius范数的性质

- $\|A\|_F = \sqrt{\text{trace}(A^T A)}$

Frobenius范数的性质

- $\|A\|_F = \sqrt{\mathrm{trace}(A^TA)}$
- $\|AB\|_F \leqslant \|A\|_2\|B\|_F$, $\|AB\|_F \leqslant \|A\|_F\|B\|_2$

Frobenius范数的性质

- $\|A\|_F = \sqrt{\mathrm{trace}(A^TA)}$
- $\|AB\|_F \leqslant \|A\|_2\|B\|_F$, $\|AB\|_F \leqslant \|A\|_F\|B\|_2$
- $\|\cdot\|_F$是矩阵范数，但它不是算子范数（为什么？）

Frobenius范数的性质

- $\|A\|_F = \sqrt{\mathrm{trace}(A^T A)}$
- $\|AB\|_F \leqslant \|A\|_2 \|B\|_F$, $\|AB\|_F \leqslant \|A\|_F \|B\|_2$
- $\|\cdot\|_F$是矩阵范数，但它不是算子范数（为什么？）
  - $\|I\|_F = \sqrt{n}$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max_{i,j} |a_{ij}| \leqslant \|A\|_2 \leqslant n \max_{i,j} |a_{ij}|$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max_{i,j} |a_{ij}| \leqslant \|A\|_2 \leqslant n \max_{i,j} |a_{ij}|$
- $\|A\|_\infty/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max\limits_{i,j} |a_{ij}| \leqslant \|A\|_2 \leqslant n \max\limits_{i,j} |a_{ij}|$
- $\|A\|_\infty/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$
- $\|A\|_1/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_1$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max_{i,j} |a_{ij}| \leqslant \|A\|_2 \leqslant n \max_{i,j} |a_{ij}|$
- $\|A\|_\infty/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$
- $\|A\|_1/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_1$
- $\|A\|_2 \leqslant \sqrt{\|A\|_1\|A\|_\infty}$

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max_{i,j} |a_{ij}| \leqslant \|A\|_2 \leqslant n \max_{i,j} |a_{ij}|$
- $\|A\|_\infty/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$
- $\|A\|_1/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_1$
- $\|A\|_2 \leqslant \sqrt{\|A\|_1\|A\|_\infty}$
- 利用$\|A\|_F$的trace表达式证明性质1; 利用$\|A\|_2$的定义，取特殊的x证明2,3,4,5等性质

矩阵范数的性质及等价性

- $\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{n}\|A\|_2$
- $\max\limits_{i,j}|a_{ij}| \leqslant \|A\|_2 \leqslant n\max\limits_{i,j}|a_{ij}|$
- $\|A\|_\infty/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_\infty$
- $\|A\|_1/\sqrt{n} \leqslant \|A\|_2 \leqslant \sqrt{n}\|A\|_1$
- $\|A\|_2 \leqslant \sqrt{\|A\|_1\|A\|_\infty}$
- 利用$\|A\|_F$的trace表达式证明性质1; 利用$\|A\|_2$的定义，取特殊的x证明2,3,4,5等性质
- 各系数是最佳的（为什么？）

证明示例：$\|A\|_2 \leqslant \sqrt{\|A\|_1 \|A\|_\infty}$

- 取$x \neq 0$, 使得$A^T Ax = \lambda^2 x$, 其中$\lambda = \|A\|_2$

证明示例：$\|A\|_2 \leqslant \sqrt{\|A\|_1\|A\|_\infty}$

- 取$x \neq 0$, 使得$A^TAx = \lambda^2 x$, 其中$\lambda = \|A\|_2$
- 那么

$$\begin{aligned}
\lambda^2\|x\|_1 &= \|A^TAx\|_1 \\
&\leqslant \|A^T\|_1\|A\|_1\|z\|_1 \\
&= \|A\|_\infty\|A\|_1\|x\|_1
\end{aligned}$$

谱半径

- 设$A \in \mathbb{C}^{n\times n}$, 则称

$$\rho(A) = \max\{|\lambda| : \lambda \in \lambda(A)\}$$

为A的谱半径，这里$\lambda(A)$表示A的特征值的全体

谱半径与矩阵范数

定理

设$A \in \mathbb{C}^{n\times n}$，则

- 对任意矩阵范数$\|\cdot\|$，$\rho(A) \leqslant \|A\|$；
- 对$\forall\varepsilon > 0$，存在算子范数$\|\cdot\|$，使得

$$\|A\| \leqslant \rho(A) + \varepsilon$$

$\rho(A) \leqslant \|A\|$的证明

- 取$x \neq 0$, $Ax = \lambda x$, $|\lambda| = \rho(A)$

$\rho(A) \leqslant \|A\|$的证明

- 取$x \neq 0$, $Ax = \lambda x$, $|\lambda| = \rho(A)$
- 则有(xe_1^T表示以x为第一列的增广矩阵)

$$\rho(A)\|xe_1^T\| = \|\lambda xe_1^T\| = \|Axe_1^T\| \leqslant \|A\|\|xe_1^T\|$$

$\rho(A) \leqslant \|A\|$的证明

- 取$x \neq 0$, $Ax = \lambda x$, $|\lambda| = \rho(A)$
- 则有(xe_1^T表示以x为第一列的增广矩阵)

$$\rho(A)\|xe_1^T\| = \|\lambda xe_1^T\| = \|Axe_1^T\| \leqslant \|A\|\|xe_1^T\|$$

- 所以

$$\rho(A) \leqslant \|A\|$$

$\|A\| \leqslant \rho(A) + \varepsilon$的证明

- Jordan分解定理：

$$X^{-1}AX = \begin{pmatrix} \lambda_1 & \delta_1 & & & \\ & \lambda_2 & \delta_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & \delta_{n-1} \\ & & & & \lambda_n \end{pmatrix}$$

其中$\delta_i = 1$或0

$\|A\| \leqslant \rho(A) + \varepsilon$的证明

- 对$\varepsilon > 0$取$D_\varepsilon = \mathrm{diag}(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{n-1})$

$$D_\varepsilon^{-1} X^{-1} A X D_\varepsilon = \begin{pmatrix} \lambda_1 & \varepsilon\delta_1 & & & \\ & \lambda_2 & \varepsilon\delta_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & \varepsilon\delta_{n-1} \\ & & & & \lambda_n \end{pmatrix}$$

$\|A\| \leqslant \rho(A) + \varepsilon$的证明

- 定义

$$\|G\|_\varepsilon = \|(XD_\varepsilon)^{-1}G(XD_\varepsilon)\|_\infty$$

这是由下述向量范数诱导出的算子范数：

$$\|x\|_{XD_\varepsilon} = \|(XD_\varepsilon)^{-1}x\|_\infty$$

诱导性的证明

- 由定义：

$$\|G\|_{XD_\varepsilon} = \max_{\|x\|_{XD_\varepsilon}=1} \|Gx\|_{XD_\varepsilon}$$

诱导性的证明

- 由定义：

$$\|G\|_{XD_\varepsilon} = \max_{\|x\|_{XD_\varepsilon}=1} \|Gx\|_{XD_\varepsilon}$$

-

$$\begin{aligned}\|Gx\|_{XD_\varepsilon} &= \|(XD_\varepsilon)^{-1}Gx\|_\infty \\ &= \|(XD_\varepsilon)^{-1}G(XD_\varepsilon)(XD_\varepsilon)^{-1}x\|_\infty \\ &\leqslant \|G\|_\varepsilon\|x\|_{XD_\varepsilon} = \|G\|_\varepsilon\end{aligned}$$

诱导性的证明

- 由定义：

$$\|G\|_{XD_\varepsilon} = \max_{\|x\|_{XD_\varepsilon}=1} \|Gx\|_{XD_\varepsilon}$$

-

$$\begin{aligned}\|Gx\|_{XD_\varepsilon} &= \|(XD_\varepsilon)^{-1}Gx\|_\infty \\ &= \|(XD_\varepsilon)^{-1}G(XD_\varepsilon)(XD_\varepsilon)^{-1}x\|_\infty \\ &\leqslant \|G\|_\varepsilon \|x\|_{XD_\varepsilon} = \|G\|_\varepsilon\end{aligned}$$

- 所以$\|G\|_{XD_\varepsilon} \leqslant \|G\|_\varepsilon$.

诱导性的证明

- 由定义：
$$\|G\|_{XD_\varepsilon} = \max_{\|x\|_{XD_\varepsilon}=1} \|Gx\|_{XD_\varepsilon}$$

-
$$\begin{aligned}\|Gx\|_{XD_\varepsilon} &= \|(XD_\varepsilon)^{-1}Gx\|_\infty \\ &= \|(XD_\varepsilon)^{-1}G(XD_\varepsilon)(XD_\varepsilon)^{-1}x\|_\infty \\ &\leqslant \|G\|_\varepsilon \|x\|_{XD_\varepsilon} = \|G\|_\varepsilon\end{aligned}$$

- 所以$\|G\|_{XD_\varepsilon} \leqslant \|G\|_\varepsilon$.
- 由矩阵∞范数的性质，可知等号成立

$\|A\| \leqslant \rho(A) + \varepsilon$的证明

- 从而有

$$\|A\|_{\varepsilon} = \max_{1 \leqslant i \leqslant n} (|\lambda_i| + |\varepsilon \delta_i|) \leqslant \rho(A) + \varepsilon$$

谱半径与收敛

定理

设$A \in \mathbb{C}^{n\times n}$,

$$\lim_{k\to\infty} A^k = 0 \iff \rho(A) < 1$$

必要性

- 取$\lambda \in \lambda(A)$满足$|\lambda| = \rho(A)$

必要性

- 取$\lambda \in \lambda(A)$满足$|\lambda| = \rho(A)$
- $\forall k,\ \lambda^k \in \lambda(A^k)$

必要性

- 取$\lambda \in \lambda(A)$满足$|\lambda| = \rho(A)$
- $\forall k,\ \lambda^k \in \lambda(A^k)$
- $(\rho(A))^k = |\lambda|^k \leqslant \rho(A^k) \leqslant \|A^k\|_2$

必要性

- 取$\lambda \in \lambda(A)$满足$|\lambda| = \rho(A)$
- $\forall k,\ \lambda^k \in \lambda(A^k)$
- $(\rho(A))^k = |\lambda|^k \leqslant \rho(A^k) \leqslant \|A^k\|_2$
- 所以有$\rho(A) < 1$

充分性

- 设$\rho(A) < 1$

充分性

- 设$\rho(A) < 1$
- 存在算子范数$\|\cdot\|$, 使得$\|A\| < 1$

充分性

- 设$\rho(A) < 1$
- 存在算子范数$\|\cdot\|$, 使得$\|A\| < 1$
- $0 \leqslant \|A^k\| \leqslant \|A\|^k \to 0, \quad k \to \infty$

充分性

- 设$\rho(A) < 1$
- 存在算子范数$\|\cdot\|$, 使得$\|A\| < 1$
- $0 \leqslant \|A^k\| \leqslant \|A\|^k \to 0, \quad k \to \infty$
- 所以$\lim\limits_{k\to\infty} A^k = 0.$

矩阵级数的收敛性

定理

- $\sum_{k=0}^{\infty} A^k$收敛当且仅当$\rho(A) < 1$
- 当$\sum_{k=0}^{\infty} A^k$收敛时我们有$\sum_{k=0}^{\infty} A^k = (I-A)^{-1}$ 且存在算子范数$\|\cdot\|$使得对$\forall m \geqslant 0$

$$\left\|(I-A)^{-1} - \sum_{k=0}^{m} A^k\right\| \leqslant \frac{\|A\|^{m+1}}{1-\|A\|}$$

推论

- 设$\|\cdot\|$是$\mathbb{C}^{n\times n}$上的矩阵范数，并假设$A\in\mathbb{C}^{n\times n}$满足$\|A\|<1$, 则$I-A$可逆

推论

- 设$\|\cdot\|$是$\mathbb{C}^{n\times n}$上的矩阵范数，并假设$A\in\mathbb{C}^{n\times n}$满足$\|A\|<1$, 则$I-A$可逆
- 进一步，若对此范数有$\|I\|=1$, 则

$$\|(I-A)^{-1}\|\leqslant\frac{1}{1-\|A\|}$$

敏度分析的问题描述

- 一个线性方程组$Ax = b$是由系数矩阵A和右端项b所确定的

敏度分析的问题描述

- 一个线性方程组$Ax = b$是由系数矩阵A和右端项b所确定的
- A或b通常是带有误差的，这些误差相对于精确值是微小的

敏度分析的问题描述

- 一个线性方程组$Ax = b$是由系数矩阵A和右端项b所确定的
- A或b通常是带有误差的，这些误差相对于精确值是微小的
- 这种微小误差对解有何影响？这就是线性方程组的敏感性问题

微小扰动

- 确实有方程组，系数的小扰动使解产生巨大变化。Mathematica2.2.1.nb

微小扰动

- 确实有方程组，系数的小扰动使解产生巨大变化。Mathematica2.2.1.nb
- 对于$Ax = b$, 经扰动后变为

$$(A + \delta A)(x + \delta x) = b + \delta b$$

微小扰动

- 确实有方程组，系数的小扰动使解产生巨大变化。Mathematica2.2.1.nb
- 对于$Ax=b$, 经扰动后变为

$$(A+\delta A)(x+\delta x)=b+\delta b$$

- 利用$Ax=b$简化后

$$(A+\delta A)\delta x=\delta b-(\delta A)x$$

- A非奇异，由前述推论，当$\|A^{-1}\|\|\delta A\| < 1$时，$I + A^{-1}\delta A$可逆，而且若$\|I\| = 1$则

$$\|(I + A^{-1}\delta A)^{-1}\| \leqslant \frac{1}{1 - \|A^{-1}\|\|\delta A\|}$$

- A非奇异，由前述推论，当$\|A^{-1}\|\|\delta A\| < 1$时，$I + A^{-1}\delta A$可逆，而且若$\|I\| = 1$则

$$\|(I + A^{-1}\delta A)^{-1}\| \leqslant \frac{1}{1 - \|A^{-1}\|\|\delta A\|}$$

- 从而$A + \delta A = A(I + A^{-1}\delta A)$非奇异，

$$\begin{aligned}\delta x &= (A + \delta A)^{-1}(\delta b - (\delta A)x) \\ &= (I + A^{-1}\delta A)^{-1}A^{-1}(\delta b - (\delta A)x)\end{aligned}$$

- 两边取范数

$$\begin{aligned}\|\delta x\| &\leqslant \|(I + A^{-1}\delta A)^{-1}\|\|A^{-1}\|(\|\delta b\| + \|\delta A\|\|x\|) \\ &\leqslant \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\delta A\|}(\|\delta b\| + \|\delta A\|\|x\|)\end{aligned}$$

- 两边取范数

$$\begin{aligned}\|\delta x\| &\leqslant \|(I+A^{-1}\delta A)^{-1}\|\|A^{-1}\|(\|\delta b\|+\|\delta A\|\|x\|)\\ &\leqslant \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}(\|\delta b\|+\|\delta A\|\|x\|)\end{aligned}$$

- 当$x\neq 0$, 注意到$\|b\|\leqslant\|A\|\|x\|$, 上式两边同除以$\|x\|$我们有

$$\frac{\|\delta x\|}{\|x\|}\leqslant\frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\right)$$

定理

定理

$\|\cdot\|$是一个满足$\|I\|=1$的矩阵范数，A非奇异，$b\neq 0$, $\|A^{-1}\|\|\delta A\|<1$. 若x和$x+\delta x$分别是$Ax=b$和$(A+\delta x)(x+\delta x)=b+\delta b$的解，则$(\kappa(A)=\|A\|\|A^{-1}\|)$

$$\frac{\|\delta x\|}{\|x\|}\leqslant\frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\right)$$

$\kappa(A)$

- 当$\|\delta A\|/\|A\|$较小时，

$$\frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}} \approx \kappa(A)$$

$\kappa(A)$

- 当$\|\delta A\|/\|A\|$较小时，

$$\frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}} \approx \kappa(A)$$

- 从而有

$$\frac{\|\delta x\|}{\|x\|} \lessapprox \kappa(A)\left(\frac{\|\delta A\|}{\|A\|}+\frac{\|\delta b\|}{\|b\|}\right)$$

条件数

- x的相对误差的上界是b和A的相对误差之和乘以$\kappa(A)$

条件数

- x的相对误差的上界是b和A的相对误差之和乘以$\kappa(A)$
- 扰动对方程组解的影响大小与$\kappa(A)$密切相关

条件数

- x的相对误差的上界是b和A的相对误差之和乘以$\kappa(A)$
- 扰动对方程组解的影响大小与$\kappa(A)$密切相关
  - 若$\kappa(A)$不大，则扰动对解的影响也不会太大

条件数

- x的相对误差的上界是b和A的相对误差之和乘以$\kappa(A)$
- 扰动对方程组解的影响大小与$\kappa(A)$密切相关
  - 若$\kappa(A)$不大，则扰动对解的影响也不会太大
  - 若$\kappa(A)$很大，则扰动对解的影响可能就很大

条件数

- x的相对误差的上界是b和A的相对误差之和乘以$\kappa(A)$
- 扰动对方程组解的影响大小与$\kappa(A)$密切相关
  - 若$\kappa(A)$不大，则扰动对解的影响也不会太大
  - 若$\kappa(A)$很大，则扰动对解的影响可能就很大
- $\kappa(A)$称为线性方程组$Ax = b$的条件数

病态与良态方程组

- 条件数在一定程度上刻画了扰动对方程组解的影响程度

病态与良态方程组

- 条件数在一定程度上刻画了扰动对方程组解的影响程度
- 若条件数很大，就称该线性方程组的求解问题是病态的，有时也直接称A是病态的

病态与良态方程组

- 条件数在一定程度上刻画了扰动对方程组解的影响程度
- 若条件数很大，就称该线性方程组的求解问题是病态的，有时也直接称A是病态的
- 若条件数很小，就称该线性方程组的求解问题是良态的，有时也直接称A是良态的

条件数与范数

- 条件数与所采用的范数有关

条件数与范数

- 条件数与所采用的范数有关
- 可以通过在条件数上加下标表明所采用的范数：$\kappa_2(A)$

条件数与范数

- 条件数与所采用的范数有关
- 可以通过在条件数上加下标表明所采用的范数：$\kappa_2(A)$
- 病态、良态性是否与范数有关呢？

条件数的等价性

- 由矩阵范数的等价性可知任两个范数意义下的条件数也是等价的

条件数的等价性

- 由矩阵范数的等价性可知任两个范数意义下的条件数也是等价的
- $\frac{1}{n}\kappa_2(A) \leqslant \kappa_1(A) \leqslant n\kappa_2(A)$

条件数的等价性

- 由矩阵范数的等价性可知任两个范数意义下的条件数也是等价的
- $\frac{1}{n}\kappa_2(A) \leqslant \kappa_1(A) \leqslant n\kappa_2(A)$
- $\frac{1}{n}\kappa_\infty(A) \leqslant \kappa_2(A) \leqslant n\kappa_\infty(A)$

条件数的等价性

- 由矩阵范数的等价性可知任两个范数意义下的条件数也是等价的
- $\frac{1}{n}\kappa_2(A) \leqslant \kappa_1(A) \leqslant n\kappa_2(A)$
- $\frac{1}{n}\kappa_\infty(A) \leqslant \kappa_2(A) \leqslant n\kappa_\infty(A)$
- $\frac{1}{n^2}\kappa_1(A) \leqslant \kappa_\infty(A) \leqslant n^2\kappa_1(A)$

矩阵求逆

- 设$\|\cdot\|$是满足$\|I\|=1$的矩阵范数，A非奇异，且$\|A^{-1}\|\|\delta A\|<1$, 则$A+\delta A$也是非奇异的，而且有

$$\frac{\|(A+\delta A)^{-1}-A^{-1}\|}{\|A^{-1}\|}\leqslant\frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}\frac{\|\delta A\|}{\|A\|}$$

矩阵求逆

- 设$\|\cdot\|$是满足$\|I\|=1$的矩阵范数，A非奇异，且$\|A^{-1}\|\|\delta A\|<1$, 则$A+\delta A$也是非奇异的，而且有

$$\frac{\|(A+\delta A)^{-1}-A^{-1}\|}{\|A^{-1}\|}\leqslant\frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}\frac{\|\delta A\|}{\|A\|}$$

  - $(A+\delta A)^{-1}-A^{-1}=(A+\delta A)^{-1}(\delta A)A^{-1}$

矩阵求逆

- 设$\|\cdot\|$是满足$\|I\|=1$的矩阵范数，A非奇异，且$\|A^{-1}\|\|\delta A\|<1$, 则$A+\delta A$也是非奇异的，而且有

$$\frac{\|(A+\delta A)^{-1}-A^{-1}\|}{\|A^{-1}\|}\leqslant\frac{\kappa(A)}{1-\kappa(A)\frac{\|\delta A\|}{\|A\|}}\frac{\|\delta A\|}{\|A\|}$$

  - $(A+\delta A)^{-1}-A^{-1}=(A+\delta A)^{-1}(\delta A)A^{-1}$

- 这表明$\kappa(A)$也可以看作矩阵求逆问题的条件数

条件数的几何意义

定理

设A非奇异，

$$\min\left\{\frac{\|\delta A\|_2}{\|A\|_2}: A+\delta A\text{奇异}\right\}=\frac{1}{\kappa_2(A)}$$

- 即在谱范数下，矩阵的条件数的倒数恰好等于该矩阵与全体奇异矩阵所成集合的相对距离

条件数的几何意义

定理

设A非奇异，

$$\min\left\{\frac{\|\delta A\|_2}{\|A\|_2} : A+\delta A\text{奇异}\right\} = \frac{1}{\kappa_2(A)}$$

- 即在谱范数下，矩阵的条件数的倒数恰好等于该矩阵与全体奇异矩阵所成集合的相对距离
- 当A十分病态时，它与一个奇异矩阵十分靠近

证明

只需证明：

$$\min\left\{\|\delta A\|_2 : A+\delta A\text{奇异}\right\} = \frac{1}{\|A^{-1}\|_2}$$

- 由推论，当$\|A^{-1}\|_2\|\delta A\|_2 < 1$时，$A+\delta A$非奇异，从而有

$$\min\left\{\|\delta A\|_2 : A+\delta A\text{奇异}\right\} \geqslant \frac{1}{\|A^{-1}\|_2}$$

证明

- 由于谱范数是向量2范数诱导出的算子范数，所以存在x, $\|x\|_2 = 1$, 使得$\|A^{-1}x\|_2 = \|A^{-1}\|_2$

证明

- 由于谱范数是向量2范数诱导出的算子范数，所以存在x, $\|x\|_2 = 1$, 使得$\|A^{-1}x\|_2 = \|A^{-1}\|_2$
- 令$y = \dfrac{A^{-1}x}{\|A^{-1}x\|_2}$, $\delta A = -\dfrac{xy^T}{\|A^{-1}\|_2}$ 则$\|y\|_2 = 1$而且

$$\begin{aligned}(A+\delta A)y &= Ay + (\delta A)y \\ &= \frac{x}{\|A^{-1}x\|_2} - \frac{x}{\|A^{-1}\|_2} \\ &= 0\end{aligned}$$

证明

- 由于谱范数是向量2范数诱导出的算子范数，所以存在x, $\|x\|_2 = 1$, 使得$\|A^{-1}x\|_2 = \|A^{-1}\|_2$
- 令$y = \dfrac{A^{-1}x}{\|A^{-1}x\|_2}$, $\delta A = -\dfrac{xy^T}{\|A^{-1}\|_2}$ 则$\|y\|_2 = 1$而且
$$\begin{aligned}(A+\delta A)y &= Ay + (\delta A)y \\ &= \frac{x}{\|A^{-1}x\|_2} - \frac{x}{\|A^{-1}\|_2} \\ &= 0\end{aligned}$$
- 这说明A的秩1校正$A+\delta A$奇异

证明

- 注意到

$$\begin{aligned}
\|\delta A\|_2 &= \max_{\|z\|_2=1} \left\| \frac{xy^T}{\|A^{-1}\|_2} z \right\|_2 \\
&= \frac{\|x\|_2}{\|A^{-1}\|_2} \max_{\|z\|_2=1} |y^T z| \\
&= \frac{1}{\|A^{-1}\|_2}
\end{aligned}$$

证明

- 注意到

$$\begin{aligned}\|\delta A\|_2 &= \max_{\|z\|_2=1}\left\|\frac{xy^T}{\|A^{-1}\|_2}z\right\|_2 \\ &= \frac{\|x\|_2}{\|A^{-1}\|_2}\max_{\|z\|_2=1}|y^Tz| \\ &= \frac{1}{\|A^{-1}\|_2}\end{aligned}$$

- 这说明$\|\delta A\|_2 = \|A^{-1}\|_2^{-1}$

浮点数

- 计算机中浮点数f可表示为

$$f = \pm w \times \beta^J, \quad L \leqslant J \leqslant U$$

其中β是机器所用浮点数的基底，
J是阶，w是尾数

浮点数

- 计算机中浮点数f可表示为

$$f = \pm w \times \beta^J, \quad L \leqslant J \leqslant U$$

其中β是机器所用浮点数的基底，J是阶，w是尾数

- 尾数的一般形式为$w = 0.d_1 d_2 \cdots d_t$，其中t是尾数位数，称为字长，$0 \leqslant d_i < \beta$.

浮点数

- 计算机中浮点数f可表示为

$$f = \pm w \times \beta^J, \quad L \leqslant J \leqslant U$$

其中β是机器所用浮点数的基底，J是阶，w是尾数

- 尾数的一般形式为$w = 0.d_1 d_2 \cdots d_t$，其中t是尾数位数，称为字长，$0 \leqslant d_i < \beta$.
- 若$d_1 \neq 0$，称该浮点数为规格化的

浮点数全体

- 计算机系统的浮点数全体是

$$\mathcal{F} = \{0\} \cup \{f : f = \pm 0.d_1 d_2 \cdots d_t \times \beta^J, \\ 0 \leqslant d_i < \beta, \\ d_1 \neq 0, L \leqslant J \leqslant U\}$$

浮点数全体

- 计算机系统的浮点数全体是

$$\mathcal{F} = \{0\} \cup \{f : f = \pm 0.d_1 d_2 \cdots d_t \times \beta^J, \quad 0 \leqslant d_i < \beta, \quad d_1 \neq 0, L \leqslant J \leqslant U\}$$

- 可以用四元数组(β, t, L, U)描述$\mathcal{F}$. 典型取值是$(2, 56, -63, 64)$

- $\mathcal{F}$中元素个数为$2(\beta-1)\beta^{t-1}(U-L+1)+1$, 它们对称地分布在$[m,M]$和$[-M,-m]$中，这里$m=\beta^{L-1}$, $M=\beta^{U}(1-\beta^{-t})$

- $\mathcal{F}$中元素个数为$2(\beta-1)\beta^{t-1}(U-L+1)+1$, 它们对称地分布在$[m, M]$和$[-M, -m]$中，这里$m=\beta^{L-1}$, $M=\beta^U(1-\beta^{-t})$
  - 最小正数为$0.1\times\beta^L=\beta^{L-1}$, 其中0.1为$\beta$进制数

- $\mathcal{F}$中元素个数为$2(\beta-1)\beta^{t-1}(U-L+1)+1$, 它们对称地分布在$[m,M]$和$[-M,-m]$中，这里$m=\beta^{L-1}$, $M=\beta^U(1-\beta^{-t})$
  - 最小正数为$0.1\times\beta^L=\beta^{L-1}$, 其中0.1为$\beta$进制数
  - 最大正数为$0.\underbrace{(\beta-1)\cdots(\beta-1)}_{t位}\times\beta^U$

- $\mathcal{F}$中元素个数为$2(\beta-1)\beta^{t-1}(U-L+1)+1$, 它们对称地分布在$[m,M]$和$[-M,-m]$中，这里$m=\beta^{L-1}$, $M=\beta^U(1-\beta^{-t})$
  - 最小正数为$0.1\times\beta^L=\beta^{L-1}$, 其中0.1为$\beta$进制数
  - 最大正数为$0.\underbrace{(\beta-1)\cdots(\beta-1)}_{t位}\times\beta^U$
- 这些数在区间中非均匀分布

浮点数表示能力的局限

- $\mathcal{F}$是有限集，所以不能表示$[m, M]$, $[-M, -m]$中所有实数

浮点数表示能力的局限

- $\mathcal{F}$是有限集，所以不能表示$[m, M]$，$[-M, -m]$中所有实数
  - 0.584635无法用4位10进制浮点数表示

浮点数表示能力的局限

- $\mathcal{F}$是有限集，所以不能表示$[m, M]$，$[-M, -m]$中所有实数
  - 0.584635无法用4位10进制浮点数表示
  - 两个t位浮点数的乘积一般需要$2t$位浮点数

浮点数表示能力的局限

- $\mathcal{F}$是有限集，所以不能表示$[m, M]$，$[-M, -m]$中所有实数
  - 0.584635无法用4位10进制浮点数表示
  - 两个t位浮点数的乘积一般需要$2t$位浮点数
- 因此舍入误差不可避免

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$
- 若$x = 0$, 则$\mathrm{fl}(x) = 0$

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$
- 若$x = 0$, 则$\mathrm{fl}(x) = 0$
- 若$m \leqslant |x| \leqslant M$,

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$
- 若$x = 0$, 则$\mathrm{fl}(x) = 0$
- 若$m \leqslant |x| \leqslant M$,
  - 舍入法：取$\mathrm{fl}(x)$为$\mathcal{F}$中最接近于x的数

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$
- 若$x = 0$, 则$\mathrm{fl}(x) = 0$
- 若$m \leqslant |x| \leqslant M$,
  - 舍入法：取$\mathrm{fl}(x)$为$\mathcal{F}$中最接近于x的数
  - 截断法：取$\mathrm{fl}(x)$为$\mathcal{F}$中满足$|f| \leqslant |x|$且最接近于x的数

浮点数的选择

- 记实数x的浮点数表示为$\mathrm{fl}(x)$
- 若$x = 0$, 则$\mathrm{fl}(x) = 0$
- 若$m \leqslant |x| \leqslant M$,
  - 舍入法：取$\mathrm{fl}(x)$为$\mathcal{F}$中最接近于x的数
  - 截断法：取$\mathrm{fl}(x)$为$\mathcal{F}$中满足$|f| \leqslant |x|$且最接近于x的数
  - 例：$(\beta, t, L, U) = (10, 3, 0, 2)$, $x = 5.45627$, 舍入法时$\mathrm{fl}(x) = 0.546 \times 10$, 截断法时$\mathrm{fl}(x) = 0.545 \times 10$

机器精度与相对误差

- 定义机器精度**u**为

$$\mathbf{u} = \begin{cases} \beta^{1-t}/2 & \text{舍入法} \\ \beta^{1-t} & \text{截断法} \end{cases}$$

机器精度与相对误差

- 定义机器精度**u**为

$$\mathbf{u} = \begin{cases} \beta^{1-t}/2 & \text{舍入法} \\ \beta^{1-t} & \text{截断法} \end{cases}$$

- 由于$t \geqslant 1$, 我们有$\mathbf{u} \in (0, 1]$

机器精度与相对误差

- 定义机器精度$\mathbf{u}$为

$$\mathbf{u} = \begin{cases} \beta^{1-t}/2 & \text{舍入法} \\ \beta^{1-t} & \text{截断法} \end{cases}$$

- 由于$t \geqslant 1$, 我们有$\mathbf{u} \in (0, 1]$

机器精度与相对误差

- 定义机器精度$\mathbf{u}$为

$$\mathbf{u} = \begin{cases} \beta^{1-t}/2 & \text{舍入法} \\ \beta^{1-t} & \text{截断法} \end{cases}$$

- 由于$t \geqslant 1$, 我们有$\mathbf{u} \in (0, 1]$

定理

设$m \leqslant |x| \leqslant M$, 则

$$\mathrm{fl}(x) = x(1+\delta), \quad |\delta| \leqslant \mathbf{u}$$

证明

- 不妨设$x > 0$, 取α使得

$$\beta^{\alpha-1} \leqslant x < \beta^{\alpha}$$

证明

- 不妨设$x > 0$, 取α使得

$$\beta^{\alpha-1} \leqslant x < \beta^{\alpha}$$

- 在$[\beta^{\alpha-1}, \beta^{\alpha})$中浮点数的阶为$\alpha$, 所以在这个区间中所有$t$位的浮点数以间距$\beta^{\alpha-t}$均匀分布

证明

- 对于舍入法，

$$|\mathrm{fl}(x)-x| \leqslant \frac{1}{2}\beta^{\alpha-t} = \frac{1}{2}\beta^{\alpha-1}\beta^{1-t} \leqslant \frac{1}{2}x\beta^{1-t}$$

从而有

$$\frac{|\mathrm{fl}(x)-x|}{x} \leqslant \frac{1}{2}\beta^{1-t}$$

证明

- 对于截断法，

$$|\operatorname{fl}(x)-x| \leqslant \beta^{\alpha-t}=\beta^{\alpha-1}\beta^{1-t} \leqslant x\beta^{1-t}$$

从而有

$$\frac{|\operatorname{fl}(x)-x|}{x} \leqslant \beta^{1-t}$$

注解

- 有时为了使用方便，我们也将上述结果表示为

$$\mathrm{fl}(x) = \frac{x}{1+\delta}, \quad |\delta| \leqslant \mathbf{u}$$

注解

- 有时为了使用方便，我们也将上述结果表示为

$$\mathrm{fl}(x) = \frac{x}{1+\delta}, \quad |\delta| \leqslant \mathbf{u}$$

- 为此，只需要在证明中考虑$|\mathrm{fl}(x) - x|$相对于$\mathrm{fl}(x)$的界即可

- 对于舍入法，

$$\beta^{\alpha-1} \leqslant x < \beta^{\alpha} \Longrightarrow \beta^{\alpha-1} \leqslant \mathrm{fl}(x) \leqslant \beta^{\alpha}$$

- 对于舍入法，
  $\beta^{\alpha-1} \leqslant x < \beta^{\alpha} \Longrightarrow \beta^{\alpha-1} \leqslant \mathrm{fl}(x) \leqslant \beta^{\alpha}$
- 从而
  $$|\,\mathrm{fl}(x) - x| \leqslant \frac{1}{2}\beta^{\alpha-t} = \frac{1}{2}\beta^{\alpha-1}\beta^{1-t} \leqslant \frac{1}{2}\,\mathrm{fl}(x)\beta^{1-t}$$
  从而有
  $$\frac{|\,\mathrm{fl}(x) - x|}{\mathrm{fl}(x)} \leqslant \frac{1}{2}\beta^{1-t}$$

- 对于截断法，
  $$\beta^{\alpha-1} \leqslant x < \beta^{\alpha} \Longrightarrow \beta^{\alpha-1} \leqslant \mathrm{fl}(x) < \beta^{\alpha}$$

- 对于截断法，
  $\beta^{\alpha-1} \leqslant x < \beta^{\alpha} \Longrightarrow \beta^{\alpha-1} \leqslant \mathrm{fl}(x) < \beta^{\alpha}$
- 从而

$$|\,\mathrm{fl}(x) - x| \leqslant \beta^{\alpha-t} = \beta^{\alpha-1}\beta^{1-t} \leqslant \mathrm{fl}(x)\beta^{1-t}$$

从而有

$$\frac{|\,\mathrm{fl}(x) - x|}{\mathrm{fl}(x)} \leqslant \beta^{1-t}$$

基本运算的舍入误差

- 设$a, b \in \mathcal{F}$

基本运算的舍入误差

- 设$a, b \in \mathcal{F}$
- 用$\circ$表示$+, -, \times, \div$中任意一种运算

基本运算的舍入误差

- 设$a, b \in \mathcal{F}$
- 用$\circ$表示$+, -, \times, \div$中任意一种运算
- $\mathrm{fl}(a \circ b)$表示先把a, b看作实数进行运算，然后再按舍入规则把结果转化为浮点数。运算时若$|a \circ b| > M$或$0 < |a \circ b| < m$, 则发生了上溢或下溢

基本运算的舍入误差

- 设$a, b \in \mathcal{F}$
- 用$\circ$表示$+, -, \times, \div$中任意一种运算
- $\mathrm{fl}(a \circ b)$表示先把a, b看作实数进行运算，然后再按舍入规则把结果转化为浮点数。运算时若$|a \circ b| > M$或$0 < |a \circ b| < m$, 则发生了上溢或下溢
- 在不发生溢出时，由前述定理可有

$$\mathrm{fl}(a \circ b) = (a \circ b)(1 + \delta), \quad |\delta| \leqslant \mathbf{u}$$

例

设x, y是两个由浮点数构成的n维向量。试估计$|\mathrm{fl}(x^Ty) - x^Ty|$的上界

- 令$S_k = \mathrm{fl}\left(\sum_{i=1}^{k} x_i y_i\right)$

$$\begin{aligned} S_1 &= x_1y_1(1+\gamma_1), \quad |\gamma_1| \leqslant \mathbf{u} \\ S_k &= \mathrm{fl}(S_{k-1} + \mathrm{fl}(x_ky_k)) \\ &= (S_{k-1} + x_ky_k(1+\gamma_k))(1+\delta_k), \\ &\qquad |\delta_k|, |\gamma_k| \leqslant \mathbf{u} \end{aligned}$$

- 于是有

$$\mathrm{fl}(x^Ty) = S_n = \sum_{i=1}^{n} x_iy_i(1+\gamma_i)\prod_{j=i}^{n}(1+\delta_j)$$
$$= \sum_{i=1}^{n}(1+\varepsilon_i)x_iy_i,$$

其中

$$1+\varepsilon_i = (1+\gamma_i)\prod_{j=i}^{n}(1+\delta_j), \delta_1 = 0$$

- 所以我们得到

$$
\begin{aligned}
|\mathrm{fl}(x^Ty) - x^Ty| &\leqslant \sum_{i=1}^{n} |\varepsilon_i||x_iy_i| \\
&\leqslant 1.01n\mathbf{u} \sum_{i=1}^{n} |x_iy_i|
\end{aligned}
$$

这里我们用到了当给定条件$n\mathbf{u} \leqslant 0.01$时，后面马上要证明的结论

- 所以我们得到

$$
\begin{aligned}
|\mathrm{fl}(x^T y) - x^T y| &\leqslant \sum_{i=1}^{n} |\varepsilon_i||x_i y_i| \\
&\leqslant 1.01n\mathbf{u} \sum_{i=1}^{n} |x_i y_i|
\end{aligned}
$$

这里我们用到了当给定条件$n\mathbf{u} \leqslant 0.01$时，后面马上要证明的结论

- 上式表明若$|x^T y| \ll \sum_{i=1}^{n} |x_i y_i|$, 则$\mathrm{fl}(x^T y)$的相对误差可能会很大

定理

若$|\delta_i| \leqslant \mathbf{u}$且$n\mathbf{u} \leqslant 0.01$, 那么

$$1 - 1.01n\mathbf{u} \leqslant \prod_{i=1}^{n}(1+\delta_i) \leqslant 1 + 1.01n\mathbf{u}$$

或写成

$$\prod_{i=1}^{n}(1+\delta_i) = 1+\delta, \quad |\delta| \leqslant 1.01n\mathbf{u}$$

证明

- 由于$|\delta_i| \leqslant \mathbf{u}$, 所以

$$(1-\mathbf{u})^n \leqslant \prod_{i=1}^{n}(1+\delta_i) \leqslant (1+\mathbf{u})^n$$

证明

- 由于$|\delta_i| \leqslant \mathbf{u}$, 所以

$$(1-\mathbf{u})^n \leqslant \prod_{i=1}^{n}(1+\delta_i) \leqslant (1+\mathbf{u})^n$$

- 由于当$x \in (0,1)$时$1-nx \leqslant (1-x)^n$, 所以

$$1-1.01n\mathbf{u} \leqslant 1-n\mathbf{u} \leqslant (1-\mathbf{u})^n$$

- 注意到e^x的Taylor展开式

$$
\begin{aligned}
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\
&= 1 + x + \frac{x}{2} \cdot x \cdot \left(1 + \frac{x}{3} + \frac{2x^2}{4!} + \cdots\right)
\end{aligned}
$$

- 注意到e^x的Taylor展开式

$$\begin{aligned} e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\ &= 1 + x + \frac{x}{2} \cdot x \cdot \left(1 + \frac{x}{3} + \frac{2x^2}{4!} + \cdots\right) \end{aligned}$$

- 所以当$x \in [0, 0.01]$时，(注：$e^{0.01} \approx 1.01$)

$$1 + x \leqslant e^x \leqslant 1 + x + \frac{0.01}{2} x e^x \leqslant 1 + 1.01x$$

- 在左端不等式中取$x = \mathbf{u}$, 得到

$$(1+u)^n \leqslant e^{n\mathbf{u}}$$

- 在左端不等式中取$x = \mathbf{u}$, 得到

$$(1+u)^n \leqslant e^{n\mathbf{u}}$$

- 在右端不等式中取$x = n\mathbf{u}$,

$$e^{n\mathbf{u}} \leqslant 1 + 1.01n\mathbf{u}$$

- 在左端不等式中取$x = \mathbf{u}$, 得到

$$(1+u)^n \leqslant e^{n\mathbf{u}}$$

- 在右端不等式中取$x = n\mathbf{u}$,

$$e^{n\mathbf{u}} \leqslant 1 + 1.01n\mathbf{u}$$

- 综合得到$(1+\mathbf{u})^n \leqslant 1 + 1.01n\mathbf{u}$

矩阵计算

- 取n阶方阵E, 定义记号$|E| = (|e_{ij}|)$

矩阵计算

- 取n阶方阵E, 定义记号$|E| = (|e_{ij}|)$
- 规定$|E| \leqslant |F| \iff |e_{ij}| \leqslant |f_{ij}|$, $i, j = 1, \ldots, n$

矩阵计算

- 取n阶方阵E, 定义记号$|E| = (|e_{ij}|)$
- 规定$|E| \leqslant |F| \iff |e_{ij}| \leqslant |f_{ij}|$, $i, j = 1, \ldots, n$
- 设$A, B \in \mathcal{F}^{n\times n}$, $\alpha \in \mathcal{F}$, 则

$$\mathrm{fl}(\alpha A) = \alpha A + E, |E| \leqslant \mathbf{u}|\alpha A|,$$

$$\mathrm{fl}(A + B) = (A + B) + E, |E| \leqslant \mathbf{u}|A + B|$$

矩阵乘积

- 基于前面的例子，我们可有

$$\mathrm{fl}(AB) = AB + E, |E| \leqslant 1.01n\mathbf{u}|A||B|$$

矩阵乘积

- 基于前面的例子，我们可有

$$\mathrm{fl}(AB) = AB + E, |E| \leqslant 1.01n\mathbf{u}|A||B|$$

- 注意$|AB|$可能比$|A||B|$小得多

矩阵乘积

- 基于前面的例子，我们可有

$$\mathrm{fl}(AB) = AB + E, |E| \leqslant 1.01n\mathbf{u}|A||B|$$

- 注意$|AB|$可能比$|A||B|$小得多
- 计算内积时，我们通常会先用双精度进行计算，然后再舍入到单精度数

向前误差分析

- 上述舍入误差界，是通过估计计算解与精确解之间的误差得到

向前误差分析

- 上述舍入误差界，是通过估计计算解与精确解之间的误差得到
- 界与精确解有关

向前误差分析

- 上述舍入误差界，是通过估计计算解与精确解之间的误差得到
- 界与精确解有关
- 这种误差分析的方法称为向前误差分析法

再议矩阵乘法

- 假设A, B是两个2×2的上三角阵，则

$$\mathrm{fl}(AB) = \begin{pmatrix} a_{11}b_{11}(1+\varepsilon_1) & \tilde{a}_{12} \\ 0 & a_{22}b_{22}(1+\varepsilon_5) \end{pmatrix}$$

其中

$$\tilde{a}_{12} = (a_{11}b_{12}(1+\varepsilon_2) + a_{12}b_{22}(1+\varepsilon_3))(1+\varepsilon_4),$$

$$|\varepsilon_i| \leqslant \mathbf{u}, \quad i = 1, 2, 3, 4, 5$$

- 令

$$\tilde{A} = \begin{pmatrix} a_{11} & a_{12}(1+\varepsilon_3)(1+\varepsilon_4) \\ 0 & a_{22}(1+\varepsilon_5) \end{pmatrix}$$

$$\tilde{B} = \begin{pmatrix} b_{11}(1+\varepsilon_1) & b_{12}(1+\varepsilon_2)(1+\varepsilon_4) \\ 0 & b_{22} \end{pmatrix}$$

- 令

$$\tilde{A} = \begin{pmatrix} a_{11} & a_{12}(1+\varepsilon_3)(1+\varepsilon_4) \\ 0 & a_{22}(1+\varepsilon_5) \end{pmatrix}$$

$$\tilde{B} = \begin{pmatrix} b_{11}(1+\varepsilon_1) & b_{12}(1+\varepsilon_2)(1+\varepsilon_4) \\ 0 & b_{22} \end{pmatrix}$$

- 则可证$\mathrm{fl}(AB) = \tilde{A}\tilde{B}$, 其中$\tilde{A} = A + E$, $|E| \leqslant 3\mathbf{u}|A|$; $\tilde{B} = B + F$, $|F| \leqslant 3\mathbf{u}|B|$

向后误差分析

- $\mathrm{fl}(AB)$是有了微小扰动的两个矩阵$\tilde{A}$, $\tilde{B}$的精确乘积

向后误差分析

- $\mathrm{fl}(AB)$是有了微小扰动的两个矩阵$\tilde{A}$, $\tilde{B}$的精确乘积
- 这种把计算过程产生的误差归结为具有误差的原始数据的精确运算的分析方法称为向后误差分析法

向后误差分析

- $\mathrm{fl}(AB)$是有了微小扰动的两个矩阵$\tilde{A}$, $\tilde{B}$的精确乘积
- 这种把计算过程产生的误差归结为具有误差的原始数据的精确运算的分析方法称为向后误差分析法
- 这种方法把浮点数的运算转化为实数的精确运算，从而可以在分析过程中便利地应用实数的代数运算法则

列主元Gauss消去法的舍入误差分析

本节我们利用浮点数基本运算的舍入误差理论对求解线性方程组的列主元Gauss消去法进行向后误差分析

- 将证明用列主元Gauss消去法求解$Ax = b$时，它的计算解$\tilde{x}$满足

$$(A + E)\tilde{x} = b$$

列主元Gauss消去法的舍入误差分析

本节我们利用浮点数基本运算的舍入误差理论对求解线性方程组的列主元Gauss消去法进行向后误差分析

- 将证明用列主元Gauss消去法求解$Ax = b$时，它的计算解$\tilde{x}$满足

$$(A + E)\tilde{x} = b$$

- 并且给出误差矩阵E的上界估计

三角分解时的舍入误差

定理

设$n \times n$浮点数矩阵$A = (a_{ij})$有三角分解，且$1.01n\mathbf{u} \leqslant 0.01$，则用Gauss消去法计算得到的单位下三角阵$\tilde{L}$和上三角阵$\tilde{U}$满足

$$\tilde{L}\tilde{U} = A + E,$$

其中$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$

证明之$\widetilde{U}$的计算

- 设$\tilde{U} = (\tilde{u}_{ij})$, $\tilde{L} = (\tilde{\ell}_{ij})$, 则$a_{ij} = \sum_{k=1}^{\min(i,j)} \ell_{ik} u_{kj}$

证明之$\widetilde{U}$的计算

- 设$\tilde{U}=(\tilde{u}_{ij})$, $\tilde{L}=(\tilde{\ell}_{ij})$, 则$a_{ij}=\sum_{k=1}^{\min(i,j)}\ell_{ik}u_{kj}$
- 由Gauss消去法可知$\tilde{u}_{ij}(i\leqslant j)$是从$a_{ij}$中依次减去$\tilde{\ell}_{ik}\tilde{u}_{kj}$而得到的，$k=1,\ldots,i-1$,即

$$
\begin{aligned}
a_{ij}^{(0)} &= a_{ij},\\
a_{ij}^{(k)} &= \mathrm{fl}(a_{ij}^{(k-1)}-\mathrm{fl}(\tilde{\ell}_{ik}\tilde{u}_{kj})), k=1,\ldots,i-2,\\
\tilde{u}_{ij} &= a_{ij}^{(i-1)}=\mathrm{fl}(a_{ij}^{(i-2)}-\mathrm{fl}(\tilde{\ell}_{i,i-1}\tilde{u}_{i-1,i}))
\end{aligned}
$$

证明之$\widetilde{U}$的计算

- 由基本运算舍入误差分析的结果可知

$$a_{ij}^{(k)} = (a_{ij}^{(k-1)} - (\tilde{\ell}_{i,k}\tilde{u}_{k,j})(1+\gamma_k))(1+\varepsilon_k)$$

其中$|\gamma_k|, |\varepsilon_k| \leqslant \mathbf{u}$

证明之$\widetilde{U}$的计算

- 由基本运算舍入误差分析的结果可知

$$a_{ij}^{(k)} = (a_{ij}^{(k-1)} - (\tilde{\ell}_{i,k}\tilde{u}_{k,j})(1+\gamma_k))(1+\varepsilon_k)$$

其中$|\gamma_k|, |\varepsilon_k| \leqslant \mathbf{u}$

- 从而仿照计算内积的例题所示，我们有

$$\tilde{u}_{ij} = a_{ij}(1+\delta_i) - \sum_{k=1}^{i-1}(\tilde{\ell}_{ik}\tilde{u}_{kj})(1+\delta_k)$$

其中$|\delta_k| \leqslant 1.01n\mathbf{u}$, $k=1,\dots,i$

证明之$\widetilde{U}$的计算

- 由前式解出a_{ij}：

$$a_{ij} = \frac{\tilde{u}_{ij}}{1+\delta_i} + \sum_{k=1}^{i-1}(\tilde{\ell}_{ik}\tilde{u}_{kj})\frac{1+\delta_k}{1+\delta_i}$$

$$= \sum_{k=1}^{i}\tilde{\ell}_{ik}\tilde{u}_{kj} - e_{ij}$$

其中$\tilde{\ell}_{ii} = 1$,

$$e_{ij} = (\tilde{\ell}_{ii}\tilde{u}_{ij})\frac{\delta_i}{1+\delta_i} + \sum_{k=1}^{i-1}(\tilde{\ell}_{ik}\tilde{u}_{kj})\frac{\delta_i-\delta_k}{1+\delta_i}$$

证明之$\widetilde{U}$的计算

- 注意到$|\delta_k| \leqslant 1.01n\mathbf{u} < 0.01$，我们有

$$
\begin{aligned}
|e_{ij}| &\leqslant \frac{2.02n\mathbf{u}}{1-0.01}\sum_{k=1}^{i}|\tilde{\ell}_{ik}||\tilde{u}_{kj}| \\
&\leqslant 2.05n\mathbf{u}\sum_{k=1}^{i}|\tilde{\ell}_{ik}||\tilde{u}_{kj}|
\end{aligned}
$$

证明之$\tilde{L}$的计算

- 由Gauss消去法的具体实现可知，$\tilde{\ell}_{ij}(i > j)$的计算过程为

$$
\begin{aligned}
a_{ij}^{(0)} &= a_{ij}, \\
a_{ij}^{(k)} &= \mathrm{fl}(a_{ij}^{(k-1)} - \mathrm{fl}(\tilde{\ell}_{ik}\tilde{u}_{kj})), k = 1, \ldots, j-1 \\
\tilde{\ell}_{ij} &= \mathrm{fl}(a_{ij}^{(j-1)}/\tilde{u}_{jj})
\end{aligned}
$$

证明之$\widetilde{L}$的计算

- 由基本运算舍入误差分析的结果可知

$$a_{ij}^{(k)} = (a_{ij}^{(k-1)} - (\tilde{\ell}_{ik}\tilde{u}_{kj})(1+\gamma_k))(1+\varepsilon_k)$$
$$\tilde{\ell}_{ij} = a_{ij}^{(j-1)}/(\tilde{u}_{jj}(1+\delta))$$

其中$|\delta|, |\gamma_k|, |\varepsilon_k| \leqslant \mathbf{u}$

证明之$\widetilde{L}$的计算

- 由此可类似证明

$$a_{ij} = \sum_{k=1}^{j} \tilde{\ell}_{ik}\tilde{u}_{kj} - e_{ij},$$

其中

$$|e_{ij}| \leqslant 2.05 n\mathbf{u} \sum_{k=1}^{j} |\tilde{\ell}_{ik}||\tilde{u}_{kj}|$$

有行列交换时的情形

定理

设$n\times n$浮点数矩阵$A=(a_{ij})$非奇异，且$1.01n\mathbf{u}\leqslant 0.01$，则用列主元Gauss消去法计算得到的单位下三角阵$\tilde{L}$，上三角阵$\tilde{U}$以及排列方阵$\tilde{P}$满足$\tilde{L}\tilde{U}=\tilde{P}A+E$，其中$|E|\leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$

注意：行列交换并不引入舍入误差

求解三角形方程组的舍入误差

引理

设$n \times n$浮点数三角阵S非奇异，且$1.01n\mathbf{u} \leqslant 0.01$，则用前一章方法求解$Sx = b$得到的计算解$\tilde{x}$满足

$$(S + H)\tilde{x} = b$$

其中$|H| \leqslant 1.01n\mathbf{u}|S|$

证明

- 不妨设S是下三角阵

证明

- 不妨设S是下三角阵
- 采用归纳法：对n进行归纳

证明

- 不妨设S是下三角阵
- 采用归纳法：对n进行归纳
- $n = 1$时成立

证明

- 不妨设S是下三角阵
- 采用归纳法：对n进行归纳
- $n = 1$时成立
- 假设$n-1$阶时结论成立

证明

- 采用前代法求解$Lx = b$，得到计算解$\tilde{x}$

证明

- 采用前代法求解$Lx = b$，得到计算解$\tilde{x}$
- 对b, $\tilde{x}$进行分块：$b = (b_1, c^T)^T$, $\tilde{x} = (\tilde{x}_1, \tilde{y}^T)^T$, 其中$b_1$, $\tilde{x}_1 \in \mathcal{F}$, $c, \tilde{y}$为列向量;

证明

线性方程组的敏度分析与消去法的舍入误差分析

邓建松

向量与矩阵的范数

线性方程组的敏度分析

基本运算的舍入误差分析

列主元Gauss消去法的舍入误差分析

计算解的精度估计和迭代改进

- 采用前代法求解$Lx=b$，得到计算解$\tilde{x}$
- 对b, $\tilde{x}$进行分块：$b=(b_1,c^T)^T$, $\tilde{x}=(\tilde{x}_1,\tilde{y}^T)^T$, 其中$b_1$, $\tilde{x}_1\in\mathcal{F}$, $c,\tilde{y}$为列向量;
- 对L进行相应分块：

$$L=\begin{pmatrix} \ell_{11} & 0 \\ \ell_1 & L_1 \end{pmatrix}$$

- 我们有

$$\tilde{x}_1 = \mathrm{fl}(b_1/\ell_{11}) = \frac{b_1}{\ell_{11}(1+\delta_1)}, |\delta_1| \leqslant \mathbf{u}$$

- 我们有

$$\tilde{x}_1 = \mathrm{fl}(b_1/\ell_{11}) = \frac{b_1}{\ell_{11}(1+\delta_1)}, |\delta_1| \leqslant \mathbf{u}$$

- 而$\tilde{y}$是用前代法求解下述$n-1$阶线性方程组所得到的计算解

$$L_1 y = \mathrm{fl}(c - \tilde{x}_1 \ell_1)$$

- 我们有

$$\tilde{x}_1 = \mathrm{fl}(b_1/\ell_{11}) = \frac{b_1}{\ell_{11}(1+\delta_1)}, |\delta_1| \leqslant \mathbf{u}$$

- 而$\tilde{y}$是用前代法求解下述$n-1$阶线性方程组所得到的计算解

$$L_1 y = \mathrm{fl}(c - \tilde{x}_1 \ell_1)$$

- 由归纳假设，$(L_1 + H_1)\tilde{y} = \mathrm{fl}(c - \tilde{x}_1\ell_1)$，这里$|H_1| \leqslant 1.01(n-1)\mathbf{u}|L_1|$

- 根据基本运算的舍入误差结果，我们有

$$\begin{aligned} \mathrm{fl}(c - \tilde{x}_1\ell_1) &= \mathrm{fl}(c - \mathrm{fl}(\tilde{x}_1\ell_1)) \\ &= (I + D_\gamma)^{-1}(c - \tilde{x}_1\ell_1 - \tilde{x}_1 D_\delta \ell_1), \end{aligned}$$

  其中

$$\begin{aligned} D_\gamma &= \mathrm{diag}(\gamma_2, \ldots, \gamma_n), \\ D_\delta &= \mathrm{diag}(\delta_2, \ldots, \delta_n), \\ |\gamma_i|, |\delta_i| &\leqslant \mathbf{u}, \quad i = 2, 3, \ldots, n \end{aligned}$$

- 于是

$$\tilde{x}_1\ell_1 + \tilde{x}_1 D_\delta \ell_1 + (I + D_\gamma)(L_1 + H_1)\tilde{y} = c$$

- 于是

$$\tilde{x}_1\ell_1 + \tilde{x}_1 D_\delta \ell_1 + (I + D_\gamma)(L_1 + H_1)\tilde{y} = c$$

- 从而有

$$(L + H)\tilde{x} = b$$

其中

$$H = \begin{pmatrix} \delta_1 \ell_{11} & 0 \\ D_\delta \ell_1 & H_1 + D_\gamma(L_1 + H_1) \end{pmatrix}$$

$$
\begin{aligned}
|H| &\leqslant \begin{pmatrix} |\delta_1||\ell_{11}| & 0 \\ |D_\delta||\ell_1| & |H_1| + |D_\gamma|(|L_1| + |H_1|) \end{pmatrix} \\
&\leqslant \begin{pmatrix} \mathbf{u}|\ell_{11}| & 0 \\ \mathbf{u}|\ell_1| & |H_1| + \mathbf{u}(|L_1| + |H_1|) \end{pmatrix} \\
&\leqslant \mathbf{u} \begin{pmatrix} |\ell_{11}| & 0 \\ |\ell_1| & (1.01(n-1) + 1 + 1.01(n-1)\mathbf{u})|L_1| \end{pmatrix} \\
&\leqslant 1.01 n\mathbf{u}|L|
\end{aligned}
$$

最后一个不等式中用到了条件$1.01n\mathbf{u} \leqslant 0.01$

完整求解

- 把上述引理应用于三角形方程组

$$\tilde{L}y = \tilde{P}b, \quad \tilde{U}x = y$$

可知最后得到的计算解$\tilde{x}$应满足

$$(\tilde{L} + F)(\tilde{U} + G)\tilde{x} = \tilde{P}b$$

即

$$(\tilde{L}\tilde{U} + F\tilde{U} + \tilde{L}G + FG)\tilde{x} = \tilde{P}b$$

其中$|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|$, $|G| \leqslant 1.01n\mathbf{u}|\tilde{U}|$

舍入误差

- 把$\tilde{L}\tilde{U} = \tilde{P}A + E$代入得

$$(A + \delta A)\tilde{x} = b$$

其中

$$\delta A = \tilde{P}^T(E + F\tilde{U} + \tilde{L}G + FG)$$

舍入误差

- 把$\tilde{L}\tilde{U} = \tilde{P}A + E$代入得

$$(A + \delta A)\tilde{x} = b$$

其中

$$\delta A = \tilde{P}^T(E + F\tilde{U} + \tilde{L}G + FG)$$

- 由已有结论，我们有

$$|\delta A| \leqslant 4.09n\mathbf{u}\tilde{P}^T|\tilde{L}||\tilde{U}|$$

4.09的来源

条件$1.01n\mathbf{u} \leqslant 0.01$

- E来源于$\tilde{L}\tilde{U} = \tilde{P}A + E$, 而且$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$

4.09的来源

条件$1.01n\mathbf{u} \leqslant 0.01$

- E来源于$\tilde{L}\tilde{U} = \tilde{P}A + E$, 而且$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$
- $|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|$

4.09的来源

条件$1.01n\mathbf{u} \leqslant 0.01$

- E来源于$\tilde{L}\tilde{U} = \tilde{P}A + E$, 而且$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$
- $|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|$
- $|G| \leqslant 1.01n\mathbf{u}|\tilde{U}|$

4.09的来源

条件$1.01n\mathbf{u} \leqslant 0.01$

- E来源于$\tilde{L}\tilde{U} = \tilde{P}A + E$, 而且$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$
- $|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|$
- $|G| \leqslant 1.01n\mathbf{u}|\tilde{U}|$
- $|FG| \leqslant 0.01(1.01n\mathbf{u})|\tilde{L}||\tilde{U}| \leqslant 0.02n\mathbf{u}|\tilde{L}||\tilde{U}|$

4.09的来源

条件$1.01n\mathbf{u} \leqslant 0.01$

- E来源于$\tilde{L}\tilde{U} = \tilde{P}A + E$, 而且$|E| \leqslant 2.05n\mathbf{u}|\tilde{L}||\tilde{U}|$
- $|F| \leqslant 1.01n\mathbf{u}|\tilde{L}|$
- $|G| \leqslant 1.01n\mathbf{u}|\tilde{U}|$
- $|FG| \leqslant 0.01(1.01n\mathbf{u})|\tilde{L}||\tilde{U}| \leqslant 0.02n\mathbf{u}|\tilde{L}||\tilde{U}|$
- $2.05 + 1.01 + 1.01 + 0.02 = 4.09$

增长因子

- 由于$\tilde{L}$的元素的绝对值不超过1，所以$\|\tilde{L}\|_\infty \leqslant n$

增长因子

- 由于$\tilde{L}$的元素的绝对值不超过1，所以$\|\tilde{L}\|_\infty \leqslant n$
- 为了估计$\|\tilde{U}\|_\infty$, 定义

$$\rho = \max_{i,j} |\tilde{u}_{ij}| / \max_{i,j} |a_{ij}|$$

称之为列主元Gauss消去法的增长因子

增长因子

- 由于$\tilde{L}$的元素的绝对值不超过1，所以$\|\tilde{L}\|_\infty \leqslant n$
- 为了估计$\|\tilde{U}\|_\infty$, 定义

$$\rho = \max_{i,j} |\tilde{u}_{ij}| / \max_{i,j} |a_{ij}|$$

称之为列主元Gauss消去法的增长因子

- 于是我们有

$$\|\tilde{U}\|_\infty \leqslant n \max_{i,j} |\tilde{u}_{ij}| = n\rho \max_{i,j} |a_{ij}| \leqslant n\rho \|A\|_\infty$$

Gauss消去法的舍入误差定理

定理

设$n \times n$浮点数矩阵A非奇异，且$1.01n\mathbf{u} \leqslant 0.01$，则用列主元Gauss消去法求解线性方程组$Ax = b$所得到的计算解$\tilde{x}$满足

$$(A + \delta A)\tilde{x} = b$$

其中

$$\|\delta A\|_\infty / \|A\|_\infty \leqslant 4.09 n^3 \rho \mathbf{u}$$

注解

- 求解过程中引入舍入误差产生的计算解相当于系数矩阵作某些扰动而得到的扰动方程组的精确解

注解

- 求解过程中引入舍入误差产生的计算解相当于系数矩阵作某些扰动而得到的扰动方程组的精确解
- 通常δA的元素比A的元素的初始误差小得多，从这个意义上说，列主元Gauss消去法是数值稳定的

注解

- 理论上可以证明$\rho \leqslant 2^{n-1}$, 而且上界可以达到；实际中常遇到的情形是ρ不会超过n

注解

- 理论上可以证明$\rho \leqslant 2^{n-1}$, 而且上界可以达到；实际中常遇到的情形是ρ不会超过n
- 定理中的上界比真正的$\|\delta A\|_\infty / \|A\|_\infty$大很多，而实际计算中几乎都有$\|\delta A\|_\infty / \|A\|_\infty \approx \mathbf{u}$

精度估计

- 设采用某种方法求解$Ax = b$得到的计算解为$\tilde{x}$

精度估计

- 设采用某种方法求解$Ax = b$得到的计算解为$\tilde{x}$
- 令$r = b - A\tilde{x}$, 则有

$$r = Ax - A\tilde{x} = A(x - \tilde{x})$$

精度估计

- 设采用某种方法求解$Ax = b$得到的计算解为$\tilde{x}$
- 令$r = b - A\tilde{x}$, 则有

$$r = Ax - A\tilde{x} = A(x - \tilde{x})$$

- 从而

$$\|x - \tilde{x}\| = \|A^{-1}r\| \leqslant \|A^{-1}\|\|r\|$$

精度估计

- 注意到$\|b\| \leqslant \|A\|\|x\|$, 所以我们有

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leqslant \|A^{-1}\|\|A\|\frac{\|r\|}{\|b\|}$$

精度估计

- 注意到$\|b\| \leqslant \|A\|\|x\|$, 所以我们有

$$\frac{\|x-\tilde{x}\|}{\|x\|} \leqslant \|A^{-1}\|\|A\|\frac{\|r\|}{\|b\|}$$

- 在上式中取∞范数，有

$$\frac{\|x-\tilde{x}\|_\infty}{\|x\|_\infty} \leqslant \kappa_\infty(A)\frac{\|r\|_\infty}{\|b\|_\infty}$$

精度估计公式

- 通过计算$\kappa_\infty(A)\frac{\|r\|_\infty}{\|b\|_\infty}$估计计算解的精度

精度估计公式

- 通过计算$\kappa_\infty(A)\dfrac{\|r\|_\infty}{\|b\|_\infty}$估计计算解的精度
- 这个公式中除$\kappa_\infty(A)$外，其余量都容易计算

精度估计公式

- 通过计算$\kappa_\infty(A)\dfrac{\|r\|_\infty}{\|b\|_\infty}$估计计算解的精度
- 这个公式中除$\kappa_\infty(A)$外，其余量都容易计算
- 而$\|A\|_\infty$是容易计算的，问题就余下了如何计算$\|A^{-1}\|_\infty$，或者如何计算$\|A^{-T}\|_1$

多元函数极值问题

- 为了计算$\|B\|_1$, 我们在区域$D = \{x \in \mathbb{R}^n \colon \|x\|_1 \leqslant 1\}$上考虑多元函数极值问题

$$\max f(x) = \|Bx\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} b_{ij} x_j \right|$$

多元函数极值问题

- 为了计算$\|B\|_1$, 我们在区域$\mathcal{D} = \{x \in \mathbb{R}^n\colon \|x\|_1 \leqslant 1\}$上考虑多元函数极值问题

$$\max f(x) = \|Bx\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} b_{ij} x_j \right|$$

- 易证f是凸函数，$\mathcal{D}$是凸集。从而最大值一定在边界上达到，但可能是局部极大值

多元函数极值问题

- 为了计算$\|B\|_1$, 我们在区域$\mathcal{D} = \{x \in \mathbb{R}^n : \|x\|_1 \leqslant 1\}$上考虑多元函数极值问题

$$\max f(x) = \|Bx\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} b_{ij} x_j \right|$$

- 易证f是凸函数，$\mathcal{D}$是凸集。从而最大值一定在边界上达到，但可能是局部极大值
- 可以应用著名的“盲人爬山法”求解这个优化问题

梯度

- 对于$x_0 = (x_j^{(0)}) \in \mathbf{R}^n$, $\|x_0\|_1 = 1$满足

$$\sum_{j=1}^{n} b_{ij} x_j^{(0)} \neq 0, \quad i = 1, \ldots, n$$

则f在$x = x_0$点的梯度$\nabla f(x_0)$存在

梯度

- 对于$x_0 = (x_j^{(0)}) \in \mathbf{R}^n$, $\|x_0\|_1 = 1$满足
$$\sum_{j=1}^{n} b_{ij} x_j^{(0)} \neq 0, \quad i = 1, \ldots, n$$
则f在$x = x_0$点的梯度$\nabla f(x_0)$存在
- 由凸函数的性质，
$$f(y) \geqslant f(x_0) + \nabla f(x_0)(y - x), y \in \mathbf{R}^n$$

梯度计算

- 取$\xi_i = \operatorname{sgn}\left(\sum_{j=1}^{n} b_{ij} x_j^{(0)}\right)$,
  $v = (\xi_1, \ldots, \xi_n)^T$

梯度计算

- 取$\xi_i = \mathrm{sgn}\left(\sum_{j=1}^{n} b_{ij} x_j^{(0)}\right)$,
  $v = (\xi_1, \ldots, \xi_n)^T$
- 在x_0附近$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} \xi_i b_{ij} x_j$

梯度计算

- 取$\xi_i = \operatorname{sgn}\left(\sum_{j=1}^{n} b_{ij} x_j^{(0)}\right)$,
  $v = (\xi_1, \ldots, \xi_n)^T$
- 在x_0附近$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} \xi_i b_{ij} x_j$
- $\nabla f(x_0) = v^T B = (B^T v)^T$

梯度计算

- 取$\xi_i = \operatorname{sgn}\left(\sum_{j=1}^{n} b_{ij} x_j^{(0)}\right)$,
  $v = (\xi_1, \ldots, \xi_n)^T$
- 在x_0附近$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} \xi_i b_{ij} x_j$
- $\nabla f(x_0) = v^T B = (B^T v)^T$
- 定义$w = Bx_0$, $z = B^T v$

定理

定理

B, x_0, v, w, z如前所述。则有

- 若$\|z\|_\infty \leqslant z^T x_0$, 则$\|w\|_1 = \|Bx_0\|_1$是$f(x)$在$\mathcal{D}$中的局部极大值
- 若$\|z\|_\infty > z^T x_0$, 则$\|Be_j\|_1 > \|Bx_0\|_1$, 其中j满足$|z_j| = \|z\|_\infty$

证明：当$\|z\|_\infty \leqslant z^T x_0$时

- 由于在x_0附近f是线性函数，所以

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0)$$

证明：当$\|z\|_\infty \leqslant z^T x_0$时

- 由于在x_0附近f是线性函数，所以

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0)$$

- 因此我们只需证在x_0附近

$$\nabla f(x_0)(x - x_0) = z^T (x - x_0) \leqslant 0$$

证明：当$\|z\|_\infty \leqslant z^T x_0$时

- 由于在x_0附近f是线性函数，所以

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0)$$

- 因此我们只需证在x_0附近

$$\nabla f(x_0)(x - x_0) = z^T(x - x_0) \leqslant 0$$

- 实际上对于$\|x\|_1 \leqslant 1$我们有

$$\begin{aligned} z^T(x - x_0) &= z^T x - z^T x_0 \leqslant \|z\|_\infty \|x\|_1 - z^T x_0 \\ &\leqslant \|z\|_\infty - z^T x_0 \leqslant 0 \end{aligned}$$

证明：当$\|z\|_\infty > z^T x_0$时

- 取$\tilde{x} = e_j \operatorname{sgn}(z_j)$, 则有

$$
\begin{aligned}
\|Be_j\|_1 = \|B\tilde{x}\|_1 &= f(\tilde{x}) \\
&\geqslant f(x_0) + \nabla f(x_0)(\tilde{x} - x_0) \\
&= \|Bx_0\|_1 + z^T\tilde{x} - z^T x_0 \\
&= \|Bx_0\|_1 + |z_j| - z^T x_0 \\
&= \|Bx_0\|_1 + \|z\|_\infty - z^T x_0 \\
&> \|Bx_0\|_1
\end{aligned}
$$

$\|A^{-1}\|_\infty$的估算

- 设已有A的列主元三角分解$PA = LU$

$\|A^{-1}\|_\infty$的估算

- 设已有A的列主元三角分解$PA = LU$
- 对$B = A^{-T}$应用上述定理，其中$w = Bx$和$z = B^T v$相当于求解方程组$A^T w = x$和$Az = v$

$\|A^{-1}\|_\infty$的估算

- 设已有A的列主元三角分解$PA = LU$
- 对$B = A^{-T}$应用上述定理，其中$w = Bx$和$z = B^T v$相当于求解方程组$A^T w = x$和$Az = v$
- 如此仅用$O(n^2)$就可以完成估计

$\|A^{-1}\|_\infty$的估算

- 设已有A的列主元三角分解$PA = LU$
- 对$B = A^{-T}$应用上述定理，其中$w = Bx$和$z = B^T v$相当于求解方程组$A^T w = x$和$Az = v$
- 如此仅用$O(n^2)$就可以完成估计
- 初值x可以任取，如$x_i = 1/n$

精度估算过程

- 估计$\|A^{-1}\|_\infty$的一个近似值$\tilde{\nu}$

精度估算过程

- 估计$\|A^{-1}\|_\infty$的一个近似值$\tilde{\nu}$
- 分别计算$\|r\|_\infty$, $\|b\|_\infty$, $\|A\|_\infty$的近似值$\tilde{\gamma}$, $\tilde{\beta}$, $\tilde{\mu}$

精度估算过程

- 估计$\|A^{-1}\|_\infty$的一个近似值$\tilde{\nu}$
- 分别计算$\|r\|_\infty$, $\|b\|_\infty$, $\|A\|_\infty$的近似值$\tilde{\gamma}$, $\tilde{\beta}$, $\tilde{\mu}$
- 计算$\tilde{\rho} = \tilde{\nu}\tilde{\mu}\tilde{\gamma}/\tilde{\beta}$得到计算解$\tilde{x}$的相对误差的一个估计

注解

- 大多数问题中这一方法可以得到相当好的估计

注解

- 大多数问题中这一方法可以得到相当好的估计
- 有一些特殊问题该方法得到的$\tilde{\rho}$远远小于计算解的实际相对误差，因为

注解

- 大多数问题中这一方法可以得到相当好的估计
- 有一些特殊问题该方法得到的$\tilde{\rho}$远远小于计算解的实际相对误差，因为
  - 由于舍入误差的影响使得$\tilde{\gamma}$远远小于$\|r\|_\infty$的真值

注解

- 大多数问题中这一方法可以得到相当好的估计
- 有一些特殊问题该方法得到的$\tilde{\rho}$远远小于计算解的实际相对误差，因为
  - 由于舍入误差的影响使得$\tilde{\gamma}$远远小于$\|r\|_\infty$的真值
  - 当A十分病态时，三角分解已相当不准确，从而使得应用它估计出的$\tilde{\nu}$ 要比$\|A^{-1}\|_\infty$的真值小得多

注解

- 大多数问题中这一方法可以得到相当好的估计
- 有一些特殊问题该方法得到的$\tilde{\rho}$远远小于计算解的实际相对误差，因为
  - 由于舍入误差的影响使得$\tilde{\gamma}$远远小于$\|r\|_\infty$的真值
  - 当A十分病态时，三角分解已相当不准确，从而使得应用它估计出的$\tilde{\nu}$ 要比$\|A^{-1}\|_\infty$的真值小得多
- 如何估计得更好，仍是一个有待深入研究的问题

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度
- 具体过程如下：

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度
- 具体过程如下：
  - 应用双精度和原始矩阵计算$r = b - A\tilde{x}$

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度
- 具体过程如下：
  - 应用双精度和原始矩阵计算$r = b - A\tilde{x}$
  - 利用A的三角分解求解$Az = r$

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度
- 具体过程如下：
  - 应用双精度和原始矩阵计算$r = b - A\tilde{x}$
  - 利用A的三角分解求解$Az = r$
  - 计算$x = \tilde{x} + z$

迭代改进

- 若$\tilde{x}$的精度太低，可以把$\tilde{x}$作为初值，对函数$f(x) = Ax - b$应用Newton迭代法改进其精度
- 具体过程如下：
  - 应用双精度和原始矩阵计算$r = b - A\tilde{x}$
  - 利用A的三角分解求解$Az = r$
  - 计算$x = \tilde{x} + z$
  - 若$\dfrac{\|x - \tilde{x}\|_\infty}{\|x\|_\infty} \leqslant \varepsilon$ 则结束；否则令$\tilde{x} = x$ 转第一步

最小二乘问题的求解

邓建松

2018 年 10 月 19 日

复习：数学分析中一个例题

- 题目：给定平面上n个数据点(x_i, y_i), $i = 1, 2, \ldots, n$, 求一条直线$y = ax + b$使得偏差

$$\varphi(a, b) = \sum_{i=1}^{n}(ax_i + b - y_i)^2$$

最小

复习：数学分析中一个例题

- 题目：给定平面上n个数据点(x_i, y_i), $i = 1, 2, \ldots, n$, 求一条直线$y = ax + b$使得偏差

$$\varphi(a, b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

最小

- 通过对$\varphi(a, b)$关于a, b求偏导，并令其等于0，得到关于a, b的线性方程组

复习：数学分析中一个例题

- 题目：给定平面上n个数据点(x_i, y_i), $i = 1, 2, \ldots, n$, 求一条直线$y = ax + b$使得偏差

$$\varphi(a, b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

最小

- 通过对$\varphi(a, b)$关于a, b求偏导，并令其等于0，得到关于a, b的线性方程组
- 系数矩阵可证当x_i互不相等时是非奇异的

答案

- a, b为下述方程组的解：

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i,$$

$$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i.$$

答案

- a, b为下述方程组的解：

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i,$$

$$a\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i.$$

- 所求直线的方程是

$$\begin{vmatrix} x & y & 1 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i & n \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i \end{vmatrix} = 0.$$

数据拟合

- 最小二乘问题多产生于数据拟合

数据拟合

- 最小二乘问题多产生于数据拟合
  - 给定m个点$t_1, \dots, t_m$和这m个点上的实验或观测数据$y_1, \dots, y_m$

数据拟合

- 最小二乘问题多产生于数据拟合
  - 给定m个点$t_1, \dots, t_m$和这m个点上的实验或观测数据$y_1, \dots, y_m$
  - 给定在t_i上取值的n个已知函数

$$\psi_1(t), \dots, \psi_n(t)$$

数据拟合

- 最小二乘问题多产生于数据拟合
  - 给定m个点$t_1, \ldots, t_m$和这m个点上的实验或观测数据$y_1, \ldots, y_m$
  - 给定在t_i上取值的n个已知函数
$$\psi_1(t), \ldots, \psi_n(t)$$
  - 考虑ψ_i的线性组合
$$f(x; t) = x_1\psi_1(t) + \cdots + x_n\psi_n(t)$$

数据拟合

- 最小二乘问题多产生于数据拟合
  - 给定m个点$t_1, \ldots, t_m$和这m个点上的实验或观测数据$y_1, \ldots, y_m$
  - 给定在t_i上取值的n个已知函数
  $$\psi_1(t), \ldots, \psi_n(t)$$
  - 考虑ψ_i的线性组合
  $$f(x; t) = x_1\psi_1(t) + \cdots + x_n\psi_n(t)$$
  - 我们希望在$t_1, \ldots, t_m$上$f(x; t)$能最佳地逼近$y_1, \ldots, y_m$

残量与最佳逼近

- 定义残量

$$r_i(x) = y_i - \sum_{j=1}^{n} x_j \psi_j(t_i), i = 1, \ldots, m$$

其中$x = (x_1, \ldots, x_n)^T$

残量与最佳逼近

- 定义残量

$$r_i(x) = y_i - \sum_{j=1}^{n} x_j \psi_j(t_i), i = 1, \ldots, m$$

其中$x = (x_1, \ldots, x_n)^T$

- 此问题转化为：估计参数$x_1, \ldots, x_n$，使残量$r_1, \ldots, r_m$尽可能得小

矩阵-向量形式

- 上式的矩阵-向量形式为$r(x) = b - Ax$, 其中

$$
\begin{aligned}
A &= (\psi_j(t_i))_{m\times n} \\
b &= (y_1, \dots, y_m)^T \\
x &= (x_1, \dots, x_n)^T \\
r(x) &= (r_1(x), \dots, r_m(x))^T
\end{aligned}
$$

求解

- $m = n$时，我们可以要求$r(x) = 0$，从而可以用第一章中的方法处理

求解

- $m = n$时，我们可以要求$r(x) = 0$，从而可以用第一章中的方法处理
- 当$m > n$时，一般不可能所有残量都为零，但可以要求$r(x)$在某种范数意义下最小

求解

- $m = n$时，我们可以要求$r(x) = 0$, 从而可以用第一章中的方法处理
- 当$m > n$时，一般不可能所有残量都为零，但可以要求$r(x)$在某种范数意义下最小
- 最小二乘问题就是求x使得$r(x)$在2范数意义下最小

定义

- 给定矩阵$A \in \mathbb{R}^{m\times n}$及向量$b \in \mathbb{R}^m$，确定$x \in \mathbb{R}^n$，使得

$$\|b - Ax\|_2 = \|r(x)\|_2 = \min_{y\in\mathbb{R}^n} \|r(y)\|_2 = \min_{y\in\mathbb{R}^n} \|Ay - b\|_2$$

这就是最小二乘问题，简记为LS(Least-Squares)问题，其中$r(x)$称为残向量

最小二乘解

- 最小二乘问题的解x又称作线性方程组

$$Ax = b, \quad A \in \mathbb{R}^{m \times n}$$

的最小二乘解

最小二乘解

- 最小二乘问题的解x又称作线性方程组

$$Ax = b, \quad A \in \mathbb{R}^{m\times n}$$

的最小二乘解

- 当$m > n$时，方程组称为超定方程组或矛盾方程组

最小二乘解

- 最小二乘问题的解x又称作线性方程组

$$Ax = b, \quad A \in \mathbb{R}^{m \times n}$$

的最小二乘解

- 当$m > n$时，方程组称为超定方程组或矛盾方程组
- 当$m < n$时，方程组称为欠定方程组

LS问题的种类

1. $m = n$:

LS问题的种类

1. $m = n$:
   1. $\operatorname{rank} A = m = n$

LS问题的种类

1. $m = n$:
   1. $\operatorname{rank} A = m = n$
   2. $\operatorname{rank} A = k < m = n$

LS问题的种类

1. $m = n$:
   1. $\operatorname{rank} A = m = n$
   2. $\operatorname{rank} A = k < m = n$
2. $m > n$:

LS问题的种类

1. $m = n$:
   1. $\operatorname{rank} A = m = n$
   2. $\operatorname{rank} A = k < m = n$
2. $m > n$:
   1. $\operatorname{rank} A = n < m$

LS问题的种类

1. $m = n$:
   1. $\operatorname{rank} A = m = n$
   2. $\operatorname{rank} A = k < m = n$
2. $m > n$:
   1. $\operatorname{rank} A = n < m$
   2. $\operatorname{rank} A = k < n < m$

LS问题的种类

1. $m = n$:
   1. $\operatorname{rank} A = m = n$
   2. $\operatorname{rank} A = k < m = n$
2. $m > n$:
   1. $\operatorname{rank} A = n < m$
   2. $\operatorname{rank} A = k < n < m$
3. $m < n$:

LS问题的种类

1. $m = n$:
   1. $\operatorname{rank} A = m = n$
   2. $\operatorname{rank} A = k < m = n$
2. $m > n$:
   1. $\operatorname{rank} A = n < m$
   2. $\operatorname{rank} A = k < n < m$
3. $m < n$:
   1. $\operatorname{rank} A = m < n$

LS问题的种类

1. $m = n$:
   1. $\operatorname{rank} A = m = n$
   2. $\operatorname{rank} A = k < m = n$
2. $m > n$:
   1. $\operatorname{rank} A = n < m$
   2. $\operatorname{rank} A = k < n < m$
3. $m < n$:
   1. $\operatorname{rank} A = m < n$
   2. $\operatorname{rank} A = k < m < n$

概念与记号

- 本章主要讨论(2-1)情形

概念与记号

- 本章主要讨论(2-1)情形
- 设$A \in \mathbb{R}^{m\times n}$, A的值域定义为

$$\mathcal{R}(A) = \{y \in \mathbb{R}^m : y = Ax, x \in \mathbb{R}^n\}$$

概念与记号

- 本章主要讨论(2-1)情形
- 设$A \in \mathbb{R}^{m\times n}$, A的值域定义为

$$\mathcal{R}(A) = \{y \in \mathbb{R}^m : y = Ax, x \in \mathbb{R}^n\}$$

- $\mathcal{R}(A) = \mathrm{span}(a_1, \ldots, a_n)$, 其中$a_i$为$A$的列向量

概念与记号

- 本章主要讨论(2-1)情形
- 设$A \in \mathbb{R}^{m\times n}$, A的值域定义为

$$\mathcal{R}(A) = \{y \in \mathbb{R}^m : y = Ax, x \in \mathbb{R}^n\}$$

- $\mathcal{R}(A) = \operatorname{span}(a_1, \ldots, a_n)$, 其中$a_i$为$A$的列向量
- A的零空间定义为

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}$$

其维数记为$\operatorname{null}(A)$

解的存在性

- 一个子空间$S \subset \mathbb{R}^n$的正交补定义为

$$\mathcal{S}^{\perp} = \{y \in \mathbb{R}^n : y^T x = 0, \forall x \in \mathcal{S}\}$$

解的存在性

- 一个子空间$S \subset \mathbb{R}^n$的正交补定义为

$$\mathcal{S}^\perp = \{y \in \mathbb{R}^n : y^T x = 0, \forall x \in \mathcal{S}\}$$

- 方程组$Ax = b$的解存在的充分必要条件是

$$\operatorname{rank} A = \operatorname{rank}([A, b])$$

非齐次方程的全部解

- 假设$Ax = b$的解存在，x是其任一给定的解，则方程组的全部解是$x + \mathcal{N}(A)$

非齐次方程的全部解

- 假设$Ax=b$的解存在，x是其任一给定的解，则方程组的全部解是$x+\mathcal{N}(A)$
- 方程组$Ax=b$的解唯一的充分必要条件是$\mathcal{N}(A)=\{0\}$

- $m = 3$, $\operatorname{rank} A = 2$, 则$\mathcal{R}(A)$可以用一张平面表示

- $m = 3$, $\operatorname{rank} A = 2$, 则$\mathcal{R}(A)$可以用一张平面表示
- 当x取遍$\mathbb{R}^n$时，$y = Ax$就取遍整个$\mathcal{R}(A)$

- $m=3$, $\operatorname{rank} A=2$, 则$\mathcal{R}(A)$可以用一张平面表示
- 当x取遍$\mathbb{R}^n$时，$y=Ax$就取遍整个$\mathcal{R}(A)$
- LS问题等价于求$y_{\min}\in\mathcal{R}(A)$, 使得

$$\|b-y_{\min}\|_2=\min\{\|b-y\|_2, y\in\mathcal{R}(A)\}$$

- 注意到b有分解：$b = b_1 + b_2$, $b_1 \in \mathcal{R}(A)$, $b_2 \in \mathcal{R}(A)^\perp$

- 注意到b有分解：$b = b_1 + b_2$, $b_1 \in \mathcal{R}(A)$, $b_2 \in \mathcal{R}(A)^{\perp}$
- 当$b - y$垂直于$\mathcal{R}(A)$时，$\|b - y\|_2$达到极小

- 注意到b有分解：$b = b_1 + b_2$, $b_1 \in \mathcal{R}(A)$, $b_2 \in \mathcal{R}(A)^{\perp}$
- 当$b - y$垂直于$\mathcal{R}(A)$时，$\|b - y\|_2$达到极小
- 这时$y_{\min} = b_1$, 然后利用$Ax = y_{\min}$解出x即得到最小二乘解

解的存在性定理

定理

$Ax = b$对应的线性最小二乘问题的解总是存在的，而且其解唯一当且仅当

$$\mathcal{N}(A) = \{0\}$$

证明

- $\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$

证明

- $\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$
- 所以b具有唯一分解$b = b_1 + b_2$, $b_1 \in \mathcal{R}(A)$, $b_2 \in \mathcal{R}(A)^\perp$

证明

- $\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{R}(A)^\perp$
- 所以b具有唯一分解$b = b_1 + b_2$, $b_1 \in \mathcal{R}(A)$, $b_2 \in \mathcal{R}(A)^\perp$
- 从而对$\forall x \in \mathbb{R}^n$, $b_1 - Ax \in \mathcal{R}(A)$且与$b_2$正交

证明

- 从而有

$$\begin{aligned}\|r(x)\|_2^2 = \|b - Ax\|_2^2 &= \|(b_1 - Ax) + b_2\|_2^2 \\ &= \|b_1 - Ax\|_2^2 + \|b_2\|^2\end{aligned}$$

证明

- 从而有

$$\|r(x)\|_2^2 = \|b - Ax\|_2^2 = \|(b_1 - Ax) + b_2\|_2^2$$
$$= \|b_1 - Ax\|_2^2 + \|b_2\|^2$$

- 所以$\|r(x)\|_2^2$达到极小当且仅当$\|b_1 - Ax\|_2^2$达到极小

证明

- 从而有

$$\begin{aligned}\|r(x)\|_2^2 = \|b - Ax\|_2^2 &= \|(b_1 - Ax) + b_2\|_2^2 \\ &= \|b_1 - Ax\|_2^2 + \|b_2\|^2\end{aligned}$$

- 所以$\|r(x)\|_2^2$达到极小当且仅当$\|b_1 - Ax\|_2^2$达到极小
- 由于$b_1 \in \mathcal{R}(A)$, 所以$\|b_1 - Ax\|_2^2$达到极小当且仅当$Ax = b_1$

解集

- 最小二乘问题的解集记为χ_{LS}

解集

- 最小二乘问题的解集记为χ_{LS}
- 根据前面的定理，$\chi_{\mathrm{LS}} \neq \emptyset$

解集

- 最小二乘问题的解集记为χ_{LS}
- 根据前面的定理，$\chi_{\mathrm{LS}} \neq \emptyset$
- $\sharp\chi_{\mathrm{LS}} = 1 \iff A$的列线性无关

解集

- 最小二乘问题的解集记为χ_{LS}
- 根据前面的定理，$\chi_{\mathrm{LS}} \neq \emptyset$
- $\sharp\chi_{\mathrm{LS}} = 1 \Longleftrightarrow A$的列线性无关
- χ_{LS}中有且仅有一个解其2范数最小（为什么？），这称为**最小2范数解**，用x_{LS}表示

解集

- 最小二乘问题的解集记为χ_{LS}
- 根据前面的定理，$\chi_{\mathrm{LS}} \neq \emptyset$
- $\sharp\chi_{\mathrm{LS}} = 1 \iff A$的列线性无关
- χ_{LS}中有且仅有一个解其2范数最小（为什么？），这称为最小2范数解，用x_{LS}表示
  - 点集的凸性以及范数的严格凸性

定理

定理

$x \in \chi_{\mathrm{LS}}$ 当且仅当

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

证明：必要性

- $x \in \chi_{\mathrm{LS}} \Longrightarrow Ax = b_1$, 其中$b_1 \in \mathcal{R}(A)$

证明：必要性

- $x \in \chi_{\mathrm{LS}} \Longrightarrow Ax = b_1$, 其中 $b_1 \in \mathcal{R}(A)$
- $r(x) = b - Ax = b - b_1 = b_2 \in \mathcal{R}(A)^{\perp}$

证明：必要性

- $x \in \chi_{\mathrm{LS}} \Longrightarrow Ax = b_1$, 其中$b_1 \in \mathcal{R}(A)$
- $r(x) = b - Ax = b - b_1 = b_2 \in \mathcal{R}(A)^{\perp}$
- $b_2 \in \mathcal{R}(A)^{\perp}$意味着对任意$x \in \mathbb{R}^n$, $b^T Ax = 0$, 所以$(b^T A)^T = A^T b$是$\mathbb{R}^n$中的零向量

证明：必要性

- $x \in \chi_{\text{LS}} \Longrightarrow Ax = b_1$, 其中$b_1 \in \mathcal{R}(A)$
- $r(x) = b - Ax = b - b_1 = b_2 \in \mathcal{R}(A)^\perp$
- $b_2 \in \mathcal{R}(A)^\perp$意味着对任意$x \in \mathbb{R}^n$, $b^T Ax = 0$, 所以$(b^T A)^T = A^T b$是$\mathbb{R}^n$中的零向量
- 从而$A^T r(x) = A^T b_2 = 0$

证明：必要性

- $x \in \chi_{\mathrm{LS}} \Longrightarrow Ax = b_1$, 其中$b_1 \in \mathcal{R}(A)$
- $r(x) = b - Ax = b - b_1 = b_2 \in \mathcal{R}(A)^\perp$
- $b_2 \in \mathcal{R}(A)^\perp$意味着对任意$x \in \mathbb{R}^n$, $b^T Ax = 0$, 所以$(b^T A)^T = A^T b$是$\mathbb{R}^n$中的零向量
- 从而$A^T r(x) = A^T b_2 = 0$
- 把$r(x) = b - Ax$代入即得$A^T Ax = A^T b$

证明：充分性

- 设$x \in \mathbb{R}^n$满足$A^TAx = A^Tb$

证明：充分性

- 设$x \in \mathbb{R}^n$满足$A^TAx = A^Tb$
- 则对$\forall y \in \mathbb{R}^n$,

$$
\begin{aligned}
&\|b - A(x+y)\|_2^2 \\
=&\|b - Ax\|_2^2 - 2y^TA^T(b - Ax) + \|Ay\|_2^2 \\
=&\|b - Ax\|_2^2 + \|Ay\|_2^2 \\
\geqslant&\|b - Ax\|_2^2
\end{aligned}
$$

这就证明了$x \in \chi_{\mathrm{LS}}$

正则化方程组

- $A^TAx = A^Tb$称为LS问题的正则化方程组或者法方程组

正则化方程组

- $A^TAx = A^Tb$称为LS问题的正则化方程组或者法方程组
- 它一般是一个含有n个变量和n个方程的线性方程组

正则化方程组

- $A^TAx = A^Tb$称为LS问题的正则化方程组或者法方程组
- 它一般是一个含有n个变量和n个方程的线性方程组
- 如果A的列向量线性无关，那么A^TA对称正定，从而可以采用平方根法求解方程组

正则化方法

求解LS问题的最古老的算法：

- 计算$C = A^T A$, $d = A^T b$

正则化方法

求解LS问题的最古老的算法：

- 计算$C = A^TA$, $d = A^Tb$
- 用平方根法计算C的Cholesky分解：$C = LL^T$

正则化方法

求解LS问题的最古老的算法：

- 计算$C = A^T A$, $d = A^T b$
- 用平方根法计算C的Cholesky分解：$C = LL^T$
- 求解三角方程组$Ly = d$和$L^T x = y$

注解

- 在A^TA的计算中，如果不使用足够的精度，A中的一些精度可能会丢失

注解

- 在A^TA的计算中，如果不使用足够的精度，A中的一些精度可能会丢失
- 例：

$$A = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix}, A^TA = \begin{pmatrix} c & 1 & 1 \\ 1 & c & 1 \\ 1 & 1 & c \end{pmatrix}$$

其中$c = 1 + \varepsilon^2$

Moore-Penrose广义逆

- 正则化方程组的解可以写为$x = (A^TA)^{-1}A^Tb$

Moore-Penrose广义逆

- 正则化方程组的解可以写为$x = (A^T A)^{-1} A^T b$
- 定义$A^\dagger = (A^T A)^{-1} A^T$

Moore-Penrose广义逆

- 正则化方程组的解可以写为$x = (A^TA)^{-1}A^Tb$
- 定义$A^\dagger = (A^TA)^{-1}A^T$
- 则LS问题的解可以写为$x = A^\dagger b$

Moore-Penrose广义逆

- 正则化方程组的解可以写为$x = (A^TA)^{-1}A^Tb$
- 定义$A^\dagger = (A^TA)^{-1}A^T$
- 则LS问题的解可以写为$x = A^\dagger b$
- $n \times m$阶矩阵$A^\dagger$就是A的Moore-Penrose广义逆

回忆：Moore-Penrose广义逆

- 若$X \in \mathbb{R}^{m\times n}$满足

$$AXA = A, XAX = X,$$
$$(AX)^T = AX, (XA)^T = XA$$

则X就是A的Moore-Penrose广义逆

回忆：Moore-Penrose广义逆

- 若$X \in \mathbb{R}^{m \times n}$满足

$$AXA = A, XAX = X,$$
$$(AX)^T = AX, (XA)^T = XA$$

则X就是A的Moore-Penrose广义逆

- 通常记作$A^\dagger$

扰动对解的影响

- 设b有扰动δb, 且x和$x+\delta x$分别是最小二乘问题

$$\min \|b - Ax\|_2 \text{和} \min \|(b+\delta b) - Ax\|_2$$

的解，即

$$x = A^{\dagger}b,$$
$$x + \delta x = A^{\dagger}(b+\delta b) = A^{\dagger}\tilde{b}$$

其中$\tilde{b} = b + \delta b$

定理

定理

设b_1和$\tilde{b}_1$分别是b和$\tilde{b}$在$\mathcal{R}(A)$上的正交投影。

若$b_1 \neq 0$，则

$$\frac{\|\delta x\|_2}{\|x\|_2} \leqslant \kappa_2(A)\frac{\|b_1 - \tilde{b}_1\|_2}{\|b_1\|_2}$$

其中$\kappa_2(A) = \|A\|_2\|A^\dagger\|_2$

注：A非方阵，其范数与方阵的算子范数定义相同，从而满足对向量乘法的相容性；A的2范数等于A的最大奇异值

证明

- 设b在$\mathcal{R}^{\perp}$上的正交投影为b_2，则$A^T b_2 = 0$

证明

- 设b在$\mathcal{R}^\perp$上的正交投影为b_2，则$A^T b_2 = 0$
- 由$b = b_1 + b_2$可有

$$
\begin{aligned}
A^\dagger b &= A^\dagger b_1 + A^\dagger b_2 \\
&= A^\dagger b_1 + (A^T A)^{-1} A^T b_2 = A^\dagger b_1
\end{aligned}
$$

证明

- 设b在$\mathcal{R}^{\perp}$上的正交投影为b_2，则$A^T b_2 = 0$
- 由$b = b_1 + b_2$可有

$$\begin{aligned} A^{\dagger} b &= A^{\dagger} b_1 + A^{\dagger} b_2 \\ &= A^{\dagger} b_1 + (A^T A)^{-1} A^T b_2 = A^{\dagger} b_1 \end{aligned}$$

- 同理$A^{\dagger}\tilde{b} = A^{\dagger}\tilde{b}_1$

证明

- 所以

$$\begin{aligned}\|\delta x\|_2 &= \|A^\dagger b - A^\dagger \tilde{b}\|_2 = \|A^\dagger (b_1 - \tilde{b}_1)\|_2 \\ &\leqslant \|A^\dagger\|_2 \|b_1 - \tilde{b}_1\|_2\end{aligned}$$

证明

- 所以

$$\begin{aligned}\|\delta x\|_2 &= \|A^\dagger b - A^\dagger \tilde{b}\|_2 = \|A^\dagger(b_1 - \tilde{b}_1)\|_2 \\ &\leqslant \|A^\dagger\|_2\|b_1 - \tilde{b}_1\|_2\end{aligned}$$

- 由$Ax = b_1$得$\|b_1\|_2 \leqslant \|A\|_2\|x\|_2$

证明

- 所以

$$\|\delta x\|_2 = \|A^\dagger b - A^\dagger \tilde{b}\|_2 = \|A^\dagger(b_1 - \tilde{b}_1)\|_2 \leqslant \|A^\dagger\|_2 \|b_1 - \tilde{b}_1\|_2$$

- 由$Ax = b_1$得$\|b_1\|_2 \leqslant \|A\|_2 \|x\|_2$
- 根据上述两式立得结论

注解

- 若b有变化，只有它在$\mathcal{R}(A)$上的投影对x的相对误差产生影响

注解

- 若b有变化，只有它在$\mathcal{R}(A)$上的投影对x的相对误差产生影响
- LS问题解的敏感性依赖于$\kappa_2(A)$的大小

注解

- 若b有变化，只有它在$\mathcal{R}(A)$上的投影对x的相对误差产生影响
- LS问题解的敏感性依赖于$\kappa_2(A)$的大小
- 我们称$\kappa_2(A)$为LS问题的条件数

注解

- 若b有变化，只有它在$\mathcal{R}(A)$上的投影对x的相对误差产生影响
- LS问题解的敏感性依赖于$\kappa_2(A)$的大小
- 我们称$\kappa_2(A)$为LS问题的条件数
- 若$\kappa_2(A)$很大，则称LS问题是病态的；否则称为良态的

注解

- 若b有变化，只有它在$\mathcal{R}(A)$上的投影对x的相对误差产生影响
- LS问题解的敏感性依赖于$\kappa_2(A)$的大小
- 我们称$\kappa_2(A)$为LS问题的条件数
- 若$\kappa_2(A)$很大，则称LS问题是病态的；否则称为良态的
- 同时考虑A和b的扰动对解的影响就非常复杂，我们在此不讨论

条件数

定理

设A的列向量线性无关，则

$$\kappa_2(A)^2 = \kappa_2(A^TA)$$

证明：

- 根据定义，我们有

$$\|A\|_2^2 = \|A^TA\|_2,$$
$$\|A^\dagger\|_2^2 = \|A^\dagger(A^\dagger)^T\|_2 = \|(A^TA)^{-1}\|_2$$

证明

- 于是我们得到

$$
\begin{aligned}
\kappa_2(A)^2 &= \|A\|_2^2 \|A^\dagger\|_2^2 \\
&= \|A^TA\|_2 \|(A^TA)^{-1}\|_2 \\
&= \kappa_2(A^TA)
\end{aligned}
$$

注解

- 最小二乘问题在化为正则化方程组后，条件数是原来的平方

注解

- 最小二乘问题在化为正则化方程组后，条件数是原来的平方
- 这就使得求解过程增加了对舍入误差的敏感性

注解

- 最小二乘问题在化为正则化方程组后，条件数是原来的平方
- 这就使得求解过程增加了对舍入误差的敏感性
- 在使用正则化方法时，一定要注意这一点

更实用的算法

- 为了给出求解最小二乘问题的更实用的算法，我们本节介绍初等正交变换

更实用的算法

- 为了给出求解最小二乘问题的更实用的算法，我们本节介绍初等正交变换
  - 因为正交变换不改变向量和矩阵的2范数

更实用的算法

- 为了给出求解最小二乘问题的更实用的算法，我们本节介绍初等正交变换
    - 因为正交变换不改变向量和矩阵的2范数
- 第一种是Householder变换

更实用的算法

- 为了给出求解最小二乘问题的更实用的算法，我们本节介绍初等正交变换
  - 因为正交变换不改变向量和矩阵的2范数
- 第一种是Householder变换
- 第二种是Givens变换

更实用的算法

- 为了给出求解最小二乘问题的更实用的算法，我们本节介绍初等正交变换
  - 因为正交变换不改变向量和矩阵的2范数
- 第一种是Householder变换
- 第二种是Givens变换
- 它们是数值线性代数中许多重要算法的基础

更实用的算法

- 为了给出求解最小二乘问题的更实用的算法，我们本节介绍初等正交变换
  - 因为正交变换不改变向量和矩阵的2范数
- 第一种是Householder变换
- 第二种是Givens变换
- 它们是数值线性代数中许多重要算法的基础
  - 例：在计算矩阵特征值和特征向量的QR方法中，就大量应用上述两种变换

回忆：初等变换

- 应用Gauss变换可以把一个矩阵约化为上三角形式

回忆：初等变换

- 应用Gauss变换可以把一个矩阵约化为上三角形式
- 这是基于事实：对任意向量x，可以构造一个初等下三角阵L，使得$Lx = \alpha e_1$

回忆：初等变换

- 应用Gauss变换可以把一个矩阵约化为上三角形式
- 这是基于事实：对任意向量x，可以构造一个初等下三角阵L，使得$Lx = \alpha e_1$
- 本节我们讨论如何求一个初等正交矩阵，使其具有L同样的功能

镜像对称向量的计算

- 在$\mathbb{R}^n$中给定一个向量x和一张单位法向量为w的超平面π，那么x关于π的镜像对称向量是什么？

镜像对称向量的计算

- 在$\mathbb{R}^n$中给定一个向量x和一张单位法向量为w的超平面π，那么x关于π的镜像对称向量是什么？
- 显然x在单位法向量上的投影向量为$(x \cdot w)w = ww^T x$

镜像对称向量的计算

- 在$\mathbb{R}^n$中给定一个向量x和一张单位法向量为w的超平面π，那么x关于π的镜像对称向量是什么？
- 显然x在单位法向量上的投影向量为$(x \cdot w)w = ww^T x$
- 所以对称向量是

$$x - 2ww^T x = (I - 2ww^T)x$$

Householder变换

- 设$w \in \mathbb{R}^n$满足$\|w\|_2 = 1$. 定义$n \times n$矩阵

$$H = I - 2ww^T$$

其称为Householder变换

Householder变换

- 设$w \in \mathbb{R}^n$满足$\|w\|_2 = 1$. 定义$n \times n$矩阵

$$H = I - 2ww^T$$

其称为Householder变换

- 也称为初等反射矩阵或镜像矩阵

Householder变换

- 设$w \in \mathbb{R}^n$满足$\|w\|_2 = 1$. 定义$n \times n$矩阵

$$H = I - 2ww^T$$

  其称为Householder变换
- 也称为初等反射矩阵或镜像矩阵
- 这一变换最早是由A.C. Aitken在1932年提出，后由数值分析专家Alston S. Householder在1958年应用到数值线性代数中

变换的性质

- 对称性： $H^T = H$

变换的性质

- 对称性：$H^T = H$
- 正交性：$H^T H = I$

变换的性质

- 对称性：$H^T = H$
- 正交性：$H^T H = I$
- 对合性：$H^2 = I$

变换的性质

- 对称性：$H^T = H$
- 正交性：$H^T H = I$
- 对合性：$H^2 = I$
- 反射性：对$\forall x \in \mathbb{R}^n$，$Hx$是$x$关于$w$的垂直超平面$\mathrm{span}\{w\}^{\perp}$的镜像反射

证明

- 第一条显然

证明

- 第一条显然
- 后两条可由第一条导出。事实上，

$$
\begin{aligned}
H^T H = H^2 &= (I - 2ww^T)(I - 2ww^T) \\
&= I - 4ww^T + 4ww^T ww^T = I
\end{aligned}
$$

证明

- 设$x \in \mathbb{R}^n$写为$x = u + \alpha w$, 其中$u \in \operatorname{span}\{w\}^\perp$, $\alpha \in \mathbb{R}$

证明

- 设$x \in \mathbb{R}^n$写为$x = u + \alpha w$，其中$u \in \mathrm{span}\{w\}^{\perp}$，$\alpha \in \mathbb{R}$
- 由$u^T w = 0$，$w^T w = 1$可得

$$\begin{aligned} Hx &= (I - 2ww^T)(u + \alpha w) \\ &= u + \alpha w - 2ww^T u - 2\alpha ww^T w \\ &= u - \alpha w \end{aligned}$$

定理

定理

设$0 \neq x \in \mathbb{R}^n$，可以构造单位向量$w \in \mathbb{R}^n$使得$Householder$变换$H$满足

$$Hx = \alpha e_1$$

其中$\alpha = \pm \|x\|_2$

证明

- 注意到

$$Hx = (I - 2ww^T)x = x - 2(w^Tx)w$$

证明

- 注意到

$$Hx = (I - 2ww^T)x = x - 2(w^T x)w$$

- 为使$Hx = \alpha e_1$, 则w应取为

$$w = \frac{x - \alpha e_1}{\|x - \alpha e_1\|_2}$$

证明

- 注意到

$$Hx = (I - 2ww^T)x = x - 2(w^Tx)w$$

- 为使$Hx = \alpha e_1$, 则w应取为

$$w = \frac{x - \alpha e_1}{\|x - \alpha e_1\|_2}$$

- 当$\alpha = \pm\|x\|_2$时，可以直接验证如此定义的w满足定理的要求

验证

- $\alpha^2 = x^T x$

验证

- $\alpha^2 = x^T x$
- 分母为

$$\begin{aligned} &\|x - \alpha e_1\|_2^2 \\ =&x^T x - 2\alpha x^T e_1 + \alpha^2 \\ =&2(x^T x - \alpha x^T e_1) \end{aligned}$$

验证

- $\alpha^2 = x^T x$
- 分母为

$$\begin{aligned}&\|x-\alpha e_1\|_2^2\\=&x^T x - 2\alpha x^T e_1 + \alpha^2\\=&2(x^T x - \alpha x^T e_1)\end{aligned}$$

- 分母即为$2(x^T - \alpha e_1^T)x$，由此易得

$$2(w^T x)w = x - \alpha e_1$$

注解

- 定理告诉我们，对于$\forall x \in \mathbf{R}^n$（$x \neq 0$），我们都可以构造出Householder变换$H$，使得$Hx$的后$n-1$个分量为零

注解

- 定理告诉我们，对于$\forall x \in \mathbf{R}^n$（$x \neq 0$），我们都可以构造出Householder变换$H$，使得$Hx$的后$n-1$个分量为零
- 证明步骤同时告诉我们w的构造方法如下：

注解

- 定理告诉我们，对于$\forall x \in \mathbf{R}^n$ ($x \neq 0$)，我们都可以构造出Householder变换H，使得Hx的后$n-1$个分量为零
- 证明步骤同时告诉我们w的构造方法如下：
    - 计算$v = x \pm \|x\|_2 e_1$

注解

- 定理告诉我们，对于$\forall x \in \mathbf{R}^n$（$x \neq 0$），我们都可以构造出Householder变换$H$，使得$Hx$的后$n-1$个分量为零
- 证明步骤同时告诉我们w的构造方法如下：
  - 计算$v = x \pm \|x\|_2 e_1$
  - 计算$w = v/\|v\|_2$

符号的选择

- 为了使变换后得到的α为正数，我们应取$v = x - \|x\|_2 e_1$

符号的选择

- 为了使变换后得到的α为正数，我们应取$v = x - \|x\|_2 e_1$
- 问题：如果x是一个很接近于e_1的向量，计算$v_1 = x_1 - \|x\|_2$时会出现两个相近的数相减，从而严重地损失有效数字

符号的选择

- 为了使变换后得到的α为正数，我们应取$v = x - \|x\|_2 e_1$
- 问题：如果x是一个很接近于e_1的向量，计算$v_1 = x_1 - \|x\|_2$时会出现两个相近的数相减，从而严重地损失有效数字
- 变形以避免这一问题：($x_1 > 0$)

$$v_1 = x_1 - \|x\|_2 = \frac{x_1^2 - \|x\|_2^2}{x_1 + \|x\|_2} = \frac{-(x_2^2 + \cdots + x_n^2)}{x_1 + \|x\|_2}$$

w不需要计算

- 由于

$$H = I - 2ww^T = I - \frac{2}{v^Tv}vv^T = I - 2\beta vv^T$$

其中$\beta = 2/(v^Tv)$, 因此我们不必求出w，而只需求出β和v, 从而避免了开方运算

w不需要计算

- 由于

$$H = I - 2ww^T = I - \frac{2}{v^T v} vv^T = I - 2\beta vv^T$$

其中$\beta = 2/(v^T v)$, 因此我们不必求出w，而只需求出β和v, 从而避免了开方运算

- 在实际计算时，可以把v规范化为第一个分量为1（第一个分量原值肯定不为零），这样可以恰好把v的后$n-1$分量放在x的后$n-1$个化为零的分量位置上

下溢和上溢

- 当下溢发生时，计算机有可能把结果置为零，这可能会出现$v^T v$为零的情形

下溢和上溢

- 当下溢发生时，计算机有可能把结果置为零，这可能会出现$v^T v$为零的情形
- 如果x的分量太大，那么该分量平方时，会出现上溢

下溢和上溢

- 当下溢发生时，计算机有可能把结果置为零，这可能会出现v^Tv为零的情形
- 如果x的分量太大，那么该分量平方时，会出现上溢
- 由于$\forall\alpha$, αv和v的单位化向量相同，因此为了避免溢出现象的出现，我们可以用$x/\|x\|_\infty$代替x来构造v, 这相当于在原来的v之前乘了常数$1/\|x\|_\infty$

化其它元素为零

- Householder变换结果的形式并不需要局限于αe_1

化其它元素为零

- Householder变换结果的形式并不需要局限于αe_1
- 它可以把向量中任何若干相邻的元素化为零

化其它元素为零

- Householder变换结果的形式并不需要局限于αe_1
- 它可以把向量中任何若干相邻的元素化为零
- 例如，欲在$x \in \mathbb{R}^n$中从$k+1$至j位置引入零元素，只要定义

$$v = (0, \ldots, 0, x_k - \alpha, x_{k+1}, \ldots, x_j, 0, \ldots, 0)$$

即可，其中$\alpha^2 = \sum_{i=k}^{j} x_i^2$

HA的计算

- 应用Householder变换，主要的工作量是计算矩阵乘法HA

$$HA = (I - \beta vv^T)A = A - vw^T$$

其中$w = \beta A^T v$

HA的计算

- 应用Householder变换，主要的工作量是计算矩阵乘法HA

$$HA = (I - \beta vv^T)A = A - vw^T$$

其中$w = \beta A^T v$

- 计算w的一个元素需要$n + (n-1) + 1 = 2n$次运算；从而计算w需要$2mn$次运算

HA的计算

- 应用Householder变换，主要的工作量是计算矩阵乘法HA
$$HA = (I - \beta vv^T)A = A - vw^T$$
其中$w = \beta A^T v$
- 计算w的一个元素需要$n + (n-1) + 1 = 2n$次运算；从而计算w需要$2mn$次运算
- 计算$A - vw^T$的一个元素需要两次运算

HA的计算

- 应用Householder变换，主要的工作量是计算矩阵乘法HA

$$HA = (I - \beta vv^T)A = A - vw^T$$

其中$w = \beta A^T v$

- 计算w的一个元素需要$n + (n-1) + 1 = 2n$次运算；从而计算w需要$2mn$次运算
- 计算$A - vw^T$的一个元素需要两次运算
- 所以计算HA的总运算量为$4mn$

Givens变换

- Householder变换把一个向量中许多分量化为零

Givens变换

- Householder变换把一个向量中许多分量化为零
- Givens变换则只是把向量中一个分量化为零

$$\begin{aligned} G(i,k,\theta) = I &+ s(e_i e_k^T - e_k e_i^T) \\ &+ (c-1)(e_i e_i^T + e_k e_k^T) \end{aligned}$$

其中$c = \cos\theta$, $s = \sin\theta$

Givens变换

- Householder变换把一个向量中许多分量化为零
- Givens变换则只是把向量中一个分量化为零

$$G(i,k,\theta) = I + s(e_i e_k^T - e_k e_i^T) + (c-1)(e_i e_i^T + e_k e_k^T)$$

其中$c = \cos\theta$, $s = \sin\theta$

- 这一变换是由Wallace Givens于上世纪五十年代引入到数值分析领域，也称为Jacobi变换(C.G.J. Jacobi, 1804–1851)

$G(i,k,\theta)$的结构

$$G(i,k,\theta)=\begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c & \cdots & s & & \\ & & \vdots & & \vdots & & \\ & & -s & \cdots & c & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}$$

置零时的取值

- 取$x \in \mathbb{R}$, $y = G(i,k,\theta)x$, 则

$$y_i = cx_i + sx_k, y_k = -sx_i + cx_k, y_j = x_j, j \neq i,$$

置零时的取值

- 取$x \in \mathbb{R}$, $y = G(i,k,\theta)x$, 则

$$y_i = cx_i + sx_k, y_k = -sx_i + cx_k, y_j = x_j, j \neq i,$$

- 若要$y_k = 0$, 只要取

$$c = \frac{x_i}{\sqrt{x_i^2 + x_k^2}}, s = \frac{x_k}{\sqrt{x_i^2 + x_k^2}}$$

就有$y_i = \sqrt{x_i^2 + x_k^2}$, $y_k = 0$

旋转

- 从几何上看，$G(i,k,\theta)x$是在(i,k)坐标平面内将x按顺时针方向旋转θ角

旋转

- 从几何上看，$G(i,k,\theta)x$是在(i,k)坐标平面内将x按顺时针方向旋转θ角
- 所以Givens变换也称为平面旋转变换

旋转

- 从几何上看，$G(i,k,\theta)x$是在(i,k)坐标平面内将x按顺时针方向旋转θ角
- 所以Givens变换也称为平面旋转变换
- Givens变换左（或右）乘矩阵A, 则它只改变A的第i,k行（或列），其余元素保持不变

溢出的避免

- 利用c, s的定义进行计算，有可能发生溢出

溢出的避免

- 利用c, s的定义进行计算，有可能发生溢出
- 为了防止溢出，在实现时可以采用一些小技巧，见书上算法中的描述

正交变换与LS问题

- 本节讨论LS问题求解的新方法

正交变换与LS问题

- 本节讨论LS问题求解的新方法
- 由于2范数具有正交不变性，所以对任意正交矩阵$Q \in \mathbb{R}^{m\times m}$,

$$\|Ax - b\|_2 = \|Q^T(Ax - b)\|_2$$

正交变换与LS问题

- 本节讨论LS问题求解的新方法
- 由于2范数具有正交不变性，所以对任意正交矩阵$Q \in \mathbb{R}^{m\times m}$,

$$\|Ax-b\|_2 = \|Q^T(Ax-b)\|_2$$

- 从而LS问题$\min \|Q^TAx - Q^Tb\|_2$就等价于原问题$\min \|Ax-b\|_2$

正交变换与LS问题

- 本节讨论LS问题求解的新方法
- 由于2范数具有正交不变性，所以对任意正交矩阵$Q \in \mathbb{R}^{m\times m}$,
$$\|Ax-b\|_2 = \|Q^T(Ax-b)\|_2$$
- 从而LS问题$\min\|Q^TAx - Q^Tb\|_2$就等价于原问题$\min\|Ax-b\|_2$
- 期望通过适当选取正交矩阵Q, 使原问题转化为较容易求解的LS问题。这就是正交变换法的基本思想

QR分解定理

定理

设$A \in \mathbb{R}^{m\times n}$ $(m \geqslant n)$, 则A有QR分解

$$A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$$

其中$Q \in \mathbb{R}^{m\times m}$是正交矩阵，$R \in \mathbb{R}^{n\times n}$是具有非负对角元的上三角阵；而且当$m = n$且$A$非奇异时，上述分解是唯一的

证明：存在性

- 对n进行数学归纳法

证明：存在性

- 对n进行数学归纳法
- 当$n = 1$时这就是前一节关于Householder变换的定理

证明：存在性

- 对n进行数学归纳法
- 当$n = 1$时这就是前一节关于Householder变换的定理
- 假设已经证明了定理对所有的$p \times (n-1)$矩阵成立，$p \geqslant n-1$

证明：存在性

- 对n进行数学归纳法
- 当$n=1$时这就是前一节关于Householder变换的定理
- 假设已经证明了定理对所有的$p\times(n-1)$矩阵成立，$p\geqslant n-1$
- 设$A\in\mathbb{R}^{m\times n}$的第一列是$a_1$

证明：存在性

- 由Householder变换定理，存在正交矩阵$Q_1 \in \mathbb{R}^{m\times m}$使得$Q_1^T a_1 = \|a_1\|_2 e_1$

证明：存在性

- 由Householder变换定理，存在正交矩阵$Q_1 \in \mathbb{R}^{m\times m}$使得$Q_1^T a_1 = \|a_1\|_2 e_1$
- 于是我们有

$$Q_1^T A = \begin{pmatrix} \|a_1\|_2 & v^T \\ 0 & A_1 \end{pmatrix}$$

证明：存在性

- 对$(m-1)\times(n-1)$阶矩阵A_1应用归纳假设，有

$$A_1 = Q_2 \begin{pmatrix} R_2 \\ 0 \end{pmatrix}$$

其中Q_2是$(m-1)\times(m-1)$阶正交矩阵，R_2是具有非负对角元的$(n-1)\times(n-1)$上三角阵

证明：存在性

- 令

$$Q = Q_1 \begin{pmatrix} 1 & 0 \\ 0 & Q_2 \end{pmatrix}, R = \begin{pmatrix} \|a_1\|_2 & v^T \\ 0 & R_2 \\ 0 & 0 \end{pmatrix}$$

则Q和R满足定理的要求。存在性得证

证明：唯一性

- 设$m = n$，A非奇异

证明：唯一性

- 设$m = n$，A非奇异
- 设$A = QR = \tilde{Q}\tilde{R}$，其中$Q, \tilde{Q} \in \mathbb{R}^{m\times m}$是正交矩阵，$R, \tilde{R} \in \mathbb{R}^{m\times m}$是具有非负对角元的上三角阵

证明：唯一性

- 设$m = n$, A非奇异
- 设$A = QR = \tilde{Q}\tilde{R}$, 其中$Q, \tilde{Q} \in \mathbb{R}^{m\times m}$是正交矩阵，$R, \tilde{R} \in \mathbb{R}^{m\times m}$是具有非负对角元的上三角阵
- A非奇异，所以$R, \tilde{R}$的对角元均为正数，所以
$$\tilde{Q}^T Q = \tilde{R} R^{-1}$$

证明：唯一性

- 上式左边是正交矩阵

证明：唯一性

- 上式左边是正交矩阵
- 上式右边是对角元均为正数的上三角阵

证明：唯一性

- 上式左边是正交矩阵
- 上式右边是对角元均为正数的上三角阵
- 所以两边只能是单位阵，从而必有$Q = \tilde{Q}$, $R = \tilde{R}$

LS问题的正交变换法

- 设$A \in \mathbb{R}^{m \times n}$ $(m \geqslant n)$有线性无关的列，$b \in \mathbb{R}^m$

LS问题的正交变换法

- 设$A \in \mathbb{R}^{m\times n}$ $(m \geqslant n)$有线性无关的列，$b \in \mathbb{R}^m$
- A有QR分解，并且把Q分块为$Q = (Q_1, Q_2)$, 其中Q_1有n列

LS问题的正交变换法

- 设$A \in \mathbb{R}^{m\times n}$ $(m \geqslant n)$有线性无关的列，$b \in \mathbb{R}^m$
- A有QR分解，并且把Q分块为$Q = (Q_1, Q_2)$, 其中Q_1有n列
- 令

$$Q^T b = \begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix} b = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

- 那么

$$
\begin{aligned}
\|Ax-b\|_2^2 &= \|Q^TAx-Q^Tb\|_2^2 \\
&= \|Rx-c_1\|_2^2+\|c_2\|_2^2
\end{aligned}
$$

- 那么

$$\begin{aligned}\|Ax-b\|_2^2 &= \|Q^TAx-Q^Tb\|_2^2 \\ &= \|Rx-c_1\|_2^2+\|c_2\|_2^2\end{aligned}$$

- x是原LS问题的解当且仅当x是$Rx=c_1$的解

- 那么

$$\begin{aligned}\|Ax-b\|_2^2 &= \|Q^TAx-Q^Tb\|_2^2\\ &= \|Rx-c_1\|_2^2+\|c_2\|_2^2\end{aligned}$$

- x是原LS问题的解当且仅当x是$Rx=c_1$的解
- LS问题的求解转化为很容易求解的上三角方程组求解

- 那么

$$\begin{aligned}\|Ax - b\|_2^2 &= \|Q^T Ax - Q^T b\|_2^2 \\ &= \|Rx - c_1\|_2^2 + \|c_2\|_2^2\end{aligned}$$

- x是原LS问题的解当且仅当x是$Rx = c_1$的解
- LS问题的求解转化为很容易求解的上三角方程组求解
- 问题的关键是如何实现QR分解

QR分解的Householder方法

- 用Householder方法计算QR分解与不选主元的Gauss消去法非常类似

QR分解的Householder方法

- 用Householder方法计算QR分解与不选主元的Gauss消去法非常类似
- 对一般矩阵$A \in \mathbb{R}^{m\times n}$, 假设我们已进行了$k-1$步，得到了Householder变换$H_1, \dots, H_{k-1}$, 使得

$$A_k = H_{k-1} \cdots H_1 A = \begin{pmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & A_{22}^{(k)} \end{pmatrix}$$

其中$A_{11}^{(k)}$是$k-1$阶上三角阵

第k步

- 设$A_{22}^{(k)} = (u_k, \ldots, u_n)$

第k步

- 设$A_{22}^{(k)} = (u_k, \ldots, u_n)$
- 在第k步中，先确定Householder变换：

$$\tilde{H}_k = I_{m-k+1} - \beta_k v_k v_k^T \in \mathbb{R}^{(m-k+1)\times(m-k+1)}$$

使得$\tilde{H}_k u_k = r_{kk} e_1$, 其中$r_{kk} \geqslant 0$

第k步

- 设$A_{22}^{(k)} = (u_k, \ldots, u_n)$
- 在第k步中，先确定Householder变换：

$$\tilde{H}_k = I_{m-k+1} - \beta_k v_k v_k^T \in \mathbb{R}^{(m-k+1)\times(m-k+1)}$$

使得$\tilde{H}_k u_k = r_{kk} e_1$，其中$r_{kk} \geqslant 0$

- 然后计算$\tilde{H}_k A_{22}^{(k)}$

- 令$H_k = \mathrm{diag}(I_{k-1}, \tilde{H}_k)$

- 令$H_k = \mathrm{diag}(I_{k-1}, \tilde{H}_k)$
- 则我们有

$$A_{k+1} = H_k A_k = \begin{pmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & \tilde{H}_k A_{22}^{(k)} \end{pmatrix}$$
$$= \begin{pmatrix} A_{11}^{(k+1)} & A_{12}^{(k+1)} \\ 0 & A_{22}^{(k+1)} \end{pmatrix}$$

其中$A_{11}^{(k+1)}$是上三角阵

- 如此从$k=1$出发，对A依次进行n次Householder变换，我们就可以将A约化为上三角阵

- 如此从$k=1$出发，对A依次进行n次Householder变换，我们就可以将A约化为上三角阵
- 记$R=A_{11}^{(n)}$, $Q=H_1\cdots H_n$, 则有

$$A=Q\begin{pmatrix} R \\ 0 \end{pmatrix}$$

- 如此从$k=1$出发，对A依次进行n次Householder变换，我们就可以将A约化为上三角阵
- 记$R=A_{11}^{(n)}$, $Q=H_1\cdots H_n$, 则有

$$A=Q\begin{pmatrix}R\\0\end{pmatrix}$$

- 如此得到的上三角阵R的对角元都是非负的

QR分解的存储

- 可以在A中存放Q与R

QR分解的存储

- 可以在A中存放Q与R
- 通常并不是将Q算出，而是只存放构成它的n个Householder变换H_k

QR分解的存储

- 可以在A中存放Q与R
- 通常并不是将Q算出，而是只存放构成它的n个Householder变换H_k
- 对于每个H_k，我们只需要保存v_k和β_k

QR分解的存储

- 可以在A中存放Q与R
- 通常并不是将Q算出，而是只存放构成它的n个Householder变换H_k
- 对于每个H_k, 我们只需要保存v_k和β_k
- $v_k = (1, *, \ldots, *)$，可把除首位的1外的元素存放在A的对角元以下位置上

QR分解的存储

- 可以在A中存放Q与R
- 通常并不是将Q算出，而是只存放构成它的n个Householder变换H_k
- 对于每个H_k, 我们只需要保存v_k和β_k
- $v_k = (1, *, \ldots, *)$，可把除首位的1外的元素存放在A的对角元以下位置上
- β_k存放在单独一个向量中

注解

- 算法的运算量为$2n^2(m-n/3)$

注解

- 算法的运算量为$2n^2(m-n/3)$
  - 当$m=n$时，LU分解相比于QR分解，运算量约只有一半

注解

- 算法的运算量为$2n^2(m-n/3)$
  - 当$m=n$时，LU分解相比于QR分解，运算量约只有一半
- 其数值性态良好

注解

- 算法的运算量为$2n^2(m-n/3)$
  - 当$m=n$时，LU分解相比于QR分解，运算量约只有一半
- 其数值性态良好
  - 对于正交阵，相互累积相乘时，结果矩阵的元素仍是有界的

注解

- 算法的运算量为$2n^2(m - n/3)$
  - 当$m = n$时，LU分解相比于QR分解，运算量约只有一半
- 其数值性态良好
  - 对于正交阵，相互累积相乘时，结果矩阵的元素仍是有界的
- 利用这一算法求解LS问题所得到的计算解通常要比正则化方法精确得多

注解

- 算法的运算量为$2n^2(m-n/3)$
  - 当$m=n$时，LU分解相比于QR分解，运算量约只有一半
- 其数值性态良好
  - 对于正交阵，相互累积相乘时，结果矩阵的元素仍是有界的
- 利用这一算法求解LS问题所得到的计算解通常要比正则化方法精确得多
- 当然付出的代价也是不容忽视的：$m \gg n$时

注解

- 也可以利用Givens变换或者Gram-Schimidt正交化实现QR分解

注解

- 也可以利用Givens变换或者Gram-Schimidt正交化实现QR分解
- 通常Givens变换来实现QR分解的运算量大约是Householder方法的两倍。但如果A稀疏，则使用Givens变换可能会比较有效

注解

- 也可以利用Givens变换或者Gram-Schimidt正交化实现QR分解
- 通常Givens变换来实现QR分解的运算量大约是Householder方法的两倍。但如果A稀疏，则使用Givens变换可能会比较有效
- 也可以用QR分解进行特征值求解或者解线性方程组，对病态方程组可能有效，但运算量大得多

线性方程组的古典迭代解法

邓建松

2018 年 10 月 29 日

直接法的局限

- 计算机的存储量日益增大，计算速度迅速提高，直接法（如Gauss消去法、平方根法）在计算机上可以求解的线性方程组的规模也越来越大

直接法的局限

- 计算机的存储量日益增大，计算速度迅速提高，直接法（如Gauss消去法、平方根法）在计算机上可以求解的线性方程组的规模也越来越大
- 在实际应用中，特别是偏微分方程的数值求解时，通常遇到的就是大型稀疏线性方程组

直接法的局限

- 计算机的存储量日益增大，计算速度迅速提高，直接法（如Gauss消去法、平方根法）在计算机上可以求解的线性方程组的规模也越来越大
- 在实际应用中，特别是偏微分方程的数值求解时，通常遇到的就是大型稀疏线性方程组
- 而直接法在对矩阵进行分解的时候，会破坏矩阵的稀疏性

直接法的局限

- 计算机的存储量日益增大，计算速度迅速提高，直接法（如Gauss消去法、平方根法）在计算机上可以求解的线性方程组的规模也越来越大
- 在实际应用中，特别是偏微分方程的数值求解时，通常遇到的就是大型稀疏线性方程组
- 而直接法在对矩阵进行分解的时候，会破坏矩阵的稀疏性
- 寻求能够保持稀疏性的有效算法是数值线性代数中一个重要的研究课题

求解稀疏线性方程组的方法

主要有两类

- 迭代法：按照某种规则构造一个向量序列，其极限是方程组的精确解

求解稀疏线性方程组的方法

主要有两类

- 迭代法：按照某种规则构造一个向量序列，其极限是方程组的精确解
- 稀疏直接法：是直接法与某些稀疏矩阵技巧有机结合的结果，利用矩阵的特点，使得分解结果尽可能保持稀疏性

求解稀疏线性方程组的方法

主要有两类

- 迭代法：按照某种规则构造一个向量序列，其极限是方程组的精确解
- 稀疏直接法：是直接法与某些稀疏矩阵技巧有机结合的结果，利用矩阵的特点，使得分解结果尽可能保持稀疏性
- 本课程只讲迭代法

迭代法中的问题

- 如何构造迭代序列？

迭代法中的问题

- 如何构造迭代序列？
- 构造的序列是否收敛？在什么情况下收敛？

迭代法中的问题

- 如何构造迭代序列？
- 构造的序列是否收敛？在什么情况下收敛？
- 如果收敛，收敛速度如何？（收敛速度的定量刻划）

迭代法中的问题

- 如何构造迭代序列？
- 构造的序列是否收敛？在什么情况下收敛？
- 如果收敛，收敛速度如何？（收敛速度的定量刻划）
- 迭代有限步停止。需要对近似解进行误差估计和舍入误差分析

迭代法有效性

- 方法是否有效要看得到具有某个精度的近似解而付出的代价如何

迭代法有效性

- 方法是否有效要看得到具有某个精度的近似解而付出的代价如何
- 这通常是以运算量和存储量的要求为标准

迭代法有效性

- 方法是否有效要看得到具有某个精度的近似解而付出的代价如何
- 这通常是以运算量和存储量的要求为标准
- 在这个标准下，很多时候直接法要比迭代法好

迭代法有效性

- 方法是否有效要看得到具有某个精度的近似解而付出的代价如何
- 这通常是以运算量和存储量的要求为标准
- 在这个标准下，很多时候直接法要比迭代法好
- 但对大型稀疏方程组来说，迭代法更实用

Jacobi迭代法

- 考虑非奇异线性方程组$Ax = b$

Jacobi迭代法

- 考虑非奇异线性方程组$Ax = b$
- 令$A = D - L - U$, 其中D为A的对角元构成的对角阵，$-L$为A的下三角阵，$-U$为A的上三角阵（均不含对角元）

Jacobi迭代法

- 考虑非奇异线性方程组$Ax = b$
- 令$A = D - L - U$, 其中D为A的对角元构成的对角阵，$-L$为A的下三角阵，$-U$为A的上三角阵（均不含对角元）
- 则$Ax = b$可写为$x = Bx + g$, 其中$B = D^{-1}(L + U)$, $g = D^{-1}b$

迭代格式

- 取初始向量

$$x_0 = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^T$$

迭代格式

- 取初始向量

$$x_0 = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^T$$

- 代入$x = Bx + g$的右边，得到新向量

$$x_1 = Bx_0 + g$$

迭代格式

- 取初始向量

$$x_0 = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^T$$

- 代入$x = Bx + g$的右边，得到新向量

$$x_1 = Bx_0 + g$$

- 再把x_1代入右边，又得一个新向量x_2

迭代格式

- 取初始向量

$$x_0 = (x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)})^T$$

- 代入$x = Bx + g$的右边，得到新向量

$$x_1 = Bx_0 + g$$

- 再把x_1代入右边，又得一个新向量x_2
- 依此类推，我们有$x_k = Bx_{k-1} + g$, $k = 1, 2, \ldots$

迭代矩阵

- 这就是Jacobi迭代法，由C.G.J. Jacobi提出

迭代矩阵

- 这就是Jacobi迭代法，由C.G.J. Jacobi提出
- 其中B称为Jacobi迭代法的迭代矩阵，其对角元全是零

迭代矩阵

- 这就是Jacobi迭代法，由C.G.J. Jacobi提出
- 其中B称为Jacobi迭代法的迭代矩阵，其对角元全是零
- g称为Jacobi迭代法的常数项

例

- 设$A=\begin{pmatrix} 2 & 1 \\ 5 & 7 \end{pmatrix}$, $b=\begin{pmatrix} 11 \\ 13 \end{pmatrix}$, $x_0=\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

例

- 设$A = \begin{pmatrix} 2 & 1 \\ 5 & 7 \end{pmatrix}$, $b = \begin{pmatrix} 11 \\ 13 \end{pmatrix}$, $x_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
- 则$D^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/7 \end{pmatrix}$, $L = \begin{pmatrix} 0 & 0 \\ -5 & 0 \end{pmatrix}$, $U = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}$

例

- 设$A=\begin{pmatrix}2 & 1\\ 5 & 7\end{pmatrix}$, $b=\begin{pmatrix}11\\ 13\end{pmatrix}$, $x_0=\begin{pmatrix}1\\ 1\end{pmatrix}$
- 则$D^{-1}=\begin{pmatrix}1/2 & 0\\ 0 & 1/7\end{pmatrix}$, $L=\begin{pmatrix}0 & 0\\ -5 & 0\end{pmatrix}$, $U=\begin{pmatrix}0 & -1\\ 0 & 0\end{pmatrix}$
- 迭代矩阵和常数项分别为

$$B=\begin{pmatrix}0 & -1/2\\ -5/7 & 0\end{pmatrix},\quad g=\begin{pmatrix}11/2\\ 13/7\end{pmatrix}$$

例

- 迭代得到

$$x_1 = \begin{pmatrix} 5 \\ 8/7 \end{pmatrix} \approx \begin{pmatrix} 5 \\ 1.143 \end{pmatrix},$$

$$x_2 = \begin{pmatrix} 69/14 \\ -12/7 \end{pmatrix} \approx \begin{pmatrix} 4.929 \\ -1.714 \end{pmatrix}$$

例

- 迭代得到

$$x_1 = \begin{pmatrix} 5 \\ 8/7 \end{pmatrix} \approx \begin{pmatrix} 5 \\ 1.143 \end{pmatrix},$$
$$x_2 = \begin{pmatrix} 69/14 \\ -12/7 \end{pmatrix} \approx \begin{pmatrix} 4.929 \\ -1.714 \end{pmatrix}$$

- 25次迭代后，得到$x \approx \begin{pmatrix} 7.111 \\ -3.222 \end{pmatrix}$, 这约等于方程的准确解$\begin{pmatrix} 64/9 \\ -29/9 \end{pmatrix}$

Jacobi迭代法代码片段


```
for i=1 to n
  y[i]=0.0
  for j=1 to n
    y[i]=y[i]+B[i][j]*x[j]
  y[i]=y[i]+g[i]
x=y
```

Jacobi迭代法中分量计算顺序

- 在Jacobi迭代法中各分量的计算顺序是没有关系的

Jacobi迭代法中分量计算顺序

- 在Jacobi迭代法中各分量的计算顺序是没有关系的
- 先算哪个分量，结果都不变

Jacobi迭代法中分量计算顺序

- 在Jacobi迭代法中各分量的计算顺序是没有关系的
- 先算哪个分量，结果都不变
- 我们做一下改变：在计算x_k的第一个分量用x_{k-1}的各个分量计算，但当计算x_k的后面分量时，采用已算出的新分量$x_1^{(k)}$代替$x_1^{(k-1)}$，而其它分量仍用$x_i^{(k-1)}$

新代码片段


```
for i=1 to n
  x[i]=0.0
  for j=1 to n
    x[i]=x[i]+B[i][j]*x[j]
  x[i]=x[i]+g[i]
```

新格式

- 经上述改变后，我们有

$$x_k = D^{-1}Lx_k + D^{-1}Ux_{k-1} + g, \quad k = 1, 2, \ldots$$

新格式

- 经上述改变后，我们有

$$x_k = D^{-1}Lx_k + D^{-1}Ux_{k-1} + g, \quad k = 1, 2, \ldots$$

- 这称为Gauss-Seidel迭代法，简称为G-S迭代法

新格式

- 经上述改变后，我们有

$$x_k = D^{-1}Lx_k + D^{-1}Ux_{k-1} + g, \quad k = 1, 2, \ldots$$

- 这称为Gauss-Seidel迭代法，简称为G-S迭代法
- 如此变化，编程时存储量减少了

新格式

- 经上述改变后，我们有

$$x_k = D^{-1}Lx_k + D^{-1}Ux_{k-1} + g, \quad k = 1, 2, \ldots$$

- 这称为Gauss-Seidel迭代法，简称为G-S迭代法
- 如此变化，编程时存储量减少了
- 该格式1823年C.F. Gauss在给其学生C.L. Gerling的信中提到，P.L. von Seidel在1874年发表了这一方法

迭代矩阵

- 如果$(D-L)^{-1}$存在，那么迭代格式为

$$x_k = (D-L)^{-1}Ux_{k-1} + (D-L)^{-1}b$$

迭代矩阵

- 如果$(D-L)^{-1}$存在，那么迭代格式为

$$x_k = (D-L)^{-1}Ux_{k-1} + (D-L)^{-1}b$$

- 我们称$L_1 = (D-L)^{-1}U$为G-S迭代法的迭代矩阵，其第一列全是零，$(D-L)^{-1}b$称为G-S迭代法的常数项

迭代矩阵

- 如果$(D-L)^{-1}$存在，那么迭代格式为

$$x_k = (D-L)^{-1}Ux_{k-1} + (D-L)^{-1}b$$

- 我们称$L_1 = (D-L)^{-1}U$为G-S迭代法的迭代矩阵，其第一列全是零，$(D-L)^{-1}b$称为G-S迭代法的常数项
- 此时分量的计算次序是不能改变的

例

- 设$A = \begin{pmatrix} 16 & 3 \\ 7 & -11 \end{pmatrix}$, $b = \begin{pmatrix} 11 \\ 13 \end{pmatrix}$

例

- 设$A = \begin{pmatrix} 16 & 3 \\ 7 & -11 \end{pmatrix}$, $b = \begin{pmatrix} 11 \\ 13 \end{pmatrix}$
- 则$(D-L)^{-1} = \begin{pmatrix} 1/16 & 0 \\ 7/176 & -1/11 \end{pmatrix}$

例

- 设$A = \begin{pmatrix} 16 & 3 \\ 7 & -11 \end{pmatrix}$, $b = \begin{pmatrix} 11 \\ 13 \end{pmatrix}$
- 则$(D-L)^{-1} = \begin{pmatrix} 1/16 & 0 \\ 7/176 & -1/11 \end{pmatrix}$
- 迭代矩阵和常数项分别为

$$L_1 = \begin{pmatrix} 0 & -3/16 \\ 0 & -21/176 \end{pmatrix}, \quad g = \begin{pmatrix} 11/16 \\ -131/176 \end{pmatrix}$$

例

- 取初值$x_0 = (1, 1)^T$

例

- 取初值$x_0 = (1, 1)^T$
- 迭代得到各向量依次为

$$\begin{pmatrix} 0.5 \\ -0.8636 \end{pmatrix} \to \begin{pmatrix} 0.8494 \\ -0.6413 \end{pmatrix} \to \begin{pmatrix} 0.8077 \\ -0.6678 \end{pmatrix}$$

例

- 取初值$x_0 = (1, 1)^T$
- 迭代得到各向量依次为

$$\begin{pmatrix} 0.5 \\ -0.8636 \end{pmatrix} \to \begin{pmatrix} 0.8494 \\ -0.6413 \end{pmatrix} \to \begin{pmatrix} 0.8077 \\ -0.6678 \end{pmatrix}$$

- 经六次迭代，得到向量$\begin{pmatrix} 0.8122 \\ -0.6650 \end{pmatrix}$

例

- 取初值$x_0 = (1,1)^T$
- 迭代得到各向量依次为

$$\begin{pmatrix} 0.5 \\ -0.8636 \end{pmatrix} \to \begin{pmatrix} 0.8494 \\ -0.6413 \end{pmatrix} \to \begin{pmatrix} 0.8077 \\ -0.6678 \end{pmatrix}$$

- 经六次迭代，得到向量$\begin{pmatrix} 0.8122 \\ -0.6650 \end{pmatrix}$
- 方程的准确解为

$$\begin{pmatrix} 160/197 \\ -131/197 \end{pmatrix} \approx \begin{pmatrix} 0.812183 \\ -0.664975 \end{pmatrix}$$

两种方法的共性

- 在这两种方法中，新的近似解x_k是已知近似解x_{k-1}的线性函数，并且只与x_{k-1}有关，即它们都可以写为

$$x_k = Mx_{k-1} + g$$

两种方法的共性

- 在这两种方法中，新的近似解x_k是已知近似解x_{k-1}的线性函数，并且只与x_{k-1}有关，即它们都可以写为

$$x_k = Mx_{k-1} + g$$

  - Jacobi迭代法：$M = D^{-1}(L+U)$, $g = D^{-1}b$

两种方法的共性

- 在这两种方法中，新的近似解x_k是已知近似解x_{k-1}的线性函数，并且只与x_{k-1}有关，即它们都可以写为

$$x_k = Mx_{k-1} + g$$

  - Jacobi迭代法：$M = D^{-1}(L+U)$, $g = D^{-1}b$
  - G-S迭代法：$M = (D-L)^{-1}U$, $g = (D-L)^{-1}b$

定义

- 我们把$x_k = Mx_{k-1} + g$的迭代法称为单步线性定常迭代法

定义

- 我们把$x_k = Mx_{k-1} + g$的迭代法称为单步线性定常迭代法
  - $M \in \mathbb{R}^{n\times n}$称为迭代矩阵

定义

- 我们把$x_k = Mx_{k-1} + g$的迭代法称为单步线性定常迭代法
  - $M \in \mathbb{R}^{n \times n}$称为迭代矩阵
  - $g \in \mathbb{R}^n$称为常数项

定义

- 我们把$x_k = Mx_{k-1} + g$的迭代法称为单步线性定常迭代法
  - $M \in \mathbb{R}^{n\times n}$称为迭代矩阵
  - $g \in \mathbb{R}^n$称为常数项
  - $x_0 \in \mathbb{R}^n$称为初始向量

收敛与发散

- 如果迭代格式对任意的初始向量产生的向量序列$\{x_k\}$（称为迭代序列）都有极限，则称该迭代格式是收敛的

收敛与发散

- 如果迭代格式对任意的初始向量产生的向量序列$\{x_k\}$（称为迭代序列）都有极限，则称该迭代格式是收敛的
- 否则称它是不收敛的，或者是发散的

收敛与发散

- 如果迭代格式对任意的初始向量产生的向量序列$\{x_k\}$（称为迭代序列）都有极限，则称该迭代格式是收敛的
- 否则称它是不收敛的，或者是发散的
- 若收敛，记迭代序列的极限为x_*，则有$x_* = Mx_* + g$

迭代法与方程组的相容性

- x_*是方程组$(I-M)x=g$的解

迭代法与方程组的相容性

- x_*是方程组$(I-M)x=g$的解
- 如果存在非奇异矩阵$G\in\mathbb{R}^{n\times n}$使得

$$G(I-M)=A,\quad Gg=b$$

则迭代序列也收敛到$Ax=b$的解，此时两个方程称为等价方程组，也称迭代格式与方程组$Ax=b$是相容的

迭代法与方程组的相容性

- x_*是方程组$(I-M)x=g$的解
- 如果存在非奇异矩阵$G\in\mathbb{R}^{n\times n}$使得

$$G(I-M)=A,\quad Gg=b$$

则迭代序列也收敛到$Ax=b$的解，此时两个方程称为等价方程组，也称迭代格式与方程组$Ax=b$是相容的

- 已有的两种格式是相容的

误差向量

- 设x_*是$Ax = b$的解

误差向量

- 设x_*是$Ax = b$的解
- $\{x_k\}$是由相容格式$x_k = Mx_{k-1} + g$产生的迭代序列

误差向量

- 设x_*是$Ax = b$的解
- $\{x_k\}$是由相容格式$x_k = Mx_{k-1} + g$产生的迭代序列
- 定义$y_k = x_k - x_*$，称为x_k的误差向量

误差向量

- 设x_*是$Ax=b$的解
- $\{x_k\}$是由相容格式$x_k=Mx_{k-1}+g$产生的迭代序列
- 定义$y_k=x_k-x_*$, 称为x_k的误差向量
- 那么

$$\begin{aligned} y_{k+1} = x_{k+1} - x_* &= Mx_k + g - (Mx_* + g) \\ &= My_k = M^k y_0 \end{aligned}$$

收敛条件

- 对任意y_0都有$y_k \to 0$（即$x_k \to x_*$）当且仅当$M^k \to 0$

收敛条件

- 对任意y_0都有$y_k \to 0$（即$x_k \to x_*$）当且仅当$M^k \to 0$
- 而$M^k \to 0$的充要条件是$\rho(M) < 1$

收敛条件

- 对任意y_0都有$y_k \to 0$（即$x_k \to x_*$）当且仅当$M^k \to 0$
- 而$M^k \to 0$的充要条件是$\rho(M) < 1$

定理

解方程组$Ax = b$的单步线性定常迭代法$x_k = Mx_{k-1} + g$收敛的充要条件是迭代矩阵M的谱半径小于*1*

注解

- 迭代序列收敛取决于迭代矩阵的谱半径，而与初始向量的选取和常数项无关

注解

- 迭代序列收敛取决于迭代矩阵的谱半径，而与初始向量的选取和常数项无关
- 相同的方程组，Jacobi迭代矩阵和G-S迭代矩阵的谱半径不一定相同，而且无包含关系

注解

- 迭代序列收敛取决于迭代矩阵的谱半径，而与初始向量的选取和常数项无关
- 相同的方程组，Jacobi迭代矩阵和G-S迭代矩阵的谱半径不一定相同，而且无包含关系
- 例子: `Mathematica sec4.2.1.nb`

- 用谱半径判断迭代法是否收敛，这是很不方便的：谱半径的计算相当困难

- 用谱半径判断迭代法是否收敛，这是很不方便的：谱半径的计算相当困难
- 需要一些容易计算的充分条件

- 用谱半径判断迭代法是否收敛，这是很不方便的：谱半径的计算相当困难
- 需要一些容易计算的充分条件

定理

如果范数$\|M\| = q < 1$，则我们有

$$\|x_k - x_*\| \leqslant \frac{q^k}{1-q}\|x_1 - x_0\|$$

- 用谱半径判断迭代法是否收敛，这是很不方便的：谱半径的计算相当困难
- 需要一些容易计算的充分条件

定理

如果范数$\|M\| = q < 1$，则我们有

$$\|x_k - x_*\| \leqslant \frac{q^k}{1-q}\|x_1 - x_0\|$$

- 此定理结论的右边与精确解无关

定理证明

- 在$y_k = M^k y_0$的两边取范数，则有

$$\|y_k\| \leqslant \|M\|^k \|y_0\| = q^k \|y_0\|$$

定理证明

- 在$y_k = M^k y_0$的两边取范数，则有
$$\|y_k\| \leqslant \|M\|^k \|y_0\| = q^k \|y_0\|$$
- 根据y_0的定义
$$\begin{aligned}\|y_0\| &= \|x_0 - x_*\| \leqslant \|x_0 - x_1\| + \|x_1 - x_*\| \\ &= \|x_0 - x_1\| + \|My_0\| \leqslant \|x_0 - x_1\| + q\|y_0\|\end{aligned}$$
所以有$\|y_0\| \leqslant \dfrac{1}{1-q}\|x_1 - x_0\|$

注解

- 从上述近似解的误差估计可以计算出为了达到精度要求，需要进行多少次迭代

注解

- 从上述近似解的误差估计可以计算出为了达到精度要求，需要进行多少次迭代
- 这种估计一般是偏高的

注解

- 从上述近似解的误差估计可以计算出为了达到精度要求，需要进行多少次迭代
- 这种估计一般是偏高的

定理

若$\|M\| = q < 1$, 则我们有

$$\|x_k - x_*\| \leqslant \frac{q}{1-q}\|x_{k-1} - x_k\|$$

注解

- 从上述近似解的误差估计可以计算出为了达到精度要求，需要进行多少次迭代
- 这种估计一般是偏高的

定理

若$\|M\| = q < 1$，则我们有

$$\|x_k - x_*\| \leqslant \frac{q}{1-q}\|x_{k-1} - x_k\|$$

- 此定理结论是用刚得到的两个迭代向量估计最新结果的精度

定理证明

- 因为

$$
\begin{aligned}
\|x_k - x_*\| = &\|M(x_{k-1} - x_*)\| \leqslant q\|x_{k-1} - x_*\| \\
&\leqslant q\|x_{k-1} - x_k\| + q\|x_k - x_*\|
\end{aligned}
$$

所以

$$
\|x_k - x_*\| \leqslant \frac{q}{1-q}\|x_{k-1} - x_k\|
$$

注解

- 只要M的范数不是很接近1，当相邻两次迭代向量很接近时，那么x_k与x_*也就很接近

注解

- 只要M的范数不是很接近1，当相邻两次迭代向量很接近时，那么x_k与x_*也就很接近
- 实际计算时可以用量$\|x_{k-1} - x_k\|$是否适当小来判别迭代法是否应该终止

注解

- 只要M的范数不是很接近1，当相邻两次迭代向量很接近时，那么x_k与x_*也就很接近
- 实际计算时可以用量$\|x_{k-1}-x_k\|$是否适当小来判别迭代法是否应该终止
  - 若$\|M\|=0.9$, $\|x_{k-1}-x_k\|=10^{-8}$, 则$\|x_k-x_*\|\leqslant 9\times 10^{-8}$

注解

- 只要M的范数不是很接近1，当相邻两次迭代向量很接近时，那么x_k与x_*也就很接近
- 实际计算时可以用量$\|x_{k-1}-x_k\|$是否适当小来判别迭代法是否应该终止
  - 若$\|M\|=0.9$, $\|x_{k-1}-x_k\|=10^{-8}$, 则$\|x_k-x_*\|\leqslant 9\times 10^{-8}$
- 若$\|M\|$很接近1, 则不能断定精度

注解

- 只要M的范数不是很接近1，当相邻两次迭代向量很接近时，那么x_k与x_*也就很接近
- 实际计算时可以用量$\|x_{k-1} - x_k\|$是否适当小来判别迭代法是否应该终止
  - 若$\|M\| = 0.9$, $\|x_{k-1} - x_k\| = 10^{-8}$, 则$\|x_k - x_*\| \leqslant 9 \times 10^{-8}$
- 若$\|M\|$很接近1, 则不能断定精度
- 实际一般应用1范数和∞范数：容易计算

分析

- Jacobi迭代法的迭代矩阵B容易得到，因此前面的判别法基本令人满意

分析

- Jacobi迭代法的迭代矩阵B容易得到，因此前面的判别法基本令人满意
- G-S迭代法中的迭代矩阵需要求$L_1 = (D - L)^{-1}U$, 不太容易得到

分析

- Jacobi迭代法的迭代矩阵B容易得到，因此前面的判别法基本令人满意
- G-S迭代法中的迭代矩阵需要求$L_1 = (D - L)^{-1}U$，不太容易得到
- 因此我们需要给出一些辅助判别

分析

- Jacobi迭代法的迭代矩阵B容易得到，因此前面的判别法基本令人满意
- G-S迭代法中的迭代矩阵需要求$L_1 = (D-L)^{-1}U$, 不太容易得到
- 因此我们需要给出一些辅助判别
  - 能否从B的性质判断L_1的性质？

分析

- Jacobi迭代法的迭代矩阵B容易得到，因此前面的判别法基本令人满意
- G-S迭代法中的迭代矩阵需要求$L_1 = (D - L)^{-1}U$, 不太容易得到
- 因此我们需要给出一些辅助判别
  - 能否从B的性质判断L_1的性质？
  - 能否直接利用A的性质进行判断？

B与L_1的关系

- 回忆

B与L_1的关系

- 回忆
  - Jacobi迭代：$y = D^{-1}Lx + D^{-1}Ux$

B与L_1的关系

- 回忆
  - Jacobi迭代：$y = D^{-1}Lx + D^{-1}Ux$
  - G-S迭代：$y = D^{-1}Ly + D^{-1}Ux$

B与L_1的关系

- 回忆
  - Jacobi迭代：$y = D^{-1}Lx + D^{-1}Ux$
  - G-S迭代：$y = D^{-1}Ly + D^{-1}Ux$
  - $D^{-1}L$为下三角矩阵，对角元为零；$D^{-1}U$为上三角矩阵，对角元为零

B与L_1的关系

- 回忆
  - Jacobi迭代：$y = D^{-1}Lx + D^{-1}Ux$
  - G-S迭代：$y = D^{-1}Ly + D^{-1}Ux$
  - $D^{-1}L$为下三角矩阵，对角元为零；$D^{-1}U$为上三角矩阵，对角元为零
- 所以我们完全有理由根据B的性质推断L_1的性质

∞范数

- 设$\|B\|_\infty < 1$

∞范数

- 设$\|B\|_\infty < 1$
- 定义

$$\mu = \max_i \frac{\sum_{j=i+1}^{n} |b_{ij}|}{1 - \sum_{j=1}^{i-1} |b_{ij}|}$$

∞范数

- 设$\|B\|_\infty < 1$
- 定义

$$\mu = \max_i \frac{\sum_{j=i+1}^{n} |b_{ij}|}{1 - \sum_{j=1}^{i-1} |b_{ij}|}$$

引理

$$\mu \leqslant \|B\|_\infty < 1$$

证明

- 令$\ell_i = \sum_{j=1}^{i-1} |b_{ij}|,\ u_i = \sum_{j=i+1}^{n} |b_{ij}|$

证明

- 令$\ell_i = \sum_{j=1}^{i-1} |b_{ij}|,\ u_i = \sum_{j=i+1}^{n} |b_{ij}|$
- 则$\ell_i + u_i \leqslant \|B\|_\infty < 1$

证明

- 令$\ell_i = \sum_{j=1}^{i-1} |b_{ij}|$, $u_i = \sum_{j=i+1}^{n} |b_{ij}|$
- 则$\ell_i + u_i \leqslant \|B\|_\infty < 1$
- 注意到$\forall i$, $b_{ii} = 0$, 所以存在i, $\ell_i + u_i = \|B\|_\infty$

- 注意到

$$\begin{aligned}\ell_i + u_i - \frac{u_i}{1-\ell_i} &= \frac{1}{1-\ell_i}((\ell_i+u_i)(1-\ell_i)-u_i) \\ &= \frac{\ell_i}{1-\ell_i}(1-\ell_i-u_i) \geqslant 0\end{aligned}$$

- 注意到

$$\ell_i + u_i - \frac{u_i}{1-\ell_i} = \frac{1}{1-\ell_i}((\ell_i+u_i)(1-\ell_i)-u_i) = \frac{\ell_i}{1-\ell_i}(1-\ell_i-u_i) \geqslant 0$$

- 所以我们有 $\dfrac{u_i}{1-\ell_i} \leqslant \ell_i + u_i$

- 注意到

$$\begin{aligned}\ell_i + u_i - \frac{u_i}{1-\ell_i} &= \frac{1}{1-\ell_i}((\ell_i+u_i)(1-\ell_i)-u_i)\\ &= \frac{\ell_i}{1-\ell_i}(1-\ell_i-u_i) \geqslant 0\end{aligned}$$

- 所以我们有$\dfrac{u_i}{1-\ell_i} \leqslant \ell_i + u_i$
- 两边对i取最大值，我们得到

$$\mu = \max_i \frac{u_i}{1-\ell_i} \leqslant \max_i(\ell_i+u_i) = \|B\|_\infty < 1$$

$\|B\|_\infty$与$\|L_1\|_\infty$

定理

设$\|B\|_\infty < 1$，则$\|L_1\|_\infty \leqslant \|B\|_\infty < 1$，而且由G-S迭代法产生的近似解$x_k$与准确解$x_*$之间满足

$$\|x_k - x_*\|_\infty \leqslant \frac{\mu^k}{1-\mu}\|x_1 - x_0\|_\infty$$

定理证明

- 存在满足$\|x\|_\infty = 1$的$x \in \mathbb{R}^n$使得$\|L_1\|_\infty = \|L_1 x\|_\infty$

定理证明

- 存在满足$\|x\|_\infty = 1$的$x \in \mathbb{R}^n$使得$\|L_1\|_\infty = \|L_1 x\|_\infty$
- 令$y = L_1 x$, $|y_i| = \|y\|_\infty$

定理证明

- 存在满足$\|x\|_\infty = 1$的$x \in \mathbb{R}^n$使得$\|L_1\|_\infty = \|L_1 x\|_\infty$
- 令$y = L_1 x$, $|y_i| = \|y\|_\infty$
- 根据$L_1 = (D-L)^{-1}U$的定义，我们有$y = D^{-1}Ly + D^{-1}Ux$

定理证明

- 存在满足$\|x\|_\infty = 1$的$x \in \mathbb{R}^n$使得$\|L_1\|_\infty = \|L_1 x\|_\infty$
- 令$y = L_1 x$, $|y_i| = \|y\|_\infty$
- 根据$L_1 = (D-L)^{-1}U$的定义，我们有$y = D^{-1}Ly + D^{-1}Ux$
- 由于$B = D^{-1}L + D^{-1}U$, 所以$D^{-1}L$为B的下三角部分，$D^{-1}U$为B的上三角部分

- 比较两边的第i个分量，我们有

$$y_i = \sum_{j=1}^{i-1} b_{ij} y_j + \sum_{j=i+1}^{n} b_{ij} x_j$$

- 比较两边的第i个分量，我们有

$$y_i = \sum_{j=1}^{i-1} b_{ij} y_j + \sum_{j=i+1}^{n} b_{ij} x_j$$

- 两边取绝对值，可得

$$\|y\|_\infty \leqslant \|y\|_\infty \ell_i + u_i$$

- 比较两边的第i个分量，我们有

$$y_i = \sum_{j=1}^{i-1} b_{ij} y_j + \sum_{j=i+1}^{n} b_{ij} x_j$$

- 两边取绝对值，可得

$$\|y\|_\infty \leqslant \|y\|_\infty \ell_i + u_i$$

- 由此可得

$$\|L_1\|_\infty = \|y\|_\infty \leqslant \frac{u_i}{1-\ell_i} \leqslant \mu < 1$$

估计式的证明

- 根据$\|L_1\|_\infty \leqslant \mu < 1$可得

$$\begin{aligned}\|x_k - x_*\|_\infty &\leqslant \frac{\|L_1\|_\infty^k}{1-\|L_1\|_\infty}\|x_1 - x_0\|_\infty \\ &\leqslant \frac{\mu^k}{1-\mu}\|x_1 - x_0\|_\infty\end{aligned}$$

1范数

- 设$\|B\|_1 < 1$

线性方程组的古典迭代解法

邓建松

单步线性定常迭代法

Jacobi迭代法

Gauss-Seidel迭代法

形式推广

收敛性理论

收敛的充要条件

收敛的充分条件及误差估计

Jacobi迭代法和G-S迭代法的收敛性

收敛速度

平均收敛速度和渐近收敛速度

模型问题

Jacobi迭代法和G-S迭代法的渐近收敛速度

1范数

- 设$\|B\|_1 < 1$
- 定义

$$\tilde{\mu} = \max_j \frac{\sum_{i=1}^{j-1} |b_{ij}|}{1 - \sum_{i=j+1}^{n} |b_{ij}|}$$

1范数

- 设$\|B\|_1 < 1$
- 定义

$$\tilde{\mu} = \max_j \frac{\sum_{i=1}^{j-1} |b_{ij}|}{1 - \sum_{i=j+1}^{n} |b_{ij}|}$$

- 则类似前面对$\mu < 1$的证明可证$\tilde{\mu} \leqslant \|B\|_1 < 1$

$\|B\|_1$与$\rho(L_1)$

定理

若$\|B\|_1 < 1$，则$\rho(L_1) \leqslant \|B\|_1 < 1$，而且由G-S迭代法所得近似解$x_k$与准确解$x_*$满足

$$\|x_k - x_*\|_1 \leqslant \frac{\tilde{\mu}^k}{(1-\tilde{\mu})(1-s)}\|x_1 - x_0\|_1$$

其中$s = \max_j \sum_{i=j+1}^{n} |b_{ij}|$，这是$B$的下三角阵$D^{-1}L$的1范数

$\rho(L_1) < 1$的证明

- 因为

$$(D-L)^{-1}U = (I-D^{-1}L)^{-1}(D^{-1}U(I-D^{-1}L)^{-1})(I-D^{-1}L)$$

所以L_1与$\tilde{L}_1 = D^{-1}U(I-D^{-1}L)^{-1}$相似，从而$\rho(L_1) = \rho(\tilde{L}_1)$

$\rho(L_1) < 1$的证明

- 因为
$$(D-L)^{-1}U = (I-D^{-1}L)^{-1}(D^{-1}U(I-D^{-1}L)^{-1})(I-D^{-1}L)$$
所以L_1与$\tilde{L}_1 = D^{-1}U(I-D^{-1}L)^{-1}$相似，从而$\rho(L_1) = \rho(\tilde{L}_1)$
- 类似∞范数时的情况可证$\|\tilde{L}_1^T\|_\infty \leqslant \tilde{\mu} < 1$, 从而有
$$\rho(L_1) \leqslant \|\tilde{L}_1^T\|_\infty \leqslant \tilde{\mu} < 1$$

- 实际上，令$C = B^T$. $\|B\|_1 < 1 \Longrightarrow \|C\|_\infty < 1$

- 实际上，令$C = B^T$. $\|B\|_1 < 1 \Longrightarrow \|C\|_\infty < 1$
- C的下三角部分为$(D^{-1}U)^T = U^TD^{-1}$, 上三角部分为$(D^{-1}L)^T = L^TD^{-1}$

- 实际上，令$C = B^T$. $\|B\|_1 < 1 \Longrightarrow \|C\|_\infty < 1$
- C的下三角部分为$(D^{-1}U)^T = U^T D^{-1}$, 上三角部分为$(D^{-1}L)^T = L^T D^{-1}$
- $\tilde{L}_1^T = (D^{-1}U(I - D^{-1}L)^{-1})^T = (I - L^T D^{-1})^{-1} U^T D^{-1}$

- 实际上，令$C = B^T$. $\|B\|_1 < 1 \Longrightarrow \|C\|_\infty < 1$
- C的下三角部分为$(D^{-1}U)^T = U^T D^{-1}$, 上三角部分为$(D^{-1}L)^T = L^T D^{-1}$
- $\tilde{L}_1^T = (D^{-1}U(I - D^{-1}L)^{-1})^T = (I - L^T D^{-1})^{-1} U^T D^{-1}$
- $\tilde{L}_1^T x = y \Longrightarrow y = L^T D^{-1} y + U^T D^{-1} x$

误差估计的证明

- 由G-S迭代法的格式，可得

$$x_k - x_{k-1} = D^{-1}L(x_k - x_{k-1}) + D^{-1}U(x_{k-1} - x_{k-2})$$

误差估计的证明

- 由G-S迭代法的格式，可得

$$x_k - x_{k-1} = D^{-1}L(x_k - x_{k-1}) + D^{-1}U(x_{k-1} - x_{k-2})$$

- 分量表示即为

$$x_i^{(k)} - x_i^{(k-1)} = \sum_{j=1}^{i-1} b_{ij}(x_j^{(k)} - x_j^{(k-1)}) + \sum_{j=i+1}^{n} b_{ij}(x_j^{(k-1)} - x_j^{(k-2)})$$

- 两边取绝对值后对i求和，再交换求和顺序

$$\begin{aligned}
\sum_{i=1}^{n}\left|x_i^{(k)}-x_i^{(k-1)}\right| &\leqslant \sum_{i=1}^{n}\left(\sum_{j=1}^{i-1}|b_{ij}|\left|x_j^{(k)}-x_j^{(k-1)}\right|\right.\\
&\qquad\left.+\sum_{j=i+1}^{n}|b_{ij}|\left|x_j^{(k-1)}-x_j^{(k-2)}\right|\right)\\
&\leqslant \sum_{j=1}^{n}\left(\sum_{i=j+1}^{n}|b_{ij}|\left|x_j^{(k)}-x_j^{(k-1)}\right|\right.\\
&\qquad\left.+\sum_{i=1}^{j-1}|b_{ij}|\left|x_j^{(k-1)}-x_j^{(k-2)}\right|\right)
\end{aligned}$$

- 令

$$\tilde{u}_j = \sum_{i=j+1}^{n} |b_{ij}|, \quad \tilde{\ell}_j = \sum_{i=1}^{j-1} |b_{ij}|$$

- 令

$$\tilde{u}_j = \sum_{i=j+1}^{n} |b_{ij}|, \quad \tilde{\ell}_j = \sum_{i=1}^{j-1} |b_{ij}|$$

- 则接前推导我们有

$$\sum_{i=1}^{n} \left| x_i^{(k)} - x_i^{(k-1)} \right| \leqslant \sum_{j=1}^{n} \left(\tilde{u}_j \left| x_j^{(k)} - x_j^{(k-1)} \right| + \tilde{\ell}_j \left| x_j^{(k-1)} - x_j^{(k-2)} \right| \right)$$

从而有

$$
\begin{aligned}
\sum_{j=1}^{n}(1-\tilde{u}_j)\left|x_j^{(k)}-x_j^{(k-1)}\right| &\leqslant \sum_{j=1}^{n}\tilde{\ell}_j\left|x_j^{(k-1)}-x_j^{(k-2)}\right| \\
&\leqslant \tilde{\mu}\sum_{j=1}^{n}(1-\tilde{u}_j)\left|x_j^{(k-1)}-x_j^{(k-2)}\right| \\
&\leqslant \cdots\cdots \\
&\leqslant \tilde{\mu}^{k-1}\sum_{j=1}^{n}(1-\tilde{u}_j)\left|x_j^{(1)}-x_j^{(0)}\right|
\end{aligned}
$$

- 根据$\tilde{\mu}$和s的定义，$1 - s \leqslant 1 - \tilde{u}_j < 1$

- 根据$\tilde{\mu}$和s的定义，$1-s \leqslant 1-\tilde{u}_j < 1$
- 所以

$$(1-s)\sum_{j=1}^{n}\left|x_j^{(k)}-x_j^{(k-1)}\right| \leqslant \tilde{\mu}^{k-1}\sum_{j=1}^{n}\left|x_j^{(1)}-x_j^{(0)}\right|$$

- 根据$\tilde{\mu}$和s的定义，$1-s \leqslant 1-\tilde{u}_j < 1$
- 所以

$$(1-s)\sum_{j=1}^{n}\left|x_j^{(k)}-x_j^{(k-1)}\right| \leqslant \tilde{\mu}^{k-1}\sum_{j=1}^{n}\left|x_j^{(1)}-x_j^{(0)}\right|$$

- 这就是$(1-s)\|x_k-x_{k-1}\|_1 \leqslant \tilde{\mu}^{k-1}\|x_1-x_0\|_1$

- 注意到$x_k - x_* = \sum_{i=k}^{\infty}(x_i - x_{i+1})$

- 注意到$x_k - x_* = \sum_{i=k}^{\infty}(x_i - x_{i+1})$
- 所以

$$
\begin{aligned}
\|x_k - x_*\|_1 &\leqslant \sum_{i=k}^{\infty} \|x_i - x_{i+1}\|_1 \\
&\leqslant \frac{1}{1-s}\sum_{i=k}^{\infty} \tilde{\mu}^i \|x_1 - x_0\|_1 \\
&\leqslant \frac{\tilde{\mu}^k}{(1-\tilde{\mu})(1-s)} \|x_1 - x_0\|_1
\end{aligned}
$$

从A直接判断：正定矩阵

定理

如果A是对称的，而且对角元全是正数，则*Jacobi*迭代法收敛的充要条件是A和$2D-A$都正定

定理证明

- 迭代矩阵$B = D^{-1}(L+U) = I - D^{-1}A$

定理证明

- 迭代矩阵$B = D^{-1}(L + U) = I - D^{-1}A$
- 由于对角元全是正数，所以

$$B = D^{-1/2}(I - D^{-1/2}AD^{-1/2})D^{1/2}$$

定理证明

- 迭代矩阵$B = D^{-1}(L + U) = I - D^{-1}A$
- 由于对角元全是正数，所以

$$B = D^{-1/2}(I - D^{-1/2}AD^{-1/2})D^{1/2}$$

- A对称，所以$I - D^{-1/2}AD^{-1/2}$也是对称的，而且它与B相似，所以B的特征值全为实数

- Jacobi迭代法收敛$\Longleftrightarrow \rho(B) < 1$. 而$B$的特征值全为实数，所以这又等价于$I - B$和$I + B$的特征值均为正实数

- Jacobi迭代法收敛$\Longleftrightarrow \rho(B) < 1$. 而$B$的特征值全为实数，所以这又等价于$I-B$和$I+B$的特征值均为正实数
- 注意到

$$I - B = D^{-1/2}(D^{-1/2}AD^{-1/2})D^{1/2}$$

$$I + B = D^{-1/2}(2I - D^{-1/2}AD^{-1/2})D^{1/2}$$

线性方程组的古典迭代解法

邓建松

单步线性定常迭代法

Jacobi迭代法

Gauss-Seidel迭代法

形式推广

收敛性理论

收敛的充要条件

收敛的充分条件及误差估计

Jacobi迭代法和G-S迭代法的收敛性

收敛速度

平均收敛速度和渐近收敛速度

模型问题

Jacobi迭代法和G-S迭代法的渐近收敛速度

- Jacobi迭代法收敛$\Longleftrightarrow \rho(B) < 1$. 而$B$的特征值全为实数，所以这又等价于$I - B$和$I + B$的特征值均为正实数
- 注意到

$$I - B = D^{-1/2}(D^{-1/2}AD^{-1/2})D^{1/2}$$

$$I + B = D^{-1/2}(2I - D^{-1/2}AD^{-1/2})D^{1/2}$$

- 所以$\rho(B) < 1$等价于A与$2D - A$均正定

- Jacobi迭代法收敛$\Longleftrightarrow \rho(B) < 1$. 而$B$的特征值全为实数，所以这又等价于$I - B$和$I + B$的特征值均为正实数
- 注意到

$$I - B = D^{-1/2}(D^{-1/2}AD^{-1/2})D^{1/2}$$

$$I + B = D^{-1/2}(2I - D^{-1/2}AD^{-1/2})D^{1/2}$$

- 所以$\rho(B) < 1$等价于A与$2D - A$均正定
    - $D^{-1/2}AD^{-1/2}$相合于A;
      $2I - D^{-1/2}AD^{-1/2}$相合于$2D - A$

对称正定：G-S迭代

- 正定理

对称正定：G-S迭代

- 正定理

定理

若系数矩阵A对称正定，则G-S迭代法收敛

对称正定：G-S迭代

- 正定理

定理

若系数矩阵A对称正定，则G-S迭代法收敛

- 定理证明见第4节

对称正定：G-S迭代

- 正定理

定理

若系数矩阵A对称正定，则G-S迭代法收敛

  - 定理证明见第4节

- 逆定理(习题第7题)

对称正定：G-S迭代

- 正定理

定理

若系数矩阵A对称正定，则G-S迭代法收敛

  - 定理证明见第4节

- 逆定理(习题第7题)

定理

若系数矩阵A是具有正对角元的对称矩阵，G-S迭代法对任意初值收敛，则A必是正定的

对角占优矩阵

- 矩阵$A=(a_{ij})\in\mathbb{R}^{n\times n}$. 若对所有的$i=1,\ldots,n$,

$$|a_{ii}|\geqslant\sum_{j=1,j\neq i}^{n}|a_{ij}|$$

并且至少有一个对i有严格不等号成立，则称A是弱严格对角占优的。如果对所有i都有严格不等号成立，则称A是严格对角占优的

可约和不可约矩阵

- 矩阵$A=(a_{ij})\in\mathbb{R}^{n\times n}$. 若存在$n$阶排列方阵$P$使得

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

其中A_{11}是r阶方阵，A_{22}是$n-r$阶方阵，则称A是可约的或可分的；反之称为不可约的或不可分的

可约矩阵的意义

- 如果A可约，可以把$Ax=b$化为

$$PAP^TPx = Pb$$

可约矩阵的意义

- 如果A可约，可以把$Ax=b$化为

$$PAP^TPx=Pb$$

- 记$Px=y$, $Pb=f$, 则有

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

可约矩阵的意义

- 如果A可约，可以把$Ax=b$化为

$$PAP^TPx=Pb$$

- 记$Px=y$, $Pb=f$, 则有

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

- 如此我们可以先求解r阶方程组$A_{11}y_1=f_1$, 得到y_1, 代入$A_{12}y_1+A_{22}y_2=f_2$ 后就可以解出y_2

可约矩阵的等价定义

- 设A为n $(n \geqslant 2)$阶方阵，$\mathcal{W} = \{1, \dots, n\}$. 如果存在$\mathcal{W}$的两个非空的子集$\mathcal{S}$和$\mathcal{T}$满足

$$\mathcal{S} \cup \mathcal{T} = \mathcal{W}, \mathcal{S} \cap \mathcal{T} = \emptyset$$

使得$a_{ij} = 0$, $(i \in \mathcal{S}, j \in \mathcal{T})$, 则称$A$为可约的；否则称$A$是不可约的

不可约对角占优

- 如果一个矩阵不可约，并且是弱严格对角占优的，则称该矩阵为不可约对角占优的

不可约对角占优

- 如果一个矩阵不可约，并且是弱严格对角占优的，则称该矩阵为不可约对角占优的
- 下述三对角阵为不可约对角占优矩阵

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}$$

不可约对角占优

- 如果一个矩阵不可约，并且是弱严格对角占优的，则称该矩阵为不可约对角占优的
- 下述三对角阵为不可约对角占优矩阵

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}$$

- 弱严格对角占优矩阵，有可能一行全是零；而不可约对角占优，就排除了这种情形，从而任意对角元非零

定理

- 对角占优与非奇异的关系：

定理

- 对角占优与非奇异的关系：

定理

如果矩阵A是严格对角占优的，或者不可约对角占优的，则A非奇异

定理

- 对角占优与非奇异的关系：

定理

如果矩阵A是严格对角占优的，或者不可约对角占优的，则A非奇异

- 可约对角占优的，矩阵有可能奇异：

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 3 \end{pmatrix}$$

严格对角占优情形的证明

- 若A奇异，则$Ax = 0$有非零解x. 不妨设$|x_i| = \|x\|_\infty = 1$, 则有

$$|a_{ii}| = |a_{ii}x_i| = \left|\sum_{j=1,j\neq i}^{n} a_{ij}x_j\right| \leqslant \sum_{j=1,j\neq i}^{n} |a_{ij}|$$

严格对角占优情形的证明

- 若A奇异，则$Ax=0$有非零解x. 不妨设$|x_i|=\|x\|_\infty=1$, 则有

$$|a_{ii}|=|a_{ii}x_i|=\left|\sum_{j=1,j\neq i}^{n}a_{ij}x_j\right|\leqslant\sum_{j=1,j\neq i}^{n}|a_{ij}|$$

- 这与A严格对角占优矛盾

不可约对角占优情形的证明

- 同样，若A奇异，则存在x满足$\|x\|_\infty = 1$使得$Ax = 0$

不可约对角占优情形的证明

- 同样，若A奇异，则存在x满足$\|x\|_\infty = 1$使得$Ax = 0$
- 定义$\mathcal{S} = \{i : |x_i| = 1\}$, $\mathcal{T} = \{k : |x_k| < 1\}$

不可约对角占优情形的证明

- 同样，若A奇异，则存在x满足$\|x\|_\infty = 1$使得$Ax = 0$
- 定义$\mathcal{S} = \{i : |x_i| = 1\}$, $\mathcal{T} = \{k : |x_k| < 1\}$
- 显然$\mathcal{S} \cup \mathcal{T} = \mathcal{W}$, $\mathcal{S} \cap \mathcal{T} = \emptyset$, 而且$\mathcal{S}$非空

T为空集

- 假设T为空集

T为空集

- 假设T为空集
- 那么x的各分量的绝对值均为1，从而$\forall i \in S$,

$$|a_{ii}| \leqslant \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

T为空集

- 假设T为空集
- 那么x的各分量的绝对值均为1，从而$\forall i \in \mathcal{S}$,

$$|a_{ii}| \leqslant \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

- 这与A弱严格对角占优矛盾

T非空：利用不可约完成证明

- 因为A不可约，所以存在$i \in S$, $k \in T$使得$a_{ik} \neq 0$

$\mathcal{T}$非空：利用不可约完成证明

- 因为A不可约，所以存在$i \in \mathcal{S}$, $k \in \mathcal{T}$使得$a_{ik} \neq 0$
- 于是$|a_{ik}x_k| < |a_{ik}|$, 并且

$$
\begin{aligned}
|a_{ii}| = |a_{ii}x_i| &\leqslant \sum_{j\in\mathcal{S}, j\neq i} |a_{ij}||x_j| + \sum_{j\in\mathcal{T}} |a_{ij}||x_j| \\
&< \sum_{j\in\mathcal{S}, j\neq i} |a_{ij}| + \sum_{j\in\mathcal{T}} |a_{ij}| \\
&= \sum_{j\neq i} |a_{ij}|
\end{aligned}
$$

这也与A为弱严格对角占优矛盾

推论

推论

若A是严格对角占优的或者不可约对角占优的对称矩阵，且A的对角线元素均为正数，则A正定

证明

- A对称，所以特征值均为实数。为证A正定，只需要证明所有特征值均为正数

证明

- A对称，所以特征值均为实数。为证A正定，只需要证明所有特征值均为正数
- 若有一个特征值$\lambda \leqslant 0$, 考虑矩阵$A - \lambda I$

证明

- A对称，所以特征值均为实数。为证A正定，只需要证明所有特征值均为正数
- 若有一个特征值$\lambda \leqslant 0$, 考虑矩阵$A - \lambda I$
- 矩阵$A - \lambda I$只是在A的对角线元素上增加了一些，所以$A - \lambda I$和A一样是严格对角占优或者不可约对角占优的，从而是非奇异的，这与λ为A的特征值矛盾

定理

定理

若A是严格对角占优的，或者不可约对角占优的，则Jacobi迭代法和G-S迭代法都收敛

证明：Jacobi迭代

- $\forall i,\ a_{ii} \neq 0$

证明：Jacobi迭代

- $\forall i,\ a_{ii} \neq 0$
- 所以矩阵D可逆

证明：Jacobi迭代

- $\forall i,\ a_{ii} \neq 0$
- 所以矩阵D可逆
- 假设Jacobi迭代法的迭代矩阵B的某个特征值λ的模长不小于1

证明：Jacobi迭代

- $\forall i,\ a_{ii} \neq 0$
- 所以矩阵D可逆
- 假设Jacobi迭代法的迭代矩阵B的某个特征值λ的模长不小于1
- 考虑矩阵$\lambda D - L - U$, 这也是严格对角占优的或者不可约对角占优的，从而非奇异

证明：Jacobi迭代

- $\forall i,\ a_{ii} \neq 0$
- 所以矩阵D可逆
- 假设Jacobi迭代法的迭代矩阵B的某个特征值λ的模长不小于1
- 考虑矩阵$\lambda D - L - U$, 这也是严格对角占优的或者不可约对角占优的，从而非奇异
- 由于$\lambda I - B = D^{-1}(\lambda D - L - U)$, 所以$\det(\lambda I - B) \neq 0$, 矛盾

证明：G-S迭代

- G-S迭代法的迭代矩阵$L_1 = (D - L)^{-1}U$

证明：G-S迭代

- G-S迭代法的迭代矩阵$L_1 = (D-L)^{-1}U$
- $\lambda I - L_1 = (D-L)^{-1}(\lambda D - \lambda L - U)$

证明：G-S迭代

- G-S迭代法的迭代矩阵$L_1 = (D-L)^{-1}U$
- $\lambda I - L_1 = (D-L)^{-1}(\lambda D - \lambda L - U)$
- 若$|\lambda| \geqslant 1$, 则$D-L$和$\lambda D - \lambda L - U$也是严格对角占优的，或者不可约对角占优的，从而它们均可逆

证明：G-S迭代

- G-S迭代法的迭代矩阵$L_1 = (D-L)^{-1}U$
- $\lambda I - L_1 = (D-L)^{-1}(\lambda D - \lambda L - U)$
- 若$|\lambda| \geqslant 1$, 则$D-L$和$\lambda D - \lambda L - U$也是严格对角占优的，或者不可约对角占优的，从而它们均可逆
- 这就证明了G-S迭代法收敛

平均收敛速度

- 单步线性定常迭代法：$x_k = Mx_{k-1} + g$

平均收敛速度

- 单步线性定常迭代法：$x_k = Mx_{k-1} + g$
- 误差向量$y_k = x_k - x_*$满足$y_k = M^k y_0$, 从而有

$$\|y_k\| \leqslant \|M^k\| \|y_0\|$$

平均收敛速度

- 单步线性定常迭代法：$x_k = Mx_{k-1} + g$
- 误差向量$y_k = x_k - x_*$满足$y_k = M^k y_0$，从而有

$$\|y_k\| \leqslant \|M^k\| \|y_0\|$$

- 初始误差$\|y_0\|$一般不知道，通常用$\|M^k\|$的大小刻画迭代法收敛的速度

平均收敛速度

- 单步线性定常迭代法：$x_k = Mx_{k-1} + g$
- 误差向量$y_k = x_k - x_*$满足$y_k = M^k y_0$，从而有

$$\|y_k\| \leqslant \|M^k\| \|y_0\|$$

- 初始误差$\|y_0\|$一般不知道，通常用$\|M^k\|$的大小刻画迭代法收敛的速度
- 定义$R_k(M) = \dfrac{-\ln \|M^k\|}{k}$为$k$次迭代的平均收敛速度

误差缩减的比例因子

- 若迭代法收敛，则当$k \to \infty$时，$\|M^k\| \to 0$

误差缩减的比例因子

- 若迭代法收敛，则当$k \to \infty$时，$\|M^k\| \to 0$
- 所以当k充分大时，总有$R_k(M) > 0$.

误差缩减的比例因子

- 若迭代法收敛，则当$k \to \infty$时，$\|M^k\| \to 0$
- 所以当k充分大时，总有$R_k(M) > 0$.
- 设$R_k = R_k(M) > 0$, 数量

$$\sigma = \left(\frac{\|y_k\|}{\|y_0\|}\right)^{1/k} \approx \|M^k\|^{1/k} = e^{-R_k}$$

就表示误差范数在k次迭代中平均每次迭代所缩减的比例因子

对称矩阵情形

- 若M是对称矩阵，或者Hermite矩阵，或者正规矩阵，则显然有

$$\|M^k\|_2 = (\rho(M))^k$$

对称矩阵情形

- 若M是对称矩阵，或者Hermite矩阵，或者正规矩阵，则显然有

$$\|M^k\|_2 = (\rho(M))^k$$

- 所以$R_k(M) = -\ln \rho(M)$, 这是与k无关的量

对称矩阵情形

- 若M是对称矩阵，或者Hermite矩阵，或者正规矩阵，则显然有

$$\|M^k\|_2 = (\rho(M))^k$$

- 所以$R_k(M) = -\ln \rho(M)$, 这是与k无关的量
- 但在一般情况下，R_k是依赖于k的，这时计算R_k就非常复杂

收敛速度比较

- 如果两个迭代矩阵G和H满足$R_k(H) > R_k(M) > 0$，我们就说，对于k次迭代来讲，对应于H的迭代法比对应于G的迭代法的收敛速度快

收敛速度比较

- 如果两个迭代矩阵G和H满足$R_k(H) > R_k(M) > 0$，我们就说，对于k次迭代来讲，对应于H的迭代法比对应于G的迭代法的收敛速度快
- 有可能对某些k，$R_k(G) < R_k(H)$，而对另一些k，$R_k(G) > R_k(H)$

收敛速度比较

- 如果两个迭代矩阵G和H满足$R_k(H) > R_k(M) > 0$，我们就说，对于k次迭代来讲，对应于H的迭代法比对应于G的迭代法的收敛速度快
- 有可能对某些k，$R_k(G) < R_k(H)$，而对另一些k，$R_k(G) > R_k(H)$
- 所以我们考虑$k \to \infty$时$R_k(M)$的极限

渐近收敛速度

- 为了刻画整个迭代过程的收敛速度，我们定义

$$R_\infty(M) = \lim_{k\to\infty} R_k(M)$$

这称为渐近收敛速度

渐近收敛速度

- 为了刻画整个迭代过程的收敛速度，我们定义

$$R_\infty(M) = \lim_{k\to\infty} R_k(M)$$

  这称为渐近收敛速度

- 我们有

渐近收敛速度

- 为了刻画整个迭代过程的收敛速度，我们定义

$$R_\infty(M) = \lim_{k\to\infty} R_k(M)$$

  这称为渐近收敛速度

- 我们有

定理

$$R_\infty(M) = -\ln \rho(M)$$

渐近收敛速度

- 为了刻画整个迭代过程的收敛速度，我们定义

$$R_\infty(M) = \lim_{k\to\infty} R_k(M)$$

这称为渐近收敛速度

- 我们有

定理

$$R_\infty(M) = -\ln \rho(M)$$

- 定理的成立不依赖于范数的选取

证明

- 只需证明 $\lim_{k\to\infty} \|M^k\|^{1/k} = \rho(M)$ 即可

证明

- 只需证明 $\lim_{k\to\infty}\|M^k\|^{1/k}=\rho(M)$ 即可
- 因为$(\rho(M))^k=\rho(M^k)\leqslant\|M^k\|$, 所以$\rho(M)\leqslant\|M^k\|^{1/k}$

证明

- 只需证明 $\lim_{k\to\infty} \|M^k\|^{1/k} = \rho(M)$ 即可
- 因为$(\rho(M))^k = \rho(M^k) \leqslant \|M^k\|$, 所以$\rho(M) \leqslant \|M^k\|^{1/k}$
- 另一方面，对$\forall \varepsilon > 0$, 考虑矩阵

$$B_\varepsilon = \frac{1}{\rho(M) + \varepsilon} M$$

证明

- 只需证明 $\lim_{k\to\infty}\|M^k\|^{1/k}=\rho(M)$ 即可
- 因为$(\rho(M))^k=\rho(M^k)\leqslant\|M^k\|$, 所以$\rho(M)\leqslant\|M^k\|^{1/k}$
- 另一方面，对$\forall\varepsilon>0$, 考虑矩阵

$$B_\varepsilon=\frac{1}{\rho(M)+\varepsilon}M$$

- 显然$\rho(B_\varepsilon)<1$

- 于是 $\lim_{k\to\infty} B_\varepsilon^k = 0$

- 于是 $\lim_{k\to\infty} B_\varepsilon^k = 0$
- 所以存在自然数K，当$k \geqslant K$时有$\|B_\varepsilon^k\| \leqslant 1$

- 于是 $\lim\limits_{k\to\infty} B_\varepsilon^k = 0$
- 所以存在自然数K，当$k \geqslant K$时有$\|B_\varepsilon^k\| \leqslant 1$
- 这就是$\|M^k\| \leqslant (\rho(M) + \varepsilon)^k$

- 于是$\lim\limits_{k\to\infty} B_\varepsilon^k = 0$
- 所以存在自然数K，当$k \geqslant K$时有$\|B_\varepsilon^k\| \leqslant 1$
- 这就是$\|M^k\| \leqslant (\rho(M)+\varepsilon)^k$
- 至此我们证明了：对任意的$\varepsilon > 0$，存在自然数K，当$k \geqslant K$时，

$$\rho(M) \leqslant \|M^k\|^{1/k} \leqslant \rho(M)+\varepsilon$$

这就完成了证明

模型问题

- 为了具体说明迭代法的收敛速度，我们考虑用五点差分格式求解$[0,1]^2$上的Poisson方程第一边值问题

$$\begin{cases} \triangle u = \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = -f(x,y), (x,y) \in (0,1)^2 \\ u|_{\Gamma} = 0 \end{cases}$$

这里Γ表示正方形定义域的边界

离散化

- 把正方形的每边n等分，令$h = 1/n$, 用等分线把正方形$[0, 1]^2$分割成n^2个小正方形

离散化

- 把正方形的每边n等分，令$h = 1/n$, 用等分线把正方形$[0, 1]^2$分割成n^2个小正方形
- 记小正方形的顶点为(x_i, y_j)

离散化

- 把正方形的每边n等分，令$h = 1/n$, 用等分线把正方形$[0,1]^2$分割成n^2个小正方形
- 记小正方形的顶点为(x_i, y_j)
- 用二阶差商

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{(x_i,y_j)} = \frac{1}{h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) + O(h^2)$$

$$\left.\frac{\partial^2 u}{\partial y^2}\right|_{(x_i,y_j)} = \frac{1}{h^2}(u_{i,j-1} - 2u_{i,j} + u_{i,j+1}) + O(h^2)$$

代替二阶偏微分，这里$u_{i,j} = u(x_i, y_j)$

矩阵形式

- 如此得到方程组

$$
\begin{cases}
4u_{i,j} - (u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}) = h^2 f_{ij} \\
\qquad i, j = 1, \ldots, n-1 \\
u_{i,0} = u_{i,n} = u_{0,j} = u_{n,j} = 0, \quad i, j = 0, \ldots, n
\end{cases}
$$

矩阵形式

- 如此得到方程组

$$\begin{cases} 4u_{i,j}-(u_{i+1,j}+u_{i-1,j}+u_{i,j+1}+u_{i,j-1})=h^2f_{ij} \\ \qquad i,j=1,\ldots,n-1 \\ u_{i,0}=u_{i,n}=u_{0,j}=u_{n,j}=0, \quad i,j=0,\ldots,n \end{cases}$$

- 写成矩阵形式：

$$T_{n-1}U+UT_{n-1}=h^2F,$$

其中$U=(u_{i,j})$, $F=(f_{ij})$, T_{n-1}为$n-1$阶三对角矩阵矩阵，主对角元素为2, 上下次对角元素为-1

矩阵拉直：自然顺序排列

- 为了得到通常的$Ax = b$形式的线性方程组，我们对U和F中的元素进行“拉直”：先按j由小到大排列，j相同的按i由小到大排列

矩阵拉直：自然顺序排列

- 为了得到通常的$Ax=b$形式的线性方程组，我们对U和F中的元素进行“拉直”：先按j由小到大排列，j相同的按i由小到大排列
- 如此我们得到方程组

$$Au=h^2f$$

其中向量u和f是矩阵U和F拉直的结果

系数矩阵A

$$A = \begin{pmatrix} T_{n-1}+2I_{n-1} & -I_{n-1} & & \\ -I_{n-1} & \ddots & \ddots & \\ & \ddots & \ddots & -I_{n-1} \\ & & -I_{n-1} & T_{n-1}+2I_{n-1} \end{pmatrix}$$

- A是$(n-1)^2$阶块三对角阵，五条对角线上有非零元；它也是不可约对角占优的对称正定阵

系数矩阵A

$$A = \begin{pmatrix} T_{n-1} + 2I_{n-1} & -I_{n-1} & & \\ -I_{n-1} & \ddots & \ddots & \\ & \ddots & \ddots & -I_{n-1} \\ & & -I_{n-1} & T_{n-1} + 2I_{n-1} \end{pmatrix}$$

- A是$(n-1)^2$阶块三对角阵，五条对角线上有非零元；它也是不可约对角占优的对称正定阵
- 每个对角元的左、右各有两个非零元素，据离对角元远近，分别对应对角元上下和左右邻居

A的特征值和特征向量

- T_{n-1}的特征值为$\lambda_j = 2 - 2\cos\frac{j\pi}{n}$, 对应的单位特征向量为

$$z_j = \left(\sqrt{\frac{2}{n}}\sin\frac{j\pi}{n}, \sqrt{\frac{2}{n}}\sin\frac{2j\pi}{n}, \dots, \sqrt{\frac{2}{n}}\sin\frac{(n-1)j\pi}{n}\right)^T$$

A的特征值和特征向量

- T_{n-1}的特征值为$\lambda_j = 2 - 2\cos\dfrac{j\pi}{n}$, 对应的单位特征向量为

$$z_j = \left(\sqrt{\frac{2}{n}}\sin\frac{j\pi}{n}, \sqrt{\frac{2}{n}}\sin\frac{2j\pi}{n}, \dots, \sqrt{\frac{2}{n}}\sin\frac{(n-1)j\pi}{n}\right)^T$$

- 利用"拉直"操作，可以证明A的特征值为$\lambda_{pq} = \lambda_p + \lambda_q$, 对应特征向量为$z_p z_q^T$ "拉直"的结果

Jacobi迭代法

- 针对上述模型问题，Jacobi迭代法的迭代矩阵为$B = D^{-1}(L+U) = I - \frac{1}{4}A$

Jacobi迭代法

- 针对上述模型问题，Jacobi迭代法的迭代矩阵为$B = D^{-1}(L+U) = I - \dfrac{1}{4}A$
- B的对角元为0，其左右分别有两个非零元素（值为1/4），对应于对角元的左右和上下邻居

Jacobi迭代法

- 针对上述模型问题，Jacobi迭代法的迭代矩阵为$B=D^{-1}(L+U)=I-\frac{1}{4}A$
- B的对角元为0，其左右分别有两个非零元素（值为1/4），对应于对角元的左右和上下邻居
- 所以迭代格式为

$$u_{ij}^{(k)}=\frac{1}{4}\left(u_{i+1,j}^{(k-1)}+u_{i-1,j}^{(k-1)}+u_{i,j+1}^{(k-1)}+u_{i,j-1}^{(k-1)}\right)+\frac{h^2}{4}f_{ij},$$

$$u_{i0}=u_{in}=u_{0j}=u_{nj}=0,\ i,j=1,\dots,n-1$$

Jacobi迭代法的渐近收敛速度

- B的特征值为$\mu_{pq} = 1 - \frac{1}{4}\lambda_{pq}$

Jacobi迭代法的渐近收敛速度

- B的特征值为$\mu_{pq} = 1 - \dfrac{1}{4}\lambda_{pq}$
  - 从而若μ为B的特征值，则$-\mu$也是B的特征值

Jacobi迭代法的渐近收敛速度

- B的特征值为$\mu_{pq} = 1 - \dfrac{1}{4}\lambda_{pq}$
  - 从而若μ为B的特征值，则$-\mu$也是B的特征值
- 所以$\rho(B) = \cos\dfrac{\pi}{n} = \cos h\pi$, 从而可知渐近收敛速度为

$$R_\infty(B) = -\ln\rho(B) = -\ln\cos h\pi$$

Jacobi迭代法的渐近收敛速度

- B的特征值为$\mu_{pq} = 1 - \frac{1}{4}\lambda_{pq}$
  - 从而若μ为B的特征值，则$-\mu$也是B的特征值
- 所以$\rho(B) = \cos\frac{\pi}{n} = \cos h\pi$，从而可知渐近收敛速度为

$$R_\infty(B) = -\ln\rho(B) = -\ln\cos h\pi$$

- 从而当$h \to 0$时，我们有

$$R_\infty(B) \sim \frac{1}{2}\pi^2 h^2$$

B对应的特征值问题

- 我们这里要基于B的特征值问题给出L_1的特征值

B对应的特征值问题

- 我们这里要基于B的特征值问题给出L_1的特征值
- B的特征值问题为$B\eta = \mu\eta$

B对应的特征值问题

- 我们这里要基于B的特征值问题给出L_1的特征值
- B的特征值问题为$B\eta = \mu\eta$
- 回忆：B的对角元全是零，每个对角元的左、右各有两个非零元素（值为1/4），所以采用二重指标表示B对应的特值问题为

$$\begin{cases} \mu\eta_{ij} = \dfrac{1}{4}(\eta_{i+1,j} + \eta_{i-1,j} + \eta_{i,j+1} + \eta_{i,j-1}) \\ \eta_{i0} = \eta_{in} = \eta_{0j} = \eta_{nj} = 0 \end{cases}$$

L_1对应的特征值问题

- L_1对应的特征值问题为

$$L_1\xi = \lambda\xi \to \lambda D\xi = \lambda L\xi + U\xi$$

L_1对应的特征值问题

- L_1对应的特征值问题为

$$L_1\xi = \lambda\xi \to \lambda D\xi = \lambda L\xi + U\xi$$

- 所以采用二重指标表示L_1的特征值问题为

$$\begin{cases} \lambda\xi_{ij} = \dfrac{1}{4}(\lambda\xi_{i-1,j} + \lambda\xi_{i,j-1} + \xi_{i,j+1} + \xi_{i+1,j}) \\ \xi_{i0} = \xi_{in} = \xi_{0j} = \xi_{nj} = 0 \end{cases}$$

- 设$\lambda \neq 0$, 作变换$\xi_{ij} = \lambda^{(i+j)/2}\eta_{ij}$ 得

$$
\begin{aligned}
\lambda^{1+\frac{i+j}{2}}\eta_{ij} =& \frac{1}{4}\Big(\lambda^{1+(i+j-1)/2}\eta_{i-1,j} + \lambda^{1+(i+j-1)/2}\eta_{i,j-1} \\
& + \lambda^{(i+j+1)/2}\eta_{i+1,j} + \lambda^{(i+j+1)/2}\eta_{i,j+1}\Big) \\
=& \frac{1}{4}\lambda^{\frac{i+j+1}{2}}\left(\eta_{i-1,j} + \eta_{i,j-1} + \eta_{i+1,j} + \eta_{i,j+1}\right)
\end{aligned}
$$

- 设$\lambda \neq 0$, 作变换$\xi_{ij} = \lambda^{(i+j)/2}\eta_{ij}$ 得

$$\begin{aligned}\lambda^{1+\frac{i+j}{2}}\eta_{ij} =& \frac{1}{4}\Big(\lambda^{1+(i+j-1)/2}\eta_{i-1,j} + \lambda^{1+(i+j-1)/2}\eta_{i,j-1} \\ & +\lambda^{(i+j+1)/2}\eta_{i+1,j} + \lambda^{(i+j+1)/2}\eta_{i,j+1}\Big) \\ =& \frac{1}{4}\lambda^{\frac{i+j+1}{2}}\left(\eta_{i-1,j} + \eta_{i,j-1} + \eta_{i+1,j} + \eta_{i,j+1}\right)\end{aligned}$$

- 由边值为零，可知若λ是L_1的非零特征值当且仅当$\lambda^{1/2}$是B的非零特征值

- 设$\lambda \neq 0$, 作变换$\xi_{ij} = \lambda^{(i+j)/2}\eta_{ij}$ 得

$$\begin{aligned}
\lambda^{1+\frac{i+j}{2}}\eta_{ij} =& \frac{1}{4}\Big(\lambda^{1+(i+j-1)/2}\eta_{i-1,j} + \lambda^{1+(i+j-1)/2}\eta_{i,j-1} \\
&+ \lambda^{(i+j+1)/2}\eta_{i+1,j} + \lambda^{(i+j+1)/2}\eta_{i,j+1}\Big) \\
=& \frac{1}{4}\lambda^{\frac{i+j+1}{2}}\left(\eta_{i-1,j} + \eta_{i,j-1} + \eta_{i+1,j} + \eta_{i,j+1}\right)
\end{aligned}$$

- 由边值为零，可知若λ是L_1的非零特征值当且仅当$\lambda^{1/2}$是B的非零特征值
- 因此L_1的特征值非负

G-S迭代法的渐近收敛速度

- 从而我们可知

$$R_\infty(L_1) = -\ln\rho(L_1) = -2\ln\rho(B) \sim \pi^2 h^2, \quad h \to 0$$

G-S迭代法的渐近收敛速度

- 从而我们可知

$$R_\infty(L_1) = -\ln\rho(L_1) = -2\ln\rho(B) \sim \pi^2 h^2, \quad h \to 0$$

- 这说明G-S迭代法的渐近收敛速度是Jacobi迭代法的渐近收敛速度的两倍

超松弛迭代法

- 这是一种新的迭代法

超松弛迭代法

- 这是一种新的迭代法
- 它是G-S迭代法的引申和推广

超松弛迭代法

- 这是一种新的迭代法
- 它是G-S迭代法的引申和推广
- 也可以看作是G-S迭代法的加速

迭代格式的重新认识

- 在迭代格式中令$x_{k+1} - x_k = \Delta x$，即$x_{k+1} = x_k + \Delta x$

迭代格式的重新认识

- 在迭代格式中令$x_{k+1} - x_k = \Delta x$，即$x_{k+1} = x_k + \Delta x$
- 这可以看作是在向量x_k上加上修正项Δx得到x_{k+1}

迭代格式的重新认识

- 在迭代格式中令$x_{k+1} - x_k = \Delta x$，即$x_{k+1} = x_k + \Delta x$
- 这可以看作是在向量x_k上加上修正项Δx得到x_{k+1}
- 对于已有的迭代格式，Δx是由格式完全确定的

迭代格式的重新认识

- 在迭代格式中令$x_{k+1}-x_k=\Delta x$，即$x_{k+1}=x_k+\Delta x$
- 这可以看作是在向量x_k上加上修正项Δx得到x_{k+1}
- 对于已有的迭代格式，Δx是由格式完全确定的
- 我们可以在修正项的前面加上一个参数ω以得到松弛迭代格式

G-S迭代法对应的松弛格式

- G-S迭代格式：

  $$x_{k+1} = D^{-1}Lx_{k+1} + D^{-1}Ux_k + D^{-1}b$$

G-S迭代法对应的松弛格式

- G-S迭代格式:
  $$x_{k+1} = D^{-1}Lx_{k+1} + D^{-1}Ux_k + D^{-1}b$$
- $\Delta x = D^{-1}Lx_{k+1} + (D^{-1}U - I)x_k + D^{-1}b$

G-S迭代法对应的松弛格式

- G-S迭代格式:

  $$x_{k+1} = D^{-1}Lx_{k+1} + D^{-1}Ux_k + D^{-1}b$$

- $\Delta x = D^{-1}Lx_{k+1} + (D^{-1}U - I)x_k + D^{-1}b$
- 松弛迭代格式

  $$\begin{aligned} x_{k+1} &= x_k + \omega\Delta x \\ &= (1-\omega)x_k + \omega(D^{-1}Lx_{k+1} + D^{-1}Ux_k + D^{-1}b) \end{aligned}$$

术语

- ω叫做松弛因子

术语

- ω叫做松弛因子
- 当$\omega > 1$时，对应的格式叫做超松弛迭代法，简记为SOR迭代法

术语

- ω叫做松弛因子
- 当$\omega > 1$时，对应的格式叫做超松弛迭代法，简记为SOR迭代法
  - SOR — Successive Over–Relaxation

术语

- ω叫做松弛因子
- 当$\omega > 1$时，对应的格式叫做超松弛迭代法，简记为SOR迭代法
  - SOR — Successive Over–Relaxation
- 当$\omega < 1$时，对应的格式叫做低松弛迭代法

术语

- ω叫做松弛因子
- 当$\omega > 1$时，对应的格式叫做超松弛迭代法，简记为SOR迭代法
  - SOR — Successive Over–Relaxation
- 当$\omega < 1$时，对应的格式叫做低松弛迭代法
- 当$\omega = 1$时，对应的格式就是G-S迭代法

迭代矩阵

- 如果$(I-\omega D^{-1}L)^{-1}$存在，松弛迭代格式可以写为
$$x_{k+1} = L_\omega x_k + \omega(D-\omega L)^{-1}b$$
其中
$$L_\omega = (D-\omega L)^{-1}((1-\omega)D+\omega U)$$
称为松弛迭代格式的迭代矩阵

迭代矩阵

- 如果$(I-\omega D^{-1}L)^{-1}$存在，松弛迭代格式可以写为

$$x_{k+1}=L_\omega x_k+\omega(D-\omega L)^{-1}b$$

其中

$$L_\omega=(D-\omega L)^{-1}((1-\omega)D+\omega U)$$

称为松弛迭代格式的**迭代矩阵**

- 注意：$(D-\omega L)^{-1}$为下三角矩阵，$((1-\omega)D+\omega U$为上三角矩阵

收敛定理

定理

*SOR*迭代法收敛的充分必要条件是

$$\rho(L_\omega) < 1$$

定理

*SOR*迭代法收敛的必要条件是$0 < \omega < 2$

第一个定理显然。下面证明第二个

$0 < \omega < 2$的证明

- 迭代法收敛，所以$\rho(L_\omega) < 1$

$0 < \omega < 2$的证明

- 迭代法收敛，所以$\rho(L_\omega) < 1$
- 若λ_i为L_ω的n个特征值，则

$$|\det L_\omega| = |\lambda_1 \cdots \lambda_n| < 1$$

$0 < \omega < 2$的证明

- 迭代法收敛，所以$\rho(L_\omega) < 1$
- 若λ_i为L_ω的n个特征值，则

$$|\det L_\omega| = |\lambda_1 \cdots \lambda_n| < 1$$

- 注意到

$$L_\omega = (I - \omega D^{-1}L)^{-1}((1-\omega)I + \omega D^{-1}U),$$
$$\det((1-\omega)I + \omega D^{-1}U) = (1-\omega)^n,$$
$$\det(I - \omega D^{-1}L)^{-1} = 1$$

- 从而有

$$|\det L_\omega| = |(1-\omega)^n| < 1$$

- 从而有
$$|\det L_\omega| = |(1-\omega)^n| < 1$$
- 即$|1-\omega| < 1$, 也就是$0 < \omega < 2$

- 从而有
$$|\det L_\omega| = |(1-\omega)^n| < 1$$
- 即$|1-\omega|<1$，也就是$0<\omega<2$
- 这个结果说明，要使SOR迭代法收敛，必须选取收敛因子$\omega\in(0,2)$

充分条件：对角占优

定理

若系数矩阵A是严格对角占优的，或者不可约对角占优的，且松弛因子$\omega \in (0, 1)$，则SOR迭代法收敛

证明关键

- 取λ, $|\lambda| \geqslant 1$,

$$\lambda I - L_\omega = (D - \omega L)^{-1}((\lambda + \omega - 1)D - \lambda\omega L - \omega U)$$

证明关键

- 取λ, $|\lambda| \geqslant 1$,

$$\lambda I - L_\omega = (D - \omega L)^{-1}((\lambda + \omega - 1)D - \lambda\omega L - \omega U)$$

- 根据三角不等式:

$$\begin{aligned}|\lambda + \omega - 1| &\geqslant |\lambda| - (1 - \omega)\\ &= |\lambda|\omega + (|\lambda| - 1)(1 - \omega) \geqslant |\lambda|\omega\end{aligned}$$

证明关键

- 取λ, $|\lambda| \geqslant 1$,

$$\lambda I - L_\omega = (D - \omega L)^{-1}((\lambda + \omega - 1)D - \lambda\omega L - \omega U)$$

- 根据三角不等式:

$$\begin{aligned} |\lambda + \omega - 1| &\geqslant |\lambda| - (1 - \omega) \\ &= |\lambda|\omega + (|\lambda| - 1)(1 - \omega) \geqslant |\lambda|\omega \end{aligned}$$

- 所以$D - \omega L$, $(\lambda + \omega - 1)D - \lambda\omega L - \omega U$ 都是严格对角占优的，或者不可约对角占优的，从而是非奇异的

充分条件：对称正定

定理

若系数矩阵A是实对称的正定矩阵，则当$0 < \omega < 2$时，SOR方法收敛

根据该定理可知，当A对称正定时，G-S迭代法收敛

证明

- 设λ是L_ω的任一特征值（可以是复数），x为对应的特征向量，则(注意$L^T = U$)

$$((1-\omega)D + \omega L^T)x = \lambda(D - \omega L)x$$

证明

- 设λ是L_ω的任一特征值（可以是复数），x为对应的特征向量，则(注意$L^T = U$)

$$((1-\omega)D + \omega L^T)x = \lambda(D - \omega L)x$$

- 左乘x的共轭转置向量x^*:

$$x^*((1-\omega)D + \omega L^T)x = \lambda x^*(D - \omega L)x$$

证明

- 设λ是L_ω的任一特征值（可以是复数），x为对应的特征向量，则(注意$L^T = U$)

$$((1-\omega)D + \omega L^T)x = \lambda(D - \omega L)x$$

- 左乘x的共轭转置向量x^*:

$$x^*((1-\omega)D + \omega L^T)x = \lambda x^*(D - \omega L)x$$

- 令$x^*Dx = \delta$, $x^*Lx = \alpha + \mathrm{i}\beta$, 则有$x^*L^Tx = \alpha - \mathrm{i}\beta$

- 我们有

$$(1-\omega)\delta+\omega(\alpha-\mathrm{i}\beta)=\lambda(\delta-\omega(\alpha+\mathrm{i}\beta))$$

- 我们有

$$(1-\omega)\delta+\omega(\alpha-\mathrm{i}\beta)=\lambda(\delta-\omega(\alpha+\mathrm{i}\beta))$$

- 两边取模得到

$$|\lambda|^2=\frac{((1-\omega)\delta+\omega\alpha)^2+\omega^2\beta^2}{(\delta-\omega\alpha)^2+\omega^2\beta^2}$$

- 我们有
$$(1-\omega)\delta+\omega(\alpha-\mathrm{i}\beta)=\lambda(\delta-\omega(\alpha+\mathrm{i}\beta))$$
- 两边取模得到
$$|\lambda|^2=\frac{((1-\omega)\delta+\omega\alpha)^2+\omega^2\beta^2}{(\delta-\omega\alpha)^2+\omega^2\beta^2}$$
- 右侧的分子和分母相减，得到$\omega\delta(\delta-2\alpha)(\omega-2)$

- 我们有
$$(1-\omega)\delta+\omega(\alpha-\mathrm{i}\beta)=\lambda(\delta-\omega(\alpha+\mathrm{i}\beta))$$
- 两边取模得到
$$|\lambda|^2=\frac{((1-\omega)\delta+\omega\alpha)^2+\omega^2\beta^2}{(\delta-\omega\alpha)^2+\omega^2\beta^2}$$
- 右侧的分子和分母相减，得到$\omega\delta(\delta-2\alpha)(\omega-2)$
- 当A正定时$\delta>0$, $\delta-2\alpha=x^*Ax>0$, 所以当$0<\omega<2$时$|\lambda|<1$, 即SOR迭代法收敛

最佳收敛因子

- SOR迭代法的谱半径依赖于ω

最佳收敛因子

- SOR迭代法的谱半径依赖于ω
- 如何选取恰当的ω，从而使得收敛速度最快？

最佳收敛因子

- SOR迭代法的谱半径依赖于ω
- 如何选取恰当的ω，从而使得收敛速度最快？
- 为此我们考虑特征值问题$L_\omega x = \lambda x$，即

$$((\lambda - 1 + \omega)D - \lambda\omega L - \omega U)x = 0$$

最佳收敛因子

- SOR迭代法的谱半径依赖于ω
- 如何选取恰当的ω，从而使得收敛速度最快？
- 为此我们考虑特征值问题$L_\omega x = \lambda x$，即

$$((\lambda - 1 + \omega)D - \lambda\omega L - \omega U)x = 0$$

- 回忆：L_ω的所有特征值相乘等于$(1-\omega)^n$，所以当$\omega \neq 1$时无零特征值

仍是那个模型问题

- 对于模型问题，上述问题为

$$(\lambda+\omega-1)u_{ij}=\frac{\omega}{4}(\lambda u_{i-1,j}+\lambda u_{i,j-1}+u_{i+1,j}+u_{i,j+1}$$

$$u_{i0}=u_{in}=u_{0j}=u_{nj}=0,$$

仍是那个模型问题

- 对于模型问题，上述问题为

$$(\lambda+\omega-1)u_{ij}=\frac{\omega}{4}(\lambda u_{i-1,j}+\lambda u_{i,j-1}+u_{i+1,j}+u_{i,j+1}$$

$$u_{i0}=u_{in}=u_{0j}=u_{nj}=0,$$

- 下面分两步讨论问题：

仍是那个模型问题

- 对于模型问题，上述问题为

$$(\lambda + \omega - 1)u_{ij} = \frac{\omega}{4}(\lambda u_{i-1,j} + \lambda u_{i,j-1} + u_{i+1,j} + u_{i,j+1}$$

$$u_{i0} = u_{in} = u_{0j} = u_{nj} = 0,$$

- 下面分两步讨论问题：
  - B与L_ω的特征值之间的关系

仍是那个模型问题

- 对于模型问题，上述问题为

$$(\lambda+\omega-1)u_{ij}=\frac{\omega}{4}(\lambda u_{i-1,j}+\lambda u_{i,j-1}+u_{i+1,j}+u_{i,j+1}$$

$$u_{i0}=u_{in}=u_{0j}=u_{nj}=0,$$

- 下面分两步讨论问题：
  - B与L_ω的特征值之间的关系
  - $\rho(L_\omega)$随ω变化的情况

B与L_ω的特征值之间的关系

- 当$\lambda \neq 0$时，作变换$u_{ij} = (\pm\lambda^{1/2})^{i+j} v_{ij}$, 则有

$$\mu v_{ij} - \frac{1}{2}(v_{i-1,j} + v_{i,j-1} + v_{i+1,j} + v_{i,j+1}) = 0,$$

其中$\mu = \pm\dfrac{\lambda + \omega - 1}{\omega\lambda^{1/2}}$

B与L_ω的特征值之间的关系

- 当$\lambda \neq 0$时，作变换$u_{ij} = (\pm\lambda^{1/2})^{i+j} v_{ij}$, 则有

$$\mu v_{ij} - \frac{1}{2}(v_{i-1,j} + v_{i,j-1} + v_{i+1,j} + v_{i,j+1}) = 0,$$

其中$\mu = \pm\dfrac{\lambda + \omega - 1}{\omega\lambda^{1/2}}$

- 当$\omega \neq 1$时，若λ是L_ω的特征值，则由

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

确定的两个μ都是B的特征值；反之亦然

- 当$\omega = 1$时，前式简化为$\lambda^2 = \mu^2\lambda$

- 当$\omega = 1$时，前式简化为$\lambda^2 = \mu^2\lambda$
- 当B的特征值为$\pm\mu_i$时，对应于L_1的特征值为0和μ_i^2

- 当$\omega = 1$时，前式简化为$\lambda^2 = \mu^2\lambda$
- 当B的特征值为$\pm\mu_i$时，对应于L_1的特征值为0和μ_i^2
- 回忆：若μ是B的特征值，则$-\mu$也是B的特征值

$\rho(L_\omega)$随ω变化的情况

- 设$0 \leqslant \mu < 1$是B的一个特征值，$0 < \omega < 2$, 则由方程
$$\lambda + \omega - 1 = \pm \mu\omega\lambda^{1/2}$$
可得L_ω的两个特征值分别是
$$\lambda_+(\omega,\mu) = \left(\frac{\mu\omega}{2} + \sqrt{\left(\frac{\mu\omega}{2}\right)^2 - (\omega - 1)}\right)^2,$$
$$\lambda_-(\omega,\mu) = \left(\frac{\mu\omega}{2} - \sqrt{\left(\frac{\mu\omega}{2}\right)^2 - (\omega - 1)}\right)^2,$$

- 为了研究$\rho(L_\omega)$随ω的变化，我们考虑

$$M(\omega,\mu)=\max(|\lambda_+(\omega,\mu)|,|\lambda_-(\omega,\mu)|)$$

- 为了研究$\rho(L_\omega)$随ω的变化，我们考虑

$$M(\omega,\mu)=\max(|\lambda_+(\omega,\mu)|,|\lambda_-(\omega,\mu)|)$$

- 分析细节忽略

- 为了研究$\rho(L_\omega)$随ω的变化，我们考虑

$$M(\omega,\mu)=\max(|\lambda_+(\omega,\mu)|,|\lambda_-(\omega,\mu)|)$$

- 分析细节忽略
- 结论：随着ω从0开始增加，$\rho(L_\omega)$减少，直至

$$\omega=\frac{2}{1+\sqrt{1-\rho(B)^2}}$$

时$\rho(L_\omega)$达到极小，然后开始变大。这个ω称为**最佳松弛因子**

- 当ω取上述最佳松弛因子时，我们有

$$\begin{aligned}
R_\infty(L_\omega) &= -\ln\rho(L_\omega) \\
&= -\ln\frac{1-\sqrt{1-\rho(B)^2}}{1+\sqrt{1-\rho(B)^2}} \\
&= -\ln\frac{1-\sin h\pi}{1+\sin h\pi} \sim 2h\pi,\ h\to 0
\end{aligned}$$

- 当ω取上述最佳松弛因子时，我们有

$$
\begin{aligned}
R_\infty(L_\omega) &= -\ln \rho(L_\omega) \\
&= -\ln \frac{1-\sqrt{1-\rho(B)^2}}{1+\sqrt{1-\rho(B)^2}} \\
&= -\ln \frac{1-\sin h\pi}{1+\sin h\pi} \sim 2h\pi,\ h \to 0
\end{aligned}
$$

- SOR迭代法要比Jacobi迭代法和G-S迭代法快得多！

共轭梯度法

邓建松

2018 年 11 月 9 日

松弛迭代法的局限性

- SOR迭代法中如能取得最佳松弛因子，算法的效率会得到数量级上的提高

松弛迭代法的局限性

- SOR迭代法中如能取得最佳松弛因子，算法的效率会得到数量级上的提高
- 而最佳松弛因子只在系数矩阵具有较好性质时才有可能找到

松弛迭代法的局限性

- SOR迭代法中如能取得最佳松弛因子，算法的效率会得到数量级上的提高
- 而最佳松弛因子只在系数矩阵具有较好性质时才有可能找到
- 而且上节在计算最佳松弛因子时，还用到了对应的Jacobi迭代矩阵的谱半径

共轭梯度法

- 这是一种不需要确定任何参数的求解对称正定线性方程组的方法

共轭梯度法

- 这是一种不需要确定任何参数的求解对称正定线性方程组的方法
- 它是上世纪50年代初期由M.R. Hestenes和E. Stiefel首先提出的

共轭梯度法

- 这是一种不需要确定任何参数的求解对称正定线性方程组的方法
- 它是上世纪50年代初期由M.R. Hestenes和E. Stiefel首先提出的
- 自后得到了长足的发展，成为求解大型稀疏线性方程组最受欢迎的一类方法

共轭梯度法

- 这是一种不需要确定任何参数的求解对称正定线性方程组的方法
- 它是上世纪50年代初期由M.R. Hestenes和E. Stiefel首先提出的
- 自后得到了长足的发展，成为求解大型稀疏线性方程组最受欢迎的一类方法
- 它也是求解大型非线性优化问题的主要方法之一

线性方程组与对应的二次泛函

- 设A为对称正定矩阵

线性方程组与对应的二次泛函

- 设A为对称正定矩阵
- 考虑线性方程组$Ax = b$的求解

线性方程组与对应的二次泛函

- 设A为对称正定矩阵
- 考虑线性方程组$Ax = b$的求解
- 为此我们定义二次函数

$$\varphi(x) = x^T Ax - 2b^T x$$

定理

定理

设A对称正定，求方程组$Ax = b$的解等价于求二次函数$\varphi(x)$的极小值点

定理证明

- 直接计算可得

$$\frac{\partial \varphi}{\partial x_i} = 2(a_{i1}x_1 + \cdots + a_{in}x_n) - 2b_i$$

定理证明

- 直接计算可得

$$\frac{\partial \varphi}{\partial x_i} = 2(a_{i1}x_1 + \cdots + a_{in}x_n) - 2b_i$$

- 所以

$$\nabla \varphi(x) = 2(Ax - b)$$

定理证明

- 直接计算可得

$$\frac{\partial \varphi}{\partial x_i} = 2(a_{i1}x_1 + \cdots + a_{in}x_n) - 2b_i$$

- 所以

$$\nabla \varphi(x) = 2(Ax - b)$$

- 若$\varphi(x)$在某点x_*达到极小，则必有$\nabla \varphi(x_*) = 0$，从而$Ax_* = b$

定理证明

- 直接计算可得
$$\frac{\partial \varphi}{\partial x_i} = 2(a_{i1}x_1 + \cdots + a_{in}x_n) - 2b_i$$
- 所以
$$\nabla\varphi(x) = 2(Ax - b)$$
- 若$\varphi(x)$在某点x_*达到极小，则必有$\nabla\varphi(x_*) = 0$，从而$Ax_* = b$
- 若x_*是$Ax = b$的解，而$\varphi(x)$的Hessian阵是A，对称正定，从而x_*是$\varphi(x)$ 的极小值点

求解方法：盲人下山

- 为了求解线性方程组，我们可以计算二次函数$\varphi(x)$的极小值

求解方法：盲人下山

- 为了求解线性方程组，我们可以计算二次函数$\varphi(x)$的极小值
- 为了求二次函数的极小值，我们可以模拟盲人下山：先任意给定一个初始点x_0，确定一个下山的方向p_0，沿着经过点x_0而方向为p_0的直线$x = x_0 + \alpha p_0$上找一点x_1使$\varphi(x)$达到极小

重复：迈更多步

- 第一步走到x_1

重复：迈更多步

- 第一步走到x_1
- 然后在x_1点，再找一个下山的方向p_1, 沿直线$x = x_1 + \alpha p_1$再跨出一步

重复：迈更多步

- 第一步走到x_1
- 然后在x_1点，再找一个下山的方向p_1，沿直线$x = x_1 + \alpha p_1$再跨出一步
- ……

重复：迈更多步

- 第一步走到x_1
- 然后在x_1点，再找一个下山的方向p_1, 沿直线$x = x_1 + \alpha p_1$再跨出一步
- ……
- 这样就得到一串参数α_i和方向p_i.

重复：迈更多步

- 第一步走到x_1
- 然后在x_1点，再找一个下山的方向p_1，沿直线$x = x_1 + \alpha p_1$再跨出一步
- ……
- 这样就得到一串参数α_i和方向p_i.
- p_i称为搜索方向，α_k为步长

重复：迈更多步

- 第一步走到x_1
- 然后在x_1点，再找一个下山的方向p_1, 沿直线$x = x_1 + \alpha p_1$再跨出一步
- ……
- 这样就得到一串参数α_i和方向p_i.
- p_i称为搜索方向，α_k为步长
- 不同的确定搜索方向和步长的方法，就得出不同的算法

步长的确定

- 设x_k已确定，下山方向p_k也确定

步长的确定

- 设x_k已确定，下山方向p_k也确定
- 任务：在直线$x = x_k + \alpha p_k$上确定α_k使得$\varphi(x)$在$x_{k+1} = x_k + \alpha_k p_k$处达到极小

步长的确定

- 设x_k已确定，下山方向p_k也确定
- 任务：在直线$x = x_k + \alpha p_k$上确定α_k使得$\varphi(x)$在$x_{k+1} = x_k + \alpha_k p_k$处达到极小
- 令$f(\alpha) = \varphi(x_k + \alpha p_k)$, 则

$$
\begin{aligned}
f(\alpha) &= (x_k + \alpha p_k)^T A(x_k + \alpha p_k) - 2b^T(x_k + \alpha p_k) \\
&= \alpha^2 p_k^T A p_k - 2\alpha r_k^T p_k + \varphi(x_k)
\end{aligned}
$$

其中$r_k = b - Ax_k$为$\varphi(x)$在$x = x_k$的负梯度方向

求导确定α_k

- 计算$f(\alpha)$的导数：

$$f'(\alpha) = 2\alpha p_k^T A p_k - 2 r_k^T p_k$$

求导确定α_k

- 计算$f(\alpha)$的导数：

$$f'(\alpha) = 2\alpha p_k^T A p_k - 2 r_k^T p_k$$

- 令$f'(\alpha) = 0$即得$\alpha_k = \dfrac{r_k^T p_k}{p_k^T A p_k}$，由此算出$x_{k+1}$

求导确定α_k

- 计算$f(\alpha)$的导数：

$$f'(\alpha) = 2\alpha p_k^T A p_k - 2 r_k^T p_k$$

- 令$f'(\alpha) = 0$即得$\alpha_k = \dfrac{r_k^T p_k}{p_k^T A p_k}$，由此算出$x_{k+1}$
- 验证:

$$\begin{aligned}\varphi(x_{k+1}) - \varphi(x_k) &= \alpha_k^2 p_k^T A p_k - 2\alpha_k r_k^T p_k \\ &= -\frac{(r_k^T p_k)^2}{p_k^T A p_k}\end{aligned}$$

因此只要$r_k^T p_k \neq 0$，我们就有$\varphi(x_{k+1}) < \varphi(x_k)$

新到达点原搜索方向与新梯度方向垂直

- 在第k步确定了搜索方向p_k后，按照前述公式确定步长α_k，那么到达了新点$x_{k+1} = x_k + \alpha_k p_k$

新到达点原搜索方向与新梯度方向垂直

- 在第k步确定了搜索方向p_k后，按照前述公式确定步长α_k，那么到达了新点$x_{k+1} = x_k + \alpha_k p_k$
- 显然p_k与等值线$\varphi(x) = \varphi(x_{k+1})$相切于点$x = x_{k+1}$

新到达点原搜索方向与新梯度方向垂直

- 在第k步确定了搜索方向p_k后，按照前述公式确定步长α_k，那么到达了新点$x_{k+1} = x_k + \alpha_k p_k$
- 显然p_k与等值线$\varphi(x) = \varphi(x_{k+1})$相切于点$x = x_{k+1}$
- $r_{k+1} = b - Ax_{k+1}$是上述等值线在$x = x_{k+1}$处的法向量

新到达点原搜索方向与新梯度方向垂直

- 在第k步确定了搜索方向p_k后，按照前述公式确定步长α_k, 那么到达了新点$x_{k+1} = x_k + \alpha_k p_k$
- 显然p_k与等值线$\varphi(x) = \varphi(x_{k+1})$相切于点$x = x_{k+1}$
- $r_{k+1} = b - Ax_{k+1}$是上述等值线在$x = x_{k+1}$处的法向量
- 所以$r_{k+1}^T p_k = 0$

新到达点原搜索方向与新梯度方向垂直

- 在第k步确定了搜索方向p_k后，按照前述公式确定步长α_k, 那么到达了新点$x_{k+1} = x_k + \alpha_k p_k$
- 显然p_k与等值线$\varphi(x) = \varphi(x_{k+1})$相切于点$x = x_{k+1}$
- $r_{k+1} = b - Ax_{k+1}$是上述等值线在$x = x_{k+1}$处的法向量
- 所以$r_{k+1}^T p_k = 0$
- 后面我们用代数方法证明更一般性的结论

下山方向的确定：最速下降法

- $\varphi(x)$增加最快的方向是梯度方向

下山方向的确定：最速下降法

- $\varphi(x)$增加最快的方向是梯度方向
- 因此负梯度方向应该是$\varphi(x)$减小最快的方向

下山方向的确定：最速下降法

- $\varphi(x)$增加最快的方向是梯度方向
- 因此负梯度方向应该是$\varphi(x)$减小最快的方向
- 所以我们取p_k为负梯度方向

$$r_k = b - Ax_k$$

收敛定理的准备：一个引理

引理

设A的特征值为$0<\lambda_1\leqslant\cdots\leqslant\lambda_n$, $P(t)$是一个关于t的多项式，则对$\forall x\in\mathbb{R}^n$

$$\|P(A)x\|_A\leqslant\max_{1\leqslant i\leqslant n}|P(\lambda_i)|\|x\|_A$$

其中$\|x\|_A=\sqrt{x^TAx}$

引理的证明

- 取由A对应于$\lambda_1, \ldots, \lambda_n$的特征向量$y_1, \ldots, y_n$，其构成$\mathbb{R}^n$的一组标准正交基

引理的证明

- 取由A对应于$\lambda_1, \ldots, \lambda_n$的特征向量$y_1, \ldots, y_n$，其构成$\mathbb{R}^n$的一组标准正交基
- 对任意$x \in \mathbb{R}^n$, $x = \sum_{i=1}^{n} \beta_i y_i$, 从而有

$$
\begin{aligned}
x^T P(A)AP(A)x &= \left(\sum_{i=1}^{n} \beta_i P(\lambda_i) y_i\right)^T A \left(\sum_{i=1}^{n} \beta_i P(\lambda_i) y_i\right) \\
&= \sum_{i=1}^{n} \lambda_i \beta_i^2 P^2(\lambda_i) \leqslant \max_{1 \leqslant i \leqslant n} P^2(\lambda_i) \sum_{i=1}^{n} \lambda_i \beta_i^2 \\
&= \max_{1 \leqslant i \leqslant n} P^2(\lambda_i) x^T A x
\end{aligned}
$$

收敛定理

定理

设A的特征值为$0 < \lambda_1 \leqslant \cdots \leqslant \lambda_n$，则由最速下降法产生的序列$\{x_k\}$满足

$$\|x_k - x_*\|_A \leqslant \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^k \|x_0 - x_*\|_A$$

其中$x_* = A^{-1}b$

定理证明

- 展开$(x-x_*)^TA(x-x_*)$，并利用$Ax_*=b$可得

$$\varphi(x)+x_*^TAx_*=(x-x_*)^TA(x-x_*)$$

定理证明

- 展开$(x-x_*)^TA(x-x_*)$，并利用$Ax_*=b$可得

$$\varphi(x)+x_*^TAx_*=(x-x_*)^TA(x-x_*)$$

- 根据x_k的构造方法，我们有

$$\varphi(x_k)\leqslant\varphi(x_{k-1}+\alpha r_{k-1}),\quad\forall\alpha\in\mathbb{R}$$

定理证明

- 展开$(x-x_*)^TA(x-x_*)$，并利用$Ax_*=b$可得

$$\varphi(x)+x_*^TAx_*=(x-x_*)^TA(x-x_*)$$

- 根据x_k的构造方法，我们有

$$\varphi(x_k)\leqslant\varphi(x_{k-1}+\alpha r_{k-1}),\quad \forall\alpha\in\mathbb{R}$$

- 所以

$$\begin{aligned}&(x_k-x_*)^TA(x_k-x_*)\\ \leqslant&(x_{k-1}+\alpha r_{k-1}-x_*)^TA(x_{k-1}+\alpha r_{k-1}-x_*)\end{aligned}$$

定理证明

- 根据构造方法，

  $$r_{k-1} = b - Ax_{k-1} = A(x_* - x_{k-1})$$

定理证明

- 根据构造方法，

  $r_{k-1} = b - Ax_{k-1} = A(x_* - x_{k-1})$

- 所以我们有

$$
\begin{aligned}
&(x_k - x_*)^T A(x_k - x_*) \\
\leqslant&(x_{k-1} + \alpha r_{k-1} - x_*)^T A(x_{k-1} + \alpha r_{k-1} - x_*) \\
=&((I - \alpha A)(x_{k-1} - x_*))^T A((I - \alpha A)(x_{k-1} - x_*))
\end{aligned}
$$

- 取$P_\alpha(t) = 1 - \alpha t$, 则由引理对$\forall \alpha \in \mathbb{R}$有

$$\begin{aligned}\|x_k - x_*\|_A &\leqslant \|P_\alpha(A)(x_{k-1} - x_*)\|_A \\ &\leqslant \max_{1\leqslant i\leqslant n} |P_\alpha(\lambda_i)| \|x_{k-1} - x_*\|_A\end{aligned}$$

- 取$P_\alpha(t) = 1 - \alpha t$, 则由引理对$\forall \alpha \in \mathbb{R}$有

$$\begin{aligned}\|x_k - x_*\|_A &\leqslant \|P_\alpha(A)(x_{k-1} - x_*)\|_A \\ &\leqslant \max_{1\leqslant i\leqslant n} |P_\alpha(\lambda_i)| \|x_{k-1} - x_*\|_A\end{aligned}$$

- 根据Chebyshev多项式的性质，

$$\min_\alpha \max_{\lambda_1\leqslant t\leqslant\lambda_n} |1 - \alpha t|$$

应在$1 - \alpha\lambda_1$与$1 - \alpha\lambda_n$互为相反数时达到，此时$\alpha = \dfrac{2}{\lambda_1 + \lambda_n}$, 极值为$\dfrac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}$

注解

- 上述定理表明，从任一初始向量x_0出发，由最速下降法产生的点列总是收敛到方程组的解

注解

- 上述定理表明，从任一初始向量x_0出发，由最速下降法产生的点列总是收敛到方程组的解
- 收敛速度由$(\lambda_n - \lambda_1)/(\lambda_n + \lambda_1)$的大小决定的

注解

- 上述定理表明，从任一初始向量x_0出发，由最速下降法产生的点列总是收敛到方程组的解
- 收敛速度由$(\lambda_n-\lambda_1)/(\lambda_n+\lambda_1)$的大小决定的
- 最速下降法简单易实现，而且可以充分利用A的稀疏性，但在$\lambda_1 \ll \lambda_n$时速度变得非常之慢

注解

- 上述定理表明，从任一初始向量x_0出发，由最速下降法产生的点列总是收敛到方程组的解
- 收敛速度由$(\lambda_n - \lambda_1)/(\lambda_n + \lambda_1)$的大小决定的
- 最速下降法简单易实现，而且可以充分利用A的稀疏性，但在$\lambda_1 \ll \lambda_n$时速度变得非常之慢
- 在求解线性方程组时很少用它，但它的想法很重要，并且在非线性优化求解中有大量的应用和拓展

注解

- 上述定理表明，从任一初始向量x_0出发，由最速下降法产生的点列总是收敛到方程组的解
- 收敛速度由$(\lambda_n - \lambda_1)/(\lambda_n + \lambda_1)$的大小决定的
- 最速下降法简单易实现，而且可以充分利用A的稀疏性，但在$\lambda_1 \ll \lambda_n$时速度变得非常之慢
- 在求解线性方程组时很少用它，但它的想法很重要，并且在非线性优化求解中有大量的应用和拓展

共轭梯度法的动机

- 从局部来看，负梯度方向确实是最佳的下山方向

共轭梯度法的动机

- 从局部来看，负梯度方向确实是最佳的下山方向
- 但从整体看看，它并非最佳：迭代得到的各点连线具有明显的锯齿形状

共轭梯度法的动机

- 从局部来看，负梯度方向确实是最佳的下山方向
- 但从整体看看，它并非最佳：迭代得到的各点连线具有明显的锯齿形状
- 我们要寻找更好的下山方向，而且在方向寻找上付出的代价不要太大

共轭梯度法的计算过程

- 给定初始点x_0, 第一步仍然选负梯度方向为下山方向，即$p_0 = r_0$, 于是有

$$\alpha_0 = \frac{r_0^T r_0}{r_0^T A r_0}, x_1 = x_0 + \alpha_0 p_0, r_1 = b - Ax_1$$

共轭梯度法的计算过程

- 给定初始点x_0，第一步仍然选负梯度方向为下山方向，即$p_0 = r_0$，于是有

$$\alpha_0 = \frac{r_0^T r_0}{r_0^T A r_0}, x_1 = x_0 + \alpha_0 p_0, r_1 = b - Ax_1$$

- 在第$k+1 (k \geqslant 1)$步，下山方向不再简单地取r_k，而是在过点x_k由向量r_k，p_{k-1}所张成的二维平面π_2内找出使函数φ下降最快的方向作为p_k

共轭梯度法的计算过程

- 给定初始点x_0，第一步仍然选负梯度方向为下山方向，即$p_0 = r_0$，于是有

$$\alpha_0 = \frac{r_0^T r_0}{r_0^T A r_0}, x_1 = x_0 + \alpha_0 p_0, r_1 = b - Ax_1$$

- 在第$k+1 (k \geqslant 1)$步，下山方向不再简单地取r_k，而是在过点x_k由向量r_k，p_{k-1}所张成的二维平面π_2内找出使函数φ下降最快的方向作为p_k
  - 注意：$r_k \perp p_{k-1}$

新下山方向的计算

- 考虑φ在π_2上的限制：

$$
\begin{aligned}
\psi(\xi,\eta) &= \varphi(x_k+\xi r_k+\eta p_{k-1}) \\
&= (x_k+\xi r_k+\eta p_{k-1})^T A(x_k+\xi r_k+\eta p_{k-1}) \\
&\quad - 2b^T(x_k+\xi r_k+\eta p_{k-1})
\end{aligned}
$$

新下山方向的计算

- 考虑φ在π_2上的限制：

$$
\begin{aligned}
\psi(\xi,\eta) &= \varphi(x_k+\xi r_k+\eta p_{k-1}) \\
&= (x_k+\xi r_k+\eta p_{k-1})^T A(x_k+\xi r_k+\eta p_{k-1}) \\
&\quad - 2b^T(x_k+\xi r_k+\eta p_{k-1})
\end{aligned}
$$

- 分别对ξ,η求偏导，得到局部下降最快的方向

新下山方向的计算

- 考虑φ在π_2上的限制：

$$\begin{aligned}\psi(\xi,\eta) &= \varphi(x_k+\xi r_k+\eta p_{k-1}) \\ &= (x_k+\xi r_k+\eta p_{k-1})^T A(x_k+\xi r_k+\eta p_{k-1}) \\ &\quad - 2b^T(x_k+\xi r_k+\eta p_{k-1})\end{aligned}$$

- 分别对ξ,η求偏导，得到局部下降最快的方向
  - 实际上直接求出的是在π_2中达到最小值的点

新下山方向的计算

- 求偏导：

$$
\begin{aligned}
\frac{\partial\psi}{\partial\xi} &= 2r_k^T A(x_k+\xi r_k+\eta p_{k-1}) - 2b^T r_k \\
&= 2(\xi r_k^T A r_k + \eta r_k^T A p_{k-1} - r_k^T r_k) \\
\frac{\partial\psi}{\partial\eta} &= 2p_{k-1}^T A(x_k+\xi r_k+\eta p_{k-1}) - 2b^T p_{k-1} \\
&= 2(\xi r_k^T A p_{k-1} + \eta p_{k-1}^T A p_{k-1})
\end{aligned}
$$

这里利用了$r_k^T p_{k-1} = 0$

- 由此得唯一极值点$\tilde{x} = x_k + \xi_0 r_k + \eta_0 p_{k-1}$，其中$\xi_0$和$\eta_0$满足

$$\begin{cases} \xi_0 r_k^T A r_k + \eta_0 r_k^T A p_{k-1} = r_k^T r_k \\ \xi_0 r_k^T A p_{k-1} + \eta_0 p_{k-1}^T A p_{k-1} = 0 \end{cases}$$

- 由此得唯一极值点$\tilde{x} = x_k + \xi_0 r_k + \eta_0 p_{k-1}$，其中$\xi_0$和$\eta_0$满足

$$\begin{cases} \xi_0 r_k^T A r_k + \eta_0 r_k^T A p_{k-1} = r_k^T r_k \\ \xi_0 r_k^T A p_{k-1} + \eta_0 p_{k-1}^T A p_{k-1} = 0 \end{cases}$$

- 由上式可知若$r_k \neq 0$, 则必有$\xi_0 \neq 0$ (为什么？)因此可取新的下山方向为

$$p_k = \frac{1}{\xi_0}(\tilde{x} - x_k) = r_k + \frac{\eta_0}{\xi_0} p_{k-1}$$

- 由此得唯一极值点$\tilde{x} = x_k + \xi_0 r_k + \eta_0 p_{k-1}$，其中$\xi_0$和$\eta_0$满足

$$\begin{cases} \xi_0 r_k^T A r_k + \eta_0 r_k^T A p_{k-1} = r_k^T r_k \\ \xi_0 r_k^T A p_{k-1} + \eta_0 p_{k-1}^T A p_{k-1} = 0 \end{cases}$$

- 由上式可知若$r_k \neq 0$, 则必有$\xi_0 \neq 0$ (为什么？)因此可取新的下山方向为

$$p_k = \frac{1}{\xi_0}(\tilde{x} - x_k) = r_k + \frac{\eta_0}{\xi_0} p_{k-1}$$

- η_0是否可以等于0呢？

- 令$\beta_{k-1}=\eta_0/\xi_0$，则由ξ_0和η_0满足的第二个方程可知

$$\beta_{k-1}=-\frac{r_k^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}$$

- 令$\beta_{k-1}=\eta_0/\xi_0$，则由ξ_0和η_0满足的第二个方程可知

$$\beta_{k-1}=-\frac{r_k^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}$$

- 如此确定的p_k满足

$$p_k^T A p_{k-1}=\left(r_k-\frac{r_k^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}p_{k-1}\right)^T A p_{k-1}=0$$

即p_k与p_{k-1}是关于A相互共轭的

公式初步梳理

- p_k确定以后，可以采用前面的方法确定α_k

公式初步梳理

- p_k确定以后，可以采用前面的方法确定α_k
- 总结公式为

$$\alpha_k = \frac{r_k^T p_k}{p_k^T A p_k}, \quad x_{k+1} = x_k + \alpha_k p_k$$

$$r_{k+1} = b - Ax_{k+1}$$

$$\beta_k = -\frac{r_{k+1}^T A p_k}{p_k^T A p_k}, \quad p_{k+1} = r_{k+1} + \beta_k p_k$$

r_{k+1}的简化

- 根据r_{k+1}的定义，

$$\begin{aligned} r_{k+1} &= b - Ax_{k+1} = b - A(x_k + \alpha_k p_k) \\ &= r_k - \alpha_k A p_k \end{aligned}$$

r_{k+1}的简化

- 根据r_{k+1}的定义，

$$
\begin{aligned}
r_{k+1} &= b - Ax_{k+1} = b - A(x_k + \alpha_k p_k) \\
&= r_k - \alpha_k A p_k
\end{aligned}
$$

- Ap_k在计算α_k时已求出，所以计算r_{k+1}时就可以用上述递推公式得到

r_{k+1}的简化

- 根据r_{k+1}的定义，

$$
\begin{aligned}
r_{k+1} &= b - Ax_{k+1} = b - A(x_k + \alpha_k p_k) \\
&= r_k - \alpha_k A p_k
\end{aligned}
$$

- Ap_k在计算α_k时已求出，所以计算r_{k+1}时就可以用上述递推公式得到
- 由上式可得

$$
Ap_k = \frac{1}{\alpha_k}(r_k - r_{k+1})
$$

α_k和β_k的简化

- 注意等式（证明后面给出）

$$r_k^T r_{k+1} = r_k^T p_{k-1} = r_{k+1}^T p_k = 0, k = 1, 2, \ldots$$

α_k和β_k的简化

- 注意等式（证明后面给出）

$$r_k^T r_{k+1} = r_k^T p_{k-1} = r_{k+1}^T p_k = 0, k = 1, 2, \ldots$$

- 从而我们有

$$\begin{aligned} r_{k+1}^T A p_k &= \frac{1}{\alpha_k} r_{k+1}^T (r_k - r_{k+1}) = -\frac{1}{\alpha_k} r_{k+1}^T r_{k+1} \\ p_k^T A p_k &= \frac{1}{\alpha_k} p_k^T (r_k - r_{k+1}) = \frac{1}{\alpha_k} p_k^T r_k \\ &= \frac{1}{\alpha_k} r_k^T (r_k + \beta_{k-1} p_{k-1}) = \frac{1}{\alpha_k} r_k^T r_k \end{aligned}$$

α_k和β_k的简化

- 回忆：

$$\alpha_k = \frac{r_k^T p_k}{p_k^T A p_k} \quad \beta_k = -\frac{r_{k+1}^T A p_k}{p_k^T A p_k}$$

α_k和β_k的简化

- 回忆：

$$\alpha_k = \frac{r_k^T p_k}{p_k^T A p_k} \quad \beta_k = -\frac{r_{k+1}^T A p_k}{p_k^T A p_k}$$

- 对α_k的分式进行简化；并且把前页两式相除，可对β_k进行简化：

$$\alpha_k = \frac{r_k^T r_k}{p_k^T A p_k} \quad \beta_k = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$$

基本性质

由共轭梯度法得到的向量组$\{r_i\}$和$\{p_i\}$具有下面的性质：

1. $p_i^T r_j = 0,\ 0 \leqslant i < j \leqslant k$

基本性质

由共轭梯度法得到的向量组$\{r_i\}$和$\{p_i\}$具有下面的性质：

1. $p_i^T r_j = 0,\ 0 \leqslant i < j \leqslant k$
2. $r_i^T r_j = 0,\ i \neq j,\ 0 \leqslant i, j \leqslant k$

基本性质

共轭梯度法
邓建松
基本框架
步长的确定
最速下降法
共轭梯度法及其基本性质
共轭梯度法
基本性质
实用共轭梯度法及其收敛性
实用共轭梯度法
收敛性分析

由共轭梯度法得到的向量组$\{r_i\}$和$\{p_i\}$具有下面的性质：

1. $p_i^T r_j = 0,\ 0 \leqslant i < j \leqslant k$
2. $r_i^T r_j = 0,\ i \neq j,\ 0 \leqslant i, j \leqslant k$
3. $p_i^T A p_j = 0,\ i \neq j,\ 0 \leqslant i, j \leqslant k$

基本性质

由共轭梯度法得到的向量组$\{r_i\}$和$\{p_i\}$具有下面的性质：

1. $p_i^T r_j = 0,\ 0 \leqslant i < j \leqslant k$
2. $r_i^T r_j = 0,\ i \neq j,\ 0 \leqslant i, j \leqslant k$
3. $p_i^T A p_j = 0,\ i \neq j,\ 0 \leqslant i, j \leqslant k$
4. 定义$\mathcal{K}(A, r_0, k+1) = \mathrm{span}\{r_0, Ar_0, \ldots, A^k r_0\}$，称为Krylov子空间，则$\mathrm{span}\{r_0, \ldots, r_k\} = \mathrm{span}\{p_0, \ldots, p_k\} = \mathcal{K}(A, r_0, k+1)$

$k = 1$时性质的证明

- 对k进行归纳证明

$k = 1$时性质的证明

- 对k进行归纳证明
- 当$k = 1$时，因为

$$
\begin{aligned}
& p_0 = r_0, r_1 = r_0 - \alpha_0 A p_0, p_1 = r_1 + \beta_0 p_0, \\
& p_0^T r_1 = r_1^T r_0 = r_0^T (r_0 - \alpha_0 A r_0) \\
& \qquad = r_0^T r_0 - \alpha_0 r_0^T A r_0 = 0, \\
& p_1^T A p_0 = (r_1 + \beta_0 r_0)^T A r_0 \\
& \qquad = r_1^T A r_0 - \frac{r_1^T A r_0}{r_0^T A r_0} r_0^T A r_0 = 0
\end{aligned}
$$

所以性质成立

性质(1)

假设性质在k时成立，我们证明在$k+1$时也成立

1. 由于$r_{k+1}=r_k-\alpha_k Ap_k$以及归纳假设，我们有

$$p_i^T r_{k+1}=p_i^T r_k-\alpha_k p_i^T Ap_k=0, 0\leqslant i\leqslant k-1$$

又由于

$$p_k^T r_{k+1}=p_k^T r_k-\frac{p_k^T r_k}{p_k^T Ap_k}p_k^T Ap_k=0$$

所以性质(1)在$k+1$时成立

性质(2)

2. 由归纳假设

$$\mathrm{span}\{r_0, \ldots, r_k\} = \mathrm{span}\{p_0, \ldots, p_k\}$$

则由性质(1)可知r_{k+1}与上述空间正交，从而性质(2)在$k+1$时成立

性质(3)

3. 根据归纳假设，当$i = 0, 1, \ldots, k-1$时

$$\begin{aligned}
p_i^T A p_{k+1} &= p_i^T A(r_{k+1} + \beta_k p_k) = r_{k+1}^T A p_i \\
&= \frac{1}{\alpha_i} r_{k+1}^T (r_i - r_{i+1}) = 0 \\
p_{k+1}^T A p_k &= (r_{k+1} + \beta_k p_k)^T A p_k \\
&= r_{k+1}^T A p_k - \frac{r_{k+1}^T A p_k}{p_k^T A p_k} p_k^T A p_k = 0
\end{aligned}$$

所以性质(3)成立

性质(4)

4. 由归纳假设可知

$$r_k, p_k \in \mathcal{K}(A, r_0, k+1) = \mathrm{span}\{r_0, Ar_0, \ldots, A^k r_0\}$$

于是

$$\begin{aligned} r_{k+1} &= r_k - \alpha_k A p_k \in \mathcal{K}(A, r_0, k+2), \\ p_{k+1} &= r_k + \beta_k A p_k \in \mathcal{K}(A, r_0, k+2), \end{aligned}$$

而根据性质(2),(3)，$r_0, \ldots, r_{k+1}$和$p_0, \ldots, p_{k+1}$都是线性无关的，所以性质(4)成立

Krylov子空间

- 前述性质表明，向量组$r_0, \ldots, r_k$和$p_0, \ldots, p_k$分别是Krylov子空间$\mathcal{K}(A, r_0, k+1)$的正交基和共轭正交基

Krylov子空间

- 前述性质表明，向量组$r_0, \dots, r_k$和$p_0, \dots, p_k$分别是Krylov子空间$\mathcal{K}(A, r_0, k+1)$的正交基和共轭正交基
- 所以采用共轭梯度法至多n步就得到方程组的解x_*

Krylov子空间

- 前述性质表明，向量组$r_0, \ldots, r_k$和$p_0, \ldots, p_k$分别是Krylov子空间$\mathcal{K}(A, r_0, k+1)$的正交基和共轭正交基
- 所以采用共轭梯度法至多n步就得到方程组的解x_*
- 理论上讲，共轭梯度法是直接法

精度估计

定理

用共轭梯度法计算得到的近似解x_k满足

$$\varphi(x_k) = \min\{\varphi(x) : x \in x_0 + \mathcal{K}(A, r_0, k)\}$$

或者等价地表示为

$$\|x_k - x_*\|_A = \min\{\|x - x_*\|_A : x \in x_0 + \mathcal{K}(A, r_0, k)\}$$

定理证明

- 由$\varphi(x) + x_*^T A x_* = (x - x_*)^T A (x - x_*)$可知要证的两式是等价的。下面只证第二式成立

定理证明

- 由$\varphi(x) + x_*^T A x_* = (x - x_*)^T A(x - x_*)$可知要证的两式是等价的。下面只证第二式成立
- 假设共轭梯度法计算到ℓ步出现$r_\ell = 0$, 那么有

$$\begin{aligned} x_* = x_\ell &= x_{\ell-1} + \alpha_{\ell-1} p_{\ell-1} \\ &= x_{\ell-2} + \alpha_{\ell-2} p_{\ell-2} + \alpha_{\ell-1} p_{\ell-1} \\ &= \cdots\cdots \\ &= x_0 + \alpha_0 p_0 + \alpha_1 p_1 + \cdots + \alpha_{\ell-1} p_{\ell-1} \end{aligned}$$

- 而对计算过程中任一步$k < \ell$, 我们有

$$x_k = x_0 + \sum_{j=0}^{k-1} \alpha_j p_j \in x_0 + \mathcal{K}(A, r_0, k)$$

- 而对计算过程中任一步$k < \ell$, 我们有

$$x_k = x_0 + \sum_{j=0}^{k-1} \alpha_j p_j \in x_0 + \mathcal{K}(A, r_0, k)$$

- 设$x \in \mathcal{K}(A, r_0, k)$为任一向量，则$x$有表示

$$x = x_0 + \sum_{j=0}^{k-1} \gamma_j p_j$$

- 而对计算过程中任一步$k < \ell$, 我们有

$$x_k = x_0 + \sum_{j=0}^{k-1} \alpha_j p_j \in x_0 + \mathcal{K}(A, r_0, k)$$

- 设$x \in \mathcal{K}(A, r_0, k)$为任一向量，则$x$有表示

$$x = x_0 + \sum_{j=0}^{k-1} \gamma_j p_j$$

- 于是

$$x_* - x = \sum_{j=0}^{k-1} (\alpha_j - \gamma_j) p_j + \sum_{j=k}^{\ell-1} \alpha_j p_j$$

- 而$x_* - x_k = \sum_{j=k}^{\ell} \alpha_j p_j$，根据共轭梯度法的性质(3)可得

$$\|x_* - x\|_A^2 = \left\| \sum_{j=0}^{k-1} (\alpha_j - \gamma_j) p_j \right\|_A^2 + \left\| \sum_{j=k}^{\ell-1} \alpha_j p_j \right\|_A^2$$

$$\geqslant \left\| \sum_{j=k}^{\ell-1} \alpha_j p_j \right\|_A^2 = \|x_* - x_k\|_A^2$$

实用共轭梯度法

- 共轭梯度法在理论上确保至多n步得到方程组的精确解

实用共轭梯度法

- 共轭梯度法在理论上确保至多n步得到方程组的精确解
- 在实际使用时由于误差的存在，使得r_k之间的正交性很快损失，所以有限步终止性不再成立

实用共轭梯度法

- 共轭梯度法在理论上确保至多n步得到方程组的精确解
- 在实际使用时由于误差的存在，使得r_k之间的正交性很快损失，所以有限步终止性不再成立
- 而且在实际应用中，由于n一般很大，迭代n次迭代所耗费的计算时间令人无法接受

实用共轭梯度法

- 共轭梯度法在理论上确保至多n步得到方程组的精确解
- 在实际使用时由于误差的存在，使得r_k之间的正交性很快损失，所以有限步终止性不再成立
- 而且在实际应用中，由于n一般很大，迭代n次迭代所耗费的计算时间令人无法接受
- 所以通常仍把共轭梯度法作为一种迭代法使用，当$\|r_k\|$足够小或者达到指定迭代次数时终止

共轭梯度法的优点

- 在算法中，系数矩阵A仅仅用来由已知向量p产生向量Ap, 因此可以充分利用A的稀疏性，而且对某些提供矩阵A困难，而可以方便由p产生向量Ap的应用问题，这种方法十分有用

共轭梯度法的优点

- 在算法中，系数矩阵A仅仅用来由已知向量p产生向量Ap, 因此可以充分利用A的稀疏性，而且对某些提供矩阵A困难，而可以方便由p产生向量Ap的应用问题，这种方法十分有用
- 不需要预先估计任何参数就可以计算

共轭梯度法的优点

- 在算法中，系数矩阵A仅仅用来由已知向量p产生向量Ap, 因此可以充分利用A的稀疏性，而且对某些提供矩阵A困难，而可以方便由p产生向量Ap的应用问题，这种方法十分有用
- 不需要预先估计任何参数就可以计算
- 每次迭代的主要计算就是向量之间的运算，因此便于并行化

作为迭代法的收敛性估计

若系数矩阵与单位矩阵的差是一个秩为r的矩阵，而且r又很小的话，那么共轭梯度法收敛得很快

定理

如果$A = I + B$, $\operatorname{rank} B = r$, 那么共轭梯度法至多迭代$r + 1$步即可得到方程组$Ax = b$的精确解

定理证明

- 注意到$\operatorname{rank} B = r$蕴涵着

  $$\operatorname{span}\{r_0, Ar_0, \dots, A^k r_0\} = \operatorname{span}\{r_0, Br_0, \dots, B^k r_0\}$$

  的维数不会超过$r+1$, 因此定理成立。

误差估计

定理

用共轭梯度法求得的r_k有如下的误差估计：

$$\|x_k - x_*\|_A \leqslant 2\left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\right)^k \|x_0 - x_*\|_A$$

其中$\kappa_2 = \kappa_2(A) = \|A\|_2\|A^{-1}\|_2$

定理证明

- 由共轭梯度法的性质可知，对任意的$x \in x_0 + \mathcal{K}(A, r_0, k)$有

$$x_* - x = x_* - x_0 + \sum_{j=0}^{k-1} a_{k,j+1} A^j r_0$$
$$= \left(I + \sum_{j=1}^{k} a_{kj} A^j \right) A^{-1} r_0 = P_k(A) A^{-1} r_0$$

其中$P_k(\lambda) = 1 + \sum_{j=1}^{k} a_{kj} \lambda^j$

- 令$\mathcal{P}_k$为所有满足$P_k(0)=1$，且次数不超过k的实系数多项式全体，则（第一个不等号应用了“最速下降法”一节中已证的引理）

$$\begin{aligned}\|x_*-x_k\|_A &= \min\{\|x-x_*\|_A : x\in x_0+\mathcal{K}(A,r_0,k)\}\\ &= \min_{P_k\in\mathcal{P}_k}\|P_k(A)A^{-1}r_0\|_A\\ &\leqslant \min_{P_k\in\mathcal{P}_k}\max_{1\leqslant i\leqslant n}|P_k(\lambda_i)|\|A^{-1}r_0\|_A\\ &\leqslant \min_{P_k\in\mathcal{P}_k}\max_{a\leqslant\lambda\leqslant b}|P_k(\lambda)|\|x_*-x_0\|_A\end{aligned}$$

其中$0<a=\lambda_0\leqslant\cdots\leqslant\lambda_n=b$是$A$的特征值

- 根据Chebyshev多项式的性质，最优化问题$\min\limits_{P_k\in\mathcal{P}_k}\max\limits_{a\leqslant\lambda\leqslant b}|P_k(\lambda)|$有唯一解

$$\tilde{P}_k(\lambda)=\frac{T_k\left(\dfrac{b+a-2\lambda}{b-a}\right)}{T_k\left(\dfrac{b+a}{b-a}\right)}$$

其中$T_k(x)$是k次Chebyshev多项式

暂停：Chebyshev多项式

- Chebyshev多项式：$T_n(x) = cos(n \arccos x)$，它是定义在$[-1, 1]$上，在所有同首项系数的$n$次多项式中，它在$[-1, 1]$上的绝对最大值最小

暂停：Chebyshev多项式

- Chebyshev多项式：$T_n(x) = cos(n \arccos x)$，它是定义在$[-1, 1]$上，在所有同首项系数的$n$次多项式中，它在$[-1, 1]$上的绝对最大值最小
- 在$[-1, 1]$外用多项式形式直接延拓

暂停：Chebyshev多项式

- Chebyshev多项式：$T_n(x) = cos(n \arccos x)$，它是定义在$[-1, 1]$上，在所有同首项系数的$n$次多项式中，它在$[-1, 1]$上的绝对最大值最小
- 在$[-1, 1]$外用多项式形式直接延拓
- 递推公式：$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x)$，$T_0(x) = 1$，$T_1(x) = x$

暂停：Chebyshev多项式

- Chebyshev多项式：$T_n(x) = cos(n \arccos x)$，它是定义在$[-1,1]$上，在所有同首项系数的$n$次多项式中，它在$[-1,1]$上的绝对最大值最小
- 在$[-1,1]$外用多项式形式直接延拓
- 递推公式：$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x)$, $T_0(x) = 1$, $T_1(x) = x$
- 基于上述公式，可以证明当$|x| \geqslant 1$时有
$$T_n(x) = \frac{1}{2}\left(\left(x - \sqrt{x^2-1}\right)^n + \left(x + \sqrt{x^2-1}\right)^n\right)$$

- 当$\gamma > 1$, $x = (\gamma+1)/(\gamma-1)$时

$$x \pm \sqrt{x^2 - 1} = \frac{(\sqrt{\gamma} \pm 1)^2}{\gamma - 1},$$

- 当$\gamma > 1$, $x = (\gamma+1)/(\gamma-1)$时

$$x \pm \sqrt{x^2-1} = \frac{(\sqrt{\gamma}\pm 1)^2}{\gamma - 1},$$

- 从而有

$$T_n\left(\frac{\gamma+1}{\gamma-1}\right) = \frac{(\sqrt{\gamma}+1)^{2n} + (\sqrt{\gamma}-1)^{2n}}{2(\gamma-1)^n} \geqslant \frac{(\sqrt{\gamma}+1)^{2n}}{2(\gamma-1)^n}$$

重回证明

- 根据Chebyshev多项式的性质，我们有

$$
\begin{aligned}
\max_{a\leqslant x\leqslant b}|\tilde{P}_k(\lambda)| &= \frac{1}{T_k\left(\dfrac{b+a}{b-a}\right)} \\
&\leqslant \frac{2(b-a)^k}{(\sqrt{b}+\sqrt{a})^{2k}} \\
&= 2\left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\right)^k
\end{aligned}
$$

这就完成了证明($\kappa_2 = b/a$)

注解

- 上述估计是十分粗糙的

注解

- 上述估计是十分粗糙的
- 实际收敛速度往往比这个估计快得多

注解

- 上述估计是十分粗糙的
- 实际收敛速度往往比这个估计快得多
- 但这个结果告诉我们，只要系数矩阵是十分良态的（即$\kappa_2 \approx 1$），那么共轭梯度法就会收敛得很快

注解

- 上述估计是十分粗糙的
- 实际收敛速度往往比这个估计快得多
- 但这个结果告诉我们，只要系数矩阵是十分良态的（即$\kappa_2 \approx 1$），那么共轭梯度法就会收敛得很快
- 对比于最速下降法收敛估计中的因子，我们有

$$\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \geqslant \frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}}$$

非对称特征值问题计算方法

邓建松

2018 年 11 月 28 日

问题分析

- 求一个矩阵的特征值的问题实质上是求一个多项式的根的问题

问题分析

- 求一个矩阵的特征值的问题实质上是求一个多项式的根的问题
- 五阶及五阶以上的多项式的根一般不能用有限次代数运算求得

问题分析

- 求一个矩阵的特征值的问题实质上是求一个多项式的根的问题
- 五阶及五阶以上的多项式的根一般不能用有限次代数运算求得
- 所以矩阵特征值的计算方法本质上都是迭代的

问题分析

- 求一个矩阵的特征值的问题实质上是求一个多项式的根的问题
- 五阶及五阶以上的多项式的根一般不能用有限次代数运算求得
- 所以矩阵特征值的计算方法本质上都是迭代的
- 已有不少非常成熟的数值方法计算矩阵的全部或部分特征值和特征向量

问题分析

- 求一个矩阵的特征值的问题实质上是求一个多项式的根的问题
- 五阶及五阶以上的多项式的根一般不能用有限次代数运算求得
- 所以矩阵特征值的计算方法本质上都是迭代的
- 已有不少非常成熟的数值方法计算矩阵的全部或部分特征值和特征向量
- 本节只是介绍几种最常用的基本方法

基本概念

- 设$A \in \mathbb{C}^{n\times n}$, 复数$\lambda$称为$A$的一个特征值是指存在非零向量$x \in \mathbb{C}^n$使得$Ax = \lambda x$

基本概念

- 设$A \in \mathbb{C}^{n\times n}$，复数$\lambda$称为$A$的一个特征值是指存在非零向量$x \in \mathbb{C}^n$使得$Ax = \lambda x$
- 此时x称做A的属于λ的一个特征向量

基本概念

- 设$A \in \mathbb{C}^{n\times n}$, 复数$\lambda$称为$A$的一个特征值是指存在非零向量$x \in \mathbb{C}^n$使得$Ax = \lambda x$
- 此时x称做A的属于λ的一个特征向量
- λ是A的一个特征向量当且仅当$\det(\lambda I - A) = 0$

基本概念

- 设$A \in \mathbb{C}^{n\times n}$, 复数$\lambda$称为$A$的一个特征值是指存在非零向量$x \in \mathbb{C}^n$使得$Ax = \lambda x$
- 此时x称做A的属于λ的一个特征向量
- λ是A的一个特征向量当且仅当$\det(\lambda I - A) = 0$
- 多项式$p_A(\lambda) = \det(\lambda I - A)$称为$A$的特征多项式

特征多项式

- $p_A(\lambda)$是一个首一的n次多项式

特征多项式

- $p_A(\lambda)$是一个首一的n次多项式
- 由代数基本定理知$p_A(\lambda)$有n个根，即A有n个特征值

特征多项式

- $p_A(\lambda)$是一个首一的n次多项式
- 由代数基本定理知$p_A(\lambda)$有n个根，即A有n个特征值
- 记A的特征值全体为$\lambda(A)$, 称之为A的谱集

特征多项式

- $p_A(\lambda)$是一个首一的n次多项式
- 由代数基本定理知$p_A(\lambda)$有n个根，即A有n个特征值
- 记A的特征值全体为$\lambda(A)$, 称之为A的谱集
- 假设$p_A(\lambda) = \prod_{j=1}^{r}(\lambda - \lambda_j)^{n_j}$, 其中$n_1 + \cdots + n_r = n$, $\lambda_i \neq \lambda_j$ $(i \neq j)$, 则称n_i为λ_i的代数重数, 而称$m_i = n - \mathrm{rank}(\lambda_i I - A)$为$\lambda_i$的几何重数

- 显然$m_i \leqslant n_i$

- 显然$m_i \leqslant n_i$
- 如果$n_i = 1$, 则称λ_i为A的一个单特征值；否则称λ_i是A的一个重特征值

- 显然$m_i \leqslant n_i$
- 如果$n_i = 1$, 则称λ_i为A的一个单特征值；否则称λ_i是A的一个重特征值
- 如果$n_i = m_i$, 则称λ_i为A的一个半单特征值

- 显然$m_i \leqslant n_i$
- 如果$n_i = 1$, 则称λ_i为A的一个单特征值；否则称λ_i是A的一个重特征值
- 如果$n_i = m_i$, 则称λ_i为A的一个半单特征值
- 如果A的所有特征值都是半单的，则称A是非亏损的

- 显然$m_i \leqslant n_i$
- 如果$n_i = 1$, 则称λ_i为A的一个单特征值；否则称λ_i是A的一个重特征值
- 如果$n_i = m_i$, 则称λ_i为A的一个半单特征值
- 如果A的所有特征值都是半单的，则称A是非亏损的
- A是非亏损的当且仅当A有n个线性无关的特征向量，即A是可对角化的

相似变换

- 设$A, B \in \mathbb{C}^{n\times n}$. 若存在非奇异阵$X \in \mathbb{C}^{n\times n}$使得$B = XAX^{-1}$, 则称$A$与$B$是相似的，而上述变换称为相似变换

相似变换

- 设$A, B \in \mathbb{C}^{n\times n}$. 若存在非奇异阵$X \in \mathbb{C}^{n\times n}$使得$B = XAX^{-1}$, 则称$A$与$B$是相似的，而上述变换称为相似变换
- 若A与B相似，则它们有相同的特征值，而且x是A的一个特征向量当且仅当$y = Xx$是B 的一个特征向量

相似变换

- 设$A, B \in \mathbb{C}^{n\times n}$. 若存在非奇异阵$X \in \mathbb{C}^{n\times n}$使得$B = XAX^{-1}$, 则称$A$与$B$是相似的，而上述变换称为相似变换
- 若A与B相似，则它们有相同的特征值，而且x是A的一个特征向量当且仅当$y = Xx$是B 的一个特征向量
- 矩阵特征值问题的基本思想：把给定矩阵相似变换为易于计算特征值和特征向量的矩阵

Jordan分解定理

- 设$A \in \mathbb{C}^{n\times n}$有$r$个互不相同的特征值$\lambda_1, \ldots, \lambda_r$，其（代数）重数分别为$n(\lambda_1), \ldots, n(\lambda_r)$，则存在非奇异矩阵$P \in \mathbb{C}^{n\times n}$

$$P^{-1}AP = \mathrm{diag}(J(\lambda_1), \ldots, J(\lambda_r))$$

其中

$$J(\lambda_i) = \mathrm{diag}(J_1(\lambda_i), \ldots, J_{k_i}(\lambda_i)) \in \mathbb{C}^{n(\lambda_i)\times n(\lambda_i)}$$

Jordan块

$$J_j(\lambda_i) = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{n_j(\lambda_i)\times n_j(\lambda_i)}$$

其中

$$n_1(\lambda_i) + \cdots + n_{k_i}(\lambda_i) = n(\lambda_i)$$

Schur分解定理

- 设$A \in \mathbb{C}^{n\times n}$，则存在酉矩阵$U \in \mathbb{C}^{n\times n}$ 使得

$$U^* A U = T$$

其中T是上三角阵。适当选取U可使T的对角元按任意指定的顺序排列

Schur分解定理

- 设$A \in \mathbb{C}^{n\times n}$，则存在酉矩阵$U \in \mathbb{C}^{n\times n}$ 使得

  $$U^* AU = T$$

  其中T是上三角阵。适当选取U可使T的对角元按任意指定的顺序排列
- 特征值问题求解的QR方法就是基于这一定理而设计的

Gerschgorin圆盘定理

- 设$A = (a_{ij}) \in \mathbb{C}^{n\times n}$，令

$$G_i(A) = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leqslant \sum_{j\neq i} |a_{ij}| \right\}$$

则有

$$\lambda(A) = \bigcup_{j=1}^{n} G_j(A)$$

左、右特征向量

- 设λ是矩阵$A \in \mathbb{C}^{n\times n}$的一个特征值

左、右特征向量

- 设λ是矩阵$A \in \mathbb{C}^{n\times n}$的一个特征值
- 则关于$x \in \mathbb{C}^n$的方程$Ax = \lambda x$有解，关于$y \in \mathbb{C}^n$的方程$y^T A = \lambda y^T$也有解

左、右特征向量

- 设λ是矩阵$A \in \mathbb{C}^{n\times n}$的一个特征值
- 则关于$x \in \mathbb{C}^n$的方程$Ax = \lambda x$有解，关于$y \in \mathbb{C}^n$的方程$y^TA = \lambda y^T$也有解
- 显然x是前面所定义的相应于λ的特征向量

左、右特征向量

- 设λ是矩阵$A \in \mathbb{C}^{n\times n}$的一个特征值
- 则关于$x \in \mathbb{C}^n$的方程$Ax = \lambda x$有解，关于$y \in \mathbb{C}^n$的方程$y^T A = \lambda y^T$也有解
- 显然x是前面所定义的相应于λ的特征向量
- 我们称x是相应于λ的右特征向量；而y是相应于λ的左特征向量

左、右特征向量

- 设λ是矩阵$A \in \mathbb{C}^{n\times n}$的一个特征值
- 则关于$x \in \mathbb{C}^n$的方程$Ax = \lambda x$有解，关于$y \in \mathbb{C}^n$的方程$y^TA = \lambda y^T$也有解
- 显然x是前面所定义的相应于λ的特征向量
- 我们称x是相应于λ的右特征向量；而y是相应于λ的左特征向量
- y是矩阵A^T相应于λ的右特征向量

左、右特征向量

- 设x_i是A对应于特征值λ_i的右特征向量，y_j是A对应于特征值λ_j的左特征向量

左、右特征向量

- 设x_i是A对应于特征值λ_i的右特征向量，y_j是A对应于特征值λ_j的左特征向量
- 若$\lambda_i \neq \lambda_j$，则必有$y_j^T x_i = 0$. 注意：即使是在复数域上进行运算，这里也只是进行矩阵转置，没有共轭(对于非零复向量x, 可能有$x^T x = 0$)

左、右特征向量

- 设x_i是A对应于特征值λ_i的右特征向量，y_j是A对应于特征值λ_j的左特征向量
- 若$\lambda_i \neq \lambda_j$，则必有$y_j^T x_i = 0$. 注意：即使是在复数域上进行运算，这里也只是进行矩阵转置，没有共轭(对于非零复向量x, 可能有$x^T x = 0$)
- 实际上，在$Ax_i = \lambda_i x_i$上左乘y_j^T; 在$y_j^T A = \lambda_j y_j^T$上右乘$x_i$, 然后两式相减，即得

$$(\lambda_i - \lambda_j) y_j^T x_i = 0$$

左、右特征向量

- 设x_i是A对应于特征值λ_i的右特征向量，y_j是A对应于特征值λ_j的左特征向量
- 若$\lambda_i \neq \lambda_j$，则必有$y_j^T x_i = 0$. 注意：即使是在复数域上进行运算，这里也只是进行矩阵转置，没有共轭(对于非零复向量x, 可能有$x^T x = 0$)
- 实际上，在$Ax_i = \lambda_i x_i$上左乘y_j^T；在$y_j^T A = \lambda_j y_j^T$上右乘$x_i$, 然后两式相减，即得
$$(\lambda_i - \lambda_j) y_j^T x_i = 0$$
- 若$\lambda_i = \lambda_j$，结论如何？

多项式的友阵

- 给定首一n次多项式

$$p(x) = x^n + p_{n-1}x^{n-1} + \cdots + p_1 x + p_0$$

多项式的友阵

- 给定首一n次多项式

$$p(x) = x^n + p_{n-1}x^{n-1} + \cdots + p_1 x + p_0$$

- 容易验证矩阵

$$C = \begin{pmatrix} -p_{n-1} & -p_{n-2} & \cdots & -p_1 & -p_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

的特征多项式就是$p(x)$

友阵

- C称为$p(x)$的友阵

友阵

- C称为$p(x)$的友阵
- 那么C的Jordan标准型是什么？

友阵

- C称为$p(x)$的友阵
- 那么C的Jordan标准型是什么？
- 或者问：当矩阵A的特征多项式是$p(x)$时，何时A与C是相似的？

友阵

- C称为$p(x)$的友阵
- 那么C的Jordan标准型是什么？
- 或者问：当矩阵A的特征多项式是$p(x)$时，何时A与C是相似的？
- 结论：若对应于不同特征值只有一个Jordan块，或者说只有一个特征向量时，矩阵A相似于C（参考：J.H. Wilkinson的著作【代数特征值问题】）

特征值和特值向量敏感性分析

- 从数值计算的角度来看，我们应首先弄清要计算的特征值和特征向量是否是病态的，即矩阵元素的小变化，是否会引起所关心的特征值和特征向量的巨大变化

特征值和特值向量敏感性分析

- 从数值计算的角度来看，我们应首先弄清要计算的特征值和特征向量是否是病态的，即矩阵元素的小变化，是否会引起所关心的特征值和特征向量的巨大变化
- 对一般矩阵而言，这一问题非常复杂

特征值和特值向量敏感性分析

- 从数值计算的角度来看，我们应首先弄清要计算的特征值和特征向量是否是病态的，即矩阵元素的小变化，是否会引起所关心的特征值和特征向量的巨大变化
- 对一般矩阵而言，这一问题非常复杂
- 我们这里只介绍一个简单而又非常重要的结果

- 假设λ是A的一个单特征值，x是相应的特征向量，而且$\|x\|_2 = 1$

- 假设λ是A的一个单特征值，x是相应的特征向量，而且$\|x\|_2 = 1$
- 令$U = (x, U_2) \in \mathbb{C}^{n\times n}$是酉阵，则我们有

$$U^* AU = \begin{pmatrix} \lambda & x^* AU_2 \\ 0 & A_2 \end{pmatrix}$$

- 假设λ是A的一个单特征值，x是相应的特征向量，而且$\|x\|_2 = 1$
- 令$U = (x, U_2) \in \mathbb{C}^{n\times n}$是酉阵，则我们有

$$U^* AU = \begin{pmatrix} \lambda & x^* AU_2 \\ 0 & A_2 \end{pmatrix}$$

- 由于λ是A的单特征值，所以

$$\delta = \min_{\mu\in\lambda(A_2)} |\lambda - \mu| > 0$$

- 定义$\Sigma^{\perp} = U_2(\lambda I - A_2)^{-1} U_2^*$

- 定义$\Sigma^\perp = U_2(\lambda I - A_2)^{-1}U_2^*$
- 取y为A的属于λ的左特征向量(即$y^TA = \lambda y^T$)

- 定义$\Sigma^{\perp} = U_2(\lambda I - A_2)^{-1}U_2^*$
- 取y为A的属于λ的左特征向量(即$y^TA = \lambda y^T$)
- 由于λ是单特征值，所以$y^Tx \neq 0$. (为什么？)
  从而可以取y使得$y^Tx = 1$

- 定义$\Sigma^{\perp} = U_2(\lambda I - A_2)^{-1}U_2^*$
- 取y为A的属于λ的左特征向量(即$y^TA = \lambda y^T$)
- 由于λ是单特征值，所以$y^Tx \neq 0$. (为什么？)
  从而可以取y使得$y^Tx = 1$
  - 利用Jordan标准形考虑特征向量的形式

- 定义$\Sigma^{\perp} = U_2(\lambda I - A_2)^{-1}U_2^*$
- 取y为A的属于λ的左特征向量(即$y^TA = \lambda y^T$)
- 由于λ是单特征值，所以$y^Tx \neq 0$. (为什么？)
  从而可以取y使得$y^Tx = 1$
  - 利用Jordan标准形考虑特征向量的形式
- 给矩阵A以微小的扰动使其变为$\tilde{A}$,
  记$\varepsilon = \|\tilde{A} - A\|_2$, 则存在$\tilde{A}$的一个特征值$\tilde{\lambda}$和对应的特征向量$\tilde{x}$, 使得(证明略)
  $$|\tilde{\lambda} - \lambda| \leqslant \|y\|_2\varepsilon + O(\varepsilon^2), \|\tilde{x} - x\|_2 \leqslant \|\Sigma^{\perp}\|_2\varepsilon + O(\varepsilon^2)$$

条件数

- 上式表明λ和x的敏感性分别与$\|y\|_2$和$\|\Sigma^{\perp}\|_2$的大小有关

条件数

- 上式表明λ和x的敏感性分别与$\|y\|_2$和$\|\Sigma^{\perp}\|_2$的大小有关
- 我们分别称$\|y\|_2$和$\|\Sigma^{\perp}\|_2$为特征值λ和特征向量x的条件数. 记做

$$\mathrm{cond}(\lambda) = \|y\|_2, \quad \mathrm{cond}(x) = \|\Sigma^{\perp}\|_2$$

幂法：初步假设

- 幂法是计算矩阵模最大特征值和对应特征向量的一种迭代方法

幂法：初步假设

- 幂法是计算矩阵模最大特征值和对应特征向量的一种迭代方法
- 为了说明基本想法，我们先假定$A \in \mathbb{C}^{n\times n}$可对角化，即$A$有如下分解

$$A = X\Lambda X^{-1}$$

其中$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$,
$X = (x_1, \ldots, x_n) \in \mathbb{C}^{n\times n}$ 非奇异

幂法：初步假设

- 幂法是计算矩阵模最大特征值和对应特征向量的一种迭代方法
- 为了说明基本想法，我们先假定$A \in \mathbb{C}^{n\times n}$可对角化，即$A$有如下分解

$$A = X\Lambda X^{-1}$$

其中$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$,

$X = (x_1, \ldots, x_n) \in \mathbb{C}^{n\times n}$ 非奇异

- 再假定$|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$

开始迭代

- 任取一向量$u_0 \in \mathbb{C}^n$

开始迭代

- 任取一向量$u_0 \in \mathbb{C}^n$
- 由于X的列向量构成$\mathbb{C}^n$的一组基，所以有

$$u_0 = \sum_{j=1}^{n} \alpha_i x_i, \quad \alpha_i \in \mathbb{C}$$

开始迭代

- 任取一向量$u_0 \in \mathbb{C}^n$
- 由于X的列向量构成$\mathbb{C}^n$的一组基，所以有

$$u_0 = \sum_{j=1}^{n} \alpha_i x_i, \quad \alpha_i \in \mathbb{C}$$

- 从而我们有

$$\begin{aligned} A^k u_0 &= \sum_{j=1}^{n} \alpha_j A^k x_j = \sum_{j=1}^{n} \alpha_j \lambda_j^k x_j \\ &= \lambda_1^k \left(\alpha_1 x_1 + \sum_{j=2}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k x_j \right) \end{aligned}$$

- 由此即知

$$\lim_{k\to\infty}\frac{A^k u_0}{\lambda_1^k}=\alpha_1 x_1$$

- 由此即知
$$\lim_{k\to\infty}\frac{A^k u_0}{\lambda_1^k}=\alpha_1 x_1$$
- 这表明当$\alpha_1\neq 0$且k充分大时，向量$u_k=A^k u_0/\lambda_1^k$就是A的一个很好的近似特征向量

实际应用的难处

- 我们事先并不知道A的特征值λ_1

实际应用的难处

- 我们事先并不知道A的特征值λ_1
- 对充分大的k, A^k的计算工作量太大

实际应用的难处

- 我们事先并不知道A的特征值λ_1
- 对充分大的k, A^k的计算工作量太大
- 因此直接应用上式计算A的近似特征值有困难

变通

- 在上式中λ_1^k只是改变向量的长度，但不改方向，因此我们不必用λ_1^k约化$A^k u_0$

变通

- 在上式中λ_1^k只是改变向量的长度，但不改方向，因此我们不必用λ_1^k约化$A^k u_0$
- 但这种约化也是必要的，以防止溢出

变通

- 在上式中λ_1^k只是改变向量的长度，但不改方向，因此我们不必用λ_1^k约化$A^k u_0$
- 但这种约化也是必要的，以防止溢出
- 在计算$A^k u_0$时也不必先算出A^k，只需迭代地进行就可以了

幂法的迭代格式

- 任意取定$u_0 \in \mathbb{C}^n$为初始向量，通常要求$\|u_0\|_\infty = 1$; $k \leftarrow 1$

幂法的迭代格式

- 任意取定$u_0 \in \mathbb{C}^n$为初始向量，通常要求$\|u_0\|_\infty = 1$; $k \leftarrow 1$
- $y_k = Au_{k-1}$

幂法的迭代格式

- 任意取定$u_0 \in \mathbb{C}^n$为初始向量，通常要求$\|u_0\|_\infty = 1$; $k \leftarrow 1$
- $y_k = Au_{k-1}$
- $\mu_k = \zeta_j^{(k)}$, $\zeta_j^{(k)}$是y_k的模最大分量

幂法的迭代格式

- 任意取定$u_0 \in \mathbb{C}^n$为初始向量，通常要求$\|u_0\|_\infty = 1$; $k \leftarrow 1$
- $y_k = Au_{k-1}$
- $\mu_k = \zeta_j^{(k)}$, $\zeta_j^{(k)}$是y_k的模最大分量
- $u_k = y_k / \mu_k$

幂法的迭代格式

非对称特征值问题计算方法

邓建松

基本概念与性质

幂法

反幂法

QR方法

基本迭代和收敛性

实Schur标准形

上Hessenberg化

带原点位移的QR迭代

双重步位移的QR迭代

隐式QR算法

- 任意取定$u_0 \in \mathbb{C}^n$为初始向量，通常要求$\|u_0\|_\infty = 1$; $k \leftarrow 1$
- $y_k = Au_{k-1}$
- $\mu_k = \zeta_j^{(k)}$, $\zeta_j^{(k)}$是y_k的模最大分量
- $u_k = y_k/\mu_k$
- $k \leftarrow k+1$, 返回第二步

幂法的迭代格式

- 任意取定$u_0 \in \mathbb{C}^n$为初始向量，通常要求$\|u_0\|_\infty = 1$; $k \leftarrow 1$
- $y_k = Au_{k-1}$
- $\mu_k = \zeta_j^{(k)}$, $\zeta_j^{(k)}$是y_k的模最大分量
- $u_k = y_k/\mu_k$
- $k \leftarrow k+1$, 返回第二步
- 这一迭代方法称做幂法

幂法的收敛性定理

定理

设$A \in \mathbb{C}^{n\times n}$有$r$个互不相同的特征值，且满足$|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_r|$, 而且$\lambda_1$是半单的。如果初始向量$u_0$在$\lambda_1$的特征子空间上的投影不为零，则由幂法产生的向量序列$\{u_k\}$收敛到λ_1的一个特征向量x_1（极限平行于x_0在特征子空间上投影向量），而且$\{\mu_k\}$收敛到λ_1

收敛定理的证明

- 根据假设，A的Jordan分解为

$$A = X\,\mathrm{diag}(J_1, \ldots, J_r)X^{-1}$$

其中$X \in \mathbb{C}^{n\times n}$非奇异，$J_i$是由属于$\lambda_i$的Jordan块构成的上三角阵，$n_1 + \cdots + n_r = n$. 由于$\lambda_1$是半单的，所以$J_1 = \lambda_1 I_{n_1}$

收敛定理的证明

- 根据假设，A的Jordan分解为

$$A = X\,\mathrm{diag}(J_1, \ldots, J_r)X^{-1}$$

其中$X \in \mathbb{C}^{n\times n}$非奇异，$J_i$是由属于$\lambda_i$的Jordan块构成的上三角阵，$n_1 + \cdots + n_r = n$. 由于$\lambda_1$是半单的，所以$J_1 = \lambda_1 I_{n_1}$

- 令$y = X^{-1}u_0$, 并将y与X分块：

$$y = (y_1^T, y_2^T, \ldots, y_{n_r}^T)^T,$$

$$X = (X_1, X_2, \ldots, X_r)$$

其中y_i^T有n_i个元素，X_i有n_i列

- 从而我们有

$$\begin{aligned} A^k u_0 &= X \operatorname{diag}(J_1^k, \ldots, J_r^k) X^{-1} u_0 \\ &= X_1 J_1^k y_1 + X_2 J_2^k y_2 + \cdots + X_r J_r^k y_r \\ &= \lambda_1^k X_1 y_1 + X_2 J_2^k y_2 + \cdots + X_r J_r^k y_r \\ &= \lambda_1^k \left(X_1 y_1 + \sum_{j=2}^{r} X_j \left(\frac{J_j}{\lambda_1} \right)^k y_j \right) \end{aligned}$$

- 注意到$\lambda_1^{-1}J_i$的谱半径为$\rho(\lambda_1^{-1}J_i) = |\lambda_i|/|\lambda_1| < 1$, $i = 2, \ldots, r$, 所以我们有

$$\lim_{k\to\infty} \frac{1}{\lambda_1^k} A^k u_0 = X_1 y_1$$

- 注意到$\lambda_1^{-1}J_i$的谱半径为$\rho(\lambda_1^{-1}J_i) = |\lambda_i|/|\lambda_1| < 1$, $i = 2, \dots, r$, 所以我们有
$$\lim_{k\to\infty} \frac{1}{\lambda_1^k} A^k u_0 = X_1 y_1$$
- 根据假设（u_0在λ_1的特征子空间上的投影不为零），我们有$X_1 y_1 \neq 0$

- 注意到$\lambda_1^{-1}J_i$的谱半径为$\rho(\lambda_1^{-1}J_i) = |\lambda_i|/|\lambda_1| < 1$, $i = 2, \ldots, r$, 所以我们有

$$\lim_{k\to\infty} \frac{1}{\lambda_1^k} A^k u_0 = X_1 y_1$$

- 根据假设（u_0在λ_1的特征子空间上的投影不为零），我们有$X_1 y_1 \neq 0$
  - $y = X^{-1}u_0 \Rightarrow u_0 = Xy$, 所以$y_1$即为$u_0$在$\lambda_1$的特征子空间上的组合系数，基向量就是$X_1$中各列向量，从而$X_1 y_1$就是$u_0$在这个特征子空间中的投影向量

- 由幂法产生的$\{u_k\}$满足

$$u_k = \frac{Au_{k-1}}{\mu_k} = \frac{A^k u_0}{\mu_k \mu_{k-1} \cdots \mu_1}$$

- 由幂法产生的$\{u_k\}$满足

$$u_k = \frac{Au_{k-1}}{\mu_k} = \frac{A^k u_0}{\mu_k \mu_{k-1} \cdots \mu_1}$$

- $\|u_k\|_\infty = 1$, 即u_k至少有一个分量为1, 所以$\zeta_k = \mu_k \mu_{k-1} \cdots \mu_1$必为$A^k u_0$的一个模最大分量

- 由幂法产生的$\{u_k\}$满足

$$u_k = \frac{Au_{k-1}}{\mu_k} = \frac{A^k u_0}{\mu_k \mu_{k-1} \cdots \mu_1}$$

- $\|u_k\|_\infty = 1$, 即u_k至少有一个分量为1, 所以$\zeta_k = \mu_k \mu_{k-1} \cdots \mu_1$必为$A^k u_0$的一个模最大分量
- 从而ζ_k/λ_1^k就是$A^k u_0/\lambda_1^k$的一个模最大分量

- 由幂法产生的$\{u_k\}$满足

$$u_k = \frac{Au_{k-1}}{\mu_k} = \frac{A^k u_0}{\mu_k \mu_{k-1} \cdots \mu_1}$$

- $\|u_k\|_\infty = 1$, 即u_k至少有一个分量为1, 所以$\zeta_k = \mu_k \mu_{k-1} \cdots \mu_1$必为$A^k u_0$的一个模最大分量
- 从而ζ_k/λ_1^k就是$A^k u_0/\lambda_1^k$的一个模最大分量
- 这样由前述极限等式我们知下述极限存在：

$$\zeta = \lim_{k\to\infty} \frac{\zeta_k}{\lambda_1^k}$$

- 这样我们就有

$$\lim_{k\to\infty} u_k = \lim_{k\to\infty} \frac{A^k u_0}{\zeta_k} = \lim_{k\to\infty} \frac{A^k u_0/\lambda_1^k}{\zeta_k/\lambda_1^k} = \frac{X_1 y_1}{\zeta} = x_1$$

- 这样我们就有

$$\lim_{k\to\infty} u_k = \lim_{k\to\infty} \frac{A^k u_0}{\zeta_k} = \lim_{k\to\infty} \frac{A^k u_0/\lambda_1^k}{\zeta_k/\lambda_1^k} = \frac{X_1 y_1}{\zeta} = x_1$$

- 由于$\tilde{y} = (y_1^T, 0, \ldots, 0)^T$是$\mathrm{diag}(J_1, \ldots, J_r)$相应于$\lambda_1$的特征向量，所以$X_1 y_1 = X\tilde{y}$是$A$相应于$\lambda_1$的特征向量，这就说明$x_1$是属于$\lambda_1$的一个特征向量

- 这样我们就有

$$\lim_{k\to\infty} u_k = \lim_{k\to\infty} \frac{A^k u_0}{\zeta_k} = \lim_{k\to\infty} \frac{A^k u_0/\lambda_1^k}{\zeta_k/\lambda_1^k} = \frac{X_1 y_1}{\zeta} = x_1$$

- 由于$\tilde{y} = (y_1^T, 0, \ldots, 0)^T$是$\mathrm{diag}(J_1, \ldots, J_r)$相应于$\lambda_1$的特征向量，所以$X_1 y_1 = X\tilde{y}$是$A$相应于$\lambda_1$的特征向量，这就说明$x_1$是属于$\lambda_1$的一个特征向量
- 由$Au_{k-1} = \mu_k u_k$, $x_1 \neq 0$可知$\mu_k \to \lambda_1$

注解

- 若定理条件不满足，则由幂法产生的序列的收敛性分析变得非常复杂

注解

- 若定理条件不满足，则由幂法产生的序列的收敛性分析变得非常复杂
- 这时$\{u_k\}$可能有若干个收敛于不同向量的子列

注解

- 若定理条件不满足，则由幂法产生的序列的收敛性分析变得非常复杂
- 这时$\{u_k\}$可能有若干个收敛于不同向量的子列
- 例：$A = XDX^{-1}$, $D = \mathrm{diag}(3, 2, 1, -3)$,

$$X = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & 2 & 1 & 1 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

收敛速度分析

- 幂法的收敛速度主要取决于$|\lambda_2|/|\lambda_1|$的大小

收敛速度分析

- 幂法的收敛速度主要取决于$|\lambda_2|/|\lambda_1|$的大小
- 这个数总是小于1的，它越小收敛也就越快

收敛速度分析

- 幂法的收敛速度主要取决于$|\lambda_2|/|\lambda_1|$的大小
- 这个数总是小于1的，它越小收敛也就越快
- 为了加快收敛，我们可以采用位移的方法，即对$A-\mu I$应用幂法，这里选取恰当的μ，使$A-\mu I$的模最大特征值与其它特征值之模的距离更大

收敛速度分析

- 幂法的收敛速度主要取决于$|\lambda_2|/|\lambda_1|$的大小
- 这个数总是小于1的，它越小收敛也就越快
- 为了加快收敛，我们可以采用位移的方法，即对$A-\mu I$应用幂法，这里选取恰当的μ，使$A-\mu I$的模最大特征值与其它特征值之模的距离更大
- 对前例可取$\mu=3$, 从而收敛到-3对应的特征向量

求模第二大的特征值和对应特征向量

- 为了求模第二大的特征值和对应的特征向量，那么直接迭代是不行的

求模第二大的特征值和对应特征向量

- 为了求模第二大的特征值和对应的特征向量，那么直接迭代是不行的
- 需要对原矩阵进行降阶：即在知道了λ_1和x_1的前提下，把矩阵降低一阶，使它只包含A 的特征值$\lambda_2, \ldots, \lambda_n$，这一方法通常称为收缩技巧

求模第二大的特征值和对应特征向量

- 为了求模第二大的特征值和对应的特征向量，那么直接迭代是不行的
- 需要对原矩阵进行降阶：即在知道了λ_1和x_1的前提下，把矩阵降低一阶，使它只包含A 的特征值$\lambda_2, \ldots, \lambda_n$，这一方法通常称为收缩技巧
- 最简单实用的收缩技巧是利用正交变换

基于正交变换的收缩技巧

- 假设$Ax_1 = \lambda_1 x_1$

基于正交变换的收缩技巧

- 假设$Ax_1 = \lambda_1 x_1$
- 取酉矩阵P使得$Px_1 = \alpha e_1$ (可用Householder变换实现)

基于正交变换的收缩技巧

- 假设$Ax_1 = \lambda_1 x_1$
- 取酉矩阵P使得$Px_1 = \alpha e_1$ (可用Householder变换实现)
- 从而$PAx_1 = \lambda_1 Px_1 = \lambda_1 \alpha e_1$

基于正交变换的收缩技巧

- 假设$Ax_1 = \lambda_1 x_1$
- 取酉矩阵P使得$Px_1 = \alpha e_1$ (可用Householder变换实现)
- 从而$PAx_1 = \lambda_1 Px_1 = \lambda_1 \alpha e_1$
- 而$x_1 = \alpha P^* e_1$, 所以$PAP^* e_1 = \lambda_1 e_1$, 即

$$PAP^* = \begin{pmatrix} \lambda_1 & * \\ 0 & B_1 \end{pmatrix}$$

基于正交变换的收缩技巧

- 假设$Ax_1 = \lambda_1 x_1$
- 取酉矩阵P使得$Px_1 = \alpha e_1$ (可用Householder变换实现)
- 从而$PAx_1 = \lambda_1 Px_1 = \lambda_1 \alpha e_1$
- 而$x_1 = \alpha P^* e_1$, 所以$PAP^* e_1 = \lambda_1 e_1$, 即

$$PAP^* = \begin{pmatrix} \lambda_1 & * \\ 0 & B_1 \end{pmatrix}$$

- $n-1$阶矩阵B_1的特征值就是$\lambda_2, \ldots, \lambda_n$，对其应用幂法即可

注解

- 幂法的计算严重依赖于矩阵特征值的分布情况，因此实际用起来很不方便，特别是不适用于自动计算

注解

- 幂法的计算严重依赖于矩阵特征值的分布情况，因此实际用起来很不方便，特别是不适用于自动计算
- 当矩阵阶数非常高，无法利用其他更有效的算法时才用幂法计算少数几个模最大的特征值和相应的特征向量

注解

- 幂法的计算严重依赖于矩阵特征值的分布情况，因此实际用起来很不方便，特别是不适用于自动计算
- 当矩阵阶数非常高，无法利用其他更有效的算法时才用幂法计算少数几个模最大的特征值和相应的特征向量
- 基于幂法可以诱导出一些更有效的算法

反幂法

- 反幂法也称为**反迭代法**，它对A^{-1}应用幂法，从而求出A的模最小特征值和对应的特征向量

反幂法

- 反幂法也称为反迭代法，它对A^{-1}应用幂法，从而求出A的模最小特征值和对应的特征向量
- 其基本格式为

$$
\begin{aligned}
Ay_k &= z_{k-1}, \\
\mu_k &= \zeta_i, \ \zeta_i\text{是}y_k\text{的模最大分量} \\
z_k &= y_k/\mu_k
\end{aligned}
$$

迭代结果

- 若A的特征值为$|\lambda_n| < |\lambda_{n-1}| \leqslant \cdots \leqslant |\lambda_1|$，则$\{z_k\}$收敛到$A$的对应于$\lambda_n$的一个特征向量，而$\{\mu_k\}$收敛于$\lambda_n^{-1}$

迭代结果

- 若A的特征值为$|\lambda_n| < |\lambda_{n-1}| \leqslant \cdots \leqslant |\lambda_1|$，则$\{z_k\}$收敛到$A$的对应于$\lambda_n$的一个特征向量，而$\{\mu_k\}$收敛于$\lambda_n^{-1}$
- 收敛速度由$|\lambda_n|/|\lambda_{n-1}|$的大小决定

应用场合

- 实际应用中，反幂法主要是用来求特征向量的，是在用某种方法求得A的某个特征值λ_i的近似值$\tilde{\lambda}_i$之后，应用反幂法于$A-\tilde{\lambda}_i I$上

应用场合

- 实际应用中，反幂法主要是用来求特征向量的，是在用某种方法求得A的某个特征值λ_i的近似值$\tilde{\lambda}_i$之后，应用反幂法于$A-\tilde{\lambda}_i I$上
- 即实际计算中常用的是带位移的反幂法，此时迭代格式为：

$$(A-\mu I)v_k = z_{k-1},$$

$$z_k = v_k/\|v\|_2$$

格式分析

- 反幂法每迭代一次就需要解一个线性方程组，这要比幂法的运算量大得多

格式分析

- 反幂法每迭代一次就需要解一个线性方程组，这要比幂法的运算量大得多
- 由于方程组的系数矩阵不随k的变化而变化，所以可以事先对它进行列主元的LU分解，然后每次迭代就只需要解两个三角形方程组即可

格式分析

- 反幂法每迭代一次就需要解一个线性方程组，这要比幂法的运算量大得多
- 由于方程组的系数矩阵不随k的变化而变化，所以可以事先对它进行列主元的LU分解，然后每次迭代就只需要解两个三角形方程组即可
- 这里采用$\|\cdot\|_2$进行规范化，只是为了下面的分析方便。实际应用中完全可以采用$\|\cdot\|_\infty$进行规范化

收敛速度分析

- 迭代收敛速度取决于$|\lambda_1 - \mu|/|\lambda_2 - \mu|$的大小

收敛速度分析

- 迭代收敛速度取决于$|\lambda_1 - \mu|/|\lambda_2 - \mu|$的大小
- 从收敛速度的角度来看，μ取得越靠近A的某个特征值越好

收敛速度分析

- 迭代收敛速度取决于$|\lambda_1 - \mu|/|\lambda_2 - \mu|$的大小
- 从收敛速度的角度来看，μ取得越靠近A的某个特征值越好
- 但是当μ与A的某个特征值很靠近时，$A - \mu I$就与一个奇异矩阵很靠近，迭代时就需要求解一个非常病态的方程组

收敛速度分析

- 迭代收敛速度取决于$|\lambda_1-\mu|/|\lambda_2-\mu|$的大小
- 从收敛速度的角度来看，μ取得越靠近A的某个特征值越好
- 但是当μ与A的某个特征值很靠近时，$A-\mu I$就与一个奇异矩阵很靠近，迭代时就需要求解一个非常病态的方程组
- 实际计算的经验和理论分析的结果表明：$A-\mu I$的病态性并不影响其收敛速度，而且当μ与A的某个特征值很靠近时，常常只迭代一次就可以了

现象分析

- 设λ是A的一个单特征值，x是属于λ的单位特征向量

现象分析

- 设λ是A的一个单特征值，x是属于λ的单位特征向量
- 假定带位移的迭代格式中位移μ与λ十分靠近，且x是良态的，即$\mathrm{cond}(x)$不太大

现象分析

- 设λ是A的一个单特征值，x是属于λ的单位特征向量
- 假定带位移的迭代格式中位移μ与λ十分靠近，且x是良态的，即$\mathrm{cond}(x)$不太大
  - 设(x, U_2)是酉阵

现象分析

- 设λ是A的一个单特征值，x是属于λ的单位特征向量
- 假定带位移的迭代格式中位移μ与λ十分靠近，且x是良态的，即$\mathrm{cond}(x)$不太大
  - 设(x, U_2)是酉阵
  - 由条件数的定义知

$$\begin{aligned}\mathrm{cond}(x) &= \|U_2(\lambda I - A_2)^{-1}U_2^*\|_2 \\ &= \|(\lambda I - A_2)^{-1}U_2^*\|_2\end{aligned}$$

其中$A_2 = U_2^* A U_2$

- 假定给定z_0之后，我们用列主元Gauss消去法求解迭代格式中的线性方程组$(A-\mu I)v_1 = z_0$，得到计算解$\hat{v}_1$

- 假定给定z_0之后，我们用列主元Gauss消去法求解迭代格式中的线性方程组$(A-\mu I)v_1=z_0$，得到计算解$\hat{v}_1$
- 由Gauss消去法的误差分析知$\hat{v}_1$满足

$$(A-\mu I-E)\hat{v}_1=z_0$$

其中E与$A-\mu I$和z_0有关，但$\|E\|_2$有一致的上界，通常差不多就是机器精度

- 记$e = \hat{v}_1 - v_1 = (A - \mu I)^{-1} E \hat{v}_1$, 并将$e$分解为

$$e = x_1 + x_2$$

其中$x_1 \in \mathrm{span}\{x\}$, $x_2 \in \mathrm{span}\{x\}^{\perp}$, 则存在$\alpha \in \mathbb{C}$和$y \in \mathbb{C}^{n-1}$使得

$$x_1 = \alpha x, \quad x_2 = U_2 y$$

- 另一方面，我们有

$$A - \mu I = (x, U_2) \begin{pmatrix} \lambda - \mu & x^* A U_2 \\ 0 & A_2 - \mu I \end{pmatrix} \begin{pmatrix} x^* \\ U_2^* \end{pmatrix}$$

- 另一方面，我们有

$$A - \mu I = (x, U_2) \begin{pmatrix} \lambda - \mu & x^* A U_2 \\ 0 & A_2 - \mu I \end{pmatrix} \begin{pmatrix} x^* \\ U_2^* \end{pmatrix}$$

- 因而

$$\begin{aligned} &(A - \mu I)^{-1} \\ &= (x, U_2) \begin{pmatrix} \frac{1}{\lambda - \mu} & \frac{-1}{\lambda - \mu} x^* A U_2 (A_2 - \mu I)^{-1} \\ 0 & (A_2 - \mu I)^{-1} \end{pmatrix} \begin{pmatrix} x^* \\ U_2^* \end{pmatrix} \end{aligned}$$

- 注意到$x^*xx^* = x^*$, $x^*U_2 = 0$

- 注意到$x^*xx^* = x^*$, $x^*U_2 = 0$
- 这样我们就有

$$
\begin{aligned}
(A-\mu I)^{-1} &= \frac{xx^*}{\lambda-\mu}(I - AU_2(A_2-\mu I)^{-1}U_2^*) \\
&\quad + U_2(A_2-\mu I)^{-1}U_2^* \\
\alpha = x^*e &= x^*(A-\mu I)^{-1}E\hat{v}_1 \\
&= \frac{x^*}{\lambda-\mu}(I - AU_2(A_2-\mu I)^{-1}U_2^*)E\hat{v}_1 \\
y = U_2^*e &= U_2^*(A-\mu I)^{-1}E\hat{v}_1 \\
&= (A_2-\mu I)^{-1}U_2^*E\hat{v}_1
\end{aligned}
$$

- 注意到

$$(A_2-\mu I)^{-1}=(I+(\lambda-\mu)(A_2-\lambda I)^{-1})^{-1}(A_2-\lambda I)^{-1}$$

- 注意到

$$(A_2-\mu I)^{-1} = (I+(\lambda-\mu)(A_2-\lambda I)^{-1})^{-1}(A_2-\lambda I)^{-1}$$

- 所以当λ与μ很近时

$$s = \|(A_2-\mu I)^{-1}U_2^*\|_2 \approx \|(A_2-\lambda I)^{-1}U_2^*\|_2 = \mathrm{cond}(x$$

- 注意到

$$(A_2-\mu I)^{-1} = (I+(\lambda-\mu)(A_2-\lambda I)^{-1})^{-1}(A_2-\lambda I)^{-1}$$

- 所以当λ与μ很近时

$$s = \|(A_2-\mu I)^{-1}U_2^*\|_2 \approx \|(A_2-\lambda I)^{-1}U_2^*\|_2 = \mathrm{cond}(x$$

- 因此当x良态时，s也不会太大，于是

$$\|x_2\|_2 = \|y\|_2 \leqslant s\|E\hat{v}_1\|_2$$

就是一个不太大的量

- 另一方面，

$$|\alpha| \leqslant \frac{1}{|\lambda| - |\mu|}(1 + \|A\|_2 s)\|E\hat{v}_1\|_2$$

却是一个很大的量

- 另一方面，
$$|\alpha| \leqslant \frac{1}{|\lambda| - |\mu|}(1 + \|A\|_2 s)\|E\hat{v}_1\|_2$$
却是一个很大的量
- 这就是说求解方程组所引起的误差主要对其解在特征子空间span$\{x\}$上投影的长度有影响

- 另一方面，
$$|\alpha| \leqslant \frac{1}{|\lambda| - |\mu|}(1 + \|A\|_2 s)\|E\hat{v}_1\|_2$$
却是一个很大的量
- 这就是说求解方程组所引起的误差主要对其解在特征子空间span$\{x\}$上投影的长度有影响
- 而这对于我们要计算λ的近似特征向量是十分有利的

- 另一方面，
$$|\alpha| \leqslant \frac{1}{|\lambda| - |\mu|}(1 + \|A\|_2 s)\|E\hat{v}_1\|_2$$
却是一个很大的量
- 这就是说求解方程组所引起的误差主要对其解在特征子空间span$\{x\}$上投影的长度有影响
- 而这对于我们要计算λ的近似特征向量是十分有利的
- 因为我们关心的主要是所得向量的方向而不是它的长度

- 上面的分析实质上也表明，当μ靠近λ且x良态时，只需应用一次反幂法就可得到λ的较好近似特征向量

- 上面的分析实质上也表明，当μ靠近λ且x良态时，只需应用一次反幂法就可得到λ的较好近似特征向量
- 为此我们引进一些概念

达到机器精度

- 设机器精度为**u**

达到机器精度

- 设机器精度为$\mathbf{u}$
- 对于给定的$\mu \in \mathbb{C}$, 如果存在$E \in \mathbb{C}^{n\times n}$使得

$$\det(A + E - \mu I) = 0, \|E\|_2 = O(\mathbf{u})$$

我们称μ是A的一个**达到机器精度的近似特征值**

达到机器精度

- 设机器精度为$\mathbf{u}$
- 对于给定的$\mu \in \mathbb{C}$, 如果存在$E \in \mathbb{C}^{n\times n}$使得
$$\det(A + E - \mu I) = 0, \|E\|_2 = O(\mathbf{u})$$
我们称μ是A的一个达到机器精度的近似特征值
- 对于一个给定的$x \in \mathbb{C}^n$, 如果存在$F \in \mathbb{C}^{n\times n}$满足$\|F\|_2 = O(\mathbf{u})$, 使得$x$是$A + F$的特征向量，我们称$x$是$A$的一个达到机器精度的近似特征向量

两者的关系

- 若μ是A的一个达到机器精度的近似特征值，则存在$E \in \mathbb{C}^{n\times n}$满足$\|E\|_2 = O(\mathbf{u})$使得$(A+E-\mu I)y = 0$有非零解

两者的关系

- 若μ是A的一个达到机器精度的近似特征值，则存在$E \in \mathbb{C}^{n\times n}$ 满足$\|E\|_2 = O(\mathbf{u})$使得$(A + E - \mu I)y = 0$有非零解
- 设$y \in \mathbb{C}^n$满足$(A + E - \mu I)y = 0$, $\|y\|_2 = 1$, $\|E\|_2 = O(\mathbf{u})$, 则我们有

$$(A + E)y = \mu y$$

即y是A的一个达到机器精度的近似特征向量

- 若我们在迭代格式中取$z_0 = (A - \mu I)y$, 那么在精确计算的前提下，只需迭代一次就可得到A的达到机器精度的特征向量

- 若我们在迭代格式中取$z_0 = (A - \mu I)y$, 那么在精确计算的前提下，只需迭代一次就可得到A的达到机器精度的特征向量
- 在实际计算时我们无法按这种方式取初始向量，但这说明了在恰当选取初始向量之后，反幂法具有“一次迭代”性

λ病态时的情形

- 还需要指出的是，当λ比较病态时，利用反幂法再进行迭代，一般不会得到更好的近似特征向量

λ病态时的情形

- 还需要指出的是，当λ比较病态时，利用反幂法再进行迭代，一般不会得到更好的近似特征向量
- 例

$$A = \begin{pmatrix} 1 & 1 \\ 10^{-10} & 1 \end{pmatrix}$$

其特征值$\lambda_1 = 0.99999$, $\lambda_2 = 1.00001$, 对应的特征向量$x_1 = (1, -10^{-5})^T$, $x_2 = (1, 10^{-5})^T$. 取$\mu = 1$, $z_0 = (0, 1)^T$,两次迭代结果更差

初始向量的选取方法

- 利用随机数发生子程序随机地选取向量，规范化后作为初始向量

初始向量的选取方法

- 利用随机数发生子程序随机地选取向量，规范化后作为初始向量
- 半次迭代法：首先对$A-\mu I$进行LU分解，得到$A-\mu I=LU$, 第一次迭代为$LUv_1=z_0$. 选$z_0=Le$，其中$e=(1,\ldots,1)^T$, 则为了求v_1只需求解$Uv_1=e$

QR方法简介

- QR方法是自计算机问世以来矩阵计算的重大进展之一，也是目前计算一般矩阵的全部特征值和特征向量的最有效方法之一

QR方法简介

- QR方法是自计算机问世以来矩阵计算的重大进展之一，也是目前计算一般矩阵的全部特征值和特征向量的最有效方法之一
- QR方法是一种迭代方法，它利用正交相似变换把矩阵逐步约化为上三角阵或者拟上三角阵

QR方法简介

- QR方法是自计算机问世以来矩阵计算的重大进展之一，也是目前计算一般矩阵的全部特征值和特征向量的最有效方法之一
- QR方法是一种迭代方法，它利用正交相似变换把矩阵逐步约化为上三角阵或者拟上三角阵
- 基本收敛速度是二次的。当原矩阵是实对称时，可达到三次收敛

The top 10 algorithms of 1900s

- Metropolis Algorithm for Monte Carlo

The top 10 algorithms of 1900s

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming

The top 10 algorithms of 1900s

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods

The top 10 algorithms of 1900s

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations

The top 10 algorithms of 1900s

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations
- The Fortran Optimizing Compiler

The top 10 algorithms of 1900s

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations
- The Fortran Optimizing Compiler
- QR Algorithm for Computing Eigenvalues

The top 10 algorithms of 1900s

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations
- The Fortran Optimizing Compiler
- QR Algorithm for Computing Eigenvalues
- Quicksort Algorithm for Sorting

The top 10 algorithms of 1900s

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations
- The Fortran Optimizing Compiler
- QR Algorithm for Computing Eigenvalues
- Quicksort Algorithm for Sorting
- Fast Fourier Transform

The top 10 algorithms of 1900s

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations
- The Fortran Optimizing Compiler
- QR Algorithm for Computing Eigenvalues
- Quicksort Algorithm for Sorting
- Fast Fourier Transform
- Integer Relation Detection

The top 10 algorithms of 1900s

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations
- The Fortran Optimizing Compiler
- QR Algorithm for Computing Eigenvalues
- Quicksort Algorithm for Sorting
- Fast Fourier Transform
- Integer Relation Detection
- Fast Multipole Method

历史

- J.G.F. Francis与V.N. Kublanovskaya在上世纪50年代独立发展出QR方法：基于QR分解

历史

- J.G.F. Francis与V.N. Kublanovskaya在上世纪50年代独立发展出QR方法：基于QR分解
- 经过几十年的发展，QR方法的现代版本称为隐式QR算法

历史

- J.G.F. Francis与V.N. Kublanovskaya在上世纪50年代独立发展出QR方法：基于QR分解
- 经过几十年的发展，QR方法的现代版本称为隐式QR算法
- 但在隐式QR算法中没有显式地进行QR分解，因此有人建议称之为Francis算法

历史

- J.G.F. Francis与V.N. Kublanovskaya在上世纪50年代独立发展出QR方法：基于QR分解
- 经过几十年的发展，QR方法的现代版本称为隐式QR算法
- 但在隐式QR算法中没有显式地进行QR分解，因此有人建议称之为Francis算法
- 在QR算法之前，曾出现过LR算法，它是基于LU分解的。LR算法是由H. Rutishauser在1950年代发展的

QR方法的基本迭代

- 对给定的$A_0 = A \in \mathbb{C}^{n\times n}$，QR方法的基本迭代格式如下：

$$
\begin{aligned}
A_{m-1} &= Q_m R_m, \quad (A_{m-1}\text{的QR分解}) \\
A_m &= R_m Q_m
\end{aligned}
$$

这里Q_m为酉阵，R_m为上三角阵

QR方法的基本迭代

- 对给定的$A_0 = A \in \mathbb{C}^{n\times n}$，QR方法的基本迭代格式如下：

$$\begin{aligned} A_{m-1} &= Q_m R_m, \quad (A_{m-1}\text{的QR分解}) \\ A_m &= R_m Q_m \end{aligned}$$

  这里Q_m为酉阵，R_m为上三角阵

- 课本例6.4.1和其它两个例子的程序演示

QR方法的基本迭代

- 对给定的$A_0 = A \in \mathbb{C}^{n\times n}$，QR方法的基本迭代格式如下：

$$
\begin{aligned}
A_{m-1} &= Q_m R_m, \quad (A_{m-1}\text{的QR分解}) \\
A_m &= R_m Q_m
\end{aligned}
$$

这里Q_m为酉阵，R_m为上三角阵

- 课本例6.4.1和其它两个例子的程序演示
- 为了下面理论分析方便，我们暂且要求R_m的对角元都是非负的

迭代过程分析

- 根据迭代格式，我们有$A_m = Q_m^* A_{m-1} Q_m$, 所以矩阵序列$\{A_m\}$中每一个矩阵都与A相似

迭代过程分析

- 根据迭代格式，我们有$A_m = Q_m^* A_{m-1} Q_m$，所以矩阵序列$\{A_m\}$中每一个矩阵都与A相似
- 从而有$A_m = \tilde{Q}_m^* A \tilde{Q}_m$，其中$\tilde{Q}_m = Q_1 Q_2 \cdots Q_m$

迭代过程分析

- 根据迭代格式，我们有$A_m = Q_m^* A_{m-1} Q_m$，所以矩阵序列$\{A_m\}$中每一个矩阵都与A相似
- 从而有$A_m = \tilde{Q}_m^* A \tilde{Q}_m$，其中$\tilde{Q}_m = Q_1 Q_2 \cdots Q_m$
- 即$A\tilde{Q}_m = \tilde{Q}_m A_m = \tilde{Q}_m Q_{m+1} R_{m+1}$

迭代过程分析

- 根据迭代格式，我们有$A_m = Q_m^* A_{m-1} Q_m$，所以矩阵序列$\{A_m\}$中每一个矩阵都与A相似
- 从而有$A_m = \tilde{Q}_m^* A \tilde{Q}_m$，其中$\tilde{Q}_m = Q_1 Q_2 \cdots Q_m$
- 即$A\tilde{Q}_m = \tilde{Q}_m A_m = \tilde{Q}_m Q_{m+1} R_{m+1}$
- 定义$\tilde{R}_k = R_k R_{k-1} \cdots R_1$，我们有

$$\tilde{Q}_{m+1} \tilde{R}_{m+1} = A \tilde{Q}_m \tilde{R}_m$$

迭代过程分析

- 根据迭代格式，我们有$A_m = Q_m^* A_{m-1} Q_m$，所以矩阵序列$\{A_m\}$中每一个矩阵都与A相似
- 从而有$A_m = \tilde{Q}_m^* A \tilde{Q}_m$，其中$\tilde{Q}_m = Q_1 Q_2 \cdots Q_m$
- 即$A\tilde{Q}_m = \tilde{Q}_m A_m = \tilde{Q}_m Q_{m+1} R_{m+1}$
- 定义$\tilde{R}_k = R_k R_{k-1} \cdots R_1$，我们有

$$\tilde{Q}_{m+1} \tilde{R}_{m+1} = A \tilde{Q}_m \tilde{R}_m$$

- 由此可归纳证明$A^m = \tilde{Q}_m \tilde{R}_m$，这是$A^m$的QR分解

QR方法与幂法的关系

- 记$\tilde{R}_m$的元素为γ_{ij}，$\tilde{Q}_m$的第一列为$q_1^{(m)}$，则由$A^m = \tilde{Q}_m \tilde{R}_m$可得

$$A^m e_1 = \gamma_{11} q_1^{(m)}$$

QR方法与幂法的关系

- 记$\tilde{R}_m$的元素为γ_{ij}, $\tilde{Q}_m$的第一列为$q_1^{(m)}$, 则由$A^m = \tilde{Q}_m\tilde{R}_m$可得

$$A^m e_1 = \gamma_{11} q_1^{(m)}$$

- 所以$q_1^{(m)}$可以看作是对A用e_1作初始向量的幂法所得到的向量

QR方法与幂法的关系

- 记$\tilde{R}_m$的元素为γ_{ij}，$\tilde{Q}_m$的第一列为$q_1^{(m)}$，则由$A^m = \tilde{Q}_m \tilde{R}_m$可得

$$A^m e_1 = \gamma_{11} q_1^{(m)}$$

- 所以$q_1^{(m)}$可以看作是对A用e_1作初始向量的幂法所得到的向量
- 若A的模最大特征值λ_1与其他特征值分离，那么$q_1^{(m)}$将收敛到A的一个属于λ_1的特征向量

A_m下三角元素趋于0

定理

设$A \in \mathbb{C}^{n\times n}$的$n$个特征值满足

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$$

并设n阶方阵Y的第i行是A对应于λ_i的左特征向量。如果Y有LU分解，则由QR方法产生的矩阵A_m的对角线以下的元素趋向于0，同时第i个对角元趋向于λ_i，$i = 1, 2, \ldots, n$

定理证明框架

- A可对角化，从而A^m也可对角化

定理证明框架

- A可对角化，从而A^m也可对角化
- 基于条件，构造A^m的QR分解，其中各元素在$m \to \infty$时的极限形状已知

定理证明框架

- A可对角化，从而A^m也可对角化
- 基于条件，构造A^m的QR分解，其中各元素在$m \to \infty$时的极限形状已知
- 利用$A^m = \tilde{Q}_m \tilde{R}_m$和QR分解的唯一性得到$\tilde{Q}_m$和$\tilde{R}_m$，其在$m \to \infty$时的极限形状已知

定理证明框架

- A可对角化，从而A^m也可对角化
- 基于条件，构造A^m的QR分解，其中各元素在$m \to \infty$时的极限形状已知
- 利用$A^m = \tilde{Q}_m \tilde{R}_m$和QR分解的唯一性得到$\tilde{Q}_m$和$\tilde{R}_m$，其在$m \to \infty$时的极限形状已知
- 利用$A_m = \tilde{Q}_m^* A \tilde{Q}_m$可证$A_m$下三角元素趋向于零

定理证明

- 令$X = Y^{-1}$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, 则有$A = X\Lambda Y$

定理证明

- 令$X = Y^{-1}$, $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, 则有$A = X\Lambda Y$
- 设Y的LU分解为$Y = LU$, 其中L是单位下三角阵，U是上三角阵

定理证明

- 令$X = Y^{-1}$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, 则有$A = X\Lambda Y$
- 设Y的LU分解为$Y = LU$, 其中L是单位下三角阵，U是上三角阵
- 从而我们有

$$
\begin{aligned}
A^m = X\Lambda^m Y = X\Lambda^m LU &= X(\Lambda^m L\Lambda^{-m})\Lambda^m U \\
&= X(I + E_m)\Lambda^m U
\end{aligned}
$$

构造A^m的QR分解

- 由于L是单位下三角阵，而$|\lambda_i| < |\lambda_j|(\ i > j)$, 所以有$\lim\limits_{m\to\infty} E_m = 0$

构造A^m的QR分解

- 由于L是单位下三角阵，而$|\lambda_i| < |\lambda_j|(\ i > j)$, 所以有$\lim\limits_{m\to\infty} E_m = 0$
- 令X的QR分解为$X = QR$. 由于X非奇异，所以可要求R的对角元全是正数。这样我们有

$$A^m = QR(I + E_m)\Lambda^m U = Q(I + RE_mR^{-1})R\Lambda^m U$$

构造A^m的QR分解

- 由于L是单位下三角阵，而$|\lambda_i| < |\lambda_j|(\ i > j)$, 所以有$\lim\limits_{m\to\infty} E_m = 0$
- 令X的QR分解为$X = QR$. 由于X非奇异，所以可要求R的对角元全是正数。这样我们有

$$A^m = QR(I + E_m)\Lambda^m U = Q(I + RE_mR^{-1})R\Lambda^m U$$

- 当m充分大时，$I + RE_mR^{-1}$非奇异，故可取它的QR分解$I + RE_mR^{-1} = \hat{Q}_m\hat{R}_m$, 其中$\hat{R}_m$的对角元均正数

- 由$E_m \to 0\ (m \to \infty)$可知

$$\lim_{m\to\infty} \hat{Q}_m = \lim_{m\to\infty} \hat{R}_m = I$$

- 由$E_m \to 0\ (m \to \infty)$可知

$$\lim_{m\to\infty} \hat{Q}_m = \lim_{m\to\infty} \hat{R}_m = I$$

- 至此我们有$A^m = (Q\hat{Q}_m)(\hat{R}_m R \Lambda^m U)$

- 由$E_m \to 0$ $(m \to \infty)$可知

$$\lim_{m\to\infty} \hat{Q}_m = \lim_{m\to\infty} \hat{R}_m = I$$

- 至此我们有$A^m = (Q\hat{Q}_m)(\hat{R}_m R \Lambda^m U)$
  - 这是A^m的一个QR分解，只是对角元可能不是正数

- 由$E_m \to 0$ $(m \to \infty)$可知
$$\lim_{m\to\infty} \hat{Q}_m = \lim_{m\to\infty} \hat{R}_m = I$$
- 至此我们有$A^m = (Q\hat{Q}_m)(\hat{R}_m R \Lambda^m U)$
  - 这是A^m的一个QR分解，只是对角元可能不是正数
- 为校正这一点，设u_{ii}是U的对角元，定义
$$D_1 = \operatorname{diag}\left(\frac{\lambda_1}{|\lambda_1|}, \ldots, \frac{\lambda_n}{|\lambda_n|}\right)$$
$$D_2 = \operatorname{diag}\left(\frac{u_{11}}{|u_{11}|}, \ldots, \frac{u_{nn}}{|u_{nn}|}\right)$$

- 这样我们就有

$$A^m = (Q\hat{Q}_m D_1^m D_2)(D_2^{-1} D_1^{-m} \hat{R}_m R \Lambda^m U)$$

- 这样我们就有

$$A^m = (Q\hat{Q}_m D_1^m D_2)(D_2^{-1} D_1^{-m} \hat{R}_m R \Lambda^m U)$$

- 由于$A^m = \tilde{Q}_m \tilde{R}_m$，所以根据QR分解的唯一性，我们有

$$\tilde{Q}_m = Q\hat{Q}_m D_1^m D_2, \tilde{R}_m = D_2^{-1} D_1^{-m} \hat{R}_m R \Lambda^m U$$

- 这样我们就有

$$A^m = (Q\hat{Q}_m D_1^m D_2)(D_2^{-1} D_1^{-m} \hat{R}_m R \Lambda^m U)$$

- 由于$A^m = \tilde{Q}_m \tilde{R}_m$, 所以根据QR分解的唯一性，我们有

$$\tilde{Q}_m = Q\hat{Q}_m D_1^m D_2, \tilde{R}_m = D_2^{-1} D_1^{-m} \hat{R}_m R \Lambda^m U$$

- 所以有

$$A_m = \left(D_2^* (D_1^*)^m \hat{Q}_m^* Q^*\right) A \left(Q\hat{Q}_m D_1^m D_2\right)$$

- 而注意到$A = X\Lambda Y = X\Lambda X^{-1} = QR\Lambda R^{-1}Q^*$

- 而注意到$A = X\Lambda Y = X\Lambda X^{-1} = QR\Lambda R^{-1}Q^*$
- 从而有

$$A_m = D_2^*(D_1^*)^m \hat{Q}_m^* R\Lambda R^{-1}\hat{Q}_m D_1^m D_2$$

- 而注意到$A = X\Lambda Y = X\Lambda X^{-1} = QR\Lambda R^{-1}Q^*$
- 从而有

$$A_m = D_2^*(D_1^*)^m \hat{Q}_m^* R\Lambda R^{-1}\hat{Q}_m D_1^m D_2$$

- 这里除$\hat{Q}_m$外，其它矩阵都是上三角阵，而$\hat{Q}_m \to I$ $(m \to \infty)$, 所以A_m的下三角元趋向于零

- 而注意到$A = X\Lambda Y = X\Lambda X^{-1} = QR\Lambda R^{-1}Q^*$
- 从而有

$$A_m = D_2^*(D_1^*)^m \hat{Q}_m^* R\Lambda R^{-1}\hat{Q}_m D_1^m D_2$$

- 这里除$\hat{Q}_m$外，其它矩阵都是上三角阵，而$\hat{Q}_m \to I$ $(m \to \infty)$, 所以A_m的下三角元趋向于零
- 从而对角元趋向于A的第i个特征值

只用到实数运算的QR方法

- 实际应用中所遇到的大量特征值问题都是关于实矩阵的，因此我们希望设计出只涉及到实数运算的QR迭代

只用到实数运算的QR方法

- 实际应用中所遇到的大量特征值问题都是关于实矩阵的，因此我们希望设计出只涉及到实数运算的QR迭代
- 如此的迭代格式为

$$A_{m-1} = Q_m R_m$$
$$A_m = R_m Q_m$$

其中Q_m是正交矩阵，R_m是上三角阵

只用到实数运算的QR方法

- 实际应用中所遇到的大量特征值问题都是关于实矩阵的，因此我们希望设计出只涉及到实数运算的QR迭代
- 如此的迭代格式为
$$A_{m-1} = Q_m R_m$$
$$A_m = R_m Q_m$$
其中Q_m是正交矩阵，R_m是上三角阵
- 但由于复共轭特征值的存在，因此我们不能期望迭代格式产生的A_m仍趋向于一个上三角阵

实Schur分解

定理

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

$$Q^TAQ = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ & R_{22} & \cdots & R_{2m} \\ & & \ddots & \vdots \\ & & & R_{mm} \end{pmatrix}$$

其中R_{ii}或者是一个实数，或者是一个具有一对复共轭特征值的2阶方阵

- 实Schur分解定理中右边的拟上三角阵称为A的实Schur标准形

- 实Schur分解定理中右边的拟上三角阵称为A的实Schur标准形
- 类似可证明实QR迭代格式产生的A_k应逼近于A的实Schur标准形

- 实Schur分解定理中右边的拟上三角阵称为A的实Schur标准形
- 类似可证明实QR迭代格式产生的A_k应逼近于A的实Schur标准形
- 但目前这个版本的QR方法在实用性方面没有竞争力，因为：

- 实Schur分解定理中右边的拟上三角阵称为A的实Schur标准形
- 类似可证明实QR迭代格式产生的A_k应逼近于A的实Schur标准形
- 但目前这个版本的QR方法在实用性方面没有竞争力，因为：
  - 每次迭代的运算量太大

- 实Schur分解定理中右边的拟上三角阵称为A的实Schur标准形
- 类似可证明实QR迭代格式产生的A_k应逼近于A的实Schur标准形
- 但目前这个版本的QR方法在实用性方面没有竞争力，因为：
  - 每次迭代的运算量太大
  - 收敛速度太慢

把QR方法改造成实用算法

- 把A相似到一种特殊矩阵，其QR分解计算简单，而且对QR迭代仍保持矩阵的特殊性不变

把QR方法改造成实用算法

- 把A相似到一种特殊矩阵，其QR分解计算简单，而且对QR迭代仍保持矩阵的特殊性不变
  - 上Hessenberg矩阵

把QR方法改造成实用算法

- 把A相似到一种特殊矩阵，其QR分解计算简单，而且对QR迭代仍保持矩阵的特殊性不变
  - 上Hessenberg矩阵
- 引用原点位移方法

把QR方法改造成实用算法

- 把A相似到一种特殊矩阵，其QR分解计算简单，而且对QR迭代仍保持矩阵的特殊性不变
  - 上Hessenberg矩阵
- 引用原点位移方法
- 只采用实数运算，处理共轭复特征值情况

把QR方法改造成实用算法

- 把A相似到一种特殊矩阵，其QR分解计算简单，而且对QR迭代仍保持矩阵的特殊性不变
  - 上Hessenberg矩阵
- 引用原点位移方法
- 只采用实数运算，处理共轭复特征值情况
  - 双重步位移

把QR方法改造成实用算法

- 把A相似到一种特殊矩阵，其QR分解计算简单，而且对QR迭代仍保持矩阵的特殊性不变
  - 上Hessenberg矩阵
- 引用原点位移方法
- 只采用实数运算，处理共轭复特征值情况
  - 双重步位移
- 实用算法：隐式QR算法

把QR方法改造成实用算法

- 把A相似到一种特殊矩阵，其QR分解计算简单，而且对QR迭代仍保持矩阵的特殊性不变
  - 上Hessenberg矩阵
- 引用原点位移方法
- 只采用实数运算，处理共轭复特征值情况
  - 双重步位移
- 实用算法：隐式QR算法
  - 平均两次QR迭代就可以分离出一个特征值或2×2子矩阵，计算量为$O(n^3)$

相似约化到准上三角阵

- 为了减少每次迭代的运算量，我们先把原矩阵A经相似变换约化为一个准上三角阵

相似约化到准上三角阵

- 为了减少每次迭代的运算量，我们先把原矩阵A经相似变换约化为一个准上三角阵
- 为此，我们先看一下基于Householder变换的相似变换可以得到什么

相似约化到准上三角阵

- 为了减少每次迭代的运算量，我们先把原矩阵A经相似变换约化为一个准上三角阵
- 为此，我们先看一下基于Householder变换的相似变换可以得到什么
- 第一步我们自然选取Householder变换H_1，使得H_1A的第一列有尽可能多的零元素

相似约化到准上三角阵

- 为了减少每次迭代的运算量，我们先把原矩阵A经相似变换约化为一个准上三角阵
- 为此，我们先看一下基于Householder变换的相似变换可以得到什么
- 第一步我们自然选取Householder变换H_1，使得H_1A的第一列有尽可能多的零元素
- 最多可以得到$n-1$个零。这能做到么？

相似变换的妥协

- 在对$A = (a_{ij})$左乘H_1进行行变换之后，还需要对A右乘H_1进行列变换

相似变换的妥协

- 在对$A=(a_{ij})$左乘H_1进行行变换之后，还需要对A右乘H_1进行列变换
- 为了保证已在H_1A中的第一列所出现的零元素不至于因右乘H_1被破坏，我们选取H_1具有如下形式：

$$H_1=\begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix}$$

$\tilde{H}_1$的构造

- 如此我们有

$$H_1AH_1 = \begin{pmatrix} a_{11} & a_2^T\tilde{H}_1 \\ \tilde{H}_1 a_1 & \tilde{H}_1 A_{22}\tilde{H}_1 \end{pmatrix}$$

其中$a_1^T = (a_{21}, a_{31}, \ldots, a_{n1})$,

$a_2^T = (a_{12}, a_{13}, \ldots, a_{1n})$, A_{22}是A右下角的$n-1$阶主子阵

$\tilde{H}_1$的构造

- 如此我们有

$$H_1AH_1 = \begin{pmatrix} a_{11} & a_2^T\tilde{H}_1 \\ \tilde{H}_1a_1 & \tilde{H}_1A_{22}\tilde{H}_1 \end{pmatrix}$$

其中$a_1^T = (a_{21}, a_{31}, \ldots, a_{n1})$, $a_2^T = (a_{12}, a_{13}, \ldots, a_{1n})$, A_{22}是A右下角的$n-1$阶主子阵

- 所以Householder变换$\tilde{H}_1$的最佳选择应该使得$\tilde{H}_1a_1 = pe_1$

- 如此H_1AH_1的第一列中除开始两个元素外，余下的$n-2$个元素为零

- 如此H_1AH_1的第一列中除开始两个元素外，余下的$n-2$个元素为零
- 类似对后面各列进行处理，我们可以找到$n-2$个Householder变换$H_1, \ldots, H_{n-2}$使得

$$H_{n-2} \cdots H_1 A H_1 \cdots H_{n-2} = H$$

其中$H = (h_{ij})$满足$h_{ij} = 0$, $i > j+1$, 这样的矩阵称为上Hessenberg矩阵

上Hessenberg分解

- 现在令

$$Q_0 = H_1 H_2 \cdots H_{n-2}$$

则有

$$Q_0^T A Q_0 = H$$

上Hessenberg分解

- 现在令
$$Q_0 = H_1 H_2 \cdots H_{n-2}$$
则有
$$Q_0^T A Q_0 = H$$
- 一般称如此分解式为A的上Hessenberg分解

上Hessenberg分解

- 现在令
$$Q_0 = H_1 H_2 \cdots H_{n-2}$$
则有
$$Q_0^T A Q_0 = H$$
- 一般称如此分解式为A的上Hessenberg分解
- 算法运算量为$10n^3/3$; 如果要记录Q, 还需要增加运算量$4n^3/3$

上Hessenberg分解

- 现在令

$$Q_0 = H_1 H_2 \cdots H_{n-2}$$

则有

$$Q_0^T A Q_0 = H$$

- 一般称如此分解式为A的上Hessenberg分解
- 算法运算量为$10n^3/3$; 如果要记录Q, 还需要增加运算量$4n^3/3$
- 算法6.4.1的程序演示

误差分析

- 可以证明：如此算法得到的上Hessenberg矩阵$\hat{H}$满足

$$\hat{H} = Q^T(A+E)Q$$

其中Q是正交矩阵，$\|E\|_F \leqslant cn^2\|A\|_F\mathbf{u}$

误差分析

- 可以证明：如此算法得到的上Hessenberg矩阵$\hat{H}$满足
$$\hat{H} = Q^T(A+E)Q$$
其中Q是正交矩阵，$\|E\|_F \leqslant cn^2\|A\|_F\mathbf{u}$
- 我们也可以采用Givens变换(运算量增加)或者列主元的Gauss消去法将A约化为上Hessenberg矩阵（运算量少，但稳定性较差）

唯一性定理

定理

设A有如下两个上Hessenberg分解：

$$U^TAU = H, \quad V^TAV = G$$

其中$U=(u_1,\dots,u_n)$和$V=(v_1,\dots,v_n)$是n阶正交矩阵，$H=(h_{ij})$和$G=(g_{ij})$是上Hessenberg矩阵。若$u_1=v_1$，而且H的次对角元$h_{i+1,i}$均不为零，则存在对角元均为1或-1的对角阵D，使得$U=VD$，$H=DGD$

定理证明

采用归纳法证明

- 假定对某个m已证$u_j = \varepsilon_j v_j$, $j = 1, \ldots, m$, 其中$\varepsilon_1 = 1$, $\varepsilon_j = 1$或-1, $j = 2, \ldots, m$. 下面证明存在$\varepsilon_{m+1} = 1$或-1使得$u_{m+1} = \varepsilon_{m+1} v_{m+1}$

定理证明

采用归纳法证明

- 假定对某个m已证$u_j = \varepsilon_j v_j$, $j = 1, \ldots, m$, 其中$\varepsilon_1 = 1$, $\varepsilon_j = 1$或-1, $j = 2, \ldots, m$. 下面证明存在$\varepsilon_{m+1} = 1$或-1使得$u_{m+1} = \varepsilon_{m+1} v_{m+1}$
- 根据上Hessenberg分解式，我们有

$$AU = UH, \quad AV = VG$$

U, V的关系反应到H, G上

- 比例两个矩阵等式的第m列，我们有

$$Au_m = h_{1m}u_1 + \cdots + h_{mm}u_m + h_{m+1,m}u_{m+1}$$

$$Av_m = g_{1m}v_1 + \cdots + g_{mm}v_m + g_{m+1,m}v_{m+1}$$

U, V的关系反应到H, G上

- 比例两个矩阵等式的第m列，我们有

$$Au_m = h_{1m}u_1 + \cdots + h_{mm}u_m + h_{m+1,m}u_{m+1}$$

$$Av_m = g_{1m}v_1 + \cdots + g_{mm}v_m + g_{m+1,m}v_{m+1}$$

- 在上式两边分别左乘u_i^T和v_i^T，可得

$$h_{im} = u_i^T A u_m, \quad g_{im} = v_i^T A v_m, i = 1, \ldots, m$$

U, V的关系反应到H, G上

- 比例两个矩阵等式的第m列，我们有

$$Au_m = h_{1m}u_1 + \cdots + h_{mm}u_m + h_{m+1,m}u_{m+1}$$
$$Av_m = g_{1m}v_1 + \cdots + g_{mm}v_m + g_{m+1,m}v_{m+1}$$

- 在上式两边分别左乘u_i^T和v_i^T，可得

$$h_{im} = u_i^T A u_m, \quad g_{im} = v_i^T A v_m, i = 1, \ldots, m$$

- 根据归纳假设，可得$h_{im} = \varepsilon_i \varepsilon_m g_{im}$，$i = 1, \ldots, m$

第$m+1$列的关系

- 从而我们有

$$\begin{aligned} h_{m+1,m}u_{m+1} &= \varepsilon_m(Av_m - \varepsilon_1^2 g_{1m}v_1 - \cdots - \varepsilon_m^2 g_{mm}v_m) \\ &= \varepsilon_m(Av_m - g_{1m}v_1 - \cdots - g_{mm}v_m) \\ &= \varepsilon_m g_{m+1,m}v_{m+1} \end{aligned}$$

第$m+1$列的关系

- 从而我们有

$$
\begin{aligned}
h_{m+1,m}u_{m+1} &= \varepsilon_m(Av_m - \varepsilon_1^2 g_{1m}v_1 - \cdots - \varepsilon_m^2 g_{mm}v_m) \\
&= \varepsilon_m(Av_m - g_{1m}v_1 - \cdots - g_{mm}v_m) \\
&= \varepsilon_m g_{m+1,m}v_{m+1}
\end{aligned}
$$

- 由此即得$|h_{m+1,m}| = |g_{m+1,m}|$

第$m+1$列的关系

- 从而我们有

$$\begin{aligned} h_{m+1,m}u_{m+1} &= \varepsilon_m(Av_m - \varepsilon_1^2 g_{1m}v_1 - \cdots - \varepsilon_m^2 g_{mm}v_m) \\ &= \varepsilon_m(Av_m - g_{1m}v_1 - \cdots - g_{mm}v_m) \\ &= \varepsilon_m g_{m+1,m}v_{m+1} \end{aligned}$$

- 由此即得$|h_{m+1,m}| = |g_{m+1,m}|$
- 而$h_{m+1,m} \neq 0$，所以我们有$u_{m+1} = \varepsilon_{m+1}v_{m+1}$，其中$\varepsilon_{m+1} = 1$或$-1$

变换矩阵的第一列

- 根据前面给出的分解方法，我们知道

$$H_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix}$$

变换矩阵的第一列

- 根据前面给出的分解方法，我们知道

$$H_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix}$$

- 所以最终的变换矩阵也具有形式

$$H = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H} \end{pmatrix}$$

变换矩阵的第一列

- 根据前面给出的分解方法，我们知道

$$H_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix}$$

- 所以最终的变换矩阵也具有形式

$$H = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H} \end{pmatrix}$$

- 实际上我们可以把H的第一列取为任意单位向量

变换矩阵的第一列

- 根据前面给出的分解方法，我们知道

$$H_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix}$$

- 所以最终的变换矩阵也具有形式

$$H = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H} \end{pmatrix}$$

- 实际上我们可以把H的第一列取为任意单位向量
  - 根据H的第一列确定一个正交阵Q

变换矩阵的第一列

- 根据前面给出的分解方法，我们知道

$$H_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix}$$

- 所以最终的变换矩阵也具有形式

$$H = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H} \end{pmatrix}$$

- 实际上我们可以把H的第一列取为任意单位向量
  - 根据H的第一列确定一个正交阵Q
  - 那么对Q^TAQ进行前述上Hessenberg化，设变换矩阵为$\tilde{H}$, $\tilde{H}$具有上述形式，那么$H = Q\tilde{H}$ 的第一列就是所指定的单位向量

不可约上Hessenerg矩阵

- 一个上Hessenberg矩阵的下次对角元均不为零，则称为不可约的

不可约上Hesseneng矩阵

- 一个上Hessenberg矩阵的下次对角元均不为零，则称为不可约的
- 之所以如此定义，如果有一个下次对角元为零，那么可以分别考虑上下子阵的特征值问题

不可约上Hessenerg矩阵

- 一个上Hessenberg矩阵的下次对角元均不为零，则称为**不可约的**
- 之所以如此定义，如果有一个下次对角元为零，那么可以分别考虑上下子阵的特征值问题
- 唯一性定理表明：如果$Q^TAQ = H$为不可约的上Hessenberg矩阵，则Q和H在相差一个正负号的意义下完全由Q的第一列确定

不可约上Hessenerg矩阵

- 一个上Hessenberg矩阵的下次对角元均不为零，则称为不可约的
- 之所以如此定义，如果有一个下次对角元为零，那么可以分别考虑上下子阵的特征值问题
- 唯一性定理表明：如果$Q^TAQ = H$为不可约的上Hessenberg矩阵，则Q和H在相差一个正负号的意义下完全由Q的第一列确定
- 这是得以建立隐式QR方法的关键所在

上Hessenberg矩阵的QR分解

- 假设$H \in \mathbb{R}^{n\times n}$是上Hessenberg矩阵

上Hessenberg矩阵的QR分解

- 假设$H \in \mathbb{R}^{n\times n}$是上Hessenberg矩阵
- H的QR分解可以通过$n-1$个平面旋转变换完成

$$P_{n-1,n}P_{n-2,n-1}\cdots P_{12}H = R$$

上Hessenberg矩阵的QR分解

- 假设$H \in \mathbb{R}^{n\times n}$是上Hessenberg矩阵
- H的QR分解可以通过$n-1$个平面旋转变换完成

$$P_{n-1,n}P_{n-2,n-1}\cdots P_{12}H = R$$

- 令$Q = (P_{n-1,n}P_{n-2,n-1}\cdots P_{12})^T$，则$H = QR$就是$H$的QR分解

上Hessenberg矩阵的QR分解

- 假设$H \in \mathbb{R}^{n\times n}$是上Hessenberg矩阵
- H的QR分解可以通过$n-1$个平面旋转变换完成

$$P_{n-1,n}P_{n-2,n-1}\cdots P_{12}H = R$$

- 令$Q = (P_{n-1,n}P_{n-2,n-1}\cdots P_{12})^T$，则$H = QR$就是$H$的QR分解
- 为了完成一次QR迭代，还需要计算$\tilde{H} = RQ = RP_{12}^TP_{23}^T\cdots P_{n-1,n}^T$

RQ的计算结果

- 由于P_{12}是$(1,2)$坐标平面内的旋转变换，因此RP_{12}^T仅有前两列与R不同，而RP_{12}^T的前两列是由R的前两列的线性组合构成，R为上三角阵，所以RP_{12}^T的第一个下次对角元非零

RQ的计算结果

- 由于P_{12}是$(1,2)$坐标平面内的旋转变换，因此RP_{12}^T仅有前两列与R不同，而RP_{12}^T的前两列是由R的前两列的线性组合构成，R为上三角阵，所以RP_{12}^T的第一个下次对角元非零
- 因此RQ仍是一个上Hessenberg矩阵

RQ的计算结果

- 由于P_{12}是$(1,2)$坐标平面内的旋转变换，因此RP_{12}^T仅有前两列与R不同，而RP_{12}^T的前两列是由R的前两列的线性组合构成，R为上三角阵，所以RP_{12}^T的第一个下次对角元非零
- 因此RQ仍是一个上Hessenberg矩阵
- 上Hessenberg矩阵经一次QR迭代后还是上Hessenberg矩阵，计算运算量为$O(n^2)$

QR分解开始消失

- 从一个不可约上Hessenberg矩阵H_k出发，经过一次QR迭代，得到新的上Hessenberg矩阵H_{k+1}，

$$H_{k+1} = Q_k^T H_k Q_k$$

QR分解开始消失

- 从一个不可约上Hessenberg矩阵H_k出发，经过一次QR迭代，得到新的上Hessenberg矩阵H_{k+1}，

$$H_{k+1} = Q_k^T H_k Q_k$$

- 这里的QR迭代实际上就是连续$n-1$次平面旋转操作，不同于标准的QR分解

QR分解开始消失

- 从一个不可约上Hessenberg矩阵H_k出发，经过一次QR迭代，得到新的上Hessenberg矩阵H_{k+1}，

$$H_{k+1} = Q_k^T H_k Q_k$$

- 这里的QR迭代实际上就是连续$n-1$次平面旋转操作，不同于标准的QR分解
- 回忆：上Hessenberg矩阵分解的唯一性

QR分解开始消失

- 从一个不可约上Hessenberg矩阵H_k出发，经过一次QR迭代，得到新的上Hessenberg矩阵H_{k+1}，

$$H_{k+1} = Q_k^T H_k Q_k$$

- 这里的QR迭代实际上就是连续$n-1$次平面旋转操作，不同于标准的QR分解
- 回忆：上Hessenberg矩阵分解的唯一性
- 知道了Q_k的第一列，就几乎完全确定了整个矩阵Q_k

加速收敛：原点位移

- 基本的QR方法是线性收敛的

加速收敛：原点位移

- 基本的QR方法是线性收敛的
- 为了加速收敛，我们引进原点位移

加速收敛：原点位移

- 基本的QR方法是线性收敛的
- 为了加速收敛，我们引进原点位移
- 带原点位移的QR迭代格式如下：

$$
\begin{aligned}
H_m - \mu_m I &= Q_m R_m \\
H_{m+1} &= R_m Q_m + \mu_m I
\end{aligned}
$$

其中$H_0 = H$为给定的上Hessenberg矩阵

位移的选取

- H_m为上Hessenberg矩阵，故其最后一行仅有两个非零元素$h_{n,n-1}^{(m)}$，$h_{nn}^{(m)}$

位移的选取

- H_m为上Hessenberg矩阵，故其最后一行仅有两个非零元素$h_{n,n-1}^{(m)}$，$h_{nn}^{(m)}$
- 若QR算法收敛，则当m充分大时$h_{n,n-1}^{(m)}$应很小，因而$h_{nn}^{(m)}$接近于H的一个特征值

位移的选取

- H_m为上Hessenberg矩阵，故其最后一行仅有两个非零元素$h_{n,n-1}^{(m)}$，$h_{nn}^{(m)}$
- 若QR算法收敛，则当m充分大时$h_{n,n-1}^{(m)}$应很小，因而$h_{nn}^{(m)}$接近于H的一个特征值
- 因此我们可以取$\mu_m = h_{nn}^{(m)}$

位移的选取

- H_m为上Hessenberg矩阵，故其最后一行仅有两个非零元素$h_{n,n-1}^{(m)}$，$h_{nn}^{(m)}$
- 若QR算法收敛，则当m充分大时$h_{n,n-1}^{(m)}$应很小，因而$h_{nn}^{(m)}$接近于H的一个特征值
- 因此我们可以取$\mu_m = h_{nn}^{(m)}$
- 可以证明：若$h_{n,n-1}^{(m)} = \varepsilon$很小，则一次带原点位移的QR迭代后，$h_{n,n-1}^{(m+1)} = O(\varepsilon^2)$, 即收敛速度从线性收敛加速到二次收敛

位移的选取

- H_m为上Hessenberg矩阵，故其最后一行仅有两个非零元素$h_{n,n-1}^{(m)}$，$h_{nn}^{(m)}$
- 若QR算法收敛，则当m充分大时$h_{n,n-1}^{(m)}$应很小，因而$h_{nn}^{(m)}$接近于H的一个特征值
- 因此我们可以取$\mu_m = h_{nn}^{(m)}$
- 可以证明：若$h_{n,n-1}^{(m)} = \varepsilon$很小，则一次带原点位移的QR迭代后，$h_{n,n-1}^{(m+1)} = O(\varepsilon^2)$, 即收敛速度从线性收敛加速到二次收敛
- 课本例6.4.2的程序演示

双重步位移

- 单步原点位移的QR迭代具有严重的缺点：
  若A具有复共轭特值，则实位移一般并不能起到加速的作用

双重步位移

- 单步原点位移的QR迭代具有严重的缺点：若A具有复共轭特值，则实位移一般并不能起到加速的作用
- 为此，我们引入双重步位移的QR迭代，基本想法是把两步带原点位移的QR迭代合并为一步，以避免复数运算

迭代格式

- 设$A \in \mathbb{R}^{n\times n}$, 考察如下的迭代格式:

$$\begin{aligned} H_1 &= Q_0^T A Q_0 && \text{上Hessenberg分解} \\ H_k - \mu_k I &= Q_k R_k && \text{QR分解} \\ H_{k+1} &= R_k Q_k + \mu_k I && k = 1, 2, \ldots \end{aligned}$$

迭代格式

- 设$A \in \mathbb{R}^{n\times n}$, 考察如下的迭代格式:

$$\begin{aligned} H_1 &= Q_0^T A Q_0 && \text{上Hessenberg分解} \\ H_k - \mu_k I &= Q_k R_k && \text{QR分解} \\ H_{k+1} &= R_k Q_k + \mu_k I && k = 1, 2, \ldots \end{aligned}$$

- 假设迭代格式中出现的上Hessenberg矩阵都是不可约的。否则，可分别对沿对角线的上下两个子矩阵进行QR迭代

复共轭特征值的处理

- 假设H_k的尾部2×2子矩阵

$$S_k = \begin{pmatrix} h_{mm}^{(k)} & h_{mn}^{(k)} \\ h_{nm}^{(k)} & h_{nn}^{(k)} \end{pmatrix}, \quad m = n - 1$$

有一对复共轭特征值γ_1和γ_2。这时我们不能期望$h_{nn}^{(k)}$最终收敛于A的某个特征值，而且此时取$\mu_k = h_{nn}^{(k)}$也完全没有加速效果

复共轭特征值的处理

- 假设H_k的尾部2×2子矩阵

$$S_k = \begin{pmatrix} h_{mm}^{(k)} & h_{mn}^{(k)} \\ h_{nm}^{(k)} & h_{nn}^{(k)} \end{pmatrix}, \quad m = n - 1$$

  有一对复共轭特征值γ_1和γ_2。这时我们不能期望$h_{nn}^{(k)}$最终收敛于A的某个特征值，而且此时取$\mu_k = h_{nn}^{(k)}$也完全没有加速效果

- 为了加速，我们应当取γ_1或γ_2作位移，但这样就会涉及到复数运算

连续两次位移

- 为了避免复数运算，我们计划用γ_1和γ_2连续作两次位移，即进行如下迭代：

$$
\begin{aligned}
H_k - \gamma_1 I &= U_1 R_1, & G_1 &= R_1 U_1 + \gamma_1 I \\
G_1 - \gamma_2 I &= U_2 R_2, & H_{k+1} &= R_2 U_2 + \gamma_2 I
\end{aligned}
$$

连续两次位移

- 为了避免复数运算，我们计划用γ_1和γ_2连续作两次位移，即进行如下迭代：

$$H_k - \gamma_1 I = U_1 R_1, \qquad G_1 = R_1 U_1 + \gamma_1 I$$
$$G_1 - \gamma_2 I = U_2 R_2, \qquad H_{k+1} = R_2 U_2 + \gamma_2 I$$

- 由此格式，记$M = (H_k - \gamma_1 I)(H_k - \gamma_2 I)$, $Q = U_1 U_2$, $R = R_2 R_1$, 则

$$M = QR, \quad H_{k+1} = Q^* H_k Q$$

验算

$$
\begin{aligned}
QR &= U_1U_2R_2R_1 = U_1(G_1-\gamma_2 I)R_1\\
&= U_1(R_1U_1+(\gamma_1-\gamma_2)I)R_1\\
&= U_1R_1(U_1R_1+(\gamma_1-\gamma_2)I)\\
&= (H_k-\gamma_1 I)(H_k-\gamma_2 I) = M\\
U_1G_1 &= U_1(R_1U_1+\gamma_1 I) = (U_1R_1+\gamma_1 I)U_1 = H_kU_1\\
U_2H_{k+1} &= G_1U_2
\end{aligned}
$$

实矩阵M

- 根据M的定义，我们有

$$M = H_k^2 - sH_k + tI$$

其中$s = \gamma_1 + \gamma_2 = h_{mm}^{(k)} + h_{nn}^{(k)} \in \mathbb{R}$,

$t = \gamma_1\gamma_2 = \det S_k \in \mathbb{R}$

实矩阵M

- 根据M的定义，我们有

$$M = H_k^2 - sH_k + tI$$

其中$s = \gamma_1 + \gamma_2 = h_{mm}^{(k)} + h_{nn}^{(k)} \in \mathbb{R}$,

$t = \gamma_1\gamma_2 = \det S_k \in \mathbb{R}$

- 所以M是实矩阵

实矩阵M

- 根据M的定义，我们有
$$M = H_k^2 - sH_k + tI$$
其中$s = \gamma_1 + \gamma_2 = h_{mm}^{(k)} + h_{nn}^{(k)} \in \mathbb{R}$,
$t = \gamma_1\gamma_2 = \det S_k \in \mathbb{R}$
- 所以M是实矩阵
- 根据QR分解的性质，即使γ_1和γ_2均不是H_k的特征值，并假定在计算过程中R_1和R_2的对角元均取为正数，那么Q也是实矩阵

实矩阵M

- 根据M的定义，我们有
$$M = H_k^2 - sH_k + tI$$
其中$s = \gamma_1 + \gamma_2 = h_{mm}^{(k)} + h_{nn}^{(k)} \in \mathbb{R}$,
$t = \gamma_1\gamma_2 = \det S_k \in \mathbb{R}$
- 所以M是实矩阵
- 根据QR分解的性质，即使γ_1和γ_2均不是H_k的特征值，并假定在计算过程中R_1和R_2 的对角元均取为正数，那么Q也是实矩阵
- 所以H_{k+1}也是实矩阵

初步结论

- 在没有误差的情况下，用γ_1和γ_2连续作两次位移进行QR迭代产生的H_{k+1}仍是实的上Hessenberg矩阵

初步结论

- 在没有误差的情况下，用γ_1和γ_2连续作两次位移进行QR迭代产生的H_{k+1}仍是实的上Hessenberg矩阵
- 但是，实际计算时，由于舍入误差的影响，如此得到的H_{k+1}一般并不是实矩阵

初步结论

- 在没有误差的情况下，用γ_1和γ_2连续作两次位移进行QR迭代产生的H_{k+1}仍是实的上Hessenberg矩阵
- 但是，实际计算时，由于舍入误差的影响，如此得到的H_{k+1}一般并不是实矩阵
- 为了确保得到的H_{k+1}仍是实矩阵，我们对迭代格式进行修改

基于M的迭代

- 修改后的迭代格式如下：

基于M的迭代

- 修改后的迭代格式如下：
  1. 计算$M = H_k^2 - sH_k + tI$

基于M的迭代

- 修改后的迭代格式如下：
  1. 计算$M = H_k^2 - sH_k + tI$
  2. 计算M的QR分解$M = QR$;

基于M的迭代

- 修改后的迭代格式如下：
  1. 计算$M = H_k^2 - sH_k + tI$
  2. 计算M的QR分解$M = QR$;
  3. 计算$H_{k+1} = Q^T H_k Q$

基于M的迭代

- 修改后的迭代格式如下：
  1. 计算$M = H_k^2 - sH_k + tI$
  2. 计算M的QR分解$M = QR$;
  3. 计算$H_{k+1} = Q^T H_k Q$
- M的下带宽是2，即次对角元和次次对角元非零

基于M的迭代

- 修改后的迭代格式如下：
  1. 计算$M = H_k^2 - sH_k + tI$
  2. 计算M的QR分解$M = QR$;
  3. 计算$H_{k+1} = Q^T H_k Q$
- M的下带宽是2，即次对角元和次次对角元非零
- 如此计算第一步形成M的运算量就是$O(n^3)$

降低运算量的想法

- 修改后的迭代格式运算量比较大

降低运算量的想法

- 修改后的迭代格式运算量比较大
- 想法：已有结论告诉我们，无论采用何种方法求正交矩阵$\tilde{Q}$使得$\tilde{Q}^T H_k \tilde{Q} = \tilde{H}_{k+1}$为上Hessenberg矩阵，只要保证$\tilde{Q}$的第一列与$Q$的第一列相同，则$\tilde{H}_{k+1}$就与$H_{k+1}$本质上是一样的

降低运算量的想法

- 修改后的迭代格式运算量比较大
- 想法：已有结论告诉我们，无论采用何种方法求正交矩阵$\tilde{Q}$使得$\tilde{Q}^T H_k \tilde{Q} = \tilde{H}_{k+1}$为上Hessenberg矩阵，只要保证$\tilde{Q}$的第一列与$Q$的第一列相同，则$\tilde{H}_{k+1}$就与$H_{k+1}$本质上是一样的
- 而这要求H_{k+1}是不可约的

H_k与H_{k+1}不可约性的关系

定理

若H_k是不可约的上Hessenberg矩阵，且γ_1和γ_2均非H_k的特征值，则H_{k+1} 也是不可约的上Hessenberg矩阵

定理证明

采用反证法

- 记$H_{k+1} = (\tilde{h}_{ij})$, 并假定存在$r$ $(1 \leqslant r \leqslant n-1)$使得$\tilde{h}_{r+1,r} = 0$, 而$\tilde{h}_{i+1,i} \neq 0$ $(i = 1, 2, \ldots, r-1)$

定理证明

采用反证法

- 记$H_{k+1} = (\tilde{h}_{ij})$, 并假定存在$r$ $(1 \leqslant r \leqslant n-1)$使得$\tilde{h}_{r+1,r} = 0$, 而$\tilde{h}_{i+1,i} \neq 0$ $(i = 1, 2, \ldots, r-1)$
- 比较等式$H_k Q = Q H_{k+1}$两边矩阵的前r列，我们有

$$H_k q_j = \tilde{h}_{1j} q_1 + \cdots + \tilde{h}_{jj} q_j + \tilde{h}_{j+1,j} q_{j+1},$$
$$j = 1, \ldots, r-1$$
$$H_k q_r = \tilde{h}_{1r} q_1 + \tilde{h}_{2r} q_2 + \cdots + \tilde{h}_{rr} q_r$$

- 由此考虑$r+1$个向量$q_1, H_k q_1, \ldots, H_k^r q_1$，它们均可以表示为$q_1, \ldots, q_r$的线性组合，因此存在不全为零的$\alpha_i$，

$$\left(\sum_{i=0}^{r} \alpha_i H_k^i\right) q_1 = 0$$

- 由此考虑$r+1$个向量$q_1, H_k q_1, \ldots, H_k^r q_1$, 它们均可以表示为$q_1, \ldots, q_r$的线性组合，因此存在不全为零的$\alpha_i$,

$$\left(\sum_{i=0}^{r} \alpha_i H_k^i \right) q_1 = 0$$

- 注意到q_r只出现在$H_k^{r-1} q_1$和$H_k^r q_1$中，因此必有$\alpha_r \neq 0$; 否则所有系数都是零

- 由$M = QR$可得$q_1 = \dfrac{1}{r_{11}} Me_1$. 将其代入到上式，并注意$M$也是$H_k$的多项式，我们有$My = 0$, 其中

$$y = \left(\sum_{i=0}^{r} \alpha_i H_k^i \right) e_1$$

- 由$M = QR$可得$q_1 = \dfrac{1}{r_{11}} M e_1$. 将其代入到上式，并注意$M$也是$H_k$的多项式，我们有$My = 0$, 其中
$$y = \left(\sum_{i=0}^{r} \alpha_i H_k^i \right) e_1$$
- 记$H_k = (h_{ij})$, 直接计算可知y的第$r+1$个分量为
$$\alpha_r h_{21} h_{32} \cdots h_{r+1,r} \neq 0$$
这就是说方程组$My = 0$有非零解，这与γ_1和γ_2均非H_k的特征值，从而M非奇异矛盾

实现从H_k到H_{k+1}的新方法

- 由于Q的第一列是M的第一列单位化得到的

实现从H_k到H_{k+1}的新方法

- 由于Q的第一列是M的第一列单位化得到的
- 由$M = H_k^2 - sH_k + tI$可知M的第一列为

$$Me_1 = (\xi_1, \xi_2, \xi_3, 0, \ldots, 0)^T$$

其中

$$\begin{aligned} \xi_1 &= (h_{11}^{(k)})^2 + h_{12}^{(k)} h_{21}^{(k)} - s h_{11}^{(k)} + t \\ \xi_2 &= h_{21}^{(k)} (h_{11}^{(k)} + h_{22}^{(k)} - s) \\ \xi_3 &= h_{21}^{(k)} h_{32}^{(k)} \end{aligned}$$

- 在M的QR分解中，如果Householder变换P_0把Me_1变为αe_1，那么P_0的第一列与Me_1共线

- 在M的QR分解中，如果Householder变换P_0把Me_1变为αe_1，那么P_0的第一列与Me_1共线
  - 利用Householder变换的定义与意义，可直接证明（作为练习）

- 在M的QR分解中，如果Householder变换P_0把Me_1变为αe_1，那么P_0的第一列与Me_1共线
    - 利用Householder变换的定义与意义，可直接证明（作为练习）
- 这就是说P_0的第一列就可以作为Q的第一列

- 在M的QR分解中，如果Householder变换P_0把Me_1变为αe_1，那么P_0的第一列与Me_1共线
    - 利用Householder变换的定义与意义，可直接证明（作为练习）
- 这就是说P_0的第一列就可以作为Q的第一列
- 根据Householder变换的理论，可以确定P_0的具体表示

P_0的表示

- P_0可以按下述方式确定：

$$P_0 = \mathrm{diag}(\tilde{P}_0, I_{n-3})$$

其中

$$\tilde{P}_0 = I_3 - \beta v v^T, \qquad v = (\xi_1 - \alpha, \xi_2, \xi_3)^T,$$

$$\alpha = \sqrt{\xi_1^2 + \xi_2^2 + \xi_3^2}, \quad \beta = 2/(v^T v)$$

- 现令$B = P_0 H_k P_0$, 则我们只要找到第一列为e_1的正交矩阵$\tilde{Q}$, 使得$\tilde{Q}^T B \tilde{Q} = \tilde{H}$为上Hessenberg矩阵，那么$\tilde{H}$就是所期望的$H_{k+1}$

- 现令$B = P_0 H_k P_0$, 则我们只要找到第一列为e_1的正交矩阵$\tilde{Q}$, 使得$\tilde{Q}^T B \tilde{Q} = \tilde{H}$为上Hessenberg矩阵，那么$\tilde{H}$就是所期望的$H_{k+1}$
- 根据前面给出的约化一个矩阵为上Hessenberg矩阵的方法以及上Hessenberg矩阵的唯一性可知，这是很容易做到的

- 现令$B = P_0 H_k P_0$, 则我们只要找到第一列为e_1的正交矩阵$\tilde{Q}$, 使得$\tilde{Q}^T B \tilde{Q} = \tilde{H}$为上Hessenberg矩阵，那么$\tilde{H}$就是所期望的$H_{k+1}$
- 根据前面给出的约化一个矩阵为上Hessenberg矩阵的方法以及上Hessenberg矩阵的唯一性可知，这是很容易做到的
- 而且算法的运算量只是$O(n^2)$

Francis算法

- 事实上，由于用P_0对H_k进行相似变换为B，只是改变了H的前三列和前三行，因此B比上Hessenberg矩阵只是多了三个可能的非零元素

Francis算法

- 事实上，由于用P_0对H_k进行相似变换为B，只是改变了H的前三列和前三行，因此B比上Hessenberg矩阵只是多了三个可能的非零元素
- 由此我们可以构造Householder变换P_1把第一列多余的两个非零元素消去

Francis算法

- 事实上，由于用P_0对H_k进行相似变换为B，只是改变了H的前三列和前三行，因此B比上Hessenberg矩阵只是多了三个可能的非零元素
- 由此我们可以构造Householder变换P_1把第一列多余的两个非零元素消去
- 逐步递推下去，就可以把B化为上Hessenberg矩阵

Francis算法

- 事实上，由于用P_0对H_k进行相似变换为B，只是改变了H的前三列和前三行，因此B比上Hessenberg矩阵只是多了三个可能的非零元素
- 由此我们可以构造Householder变换P_1把第一列多余的两个非零元素消去
- 逐步递推下去，就可以把B化为上Hessenberg矩阵
- 由此给出的就是著名的Francis双重步位移的QR迭代算法

实用算法

- 前面的讨论解决了用QR方法求给定实矩阵的实Schur标准形的几个关键问题

实用算法

- 前面的讨论解决了用QR方法求给定实矩阵的实Schur标准形的几个关键问题
- 为了得到一个实用的算法，我们还需要给出一个有效的判定准则，确定迭代过程中所产生的上Hessneberg矩阵的次对角元何时可以忽略不计

实用算法

- 前面的讨论解决了用QR方法求给定实矩阵的实Schur标准形的几个关键问题
- 为了得到一个实用的算法，我们还需要给出一个有效的判定准则，确定迭代过程中所产生的上Hessneberg矩阵的次对角元何时可以忽略不计
- 一种简单而实用的准则是：当
$$|h_{i+1,i}| \leqslant (|h_{ii}| + |h_{i+1,i+1}|)\mathbf{u}$$
时，就将$h_{i+1,i}$看做零

隐式QR算法

- 把前面所有分析综合在一起，就是隐式QR算法

隐式QR算法

- 把前面所有分析综合在一起，就是隐式QR算法
- 该算法计算给定的n阶实矩阵A的实Schur分解：$Q^TAQ = T$, 其中Q是正交矩阵，T为拟上三角阵，即对角块为1×1或2×2方阵的块上三角阵，而且每个2×2的对角块必有一对复共轭特征值

算法注解

- 实际计算的统计表明，这一算法每分离出一个1×1或2×2子矩阵平均需要2次QR迭代

算法注解

- 实际计算的统计表明，这一算法每分离出一个1×1或2×2子矩阵平均需要2次QR迭代
- 只计算特征值，运算量平均为$10n^3$；如果还需要Q，总运算量平均约为$25n^3$

算法注解

- 实际计算的统计表明，这一算法每分离出一个1×1或2×2子矩阵平均需要2次QR迭代
- 只计算特征值，运算量平均为$10n^3$；如果还需要Q，总运算量平均约为$25n^3$
- 算法是稳定的

对称特征值问题的计算方法

邓建松

2018 年 12 月 5 日

对称矩阵的特征值问题

- 对称矩阵的特征值问题具有许多良好的性质和十分丰富又完美的数学理论

对称矩阵的特征值问题

- 对称矩阵的特征值问题具有许多良好的性质和十分丰富又完美的数学理论
- 关于它的计算方法和相应的理论成为矩阵计算中发展得最为完善的一部分

对称矩阵的特征值问题

- 对称矩阵的特征值问题具有许多良好的性质和十分丰富又完美的数学理论
- 关于它的计算方法和相应的理论成为矩阵计算中发展得最为完善的一部分
- 本章介绍其中几个最基本的数值方法

对称矩阵特征值的性质

- 实对称矩阵的特征值均为实数

对称矩阵特征值的性质

- 实对称矩阵的特征值均为实数
- 其特征向量可以构成$\mathbb{R}^n$的一组标准正交基

对称矩阵特征值的性质

- 实对称矩阵的特征值均为实数
- 其特征向量可以构成$\mathbb{R}^n$的一组标准正交基

定理

若$A \in \mathbb{R}^{n\times n}$是对称的，则存在正交矩阵$Q \in \mathbb{R}^{n\times n}$使得

$$Q^T AQ = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$$

极小极大定理

定理

设$A \in \mathbb{R}^{n\times n}$是对称的，并假定$A$的特征值为$\lambda_1 \geqslant \cdots \geqslant \lambda_n$，则有

$$\lambda_i = \max_{\mathcal{X}\in\mathcal{G}_i^n} \min_{0\neq u\in\mathcal{X}} \frac{u^TAu}{u^Tu} = \min_{\mathcal{X}\in\mathcal{G}_{n-i+1}^n} \max_{0\neq u\in\mathcal{X}} \frac{u^TAu}{u^Tu}$$

其中$\mathcal{G}_k^n$表示$\mathbb{R}^n$中所有k维子空间的全体

特征值的敏感性：Weyl定理

定理

设n阶对称矩阵A和B的特征值分别为

$$\lambda_1 \geqslant \cdots \geqslant \lambda_n \text{和} \mu_1 \geqslant \cdots \geqslant \mu_n$$

则有

$$|\lambda_i - \mu_i| \leqslant \|A - B\|_2, \quad i = 1, \ldots, n$$

特征向量的敏感性

定理

设A和$A+E$是两个n阶实对称矩阵，并假设q_1是A的一个单位特征向量，$Q=(q_1,Q_2)$是n 阶正交矩阵，矩阵分块如下：

$$Q^TAQ=\begin{pmatrix}\lambda & 0\\ 0 & D_2\end{pmatrix},\quad Q^TEQ=\begin{pmatrix}\varepsilon & e^T\\ e & E_{22}\end{pmatrix}$$

若$d=\min_{\mu\in\lambda(D_2)}|\lambda-\mu|>0$, $\|E\|_2\leqslant\frac{1}{4}d$, 则存在$A+E$的一个单位特征向量$\tilde{q}_1$使得

$$\sin\theta=\sqrt{1-|q_1^T\tilde{q}_1|^2}\leqslant\frac{4}{d}\|e\|_2\leqslant\frac{4}{d}\|E\|_2$$

其中θ是向量q_1和$\tilde{q}_1$之间所夹的锐角，即$\theta=\arccos|q_1^T\tilde{q}_1|$

SVD分解定理

- 设$A \in \mathbb{R}^{m\times n}$, 则存在正交矩阵$U \in \mathbb{R}^{m\times m}$和$V \in \mathbb{R}^{n\times n}$使得

$$U^T AV = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix}$$

其中$\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, $\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0$

SVD分解定理

- 设$A \in \mathbb{R}^{m\times n}$, 则存在正交矩阵$U \in \mathbb{R}^{m\times m}$和$V \in \mathbb{R}^{n\times n}$使得
$$U^TAV = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix}$$
其中$\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, $\sigma_1 \geqslant \cdots \geqslant \sigma_r > 0$
- 对于上述分解，我们称
$$\sigma_1 \geqslant \cdots \geqslant \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$$
为A的奇异值；U和V的列向量分别称为A的左/右奇异向量

奇异值的稳定性

定理

设$A, B \in \mathbb{R}^{m\times n}$, 并假定它们的奇异值分别为

$$\sigma_1 \geqslant \cdots \geqslant \sigma_n \text{和} \tau_1 \geqslant \cdots \geqslant \tau_n$$

则有

$$|\sigma_i - \tau_i| \leqslant \|A - B\|_2, \quad i = 1, \ldots, n$$

对称QR方法

- 对称QR方法就是求解对称特征值问题的QR方法

对称QR方法

- 对称QR方法就是求解对称特征值问题的QR方法
- 它是将QR方法应用于对称矩阵，并且充分利用了其对称性

对称QR方法

- 对称QR方法就是求解对称特征值问题的QR方法
- 它是将QR方法应用于对称矩阵，并且充分利用了其对称性
- 此时上Hessenberg矩阵就是三对角对称矩阵

三对角化

- 若A是n阶实对称矩阵，并假定A的上Hessenberg分解为

$$Q^T A Q = T$$

其中Q是正交矩阵，T是上Hessenberg矩阵，则可知T是对称三对角阵

三对角化

- 若A是n阶实对称矩阵，并假定A的上Hessenberg分解为
$$Q^TAQ = T$$
其中Q是正交矩阵，T是上Hessenberg矩阵，则可知T是对称三对角阵
- 因此当处理的是对称矩阵时，上Hessenberg化就是三对角化

三对角化

- 若A是n阶实对称矩阵，并假定A的上Hessenberg分解为

$$Q^T AQ = T$$

其中Q是正交矩阵，T是上Hessenberg矩阵，则可知T是对称三对角阵

- 因此当处理的是对称矩阵时，上Hessenberg化就是三对角化
- 因此我们约化过程中可以充分利用对称性，使运算量大减

具体分析

- 对A进行分块：

$$A = \begin{pmatrix} \alpha_1 & v_0^T \\ v_0 & A_0 \end{pmatrix}$$

其中$\alpha_1 \in \mathbb{R}$.

具体分析

- 对A进行分块：

$$A = \begin{pmatrix} \alpha_1 & v_0^T \\ v_0 & A_0 \end{pmatrix}$$

其中$\alpha_1 \in \mathbb{R}$.

- 利用Householder变换把A约化为上Hessenberg矩阵的第一步是把v_0转化为$\beta_0 e_1$，从而A_0变为新的$n-1$阶矩阵，我们对A_0进行类似分块，得到A_1，依次类推……

约化的第k步

- 计算Householder变换$\tilde{H}_k \in \mathbb{R}^{(n-k)\times(n-k)}$，使得

$$\tilde{H}_k v_{k-1} = \beta_k e_1, \quad \beta_k \in \mathbb{R}$$

约化的第k步

- 计算Householder变换$\tilde{H}_k \in \mathbb{R}^{(n-k)\times(n-k)}$，使得

$$\tilde{H}_k v_{k-1} = \beta_k e_1, \quad \beta_k \in \mathbb{R}$$

- 计算

$$\tilde{H}_k A_{k-1} \tilde{H}_k = \begin{pmatrix} \alpha_{k+1} & v_k^T \\ v_k & A_k \end{pmatrix}$$

分解结果

- 定义

$$T = \begin{pmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \beta_{n-1} & \alpha_n \end{pmatrix}$$

$$Q = H_1 \cdots H_{n-2}, \quad H_k = \operatorname{diag}(I_k, \tilde{H}_k)$$

分解结果

- 定义

$$T = \begin{pmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \beta_{n-1} & \alpha_n \end{pmatrix}$$

$$Q = H_1 \cdots H_{n-2}, \quad H_k = \mathrm{diag}(I_k, \tilde{H}_k)$$

- 则我们有$Q^TAQ = T$, 这称为A的三对角分解

运算量

- 第k步约化的主要工作是计算$\tilde{H}_k A_{k-1} \tilde{H}_k$

运算量

- 第k步约化的主要工作是计算$\tilde{H}_k A_{k-1} \tilde{H}_k$
- 设$\tilde{H}_k = I - \beta v v^T$, $v \in \mathbb{R}^{n-k}$, 由A_{k-1}的对称性，我们有

$$\tilde{H}_k A_{k-1} \tilde{H}_k = A_{k-1} - v w^T - w v^T$$

其中$w = u - \beta(v^T u)v/2$, $u = \beta A_{k-1} v$

运算量

- 第k步约化的主要工作是计算$\tilde{H}_k A_{k-1} \tilde{H}_k$
- 设$\tilde{H}_k = I - \beta v v^T$, $v \in \mathbb{R}^{n-k}$, 由A_{k-1}的对称性，我们有

$$\tilde{H}_k A_{k-1} \tilde{H}_k = A_{k-1} - v w^T - w v^T$$

其中$w = u - \beta(v^T u)v/2$, $u = \beta A_{k-1} v$

- 利用这一等式计算，运算量仅为$4(n-k)^2$, 从而总体运算量为$4n^3/3$次乘法运算

带原点位移的QR迭代

- 完成了三对角分解后，接下来就是选取适当的位移进行QR迭代

带原点位移的QR迭代

- 完成了三对角分解后，接下来就是选取适当的位移进行QR迭代
- 由于此时特征值全是实数，因此没有必要进行双重步位移

带原点位移的QR迭代

- 完成了三对角分解后，接下来就是选取适当的位移进行QR迭代
- 由于此时特征值全是实数，因此没有必要进行双重步位移
- 带原点位移的QR迭代格式为

$$
\begin{aligned}
T_k - \mu_k I &= Q_k R_k \quad (\text{QR分解}) \\
T_{k+1} &= R_k Q_k + \mu_k I, \quad k = 0, 1, \ldots
\end{aligned}
$$

其中$T_0 = T$是对称三对角阵

性质保持

- 根据QR迭代保持上Hessenberg形和对称性的特点可知T_k都是对称三对角阵

性质保持

- 根据QR迭代保持上Hessenberg形和对称性的特点可知T_k都是对称三对角阵
- 与非对称QR方法一样，我们这里也假定迭代中所出现的T_k都是不可约的

位移的选取

- 与非对称QR迭代中一样，最简单的做法是取μ_k为T_k右下角元素

位移的选取

- 与非对称QR迭代中一样，最简单的做法是取μ_k为T_k右下角元素
- 但此时有一个更好的取法，即Wilkinson位移

位移的选取

- 与非对称QR迭代中一样，最简单的做法是取μ_k为T_k右下角元素
- 但此时有一个更好的取法，即Wilkinson位移
- 设T_k右下角的2×2阶矩阵为$\begin{pmatrix} \alpha_{n-1} & \beta_{n-1} \\ \beta_{n-1} & \alpha_n \end{pmatrix}$，我们取$\mu_k$为该矩阵的两个特征值之中靠近$\alpha_n$的那一个，即

$$\mu_k = \alpha_n + \delta - \mathrm{sgn}(\delta)\sqrt{\delta^2 + \beta_{n-1}^2}$$

其中$\delta = (\alpha_{n-1} - \alpha_n)/2$

两种选取的对比

- 两种位移最终都是三次收敛的

两种选取的对比

- 两种位移最终都是三次收敛的
- Wilkinson论证了为什么后者优于前者的理由

一次对称QR迭代的实现

- 假设一次对称QR迭代的形式为

$$T - \mu I = QR, \quad \tilde{T} = RQ + \mu I$$

一次对称QR迭代的实现

- 假设一次对称QR迭代的形式为

$$T - \mu I = QR, \quad \tilde{T} = RQ + \mu I$$

- 当然可以利用Givens变换直接实现$T - \mu I$的QR分解，进而完成一步迭代

一次对称QR迭代的实现

- 假设一次对称QR迭代的形式为

$$T-\mu I=QR,\quad \tilde{T}=RQ+\mu I$$

- 当然可以利用Givens变换直接实现$T-\mu I$的QR分解，进而完成一步迭代
- 更漂亮的做法是以隐含的方式实现由T到$\tilde{T}$的变换

基本想法

- 根据上Hessenberg约化的唯一性定理，对于$\tilde{T} = Q^T TQ$，$\tilde{T}$本质上是由Q的第一列完全确定的

基本想法

- 根据上Hessenberg约化的唯一性定理，对于$\tilde{T} = Q^T T Q$，$\tilde{T}$本质上是由Q的第一列完全确定的
- 利用Givens变换对$T - \mu I$进行QR分解，那么$Qe_1 = G_1 e_1$，这里$G_1 = G(1, 2, \theta)$ 使得

$$G(1,2,\theta)\begin{pmatrix} \alpha_1 - \mu \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} \alpha_1 - \mu \\ \beta_1 \end{pmatrix}$$

$$= \begin{pmatrix} * \\ 0 \end{pmatrix}$$

后续约化

- 令$B=G_1TG_1^T$，则B的左上3×3阶矩阵非零，即仅比对称三对角阵多两个非零元

后续约化

- 令$B = G_1 T G_1^T$，则B的左上3×3阶矩阵非零，即仅比对称三对角阵多两个非零元
- 将B再用$n-1$个Givens变换约化为三对角阵$\tilde{T}$

后续约化

- 令$B = G_1 T G_1^T$, 则B的左上3×3阶矩阵非零，即仅比对称三对角阵多两个非零元
- 将B再用$n-1$个Givens变换约化为三对角阵$\tilde{T}$
- 如此即为带Wilkinson位移的隐式对称QR迭代算法

隐式对称QR算法

- 类比于非对称QR算法，综合前面的讨论，我们可以得到隐式对称QR算法

隐式对称QR算法

- 类比于非对称QR算法，综合前面的讨论，我们可以得到隐式对称QR算法
- 此时算法的输出为对角阵，元素是A的特征值

隐式对称QR算法

- 类比于非对称QR算法，综合前面的讨论，我们可以得到隐式对称QR算法
- 此时算法的输出为对角阵，元素是A的特征值
- 只计算特征值，算法运算量约为$4n^3/3$

隐式对称QR算法

- 类比于非对称QR算法，综合前面的讨论，我们可以得到隐式对称QR算法
- 此时算法的输出为对角阵，元素是A的特征值
- 只计算特征值，算法运算量约为$4n^3/3$
- 这是矩阵计算中最漂亮的算法之一，它是数值稳定的

Jacobi方法

- Jacobi方法是求实对称矩阵全部特征值和特征向量的最古老的方法之一

Jacobi方法

- Jacobi方法是求实对称矩阵全部特征值和特征向量的最古老的方法之一
- 它是由C.G.J. Jacobi于1846年首先提出的

Jacobi方法

- Jacobi方法是求实对称矩阵全部特征值和特征向量的最古老的方法之一
- 它是由C.G.J. Jacobi于1846年首先提出的
- 该方法利用了实对称矩阵可以正交相似变换约化为对角阵的性质，用一系列适当选取的平面旋转变换将给定矩阵约化为对角阵

Jacobi方法

- Jacobi方法是求实对称矩阵全部特征值和特征向量的最古老的方法之一
- 它是由C.G.J. Jacobi于1846年首先提出的
- 该方法利用了实对称矩阵可以正交相似变换约化为对角阵的性质，用一系列适当选取的平面旋转变换将给定矩阵约化为对角阵
- 它的速度相比对称QR方法要相差很远，但它编程简单，并行效率高

Jacobi方法的目标

- 设$A = (\alpha_{ij})$是$n \times n$实对称矩阵

Jacobi方法的目标

- 设$A = (\alpha_{ij})$是$n \times n$实对称矩阵
- Jacobi方法的目标就是将A的非对角“范数”$E(A)$逐步约化为零：

$$E(A) = \left(\|A\|_F^2 - \sum_{i=1}^{n} \alpha_{ii}^2 \right)^{1/2} = \left(\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \alpha_{ij}^2 \right)^{1/2}$$

Jacobi方法的目标

- 设$A = (\alpha_{ij})$是$n \times n$实对称矩阵
- Jacobi方法的目标就是将A的非对角“范数” $E(A)$逐步约化为零：

$$E(A) = \left(\|A\|_F^2 - \sum_{i=1}^{n} \alpha_{ii}^2 \right)^{1/2} = \left(\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \alpha_{ij}^2 \right)^{1/2}$$

- 所采用的基本工具就是由Givens变换定义的平面旋转变换

Jacobi变换

- Given变换$G(p,q,\theta)$也记作$J(p,q,\theta)$

$$\begin{aligned} J(p,q,\theta) = I &+ (\cos\theta - 1)(e_p e_p^T + e_q e_q^T) \\ &+ \sin\theta(e_p e_q^T - e_q e_p^T) \end{aligned}$$

Jacobi变换

- Given变换$G(p, q, \theta)$也记作$J(p, q, \theta)$

$$
\begin{aligned}
J(p, q, \theta) = I &+ (\cos\theta - 1)(e_p e_p^T + e_q e_q^T) \\
&+ \sin\theta(e_p e_q^T - e_q e_p^T)
\end{aligned}
$$

- 此时假定$p < q$

Jacobi变换

- Given变换$G(p,q,\theta)$也记作$J(p,q,\theta)$

$$\begin{aligned} J(p,q,\theta) = I + (\cos\theta - 1)(e_p e_p^T + e_q e_q^T) \\ + \sin\theta(e_p e_q^T - e_q e_p^T) \end{aligned}$$

- 此时假定$p < q$
- 这一变换也称为(p,q)平面的Jacobi变换

Jacobi方法一次约化的步骤

1. 选择旋转平面(p, q), $1 \leqslant p < q \leqslant n$

Jacobi方法一次约化的步骤

1. 选择旋转平面(p, q), $1 \leqslant p < q \leqslant n$
2. 确定旋转角θ使得

$$\begin{pmatrix} \beta_{pp} & \beta_{pq} \\ \beta_{qp} & \beta_{qq} \end{pmatrix} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}^T \begin{pmatrix} \alpha_{pp} & \alpha_{pq} \\ \alpha_{qp} & \alpha_{qq} \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$

为对角阵，其中$c = \cos\theta$, $s = \sin\theta$

Jacobi方法一次约化的步骤

1. 选择旋转平面(p, q), $1 \leqslant p < q \leqslant n$
2. 确定旋转角θ使得

$$\begin{pmatrix} \beta_{pp} & \beta_{pq} \\ \beta_{qp} & \beta_{qq} \end{pmatrix} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}^T \begin{pmatrix} \alpha_{pp} & \alpha_{pq} \\ \alpha_{qp} & \alpha_{qq} \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$

   为对角阵，其中$c = \cos\theta$, $s = \sin\theta$
3. 对A进行相似变换：$B = (\beta_{ij}) = J^T AJ$, 其中$J = J(p, q, \theta)$

A与B的关系

- 矩阵A与B只在第p行/列和第q行/列不同，它们之间有关系如下：

$$\beta_{ip} = \beta_{pi} = c\alpha_{ip} - s\alpha_{iq}, i \neq p, q$$

$$\beta_{iq} = \beta_{qi} = s\alpha_{ip} + c\alpha_{iq}, i \neq p, q$$

$$\beta_{pp} = c^2\alpha_{pp} - 2sc\alpha_{pq} + s^2\alpha_{qq}$$

$$\beta_{qq} = s^2\alpha_{pp} + 2sc\alpha_{pq} + c^2\alpha_{qq}$$

$$\beta_{pq} = \beta_{qp} = (c^2 - s^2)\alpha_{pq} + sc(\alpha_{pp} - \alpha_{qq})$$

c, s的计算

- 先假设选定了旋转平面(p, q)

c, s的计算

- 先假设选定了旋转平面(p, q)
- 下面根据$\beta_{pq} = \beta_{qp} = 0$求出$c, s$

c, s的计算

- 先假设选定了旋转平面(p, q)
- 下面根据$\beta_{pq} = \beta_{qp} = 0$求出$c, s$
- 由β_{pq}的表达式可知，这等价于计算c, s使得

$$(c^2 - s^2)\alpha_{pq} + sc(\alpha_{pp} - \alpha_{qq}) = 0$$

c, s的计算

- 先假设选定了旋转平面(p, q)
- 下面根据$\beta_{pq} = \beta_{qp} = 0$求出$c, s$
- 由β_{pq}的表达式可知，这等价于计算c, s使得

$$(c^2 - s^2)\alpha_{pq} + sc(\alpha_{pp} - \alpha_{qq}) = 0$$

- 若$\alpha_{pq} = 0$, 可取$c = 1, s = 0$

$\alpha_{pq} \neq 0$的情形

- 如果$\alpha_{pq} \neq 0$，令

$$\tau = \frac{\alpha_{qq} - \alpha_{pp}}{2\alpha_{pq}}, t = \tan\theta = \frac{s}{c}$$

$\alpha_{pq} \neq 0$的情形

- 如果$\alpha_{pq} \neq 0$，令

$$\tau = \frac{\alpha_{qq} - \alpha_{pp}}{2\alpha_{pq}}, t = \tan\theta = \frac{s}{c}$$

- 由此得到方程$t^2 + 2\tau t - 1 = 0$

$\alpha_{pq} \neq 0$的情形

- 如果$\alpha_{pq} \neq 0$，令

$$\tau = \frac{\alpha_{qq} - \alpha_{pp}}{2\alpha_{pq}}, t = \tan\theta = \frac{s}{c}$$

- 由此得到方程$t^2 + 2\tau t - 1 = 0$
- 如此t有两种选择。选择其绝对值较小的即可，从而

$$t = \frac{\mathrm{sgn}(\tau)}{|\tau| + \sqrt{1+\tau^2}}$$

- 如此选择保证了旋转角θ满足$|\theta| \leqslant \pi/4$，这对Jacobi方法的收敛性是至关重要的

- 如此选择保证了旋转角θ满足$|\theta| \leqslant \pi/4$, 这对Jacobi方法的收敛性是至关重要的
- 细节见后面的收敛性分析

- 如此选择保证了旋转角θ满足$|\theta| \leqslant \pi/4$, 这对Jacobi方法的收敛性是至关重要的
- 细节见后面的收敛性分析
- 确定了t之后，c, s可由下面的公式确定：

$$c = \frac{1}{\sqrt{1+t^2}}, \quad s = tc$$

Frobenius范数的变化

- 由于Frobenius范数对正交变换保持不变，从而$\|B\|_F = \|A\|_F$，即

$$\alpha_{pp}^2 + \alpha_{qq}^2 + 2\alpha_{pq}^2 = \beta_{pp}^2 + \beta_{qq}^2$$

Frobenius范数的变化

- 由于Frobenius范数对正交变换保持不变，从而$\|B\|_F = \|A\|_F$，即
$$\alpha_{pp}^2 + \alpha_{qq}^2 + 2\alpha_{pq}^2 = \beta_{pp}^2 + \beta_{qq}^2$$
- 如此我们有
$$\begin{aligned}
E(B)^2 &= \|B\|_F^2 - \sum_{i=1}^{n} \beta_{ii}^2 \\
&= \|A\|_F^2 - \sum_{i=1}^{n} \alpha_{ii}^2 + (\alpha_{pp}^2 + \alpha_{qq}^2 - \beta_{pp}^2 - \beta_{qq}^2) \\
&= E(A)^2 - 2\alpha_{pq}^2
\end{aligned}$$

旋转平面的选取

- 我们的目标是使$E(B)$尽可能小，因此(p,q)的最佳选择应使

$$|\alpha_{pq}| = \max_{1\leqslant i<j\leqslant n} |\alpha_{ij}|$$

即应选取非对角元中绝对值最大者所在的行列为旋转平面

经典Jacobi方法

- 根据上述原则选择旋转平面(p, q), 然后再确定c, s的方法就是经典Jacobi方法

经典Jacobi方法

- 根据上述原则选择旋转平面(p, q), 然后再确定c, s的方法就是经典Jacobi方法
- 其基本迭代格式如下：

$$A_k = (\alpha_{ij}^{(k)}) = J_k^T A_{k-1} J_k, \quad k = 1, 2, \ldots$$

其中$A_0 = A$, J_k是对A_{k-1}如前所确定的Jacobi变换

收敛定理

定理

存在A的特征值的一个排列$\lambda_1, \lambda_2, \ldots, \lambda_n$，使得

$$\lim_{k\to\infty} A_k = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

定理证明:$E(A_k) \to 0$

- 我们先证明随着迭代次数k的增加，A_k的非对角“范数”$E(A_k) \to 0$

定理证明:$E(A_k) \to 0$

- 我们先证明随着迭代次数k的增加，A_k的非对角“范数” $E(A_k) \to 0$
- 根据前面的讨论，我们有

$$E(A_k)^2 = E(A_{k-1})^2 - 2(\alpha_{pq}^{(k-1)})^2$$

其中$\alpha_{pq}^{(k-1)}$是A_{k-1}的非对角元之中绝对值最大者

定理证明:$E(A_k) \to 0$

- 注意到

$$E(A_{k-1})^2 \leqslant (n^2 - n)(\alpha_{pq}^{(k-1)})^2$$

定理证明:$E(A_k) \to 0$

- 注意到

$$E(A_{k-1})^2 \leqslant (n^2 - n)(\alpha_{pq}^{(k-1)})^2$$

- 从而有

$$E(A_k)^2 \leqslant \left(1 - \frac{2}{n^2 - n}\right) E(A_{k-1})^2$$

定理证明:$E(A_k) \to 0$

- 注意到

$$E(A_{k-1})^2 \leqslant (n^2 - n)(\alpha_{pq}^{(k-1)})^2$$

- 从而有

$$E(A_k)^2 \leqslant \left(1 - \frac{2}{n^2 - n}\right) E(A_{k-1})^2$$

- 系数为与k无关的绝对值小于1的数，由此$\lim_{k\to\infty} E(A_k) = 0$

定理的继续证明

- 下面证明$\alpha_{ii}^{(k)} \to \lambda_i$, $k \to \infty$

定理的继续证明

- 下面证明$\alpha_{ii}^{(k)} \to \lambda_i$, $k \to \infty$
- 假设A的互不相同的特征值之间最小距离为δ

定理的继续证明

- 下面证明$\alpha_{ii}^{(k)} \to \lambda_i$, $k \to \infty$
- 假设A的互不相同的特征值之间最小距离为δ
- 任取ε满足$0 < \varepsilon < \delta/4$，由$\lim\limits_{k\to\infty} E(A_k) = 0$ 知存在k_0使得当$k \geqslant k_0$时有$E(A_k) < \varepsilon < \delta/4$

- 由于$\lambda(A_{k_0}) = \lambda(A)$, 对矩阵$A_{k_0}$与其对角元作成的对角阵

$$D_{k_0} = \mathrm{diag}(\alpha_{11}^{(k_0)}, \alpha_{22}^{(k_0)}, \ldots, \alpha_{nn}^{(k_0)})$$

应用本章开始给出的Weyl定理可知，存在A的特征值的一个排列$\lambda_1, \ldots, \lambda_n$使得对$i = 1, \ldots, n$

$$\left|\lambda_i - \alpha_{ii}^{(k_0)}\right| \leqslant \|A_{k_0} - D_{k_0}\|_2 \leqslant E(A_{k_0}) < \varepsilon < \delta/4$$

- 由于$\lambda(A_{k_0}) = \lambda(A)$, 对矩阵$A_{k_0}$与其对角元作成的对角阵

$$D_{k_0} = \mathrm{diag}(\alpha_{11}^{(k_0)}, \alpha_{22}^{(k_0)}, \ldots, \alpha_{nn}^{(k_0)})$$

应用本章开始给出的Weyl定理可知，存在A的特征值的一个排列$\lambda_1, \ldots, \lambda_n$使得对$i = 1, \ldots, n$

$$\left|\lambda_i - \alpha_{ii}^{(k_0)}\right| \leqslant \|A_{k_0} - D_{k_0}\|_2 \leqslant E(A_{k_0}) < \varepsilon < \delta/4$$

- 目前来说，特征值的排列是与k_0有关的

- 实际上，可以证明这种排列对任意$k \geqslant k_0$都是一致的（暂缺）

- 实际上，可以证明这种排列对任意$k \geqslant k_0$都是一致的（暂缺）
- 也说是说，只要能证明上式蕴涵着

$$\left|\lambda_i - \alpha_{ii}^{(k_0+1)}\right| < \varepsilon, i = 1, \ldots, n$$

则由归纳法可知对一切$k \geqslant k_0$有

$$\left|\lambda_i - \alpha_{ii}^{(k)}\right| < \varepsilon, i = 1, \ldots, n$$

从而定理得证

欠缺的一环

- 由于A_{k_0+1}与A_{k_0}的对角元只可能有两个不同：$\alpha_{pp}^{(k_0+1)}$和$\alpha_{qq}^{(k_0+1)}$，所以只需要对$i = p, q$证明欠缺的蕴涵关系成立即可

欠缺的一环

- 由于A_{k_0+1}与A_{k_0}的对角元只可能有两个不同：$\alpha_{pp}^{(k_0+1)}$和$\alpha_{qq}^{(k_0+1)}$，所以只需要对$i = p, q$证明欠缺的蕴涵关系成立即可
- 由于$t = s/c$, 根据c, s计算过程，$\alpha_{pq}^{(k_0)}(c^2 - s^2) + (\alpha_{pp}^{(k_0)} - \alpha_{qq}^{(k_0)})cs = 0$, 所以$(1 - t^2)\alpha_{pq}^{(k_0)} = t(\alpha_{qq}^{(k_0)} - \alpha_{pp}^{(k_0)})$

- 根据迭代格式，我们有

$$
\begin{aligned}
\alpha_{pp}^{(k_0+1)} &= \alpha_{pp}^{(k_0)} + c^2\left(-2t\alpha_{pq}^{(k_0)} + t^2(\alpha_{qq}^{(k_0)} - \alpha_{pp}^{(k_0)})\right) \\
&= \alpha_{pp}^{(k_0)} + c^2\left(-2t\alpha_{pq}^{(k_0)} + t(1-t^2)\alpha_{pq}^{(k_0)}\right) \\
&= \alpha_{pp}^{(k_0)} - t\alpha_{pq}^{(k_0)} \\
\alpha_{qq}^{(k_0+1)} &= \alpha_{qq}^{(k_0)} + t\alpha_{pq}^{(k_0)}
\end{aligned}
$$

- 根据迭代格式，我们有

$$\begin{aligned}\alpha_{pp}^{(k_0+1)} &= \alpha_{pp}^{(k_0)} + c^2\left(-2t\alpha_{pq}^{(k_0)} + t^2(\alpha_{qq}^{(k_0)} - \alpha_{pp}^{(k_0)})\right)\\ &= \alpha_{pp}^{(k_0)} + c^2\left(-2t\alpha_{pq}^{(k_0)} + t(1-t^2)\alpha_{pq}^{(k_0)}\right)\\ &= \alpha_{pp}^{(k_0)} - t\alpha_{pq}^{(k_0)}\\ \alpha_{qq}^{(k_0+1)} &= \alpha_{qq}^{(k_0)} + t\alpha_{pq}^{(k_0)}\end{aligned}$$

- 从而对任何$\lambda_j \neq \lambda_p$有(注意$|t| \leqslant 1$)

$$\begin{aligned}\left|\alpha_{pp}^{(k_0+1)} - \lambda_j\right| &= \left|\alpha_{pp}^{(k_0)} - \lambda_p + \lambda_p - \lambda_j - t\alpha_{pq}^{(k_0)}\right|\\ &\geqslant |\lambda_p - \lambda_j| - \left|\alpha_{pp}^{(k_0)} - \lambda_p\right| - |t|E(A_{k_0})\\ &\geqslant \delta - \varepsilon - \varepsilon \geqslant 2\varepsilon\end{aligned}$$

- 由于$\lambda(A_{k_0+1}) = \lambda(A)$, $E(A_{k_0+1}) < \varepsilon$, 所以根据Weyl定理可知，$\alpha_{pp}^{(k_0+1)}$必与$A$的某个特征值之间的距离小于$\varepsilon$

- 由于$\lambda(A_{k_0+1}) = \lambda(A)$, $E(A_{k_0+1}) < \varepsilon$, 所以根据Weyl定理可知，$\alpha_{pp}^{(k_0+1)}$必与$A$的某个特征值之间的距离小于$\varepsilon$
- 据前结论，$\left|\lambda_i - \alpha_{ii}^{(k)}\right| < \varepsilon$在$i = p$时成立

- 由于$\lambda(A_{k_0+1}) = \lambda(A)$, $E(A_{k_0+1}) < \varepsilon$, 所以根据Weyl定理可知，$\alpha_{pp}^{(k_0+1)}$必与$A$的某个特征值之间的距离小于$\varepsilon$
- 据前结论，$\left|\lambda_i - \alpha_{ii}^{(k)}\right| < \varepsilon$在$i = p$时成立
- 类似可证该不等式$i = q$时成立

- 由于$\lambda(A_{k_0+1}) = \lambda(A)$, $E(A_{k_0+1}) < \varepsilon$, 所以根据Weyl定理可知，$\alpha_{pp}^{(k_0+1)}$必与$A$的某个特征值之间的距离小于$\varepsilon$
- 据前结论，$\left|\lambda_i - \alpha_{ii}^{(k)}\right| < \varepsilon$在$i = p$时成立
- 类似可证该不等式$i = q$时成立
- 定理证明完成

注解

- 从定理的证明可见，$|t| \leqslant 1$对经典Jacobi方法的收敛起了至关重要的作用。它保证了迭代产生的每一个对角元一致地趋向于A的某一个固定的特征值

注解

- 从定理的证明可见，$|t| \leqslant 1$对经典Jacobi方法的收敛起了至关重要的作用。它保证了迭代产生的每一个对角元一致地趋向于A的某一个固定的特征值
- 证明也给出了经典Jacobi方法的收敛速度的一个粗略的估计：

$$E(A_k)^2 \leqslant \left(1 - \frac{2}{n^2 - n}\right)^k E(A_0)^2$$

注解

- 从定理的证明可见，$|t| \leqslant 1$对经典Jacobi方法的收敛起了至关重要的作用。它保证了迭代产生的每一个对角元一致地趋向于A的某一个固定的特征值
- 证明也给出了经典Jacobi方法的收敛速度的一个粗略的估计：

$$E(A_k)^2 \leqslant \left(1 - \frac{2}{n^2 - n}\right)^k E(A_0)^2$$

- 这表明经典Jacobi方法是线性收敛的

扫描

- 通常将$N=(n^2-n)/2$次Jacobi迭代称为一次扫描

扫描

- 通常将$N=(n^2-n)/2$次Jacobi迭代称为一次扫描
- 可以证明经典Jacobi方法的渐近收敛速度是二次的，即存在常数$c>0$，对充分大的k

$$E(A_{k+N}) \leqslant cE(A_k)^2$$

扫描

- 通常将$N = (n^2 - n)/2$次Jacobi迭代称为一次扫描
- 可以证明经典Jacobi方法的渐近收敛速度是二次的，即存在常数$c > 0$, 对充分大的k

$$E(A_{k+N}) \leqslant cE(A_k)^2$$

- 所以每扫描一次，其非对角“范数”将以平方收敛的速度接近于零

循环Jacobi方法

- 经典Jacobi方法中每进行一次相似变换，所需的运算仅为$O(n)$, 而确定旋转平面却需要进行$(n^2 - n)/2$个元素比较

循环Jacobi方法

- 经典Jacobi方法中每进行一次相似变换，所需的运算仅为$O(n)$, 而确定旋转平面却需要进行$(n^2-n)/2$个元素比较
- 所以经典Jacobi方法的大部分时间用在了寻找最佳的旋转平面上

循环Jacobi方法

- 经典Jacobi方法中每进行一次相似变换，所需的运算仅为$O(n)$, 而确定旋转平面却需要进行$(n^2-n)/2$个元素比较
- 所以经典Jacobi方法的大部分时间用在了寻找最佳的旋转平面上
- 一种变通方法：直接按某种预定顺序对每个非对角元消去一次，这就是所谓的循环Jacobi方法

自然的遍历顺序

- 在循环Jacobi方法中，最自然的遍历非对角元的顺序是

$$(1,2),\dots,(1,n);(2,3),\dots,(2,n);\dots,(n-1,n)$$

自然的遍历顺序

- 在循环Jacobi方法中，最自然的遍历非对角元的顺序是

$$(1,2),\dots,(1,n);(2,3),\dots,(2,n);\dots,(n-1,n)$$

- 这种方法也可以证明是渐近平方收敛的，但它相比于经典Jacobi方法要快很多

变形：过关Jacobi方法

- 在实际计算中用得最多的是循环Jacobi方法的一种变体，即过关Jacobi方法

变形：过关Jacobi方法

- 在实际计算中用得最多的是循环Jacobi方法的一种变体，即过关Jacobi方法
- 在该变体中，先确定一个正数（称为关值），在一次扫描中只对那些绝对值超过关值的非对角元所在的平面进行Jacobi变换

变形：过关Jacobi方法

- 在实际计算中用得最多的是循环Jacobi方法的一种变体，即过关Jacobi方法
- 在该变体中，先确定一个正数（称为关值），在一次扫描中只对那些绝对值超过关值的非对角元所在的平面进行Jacobi变换
- 如此反复扫描，当所有的非对角元的绝对值都不超过关值时，减小关值，再进行类似扫描

变形：过关Jacobi方法

- 在实际计算中用得最多的是循环Jacobi方法的一种变体，即过关Jacobi方法
- 在该变体中，先确定一个正数（称为关值），在一次扫描中只对那些绝对值超过关值的非对角元所在的平面进行Jacobi变换
- 如此反复扫描，当所有的非对角元的绝对值都不超过关值时，减小关值，再进行类似扫描
- 直至关值充分小时结束

关值的选取

- 常用的关值是按如下方式选取的：

$$\delta_0 = E(A), \delta_k = \frac{\delta_{k-1}}{\sigma}$$

其中$\sigma \geqslant n$是一个固定的正数

关值的选取

- 常用的关值是按如下方式选取的：

$$\delta_0 = E(A), \delta_k = \frac{\delta_{k-1}}{\sigma}$$

其中$\sigma \geqslant n$是一个固定的正数

- 可以证明：针对如此选取的关值，过关Jacobi方法是收敛的

Jacobi方法中特征向量的计算

- 经过k次变换后迭代停止，则我们有

$$A_k = Q_k^T A Q_k$$

其中$Q_k = J_1 J_2 \cdots J_k$

Jacobi方法中特征向量的计算

- 经过k次变换后迭代停止，则我们有
$$A_k = Q_k^T A Q_k$$
其中$Q_k = J_1 J_2 \cdots J_k$
- 由于A_k的非对角元已非常小，那么A的对角元就是特征值的近似，Q_k的列向量就是A的特征向量近似

二分法

- 本节介绍求一个实对称三对角阵T的任意指定特征值的二分法

二分法

- 本节介绍求一个实对称三对角阵T的任意指定特征值的二分法
- 设T的对角元为α_i, $i = 1, \ldots, n$; 次对角元为β_i, $i = 2, \ldots, n$

二分法

- 本节介绍求一个实对称三对角阵T的任意指定特征值的二分法
- 设T的对角元为α_i, $i = 1, \ldots, n$; 次对角元为β_i, $i = 2, \ldots, n$
- 假定次对角元非零

顺序主子式

- 记$p_i(\lambda)$表示$T-\lambda I$的i阶顺序主子式

顺序主子式

- 记$p_i(\lambda)$表示$T-\lambda I$的i阶顺序主子式
- 则$p_i(\lambda)$满足下面的三项递推公式：

$$
\begin{aligned}
&p_0(\lambda)=1, \quad p_1(\lambda)=\alpha_1-\lambda \\
&p_i(\lambda)=(\alpha_i-\lambda)p_{i-1}(\lambda)-\beta_i^2 p_{i-2}(\lambda) \\
&\qquad i=2,\ldots,n
\end{aligned}
$$

顺序主子式

- 记$p_i(\lambda)$表示$T-\lambda I$的i阶顺序主子式
- 则$p_i(\lambda)$满足下面的三项递推公式：

$$p_0(\lambda)=1,\quad p_1(\lambda)=\alpha_1-\lambda$$
$$p_i(\lambda)=(\alpha_i-\lambda)p_{i-1}(\lambda)-\beta_i^2 p_{i-2}(\lambda)$$
$$i=2,\ldots,n$$

- 由于T实对称，所以$p_i(\lambda)$的根都是实的

例

- 设 $T = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$, 则

$$
\begin{aligned}
p_0(\lambda) &= 1 \\
p_1(\lambda) &= 1 - \lambda \\
p_2(\lambda) &= (1-\lambda)^2 - 1 \\
p_3(\lambda) &= (1-\lambda)^3 - 2(1-\lambda)
\end{aligned}
$$

$p_i(\lambda)$的性质

定理

1. 存在正数M, 使得当$\lambda > M$时，$p_i(-\lambda) > 0$, 而$p_i(\lambda)$的符号为$(-1)^i$

$p_i(\lambda)$的性质

定理

1. 存在正数M，使得当$\lambda > M$时，$p_i(-\lambda) > 0$，而$p_i(\lambda)$的符号为$(-1)^i$
2. 相邻两个多项式没有公共根

$p_i(\lambda)$的性质

定理

1. 存在正数M, 使得当$\lambda > M$时，$p_i(-\lambda) > 0$, 而$p_i(\lambda)$的符号为$(-1)^i$
2. 相邻两个多项式没有公共根
3. 若$p_i(\mu) = 0$, 则$p_{i-1}(\mu)p_{i+1}(\mu) < 0$

$p_i(\lambda)$的性质

定理

1. 存在正数M，使得当$\lambda > M$时，$p_i(-\lambda) > 0$，而$p_i(\lambda)$的符号为$(-1)^i$
2. 相邻两个多项式没有公共根
3. 若$p_i(\mu) = 0$，则$p_{i-1}(\mu)p_{i+1}(\mu) < 0$
4. $p_i(\lambda)$的根全是单重的，并且$p_i(\lambda)$的根严格分离$p_{i+1}(\lambda)$的根

性质1,2的证明

- 根据$p_i(\lambda)$的定义，其首项为$(-1)^i\lambda^i$, 所以性质1成立

性质1,2的证明

- 根据$p_i(\lambda)$的定义，其首项为$(-1)^i\lambda^i$, 所以性质1成立
- 性质2的证明：采用反证法。假设存在某个i, $p_{i-1}(\lambda)$与$p_i(\lambda)$有公共根μ, 那么由三项递推公式

$$0 = p_i(\mu) = (\alpha_i - \mu)p_{i-1}(\mu) - \beta_i^2 p_{i-2}(\mu)$$

而$\beta_i \neq 0$, 所以有$p_{i-2}(\mu) = 0$, 以此类推可得$p_0(\mu) = 0$, 矛盾

性质3的证明

- 设$p_i(\mu) = 0$

性质3的证明

- 设$p_i(\mu) = 0$
- 代入三项递推关系，可得

$$p_{i+1}(\mu) = -\beta_i^2 p_{i-1}(\mu)$$

性质3的证明

- 设$p_i(\mu) = 0$
- 代入三项递推关系，可得

$$p_{i+1}(\mu) = -\beta_i^2 p_{i-1}(\mu)$$

- 由性质2，μ不是$p_{i+1}(\lambda)$和$p_{i-1}(\lambda)$的根，从而可知$p_{i-1}(\mu)p_{i+1}(\mu) < 0$

性质4的证明

采用数学归纳法

- 当$i=1$时$p_1(\lambda)=\alpha_1-\lambda$, 即$\alpha_1$是$p_1(\lambda)$的单根

性质4的证明

采用数学归纳法

- 当$i=1$时$p_1(\lambda)=\alpha_1-\lambda$, 即$\alpha_1$是$p_1(\lambda)$的单根
- 当$i=2$时，由于$p_2(\alpha_1)=-\beta_2^2<0$, 而且根据性质1, 当$\lambda$充分大时有$p_2(\pm\lambda)>0$, 因此在$(-\infty,\alpha_1)$和$(\alpha_1,+\infty)$之内各有$p_2(\lambda)$的一个根，而且$\alpha_1$严格分隔这两个根

性质4的证明

采用数学归纳法

- 当$i=1$时$p_1(\lambda)=\alpha_1-\lambda$, 即$\alpha_1$是$p_1(\lambda)$的单根
- 当$i=2$时，由于$p_2(\alpha_1)=-\beta_2^2<0$, 而且根据性质1, 当$\lambda$充分大时有$p_2(\pm\lambda)>0$, 因此在$(-\infty,\alpha_1)$和$(\alpha_1,+\infty)$之内各有$p_2(\lambda)$的一个根，而且$\alpha_1$严格分隔这两个根
- 假设性质在$i=k$时成立，即$p_{k-1}(\lambda)$和$p_k(\lambda)$的根都是单根，并且$p_{k-1}(\lambda)$的根严格分隔$p_k(\lambda)$的根

- 设$p_{k-1}(\lambda)$和$p_k(\lambda)$的根分别为

$$\nu_1 < \nu_2 < \cdots < \nu_{k-1}, \mu_1 < \mu_2 < \cdots < \mu_k$$

则由归纳假设可知

$$\mu_1 < \nu_1 < \mu_2 < \nu_2 < \cdots < \nu_{k-1} < \mu_k$$

- 设$p_{k-1}(\lambda)$和$p_k(\lambda)$的根分别为

$$\nu_1 < \nu_2 < \cdots < \nu_{k-1}, \mu_1 < \mu_2 < \cdots < \mu_k$$

则由归纳假设可知

$$\mu_1 < \nu_1 < \mu_2 < \nu_2 < \cdots < \nu_{k-1} < \mu_k$$

- 应用三项递推公式可有

$$p_{k+1}(\mu_j) = -\beta_{k+1}^2 p_{k-1}(\mu_j), j = 1, \ldots, k$$

- 根据性质1和$p_{k-1}(\nu_j) = 0\ (1 \leqslant j \leqslant k-1)$可知$(-1)^{j-1} p_{k-1}(\mu_j) > 0,\ j = 1, 2, \ldots, k$

- 根据性质1和$p_{k-1}(\nu_j)=0\ (1\leqslant j\leqslant k-1)$可知$(-1)^{j-1}p_{k-1}(\mu_j)>0,\ j=1,2,\dots,k$
- 于是$(-1)^j p_{k+1}(\mu_j)>0,\ j=1,2,\dots,k$

- 根据性质1和$p_{k-1}(\nu_j)=0\ (1\leqslant j\leqslant k-1)$可知$(-1)^{j-1}p_{k-1}(\mu_j)>0,\ j=1,2,\dots,k$
- 于是$(-1)^j p_{k+1}(\mu_j)>0,\ j=1,2,\dots,k$
- 再注意到对充分大的正数μ有$p_{k+1}(-\mu)>0$, $(-1)^{k+1}p_{k+1}(\mu)>0$, 所以在区间$(-\infty,\mu_1)$, $(\mu_1,\mu_2),\dots,(\mu_{k-1},\mu_k)$, $(\mu_k,+\infty)$内都有$p_{k+1}(\lambda)$的根

- 根据性质1和$p_{k-1}(\nu_j) = 0$ $(1 \leqslant j \leqslant k-1)$可知$(-1)^{j-1}p_{k-1}(\mu_j) > 0$, $j = 1, 2, \ldots, k$
- 于是$(-1)^j p_{k+1}(\mu_j) > 0$, $j = 1, 2, \ldots, k$
- 再注意到对充分大的正数μ有$p_{k+1}(-\mu) > 0$, $(-1)^{k+1}p_{k+1}(\mu) > 0$, 所以在区间$(-\infty, \mu_1)$, (μ_1, μ_2), $\ldots$, (μ_{k-1}, μ_k), $(\mu_k, +\infty)$内都有$p_{k+1}(\lambda)$的根
- 这就完成了证明

变号数

- 根据性质可知，不可约三对角对称多项式的特征值都是单重实数

变号数

- 根据性质可知，不可约三对角对称多项式的特征值都是单重实数
- 对任意给定的实数μ, 定义$s_k(\mu)$表示数列$p_0(\mu), p_1(\mu), \ldots, p_k(\mu)$的变号数

变号数

- 根据性质可知，不可约三对角对称多项式的特征值都是单重实数
- 对任意给定的实数μ, 定义$s_k(\mu)$表示数列$p_0(\mu), p_1(\mu), \ldots, p_k(\mu)$的变号数
- 规定：如果$p_i(\mu) = 0$, 则$p_i(\mu)$与$p_{i-1}(\mu)$同号（注意，$p_{i-1}(\mu)$不可能也为零）

变号数

- 根据性质可知，不可约三对角对称多项式的特征值都是单重实数
- 对任意给定的实数μ, 定义$s_k(\mu)$表示数列$p_0(\mu), p_1(\mu), \ldots, p_k(\mu)$的变号数
- 规定：如果$p_i(\mu) = 0$, 则$p_i(\mu)$与$p_{i-1}(\mu)$同号（注意，$p_{i-1}(\mu)$不可能也为零）
- 例：对前面的三阶矩阵例，$\mu = 1$, 则我们有

$$p_0(1) = 1, p_1(1) = 0, p_2(1) = -1, p_3(1) = 0$$

所以变号数$s_3(1) = 1$

变号数与根的个数

定理

若T为不可约对称三对角矩阵，则$s_k(\mu)$ $(1 \leqslant k \leqslant n)$恰好是$p_k(\lambda)$在区间$(-\infty, \mu)$内根的个数

推论

若T为不可约对称三对角矩阵，则$s_n(\mu)$ 恰好是T在区间$(-\infty, \mu)$内特征值的个数

定理的证明

采用数学归纳法

- 当$k = 1$时定理显然成立

定理的证明

采用数学归纳法

- 当$k=1$时定理显然成立
- 假设当$k=\ell$时定理成立

定理的证明

采用数学归纳法

- 当$k=1$时定理显然成立
- 假设当$k=\ell$时定理成立
- 设$p_\ell(\lambda)$和$p_{\ell+1}(\lambda)$的根分别为

$$\mu_1<\mu_2<\cdots<\mu_\ell, \lambda_1<\lambda_2<\cdots<\lambda_{\ell+1}$$

则根据性质4,

$$\lambda_1<\mu_1<\lambda_2<\mu_2<\cdots<\mu_\ell<\lambda_{\ell+1}$$

- 设$s_\ell(\mu) = m$, 则由归纳假设可知$\mu_m < \mu \leqslant \mu_{m+1}$

- 设$s_\ell(\mu)=m$, 则由归纳假设可知$\mu_m<\mu\leqslant\mu_{m+1}$
- 注意到$\mu_m<\lambda_{m+1}<\mu_{m+1}$, 从而$\mu$所在的位置有两种可能性:

$$\lambda_m<\mu_m<\mu\leqslant\lambda_{m+1} \text{ 或 } \lambda_{m+1}<\mu\leqslant\mu_{m+1}$$

- 设$s_\ell(\mu) = m$, 则由归纳假设可知$\mu_m < \mu \leqslant \mu_{m+1}$
- 注意到$\mu_m < \lambda_{m+1} < \mu_{m+1}$, 从而$\mu$所在的位置有两种可能性：

$$\lambda_m < \mu_m < \mu \leqslant \lambda_{m+1} \text{ 或 } \lambda_{m+1} < \mu \leqslant \mu_{m+1}$$

- 注意到

$$p_\ell(\mu) = \prod_{i=1}^{\ell}(\mu_i - \mu), p_{\ell+1} = \prod_{i=1}^{\ell+1}(\lambda_i - \mu)$$

$\lambda_m < \mu_m < \mu \leqslant \lambda_{m+1}$时

- 当$\lambda_m < \mu_m < \mu \leqslant \lambda_{m+1}$时，可知$p_\ell(\mu)$与$p_{\ell+1}(\mu)$同号，即使$\mu = \lambda_{m+1}$按规定此两数也是同号的

$\lambda_m < \mu_m < \mu \leqslant \lambda_{m+1}$时

- 当$\lambda_m < \mu_m < \mu \leqslant \lambda_{m+1}$时，可知$p_\ell(\mu)$与$p_{\ell+1}(\mu)$同号，即使$\mu = \lambda_{m+1}$按规定此两数也是同号的
- 从而$s_{\ell+1}(\mu) = s_\ell(\mu) = m$, 这正好是$p_{\ell+1}(\lambda)$在区间$(-\infty, \mu)$内根的个数

$\lambda_{m+1} < \mu \leqslant \mu_{m+1}$时

当$\lambda_{m+1} < \mu \leqslant \mu_{m+1}$时，我们分两种情况证明$s_{\ell+1}(\mu) = m+1$

- 若$\mu < \mu_{m+1}$，则$p_\ell(\mu)$与$p_{\ell+1}(\mu)$异号，因而$s_{\ell+1}(\mu) = s_\ell(\mu) + 1 = m+1$

$\lambda_{m+1} < \mu \leqslant \mu_{m+1}$时

当$\lambda_{m+1} < \mu \leqslant \mu_{m+1}$时，我们分两种情况证明$s_{\ell+1}(\mu) = m+1$

- 若$\mu < \mu_{m+1}$，则$p_\ell(\mu)$与$p_{\ell+1}(\mu)$异号，因而$s_{\ell+1}(\mu) = s_\ell(\mu) + 1 = m+1$
- 若$\mu = \mu_{m+1}$，则此时有$p_\ell(\mu) = 0$，按约定$p_\ell(\mu)$与$p_{\ell-1}(\mu)$ 同号。由性质3可知，$p_{\ell-1}(\mu)$与$p_{\ell+1}(\mu)$异号，因而$p_{\ell+1}(\mu)$ 与$p_\ell(\mu)$异号，从而$s_{\ell+1}(\mu) = s_\ell(\mu) + 1 = m+1$

$\lambda_{m+1} < \mu \leqslant \mu_{m+1}$时

当$\lambda_{m+1} < \mu \leqslant \mu_{m+1}$时，我们分两种情况证明$s_{\ell+1}(\mu) = m+1$

- 若$\mu < \mu_{m+1}$, 则$p_\ell(\mu)$与$p_{\ell+1}(\mu)$异号，因而$s_{\ell+1}(\mu) = s_\ell(\mu)+1 = m+1$
- 若$\mu = \mu_{m+1}$, 则此时有$p_\ell(\mu) = 0$, 按约定$p_\ell(\mu)$与$p_{\ell-1}(\mu)$ 同号。由性质3可知，$p_{\ell-1}(\mu)$与$p_{\ell+1}(\mu)$异号，因而$p_{\ell+1}(\mu)$ 与$p_\ell(\mu)$异号，从而$s_{\ell+1}(\mu) = s_\ell(\mu)+1 = m+1$
- 根据归纳假设，这就完成了证明

求指定特征值

- 利用定理的推论，我们可以用二分法求T的任何一个指定的特征值

求指定特征值

- 利用定理的推论，我们可以用二分法求T的任何一个指定的特征值
- 设T的特征值为$\lambda_1 < \lambda_2 < \cdots < \lambda_n$

求指定特征值

- 利用定理的推论，我们可以用二分法求T的任何一个指定的特征值
- 设T的特征值为$\lambda_1 < \lambda_2 < \cdots < \lambda_n$
- 则必有$|\lambda_i| \leqslant \rho(T) \leqslant \|T\|_\infty$

求指定特征值

- 利用定理的推论，我们可以用二分法求T的任何一个指定的特征值
- 设T的特征值为$\lambda_1 < \lambda_2 < \cdots < \lambda_n$
- 则必有$|\lambda_i| \leqslant \rho(T) \leqslant \|T\|_\infty$
- 假定我们期望求T的第m个特征值λ_m, 我们先取

$$\ell_0 = -\|T\|_\infty, u_0 = \|T\|_\infty$$

二分法

- 根据设定，λ_m必在区间$[\ell_0, u_0]$内

二分法

- 根据设定，λ_m必在区间$[\ell_0, u_0]$内
- 取$[\ell_0, u_0]$的中点$r_1 = (\ell_0 + u_0)/2$, 计算$s_n(r_1)$

二分法

- 根据设定，λ_m必在区间$[\ell_0, u_0]$内
- 取$[\ell_0, u_0]$的中点$r_1 = (\ell_0 + u_0)/2$，计算$s_n(r_1)$
  - 若$s_n(r_1) \geqslant m$，则$\lambda \in [\ell_0, r_1]$，于是取$\ell_1 = \ell_0$，$u_1 = r_1$

二分法

- 根据设定，λ_m必在区间$[\ell_0, u_0]$内
- 取$[\ell_0, u_0]$的中点$r_1 = (\ell_0 + u_0)/2$, 计算$s_n(r_1)$
  - 若$s_n(r_1) \geqslant m$, 则$\lambda \in [\ell_0, r_1]$, 于是取$\ell_1 = \ell_0$, $u_1 = r_1$
  - 否则$\lambda \in [r_1, u_0]$, 于是取$\ell_1 = r_1$, $u_1 = u_0$

二分法

- 根据设定，λ_m必在区间$[\ell_0, u_0]$内
- 取$[\ell_0, u_0]$的中点$r_1 = (\ell_0 + u_0)/2$，计算$s_n(r_1)$
  - 若$s_n(r_1) \geqslant m$，则$\lambda \in [\ell_0, r_1]$，于是取$\ell_1 = \ell_0$，$u_1 = r_1$
  - 否则$\lambda \in [r_1, u_0]$，于是取$\ell_1 = r_1$，$u_1 = u_0$
  - 如此我们得到一个长度减少一半的区间$[\ell_1, u_1]$仍包含特征值λ_m

二分法

- 根据设定，λ_m必在区间$[\ell_0, u_0]$内
- 取$[\ell_0, u_0]$的中点$r_1 = (\ell_0 + u_0)/2$，计算$s_n(r_1)$
  - 若$s_n(r_1) \geqslant m$，则$\lambda \in [\ell_0, r_1]$，于是取$\ell_1 = \ell_0$，$u_1 = r_1$
  - 否则$\lambda \in [r_1, u_0]$，于是取$\ell_1 = r_1$，$u_1 = u_0$
  - 如此我们得到一个长度减少一半的区间$[\ell_1, u_1]$仍包含特征值λ_m
- 继续上述过程直到区间长度足够小，最后取区间中点作为λ_m的近似值

变号数的有效计算

- 二分法的主要工作是计算$s_n(\mu)$. 这显然不能直接通过计算$p_i(\mu)$的值来实现，因为高阶多项式的计算很容易出现不稳定

变号数的有效计算

- 二分法的主要工作是计算$s_n(\mu)$. 这显然不能直接通过计算$p_i(\mu)$的值来实现，因为高阶多项式的计算很容易出现不稳定
- 定义$q_i(\lambda) = \dfrac{p_i(\lambda)}{p_{i-1}(\lambda)}$, $i = 1, 2, \ldots, n$

变号数的有效计算

- 二分法的主要工作是计算$s_n(\mu)$. 这显然不能直接通过计算$p_i(\mu)$的值来实现，因为高阶多项式的计算很容易出现不稳定
- 定义$q_i(\lambda) = \dfrac{p_i(\lambda)}{p_{i-1}(\lambda)}$, $i = 1, 2, \ldots, n$
- 根据定义可知

$$q_1(\lambda) = p_1(\lambda) = \alpha_1 - \lambda$$

$$q_i(\lambda) = \alpha_i - \lambda - \frac{\beta_i^2}{q_{i-1}(\lambda)}, i = 2, \ldots, n$$

变号数的有效计算

- 二分法的主要工作是计算$s_n(\mu)$. 这显然不能直接通过计算$p_i(\mu)$的值来实现，因为高阶多项式的计算很容易出现不稳定
- 定义$q_i(\lambda) = \dfrac{p_i(\lambda)}{p_{i-1}(\lambda)}$, $i = 1, 2, \ldots, n$
- 根据定义可知

$$q_1(\lambda) = p_1(\lambda) = \alpha_1 - \lambda$$

$$q_i(\lambda) = \alpha_i - \lambda - \frac{\beta_i^2}{q_{i-1}(\lambda)}, i = 2, \ldots, n$$

- $s_n(\mu)$就是数列$q_1(\mu), \ldots, q_n(\mu)$中负数的个数

$q_{i-1}(\lambda) = 0$的处理

- 当$q_{i-1}(\lambda) = 0$时，按规定此时q_{i-1}应按正数对待

$q_{i-1}(\lambda) = 0$的处理

- 当$q_{i-1}(\lambda) = 0$时，按规定此时q_{i-1}应按正数对待
- 在算法中可用很小的正数代替$q_{i-1}(\lambda)$

运算量

- 可以事先把β_i^2算好并贮存起来，因此上述变号数计算的算法只需要$n-1$次除法运算和$2n-1$次加减运算

运算量

- 可以事先把β_i^2算好并贮存起来，因此上述变号数计算的算法只需要$n-1$次除法运算和$2n-1$次加减运算
- 如果计算一个特征值平均需要m次二分法，则用二分法求一个特征值的运算量平均为$3nm$

注解

- 二分法具有较大的灵活性，它既可以求某些指定的特征值，也可以求某个区间内的特征值，而且对各个特征值的精度要求也可以不一样

注解

- 二分法具有较大的灵活性，它既可以求某些指定的特征值，也可以求某个区间内的特征值，而且对各个特征值的精度要求也可以不一样
- 二分法是非常稳定的，而且计算精度和所需计算时间与特征值的分离程度无关

注解

- 二分法具有较大的灵活性，它既可以求某些指定的特征值，也可以求某个区间内的特征值，而且对各个特征值的精度要求也可以不一样
- 二分法是非常稳定的，而且计算精度和所需计算时间与特征值的分离程度无关
- 在算出某个特征值之后，可以应用反幂法求特征向量

奇异值分解的计算

- 对给定的$A \in \mathbb{R}^{m\times n}$ $(m \geqslant n)$，如何计算其奇异值分解，这是一个有重要应用意义的问题

奇异值分解的计算

- 对给定的$A \in \mathbb{R}^{m\times n}$ $(m \geqslant n)$，如何计算其奇异值分解，这是一个有重要应用意义的问题
- 奇异值分解与对称矩阵的谱分解密切相关，从而也相应地有计算奇异值分解的QR方法、Jacobi方法、二分法等

奇异值分解的计算

- 对给定的$A \in \mathbb{R}^{m\times n}$ $(m \geqslant n)$，如何计算其奇异值分解，这是一个有重要应用意义的问题
- 奇异值分解与对称矩阵的谱分解密切相关，从而也相应地有计算奇异值分解的QR方法、Jacobi方法、二分法等
- 本节我们只介绍计算奇异值分解的QR方法，基本想法：隐含地应用对称QR方法于A^TA上，但又希望在整个过程中并不直接计算A^TA

二对角化

- 对应于将A^TA三对角化，这里我们是将A二对角化，即计算两个正交矩阵U, V, 使得

$$U^TAV = \begin{pmatrix} B \\ 0 \end{pmatrix}, \quad B = \begin{pmatrix} \delta_1 & \gamma_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \gamma_{n-1} \\ & & & \delta_n \end{pmatrix}$$

二对角化

- 对应于将A^TA三对角化，这里我们是将A二对角化，即计算两个正交矩阵U, V, 使得

$$U^TAV = \begin{pmatrix} B \\ 0 \end{pmatrix}, \quad B = \begin{pmatrix} \delta_1 & \gamma_1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \gamma_{n-1} \\ & & & \delta_n \end{pmatrix}$$

- 从而有$V^TA^TAV = B^TB$是一个对称三对角阵，这就相当于把A^TA三对角化

二对角化的实现

- 二对角分解可以利用Householder变换实现

二对角化的实现

- 二对角分解可以利用Householder变换实现
- 首先确定一个m阶Householder变换P_1，把$P_1^T A$的第一列除首元外化为零

二对角化的实现

- 二对角分解可以利用Householder变换实现
- 首先确定一个m阶Householder变换P_1，把$P_1^T A$的第一列除首元外化为零
- 然后确定$n-1$阶Householder变换H_1使得

$$P_1^T A \begin{pmatrix} 1 & 0 \\ 0 & H_1 \end{pmatrix}$$

的第一行除前两个元素外其余元素约化为零

二对角化的实现

- 二对角分解可以利用Householder变换实现
- 首先确定一个m阶Householder变换P_1，把$P_1^T A$的第一列除首元外化为零
- 然后确定$n-1$阶Householder变换H_1使得
$$P_1^T A \begin{pmatrix} 1 & 0 \\ 0 & H_1 \end{pmatrix}$$
的第一行除前两个元素外其余元素约化为零
- 如此继续下去就可以完成二对角分解

SVD迭代：位移的选取

- 下面对三对角阵$T = B^T B$进行带位移的隐式QR迭代。同样这里的关键是不明确把T计算出来

SVD迭代：位移的选取

- 下面对三对角阵$T = B^T B$进行带位移的隐式QR迭代。同样这里的关键是不明确把T计算出来
- 位移的选取：针对T的右下角2阶矩阵

$$\begin{pmatrix} \delta_{n-1}^2 + \gamma_{n-1}^2 & \delta_{n-1}\gamma_{n-1} \\ \delta_{n-1}\gamma_{n-1} & \delta_n^2 + \gamma_{n-1}^2 \end{pmatrix}$$

我们取位移μ为这个矩阵的两个特征值中距离$\delta_n^2 + \gamma_{n-1}^2$近的那个（即Wilkinson位移）

SVD迭代：c, s的确定

- 接下来，确定Givens变换$G_1 = G(1,2,\theta)$满足

$$\begin{pmatrix} c_1 & s_1 \\ -s_1 & c_1 \end{pmatrix}^T \begin{pmatrix} \delta_1^2 - \mu \\ \delta_1\gamma_1 \end{pmatrix} = \begin{pmatrix} * \\ 0 \end{pmatrix}$$

这里$T-\mu I$的第一列是$(\delta_1^2-\mu, \delta_1\gamma_1, 0, \dots, 0)^T$

SVD迭代：正交相似变换

- 最后一步就是确定正交矩阵Q使得$Q^T(G_1^T TG_1)Q$是对称三对角阵，且$Qe_1 = e_1$

SVD迭代：正交相似变换

对称特征值问题的计算方法

邓建松

基本性质

对称QR方法

三对角化

隐式对称QR迭代

Jacobi方法

经典Jacobi方法

循环Jacobi方法及其变形

二分法

奇异值分解的计算

二对角化

SVD迭代

SVD算法

- 最后一步就是确定正交矩阵Q使得$Q^T(G_1^T TG_1)Q$是对称三对角阵，且$Qe_1 = e_1$
- 为了避免T的计算，只需计算正交矩阵P和Q使得$P^T(BG_1)Q$是二对角阵，且$Qe_1 = e_1$

SVD迭代：正交相似变换

- 最后一步就是确定正交矩阵Q使得$Q^T(G_1^T TG_1)Q$是对称三对角阵，且$Qe_1 = e_1$
- 为了避免T的计算，只需计算正交矩阵P和Q使得$P^T(BG_1)Q$是二对角阵，且$Qe_1 = e_1$
- 可以利用Givens变换实现这一点

不可约的要求

- 隐式QR迭代的前提是$T = B^T B$是不可约的

不可约的要求

- 隐式QR迭代的前提是$T = B^T B$是不可约的
- T的次对角元为$\delta_j\gamma_j$, 因此T不可约的充要条件是$\delta_j\gamma_j \neq 0$

不可约的要求

- 隐式QR迭代的前提是$T = B^T B$是不可约的
- T的次对角元为$\delta_j\gamma_j$, 因此T不可约的充要条件是$\delta_j\gamma_j \neq 0$
- 当某个$\gamma_j = 0$时，这时可以把问题分解为两个低阶问题处理

不可约的要求

- 隐式QR迭代的前提是$T = B^T B$是不可约的
- T的次对角元为$\delta_j\gamma_j$, 因此T不可约的充要条件是$\delta_j\gamma_j \neq 0$
- 当某个$\gamma_j = 0$时，这时可以把问题分解为两个低阶问题处理
- 当某个$\delta_j = 0$，而$\gamma_j \neq 0$时，我们可以通过适当的Givens变换把B的第j行元素都变零，而保持其二对角形式不变；从而实现降阶

SVD算法

- 在实际计算时，当δ_j或γ_j很小时，我们就可以把B分解为两个低阶的二对角阵

SVD算法

- 在实际计算时，当δ_j或γ_j很小时，我们就可以把B分解为两个低阶的二对角阵
- 通常使用的准则是：如果

$$|\delta_j| \leqslant \varepsilon \|B\|_\infty \text{ 或 } |\gamma_j| \leqslant (|\delta_j| + |\delta_{j+1}|)$$

就将δ_j或γ_j视为零，其中ε是一个略大于机器精度的正数

SVD算法

- 在实际计算时，当δ_j或γ_j很小时，我们就可以把B分解为两个低阶的二对角阵
- 通常使用的准则是：如果

$$|\delta_j| \leqslant \varepsilon \|B\|_\infty \text{ 或 } |\gamma_j| \leqslant (|\delta_j| + |\delta_{j+1}|)$$

  就将δ_j或γ_j视为零，其中ε是一个略大于机器精度的正数
- 将这一准则与前述算法结合，就得到了SVD算法

SVD算法

- 在实际计算时，当δ_j或γ_j很小时，我们就可以把B分解为两个低阶的二对角阵
- 通常使用的准则是：如果

$$|\delta_j| \leqslant \varepsilon \|B\|_\infty \text{ 或 } |\gamma_j| \leqslant (|\delta_j| + |\delta_{j+1}|)$$

就将δ_j或γ_j视为零，其中ε是一个略大于机器精度的正数

- 将这一准则与前述算法给合，就得到了SVD算法
- 算法的渐近收敛速度是三次的